RDACSEEE'17

27th Nov - 29th Nov 2017

National conference on

Recent Development & Advancement In Computer Science, Electrical & Electronics Engineering

Conference Proceeding

ARYAN
INSTITUTE OF ENGINEERING & TECHNOLOGY

AIET Bhubaneswar

Aryan Institute of Engineering and Technology Bhubaneswar

Organized By-

DEPARTMENT OF
COMPUTER SCIENCE & ELECTRICAL ENGINEERING
AIET, Bhubaneswar, 752050

Recent Development and Advancement in computer Science,
Electrical and Electronics Engineering

27th Nov. – 29th Nov. 2017

CONFERENCE PROCEEDING

Organized by

Department of Computer Science and Electrical Engineering
Aryan Institute of Engineering and Technology
Bhubaneswar – 752050

List of Sponsors

Citicon Engineers Pvt. Ltd

Aryan Infra Projects

Citiz Essential Services

Oltron Technology Pvt Ltd

ABOUT THE CONFERENCE

Recent Advances in Electrical, Electronics and Computer Engineering Conference aims to bring together leading academic scientists, researchers and research scholars to exchange and share their experiences and research results on all aspects of Recent Development and Advancement in computer Science, Electrical and Electronics Engineering (RDACSEEE-17) Conference. It also provides a premier interdisciplinary platform for researchers, practitioners, and educators to present and discuss the most recent innovations, trends, and concerns as well as practical challenges encountered and solutions adopted in the fields of Recent Development and Advancement in computer Science, Electrical and Electronics Engineering (RDACSEEE-17) Conference.

ABOUT THE DEPARTMENT

The Department of Computer Science & Engineering (CSE) and Electrical Engineering (EE) is a place where future leaders learn to design technology that solves problems and improves lives. The department is recognized as one of the top programs in India. We have passionate faculty, exciting research, great job prospects for our students, and a supportive community. The students are encouraged and motivated to take up challenging projects. Summer training, industrial visit and projects are carefully planned for the students to remain updated with the technology trend. Seminars and short courses are regularly organized to update the students with the latest in the education and industry trends.

ABOUT THE INSTITUTE

Established in the year 2009 , Aryan Institute of Engineering and Technology(AIET) is one of the premier engineering colleges in the self-financing category of Engineering education in eastern India. It is situated at temple city Bhubaneswar, Odisha and is a constituent member of Aryan Educational Trust. This reputed engineering college is accredited by NAAC, UGC and is affiliated to BPUT, Odisha. AIET aims to create disciplined and trained young citizens in the field of engineering and technology for holistic and national growth.

The college is committed towards enabling secure employment for its students at the end of their four year engineering degree course. (The NAAC accreditation in the year 2018 vouches for the college’s determination and dedication for a sustainable learning environment). The academic fraternity of AIET is a unique blend of faculty with industry and academic experience. This group of facilitators work with a purpose of importing quality education in the field of technical education to the aspiring students. Affordable fee structure along with approachable location in the smart city of Bhubaneswar, makes it a preferred destination for aspiring students and parents.

The Institute works with a mission to expand human knowledge beneficial to society through inclusive education, integrated with application and research. It strives to investigate on the challenging basic problems faced by Science and Technology in an Inter disciplinary atmosphere and urges to educate its students to reach their destination, making them come up qualitatively and creatively and to contribute fruitfully. This is not only its objective but also the ultimate path to move on with truth and brilliance towards success.

Organizing Committee Members

PATRON:

Dr. Madhumita Parida
Chairperson
Aryan Institute of Engineering and Technology

Director
Prof. Sasmita Parida

Convener
Asst. Prof. Prakash Kumar Dehury

Organizing Secy.
Asst. Prof. Ipsita Pahi

Jt. Organizing Secy.
Prof. A. K. Sahoo

Treasurer
Asst. Prof. Ajit Kumar Panda

Jt. Treasurer
Asst. Prof. K. K. Baral

NATIONAL ADVISORY COMMITTEE

Dr. Jaydev Mallick
Professor
Department of Computer Science Engineering
Indian Institute of Technology,
Kanpur

Dr. Dharmendra Kumar
Professor
Department of Electrical and Tele-Communication Engineering
Indian Institute of Technology, Patna

Dr. Bharat Bhusan Sharma
Professor
Department of Computer Science Engineering
Indian Institute of Technology,
Roopar

Dr. Ekta Aggarwal
Asso. Prof.
Department of Computer Science Engineering
National Institute of Technology,
Trichy

Dr. Aditya Kumar
Asso. Prof.
Department of Electrical Engineering
National Institute of Technology,
Uttarakhand

Dr. Adityanath Bhatt
Asso. Prof.
Department of Computer Science Engineering
National Institute of Technology,
Jaypur

Dr. Anuj Kumar
Asso. Prof.
Department of Electronics and Tele-Communication Engineering
National Institute of Technology, Uttarakhand

Dr. Harish Chandra
Asst. Prof.
Department of Electrical Engineering
National Institute of Technology, Warengle

Dr. Nutu Goswami
Asst. Prof.
Department of Electronics and Communication Engineering,
National Institute of Technology, Durgapur

Dr. Nilam Shukla
Asst. Prof.
Department of Computer Science Engineering
National Institute of Technology,
Raipur

LOCAL COMMITTEE MEMBERS

Prof. Ajaya Kumar Swain
Department of Electrical Engineering

Prof. Sanjay Kumar Padhi
Department of Computer Science Engineering

Prof. Srinivas
Department of Electrical Engineering

Prof. Laxmi
Department of Computer Science Engineering

Asst. Prof. Jhalaka Hota
Department of Computer Science Engineering

Asst. Prof. Somnath Mishra
Department of Electrical Engineering

Asst. Prof. Ajit Kumar Panda
Department of Electrical Engineering

Asst. Prof. Dillip Kumar Nayak
Department of Electrical and Electronics Engineering

Asst. Prof. Kumar Dasarathi Dalei
Department of Electrical and Electronics Engineering

Asst. Prof. T. R. Baitharu
Department of Computer Science Engineering

Asst. Prof. P. C. Satpathy
Department of Electronics and Communication Engineering

Asst. Prof. P. K. Rautray
Department of Computer Science Engineering

Asst. Prof. Ajanta Priyadarshinee
Department of Electrical Engineering

Asst. Prof. Aravinda Mahapatra
Department of Electrical and Electronics Engineering

Asst. Prof. Sanjeev Kumar Mishra
Department of Electrical and Electronics Engineering

Asst. Prof. S. K. Tripathy
Department of Electronics and Communications Engineering

Asst. Prof. K. P. Patro
Department of Computer Science Engineering

Asst. Prof. Ankita Panda
Department of Electronics and Communication Engineering

Conference Committee Management

1. Reception Management

- Rasmi Rekha Mahato
- Rajkumari Lopamudra
- Bidyabati Samal
- Urmila Sahoo
- Smita Digal
- Sunita Priyadarsini

2. Transit/Accommodation Management

- Krishna Naidu
- Biswanath Majhi
- Kishore Chandra Pradhan
- Duga Prasad Padhy
- Jyoti Prakash Khunti
- Deepen Kumar Jena
- Manoranjan Mahunta

3. Seminar Hall Management

- Prasannjeet Pattanaik
- Pravat Ranjan Mishra
- Priti Ranjan Jena
- Samrat Kharabela Senapati
- Bidyadhar Jena
- Chinmaya Mohapatra

4. Catering Management

- Pradyumna Badajena
- Pravat Kumar Sahoo
- Hara Prasad Mishra
- Asutosh Bal

5. Printing/Stationary Management

- Sumit Ghosh
- Akhilesh Singh
- Vikas Meher
- Vishal Gupta

6. Design Team

- Sridhar Jena
- Mahesh Nayak
- Nitesh Samal

7. Anchoring In Inauguration Ceremony

- Rudra Prasad Nanda
- Nilimashree Niharika

Conference Sub-Committee Management

- Dillip Kumar Pradhan
- Junaid Mohammad
- Chinmaya Mohapatra
- Naresh Sharma
- Pradeep Ghadei
- Dinesh Shah
- Narendra Mallick

ORAL		Session 1		Session 2		Session 3	Poster(18:00 - 19:00)	
27-Nov	Room 1	Classification and Regression Trees	Lunch	Classification and Regression Trees	Tea Break	Sensor Applications and Deployments	27-Nov	Embedded Hardware
	Room 2	Mobile and Wireless Security		Mobile and Wireless Security		Embedded Hardware		Network Reliability
	Room 3	Sensor Applications and Deployments		Sensor Applications and Deployments		Human Centered Computing		Sensor Applications and Deployments
	Room 4	Human Centered Computing		Human Centered Computing		Mobile Networks		Human Centered Computing
28-Nov	Room 1	Machine Learning		Machine Learning		Machine Learning		Machine Learning
	Room 2	Network Management		Network Management		Security Protocols		Security Protocols
	Room 3	Distributed Architectures		Distributed Architectures		World Wide Web	28-Nov	Network Management
	Room 4	Embedded Hardware		Embedded Hardware		Sensor Networks		Sensor Networks
29-Nov	Room 1	Machine Learning						Classification and Regression Trees
	Room 2	Sensor Networks						Mobile and Wireless Security
	Room 3	Supervised Learning by Classification						Supervised Learning by Classification

Prakash Kumar Dehury

CHAIRPERSON'S MESSAGE

On behalf of the Organizing Committee, it is my great pleasure to welcome you to National Conference on Recent Development and Advancement in computer Science, Electrical and Electronics Engineering (RDACSEEE-2017). In our endeavour to raise the standards of discourse, we continue to remain aware in order to meet with the changing needs of our stakeholders. The idea to host the RDACSEEE-2017 is to bring together Researchers, Scientists, Engineers, Scholars and Students in the areas of Computer Science and Electrical Engineering. The RDACSEEE-2017 Conference will foster discussions and hopes to inspire participants from a wide array of themes to initiate Research and Development and collaborations within and across disciplines for the advancement of Technology. The conference aims to bring together innovative academic experts, researchers and Faculty in Engineering and Management to provide a platform to acquaint and share new ideas. The various thematic sessions will showcase important technological advances and highlight their significance and challenges in a world of fast changes. I welcome all of you to attend the plenary sessions and invite you to interact with the conference participants. The Conference Committees will make any possible effort to make sure that your participation will be technically rewarding and a pleasurable experience.

I am looking forward to meeting you in during RDACSEEE-17 and to sharing a most pleasant, interesting and fruitful conference.

Dr. Madhumita Parida

With regards,
Dr. Madhumita Parida
Chairperson

Aryan Institute of Engineering & Technology
Arya Vihar, Bhubaneswar, Odisha

DIRECTOR’S MESSAGE

It is a great pleasure and an honor to extend to you a warm invitation to attend the National Conference on ‘Recent Development and Advancement in computer Science, Electrical and Electronics Engineering (RDACSEEE’17)’ to be held on 27th – 29th November 2017, at Aryan Institute of Engineering and Technology, Bhubaneswar. The theme of Emerging Trends in Computer Science and Electrical Engineering will underpin the need for participation in forums for collaborative Research and cooperation of individuals from a wide range of professional backgrounds. The Conference will provide a wonderful forum for you to refresh your knowledge in the technical field in Computer Science and Electrical Engineering. The Conference will strive to offer plenty of networking opportunities, providing you with the opportunity to meet and interact with the scientists and researchers, friends as well as sponsors and exhibitors.

I hope you will join us for a symphony of the outstanding conference, and spare a little time to enjoy the spectacular and unique beauty of Bhubaneswar city.

Sasmita Parida

With regards,
Prof. Sasmita Parida
Director
Aryan Institute of Engineering & Technology
Arya Vihar, Bhubaneswar, Odisha

PRINCIPAL'S MESSAGE

I am pleased to welcome you to the National Conference on "Recent Development and Advancement in computer Science, Electrical and Electronics Engineering" (RDACSEEE-2017) to be held on 27th - 29th November, 2017.

The intent of this conference is not only to discuss lively and emerging issues of a particular domain but also dissemination of the awareness among other learned people. Over the years, dramatic improvements have been made in the field of Engineering, and Technology. I hope RDACSEEE-2017 will become the most useful National Conference dedicated to bring out latest trends in Engineering, and Technology.

In order to provide an outstanding technical level for the presentations at the conference, we have invited distinguished experts to participate in the Technical Programmes. We will have technical sessions, plenary sessions by keynote speakers during three days of conference including the awards presentation during the valedictory session on the last day of the conference.

I hope RDACSEEE-2017 will make you aware of state-of-the art systems and provide a platform to discuss various emerging technologies in Computer Science and Electrical Engineering.

With regards,
Prof. (Dr.) S. S. Khuntia
Principal
Aryan Institute of Engineering & Technology
Arya Vihar, Bhubaneswar, Odisha

CONVENER'S MESSAGE

National Conference on " Recent Development and Advancement in computer Science, Electrical and Electronics Engineering" (RDACSEEE 2017) is a prestigious event jointly organized by Computer Science and Electrical Engineering Department with a motivation to share a progress in recent technologies. The objective of RDACSEEE 2017 is to present the latest research and results of scientists (preferred under graduate and post graduate students, research scholars, post-doc scientists, academicians and working professionals) related to the subjects of Computer Science and Electrical Engineering. The conference will provide with paper presentations and research paper presentation by prominent speakers who will focus on related state-of-the-art technologies in the areas of the conference.

I wish all the success to the conference RDACSEEE-2017.

Prakash Kumar Dehury

With regards,
Mr. Prakash Kumar Dehury

Asst. Prof. Computer Science Engineering
Aryan Institute of Engineering & Technology
Arya Vihar, Bhubaneswar, India

Contents

A Triple-Band Microstrip Antenna with a Monopole Impedance Converter for WLAN and 5G Applications

Bhagaban Sri Ramakrishna, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ramakrishna12@outlook.com*

Supriya Nayak, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, supriyanayak443@gmail.com*

Asheerbad Pradhan, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, asheerbadpradhan46@gmail.com*

Smruti Samantray, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, smrutisamantray23@hotmail.com*

Abstract

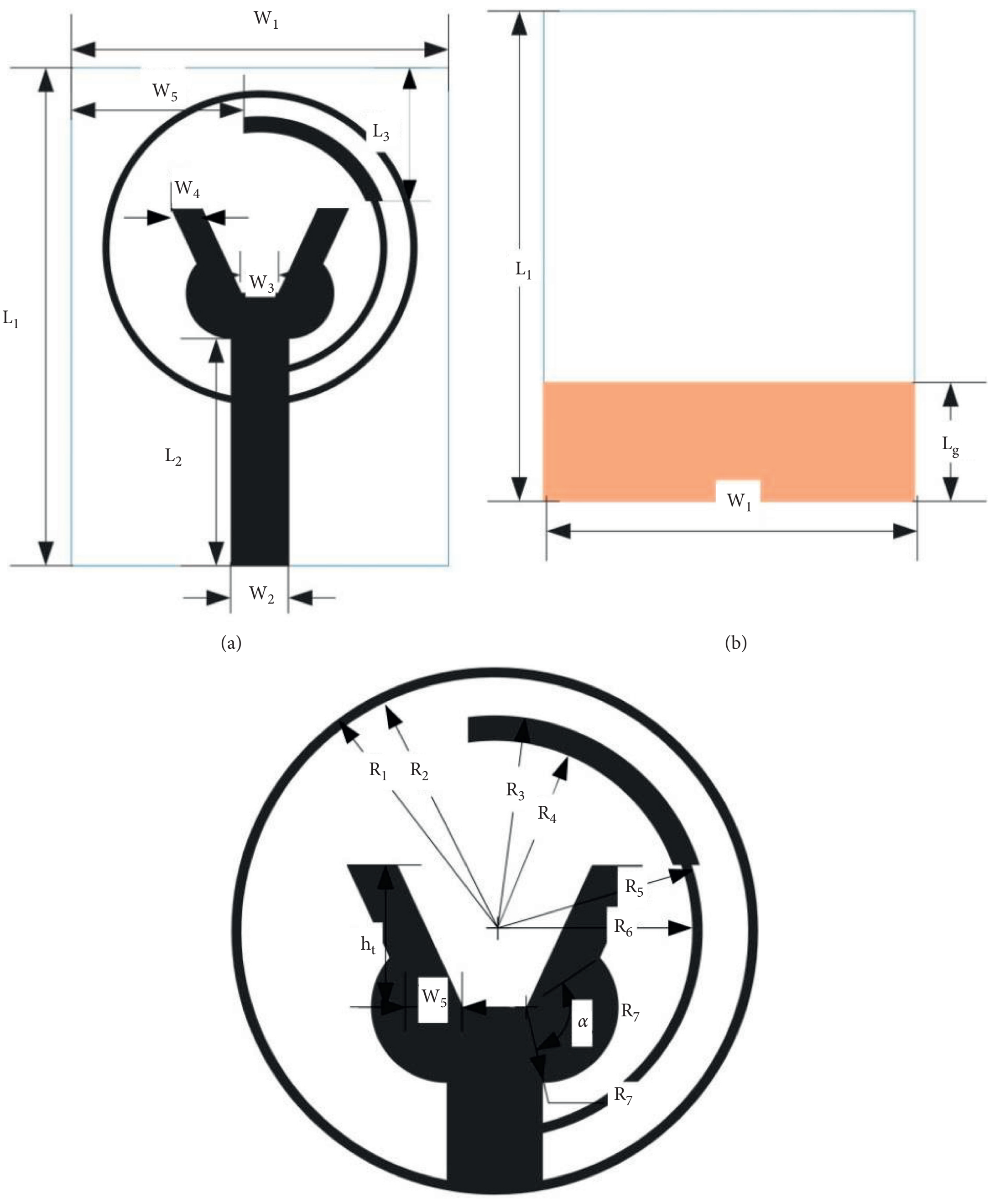
In this study, a triple-band microstrip antenna with compact size and simple structure is proposed for WLAN and 5G applications. The radiating element consists of a circular patch, a Y-shaped patch, and a monopole impedance converter. In order to obtain the desired operating bands, the monopole impedance converter is inserted between the circular patch and Y-shaped patch. The proposed antenna can work in the frequency range of 2.38~2.53 GHz, 3.29~4.11 GHz, and 4.72~5.01 GHz, with the corresponding peak gains of 4.0 9 dBi (2.4 GHz), 2.95 dBi (3.5 GHz), and 4.0 1 dBi (4.9 GHz), respectively. The measures' results are in ap-proximate agreement with the simulated values, which shows that the proposed compact antenna can offer omnidirectional radiation, appropriate gains, and sufficient bandwidths.

1. Introduction

In recent years, enhanced vehicle-to-everything, Internet of things, etc., need better data ability with high speed transmission, rapid response, and stable reliability [1]. To meet the above demands, the network technology is being shifted towards the fifth generation (5G), which aims to provide ultra-fast peak data rate, small mobility interruption time, and high transmission reliability [2]. As a terminal of global mobile data traffic, various antennas have been proposed and designed for 5G and other applications. Of all the antenna structures, microstrip antenna has some significant advantages and characteristics such as low profile, easy fabrication, mechanical stability, better compatibility, and flexible scalability.

In order to satisfy the 5G and other applications, multiple-input-multiple-output (MIMO) can highly improve the data rate, capacity, and link reliability [3]. By incorporating multiantenna systems at the transmitter as well as the receiver, $m \times n$ MIMO 5G antenna array can achieve high data throughput with good isolation and high efficiency [4–6]. In order to satisfy the low isolation degree requirement, a two-port circularly polarized antenna is designed with a rectangle ring ground plane and two T-shaped radiation patches [7]. In [8], a dual-polarization microstrip 5G MIMO antenna array is proposed with eight microstrip antenna elements arranged on the back cover of a mobile phone. Feeding networks and shorted patches' structure are proposed to realize the broadband dual-polarized filtering antenna [9]. Metamaterial structures are also used to realize the circularly polarized antenna with miniaturized size and high radiation efficiency [10]. A beam-switchable antenna is proposed with a crossover including substrate integrated waveguide and dielectric slab [11]. In addition, the fiber-reinforced plastic materials [12], the solar cells [13], airy beam at microwave frequency [14], and high-order mode in circular patch [15] are also introduced to improve the performance of the antennas.

Due to the complexity of the application environment, the pure 5G antenna cannot satisfy the demands of different users. Thus, it is required to integrate 5G frequency band with other spectrum such as WLAN, LTE, WiMAX, and



(a)

(b)

(c)

Figure 1: (a) Front view, (b) back view, and (c) radiating element of the proposed antenna.

GPS application. The multiband slotted antenna is formed by a T-shaped feed patch and a rectangular slot on top side of the substrate, which covers the GPS/WiMAX/WLAN bands [16]. A triple-band antenna consisting of a Franklin monopole strip and a rectangular patch is proposed with the operating bands at WLAN, WiMAX, and 5G, respectively [17]. In [18], a dual-band LTE-R and 5G antenna is realized by stacked configuration consisting of a double dielectric substrate, a circular patch, and five elliptical patches. By integrating a stepped patch and a ground plane [19], a low-profile multislot patch antenna is presented for LTE and 5G applications. Usually, the above multiband slot antennas have a larger size to ensure adequate current path on the radiating plane or ground plane. In order to facilitate the integration into the back cover of the handheld devices, a multiband antenna with more compact size and simple structure is desirable for WLAN and 5G application.

In this study, a compact triple-band microstrip antenna is proposed and designed for WLAN and 5G applications. The proposed antenna consists of a circular patch, a Y-shaped patch, and a monopole impedance converter. The realization principle is based on the half-wavelength

TABLE 1: The detailed dimensions of the proposed antenna.

Parameter	L_1	L_2	L_3	W_1	W_2	W_3	W_4	W_5	W_6
Size (mm)	30	15	8.8	25	3.8	2.5	2	11.4	2.6
Parameter	L_g	h_t	R_1	R_2	R_3	R_4	R_5	R_6	R_7
Size (mm)	8	5.6	10.4	10	8.7	7.7	8.4	8	3
Parameter	α								
Degree (°)	131								

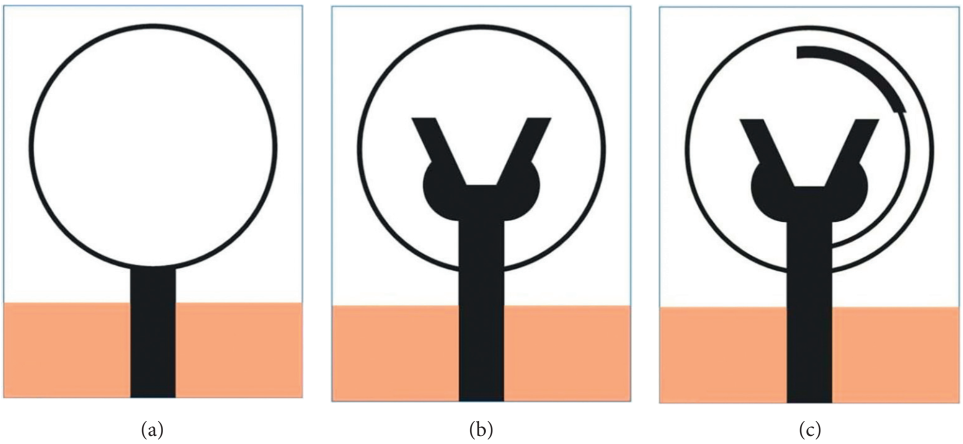

(a) (b) (c)

FIGURE 2: Design evolution of (a) antenna-1, (b) antenna-2, and (c) antenna-3.

(quarter-wavelength) resonance in the circular patch and the Y-shaped patch structure. The proposed antenna can generate the operating bands of 2.38~2.53 GHz, 3.29~4.11 GHz, and 4.72~5.01 GHz, respectively. A prototype of the antenna is manufactured and measured, and the simulation and measurement results confirm that the triple-band antenna has enough bandwidth to cover the desired WLAN and 5G applications.

2. Antenna Design Process and Parametric Analysis

Figures 1(a)~1(c) illustrate the schematic views of the proposed antenna including the front view, back view, and the radiating element. The radiating element is printed on top of the substrate, consisting of a circular patch, a Y-shape-like strip with two fan-shaped patches and a monopole impedance converter. The feed line with 50-ohm characteristics impedance is etched on the FR4 substrate, and the rectangle ground plane is printed on the backside of the dielectric substrate. In addition, the relative permittivity of the FR4 substrate is set as $\varepsilon_r = 4.4$ with the dielectric loss tangent of $\delta = 0.02$. The proposed antenna is analyzed and optimized using the HFSS Microwave Studio simulator. The detailed parameters of the proposed antenna are listed in Table 1.

Figure 2 displays the design evolution of the proposed antenna, and the corresponding simulated reflection coefficients (S_{11}) are shown in Figure 3. From Figure 2(a), it can be seen that antenna-1 is a simple monopole antenna composed of a circular ring and a rectangular ground plane. As demonstrated in Figure 3, there is only one operating band between 2.40~2.50 GHz centered at 2.45 GHz. In our design evaluation section from antenna-1 to antenna-2, we repeatedly test different kinds of radiating element to obtain another two resonant frequencies. After consulting a lot of relevant references, the Y-shape-like strip with two fan-shaped patch is selected inside the circular patch with a more compact scale. By introducing a Y-shape-like strip with two fan-shaped patches into the circular ring (antenna-2), the antenna can excite the other two operating bands, which can be used in 5G scenes. In order to obtain more compact structure, the Y-shaped monopole patch can be easily bested within the circular monopole patch and the impedance converter, which can be set as the unit cell of the MIMO antennas and easy to integrate in the back cover of the handheld devices. In order to optimize the first and third operating bands to better satisfy the requirements working frequencies, a monopole impedance converter is inserted between the Y-shaped patch and the circumscribed circle patch. It is obviously observed from Figure 3 that the operating bands of antenna-3 are 2.38~2.53 GHz, 3.29~4.11 GHz, and 4.72~5.01 GHz with the center frequencies of 2.4 GHz, 3.5 GHz, and 4.9 GHz, respectively, which satisfy the bandwidth requirements of WLAN and 5G applications.

In our proposed antenna, the realization principle is based on the half-wavelength (quarter-wavelength) resonance in the monopole radiating element with a circular patch or a Y-shaped patch structure. From Figure 3, it can be seen that the antenna-1 is a simple monopole antenna composed of a circular patch, which is fed by a 50 Ω microstrip line. The length of half of the circular patch

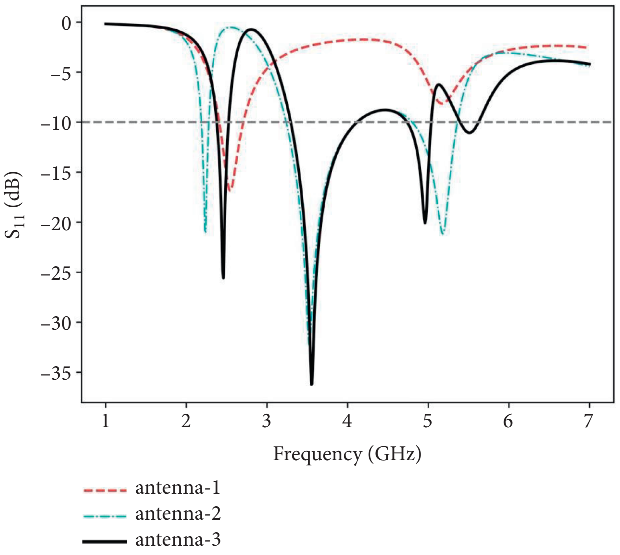


Figure 3: Simulated reflection coefficients S_{11} from antenna-1 to antenna-3.

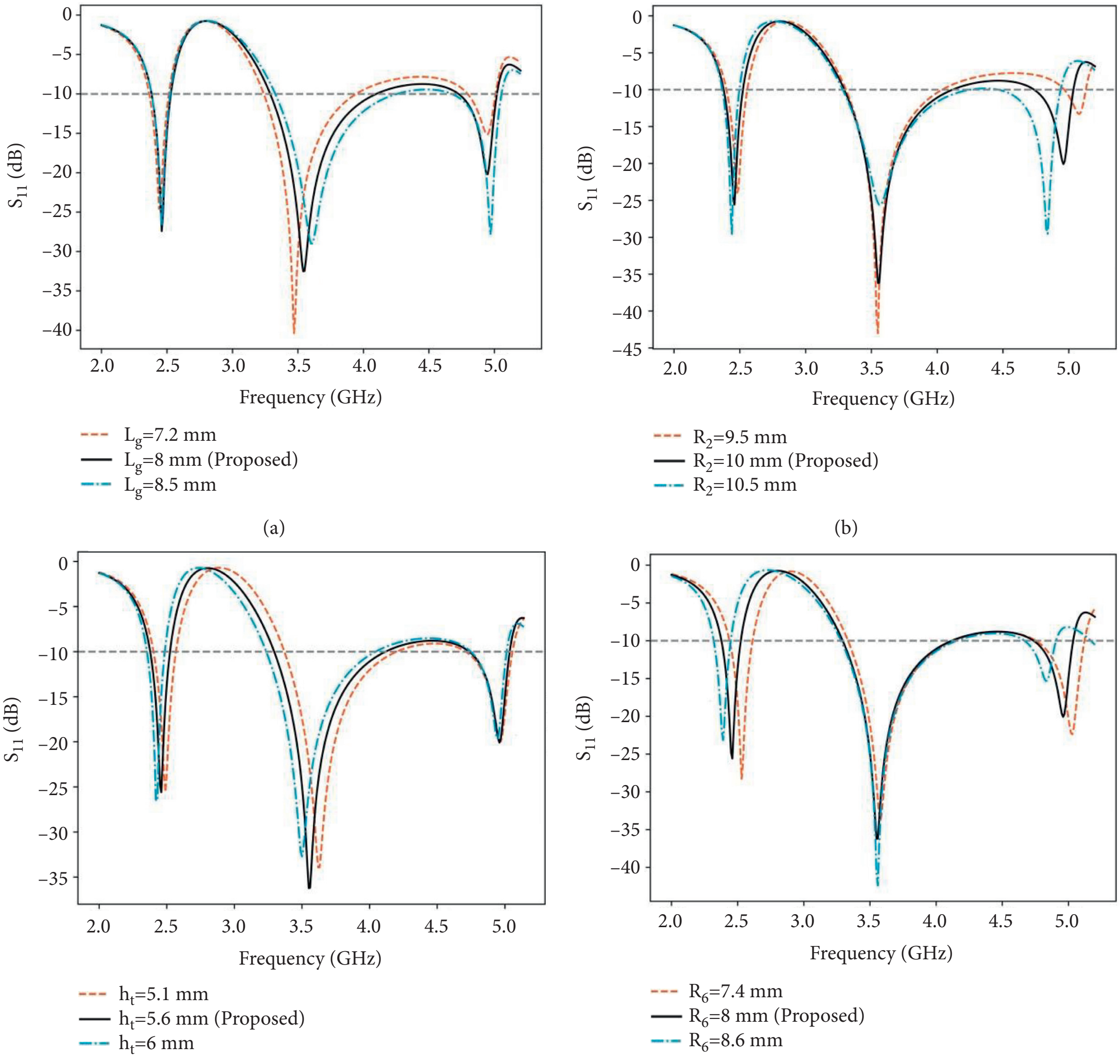


Figure 4: Simulated reflection coefficients S_{11} with different parameter values of (a) L_g, (b) W_2, (c) W_3, and (d) R_6.

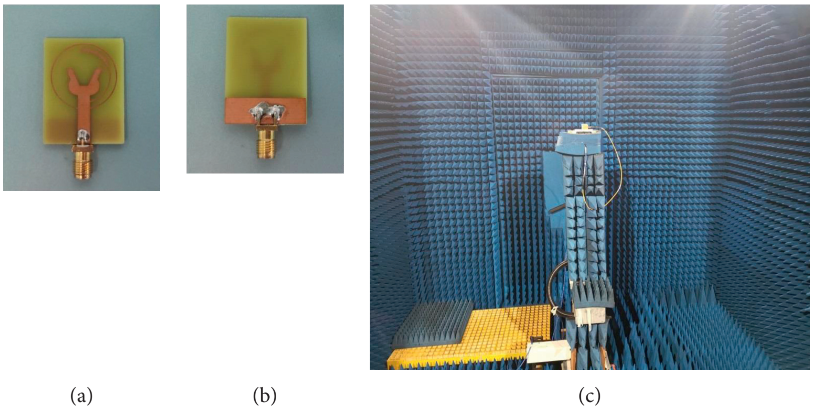


Figure 5: (a) Front and (b) back views of the fabricated antenna prototype. (c) Photograph of the measurement environment.

$(\pi * r - w_2)$ is optimized to satisfy half-wavelength of the first resonance frequency at 2.45 GHz. The corresponding resonant frequency of antenna-1 is given by

$$f_1 = \frac{c}{2 \times \sqrt{\varepsilon_{\text{eff}}} \times (\pi * r - w_2)}, \tag{1}$$

where $\varepsilon_{\text{eff}} \approx \varepsilon_r + 0.5$ and c is the speed of light. And then, antenna-2 is obtained by introducing a Y-shape-like strip with two fan-shaped patch into the circular ring. By carefully optimizing the structure parameters of the Y-shape-like strip, Antenna-2 can generate another two resonant frequencies working in 5G scenes without increasing the size of the antenna. From Figure 3, it can be found that antenna-2 has two operating bands with the resonant frequencies in the vicinity of 3.5 and 4.9 GHz. The effective length of Y-shape-like strip with two fan-shaped patch should be half (a quarter) of the guided wavelength, which can be approximately calculated by

$$f_m = \frac{c \times m}{4 \times \sqrt{\varepsilon_{eff}} \times \left(L_2 + R_7 + \sqrt{h_t^2 + W_5^2}\right)}, \quad m = 1, 2 \ldots. \tag{2}$$

In order to better analyze the operating principle of the proposed antenna, the variations of different dimensional parameters are investigated. Figure 4(a) demonstrates the variation of the simulated reflection coefficients with L_g It can be observed from the plot that the main operating frequency band of 5G is strongly dependent on the value of L_g. As the value of L_g decreases, the third resonant mode moves towards the second mode, and the bandwidth of the first resonant bandwidth becomes narrower than before. Figure 4(b) presents the simulated reflection coefficients for different values of R_2. As the figure describes, the first and second resonant mode are almost unchanged, and the third frequency demonstrates redshift obviously. From the above analysis, we can see that the desired third band for 4.9 GHz can be tuned by adjusting R_2. As described in Figure 4(c), by increasing the value of ht, the first and second operating bands have an obvious redshift while the third band changed slightly. Thus, the ranges of these two operating bands can be effectively adjusted by changing the value of ht. Figure 4(d) displays the variation of the reflection coefficients for different values of R_6. It can be observed that, with the increase of R_6, the first and third bands show redshift while the middle band keeps unchanged.

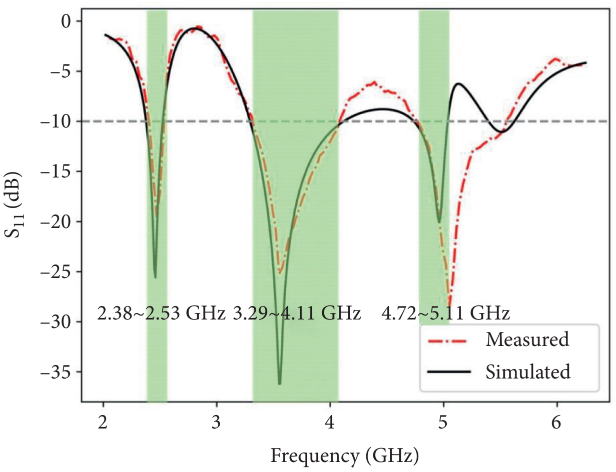
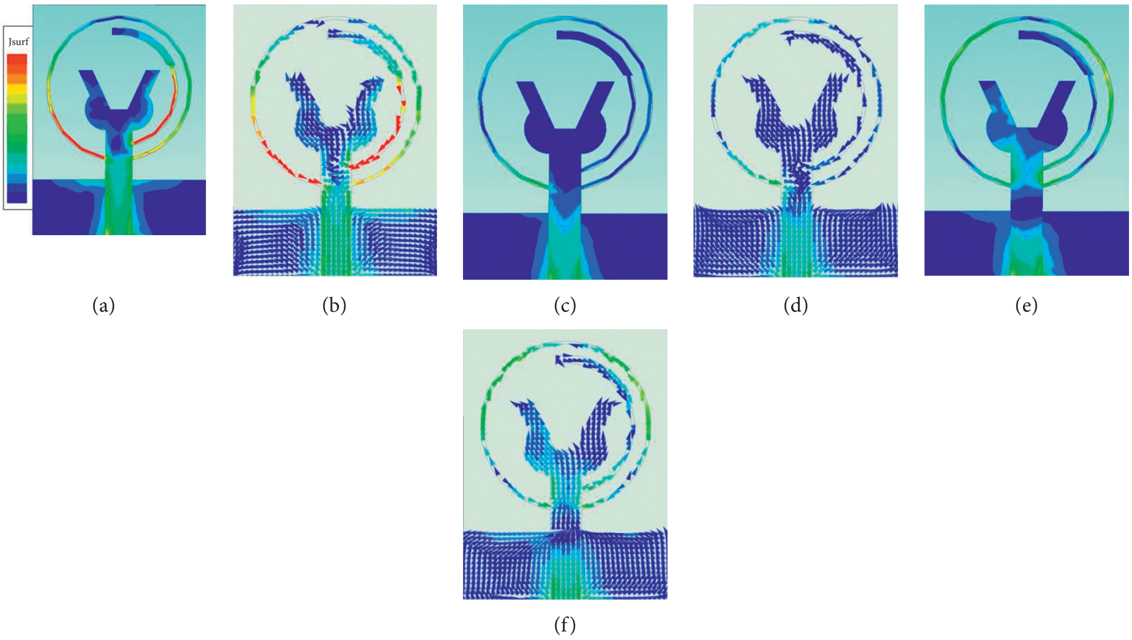


Figure 6: Simulated and measured reflection coefficients S_{11} of the proposed antenna.

3. Results and Discussion

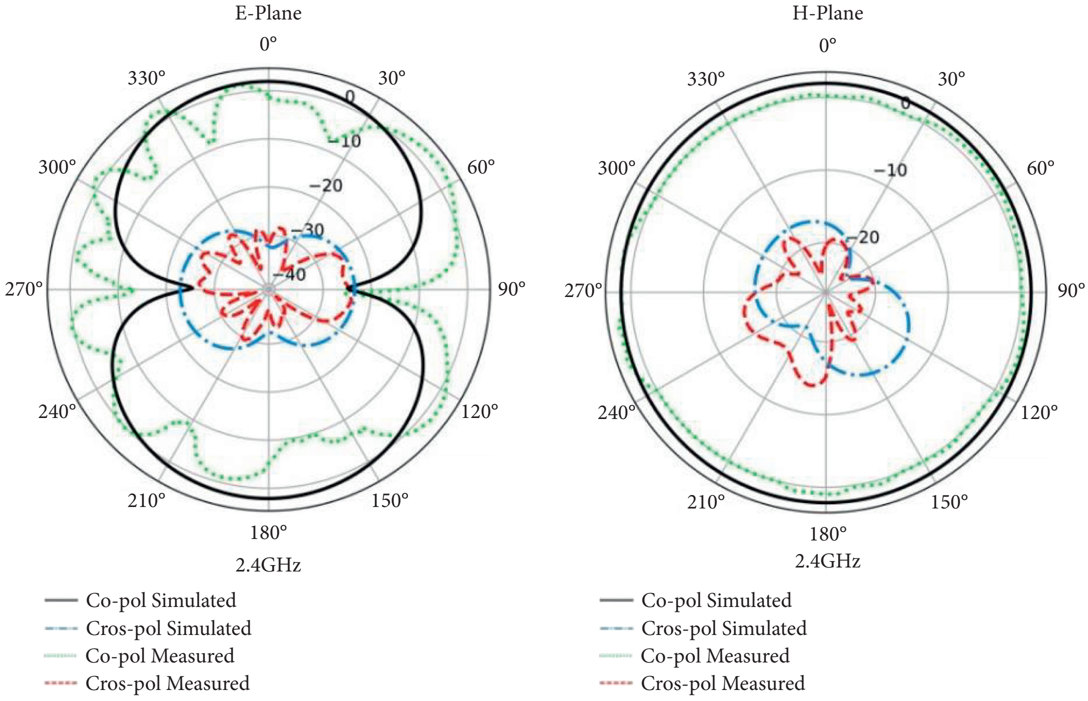
Figures 5(a) and 5(b) show the front and back views of the fabricated antenna prototype. The radiation characteristics of the antenna are measured in the anechoic chamber, and Figure 5(c) shows its actual environment for the far-field measurements. The overall size of the anechoic chamber is 8 m (length) × 6 m (width) × 3.5 m (height). During the measurement process, the distance between the auxiliary horn antenna and the antenna under test is set as 5 m, which can completely satisfy the requirements of far-field measurement. The measured reflection coefficients together with the simulated values are shown in Figure 6, from which we can see that three operating frequency bands of the antenna ($S_{11} \leq 10$ dB) are 2.38~2.53 GHz, 3.29~4.11 GHz, and 4.72~5.01 GHz, respectively. As shown in the plot, the simulated and measured results have an approximate agreement in the working frequency band of the antenna.



(a) (b) (c) (d) (e)

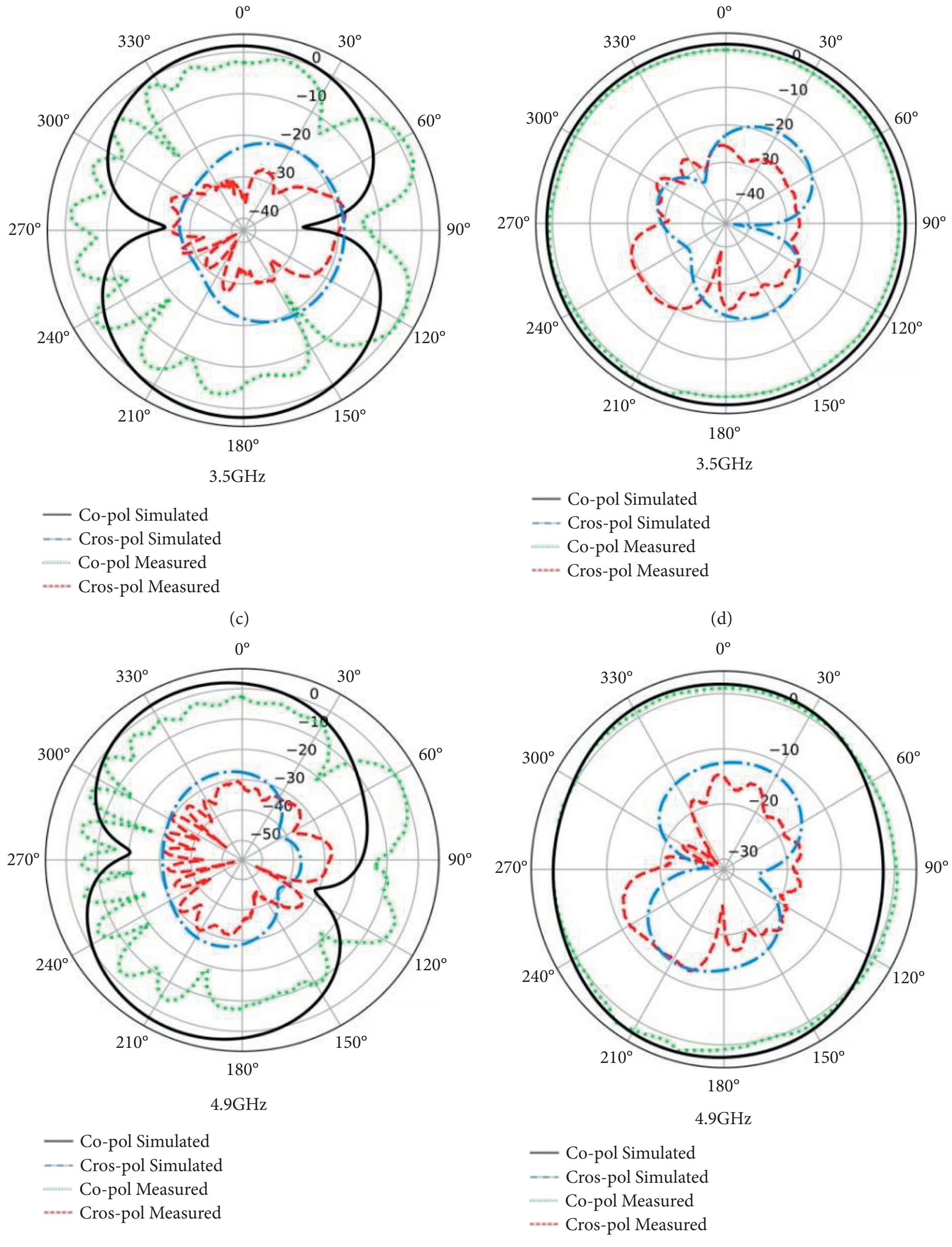
(f)

Figure 7: Simulated surface current distributions at (a) 2.4 GHz, (c) 3.5 GHz, and (e) 4.9 GHz, and the corresponding current vector distribution at (b) 2.4 GHz, (d) 3.5 GHz, and (f) 4.9 GHz.



(a) (b)

Figure 8: Continued.



(c)

(d)

(e)

(f)

Figure 8: Simulated and measured 2D (a) co-polar and (b) cross-polar radiation patterns.

The discrepancies between the simulated and measured results may be attributed to the fabrication inaccuracy, the abrasion of the patch, and the welding errors of the connector.

To better understand the working mechanism of the proposed antenna, the simulated current and current vector distributions at the resonant frequencies of 2.4 GHz, 3.5 GHz, and 4.9 GHz are shown in Figure 7. From Figures 7(a) and 7(b), it can be seen that the surface current is mainly distributed along the lower part of the circular ring and monopole impedance converter. It is illustrated that the resonant frequency of 2.4 GHz is generated by the combination between the circular ring and the monopole impedance converter. Due to the introduction of the monopole

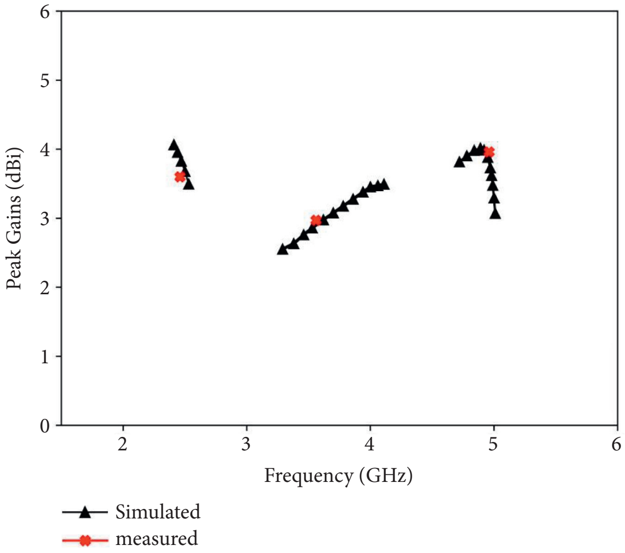


Figure 9: Simulated and measured peak gains at different operating bands.

Table 2: Comparison of the proposed antenna with other previously reported antennas.

Refs.	Operating bands (≤10 dB)	Antenna dimensions (m^3)	Peak gains (dBi)	Profile	Substrate	Applications	Omnidirectional patterns
[4]	3.4–3.8 GHz	75 × 150 × 1.6	3	Planar	FR4	5G	No
[5]	3.3–4.2 GHz	42 × 42 × 1	None	Planar	FR4	5G	No
[6]	3.3–3.6 GHz 4.8–5.0 GHz	150 × 70 × 0.8	None	Planar	FR4	5G	No
[7]	3.3–4.2 GHz	37 × 30 × 0.8	2.5	Planar	FR4	5G	No
[8]	4.37–5.5 GHz	150 × 72 × 1.2	3.7	3D	FR4	5G	No
[9]	3.3–5.1 GHz	60 × 60 × 8	8.1	3D	Rogers RO4003	5G	No
[10]	3.30–3.80 GHz	50 × 50 × 3.5	5.4	3D	F4BK	5G	No
[11]	4.25–5.82 GHz	80 × 80 × 6	9.1	3D	F4BME220	5G	No
[13]	4.8–5 GHz	80 × 80 × 3.76	10.85	3D	F4BM350RUILONG220	5G	No
[18]	0.66–0.79 GHz 3.28–3.78 GHz	180 × 60 × 1.6	2.46.1	3D	FR4	LTE5G	No
Proposed	2.38–2.53 GHz 3.29–4.11 GHz 4.72–5.01 GHz	30 × 25 × 1.59	4.092.214.01	Planar	FR4	WLAN5G	Yes

impedance converter, the distribution of the surface current is not symmetrical along the circular ring. Despite the emergence of the middle band is attributed to the introduction of the Y-shape-like strip with two fan-shaped patches, the current is mainly distributed in the left side of the circular ring, as shown in Figures 7(c) and 7(d). This is because that the currents on the monopole impedance converter and the right side of the circular ring have opposite directions, which causes the radiation energy to be cancelled. At 4.9 GHz, it is found in Figures 7(e) and 7(f) that the current is mainly focused on the circular ring and the Y-shape-like strip.

The simulated and measured 2D co-polar and cross-polar radiation patterns at the frequencies of 2.4 GHz, 3.5 GHz, and 4.9 GHz are shown in Figure 8. It can be observed from the plot that the proposed antenna has a stable omnidirectional radiation pattern. Moreover, the cross-polarized radiation distributions are significantly less than the co-polarized counterparts. The difference between the simulated and measured 2D radiation pattern in E-plane may be caused by fabrication tolerance, chamber scattering, measurement error, etc. Thus, the designed antenna exhibits stable radiation patterns over the operating band, which satisfies the requirement for 5G and other wireless applications.

As shown in Figure 9, the simulated peak gain of the proposed antenna is 3.56~4.16 dBi, 2.45~3.6 dBi, and 3.07~4.01 dBi in 2.38~2.53 GHz, 3.29~4.11 GHz, and 4.72~5.01 GHz, respectively, while the measured values are 3.60 dBi, 2.95 dBi, and 3.96 dBi at 2.4 GHz, 3.5 GHz, and 4.9 GHz, respectively. From Figure 9, it can be found that the gains of the middle band are lower than that of the upper and lower bands. Due to the opposite current vector direction on the monopole impedance converter and the right side of the circular ring at 3.5 GHz, the peak gain is relatively weaker than that at 2.4 GHz or 4.9 GHz. The gain values mainly depend on the current intensity on the radiating element in the antenna, which converts to the radiating energy of the electromagnetic field. From the simulated surface current distributions in Figure 7, we can see that the current intensity in Figure 7(d) is weaker than any other distribution in Figures 7(b) or 7(f). Table 2 displays the comparison of the performance between proposed and other previously represented antennas. Through the comparison with the references, it can be observed that our proposed antenna can provide enough bandwidths, appropriate gains, and omnidirectional radiation characteristics with more compact size.

4. Conclusions

A triple-band compact microstrip antenna is proposed for WLAN and 5G applications. The radiation element of the proposed antenna consists of a circular ring, a Y-shape-like patch, and a monopole impedance converter. With the help of the monopole impedance converter, the proposed antenna can be optimized in the operating bands of 2.38~2.53 GHz, 3.29~4.11 GHz, and 4.72~5.01 GHz, respectively. From the experimental and measured results, we can see that the antenna has stable omnidirectional patterns, appropriate gains, and enough bandwidth. By comparing the performance with the previously presented antenna, the proposed antenna is a suitable choice to be applied for the 5G antenna design and other wireless applications.

References

[1] H. Yu, H. Lee, and H. Jeon, "What is 5G? Emerging 5G mobile services and network requirements," *Sustainability, MDPI*, vol. 9, no. 10, pp. 1–12, 2017.

[2] A. Dogra, R. K. Jha, and S. Jain, "A survey on beyond 5G network with the advent of 6G: architecture and emerging technologies," *IEEE Access*, vol. 9, Article ID 67512, 2021.

[3] Q.-U.-A. Nadeem, A. Kammoun, M. Debbah, and M.-S. Alouini, "Design of 5G full dimension massive MIMO systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 726–740, 2018.

[4] N. O. Parchin, Y. I. A. Al-Yasir, A. H. Ali et al., "Eight-element dual-polarized MIMO slot antenna system for 5G smartphone applications," *IEEE Access*, vol. 7, Article ID 15612, 2019.

[5] I. R. R. Barani, K.-L. Wong, Y.-X. Zhang, and W.-Y. Li, "Low-profile wideband conjoined open-slot antennas fed by grounded coplanar waveguides for $4\times4\,\,5$ G MIMO operation," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 4, pp. 2646–2657, 2020.

[6] W. Hu, X. Liu, S. Gao et al., "Dual-band ten-element MIMO array based on dual-mode IFAs for 5G terminal applications," *IEEE Access*, vol. 7, Article ID 178476, 2019.

[7] A. K. Dwivedi, A. Sharma, A. K. Pandey, and V. Singh, "Two port circularly polarized MIMO antenna design and investigation for 5G communication systems," *Wireless Personal Communications*, vol. 120, no. 3, pp. 2085–2099, 2021.

[8] B. Cheng and Z. Du, "Dual polarization MIMO antenna for 5G mobile phone applications," *IEEE Transactions on Antennas and Propagation*, vol. 69, no. 7, pp. 4160–4165, 2021.

[9] Y. Zheng and W. Sheng, "Compact dual-polarized filtering antenna with enhanced bandwidth for 5G sub-6 GHz applications," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 31, no. 9, pp. 1–14, 2021.

[10] Z. Wang, T. Liang, and Y. Dong, "Metamaterial-based , compact, wide beam-width circularly polarized antenna for 5G indoor application," *Microwave and Optical Technology Letters*, vol. 63, no. 8, pp. 2171–2178, 2021.

[11] Y. Cheng and Y. Dong, "Wideband beam-switchable antenna loaded with dielectric slab for 5G applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 20, no. 8, pp. 1557–1561, 2021.

[12] C. M. Preddy, R. Singh, P. H. Aaen, and S. R. P. Silva, "Integrated carbon-fiber-reinforced plastic microstrip patch antennas," *IEEE Antennas and Wireless Propagation Letters*, vol. 19, no. 4, pp. 606–610, 2020.

[13] Y. Luo, J. Lai, N. Yan, W. An, and K. Ma, "Integration of aperture-coupled multipoint feed patch antenna with solar cells operating at dual compressed high-order modes," *IEEE Antennas and Wireless Propagation Letters*, vol. 20, no. 8, pp. 1468–1472, 2021.

[14] Z. Yang, G. Wen, W. Hu, D. Inserra, Y. Huang, and J. Li, "Microwave airy beam generation with microstrip patch antenna array," *IEEE Transactions on Antennas and Propagation*, vol. 69, no. 4, pp. 2290–2301, 2021.

[15] S. Radavaram, S. Naik, and M. Pour, "Stably polarized wideband circular microstrip antenna excited in TM12 mode," *IEEE Transactions on Antennas and Propagation*, vol. 69, no. 4, pp. 2370–2375, 2021.

[16] Y. F. Cao, S. W. Cheung, and T. I. Yuk, "A multiband slot antenna for GPS/WiMAX/WLAN systems," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 3, pp. 952–958, 2015.

[17] M. E. Yassin, H. A. Mohamed, E. A. F. Abdallah, and H. S. El-Hennawy, "Single-fed 4G/5G multiband 2.4/5.5/28 GHz antenna," *IET Microwaves, Antennas & Propagation*, vol. 13, no. 3, pp. 286–290, 2019.

[18] A. K. Arya, S. J. Kim, and S. Kim, "A dual-band Antenna for lte-R and 5g lower frequency operations," *Progress In Electromagnetics Research Letters*, vol. 88, pp. 113–119, 2020.

[19] R. Azim, A. M. H. Meaze, A. Affandi et al., "A multi-slotted antenna for LTE/5G Sub-6 GHz wireless communication applications," *International Journal of Microwave and Wireless Technologies*, vol. 13, no. 5, pp. 486–496, 2021.

Development of a 100 mW-Class 94 GHz High-Efficiency Single-Series Rectifier Feed by Finline for Micro-UAV Application

Pravat Kumar Subudhi, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, pk.subudhi11@outlook.com*

Madhusudan Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, madhusudandas55@hotmail.com*

Prangya Paramita Padhi, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, prangya.p.padhi@gmail.com*

Subhendu Sahoo, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, s.sahoo95@outlook.com*

Abstract

Wireless power transfer (WPT) is one solution to realize long flight times and accommodate various missions of micro-uncrewed aerial vehicles (MAVs). Reducing the constraint of power transmission distance and realizing high beam efficiency are possible because of the high directivity of WPT using millimeter wave (MMW) methods. Nevertheless, no report of the relevant literature describes an investigation of sending power to an MAV using MMW because MMW rectennas have low efficiency. The purpose of our study is to conduct fundamental research of a high-efficiency and high-power rectenna at 94 GHz aimed at MAV application using MMW. As described herein, we developed and evaluated a 100-mW-class single-diode rectifier at 94 GHz with a finline of a waveguide (WG) to a microstrip-line (MSL) transducer. With the optimum load of 150 Ω at input power of 128 mW, the output DC power and rectifying efficiency were obtained respectively as 41.7 mW and 32.5%. By comparison to an earlier study, measurement of 94 GHz rectifiers under high power input becomes more accurate through this study.

1. Introduction

In recent years, micro-UAVs (MAVs) have been developed extensively. They are used in various applications, from commercial to military. Among them, some types have less than 100 g weight, which makes them inexpensive and easy to use. However, because of the small payload, the installed batteries are a constraint, necessitating short flight times of MAVs. One solution to resolve this constraint is to send power from a ground base to a UAV using wireless power transfer (WPT). Using magnetic resonance, WPT can send sufficient power to drive UAVs stably, but it cannot realize long-distance transmission. Moreover, it restricts the free flight of MAVs. However, WPT using RF energy can reduce the distance limit. As an example of research on UAVs, a power transmission experiment from the ground to a small airplane (SHARP) [1] has been reported in Canada. In addition, the University of Kyoto conducted experiments on power transmission to model airplanes and airships [2]. In our research group, WPT to MAVs was conducted in the microwave region [3, 4]. According to beam efficiency [5], as introduced by the Friis transmission equation, the higher the operating frequency becomes, the more efficiently WPT is realized. Shimamura et al. introduced the relation between the MAV size and the received power ratio for operating MAVs [6], as shown in the following equation:

$$\eta_{\mathrm{r}} = \eta_{\mathrm{T}}\left[1 - \exp\left(-\frac{\pi k^2 S}{16}\right)\right], \quad k = \frac{f D_{\mathrm{T}}}{\mathrm{cd}}. \tag{1}$$

In this equation, η_{T}, S, f, D_{T}, c, and d respectively denote the beam efficiency, the MAV wing area, the operating frequency, the transmitting antenna diameter, the speed of light, and the transmission distance. Using the equation and an MAV (HS210; Holy Stone), we calculated and compared the received power under two operating frequencies:

5.8 GHz and 94 GHz. The parameters used and the calculated results are presented in Table 1. From the results, one can infer that higher frequencies can produce higher transmission efficiency, given the same transmission distance. That relation demonstrates the superiority of millimeter-wave (MMW) transmission. Among the MMWs available for use, 94 GHz is an atmospheric window. It is therefore possible to transmit power to MAVs efficiently at that frequency. Nevertheless, no study has examined the transmission of power to MAVs at 94 GHz because the antenna directivity is too high [7, 8] and because the rectifier efficiency is too low. Additionally, in MMW circuits, the conductor and dielectric loss in the oscillator and transmission line parts become greater than those related to microwaves. Moreover, conversion efficiency at the diode is low. Consequently, the conversion efficiency of the rectenna decreases. Research on MMW power in the field of wireless power transmission has remained inadequate compared to that for microwave power [9–12].

A millimeter-wave circuit has a short wavelength. For that reason, the circuit is small and difficult to fabricate as designed. To create such a small circuit, improvements have been made by mounting circuits with rectifying elements on a semiconductor substrate using CMOS technology [13–16]. However, the rectifiers of these technologies have low withstand voltage. Many elements must obtain the power required by an MAV or a satellite. For that reason, concerns of increased weight and cost arise. Furthermore, even when using GSG probes, which are generally used to supply power to planar circuits, large withstand voltage cannot be achieved because the maximum power that can be input is limited to several milliwatts as the frequency increases. Even in studies using a packaged GaAs Schottky diode [17], 100 mW class rectification was not realized at 94 GHz. For a study using another diode [18], a high-power oscillation source gyrotron was used to supply high power to the rectenna and thereby achieve 100 mW class input. Nevertheless, difficulties persist because the rectifier circuit and the antenna cannot be evaluated separately: accurate performance evaluation and design modification are difficult; moreover, the rectification efficiency is as low as 20%.

The purpose of our study is to conduct fundamental research of a high-efficiency and high-power rectenna at 94 GHz, aiming at MAV application using MMW. For an earlier study [19], we developed 94 GHz rectenna using a microstrip line (MSL). First, we fabricated a single-series rectifier feed using a finline, which is a waveguide- (WG-) microstrip line (MSL) transducer, as shown at the bottom of Figure 1, to achieve 100 mW-class input to the rectifier. However, the rectification efficiency of the rectifier does not match that of the rectenna in the earlier study. That is true because we used an open stub as a second harmonics notch filter and because the finline transmitting efficiency is not accurate in the study [20].

For the present study, we evaluated the finline transmitting efficiency accurately and developed a 94 GHz rectifier that realizes 100 mW-class input. The finline realizes separate evaluation of a rectifier and an antenna, which engenders improved design accuracy of tiny MMW rectennas. We select a single-series rectifier considering the low conversion efficiency at a diode and large insertion loss of a DC block in the MMW region. After evaluating the transmission efficiency of the finline, the rectifying efficiency and power are measured. The rectifying efficiency (ηrec) is calculated using equation (2). Also, P_{in}, P_{DC}, P_{o}, η_{fin}, A_{r}, A_{t}, λ, and d respectively denote the input power to the rectifier, the rectified DC power (P_{DC}) obtained by measurement, power from the oscillator, the transmission efficiency of the finline, the effective area of a receiving antenna, the effective area of a transmitting antenna, the wavelength, and the transmission distance. The rectifying efficiency was compared to the rectifying efficiency of the rectenna in an earlier study [21].

$$\eta_{\mathrm{rec}} = \frac{P_{\mathrm{DC}}}{P_{\mathrm{in}}} = \frac{P_{\mathrm{DC}}}{\eta_{\mathrm{fin}} P_{\mathrm{o}}} = \frac{(\lambda d)^2 P_{\mathrm{DC}}}{A_{\mathrm{r}} A_{\mathrm{t}} P_{\mathrm{o}}}. \qquad (2)$$

2. Design of a 94 GHz Rectifier

A microstrip line (MSL) is used as a transmission line because it is easy to fabricate on a printed circuit board (PCB) and easy to use for the design of filters and stubs. Considering the insertion loss of a DC block with a millimeter wave, a single-diode series-connected rectifier was selected and fabricated. Two filters are set to increase the rectifying efficiency in the rectifier, as presented in Figure 2. A second harmonic notch filter (filter 1) is used to prevent the second harmonic, which arises inside the diode, from flowing back to the input port. Furthermore, filter 1, which consists of a short stub, plays the role of maintaining electrical conduction between the top and bottom layer. A fundamental harmonic notch filter (filter 2) is used to prevent the fundamental harmonic from flowing out to the DC port. Filter 2 has narrower line width than other parts of the circuit, which contributes to reduction of the effect of discontinuity at the junction [22]. The MSL loss increases as the line width narrows. Therefore, choosing an appropriate line width contributes to high efficiency. We compared two versions: $W_{\mathrm{s}} = 0.1$ and 0.15 mm. We used a substrate (relative permittivity $\varepsilon_{\mathrm{r}} = 2.19$, loss tangent $\tan\delta = 0.003$ at 94 GHz, Cu thickness = 18 mm; NPC F220CJ; Nippon Pillar Packing Co., Ltd.). The Schottky barrier diode (MA4E1310; Macom) is chosen as a diode because it has durability and responsivity of high-power driving in the W-band. Several Schottky barrier diode parameters are presented in Table 2.

Filter 2 is simulated using the finite element method (FEM) with an electromagnetic simulator (EMPro; Keysight Technologies Inc.), as shown in the left panel of Figure 3. FEM realizes accurate modelling of the MMW region, but it cannot simulate nonlinear devices such as diodes. We simulated only MSL elements. The result is shown in the right panel of Figure 3. Actually, 0.1 mm shows better filtering characteristics than 0.15 mm at the operating frequency of 94 GHz. Therefore, we describe fabrication of the 0.1 mm-wide filter 2 in a later chapter.

Table 1: Calculated received power ratio for an MAV (HS210; Holy Stone) at operating frequencies of 5.8 GHz (microwave) and 94 GHz (MMW).

Parameter	Weight	Wing area, S	Transmission distance, d	Transmitting antenna diameter, D_T	Calculated result	
					5.8 GHz	94 GHz
Value	0.022 kg	0.0064 m^2	10 m	1 m	0.0028	0.43

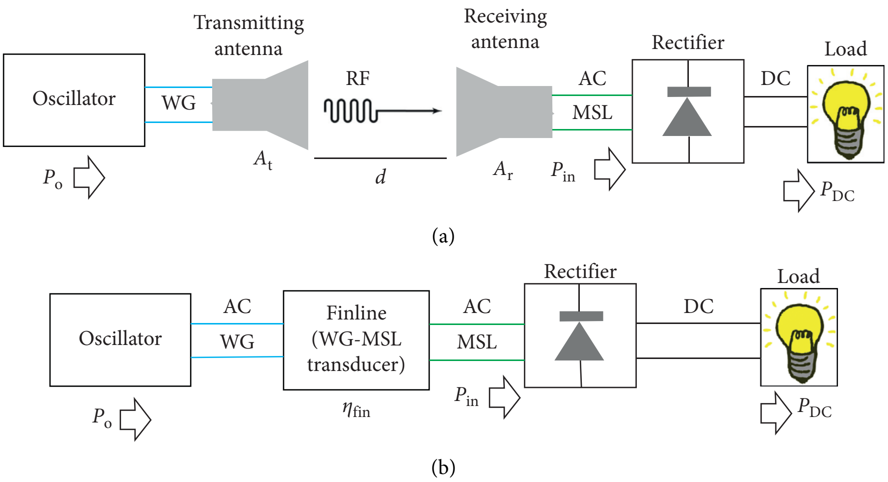


Figure 1: Rectenna (a) and rectifier (b) configurations. Each power and efficiency value used in later equations is presented in this figure. The rectifier is connected directly to the oscillator through a finline. The rectenna is fed power through free-space propagation.

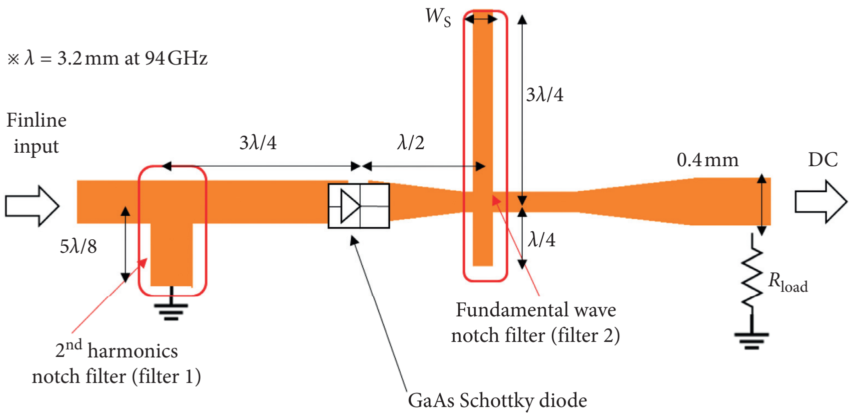


Figure 2: Design of a single-diode series-connected rectifier. A high breakdown voltage diode (MA4E1310; Macom) is selected. Two filters are used to increase the rectifying efficiency. Two widths of filter 2 are compared to obtain better filter properties.

Table 2: Several Schottky barrier diode parameters (MA4E1310; Macom).

Parameter	Junction capacitance at 0 V, 1 MHz	Forward voltage at 1 mA	Reverse breakdown voltage at 10 μA	Incident maximum RF power
Value	0.1 pF	0.7 V	7 V	20 dBm

3. Efficiency Measurements and Discussion

3.1. Finline Transmitting Efficiency. For this study, a finline [21–23] is used to input power to a rectifier. It realizes high-power operation. Although the transmission efficiency of a finline is necessary to evaluate the rectifying efficiency, it is impossible to measure the finline transmission efficiency (η_{fin}) directly because an end of a finline has MSL shape. Moreover, it is impossible to connect to the WG port of the W-band vector network analyzer (VNA). Therefore, a finline-MSL-finline (FMF) sample, as portrayed in Figure 4, was fabricated to allow that connection. Furthermore, to

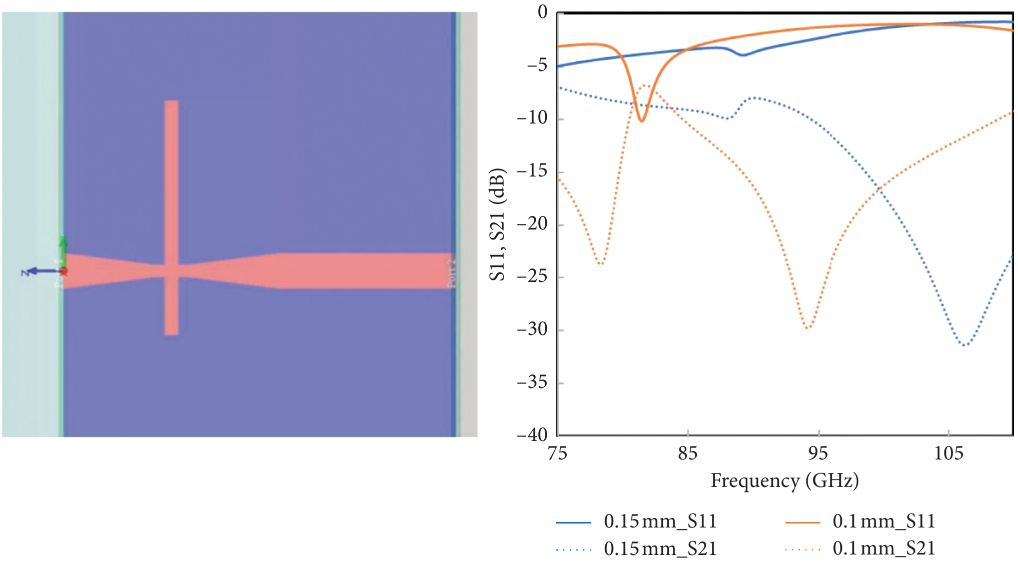


Figure 3: Configuration of the simulated part of the filter 2 in EMPro (a). Simulated result of S-parameter (b). The widths are 0.1 mm and 0.15 mm. Based on the result, the 0.1 mm-width filter shows a better notch filter property at 94 GHz.

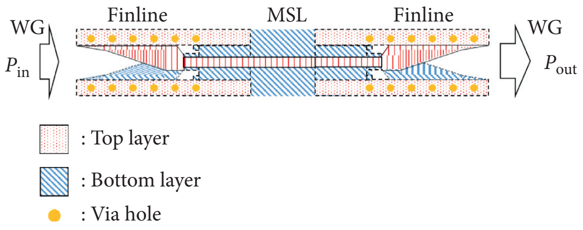
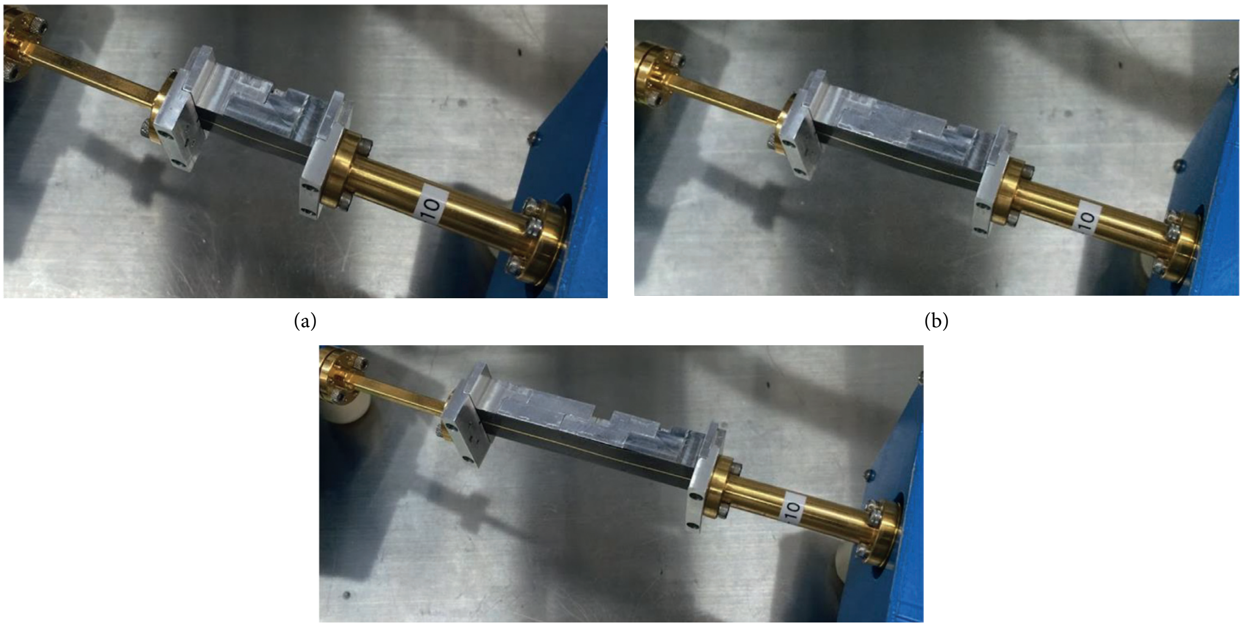


Figure 4: Configuration of a FMF sample. This sample is used for finline transmission measurements because the MSL part cannot be connected to the WG port of VNA. By changing the MSL part length and comparing them, the MSL loss per unit length is obtained. By removing the contribution of MSL, pure finline efficiency is obtainable.


(a)

(b)

(c)

Figure 5: Measurement configuration of FMF samples. Using a vector network analyzer, the transmission efficiency can be measured: 40 mm (a), 61 mm (b), and 82 mm (c). By changing the MSL part length and comparing their transmission efficiencies, the MSL loss per unit length can be calculated.

Table 3: Measurement result of FMF sample transmission efficiency and calculated result of η_{MSL} and η_{fin}.

Transmission efficiency (dB)			MSL loss (dB/mm)	MSL loss α ($L = 40$ mm) (dB)	$\eta_{\mathrm{MSL}} = 1 - \alpha$	η_{fin}
$L = 40$ mm	$L = 61$ mm	$L = 82$ mm				
−3.50	−4.90	−6.23	0.065	2.08	0.62	0.85

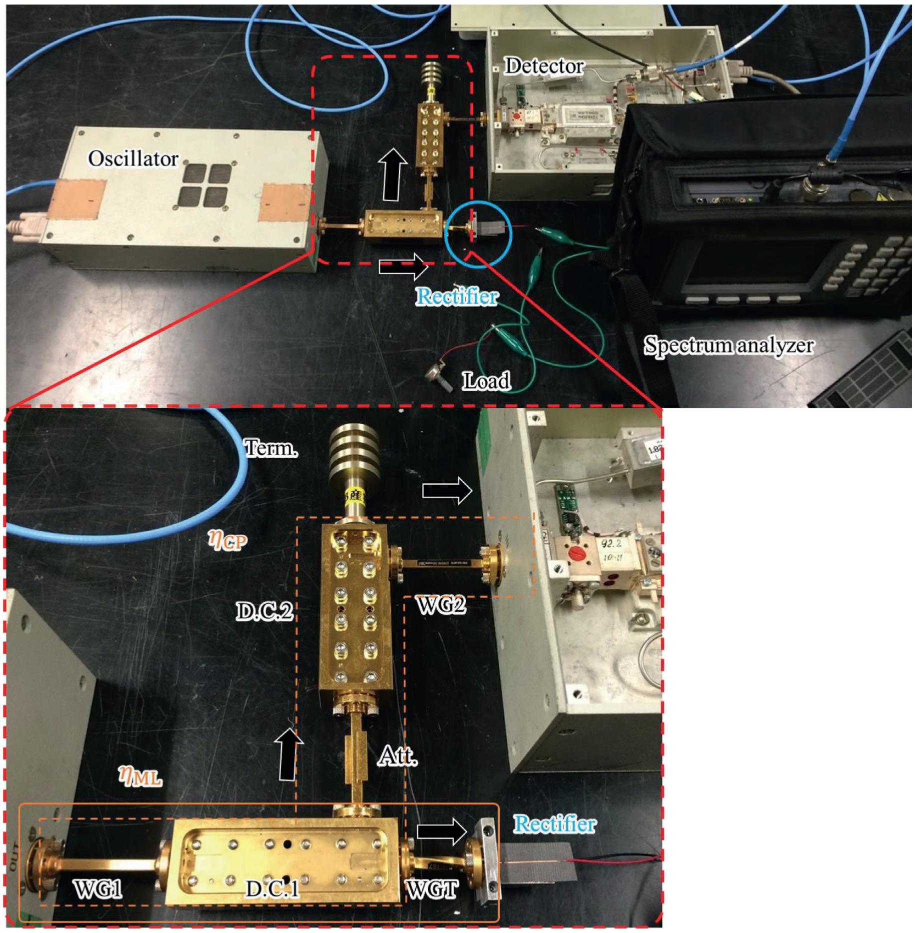


Figure 6: Rectifier measurement configuration. Power from the oscillator is divided at a directional coupler (D.C.1): one is input to the rectifier through the finline; the other is input to the detector. Voltage across the load is measured using the voltmeter.

eliminate the loss effect in MSL, we fabricated three FMF samples with different MSL lengths (40 mm, 61 mm, and 82 mm) and compared their transmission efficiencies. Results show that the MSL loss per unit length was calculated. The pure finline transmission efficiency was evaluated using equation (3).

Photographs of the FMF samples are presented in Figure 5. A W-band VNA (PNA-X N5247; Keysight Technologies Inc.) is used for FMF sample efficiency measurements. A support made of duralumin is used to connect the sample to the WG port of the VNA. To compensate the MSL length effect, the support is designed to change its length merely by inserting a joint with a length of 21 mm. The result is presented in Table 3. The measured finline transmission efficiency was 85%. This efficiency is used in rectifier efficiency measurements as described in the next section.

$$\eta_{\mathrm{fin}} = \sqrt{\frac{P_{\mathrm{out}}}{\eta_{\mathrm{MSL}} P_{\mathrm{in}}}}. \tag{3}$$

3.2. Rectifying Efficiency of a Rectifier. Figure 6 shows the measurement configuration. A 94 GHz/400 mW oscillator and a heterodyne detector (TR-10/94/x ELVA-1) are used for this measurement. The input power and the voltage at the end of the rectifier circuit are visible simultaneously using a directional coupler (D.C.1), which divides the power before input to the rectifier circuit. Another directional coupler (D.C.2) and an attenuator (Att.) are used to attenuate the power from the oscillator and to meet the upper limit of input to the detector. The main direction of D.C.2 is terminated to eliminate the effect of the open end. Output power from the oscillator is input to the circuit through the

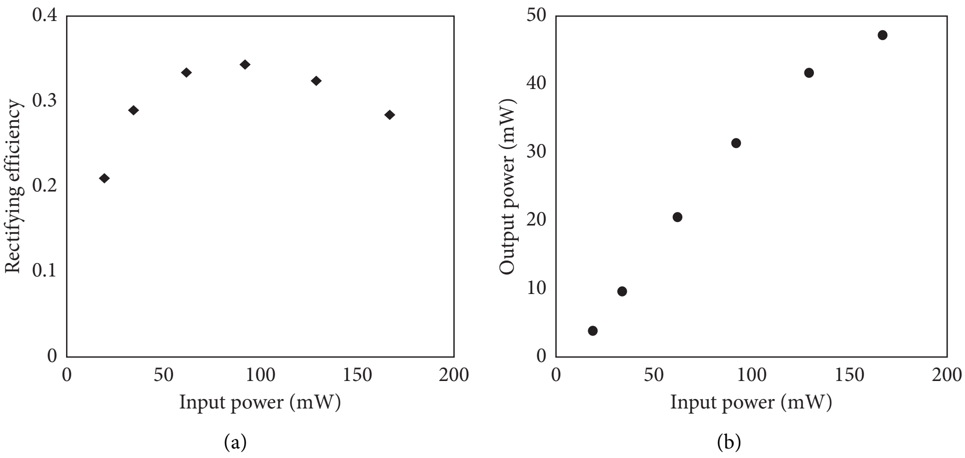


FIGURE 7: Measured rectification results at the optimum output impedance conditions: rectifying efficiency vs. input power (a) and output power vs. input power (b). At input power of 128 mW, the rectifying efficiency and output power are, respectively, 41.7 mW and 32.5%. We use this point as indicating high performance.

TABLE 4: Comparison with earlier works examining 94 GHz rectifiers.

	[13]	[18]	[17]	[14]	This work
Rectification method	CMOS Schottky diode	Mott diode (IPM RAS)	GaAs Schottky diode (VDI)	Diode-connected transistor	GaAs Schottky diode (MACOM)
Technology	130 nm CMOS	0.254 mm PTFE	0.254 mm alumina	65 nm CMOS	0.127 mm PTFE
Transmission line	FGCPW	MSL	CPW	Slotline	MSL
Power supply to a rectifier	Dual-band LTSA	Four-element MSA	Bow-tie slot	Half-wave horizontal dipole	Finline
Input RF power	2.27 mW	73 mW	3.16 mW	2.82 mW	128 mW
DC output power	0.84 mW	15 mW	1.02 mW	0.28 mW	41.7 mW
Rectifying efficiency (η_{rec})	37%	20.5	32.3%	10%	32.5%

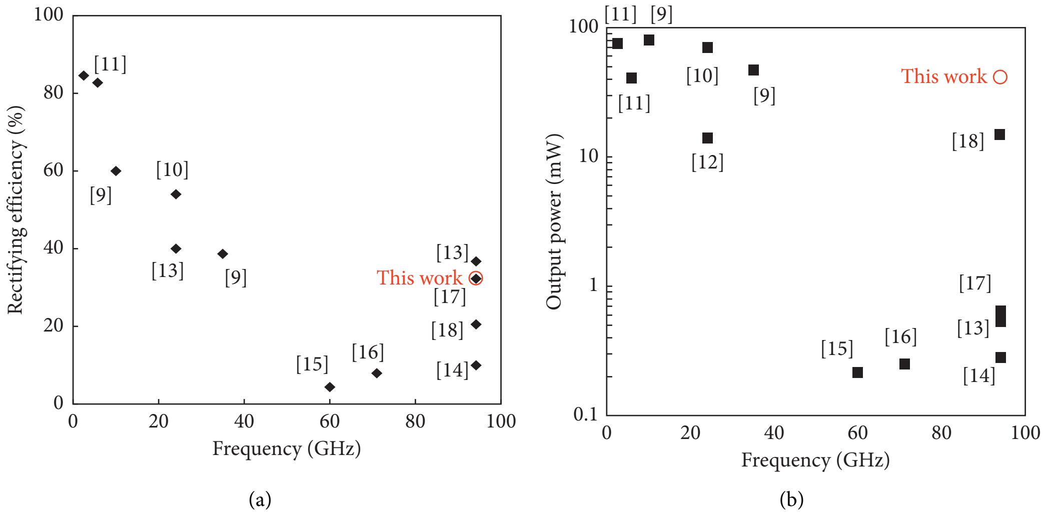


FIGURE 8: Comparison with other works from 2.45 GHz to 94 GHz: maximum rectifying efficiency vs. frequency (a); output DC power vs. maximum rectifying efficiency (b).

finline, which has transmission efficiency of 85% from measurements. Variable resistance is used as the load resistance. The output DC voltage across the resistance is measured using a voltmeter. Using equation (4), the rectifying efficiency (η_{rec}) can be calculated. The measured DC voltage (V_{out}), load resistance (R_{load}), measured power at the detector (P_{det}), finline transmission efficiency (η_{fin}), transmission efficiency of main direction (η_{ML}), and transmission efficiency of coupled direction (η_{CP}) are included in equation (4). In addition, η_{ML} and η_{CP} are measured by VNA in advance, as measured in the earlier part. Figure 7 shows η_{rec} and P_{DC} vs. P_{in}. When the load resistance is 150 Ω at input power of 128 mW, the rectifying efficiency was obtained as 32.5%. Under this condition, the output DC power was 41.7 mW. This value is approximately equal to the result obtained for the rectenna in an earlier study [19], which is 27.4%. Therefore, measurement of 94 GHz rectifiers under high power input becomes more accurate through this study.

Results obtained from this study are presented in Table 4 along with those of earlier studies that have examined this subject. At the same frequency, the efficiency was the second highest in the world; the output was found to be the highest value ever reported. Actually, the wavelength at 94 GHz is short, which can facilitate the production of small circuits. Furthermore, our highest output can engender great benefits for application to MAVs. Comparison of this work with other studies of lower frequency indicates that 94 GHz rectifiers show lower performance because of the large loss at the diode and the transmission line (Figure 8):

$$\eta_{\text{rec}} = \frac{P_{\text{DC}}}{P_{\text{in}}} = \frac{V_{\text{DC}}^2}{P_{\text{in}} R_{\text{load}}} = \frac{V_{\text{DC}}^2}{\eta_{\text{ML}} \eta_{\text{fin}} P_{\text{o}} R_{\text{load}}} = \frac{\eta_{\text{CP}} V_{\text{DC}}^2}{\eta_{\text{ML}} \eta_{\text{fin}} P_{\text{det}} R_{\text{load}}}. \tag{4}$$

4. Conclusions

We developed a 100 mW-class 94 GHz single-series rectifier aimed at MAV application using MMW. Its rectifying efficiency was evaluated before integration by fabricating FMF samples using a finline, for which the transmission efficiency was inferred as 85%. When the input RF power was 128 mW, the output DC power and rectifying efficiency of the rectifier (η_{rec}) were obtained respectively as 41.7 mW and 32.5%. The output power was the world's highest value reported to date: 94 GHz. Comparison of results of this study to an earlier study revealed that the measurement of 94 GHz rectifiers under high power input gives better accuracy.

References

[1] G. W. Jull, "Summary report on SHARP (stationary high-altitude relay platform) part A—technical feasibility of microwave-powered airplanes," CRC Report No. 1393, Communications Research Centre, Ottawa, Canada, 1985.

[2] H. Matsumoto, N. Kaya, M. Fujita, T. Fujiwara, and T. Sato, "Microwave lifted airplane experiment with active phased array antennas," MILAX Report, Kyoto University, Kyoto, Japan, 1995.

[3] S. Komatsu, K. Katsunaga, R. Ozawa, K. Komurasaki, and Y. Arakawa, "Power transmission to a micro aerial vehicle," in *Proceedings of the 45th AIAA Aerospace Sciences Meeting and Exhibit (AIAA-Paper 2007-1003)*, Reno, Nevada, January 2007.

[4] S. Nako, K. Okuda, K. Miyashiro, K. Komurasaki, and H. Koizumi, "Wireless power transfer to a microaerial vehicle with a microwave active phased array," *International Journal of Antennas and Propagation*, vol. 2014, Article ID 374543, 5 pages, 2014.

[5] N. Shinohara, "Beam efficiency of wireless power transmission via radio waves from short range to long range," *Journal of Electromagnetic Engineering and Science*, vol. 10, no. 4, pp. 224–230, 2010.

[6] K. Shimamura, H. Sawahara, A. Oda et al., "Feasibility study of microwave wireless powered flight for micro air vehicles," *Wireless Power Transfer*, vol. 4, no. 2, pp. 146–159, 2017.

[7] C.-J. Peng, S.-F. Yang, A.-C. Huang, T.-H. Huang, P.-J. Chung, and F.-M. Wu, "Harmonic enhanced location detection technique for energy harvesting receiver with resonator coupling design," in *Proceedings of the IEEE Wireless Power Transfer Conference (WPTC)*, pp. 1–3, Taipei, Taiwan, May 2017.

[8] H. Zhang, Y.-X. Guo, S.-P. Gao, and W. Wu, "Wireless power transfer antenna alignment using third harmonic," *IEEE Microwave and Wireless Components Letters*, vol. 28, no. 6, pp. 536–538, 2018.

[9] T.-W. Yoo and K. Chang, "Theoretical and experimental development of 10 and 35 GHz rectennas," *IEEE Transactions on Microwave Theory and Techniques*, vol. 40, no. 6, pp. 1259–1266, 1992.

[10] K. Hatano, "Development of 24 GHz-band MMIC rectenna,"vol. 50, pp. 199–201, in *Proceedings of the Radio And Wireless Symposium (RWS)*, vol. 50, pp. 199–201, IEEE, Austin, TX, USA, January 2013.

[11] Y. H. Suh and K. Chang, "A high efficiency dual-frequency rectenna for 2.45- and 5.8-GHz wireless power transmission," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 7, pp. 1784–1789, 2002.

[12] S. Ladan, S. Hemour, and K. Wu, "Towards millimeter-wave high-efficiency rectification for wireless energy harvesting," in *Proceedings of the 2013 IEEE International Wireless Symposium (IWS)*, pp. 7–10, Beijing, China, April 2013.

[13] H.-K. Chiou and I.-S. Chen, "High efficiency dual-band on-chip rectenna for 35- and 94-GHz wireless power transmission in 0.13-μm CMOS technology," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 12, pp. 3598–3606, 2010.

[14] N. Weissman, S. Jameson, and E. Socher, "W-band CMOS on-chip energy harvester and rectenna," in *Proceedings of the 2014 IEEE MTT-S International Microwave Symposium (IMS 2014)*, pp. 1–3, Tampa, FL, USA, June 2014.

[15] H. Gao, U. Johannsen, M. K. Matters-Kammerer et al., "A 60-GHz rectenna for monolithic wireless sensor tags," in

Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS 2013), pp. 2796–2799, Beijing, China, May 2013.

[16] H. Gao, M. K. Matters-Kamrnerer, P. Harpe et al., “A 71 GHz RF energy harvesting tag with 8% efficiency for wireless temperature sensors in 65 nm CMOS,” in *Proceedings of the 2013 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pp. 403–406, Seattle, WA, USA, June 2013.

[17] S. Hemour, C. H. P. Lorenz, and K. Wu, “Small-footprint wideband 94 GHz rectifier for swarm micro-robotics,” in *Proceedings of the 2015 IEEE MTT-S International Microwave Symposium*, pp. 5–8, Phoenix, AZ, USA, May 2015.

[18] A. Etinger, M. Pilossof, B. Litvak et al., “Characterization of a Schottky diode rectenna for millimeter wave power beaming using high power radiation sources,” in *Proceedings of the 12th Symposium of Magnetic Measurements and Modeling SMMM’ 2016*, Częstochowa-Siewierz, Poland, October 2016.

[19] K. Matsui, K. Fujiwara, Y. Okamoto et al., “Development of 94 GHz microstrip line rectenna,” in *Proceedings of the 2018 IEEE Wireless Power Transfer Conference (WPTC)*, pp. 1–4, Montreal, Canada, June 2018.

[20] K. Hatano, N. Shinohara, and T. Mitani, “Improvement of 24 GHz-band class-F load rectennas,” in *Proceedings of the 2012 IEEE MTT-S International Microwave Workshop Series on Innovative Wireless Power Transmission: Technologies, Systems, and Applications*, pp. 7–10, Kyoto, Japan, May 2012.

[21] J. H. C. van Heuven, “A new integrated waveguide-microstrip transition (short papers),” *IEEE Transactions on Microwave Theory and Techniques*, vol. 24, no. 3, pp. 144–147, 1976.

[22] K. Fujiwara and T. Kobayashi, “Low-cost W-band frequency converter with broad-band waveguide-to-microstrip transducer,” in *Proceedings of the 2016 Global Symposium on Millimeter Waves (GSMM) & ESA Workshop on Millimetre Wave Technology and Applications*, pp. 1–4, Espoo, Finland, June 2016.

[23] Japanese Patent Application No. 2016243600.

Improved TLBO for Fusion of Infrared and Visible Images

Chinmaya Ranjan Pradhan, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar,r.pradhan23@gmail.com*

Smruti Ranjan Panda, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, sr_panda@outlook.com*

Sangita Pal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sangitapal2@outlook.com*

Manoj Mohanta, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, manoj.mohanta62@outlook.com*

Abstract

Image fusion is an image enhancement method in modern artificial intelligence theory, which can reduce the pressure in data storage and obtain better image information. Due to different imaging principles, information of the infrared image and visible images' information is complementary and redundant. The infrared image can be fused with a visible image to obtain both the high-resolution texture details and the edge contour of the infrared image. In this paper, the fusion algorithm of forest sample image is studied at the feature level, which aims to accurately extract tree features through information fusion, ensure data stability and reliability, and improve the accuracy of target recognition. The main research contents of this paper are as follows: (1) teaching learning-based optimization (TLBO) algorithm was used to optimize the weighted coefficient in the fusion process, and the value range of random parameters in the model was adjusted to optimize the fusion effect. Compared with before optimization, image information increased by 2.05%, and spatial activity increased by 15.27%. (2) Experimental data show that the target recognition accuracy of feature-level fusion results was 93.6%, 13.9% higher than that of the original infrared sample image, and 18.8% higher than that of the original visible sample image. Pixel-level and feature-level fusion have their characteristics and application scopes. This method can improve the quality of the specified region in the image and is suitable for detecting intelligent information in forest regions.

1. Introduction

With the rapid development of sensor technology, single visible light mode is gradually developed into a variety of sensor modes. They differ in imagining mechanism, working environment, and requirements as well as functions. They also work in different wavelength ranges. Due to the limited information of data acquired by a single sensor, it is often difficult to meet the needs of applications. At the same time, more comprehensive and reliable information of observation targets can be obtained by using multisource data. Therefore, in order to take full advantage of increasingly complex source data, various data fusion techniques have been rapidly developed with the aim of incorporating more supplementary information into a new data set by means of more information than can be obtained from any single sensor [1]. Image fusion technology, as a very important branch of multisensor and visual information fusion, has aroused widespread concern and research upsurge in the world in the past twenty years. The main idea of image fusion is to combine multisource images from multiple sensors into a new image by using algorithms, so that the fused image has higher reliability, less uncertainty, and better comprehensibility [2].

Image fusion technology was first used in remote sensing image analysis and processing. In 1979, Daily et al. first applied the composite image of radar image and Landsat-MSS image to geological interpretation, and its processing process can be regarded as the simplest image fusion [3]. In 1981, Laner and Todd conducted a fusion experiment of Landsat-RBV and MSS image information [4]. In the middle and late 1980s, image fusion technology has been applied to the analysis and processing of remote sensing multispectral images, beginning to attract attention. It was not until the end of 1980 that people began to apply image fusion technology to general image processing (visible image, infrared image, etc.) [5]. Since the 1990s, the research of image fusion technology has been on the rise, showing great application potential in the fields of automatic recognition, computer vision, remote sensing, robotics, medical image processing, and military applications. For example, the fusion of infrared and low-light images helps soldiers see targets in the dark [6]. The fusion of CT and MRI images is helpful for doctors to diagnose diseases accurately [7]. Jin et al. extracted more accurate and reliable feature information from images by fusion of infrared and visible images, thus achieving accurate face recognition [8]. Using image fusion, Liu et al. made images with different focal lengths complement each other and improve the resolution of fusion results [9]. In recent years, image fusion has become an important and useful technique for image analysis and computer vision.

The main purpose of this paper is to find an image fusion algorithm suitable for forest environment perception, using visible light image and infrared thermal image fusion technology, to collect the image fusion processing, improve the fusion effect, accurately extract effective forest information, and obtain information for forest intelligent detection. The main research contents are as follows:

(1) The fusion background of visible and infrared images, different image processing methods, and the effects of different image fusion processing are introduced

(2) The process of fusion coefficient optimization based on teaching learning based optimization (TLBO) algorithm is introduced. The random parameters in the model are set by TLBO optimization algorithm to optimize the fusion effect. The forest images are used for image fusion experiments, and the fusion results are evaluated by objective evaluation indexes

(3) In order to enhance the search ability of the algorithm and improve the evaluation index value to a greater extent, the value range of the optimization coefficient Ri and T_f of TLBO algorithm is further set according to the entropy value, and then evaluation index is used for corresponding evaluation

2. Related Works

Multisource image fusion algorithm also has broad application prospects in the field of forestry intelligent detection. Using feature-level image fusion algorithm, Bulanona et al. extracted data information of fruits in fruit forests and monitored fruit growth status in real time in 2009 [10]. In 2013, Lei et al. identified obstacles in forest images by using the results obtained by fusion algorithm and two-dimensional laser data and intelligently and accurately distinguished trees, rocks, and animals in the images with an accuracy rate of more than 93.3% [11]. Furthermore, by improving the fusion algorithm, the data accuracy of objects such as trees in the image is improved, and the accuracy of target recognition is increased by 95.3% [12]. The quality of information fusion directly affects the accuracy of forest information detection and is an important part of research on artificial intelligence. This paper is an important branch of research on information fusion algorithm-infrared and visible image fusion algorithm. Due to different imaging principles, the information of infrared image and visible image is complementary and redundant. The target in infrared image has clear edge features and is easy to be segmented and extracted. The texture details and background information of visible image are more prominent, but the target information is difficult to extract because of complex image content. Therefore, the purpose of fusion is to synthesize complementary information, reduce redundancy, improve image quality, and express and extract useful features in images more succinctly and accurately. In this paper, the effective forest information is extracted accurately by fusion of infrared and visible images, and the obtained information is used for intelligent forest detection.

3. Materials and Methods

The detailed process of visible and near-infrared image sample fusion is shown in Figure 1. The ultimate goal is to enhance the search ability of the algorithm, improve the evaluation index value to a greater extent, and obtain the fusion image more suitable for the intelligent detection of forest information.

3.1. Data Modeling Methods

3.1.1. Pixel Level Image Fusion Algorithm

(1) Image Fusion Algorithm Based on Wavelet Transform. Wavelet transform theory was first proposed by Morlet and Gorsmsna in 1984, and its principle is developed on the basis of Fourier transform. Different from Fourier transform, wavelet transform is the local transform of frequency, which can effectively extract the signal in the image. Its advantage is to carry out multiscale analysis of the image without losing the information [13]. Stephane and Matllat proposed fast discrete wavelet and built a bridge between wavelet transform and multiscale image fusion [14].

Two-dimensional image samples after wavelet decomposition can be represented by four subband components:

$$f(x,y) = A_j f + D_j^1 f + Dj_1^2 f + D_j^3 f, \tag{1}$$

$$A_j f = \langle f(x,y), \varnothing_{j,m,n}(x,y) \rangle, \tag{2}$$

$$D_j^1 f = \left\langle f(x,y), \psi_{j,m,n}^1(x,y) \right\rangle, \tag{3}$$

$$D_j^2 f = \left\langle f(x,y), \psi_{j,m,n}^2(x,y) \right\rangle, \tag{4}$$

$$D_j^3 f = \left\langle f(x,y), \psi_{j,m,n}^3(x,y) \right\rangle, \tag{5}$$

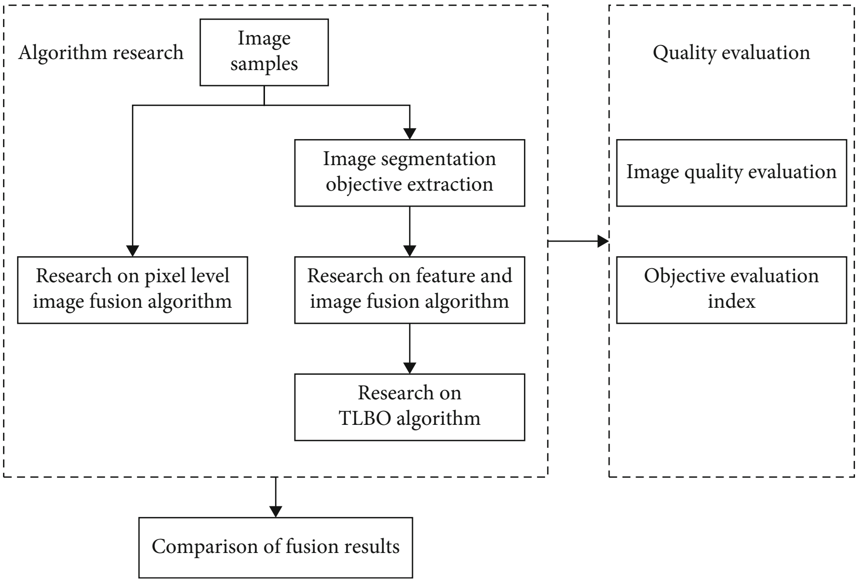

Algorithm research
Image samples
Image segmentation objective extraction
Research on pixel level image fusion algorithm
Research on feature and image fusion algorithm
Research on TLBO algorithm
Quality evaluation
Image quality evaluation
Objective evaluation index
Comparison of fusion results

FIGURE 1: Flowchart for research on infrared and visible image fusion algorithm.

where f represents the original sample, j represents the decomposition frequency of this layer, A represents the low-frequency component, and D represents the high-frequency component in different directions. $\varnothing_{j,m,n}$ is the scale coefficient that makes up the canonical orthogonal basis of the wavelet, and the wavelet function $\psi_{j,m,n}$ makes up the canonical orthogonal basis of the space.

(2) Image Fusion Algorithm Based on PCA Transform. Principal component analysis (PCA) transform, also known as principal component analysis, is a multidimensional linear transform based on the statistical characteristics of images, which has the function of centralizing variance information and compressing data volume is mathematically called $K-L$ transform.

The PCA transformation and fusion process of multisensor images is as follows:

(1) PCA was applied to the low-resolution multispectral image to obtain three principal components: P1, P2, and P3

(2) The high-resolution image was stretched and made to have the same mean and variance as the first principal component P1 of the multispectral image

(3) The stretched high-resolution image was used to replace P1 as the first principal component, and a new fusion image P was generated with components P2 and P3 through PCA inverse transformation

(3) Image Fusion Algorithm Based on Contourlet Transform. Contourlet transform, also called contourlet transform, is a multiresolution image representation method proposed by Do and Vetterli in 2002 [15]. In contourlet transform, image multiscale decomposition is realized by Laplace tower decomposition (LP) [16]. The multiscale decomposition of an approximate image can be obtained by repeated Laplacian tower decomposition [17]. However, in the process of image decomposition and reconstruction by contourlet transform, the image needs to be further sampled and upward sampled, which makes the contourlet transform lack shift-invariance (invariance) [18]. As a result, the spectrum of the signal will overlap to some extent, and the Gibbs phenomenon is obvious in the image fusion.

(4) Low-Frequency Coefficient Processing Based on PCNN. In the low-frequency domain of image fusion, Laplacian energy, as excitation input to PCNN, is processed as follows:

$$\begin{aligned}\mathrm{ML}(i,j) = {} & |2I(i,j) - I(i-\mathrm{step},j) - I(i+\mathrm{step},j)| \\ & + |2I(i,j) - I(i,j-\mathrm{step}) - I(i,j+\mathrm{step})| \\ & + |2I(i,j) - I(i-\mathrm{step},j-\mathrm{step}) - I(i+\mathrm{step},j+\mathrm{step})| \\ & + |2I(i,j) - I(i-\mathrm{step},j+\mathrm{step}) - I(i+\mathrm{step},j+\mathrm{step})|,\end{aligned} \tag{6}$$

where step represents the variable distance between the coefficients (in this paper, $\mathrm{step}=1$); $I(i,j)$ is the coefficient at point (i,j).

In order to eliminate the block effect or grayscale distortion that may be caused by the boundary discontinuity at the junction between the clear area and the fuzzy area, the sum of modified Laplacian (SML) in the field centered on point (i,j) is defined as

$$\mathrm{SML}(i,j) = \sum_p \sum_q W(p,q)[\mathrm{ML}(i+p,j+q)]^2, \tag{7}$$

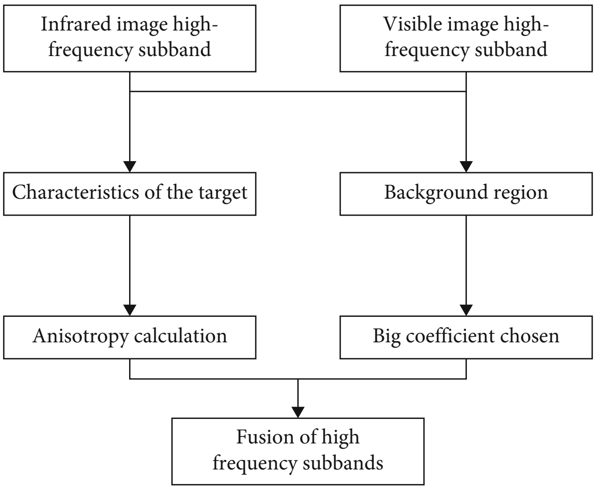

Infrared image high-frequency subband
Visible image high-frequency subband
Characteristics of the target
Background region
Anisotropy calculation
Big coefficient chosen
Fusion of high frequency subbands

Figure 2: Flowchart for high-frequency fusion.

where $W(p, q)$ is the corresponding window function. Experience shows that the best highlighting effect of the window center pixel and its changing boundary should be set as

$$W(p, q) = \frac{1}{15}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{bmatrix}. \tag{8}$$

The sum of modified Laplacian can well represent the edge details of the image, reflect the sharpness of the image, and show superior fusion performance in the fusion image.

(5) High-Frequency Coefficient Processing Based on PCNN. The high-frequency subband image represents the edge details of the image, so the coefficients decomposed by NSCT can be directly input into PCNN as excitation in the process of high-frequency coefficient fusion. The specific steps are as follows:

The PCA transformation and fusion process of multisensor images is as follows:

(1) In high-frequency subband images, the normalized gray values of each pixel are directly taken as the external input of PCNN to calculate the ignition times of each input excitation. The formula is expressed as

$$T_{ij}^k(n) = T_{ij}^k(n-1)Y_{ij}^k(n). \tag{9}$$

(2) The processing steps for the same low-frequency coefficient are as follows:

$$D_{ij,F}^k(N_1) = \begin{cases} 1, & \text{if} : T_{ij,A}^k(N_1) \geq T_{ij,B}^k(N_1), \\ 0, & \text{others}, \end{cases} \tag{10}$$

$$I_F^k(i, j) = \begin{cases} I_A^k(i, j), & \text{if} : D_{ij,F}^k(N_1) = 1, \\ I_B^k(i, j), & \text{if} : D_{ij,F}^k(N_1) = 1, \end{cases} \tag{11}$$

where $I_F^k(i, j)$, $I_A^k(i, j)$, and $I_B^k(i, j)$, respectively, represent the gray values of the fusion image and the original image A and $B(i, j)$, and k represents the NSCT decomposition of the k -layer. After each fusion subband image is obtained, the fusion image is obtained by NSCT inverse transformation.

3.1.2. Feature-Level Fusion Algorithm

(1) Low-Frequency Domain Fusion Rule Based on Fuzzy Logic. Based on fuzzy rules, fusion can also be divided into two types: spatial domain fusion and frequency domain fusion. Teng et al. fuzzified all pixel points into five fuzzy subsets based on the gray value of the image, then determined the membership degree of each fuzzy subset in the corresponding domain by a triangular membership function, and formulated fusion rules on this basis to obtain fusion results [19]. Cai and Wei first decomposed the source image into the frequency domain and then formulated the fusion rules in the low-frequency domain by using the fuzzy logic criterion to maximize the information content of the fusion sub-band image in the low-frequency domain [20]. In this paper, Gaussian membership function is used to determine the weight coefficient of image fusion, whose definition is expressed as

$$\lambda_1(i, j) = \exp\left[-\frac{(f_1(i, j) - \mu)^2}{2(k\sigma)^2}\right], \tag{12}$$

where σ is the standard deviation of the sub-band image, $f_1(\boxtimes, j)$ is the low-frequency decomposition coefficient of point (i, j), μ is the average value of the decomposition coefficient, and k is a constant.

(2) High-Frequency Domain Fusion Rules Based on Segmentation Results. The role of high-frequency fusion rules is to solve the problem that the target is not significant in the pixel-level fusion results and then improve the texture energy and other features of trees in the fusion results, so that the tree features extracted from the fusion results are more accurate. The flow chart is shown in Figure 2.

(3) Fusion Coefficient Optimization Algorithm. Teaching-learning-based optimization algorithm (TLBO) is a swarm intelligence algorithm proposed by Rao et al., an Indian scholar, in 2011 [21]. It imitates the learning process model of students and can be divided into two parts: teaching and learning phases. In 2014, Jin and Wang first applied TLBO optimization algorithm to image fusion to optimize the fusion coefficient and improve the image quality evaluation index [22].

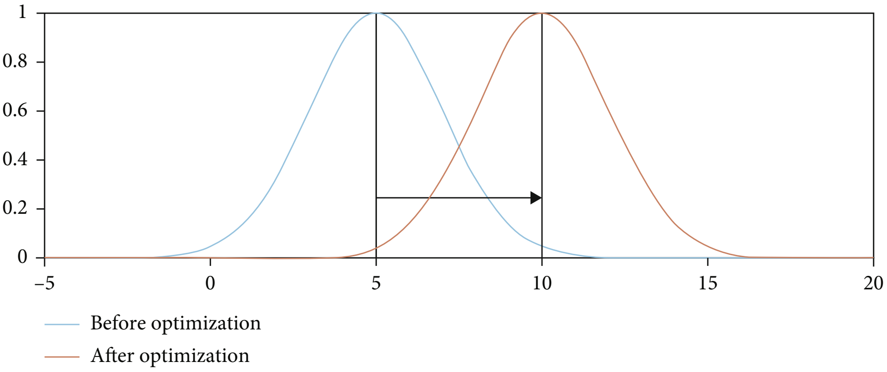


Figure 3: Schematic diagram of optimization process.

(1) Teaching phase

In the teaching phase, the overall optimization can be achieved by encouraging top students. Figure 3 shows the schematic diagram of the overall optimization process in this phase.

As shown in the curve for a group of student's overall academic record in Figure 3, the result agrees with the normal distribution, with its average representing students' overall level. In each optimization process, the teacher scored the best by students is defined, and then the level of the teacher is further optimized in the overall level of students through their influence.

(2) Learning phase

In the learning phase, the target function index can be improved through mutual learning among individuals. The process is carried out according to the following rules:

$$\begin{aligned} &\text{if } f(x_i) < f(x_j), && x_{\text{new},i} = x_{\text{old},i} + r_i(x_i - x_j), \\ &\text{else}, && x_{\text{new},i} = x_{\text{old},i} + r_j(x_j - x_i). \end{aligned} \tag{13}$$

Compared with PSO, GA, and other optimization algorithms, the coefficient in TLBO has less influence on the optimization effect, with better convergence but requires shorter optimization time.

The detailed optimization steps are as follows:

(a) The weight coefficients determined by Gaussian membership function during image fusion were converted into a row of vectors, which were used as a group of samples. Another 9 groups of vectors with the same size were randomly generated to form the model to be tested

(b) The entropy value of fusion image was selected as the objective function

(c) The model was put into the TLBO system, and the fusion coefficient group under the optimal entropy value was obtained through the cycle until the convergence of objective function

(d) The cycle was terminated

(4) Improved TLBO Parameter Optimization Algorithm. The basic TLBO algorithm can find the global optimal value when solving simple low-dimensional problems, but when solving complex multimode high-dimensional problems, it is easy to fall into the local optimal value and cannot find the values adjacent to the global optimal value. Many scholars have improved the TLBO algorithm. Rao et al. supplemented and improved the structure of TLBO algorithm. Gao et al. introduced the crossover operation of differential evolution algorithm into the algorithm to further improve the local search ability of the algorithm [23]. All the indexes of the image optimized by using the basic TLBO algorithm were improved but not quite significantly. Therefore, in order to enhance the search ability of the algorithm and improve the evaluation index value to a greater extent, the value range of optimization coefficients R_i and T_f of TLBO algorithm was further adjusted.

The detailed optimization steps of the improved TLBO are as follows:

(1) The weight coefficients determined by Gaussian membership function during image fusion were converted into a row of vectors, which are used as a group of samples. Another 9 groups of vectors with the same size were randomly generated to form the model to be tested

(2) The entropy value of fusion image was selected as the objective function

(3) The value range of T_f was kept unchanged, the range of parameter R_i was set to compare the influence of R_i in different ranges on the image entropy, and then the optimal R_i was selected

(4) The model was put into the TLBO system, and the fusion coefficient group under the optimal entropy value was obtained through the cycle until the convergence of the objective function

(5) The R_i value range was kept unchanged, the parameter T_f range was set to compare the influence of T_f

Figure 4: Fluke Ti55 infrared thermal imaging camera.

in different ranges on the image entropy value, and then the optimal T_f was selected

(6) The model was then brought into the TLBO system, and the fusion coefficient group under the optimal entropy value was obtained through the cycle until the convergence of objective function

(7) The cycle was terminated

3.1.3. The Evaluation Index. In this paper, information entropy, mean gradient, standard deviation, spatial resolution, and interactive information are selected as image evaluation indicators.

(1) Information Entropy. Information entropy is the most widely used objective evaluation index of images at present, which quantitatively describes the information contained in images, and its mathematical definition is expressed as

$$E = -\sum_{i=0}^{255} P_i \log P_i, \tag{14}$$

where E represents information entropy, and P represents the proportion of the number of pixels with gray value of I in the total pixel points. The larger the information entropy is, the more scattered the gray value of image pixels is, the richer the content is, the larger the information is, and the better the fusion effect is.

(2) Average Gradient. The average gradient reflects the difference between adjacent pixels in the image. The larger the average gradient is, the greater the image contrast is, the more obvious the edge effect of objects in the image is, and the clearer the texture details are. The mean gradient is defined as

$$\overline{\text{grad}} = \frac{1}{(m-1)(n-1)} \sum_{m-1}\sum_{n-1} \sqrt{\frac{(F(i,j)-F(i+1,j))^2 + (F(i,j)-F(i,j+1))^2}{2}}, \tag{15}$$

where $\overline{\text{grad}}$ represents the average gradient; m and n are the size of the image; $F(i,j)$ represents the pixel gray value of coordinate (i,j).

(3) Standard Deviation. Standard deviation represents the dispersion degree of pixel gray value distribution. The larger the value is, the more discrete the gray value distribution of image pixels is, and the stronger the contrast. The mathematical expression of standard deviation is defined as

$$\text{std} = \sqrt{\frac{\sum\left(F(i,j)-\bar{F}\right)^2}{(n-1)}}, \tag{16}$$

where STD stands for standard deviation, $F(i,j)$ stands for pixel value at point (i,j), and $\bar{F}$ stands for pixel mean of all pixel points.

(4) Spatial Resolution. Spatial frequency is a parameter used to represent the activity degree of images in space. The higher the value is, the higher the activity degree of images in space is and the better the quality of images is. The formula of spatial frequency is expressed as

$$\text{RF} = \sqrt{\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=2}^{N}[F(i,j)-F(i,j-1)]^2}, \tag{17}$$

$$\text{CF} = \sqrt{\frac{1}{M\times N}\sum_{i=2}^{M}\sum_{j=1}^{N}[F(i,j)-F(i-1,j)]^2}, \tag{18}$$

$$\text{SF} = \sqrt{\text{RF}^2 + \text{CF}^2}, \tag{19}$$

where SF is spatial frequency, RF and CF represent spatial column frequency and spatial row frequency, respectively. M, N represent the number of rows and columns of the image, and $F(i,j)$ is the gray value of pixel point (i,j).

(5) Interactive Information. Interactive information, also known as mutual information, is usually used to demonstrate the correlation between multiple variables. In the image quality evaluation system, it is used to evaluate the correlation between fusion results and original samples. The greater the amount of interaction information, the higher the correlation between the fusion result and the original sample, and the more information can be obtained from the original sample:

$$\text{MI}_{\text{AF}} = E(A) + E(F) - E(\text{AF}), \tag{20}$$

$$\text{MI}_{\text{BF}} = E(B) + E(F) - E(\text{BF}), \tag{21}$$

$$\text{MI} = \text{MI}_{\text{AF}} + \text{MI}_{\text{BF}}, \tag{22}$$

where MI is the interactive information, A, B represents the original sample, F is the fusion result, and E is the image entropy value.

3.2. Data Analysis Materials. Nearly 400 groups of forest infrared and visible images were collected in this study. The equipment used is Fluke TI55 infrared thermal imager. The time period selected in this experiment is the morning and evening when the temperature difference is large, and

Table 1: Technical parameters of Fluke Ti55.

	Visible lens	Infrared lens
Detector type	1280∗1024 full color pixel	320∗240 focal plane array
Calibration temperature range	-20~600°C	-20~600°C
Visual angle	—	23°∗ 17°
Spatial resolution	0.47 mrad	1.30 mrad
Minimum focus	0.6 m	0.15 m
Accuracy	2%	—
NETD	—	≤0.05°C
Spectral band	—	8 ~ 14 μm
Detector type	1280∗1024 full color pixel	320∗240 focal plane Array
Calibration temperature range	-20~600°C	-20~600°C

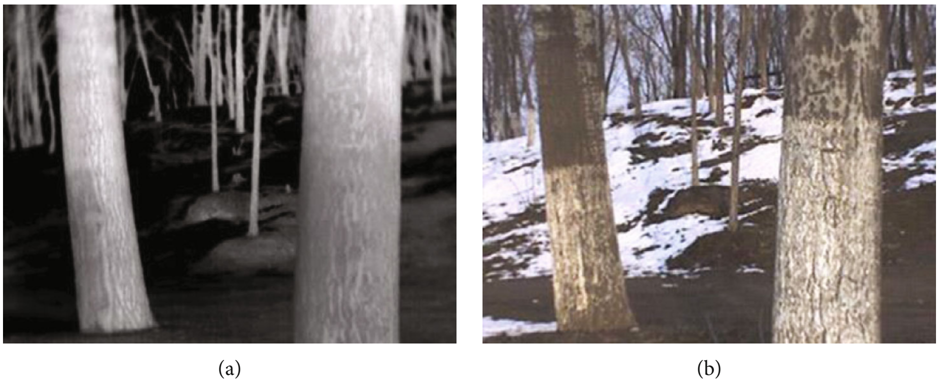


Figure 5: Examples of infrared image and visible image ((a) infrared image; (b) visible image).

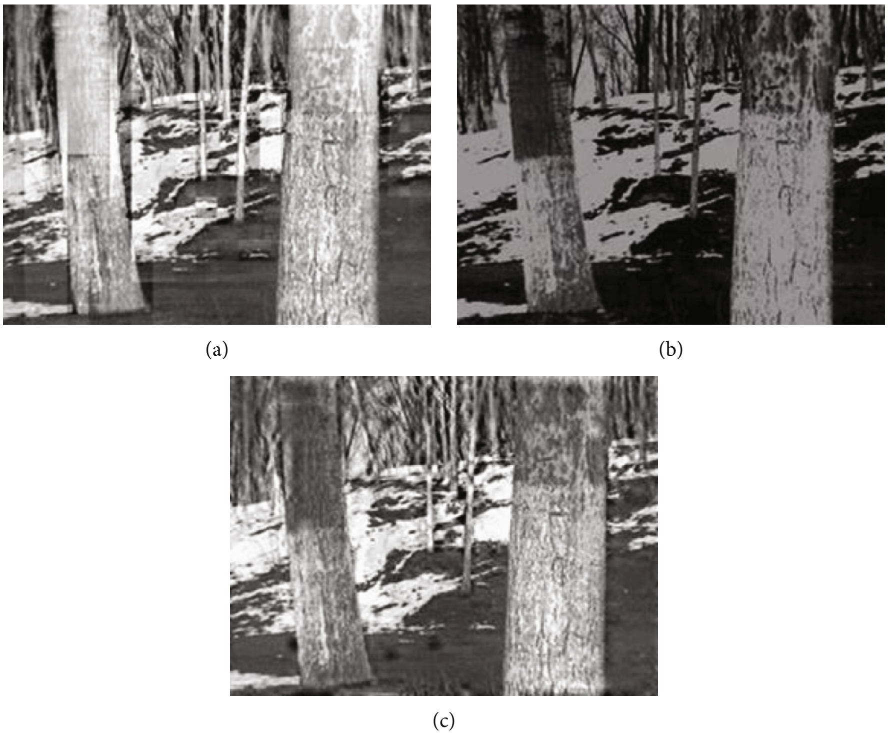


Figure 6: Pixel level fused result ((a) wavelet transform; (b) PCA; (c) contourlet and PCNN).

Table 2: Quality assessment of pixel-level fusion result.

	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Wavelet transform	7.5953	62.1048	19.8465	28.9314	5.4801
PCA transform	6.8579	55.7364	16.6849	23.3580	5.2299
Contourlet +PCNN	7.6367	56.3843	18.6997	29.2188	5.5820

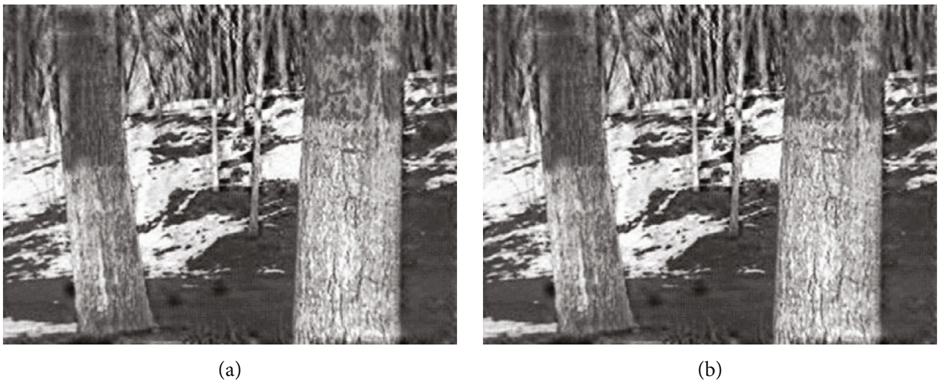


Figure 7: Fusion result of pixel level and region-based level ((a) pixel-level; (b) region-based level).

Table 3: Quality assessment of pixel-level and region-based level fusion results.

	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Pixel-level image fusion results	7.6367	56.3843	18.6997	29.2188	5.5820
Feature level image fusion results	7.3981	46.2270	15.4640	29.9097	5.8198

Table 4: Quality assessment of region-based level and TLBO optimization fusion results based on image samples.

	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Region-based level	7.3981	46.2270	15.4640	29.9097	5.8198
TLBO optimization	7.5121	54.9715	17.4434	33.1374	6.0479

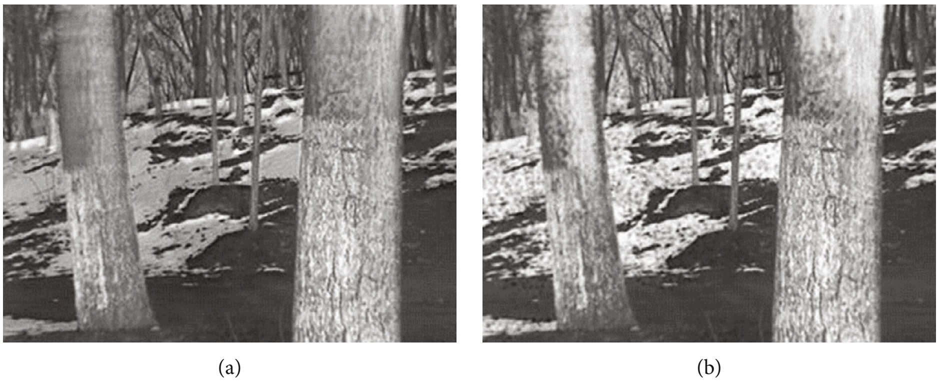


Figure 8: Fusion result of region-based level and TLBO optimization based on image samples ((a) region-based level; (b) TLBO optimization).

Table 5: Quality assessment of region-based level and TLBO optimization fusion results based on the other group images.

	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Region-based level	7.2378	38.5448	18.2632	31.9562	5.7235
TLBO optimization	7.3910	45.8079	18.8261	32.5719	5.2021

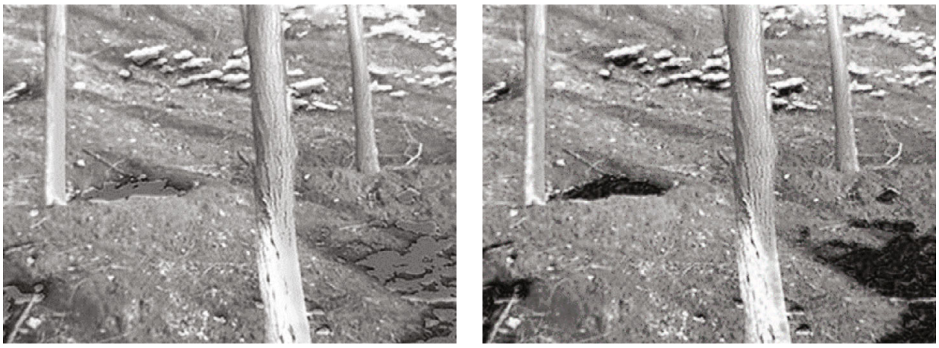


Figure 9: Fusion result of region-based level and TLBO optimization based on the other group images ((a) region-based level; (b) TLBO optimization).

the afternoon when the visual effect is easily affected. The image sample collection experiment and experimental equipment of this study are shown in Figure 4, and its technical parameters are shown in Table 1.

The infrared lens captures the spectral information in the 8-14 band, which is the middle and far infrared image. The contour of forest edge in infrared samples is obvious and thus easy to be segmented and extracted, but the accuracy of target recognition cannot be guaranteed due to the lack of details such as texture. Visible light samples contain rich texture details, but there is a very small gray difference between trees and background area without any pronounced characteristics, so it is difficult to achieve the stable extraction of tree information alone. Therefore, the purpose of this study is to improve image quality and accurately extract forest tree feature information by integrating the characteristics of infrared and visible images through fusion processing to ensure the accuracy of recognition.

4. Results

4.1. Pixel-Level Image Fusion Results. Figure 5 shows samples of infrared and visible forest images. Figure 6 shows the results of wavelet decomposition algorithm, PCA fusion algorithm, and contoulet combined with PCNN fusion algorithm, respectively. In wavelet transform, the image is decomposed into a low-frequency domain and a high-frequency domain in three directions, including horizontal high-frequency domain, vertical high-frequency domain, and oblique high-frequency domain. Wavelet decomposition can overcome the instability of Laplace decomposition and effectively reduce the influence of noise on the image. However, due to the defect of wavelet decomposition basis, jagged block error is likely to occur when processing smooth curves. PCA transformation of the principal component information is relatively high, using the gray value of panchromatic band image to replace PCA, and then inverse transformation of the enhanced multispectral band image, the information is vulnerable to loss. Contourlet decomposition +PCNN transform can avoid block effect and grayscale distortion while improving image definition and contrast, avoiding generating new noise and expanding the information of a single image.

From the perspective of subjective evaluation, the fusion results synthesize the features of the source image, expand the information content of a single image, and improve the image clarity. Based on contourlet decomposition, this algorithm is a multiscale and multidirection computing framework for discrete images. It can be regarded as the enhancement technology of contourlet decomposition, which can carry out multidirection decomposition and multiscale decomposition of images, respectively. By the improved method, contourlet decomposition +PCNN transform can eliminate the aliasing effect caused by using contourlet and provide a good and stable input signal for the subsequent fusion.

In order to quantitatively evaluate the quality of the fusion results, the image was quantitatively analyzed, as shown in Table 2. As can be seen from the table, the standard deviation and mean gradient data distortion of wavelet fusion are caused by the fact that wavelet transform cannot effectively process the smooth curve in the image, which is likely to result in the jagged noise and interferes with the statistical characteristics of the image. Contourlet decomposition +PCNN transform can avoid block effect and grayscale distortion, while improving image definition and contrast, avoiding generating new noise, and expanding the information of a single image. It can also more accurately describe the forest area of the tree information and its scene details. In terms of quantitative data analysis, the results obtained by contourlet combined with PCNN algorithm are better than those obtained by the other two algorithms

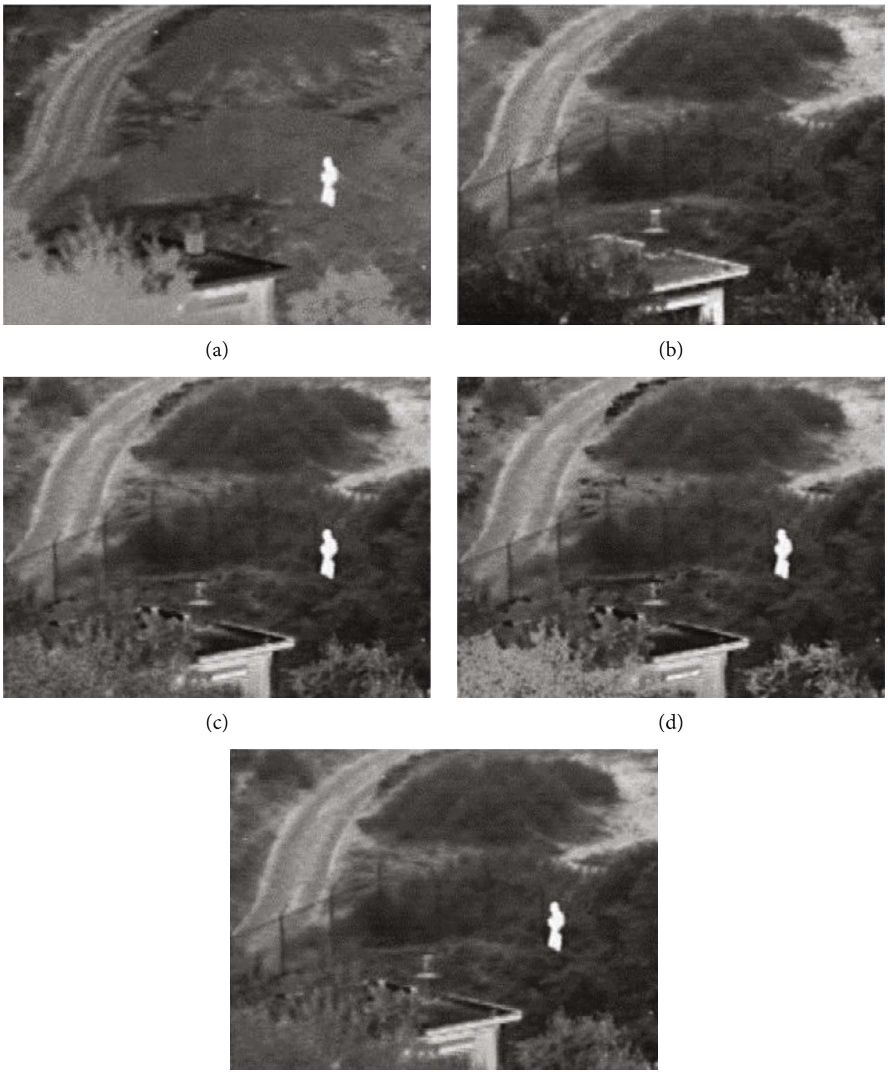


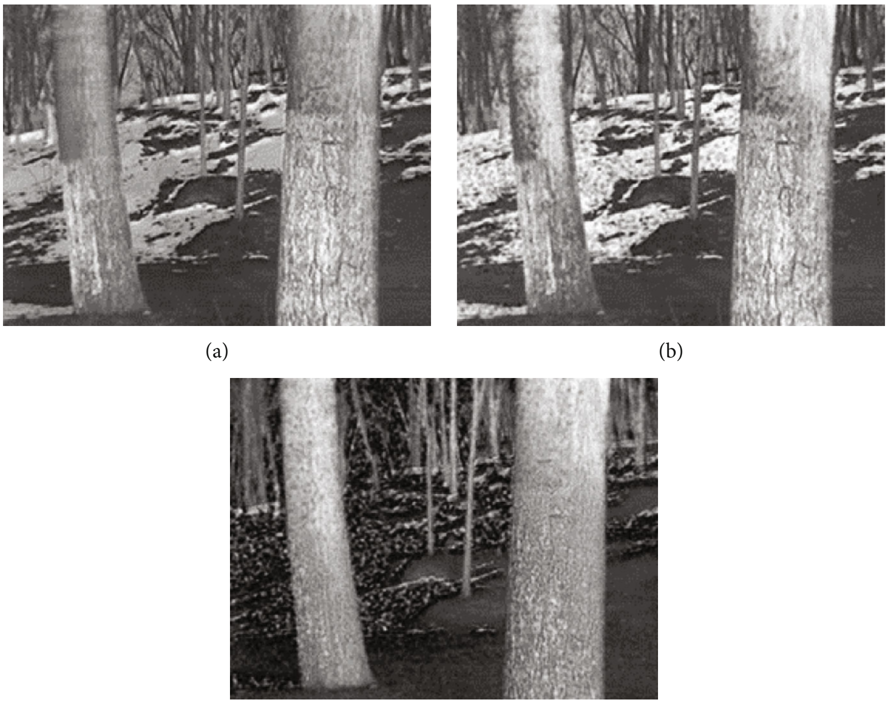
Figure 10: Comparison of fusion result based on UN-CAMP images ((a) infrared light sample; (b) visible light sample; (c) feature-level fusion; (d) original TLBO algorithm; (e) improved TLBO algorithm).

in entropy, spatial resolution, and interactive information, and they also have slightly lower standard deviation and mean gradient is slightly lower than the traditional algorithm, but higher than PCA.

Therefore, the fusion result of contourlet decomposition +PCNN transform has a larger amount of information, stronger contrast, and better visual effect. At the same time, it can effectively avoid grayscale distortion and block effect easily caused by the fusion between forest infrared and visible images while avoiding the influence of noises. Therefore, compared with common pixel-level fusion algorithms, this algorithm performs better in improving image information and sharpness.

4.2. Feature-Level Image Fusion Results. Compared with the pixel-level image fusion algorithm, the high-frequency domain fusion algorithm proposed in Figure 7 improves the visual effect of the image. In the forest image, the background area has little influence on the target area, which reduces the block effect and ringing effect behind. The contour is clearer and more information about the target area is retained. Compared with the pixel pole fusion image, the block effect is significantly reduced. As can be seen from Table 3, pixel-level fusion images have the best data in terms of entropy, mean gradient, standard deviation, and spatial frequency, indicating that pixel-level fusion images are better in terms of information content, contrast, and spatial activity. The spatial resolution and interactive information of feature-level fusion images are optimal, which indicates that the image is better than other fusion images in terms of the degree of association with the source image and noise interference prevention.

As can be seen from the table, the result obtained by the feature-level fusion algorithm has a larger amount of information and better visual effect. It effectively avoids grayscale distortion and block effect easily caused by the fusion of forest infrared and visible images while avoiding the influence of noises. Among different evaluation indexes, pixel-level



(a)

(b)

(c)

Figure 11: Comparison of fusion result based on image samples ((a) feature-level fusion; (b) original TLBO algorithm; (c) improved TLBO algorithm).

Table 6: Quality assessment of region-based level and improved TLBO optimization fusion results based on UN-CAMP images.

	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Region-based level	7.0048	35.2247	4.5529	10.0739	4.8237
TLBO optimization	7.0858	37.6909	5.3584	10.5275	4.8651
Improved TLBO optimization	7.1483	39.9000	5.8673	11.6199	4.9080

Table 7: Quality assessment of region-based level and improved TLBO optimization fusion results based on image samples.

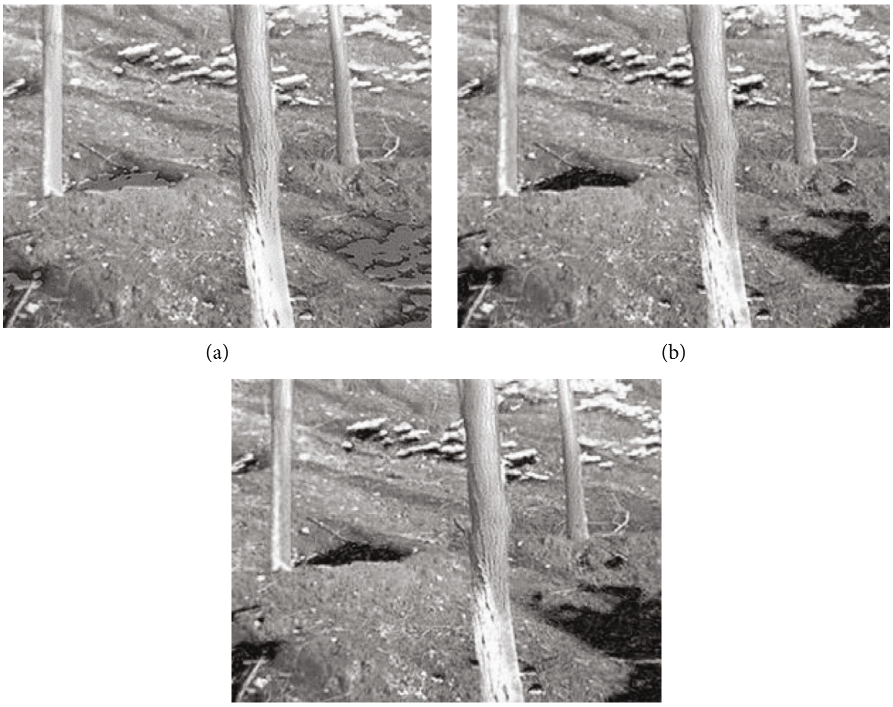
	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Region-based level	7.3981	46.2270	15.4640	29.9097	5.8198
TLBO optimization	7.5121	54.9715	17.4434	33.1374	6.0479
Improved TLBO optimization	7.6921	57.0801	17.6343	29.7725	5.5715

Table 8: Quality assessment of region-based level and improved TLBO optimization fusion results based on the other group images.

	Information entropy	Average gradient	Standard deviation	Spatial resolution	Interactive information
Region-based level	7.2378	38.5448	18.2632	31.9562	5.7235
TLBO optimization	7.3910	45.8079	18.8261	32.5719	5.2021
Improved TLBO optimization	7.3979	45.8553	18.6944	32.3447	5.2667

and feature-level fusion results are better, so different fusion methods can be adapted according to different requirements.

4.3. Results of TLBO Algorithm. The feature-level image fusion results of the infrared and visible images mentioned above are shown in Table 4. It can be seen from the table that its entropy value was 7.3981, but was 7.5121 after the optimization of the original TLBO model. Figure 8 shows the comparison between the optimized image and the original feature pole fusion image.



(a)

(b)

(c)

Figure 12: Comparison of fusion result based on the other group images ((a) feature-level fusion; (b) original TLBO algorithm; (c) improved TLBO algorithm).

The entropy value, standard deviation, and mean gradient increased by 1.16%, 7%, and 17.69%, respectively. All the indexes of the optimized image were improved, but not quite significantly.

The feature-level image fusion results of the other group of infrared and visible images are shown in Table 5. It can be seen from the table that its entropy value was 7.2378, but was 7.3910 after the optimization of the original TLBO model. Figure 9 shows the comparison between the optimized image and the fusion image of the original feature pole.

The entropy value, standard deviation, and mean gradient increased by 2.12%, 18.84%, and 3.10%, respectively. Except for interactive information, all the indexes of the optimized image were improved, but not quite significantly. The interaction information represents the relationship between the fusion image and the source image, and the larger the value is, the better the fusion effect is. However, in the fusion image, more infrared image information is needed to obtain a more obvious contour and detailed texture with a better entropy value, so the parameters of interaction information are relatively low.

4.4. Results of Improved TLBO Algorithm. Figures 10(a) and 10(b) are a group of infrared and visible light sample images named UN-CAMP, which have been applied to effect comparison in many domestic and foreign literature on image fusion algorithms. Figure 10(c) is the feature-level fusion image. Figures 10(d) and 10(e) are, respectively, the optimized results of the original TLBO algorithm and the improved TLBO algorithm after the random parameter setting. Table 6 shows the corresponding index evaluation results.

The data in the above table show that all quantitative evaluation indexes of the results after optimization of fusion parameters were improved. Compared with before optimization, the amount of information and spatial activity of images increased by 2.05% and 15.27%, respectively, and the standard deviation and mean gradient of image sharpness and visual effect increased by 13.27% and 28.87%, respectively.

For the infrared and visible image samples selected above, after the same random parameter setting process, the entropy value of the fused image reached the optimal value when the value range of the random parameter R_i and T_f were fixed at [0.4,0.9] and [0.5,1], respectively. Figure 11 shows the feature-level fusion image and the optimized results of the original TLBO algorithm and the improved TLBO algorithm after random parameter setting.

Table 7 shows the evaluation results of corresponding indicators. For this sample, when all entropy values, standard deviation, and mean gradient were improved, the spatial resolution and interactive information data decreased compared with the feature fusion results. Due to complex background information, the improved TLBO algorithm is better in terms of the information content, contrast, and spatial activity of the optimized image. However, the processing results are different from the source image in terms of visual effect due to the excessive influence of background information. In general, the algorithm is quite effective in improving the quality of fusion images and execution efficiency and in achieving better extraction results of target forest images than other algorithms.

For the infrared and visible images of the other group, after the same random parameter setting, the entropy value of the fused image reached the optimal value when the value range of the random parameter R_i and T_f were fixed at [0.3,0.8] and [0.5,1], respectively. Figure 12 shows the feature-level fusion image, the optimized results of the original TLBO algorithm, and the improved TLBO algorithm after random parameter setting.

Table 8 shows the evaluation results of corresponding indicators. As can be seen from the table, when the entropy value and standard deviation indicators were improved, the spatial resolution and mean gradient data became lower than the original optimization results, but still higher than the feature-level fusion image. In the algorithm, the amount of information and contrast of the optimized image are better. By comparing the results obtained from multiple sets of data, it can be seen that for different image samples, the algorithm has relatively optimized effects in improving the quality of fusion images and execution efficiency and could achieve better extraction results of target forest images than other algorithms.

5. Discussion and Conclusions

In the pixel-level image fusion algorithm research, the pulse coupled neural network model relying on contourlet transform is applied to avoid block effect and grayscale distortion caused by the fusion of infrared and visible images. Given the significant difference in gray level between infrared and visible forest images, a reasonable threshold value is selected in the low-frequency domain fusion processing. The points with different output pulse signals are treated differently, and the fusion rules are explicitly formulated. Thus, grayscale distortion and block effect are avoided, but the quality of the fusion image can be effectively improved, and all evaluation indexes of the image can be improved to some extent. In the research of feature-level image fusion algorithms, the PCNN model was used to eliminate the influence of noise in path optimization. The segmentation consequences are sufficient to meet the needs of feature-level fusion research even though they are disturbed by confusable issues. Based on the fuzzy logic rules, the fusion rules in the low-frequency domain are formulated by calculating the degree of dissimilarity between corresponding points of source images. The fusion rules in the high-frequency domain are determined by combining the image segmentation results. While ensuring the visual effect of the fused image, the detailed characteristic information of the target region in the image was displayed, making the research on image fusion a more targeted and purposeful algorithm into the algorithm to improve further the local search ability of the algorithm [24, 25]. The experimental results show that the feature-level image fusion algorithm ensures image quality, achieves the detailed display of the tree target area in the image, and improves quality evaluation indexes. Compared with the pixel-level fusion results, the tree texture obtained by this method is more evident, with more apparent edges.

In the research of feature-level image optimization, the teaching learning-based optimization (TLBO) parameter optimization algorithm is introduced to optimize the fusion coefficient in the fusion process to improve the fusion image's index data. In order to obtain better image results, the optimal parameter combination for different image groups to achieve the optimal effect by setting the value range of random parameters in the TLBO model and various quantitative evaluation indexes of fused images was improved. Pixel-level and feature-level fusion algorithms have appropriate advantages for different occasions. Pixel-level fusion has advantages in improving image information and sharpness, but it takes twice as long to process information as feature-level fusion. Feature-level fusion has a broader application space in forestry intelligent information detection as it can highlight the target area and reduce com-putation. The setting of algorithm parameters has an impor-tant influence on its optimization ability. The teaching factor of the basic TLBO algorithm varies only, which affects the optimization performance of the algorithm. Therefore, an improved TLBO optimization algorithm is proposed to design the teaching factor by segmenting strategy to process the image. Experimental results verify the algorithm's effectiveness, which has good searching ability and fusion image quality. In the future, this proposed method will be extended to theoretical research and practical applications including time-serial prediction and pattern recognition [16, 26–28].

References

[1] J. Kong, H. Wang, X. Wang, X. Jin, X. Fang, and S. Lin, "Multi-stream hybrid architecture based on cross-level fusion strategy for fine-grained crop species recognition in precision agriculture," *Computers and Electronics in Agriculture*, vol. 185, article 106134, 2021.

[2] J. Kong, C. Yang, J. Wang et al., "Deep-stacking network approach by multisource data mining for hazardous risk identification in IoT-based intelligent food management systems," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 1194565, 16 pages, 2021.

[3] M. I. Daily, T. Farr, and C. Elachi, "Geologic interpretation from composited radar and Lansat imagery," *Photogrammetric Engineering and Remote Sensing*, vol. 45, no. 8, pp. 1109–1116, 1979.

[4] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Patten Recognition Letters*, vol. 9, no. 4, pp. 245–253, 1989.

[5] R. A. Eggleston and C. A. Kohl, "Symbolic fusion of MMW and IR imagery," *Proceedings of SPIE*, vol. 1003, pp. 20–27, 1988.

[6] A. Toet, J. K. Ijspeert, A. M. Waxman, and M. Aguilar, "Fusion of visible and thermal imagery improves situational awareness," *Proceedings of SPIE on Enhanced and Synthetic Vision*, vol. 3088, pp. 177–188, 1997.

[7] L. P. Pappas, P. Malik, and M. Styner, "New method to assess the registration of CT·MR images of the head," *Medical Imaging 2004: Image Processing*, vol. 5370, pp. 129–136, 2004.

[8] X. B. Jin, W. Z. Zheng, J. L. Kong et al., "Deep-learning forecasting method for electric power load via attention-based encoder-decoder with Bayesian optimization," *Energies*, vol. 14, no. 6, p. 1596, 2021.

[9] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Information Fusion*, vol. 23, pp. 139–155, 2015.

[10] D. M. Bulanona, T. F. Burksa, and V. Alchanatisb, "Image fusion of visible and thermal images for fruit detection," *Biosystems Engineering*, vol. 103, no. 1, pp. 12–22, 2009.

[11] Y. Lei, D. Xiaokang, K. Jianlei, Y. Zheng, and L. Jinhao, "Parameters optimization algorithms for improving the performance of obstacles identification in forest area," *INMATEH-Agricultural Engineering*, vol. 40, no. 2, pp. 43–52, 2013.

[12] Y. Lei, D. Xiaokang, Y. Zheng, K. Jianlei, and L. Jinhao, "A novel identification method of obstacles based on multisensor data fusion in forest," *Sensors & Transducers*, vol. 155, no. 8, p. 155, 2013.

[13] C. Pohl and J. L. Van Genderen, "Review article multisensor image fusion in remote sensing: concepts, methods and applications," *International Journal of Remote Sensing*, vol. 19, no. 5, pp. 823–854, 1998.

[14] G. Stephane and A. Matllat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[15] M. Do and M. Vetterli, "Contourlets: a directional multiresolution image representation," *International Conference of Image Processing Proceedings*, vol. 1, pp. 357–369, 2002.

[16] X. B. Jin, W. Z. Zheng, J. L. Kong et al., "Deep-learning temporal predictor via bi-directional self-attentive encoder-decoder framework for IOT-based environmental sensing in intelligent greenhouse," *Agriculture*, vol. 11, no. 8, p. 802, 2021.

[17] Y. Y. Zheng, J. L. Kong, X. B. Jin, X. Y. Wang, T. L. Su, and M. Zuo, "CropDeep: the crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, p. 1058, 2019.

[18] Y. Y. Zheng, J. L. Kong, X. B. Jin, X. Y. Wang, T. L. Su, and M. Zuo, "Probability fusion decision framework of multiple deep neural networks for fine-grained visual classification," *IEEE Access*, vol. 7, pp. 122740–122757, 2019.

[19] J. Teng, S. Wang, J. Zhang, and W. Xue, "Fusion algorithm of medical images based on fuzzy logic," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 546–559, Yantai, China, 2010.

[20] W. Cai and Z. Wei, "Pii GAN: generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2019.

[21] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems," *Computer-Aided Design*, vol. 43, no. 3, pp. 303–315, 2011.

[22] H. Jin and Y. Wang, "A fusion method for visible and infrared images based on contrast pyramid with teaching learning based optimization," *Infrared Physics & Technology*, vol. 64, pp. 134–142, 2014.

[23] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information Fusion*, vol. 14, no. 2, pp. 127–135, 2013.

[24] B. Mandal and P. K. Roy, "Optimal reactive power dispatch using quasi-oppositional teaching learning based optimization," *International Journal of Electrical Power & Energy Systems*, vol. 53, pp. 123–134, 2013.

[25] R. V. Rao and V. Patel, "Comparative performance of an elitist teaching-learning-based optimization algorithm for solving unconstrained optimization problems," *International Journal of Industrial Engineering Computations*, vol. 4, no. 1, pp. 29–50, 2013.

[26] X.-B. Jin, W.-T. Gong, J.-L. Kong, Y.-T. Bai, and T.-L. Su, "PFVAE: a planar flow-based variational auto-encoder prediction model for time series data," *Mathematics*, vol. 10, no. 4, p. 610, 2022.

[27] X.-B. Jin, J.-S. Zhang, J.-L. Kong, Y.-T. Bai, and T.-L. Su, "A reversible automatic selection normalization (RASN) deep network for predicting in the smart agriculture system," *Agronomy*, vol. 12, p. 591, 2022.

[28] J.-L. Kong, H.-X. Wang, C.-C. Yang, X.-B. Jin, M. Zuo, and X. Zhang, "Fine-grained pests and diseases recognition via spatial feature-enhanced attention architecture with high-order pooling representation for precision agriculture practice," *Agriculture*, vol. 2022, article 1592804, 2022.

A Novel Passive Circuit Emulator for a Current-Controlled Memristor

Swadesh Ranjan Jena, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, swadesh.jena@yahoo.co.in*

Rudra Prasad Nanda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, rudraprasad858@gmail.com*

Laxminarayan Mishra, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, laxminarayan.s@gmail.com*

Chinmaya Ranjan Pradhan, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, cr.pradhan23@gmail.com*

Abstract

A memristor is an electrical element, which has been conjectured in 1971 to complete the lumped circuit theory. Currently, researchers use memristor emulators through diodes, inductors, and other passive (or active) elements to study circuits with possible attractors, chaos, and ways of implementing nonlinear transformations for low-voltage novel computing paradigms. However, to date, such passive memristor emulators have been voltage-controlled. In this study, a novel circuit realization of a passive current-controlled passive inductorless emulator is established. It overcomes the lack of passive current-controlled memristor commercial devices, and it can be used as part of more sophisticated circuits. Moreover, it covers a gap in the state of the art because, currently, only passive circuit voltage-controlled memristor emulators and active current-controlled emulators have been developed and used. The emulator only uses two diodes, two resistors, and one capacitance and is passive. The formal theory and simulations validate the proposed circuit, and experimental measurements were performed. The parameter conditions of numerical simulations and experiments are consistent. Simulations were performed with an input current amplitude of 15mA and frequencies of up to 3kHz and measurements were carried out with an input current amplitude of 0.74mA and frequency of 1.5kHz in order to compare with the state of the art.

1. Introduction

A memristor is an electrical two-terminal passive nonlinear resistance element that exhibits a well-known pinched hysteresis loop at the origin of the voltage-current plane when any bipolar periodic zero-mean excitatory voltage or current of any value is applied across it. However, there are disagreements regarding whether a memristor can be considered a fundamental element and whether its dynamic is purely electromagnetic, as originally conjectured, or if there are other mechanisms involved such as ionic transport. Despite the controversy surrounding the technological realization, the amount of research into the properties of the pinched hysteresis loop continues to increase. Irrespective of whether the memristor is implemented by emulating its behavior through circuitry composed of other active or passive components, new studies continue to generate and sustain optimistic expectations in the scientific community about the use and advantages of the memristor. The research interest in the memristor is motivated by its promising potential for building novel integrated circuits and computing systems, as has been proposed in [1–7]. Memristors allow the memory and time-varying processing of information through nonlinear transformations in a unique passive element.

There is neither implementation nor a proof-of-concept for a passive current-controlled memristor system-on-chip but yes for an active current-controlled memristor [8]. With an increasing need to better understand a pinched hysteresis attractor, researchers implement memristor emulators through diodes and other passive (or active) elements. Such emulators allow studying (theoretically and numerically) possible attractors and ways of implementing nonlinear transformations for novel computing paradigms. The first voltage-controlled memristor emulator has been proposed in [9] and was based on a diode bridge with a parallel $R-C$ filter as a load. Other studies [10–16] used voltage-controlled memristor emulators as vital building blocks of other circuits for the in-depth study of their bifurcations and chaotic behaviors.

To summarize, (a) background: to the best of our knowledge and after searching in database providers, we have concluded that a passive and current-controlled memristor emulator has not been published previously (only active current-controlled memristor [8]) (although there are a lot of voltage-controlled memristor emulators) and (b) motivation: design a passive and current-controlled memristor emulator for neuromorphic computing and as a candidate for mimicking synaptic functions and inductorless emulator that could be easily integrated into CMOS technology. It is similar in spirit to some circuits in the literature, which combine a rectifier with a LC low pass filter and not an RC as it is here. This memristor enables its utilization in programmable circuits and systems controlled by digital pulses.

Memristors exhibit three characteristics for any bipolar periodic signal excitation: (i) there is a pinched hysteresis loop in the voltage-current plane, (ii) the area of the hysteresis loop decreases and shrinks to a single-valued *V-I* function when the signal excitation frequency tends to infinity, and (iii) for voltage-controlled generalized memristive time-invariant systems, the following equations apply [1–7]:

$$\begin{cases} i_m = G(\mathbf{x}, V_m)V_m \text{ and } G(\mathbf{x}, 0) \neq \infty \forall \mathbf{x}, \\ \\ \dfrac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f(\mathbf{x}, V_m)\mathbf{x} \text{ represents the inner state variables,} \end{cases} \tag{1}$$

where i_m is the current across the memristor, V_m is the voltage across the terminals, $G(\mathbf{x}, V_m)$ is bounded, and $f(\mathbf{x}, i_m)$ is the equation of state, which must be also bounded to guarantee the existence of a solution $\mathbf{x}(t)$. The area of the lobes, shapes, and orientation of the hysteresis loop evolves with frequency. All of the abovementioned references developed voltage-controlled memristor emulators (i.e., equations such as equation (1)). However, in this study, we implement the first-ever built passive current-controlled memristor emulator for which the following equations apply:

$$\begin{cases} V_m = R(\mathbf{x}, i_m)i_m \text{ and } R(\mathbf{x}, 0) \neq \infty \forall \mathbf{x}, \\ \\ \dfrac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f(\mathbf{x}, i_m)\mathbf{x} \text{ represents the inner state variables.} \end{cases} \tag{2}$$

2. Proposed Current-Controlled Memristor

Novel current-controlled memristor emulators have been introduced recently in [8, 17, 18]; however, all of them are active (it means that a voltage power supply is needed). In this work, the first passive circuit realization of a current-controlled emulator is established in Figure 1. It uses two diodes, two resistors, and one capacitor.

We construct the equations by beginning with the Shockley diode equation (both diodes are equal and without considering high-frequency effects that produce unwanted dynamic effects):

$$\begin{aligned} i_{D_1} &= I_s\left(e^{2\alpha V_{jD_1}} - 1\right), \\ i_{D_2} &= I_s\left(e^{2\alpha V_{jD_2}} - 1\right), \end{aligned} \tag{3}$$

where $2\alpha = 1/nV_T$ and I_s denote the reverse saturation current, n is the emission coefficient, V_T is the thermal voltage, and V_{jD_1} and V_{jD_2} are the junction diode voltages. If we consider their series parasitic resistance R_p, the diode voltages become

$$V_{D_1} = R_p i_{D_1} + \frac{1}{2\alpha}\ln\left(1 + \frac{i_{D_1}}{I_s}\right) \text{ and } V_{D_2} = R_p i_{D_2} + \frac{1}{2\alpha}\ln\left(1 + \frac{i_{D_2}}{I_s}\right). \tag{4}$$

Therefore,

$$i_{D_1} = I_s\left(e^{2\alpha\left(V_{D_1} - R_p I_{D_1}\right)} - 1\right) \text{ and } i_{D_2} = I_s\left(e^{2\alpha\left(V_{D_2} - R_p I_{D_2}\right)} - 1\right). \tag{5}$$

According to the voltage drop,

$$V_m = Ri_1 + V_{D_1} \text{ and } V_m = Ri_2 - V_{D_2}, \tag{6}$$

$$-V_c = V_{D_1} + V_{D_2}. \tag{7}$$

The current i_m corresponds to $i_m = i_{D_1} - i_{D_2}$. Using equation (5),

$$i_m = I_s e^{\alpha\left(V_{D_1} + V_{D_2} - R_p\left(i_{D_1} + i_{D_2}\right)\right)}\left(e^{\alpha\left(V_{D_1} - V_{D_2} - R_p\left(i_{D_1} - i_{D_2}\right)\right)} - e^{-\alpha\left(V_{D_1} - V_{D_2} - R_p\left(i_{D_1} - i_{D_2}\right)\right)}\right). \tag{8}$$

Using equation (7), it becomes convenient to express

$$i_m = 2I_s e^{-\alpha V_c} e^{-\alpha R_p\left(i_{D_1} + i_{D_2}\right)} \sinh\left(\alpha\left(V_{D_1} - V_{D_2} - R_p i_m\right)\right). \tag{9}$$

However, from equation (6),

$$2V_m = Ri_m + V_{D_1} - V_{D_2}. \tag{10}$$

Then, the $\{V_m - i_m\}$ relation is provided,

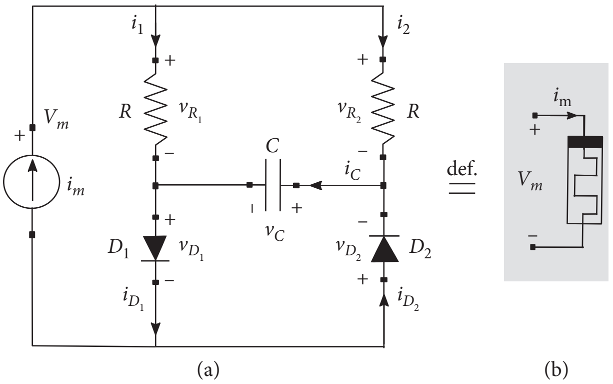


FIGURE 1: (a) Proposed current-controlled memristor circuit emulator and (b) generalized symbol of the memristor device.

$$V_m = \frac{1}{2}Ri_m + \frac{1}{2\alpha}\sinh^{-1}\left(\frac{i_m}{2I_s}e^{\alpha V_c+\alpha R_p\left(i_{D_1}+i_{D_2}\right)}\right) + \frac{1}{2}R_p i_m. \tag{11}$$

Because $i_c = i_{D_1} - i_m + i_2$ and $i_c = i_{D_2} + i_2$, we obtain $i_{D_1} + i_{D_2} = 2i_c + i_m - 2i_2$ and then,

$$V_m = \frac{1}{2}Ri_m + \frac{1}{2\alpha}\sinh^{-1}\left(\frac{i_m}{2I_s}e^{\alpha V_c+\alpha R_p\left(2i_c+i_m-2i_2\right)}\right) + \frac{1}{2}R_p i_m. \tag{12}$$

Next, we focus on the i_2 expression. By taking $V_m = V_{D_1} + V_C + Ri_2$ and $V_m = -V_{D_2} - V_C + Ri_1$, we obtain

$$V_{D_1} + V_{D_2} + 2V_C + R\left(i_2 - i_1\right) = 0. \tag{13}$$

Because $i_2 - i_1 = i_2 - 2i_1 + i_1 = i_m - 2i_1$, and $-V_c = V_{D_1} + V_{D_2}$, from equation (13), we obtain

$$i_1 = \frac{V_c}{2R} + \frac{i_m}{2} \text{ and } i_2 = -\frac{V_c}{2R} + \frac{i_m}{2}. \tag{14}$$

Finally, equation (12) becomes

$$V_m = \frac{1}{2}\left(R + R_p\right)i_m + \frac{1}{2\alpha}\sinh^{-1}\left(\frac{i_m}{2I_s}e^{\alpha V_c+\alpha R_p\left(2i_c+V_c/R\right)}\right). \tag{15}$$

Next, we focus on the state equation:

$$\begin{aligned} i_c &= C\frac{\mathrm{d}V_c}{\mathrm{d}t}. \\ &= i_{D_1} - i_1 \Longrightarrow C\frac{\mathrm{d}V_c}{\mathrm{d}t}, \\ &= i_{D_1} - \frac{V_c}{2R} - \frac{i_m}{2}. \end{aligned} \tag{16}$$

Now, to complete the state equation, we have to calculate i_{D_1}, which is straightforward (note that according to equation (2), $\mathbf{x} = V_c$):

$$\begin{aligned} \frac{V_m - V_{D_1}}{R} &= i_1, \\ &= \frac{V_c}{2R} + \frac{i_m}{2} \Rightarrow V_{D_1}, \\ &= V_m - \frac{Ri_m}{2} - \frac{V_c}{2}. \end{aligned} \tag{17}$$

Therefore, from equation (16) and $i_{D_1} = I_s\left(e^{2\alpha\left(V_{D_1} - R_p i_{D_1}\right)} - 1\right)$, we obtain

$$C\frac{\mathrm{d}V_c}{\mathrm{d}t} = I_s\left(e^{2\alpha V_m}e^{-\alpha Ri_m}e^{-\alpha V_c}e^{-2\alpha R_p i_{D_1}} - 1\right) - \frac{i_m}{2} - \frac{V_c}{2R}. \tag{18}$$

By introducing equation (15), we obtain

$$\begin{aligned} C\frac{\mathrm{d}V_c}{\mathrm{d}t} = {} & I_s e^{\sinh^{-1}\left(i_m/2I_s e^{\alpha V_c+\alpha R_p\left(2i_c+V_c/R\right)}\right)+\alpha R_p i_m - 2\alpha R_p i_{D_1} - \alpha V_c} \\ & - I_s - \frac{i_m}{2} - \frac{V_c}{2R}. \end{aligned} \tag{19}$$

Because $i_{D_1} = i_c + (V_c/2R) + (i_m/2)$, we obtain

$$C\frac{\mathrm{d}V_c}{\mathrm{d}t} = I_s e^{\sinh^{-1}\left(\left(i_m/2I_s\right)e^{\alpha V_c+\alpha R_p\left(2i_c+V_c/R\right)}\right)+\alpha R_p i_m - 2\alpha R_p\left(i_c+\left(V_c/2R\right)+\left(i_m/2\right)\right)-\alpha V_c} - I_s - \frac{i_m}{2} - \frac{V_c}{2R}. \tag{20}$$

Finally, this circuit dynamic can be written as follows (where V_{co} is the initial capacitor value, i.e., the initially configured memory):

$$\begin{cases} V_m = \frac{1}{2}\left(R + R_p\right) i_m + \frac{1}{2\alpha}\sinh^{-1}\left(\frac{i_m}{2I_s} e^{\alpha V_c + \alpha R_p\left(2i_c + \left(V_c/R\right)\right)}\right), \\ C\frac{\mathrm{d}V_c}{\mathrm{d}t} = I_s e^{\sinh^{-1}\left(\left(i_m/2I_s\right)e^{\alpha V_c + \alpha R_p\left(2i_c+\left(V_c/R\right)\right)}\right) - \alpha R_p\left(2i_c + \left(V_c/R\right)\right) - \alpha V_c} - I_s - \frac{i_m}{2} - \frac{V_c}{2R}, V_c\left(t = 0\right) = V_{co}. \end{cases} \tag{21}$$

From equation (21), one can see how the parasitic resistance R_p affects the dynamic. In order to continue, we reasonably assume $R_p \approx 0$ because most diodes datasheets report $R_p \ll 1$.

According to equation (2), V_c represents the inner state variable; however, it should be noted that the form $V_m = R(V_c, i_m) i_m$ is not achieved (i.e., it does not contain i_m proportionality). Instead, it can be naturally achieved through division by i_m as follows:

$$\begin{cases} V_m = \left[\frac{1}{2}R + \frac{1}{2\alpha i_m}\sinh^{-1}\left(\frac{i_m}{2I_s}e^{\alpha V_c}\right)\right] i_m \text{ for } i_m \neq 0, \\ V_m = 0 \text{ for } i_m = 0, \\ \frac{\mathrm{d}V_c}{\mathrm{d}t} = \frac{I_s}{C}\Bigg[e^{\sinh^{-1}\left(\left(i_m/2I_s\right)e^{\alpha V_c}\right) - \alpha V_c - 1]} - \frac{i_m}{2C} - \frac{V_c}{2RC}, V_c\left(t = 0\right) = V_{co}, \end{cases} \tag{22}$$

where according to equation (2) the term $R(V_c, i_m)$ is

$$\begin{cases} R\left(V_c, i_m\right) = \frac{1}{2}R + \frac{1}{2\alpha i_m}\sinh^{-1}\left(\frac{i_m}{2I_s}e^{\alpha V_c}\right) \text{for } i_m \neq 0, \\ R\left(V_c, i_m\right) = 0 \text{ for } i_m = 0. \end{cases} \tag{23}$$

Note that $R(V_c, i_m) \longrightarrow 0$ for $i_m \longrightarrow 0$; thus, it can be continuously extended to $i_m = 0$ (the limit exists and can be obtained by L'Hospital's rule). Figure 2 shows in advance an example of the zero-crossing $V - i$ (the next section discusses the validation by simulation).

3. Validation by Simulation

The following parameters were used to simulate the memristor circuit emulator proposed in Figure 1: $R = 270\Omega$ and $C = 0.5\mu F$. The assigned diode was 1N4148 with a PSpice-model card, model 1N4148 D (Is = 2.52 n, Rs = 0.568, N = 1.752, CJO = 4 p, M = 0.4, TT = 20 n, Iave = 200 m, Vpk = 75, mfg = OnSemi type = silicon) [19]. The initial state condition $V_c = 0$ was selected. The simulator used was LTspice [20]. For such diode models, the magnitudes of circuit parameters and the values of input signal amplitudes and frequencies are maintained similarly to compare the lobe shapes obtained in other works such as [9, 12, 14, 15]. The current-voltage characteristics obtained from PSpice simulations are shown in Figures 2 and 3.

The following observations can be made from the simulation results. The loci in the V-i plane have hysteresis loops pinched at zero in the periodic steady state. The hysteresis loop shrinks to a single-valued function when the frequency increases and decreases, and the shape of the memristor depends on the circuit parameters and also retains the odd-symmetry property. The qualitative difference in the lobe shapes between the cases (a) and (b) of Figure 3 allows us to interpret that the dynamic of the memristor can exhibit a much richer behavior at low frequencies.

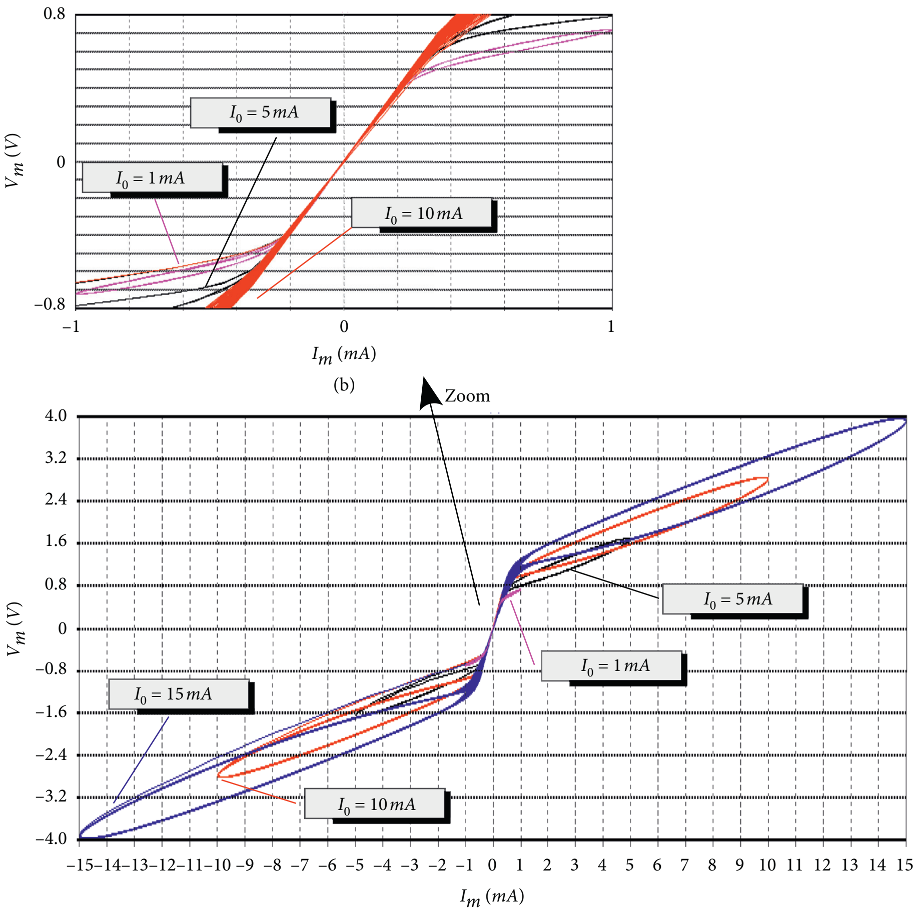
4. Measurements, Experimental Setup, and Simulation

The experimental setup is shown in Figure 4(b) and it comprises of two diodes model 1N4007, resistances $R = 270\Omega$, and the capacitor $C = 0.5\mu F$. In order to test more case studies and to confirm that the current-controller emulator behavior is independent of the diode model, we use another diode type (which is the 1N4007) with respect to the above section.

We can see the emulator under test and also the oscilloscope-waveform generator Analog Discovery 2 [21]. The experimental result is shown in Figure 4(a) and the simulated result in Figure 5.

The excitatory signal is a sinusoidal waveform of frequency f = 1500Hz and current amplitude of I_o = 0.74mA (selected as similar as possible to that shown in [8], in order to compare the resulting $v - i$ characteristic). We use a voltage-controlled current source (VCCS), i.e., Howland current source (EHCS) circuits with a maximum output current of value of 1mA. This voltage-controlled current source was built with an OpAmp MCP6004 from Microchip Inc.

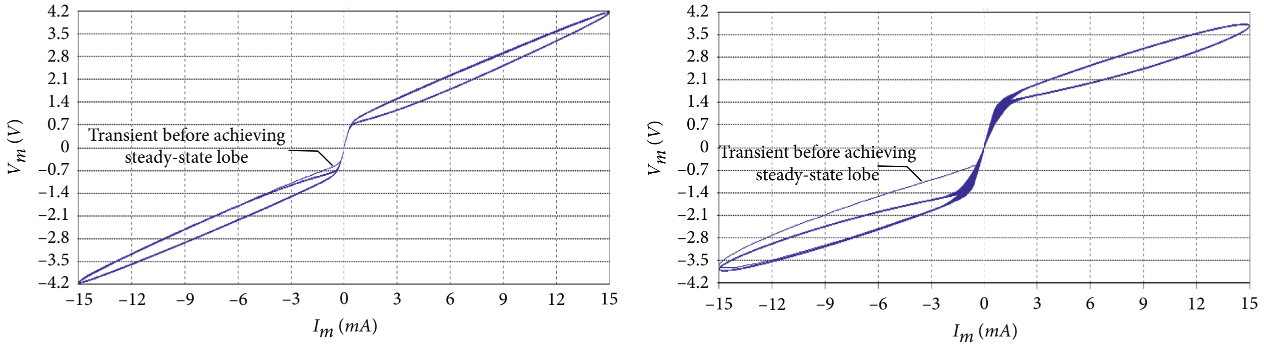
We conclude that the experimental measurement shows the same waveform pattern and the performance predicted by theory (Figure 4(a)) and by the emulator circuit simulations. Most important is to remark that the $i - v$ pattern is similar to one of the references [8] (for which the maximum current across the memristor is around 0.8mA and the voltage 1V, with lobes in the $i - v$ pattern and a straight line



(b)

(a)

Figure 2: Current-voltage characteristics obtained through the LTspice simulations of the memristor emulator driven by a current source $i_m = I_o \sin(2\pi f t)$ at a constant frequency $f = 300$Hz with different I_o values marked in (a): $I_o = 1$mA, $I_o = 5$mA, $I_o = 10$mA, and $I_o = 15$mA. (b) An amplified view of the zero-crossing $V - i$ omitting the lobe at $I_o = 15$mA to avoid the overwhelming superposition of curves.



(a)

(b)

Figure 3: Continued.

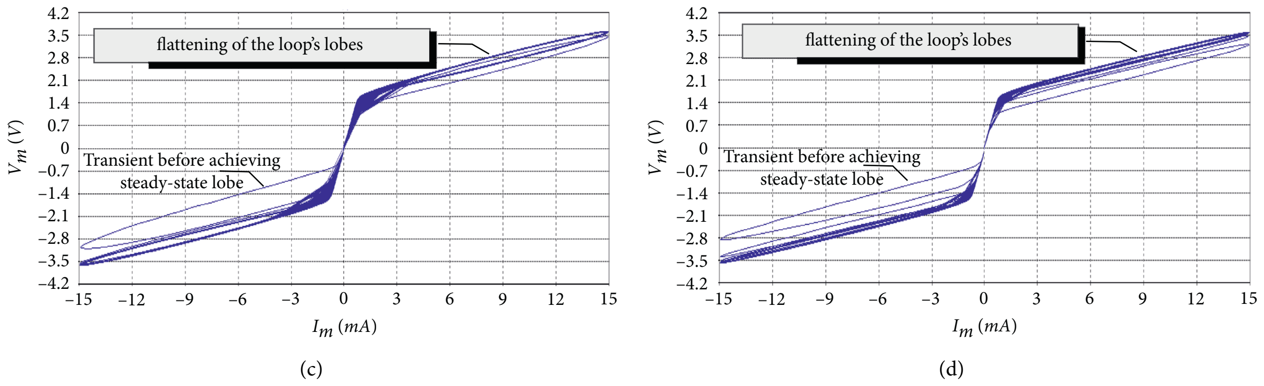


Figure 3: Simulated pinched hysteresis loop of the memristor emulator driven by a current source $i_m = I_o \sin(2\pi f t)$ with a constant amplitude $I_o = 15$mA at different f values: (a) $f = 100$Hz, (b) $f = 500$Hz, (c) $f = 1500$Hz, and (d) $f = 3000$Hz (100 cycles have been applied for each case study).

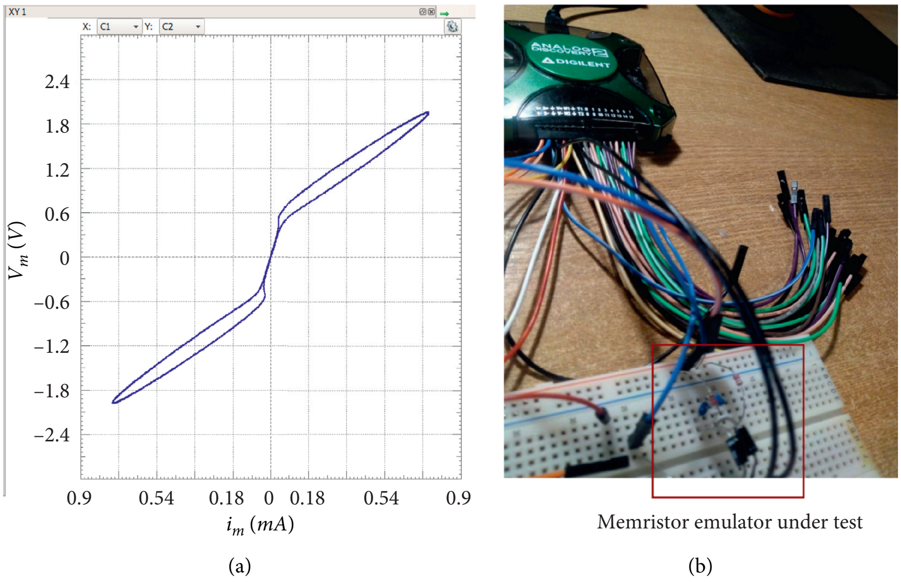


Figure 4: (a) V-i characteristics obtained through the measurements of an implemented memristor emulator driven by a current source $i_m = I_o \sin(2\pi f t)$ at a constant frequency $f = 1500$Hz with $I_o = 0.74$mA and (b) view of the memristor emulator under test.

at zero), with the difference that such reference in the state of the art develops an active current-controlled memristor that needs a DC voltage power supply of ±15V.

For completeness and comparison, the $v - i$ characteristics obtained through simulation of the measured case is shown in Figure 5, where is observed a minor difference in the slope at the origin that may be due to the diode model we use, which is model 1N4007 D (Is = 7.02767 n, Rs = 0.0341512, N = 1.80803, EG = 1.05743, XTI = 5, BV = 1000, IBV = 5e − 08, CJO = 1e − 11, VJ = 0.7, M = 0.5, FC = 0.5, TT = 1e–07, mfg = OnSemi type = silicon) (the diode model card has parasitic resistances and junction capacitance). Anyway, the experimental measurement shows the same waveform pattern and the performance predicted by theory and simulations, i.e., pinched hysteresis loop.

5. Volatility Test for Neuromorphic Computing

The emulator should not retain its value when no input signal is applied. The volatile memristor that features decay of device memory has high similarity to the biological synapses of neurons, and since neuromorphic computing is becoming more important, the decay is a desired feature.

At this point, it is important to underline that this kind of memristor emulator is not suitable for the logic-in-memory (LiM) paradigm, nor high-frequency applications; instead, our proposed circuit can be used practically for neuromorphic computing and as a candidate for mimicking biological synaptic functions. It is enough to show as example Figure 6 where an input current pulse train of 0.1mA amplitude (5 cycles) is applied. In this case study, the

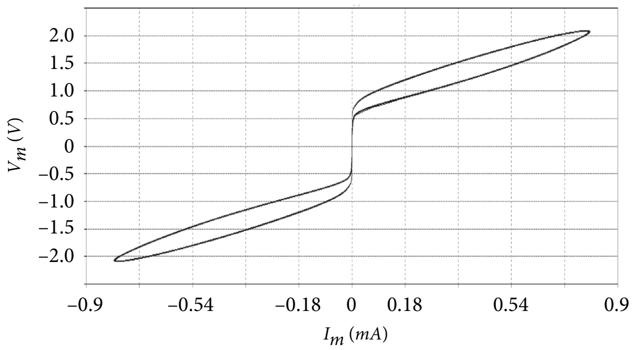


Figure 5: For completeness and comparison, the *V-i* characteristics obtained through simulation of the measured case (Figure 4).

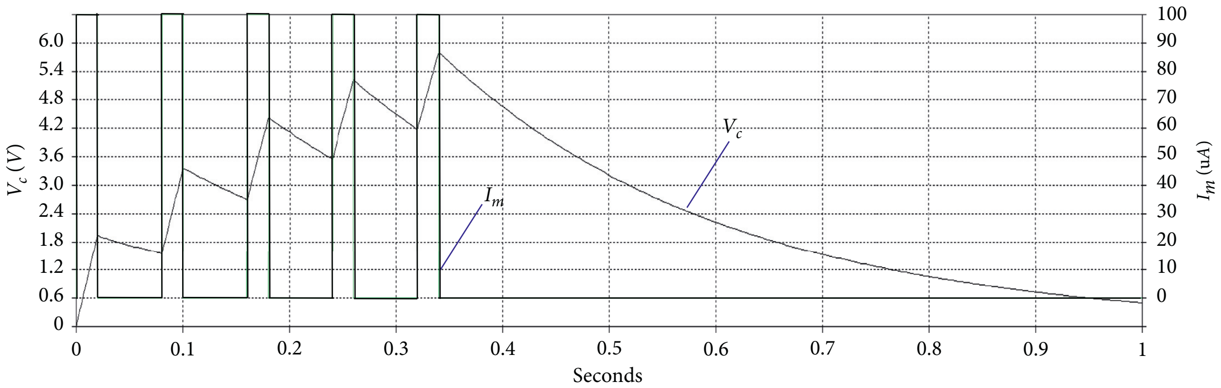


Figure 6: Variation of $V_c(t)$ (memory) with time for a given current input pulse.

emulator uses resistances $R_1 = R_2 = 270k\Omega$ and the proposed emulator shows spontaneous decay or better said volatile nature similarity to the biological nervous system.

6. Conclusion

This work shows a passive circuit current-controlled memristor emulator. It overcomes the lack of current-controlled memristor commercial devices. M oreover, it covers a gap in the state of the art because, currently, only passive circuit voltage-controlled memristor emulators have been developed and used. The mathematical model of the proposed memristor emulator was derived and verified by simulations (the mathematical treatment was simplified and the diode parasitic capacitance was not taken into account because in that case the device dynamics should be studied with intensive numerical simulations (using for instance M ATLAB) because the system is strongly nonlinear. Nevertheless, the SPICE simulator has the complete diode model with junction capacitance included; then, numerical circuital simulations are accurate). The circuit can be improved to be less symmetric in order to have a different clockwise and anticlockwise slope.

References

[1] P. Mazumder, S. M. Kang, and R. Waser, “Memristors: devices, models, and applications [scanning the issue],” *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1911–1919, 2012.

[2] S. Hamdioui, S. Kvatinsky, G. Cauwenberghs et al., “Memristor for computing: myth or reality?” in *Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE)*, pp. 722–731, Lausanne, Switzerland, March 2017.

[3] M. Selmy, H. Mostafa, and A. Dessouki, “Low power memristor based voltage controlled oscillator for electrical neural stimulation,” in *IEEE International Conference on Advanced Control Circuits and Systems and New Paradigms in Electronics and Information Technology*, pp. 344–347, Hong Kong, China, May 2017.

[4] B. Ramakrishnan, A. Durdu, K. Rajagopal, and A. Akgul, “Infinite attractors in a chaotic circuit with exponential memristor and josephson junction resonator,” *AEU-International Journal of Electronics and Communications*, vol. 123, Article ID 153319, 2020.

[5] P. Khurana, K. Singh, and A. Sharma, “A hybrid cmos-memristor based programmable wien bridge oscillator,” in *Proceedings of the 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology*, RTEICT, Bangalore, India, May 2018.

[6] M. Zidan, H. Omran, C. Smith, A. Syed, A. Radwan, and K. Salama, “A family of memristor-based reactance-less oscillators,” Engineering,” *International Journal of Circuit Theory and Applications*, vol. 42, no. 11, pp. 1103–1122, 2013.

[7] A. Mosad, M. Fouda, M. Khatib, K. Salama, and A. Radwan, “Improved memristor-based relaxation oscillator,” *Microelectronics Journal, Elseiver*, vol. 44, no. 9, 2013.

[8] G. Abdullah, J. Zainulabideen, M. Fouda, and M. H. Chowdhury, “A new simple emulator circuit for current controlled memristor,” in *Proceedings of the 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, pp. 288–291, Cairo, Egypt, December 2015.

[9] F. Corinto and A. Ascoli, “Memristive diode bridge with LCR filter,” *Electronics Letters*, vol. 48, no. 14, pp. 824-825, 2012.

[10] M. Chen, M. Li, Q. Yu, B. Bao, Q. Xu, and J. Wang, “Dynamics of self-excited attractors and hidden attractors in generalized memristor-based Chua’s circuit,” *Nonlinear Dynamics*, vol. 81, no. 1-2, pp. 215–226, 2015.

[11] Q. Xu, N. Wang, B. Bao, M. Chen, and C. Li, “A feasible memristive chua circuit via bridging a generalized memristor,” *Journal of Applied Analysis and Computation*, vol. 6, no. 4, pp. 1152–1163, 2015.

[12] M. Chen, J. Yu, Q. Yu, C. Li, and B. Bao, “A memristive diode bridge-based canonical chua’s circuit,” *Entropy*, vol. 16, no. 12, pp. 6464–6476, 2014.

[13] Z. T. Njitacke, J. kengne, H. B. Fotsin, A. N. Negou, and D. Tchiotsop, “Coexistence of multiple attractors and crisis route to chaos in a novel memristive diode bidge-based jerk circuit,” *Chaos, Solitons & Fractals*, vol. 91, pp. 180–197, 2016.

[14] Q. Xu, Q. Zhang, N. Wang, H. Wu, and B. Bao, “An improved memristive diode bridge-based band pass filter chaotic circuit,” *Mathematical Problems in Engineering*, vol. 201711 pages, 2017.

[15] B. C. Bao, P. Y. Wu, H. Bao, H. G. Wu, X. Zhang, and M. Chen, “Symmetric periodic bursting behavior and bifurcation mechanism in a third-order memristive diode bridge-based oscillator,” *Chaos, Solitons & Fractals*, vol. 109, pp. 146–153, 2018.

[16] A. Yesil, “A new grounded memristor emulator based on mosfet-c,” *AEU-International Journal of Electronics and Communications*, vol. 91, 2018.

[17] L. Zhijun, Z. Yicheng, and M. Minglin, “A novel floating memristor emulator with minimal components,” *Active and Passive Electronic Components*, vol. 2017, 2017.

[18] A. Alharbi and M. Chowdhury, “Simple current-controlled memristor emulators,” in *Memristor Emulator Circuits*-Springer, Berlin, Germany, 2020.

[19] LTwiki, “Standard.dio,” 2019, http://ltwiki.org/index.php?title=Standard.dio.

[20] A Devices, “Spice Simulation Software,” 2019, http://www.analog.com/en/design-center/design-tools-and-calculators.

[21] “analog Discovery 2,” 2020, https://www.reference.digilentinc.com/.

Indistinguishable sub-nanosecond pulse generator

Prakash Kumar Behera, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, prakash_behera21@gmail.com*

Sulochana Nanda, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, sulochanananda1@hotmail.com Bhagaban*

Sri Ramakrishna, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar,maheswarinath1@outlook.com*

Rashmi Ranjan Behera, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, rr_behera@yahoo.co.in*

ARTICLE INFO

Keywords:
Indistinguishable pulse
Ultra-short pulse
Quantum cryptography
Laser diodes
Laser driver

ABSTRACT

The indistinguishable laser pulse is predominant in the realization of qubits in decoy state-based quantum cryptography. Such cryptography in an automotive framework needs to be compact and inexpensive. However, their implementation is cumbrous and uneconomical. Significant effort has been put into reducing the form factor of quantum cryptography implementation. We report a compact, low-cost, indistinguishable, sub-nanosecond pulse generator with adjusted delay and amplitude using a Fabry–Perot laser diode. The approach was derived based on algebraic topology formulation of the electrical network, and the implementation involves time-dependent perturbation of a constant current node to generate tunable, sub-nanosecond excitation with a constant pre-bias. The simulation model of the laser diode accounted for effects of spontaneous emission and relaxation oscillation. The shortest excitation pulse thus generated was measured to have FWHM of 496 ps. Further, the indistinguishability of two laser pulses is statistically evaluated.

1. Introduction

Generation of indistinguishable excitation pulses is one of the fundamental requirements in many quantum information related experiments such as indistinguishable pulses for quantum cryptography application (Yin et al., 2020; Wei et al., 2013; Lo et al., 2005), photon number resolution using APDs (Kardynał et al., 2008). State of the art implementations for these experiments are immoderate in size and cost, to fit in automotive gateway modules and controllers. This calls for further reduction in form factor and cost.

Numerous methods have been reported towards implementing quantum cryptography, namely: electro-optic modulators in photonic ICs, driver-controlled variable attenuators. In 2015, Wabnig et al. at Nokia labs (Wabnig et al., 2015), proposed Quantum Key Distribution (QKD) based scheme using a spatial filter to generate indistinguishable pulses with Laser diode or LED. The indistinguishable pulses were defined by characteristics of the spatial filter. In 2016, to make QKD system compact, Bunandar et al. at MIT (Bunandar et al., 2016) proposed photonic integrated chip to realize transceiver. These photonic ICs comprised of ring resonators. For each resonators, delay to optical pulses was achieved using a modulator. Further, in 2017 Nordholt et al. at Los Alamos National Security (Nordholt et al., 2017) proposed an alternate photonic integrated IC-based approach. In this, a variable optical attenuator or amplitude modulator was used to reduce an average number of photons per pulse. Similarly, in 2017 Yuan et al. at Toshiba (Yuan et al., 2017) created a quantum communication system using a variable attenuator to change the intensity of the emitted laser pulses. These approaches often require a complex and dedicated driving circuits. Alternatively, direct modulation of laser can be another preferred technique to generate indistinguishable excitation (Linke and Gnauck, 1983), where the mean number of the photon in each pulses can be reduced to less than one using a constant optical attenuator, without engendering need of using dedicated driver circuit for the variable attenuator. However, state- of-the-art will be enormous, with unadjustable pulse duration in the range of nanoseconds.

In this paper, inspired by algebraic topology, by simulation and experimental measurements, we demonstrate the generation of tunable, indistinguishable, sub-nanosecond laser diode excitation. We have achieved shortest injection current FWHM of 496 ps using direct modulation of the laser diode. Further, we demonstrate the tunability of laser diode excitation pulses indistinguishable in the temporal domain. Implementation of quantum cryptography has proven challenging in terms of cost, form factor, and integration into existing infrastructure. This approach can provide low cost and small form factor for the implementation of quantum cryptography transmitter (and receiver), with greater integration into an existing automotive framework.

2. Methods

Characterization of short-duration pulses has been discussed in Williams (2004), Maan (2020). Photons emitted from periodically excited laser diode is monitored using SAP500 based linear photon detector. Fine-tuning of the excitation parameter can be programmed to generate indistinguishable pulses. Further, excitation current profile can be characterized by monitoring the anode and cathode of the laser

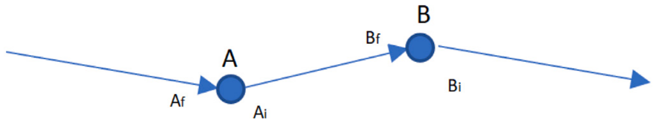


Fig. 1. Definition of boundary operator of a branch $|A_f$ is the branch with A as final node $|A_i$ is the branch with A as initial node $|B_f$ is the branch with B as final node. Similarly, B_i is the branch with B as initial node.

diode. Delay between excitation current and detected photon comprises of delay due to photon generation and time taken by photon to travel from the laser diode to the detector, as discussed in Maan (2021). Any delay thus observed can be adjusted by fine-tuning of the control parameters.

3. Theory

Let us visit some preliminaries pertaining to chain complexes:

Theorem 1 (*Boundary Operator* ∂). *If ∂_n is a boundary operator, which is a linear transformation on the sequence of vector space C_n, expressed as $\partial_n : C_n \rightarrow C_{n-1}$, we can write:*

$$\partial_n : C_n \rightarrow \partial_{n-1} : C_{n-1} \rightarrow \partial_{n-1} : C_{n-2} \dots \rightarrow \partial_0 : C_0 \rightarrow 0$$

as a chain complex if $\partial_n \cdot \partial_{n+1} = 0$. If boundary operator on a branch ω is defined as: $\partial\omega = \Omega(e - s)$ where e is the end node and s is the start node of branch Ω, then we can interpret the first operator ∂_n on 1-chain complex as nodes. Since the node does not have a boundary, so second boundary operator on the node, ∂_{n+1}, is zero.

We can now define the kernel of a boundary operator. If a k-chain has no boundary, then it is called k-cycles and can be represented as a kernel of the boundary operator (Smale, 2000):

$$Z_1 = ker \quad \partial k \tag{1}$$

if $k \in Z_1$

$$\partial k = 0 \tag{2}$$

Lemma 1. *The boundary for branches can be expressed as a sum of all the branches that leave and enter a node.*

Proof. Fig. 1 represents node A and node B, where A_f and A_i are branches with node A as final and initial node respectively. Similarly, B_f and B_i are branches with B as final and initial node respectively. If $K \in C_1$ is a set of all branches such that

$$\partial K = \{k_a, k_b \dots\} \tag{3}$$

Here $k_i \in C_0$ represents branches that enter and leave node i and can be represented as

$$k_i = k_{ei} - k_{li} \tag{4}$$

where k_{ei} and k_{li} represent branches that enter and leave node i respectively, irrespective whether it is time-independent or not.

With reference to Fig. 1:

$$\partial K = A_f(A - \Delta_i) + B_f(B - A) \tag{5}$$

where Δ_i represents arbitrary start node of the branch A_f. Consolidating all the terms pertaining to node A together:

$$\partial K = A(A_f - B_f) + B(B_f) - \Delta_i(A_f) \tag{6}$$

The first term in Eq. (6) pertains to node A, represents branch entering the node (-) branch leaving the node, and it is valid for both constant as well as time dependent branches. □

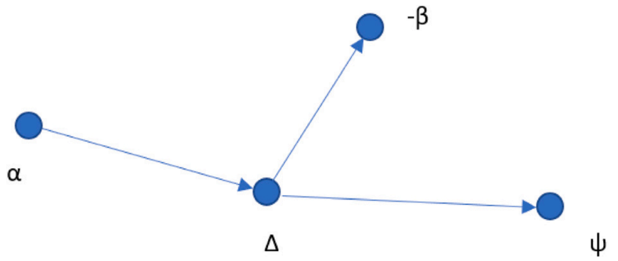


Fig. 2. Electrical sub-network for generation of excitation pulse $|\Delta$ is the constant current node, β is the negative perturbation node, α is the constant excitation node, ψ is the output node | Laser diode (or APD for voltage perturbation) is connected at ψ node.

Consider an electrical sub-network represented by: {N,B, ∂, $R_{dc}, R_{perturbation}$}, where branch $B \in C_1$ one chain vector space and node $N \in C_0$ zero chain vector space, $\partial : C_1 \rightarrow C_0$ is the boundary map, R_{dc} is the coefficient ring for dc excitation and $R_{perturbation}$ represents coefficient ring for time dependent perturbation on node Δ. If Z_1 represents kernel of boundary map, then:

$$I \in Z_1 \tag{7}$$

$$\partial I = 0 \tag{8}$$

$$\partial I = I_{\delta_\alpha} - (I_{\delta_\beta} + I_{\delta_\psi}) = 0 \tag{9}$$

$$I_{\delta_\psi} = I_{\delta_\alpha} + I_{\delta_\Delta} \tag{10}$$

where branches that enter or leave the node Δ, $\delta_i \in C_0, \delta_\beta = -\delta_\Delta$ is the time-dependent negative perturbation applied on node β and δ_α is the constant excitation on node α. Here Eq. (10) can be interpreted as a time-dependent perturbation current on node ψ with certain pre-bias current.

In a laser diode, when injection current is low, spontaneous emission due to carrier relaxation dominates and photon emission varies slowly as a function of injected current. Once the injected current approaches the lasing threshold, steady-state injected current approaches a constant value i.e. spontaneous emission is nearly constant and stimulated emission dominates. This emission in dominant mode is linearly dependent on injection current and each injected carrier results in a photon. Further, the relation between the resulting injected carrier density and optical intensity can be derived based on the laser rate equation for a single resonator mode and can be modeled in terms of resistors, capacitors, and inductors. The value of these RLC components can be calculated in terms of electron density and photon density (Katz et al., 1981). In addition, contributions of spontaneous emission and relaxation oscillation can be also expressed in terms of electrical parameters and electron-photon density (Katz et al., 1981).

$$R = \frac{R_d}{n_{photon} + 1} \tag{11}$$

$$L = \frac{R_d \tau_{photon}}{n_{photon}} \tag{12}$$

$$C = \frac{\tau_{spon}}{R_d} \tag{13}$$

$$R_{spon} = \frac{\beta R_d n_e}{n_{photon}^2} \tag{14}$$

$$R_o = \frac{-R_d \delta}{n_{sat}} \frac{1}{[1 + \frac{n_{photon}}{n_{sat}}]^2} \tag{15}$$

where $R_d = \frac{2kT}{q} \frac{1}{I_d}$ is called differential resistance of the laser diode, $n_{photon}, \tau_{photon}, \tau_{spon}$ are photon density, the lifetime of photon and spontaneous emission rate of electrons respectively. R_{sat} represents photon

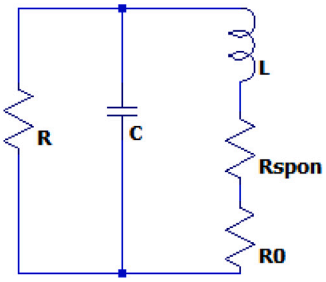


Fig. 3. Equivalent circuit model for laser diode (Katz et al., 1981)| $R = 2.555\ \Omega$, $L = 6.184$ pH, $C = 0.3557$ nF, $R_{se} = 2.811$ mΩ, $R_o = -5.511$ mΩ.

saturation density and β represents amount of spontaneous emission that couples into cavity mode of the laser diode. All the circuit components (inductor, resistance due to spontaneous emission and relaxation oscillation) related to optical phenomenon appears in series, as expected from laser rate equation. Differential resistance decreases with injection current, when below the threshold, and remains nearly constant when injected current is above the threshold.

4. Simulation

In this section, we discuss the simulation model for analyzing current injection profile into the laser diode. Additionally, using simulation, we show delay and amplitude tuning of the sub-nanosecond perturbation.

Fig. 4 represents the simulation model for generating indistinguishable excitation pulse. The shape tuning module comprises of a circuit capable of generating any arbitrary shape excitation. Shape and frequency of the excitation determines frequency and initial characteristics of each perturbation pulse. The shape-controlled excitation acts as an input for OPAMP and logic network block. The block comprises of operational amplifier and logical network to implement complementary output detector through a tunable differential delay generator. Tunability of the block further provides a selection of a wide range of pulse-width, amplitude and delay for the excitation to generate time-dependent perturbation. The constant node, α from Fig. 2, in the design is excited by a trans-conductance amplifier. Output node ψ is a time-dependent perturbation with some pre-bias current as per eqn.10. This excitation node then drives a laser diode. When driving an APD, a similar concept can be used to generate tuned voltage excitation. The design of these blocks is determined by the magnitude of current/voltage and timing requirements. OPAMP and the logic block was designed to generate perturbation of up to 25 mA of peak current. Had this requirement been higher, an additional driver stage would then be required.

Simulation model for the laser diode was obtained based on discussion in Section 2. It comprises of a parallel capacitor that arises due to carrier relaxation, differential resistance arising due to carrier injection, and inductance due to photon emission. Additional resistances are due to spontaneous emission and relaxation oscillation in the laser diode (Fig. 3). The values for the laser diode simulation model were obtained for a laser diode with threshold current value of 18.4 mA, at room temperature, and following parameters were considered: $R = 2.555\ \Omega$, $L = 6.184$ pH, $C = 0.3557$ nF, $R_{spon} = 2.811\ \Omega$, $R_o = -5.511$ mΩ.

In addition, a filter stage is designed based on the magnitude of attenuation required for filtering of the excitation pulse. Moreover, this stage can also provide feedback for control circuitry in the implementation.

Fig. 5 shows time dependent negative perturbation of around 7.5 mA generated at the constant current node, which corresponds to the node Δ in Fig. 1. The negative perturbation at β is applied at the constant current node Δ by OPAMP and Logic block through a bias tee network.

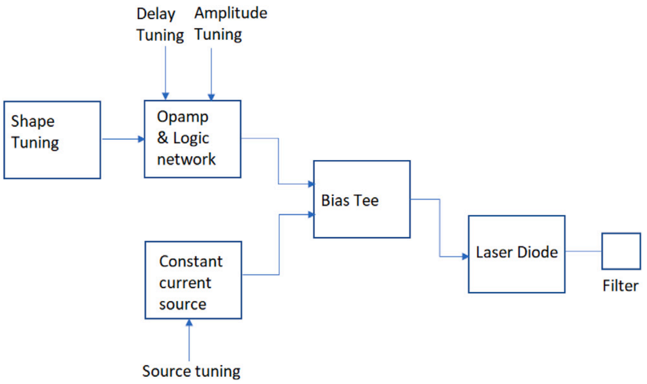


Fig. 4. Simulation block diagram | model can be adapted for APD instead of a laser diode | shape tuning circuit generates square, ramp or any arbitrary shape, OPAMP Logic network is a network comprising of operational amplifiers and FETs (can also be a digital IC) to generate tuned perturbation | Filter is used to further suppress any residual relaxation oscillation.

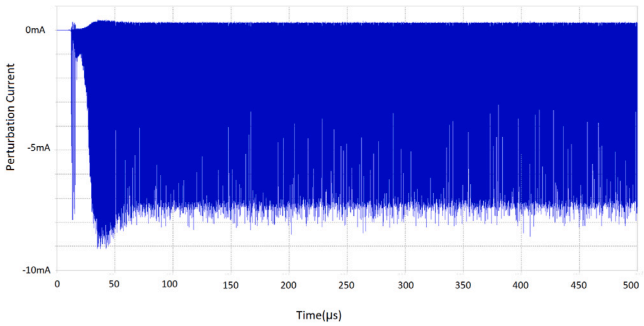


Fig. 5. Simulated negative perturbation applied at node β| the plot shows multiple sub-nanosecond perturbation.

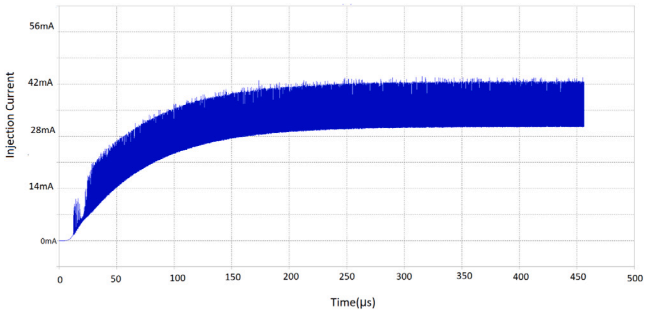


Fig. 6. Simulated current excitation profile at node ψ| The plot shows sub-nanosecond perturbation on a constant excitation of 31 mA.

Fig. 6 represents excitation current into the laser diode connected at the output node ψ in Fig. 1. The baseline current of 31 mA represents constant current excitation from node α. Fig. 7 represents zoomed-in version of the excitation shown in Fig. 6. Sub-nanosecond pulse of FWHM 600 ps can be observed. The simulation model does not completely consider parasitic effect of the PCB as the operating frequency is 100 KHz. In addition, any deviation in IC parameters from its model considered for the simulation will have an impact on the observed pulse characteristics. Fig. 8 represents tunability achieved in the laser excitation. Excitation delay was tuned by approximately 8 ns (red and blue). Moreover, amplitude of the pulse was tuned from 41.5 mA to

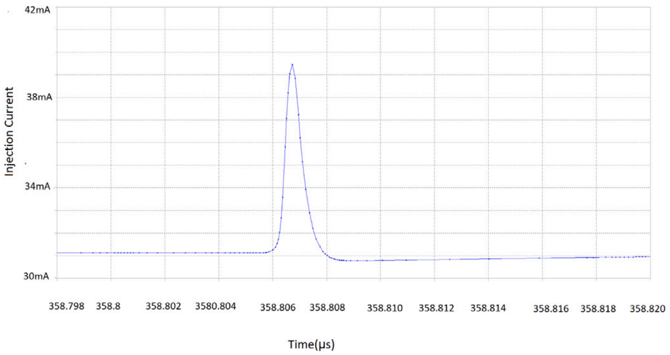


Fig. 7. Simulated sub-nanosecond pulse excitation as seen on node ψ | The plot is zoomed in version of Fig. 6.

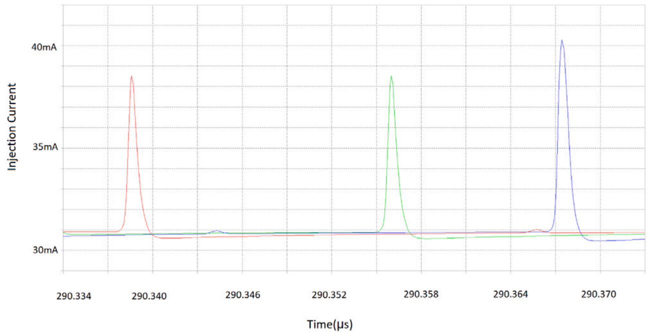


Fig. 8. Simulation of tunability in pulsed excitation | One of the controlled parameters was varied for tuning the excitation pulse. Plot in red and green demonstrate tunability in delay and plot in blue represents tunability in amplitude of the perturbation. Amplitude tuning can also be achieved by tuning the constant current excitation, which represents shifting the base current. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

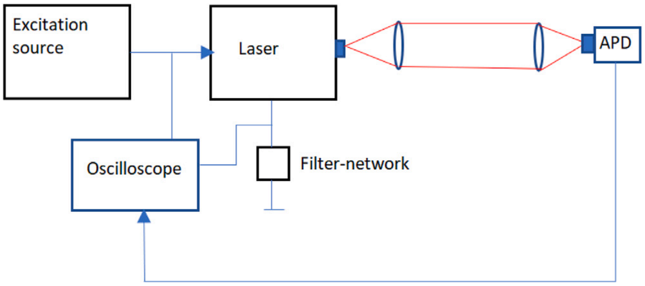


Fig. 9. Experimental setup for estimating current injection and output power of the laser diode | Excitation source represents node ψ. Differential current into the laser diode and APD, which detects focused laser (in red), is monitored using an oscilloscope. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

39.2 mA. Amplitude tuning can also be achieved by adjusting the constant excitation applied at node α.

5. Experiment

Experiment was carried out to verify sub-nanosecond, indistinguishable pulse generation from the laser diode, by measuring differential voltage across its terminals using differential probe D420-A-PB with a 4 GHz DX20 tip connected to 4 GHz Lecroy 640Zi Waverunner. Profile of the emitted optical pulses was estimated using an active probe ZS2500 with 2 GHz bandwidth and a custom built SAP500 APD detector operating in linear region. Since our receiver SAP500 APD is bandwidth limited, it does not provide accurate estimation of actual pulse width of the sub-nanosecond optical pulse. However, profile can be verified from the excitation perturbation across the laser diode. Fig. 9 represents a schematic of the experimental setup. Excitation source with pre-biased, sub-nanosecond current perturbation drives HL6748MG, a 670 nm laser diode (For cryptography application, it will be interesting to analyze coherence length and coherence time of this low-cost diode). In addition, a filter is connected in series with the laser diode to sense current and suppres any residual oscillation. Also, feedback from the current sense can be used to track any change in temperature. For our application in the automotive framework, we need extremely low-cost implementation. Since the laser driver is driven near to the threshold region, and that the pulsed excitation is less than 1 ns with frequency of 100 KHz, average case temperature of laser diode will be near to the room temperature. This way, temperature control will not be needed. Emission from the laser diode is monitored using custom- made SAP500 based detector with the APD biased in the linear region. The detector was selected based on efficiency corresponding to lasing wavelength of the laser diode. Fig. 10 characterizes differential injection pulse through the laser diode. The shortest FWHM of the excitation pulse was observed to be 496 ps. DC offset current due to pre-bias has been subtracted to underline pulse characteristics.

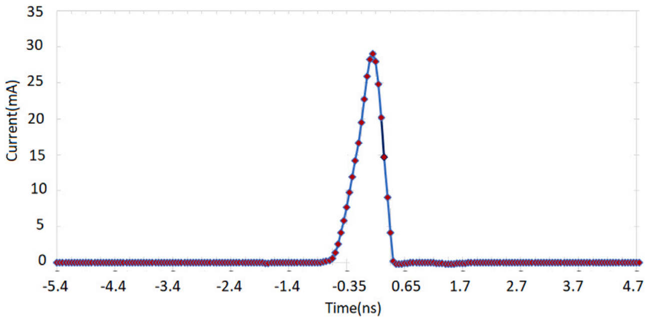


Fig. 10. Measured current excitation profile with least FWHM of 496 ps | constant current excitation has been subtracted to underline pulse characteristics.

Fig. 11 demonstrates fine-tuning of delay between current excitations by 60 ps. The voltage level of 2.346 V on zero reference of the excitation was shifted to 60 ps. Theoretically, the design should be able to attain any tuning required.

Fig. 12 represents the typical response of excited laser diode on a APD biased in the linear region. Typical linewidth as measured on the APD is 525 ps. This unexpectedly large line-width can be attributed to 500 ps of rise and fall time of the APD itself. The side oscillation is due to parasitic impedance on the PCB. Similarly, Fig. 13 represents APD response of another indistinguishable excitation with higher amplitude. Two indistinguishable pulses need to have the same characteristics in terms of shape, FWHM etc. .

Fig. 14 represents a comparative graph for amplitude normalized APD response. APD responses correspond to two different levels of excitation into the laser diode. Graph corresponding to state2 has been shifted to overlap with state1 for better comparison.

Fig. 15 represents the cumulative distribution function for state1 and state2 (Fig. 14). For two-sample Kolmogorov–Smirnov test, with D as 0.028, p-value as 0.998, for $\alpha = 0.05$. Since p-value is greater than α, it an acceptable hypothesis that both states came from the same distribution.

Since the indistinguishability in terms of waveform shape is critical for our application, and so far we have considered the indistinguishability in the temporal domain. We have not considered any distinguishability in frequency domain due to chirps. It will be interesting to study the impact on the spectral indistinguishability using the proposed approach.

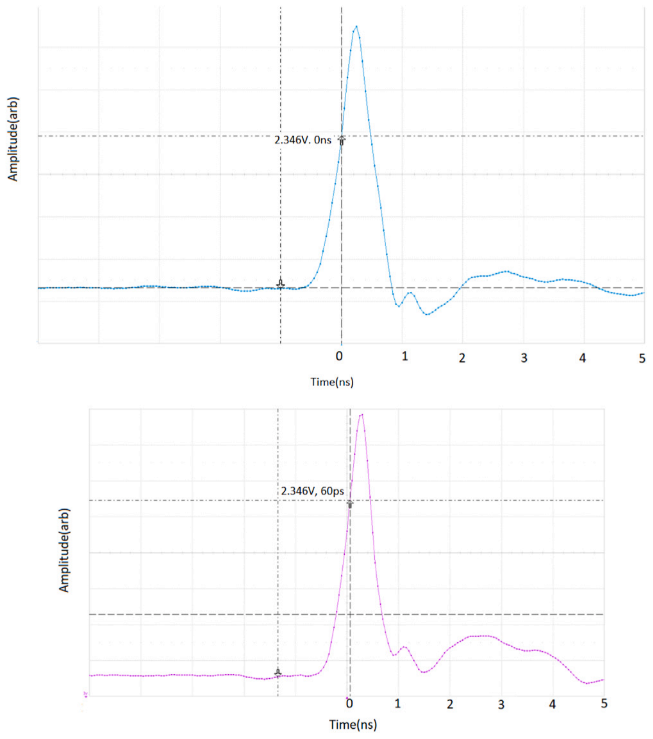


Fig. 11. Measurement of adjusted shift between two injection current profiles.

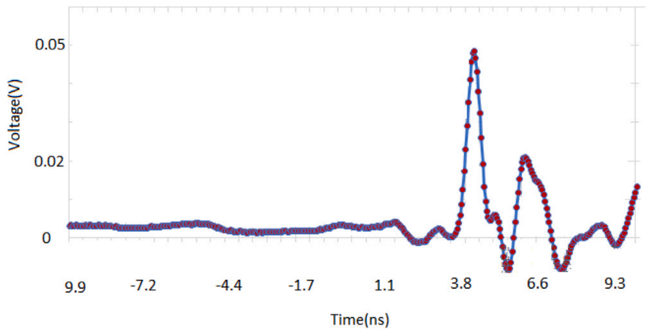


Fig. 12. Optical response of sub-nanosecond laser excitation as measured on APD (rise time and fall time 500 ps) | Additional peaks are due to parasitic impedance on the board.

6. Results

This work is based on perturbation of the constant current node to generate sub-nanosecond excitation pulse. Fig. 5 represents the negative perturbation applied on the constant current node *Δ* (Fig. 2). Consequently, positive perturbation appears at the anode of the laser diode. With controlled tuning of parameters, as long as the negative perturbation is identical, and traces between each node are maintained uniformly, excitation should be indistinguishable.

Fig. 8 represents the simulated degree of tuning in temporal domain as well as amplitude of the excitation. Depending on the delay observed in photon emission, we can always fine-tune these pulses to obtain indistinguishable photon pulses. Fig. 11 represents one such fine-tuning measured in our setup. Tuning of 60 ps is presented. In theory, we can program the tuning to achieve more minor adjustments.

It is crucial to optimize the design in terms of signal integrity. Specifically, in our setup, some oscillation due to parasitic inductance can be observed in steady-state region. On neglecting these parasitic oscillations, we can observe the indistinguishable excitation and corresponding photon emission by tuning one of our control parameters.

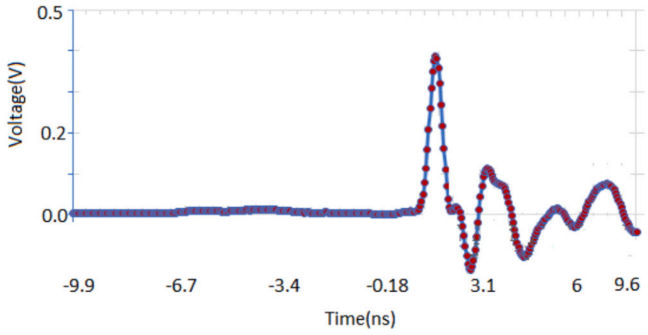


Fig. 13. The indistinguishable response measured on APD | Additional peaks are due to parasitic impedance on the board.

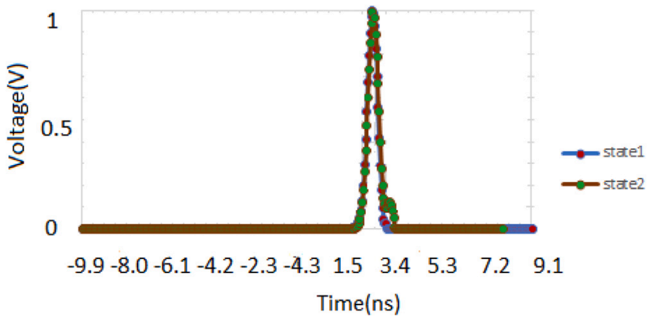


Fig. 14. Comparison of indistinguishability in the temporal domain for two different laser excitations| APD responses were generated by varying the amplitude of excitation in the laser diode| Amplitude of the graphs are normalized and shifted for better comparison. Further, oscillation due to parasitic inductance on the PCB has been removed.

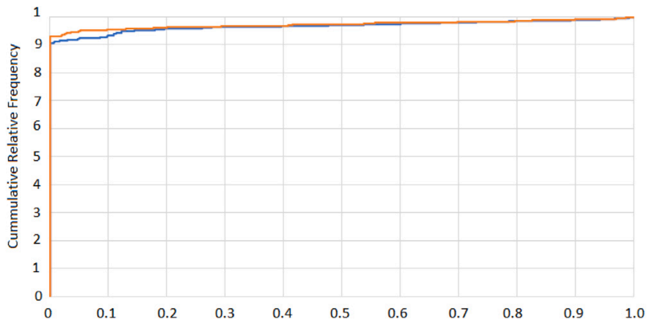


Fig. 15. The cumulative distribution function for state1 and state2 | Two-sample Kolmogorov–Smirnov maximum deviation estimated to be 0.0281, and *p*-value of 0.9979, for $\alpha = 0.05$ | since *p*-value is greater than α it is an acceptable hypothesis that state1 and state2 belong to the same distribution.

As seen from Figs. 14 and 15, from two sample Kolmogorov–Smirnov analysis, both of the photon detections belong to the same distribution. Hence, the indistinguishability. This approach can be further extended to single-photon regime for Quantum Key Distribution.

Moreover, it will be interesting to estimate the indistinguishability using this approach in the spectral domain. To be absolutely indistinguishable, the distribution must be same in temporal as well as in the spectral domain. We can extend this work to determine degree of indistinguishability in the frequency domain by fine-tuning of our design parameters.

Shortest FWHM of the sub-nanosecond pulse that we have observed is around 496 ps. Ideally, this value will be smaller if we neglect impact of bandwidth-limited oscilloscope probes. With rise time of the DX20-D420 probe system of approximately 122.5 ps, although we will obtain imprecise rise time information from the excitation profile, we can still observe oscillation peak, if any, which we can expect to be in the range of ~ 200 ps. Further, given bandwidth-limitation of our measurement system, we expect the APD measured FWHM to be much smaller than presented here.

7. Conclusion

We have designed a compact (4.5 cm × 3 cm), sub-nanosecond, tunable-indistinguishable laser pulse generator with the shortest excitation FWHM measured around 496 ps (Ideally, it will be less than 496 ps if we consider effects due to the probe). Further, this can also be used to generate shape-tunable excitation for an actively-quenched APD in single photon detection experiments, and to perform photon number resolution using APD. Indistinguishability in the spectral domain using the proposed technique should be explored.

References

Bunandar, D., et al., 2016. Apparatus and meth-ods for quantum key distribution. S10, 158, 481, B2.

Kardynał, B., Yuan, Z., Shields, A., 2008. An avalanche-photodiode-based photon-number-resolving detector. Nat. Photonics 2 (7), 425–428.

Katz, J., Margalit, S., Harder, C., Wilt, D., Yariv, A., 1981. The intrinsic electrical equivalent circuit of a laser diode. IEEE J. Quantum Electron. 17 (1), 4–7.

Linke, R.A., Gnauck, A.H., 1983. High speed laser driving circuit and gigabit modulation of injection lasers. In: Single Mode Optical Fibers, Vol. 425. International Society for Optics and Photonics, pp. 123–126.

Lo, H.-K., Ma, X., Chen, K., 2005. Decoy state quantum key distribution. Phys. Rev. Lett. 94 (23), 230504.

Maan, P., 2020. Current injection based generation of indistinguishable Glauber-state and decoy-state optical signal. US17/119, 747.

Maan, P., 2021. Towards generation of indistinguishable decoy state-Glauber state using a tuned laser diode. arxiv:2111.01303.

Nordholt, et al., 2017. Quantum communicationsystem with integrated photonic devices. US 9, 819, 418, B2.

Smale, S., 2000. On the mathematical foundations of electrical circuit theory. In: The Collected Papers of Stephen Smale: Volume 2. World Scientific, pp. 951–968.

Wabnig, J., et al., 2015. Secured wireless communication. US 9, 641, 326, B2.

Wei, Z., Wang, W., Zhang, Z., Gao, M., Ma, Z., Ma, X., 2013. Decoy-state quantum key distribution with biased basis choice. Sci. Rep. 3 (1), 1–4.

Williams, J., 2004. Signal sources, conditioners and power circuitry. Linear Technol. Corp. Appl. Note 98, 20–21.

Yin, H.-L., Zhou, M.-G., Gu, J., Xie, Y.-M., Lu, Y.-S., Chen, Z.-B., 2020. Tight security bounds for decoy-state quantum key distribution. Sci. Rep. 10 (1), 1–10.

Yuan, et al., 2017. Interference system and an interference method. US 9, 696, 133,

Surface enhanced infrared absorption spectroscopy using plasmonic nanostructures: Alternative ultrasensitive on-chip

Madhulita Mohapatra, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, madhulitamohapatra@gmail.com*

Laxminarayan Mishra, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, laxminarayan.s@gmail.com*

Prithivraj Nayakmr, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, p_nayakmr@gmail.com*

Pranay Rout, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pranayrout93@gmail.com*

ARTICLE INFO

Keywords:
Plasmonic
SEIRA
Biosensors
Nanostructures
Nanoantennas

ABSTRACT

The increasing demand for more reliable and cost effective biosensors has steered development of various novel biosensor platforms. Sensors based on optical techniques have attracted much attention because of their advantages, such as immune to electromagnetic interference, multiplex and cost effective. Surface enhanced infrared absorption (SEIRA) is a powerful vibrational label-free analytical technique that has benefited from collective excitation of surface plasmon to enhance vibrational signals of thin molecular layers. Owing to surface plasmon phenomenon to enhanced vibrational signal, SEIRA has attracted much research attention t as an alternate promising biosensor platforms and in recent years, continuous SEIRA sensing platforms are emerging. In this brief review, we outline the principle of SEIRA, in particular, theoretical models that give insight into the light interactions with plasmonic nanostructures and the coupling between nanostructures and vibrational molecules. With this, commonly used active SEIRA substrates and their optimization methods are discussed, ranging from materials, geometrical design and commonly used fabrication methods. Also discussed are advancements in analytical applications of SEIRA and brief future perspectives on SEIRA sensing platform.

1. Introduction

For decades, optical techniques have been widely used in various fields of sciences, in industries, and in other fields of applications for investigations of chemical and biological substances, monitoring of processes, and identification and detection of biomolecules (Bjerke, 2002; Hartstein, 1980; Janneh, 2016; López-Lorente, 2016; Stuart, 2004; Osawa, 2006). Sensors based on this technique is advantageous in terms of high sensitivities with short response times, immune to electromagnetic interference and to perform nondestructive measurements, also far from the sample under analysis. In particular, recent advances in different analytical transduction techniques that rely on spectroscopic methods such as surface-enhanced Raman scattering (SERS), surface-enhanced infrared absorption (SEIRA) and surface plasmon resonance (SPR) (Adato, 2013a,b; Galarreta, 2013; Michieli, 2015; Wang, 2011, 2013) have paved new research avenues.

Like SERS, SEIRA was first observed in 1980 by Hartstein et al. (Hartstein, 1980). It was shown that when a monolayer of molecule is absorbed on randomly arranged silver (Ag) or gold (Au) metasurface, the vibrational bands of the molecular monolayer can be significantly enhanced by many order of magnitude in the infrared (IR) spectrum. Since this discovery, many research groups have demonstrated innovative attempt to understand the mechanistic aspects of SEIRA with various metals (Ag, Au, Cu, etc.) employed in analysis/characterization of chemical and biological substances (Adato, 2013a,b; Krauth, 1999; Osawa, 1993). SEIRA mechanism is attributed to the electromagnetic (EM) and chemical effects, which contribute to the total enhancement. The EM effect is based on the principle of surface plasmon, in which the coupling of photons with the metastructures and dipole interactions between the megastructures and the adsorbed molecules signal can be significantly enhanced in the infrared (IR) absorption band (see Fig. 1 (a)). The enhanced absorption intensity of the molecules is proportional to the local field enhancement of incident light. Fig. 1(a) illustrates the SEIRA enhancement principle for spectra of resonant NanoAntenna (NA) before and after coating with the absorbing vibrational molecules. The sharp dips in the resonant NA correspond to the enhanced molecular absorption bands.

Several theoretical model that explain the optical properties and the

EM field enhancement on rough metal surfaces and NA have been extensively investigated (Coronado, 2003; Krauth, 1999; Feldheim, 2002; Hui, 2021), with many experimental results reported (Di Meo, 2021; Osawa, 1993; Di Meo, 2020). The chemical mechanism in SEIRA is explain by the several variations observed in the infrared vibrational fundamental, which are results of molecule-enhancer interactions that may affect the frequency, the shape of the observed infrared band. However, although the chemical mechanism is not fully understood, some experimental results assuming some chemical interactions between the surface and molecules are also reported (Osawa, 1993; Zhang, 2001; Osawa, 2006) .

The plasmon resonance frequency of metallic nanostructure ensembles such as sharp and round-edges plasmonic NA can be fine-tuned in the Near and Mid-infrared EM spectrum to amplify the local field plasmonic effect called "Hot-spot" (see Fig. 1(b)). Owing to the advances in cutting-edge fabrication techniques, enabling the realization of simple and complex plasmonic NA morphologies on SEIRA substrates at costs effective (Aksu, 2013; Bagheri, 2014, 2015; Cataldo, 2012; Kyo, 2021; Meo, 2019). In the recent years, numerous unprecedented studies on various plasmonic NA have been employed as SEIRA amplifiers for label-free detection. For example, in term of biosensors, Hui et al. used plasmonic metamaterials for rapid, label-free, and ultrasensitive detection of miR-155 (Hui, 2021). Di Meo et al, used multi-wavelength plasmonic NA for detection of vitamin D and achieved and enhancement factor of 10^5 (Di Meo, 2020). Just recently, the same group (Di Meo et al) used multispectral plasmonic nanostructures for detection of DNA and ach-ieved enhancement factor of 10^6 for the different plasmonic nano-structure pixels (Di Meo, 2021). In another work, an infrared probing of protein using SEIRA method were demonstrated (López-Lorente, 2016; Seiça, 2021). In term of single cell, Domenici et al. used plasmonic nanoparticles for monitoring single cell dynamic chemical effects (Domenici, 2019). In addition to these innovative research de-velopments, there are also many other SEIRA breakthroughs in remote environmental monitoring (Fuglerud, 2020), gas and chemical sensors (Chong, 2018; Meo, 2019), and charge transfer (Wang, 2019).

The focus of this review is to present SEIRA spectroscopy method as an alternative promising ultrasensitive on-chip biosensor. In particular, the first section provides brief understanding of SEIRA spectroscopy's mechanism. The second section is towards design and optimization of plasmonic NA as SEIRA substrates and the near and far-field enhancement factor. The third section discuses briefly analytical applications of SEIRA as suitable label-free detection method. In the final section of this review, the future development trends of SEIRA technologies are discussed.

2. SEIRA principles

Similar to SERS, SEIRA is a spectroscopic technique that exploits the EM properties of an engineered metal nanostructure to enhance the vibrational signatures of a molecular monolayer and sub-monolayer. Therefore, when a metal nanostructure is excited through the coupling of the incoming EM waves, collective oscillations of electrons at a metallic nanostructure interface occur, and this phenomenon is referred to as plasmonic resonances (Gomez, 2012; Neubrech, 2017; Zhou, 2021). With the advances in near field optical microscopy technology, local near field for surface enhanced spectroscopy are more understood and can be interpreted correctly (Neumann, 2015; Dregely, 2012). Fig. 2 (a) reports the near-field distribution of single linear NA plasmonic resonance in the infrared spectrum. The electromagnetic field confinement, also called hot spot of the NA are mainly confined around the edges of the NA (Neumann, 2015). For a half-wave dipole NA, the relationship between length of the NA and resonance frequency can be expressed accordingly to (Neubrech, 2017) as

$$\lambda = \frac{2L}{m} n a_1 + a_2 \tag{1}$$

where m is a mode number, n is the refractive index of the surrounding medium. The coefficient a_1 is related to the phase associated with the reflection at the NA edges and a_2 depends on NA's geometry and material parameters. In the visible and near-infrared EM region, the coefficients $a_{1,2}$ become very important for NA with smaller sizes (Neumann, 2015; Sönnichsen, 2002). However, for plasmonic applications in the mid-infrared EM spectrum, damping becomes more important because it can significantly influence the NA resonances. Therefore, in order to achieve an optimum SEIRA enhancement, careful design of NA geometrical structure is required (Neubrech, 2017). According to Neubrech et al. (Neubrech, 2008), when a mono-layer octadecanthiol (ODT) is absorbed on a plasmonic nanowire and excited by EM field with polarization along the long axis of the nanowire, the vibrational band of the ODT is enhanced with strong absorption modes of CH_2 group observed near 3000 cm^{-1} (black curve, see Fig. 2(b)). In another hand, when the exciting EM field is perpendicularly polarized, no enhanced absorption bands is observed (red curve, Fig. 2(b)). It explains that the vibrational bands of ODT was enhanced by nanowire.

When a NA is excited with EM field in the absence of absorbing substances, the NA spectrum however, can be presented as a Lorentzian-type line-shape, which can be expressed as follows (Neubrech, 2008; Zhou, 2021)

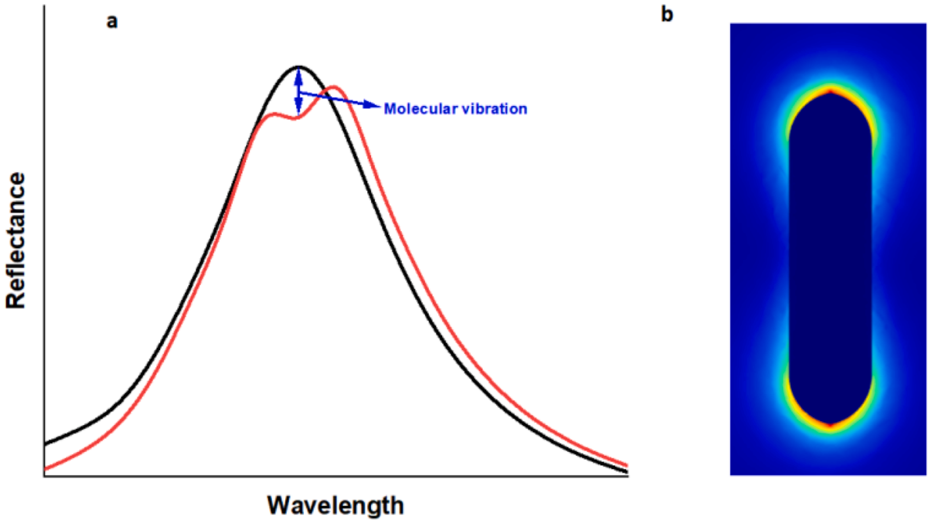


Fig. 1. (a) Illustration of SEIRA enhancement principle based on plasmonic metastructures: (red curve) vibrational bands of molecules absorbed on plasmonic nanostructure are enhanced and (black curve) is tuned plasmonic resonance to match to the vibrational molecule bands. (b) Simulated plasmonic nanostructure showing near-field confinement (Hot spot) at the edges of the NA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

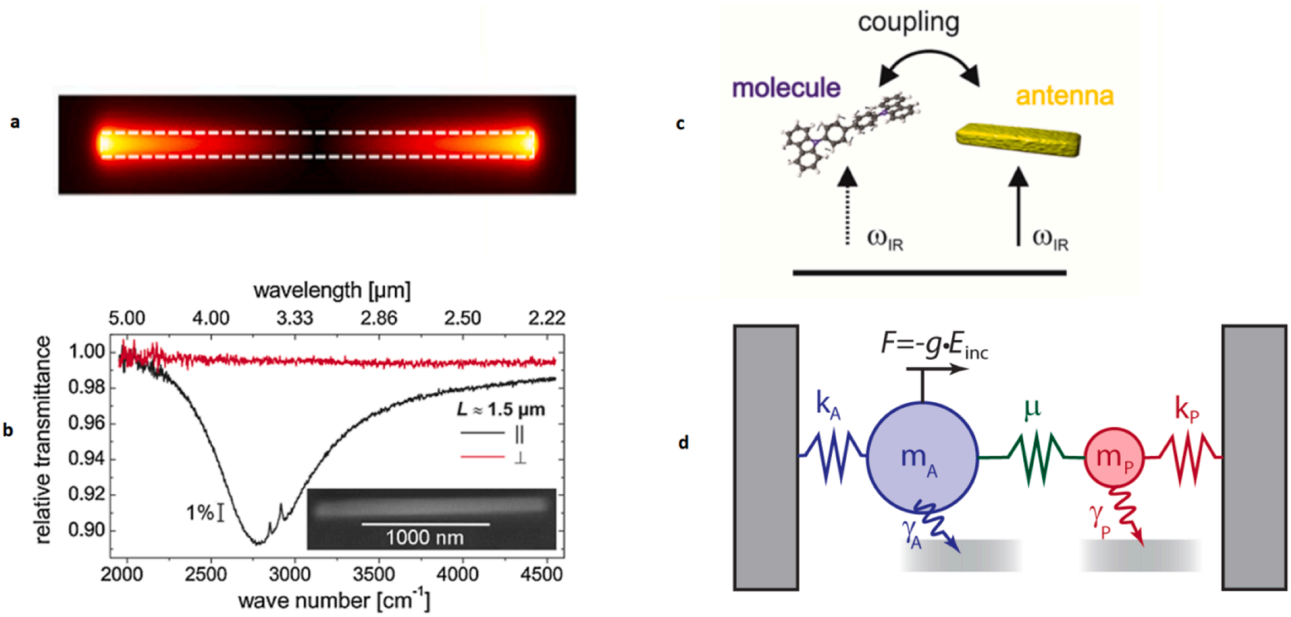


Fig. 2. Principle of SEIRA mechanism. (a) Enhanced local near-field profile of gold NA in p-polarized excitation (Neumann, 2015). (b) Relative transmission spectral of gold nanowire covered with one ODT monolayer for p-and s-polarized excitation (Neubrech, 2008). (c) Schematic illustration of excitation paths in Fano-resonances. (d) Damped harmonic oscillator conception of a coupled bright (A) and dark (P) mode (Adato, 2013; Osley, 2013; Torma, 2014). (b) Adopted from Ref (Neubrech, 2008). Copyright 2008 American Physical Society. (c) Reprinted with permission from Ref. (Neubrech, 2017). Copyright 2017 American Chemical Society. (d) Adopted with permission from Ref. (Adato, 2013). Copyright 2013 American Chemical Society.

$$L(\omega) = \frac{\alpha}{(\omega - \omega_0)^2 + \alpha^2} \tag{2}$$

where ω_0 and α are the resonance frequency and the resonance line-width, respectively. In contrast to the Lorentzian, the Fano-resonances with asymmetric line-shape can be described in the form (see Fig. 2 (c)) (Luk'yanchuk, 2010; Neubrech, 2017)

$$I\alpha \frac{(F_y + \omega - \omega_0)^2}{(\omega - \omega_0)^2 + \gamma^2} \tag{3}$$

where ω_0 and γ are standard parameters that denote the position and width of the resonance, respectively. F is the so-called Fano-parameter, which describes the degree of asymmetry. The asymmetry arises from the constructive and destructive EM field interference between the plasmonic NA and the absorbing molecule interface. The plasmonic oscillation is an effect of broad Fano-resonance, whereby, the vibrational molecules are attributed to the narrow Fano-resonance (Osley, 2013; Zhou, 2021). Another well-known model that describes the coupling between the plasmonic NA and the molecular vibration is the coupled harmonic oscillators (see Fig. 2(d)) (Adato, 2013; Osley, 2013; Torma, 2014). When plasmonic NA is excited by EM waves, the plasmonic resonance acts as a simple harmonic oscillation, referred as "bright mode (see Fig. 2(d) label A)". Whereas, because of the low molecular cross-section, the direct interaction between molecular vibration and far-field incident IR is small. As a result, the molecular vibration is referred to as "dark mode (see Fig. 2(d) label B)". The dark mode is excited by coupling to the bright mode (Torma, 2014). As reported in Fig. 2(b), the molecular vibration is like a modulator in the plasmonic resonance, instead of directly appearing in the spectrum as an absorption.

3. Optimization of SEIRA detection sensitivity

3.1. *Active SEIRA substrates*

The most common structures used for SEIRA applications are plasmonic nanostructures (Adato, 2017; Neubrech, 2017). Noble metallic structures are mainly preferred. However, semiconductors and dielectric materials have been reported as active SEIRA substrates (Law, 2013). Infrared (IR) transparent materials (CaF2, Si, MgO, BaF2, ZnS, KBr, Ge, sapphire) and non-transparent materials like silica, carbon and metals are commonly employed as supporting active SEIRA substrates. The most basic plasmonic nanostructures commonly employed as SEIRA substrates are rod-liked nanostructures such as nanorods, nanowires and nanocrosses (Neuman, 2015; Nannan, 2018) because of their design simplicity and, when their aspect ratio is optimized (i.e., length and width), large enhanced local electric fields at the edges regions (hot spots) are reported (Meo, 2019; Neubrech, 2017; Dregely, 2012).

Geometrical structure shapes with sub-nanogaps size like bowties, dimers and asymmetric split H-shape have been extensively studied as an alternative favourable design for SEIRA platforms (Dong, 2017; Huck, 2014; Wallace, 2016). Coupling effects between adjacent NA and nanogaps can extremely enhance the local near field, as their local electric field enhancement strongly depends on the nanogaps size (Mbomson, 2017). Apart from above discussed structures, SEIRA substrates based on colloidal nanostructures have been proposed and optimized as SEIRA platforms (Cataldo, 2012; Wallace, 2016; Nannan, 2018).

The physical patterns of rod-liked and nanogaps NA are commonly realized by high-resolution nanofabrication techniques like electron beam lithography (EBL) or focused-ion beam (FIB) (Zhu, 2011) that have the capability of creating complex nanostructures with lower proximity effect, high resolution and high fidelity. With these techniques, a resolution up to 10 nm can be reached. However, structures like nanogaps antennas with sizes below sub-10 nm are difficult to realize with these techniques because of their resolution limitations (Zhu, 2011). In addition, the operational cost of these instruments are expensive and time-consuming. Micro/nano fabrication techniques such as direct laser writing (DLW) (Zhao, 2014) and colloidal lithography (Ashkarran, 2013; Cataldo, 2012) have been proposed as an alternative to ELB or FIB because of their low cost and scalability. The possibility of plasmon resonances tuning during post-fabrication etching to reduce the influence of the polarizability of the substrates makes colloidal lithography very promising and competitive method. High fidelity SEIRA substrates have been realized by using this technique (Cataldo, 2012; Nannan, 2018).

3.2. *Near and far-field optimization*

The average cross-section of vibrational molecules in the infrared absorption can be extremely enhanced through the SEIRA spectroscopy for practical applications as ultrasensitive detection and identification of bimolecular fingerprints (Meo, 2019). In particular, the SEIRA enhancement mechanism strongly depends on choosing the right geometrical design (i.e., shape and size), material compositions and surface functionality, and various approaches to optimize these parameters have been reported (Cobley et al., 2009; Brehm, 2006; Cetin, 2014; Herpin, 2018; Huck, 2019; Lee, 2015; Marcellis, 2017; Schnell, 2009; Vogt, 2015; Yoo, 2018; Herpin, 2021). Noble metals such as Au and Ag are the most popular materials used in SEIRA application because of their high coupling fidelity.

The enhancement factor (EF) is used as a figure of merit (FOM) to determine the enhanced signal strengths to standard infrared reference measurements as (Meo, 2019; Neubrech, 2008; Zhou, 2021)

$$EF = \frac{I_{SEIRA}}{I_0} \bullet \frac{\mathrm{A}_0}{\mathrm{A}_{\mathrm{SEIRA}}} \qquad (4)$$

where $\mathrm{I}_{\mathrm{SEIRA}}$ is the signal enhanced due to SEIRA substrate, I_0 is the reference signal. The surface area of both the nanostructure and reference structure covered by the absorbing molecules are given as $\mathrm{A}_{\mathrm{SEIRA}}$ and A_0 respectively (Neubrech, 2008).

Several numerical simulations and experimental studies on the advantages of optimizing and tuning SEIRA NA structural parameters to increase the coupling and enhance near field have been demonstrated (Huck, 2014; Law, 2013; Neuman, 2015; Nannan, 2018). Recently, Herpin et al., engineered a multi-resonant metasurface that enables significant enhancements of absorbing biomolecules across a broad mid-IR spectrum ranging from below 1000 to above 3000 cm^{-1}. In this work, a deep neutral network (DNN) algorithm was used to classify different biomolecules fingerprints (see Fig. 3(h)) (Herpin, 2021). In another work by Hulk et al., a dimer NA configuration composed of periodic coaxial apertures with a minimum gap of 3 nm was fabricated using EBL (Fig. 3(d)) (Huck, 2014). It was demonstrated that, by decreasing the gap between the dimer NA, SEIRA enhancement factor of 10^4 can be reached with gap-size below 10 nm.

In 2017, De Marcellis et al., reported a numerical simulation method based on general parametric optimization of plasmonic NA for SEIRA enhancement (Marcellis, 2017), and later Di Moa et al., used similar technique to realize cross-NA on silicon substrate (Meo, 2019). An enhancement of 10^4 was reported, which was mainly attributed to the

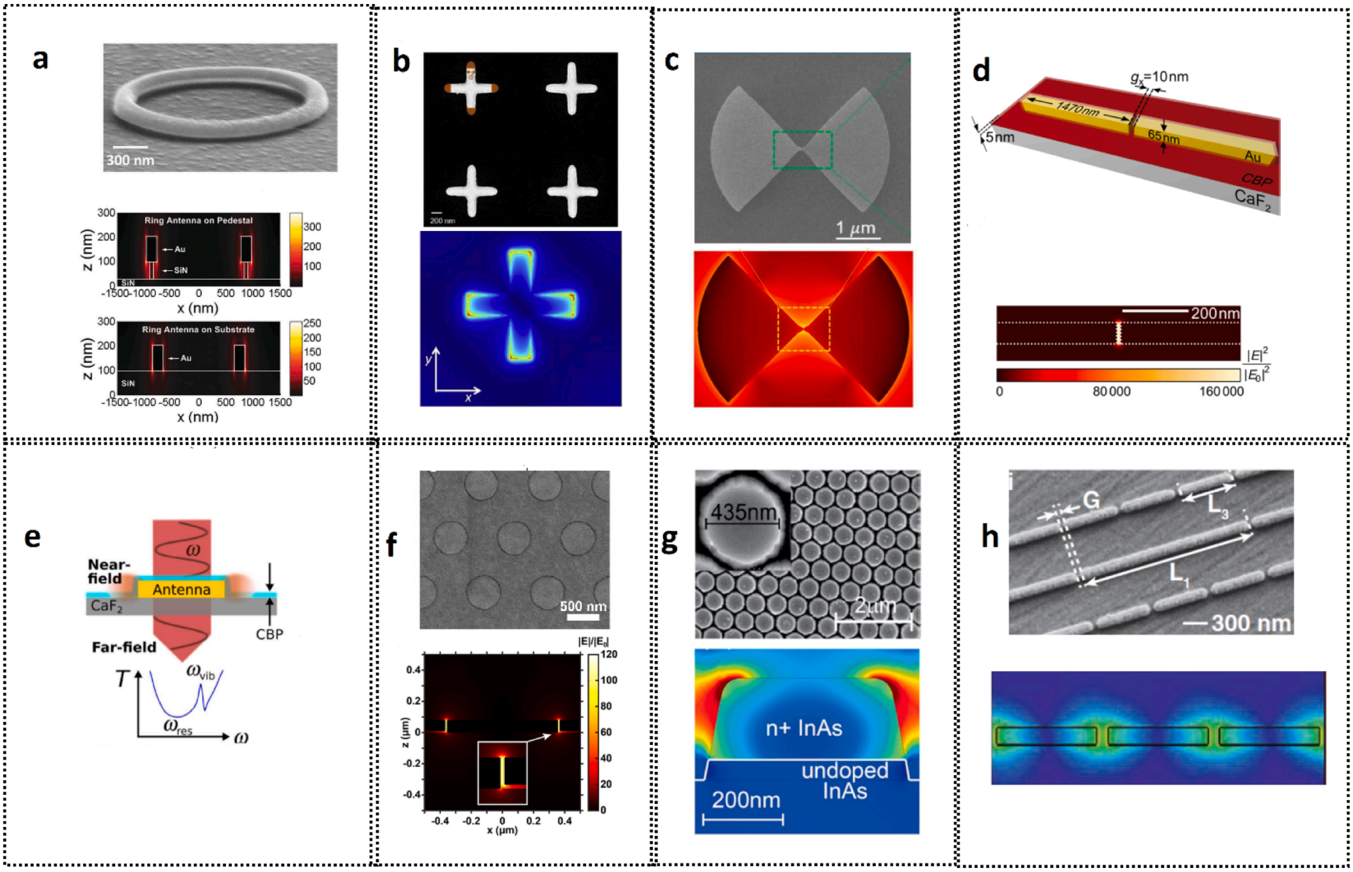


Fig. 3. Near-field enhancement for SEIRA: (a) Schematic illustration of gold nanoring antenna on the silicon nitride nanopedestal and simulated near-field intensity enhancement (Cetin, 2014). (b) SEM and simulated near field distribution of a cross-NA on silicon substrate (Meo, 2019). (c) SEM and simulated electric field distribution of a bowtie antenna (Dong, 2017). (d) Design of a dimer with a nanogaps and simulated field distribution (Huck, 2014). (e) Schematic view of resonantly excited (ω_{res}) NA of length (L) and near field enhancement of molecular vibrations (ω_{vib}) (Vogt, 2015). (f) Design view of a coaxial nanoaperture and simulated near field distributions in the coaxial nanoaperture with a 10 nm gap (Yoo, 2018). (g) SEM and simulated all semiconductor near electric field amplitudes at resonant wavelengths λ = 6 μm (Law, 2013). (h) SEM and simulated field distribution of a dual resonant array grating order-coupled nanogaps with antenna lengths L = 2.32 μm and L3 = 0.72 μm as well as inter antenna gap size G = 80 nm (Herpin, 2021). (c) Adopted with permission from Ref. (Dong, 2017). Copyright 2017 American Chemical Society. (d) Adopted with permission from Ref. (Huck, 2014). Copyright 2014 American Chemical Society. (e) Reproduced from Ref. (Vogt, 2015) with permission from the Royal Society of Chemistry. (f) Adopted with permission from Ref. (Yoo, 2018). Copyright 2018 American Chemical Society. (g) Adopted with permission from Ref. (Law, 2013). Copyright 2013 American Chemical Society.

general optimization of the NA parameters, allowing near-field lighting effect at the edges of the cross-NA (see Fig. 3(b)). To further improve the EF of SEIRA, Dong et al., fabricated a bowtie plasmonic NA with sub-nanometer gaps by using self-aligned technique and EBL fabrication method (see Fig. 3(c)) (Dong, 2017). A numerical enhancement factor of 10^7 reported strongly depends on the radius of the bowtie NA to enhance a monolayer of 4-nitrothiophenol (4-NTP) and 4-methoxythiolphenol (4-MTP) (see Fig. 3(c)). However, because of photon flux, the limit of detection of the SEIRA sensor was limited. A polarization-insensitive nanoring on a dielectric nanopedestal was fabricated by Cetin et al. (2014). As reported in Fig. 3(a), the enhancement mechanism depends on the circumference of the nanoring. Changing the circumference parameters enables tuning the plasmonic resonance and, as a result, the vibrational signal of protein was extremely enhanced.

To gain more insight on the near-field enhancement, Vogt et al., investigated the impact of near-and far-field energy shift on the SEIRA signal enhancement using Au nanorods (see Fig. 3(e)) (Vogt, 2015). By optimizing the vibrational signal enhancement for different tuning ratios between the molecular vibrational and the plasmonic resonance frequency, precise information on near-field and SEIRA enhancement was exploited.

Apart from the commonly used noble metal for SEIRA sensing, semiconductor materials also have gained more attention as suitable SEIRA substrate. Law et al., engineered all semiconductor plasmonic NA and fabricated using nanosphere lithography (see Fig. 3(g)) (Law, 2013). The author reported that when the NA's are fabricated to a length scale smaller than $\lambda_o/20$, a very weak vibrational signal can be detected. Furthermore, recent pioneer work by Yoo et al., shows that an engineered coaxial nanoaperture with gaps of 10 nm fabricated using photolithography and atomic layer deposition can enhance the vibrational bands of silk protein (Yoo, 2018). The huge near-field confinement of the coaxial apertures (see Fig. 3(f)), is as a result of coaxial apertures behaving as zero-mode resonator by allowing efficiently tunneling incident infrared light along 10-nm annular gaps resulting to an enhancement factor up to 10^5. In addition to the EF discussed above, there exist many other important benchmarks such as full width at half maxima (FWHM) optimization and quality factor (Q-factor) to characterize the SEIRA and give relevant insights about coupling between the nanostructures and vibrational molecules and suitable conditions for high SEIRA sensitivity. The limit of detection (LOD) is crucial for interpreting SEIRA sensitivity; therefore, the next section will discuss SEIRA as ultrasensitive biosensor for label free sensing.

4. Analytical application of SEIRA

The development of optical platform integrated with SEIRA spectroscopy technique has attracted much attention as a powerful and promising optical sensor technology. In particular, several biological applications have been reported using SEIRA, and it has been tested for various structural signatures of biomolecules, including immunoassay (Brehm, 2006; Dovbeshko, 2001). The amide fingerprint of a protein contains conformational information that is important for understanding its function in health and disease. SEIRA-based biosensors developed using star-shape Au was able to detect Vitamin D (25 (OH) D3 (calcifediol)) using FTIR measurement techniques, and a detection limit as high as 86 pmol/L reported (Di Meo, 2020). The same group recently demonstrated SEIRA data on ssDNA interactions with multispectral metasurface (Di Meo, 2021), using a single-stranded PNA (ssPNA) as recognition layer that supports the DNA base pairing mechanisms with a detection limit of 50 fM. The results reported by Hui et al. are very interesting (Hui, 2021). By using perfect absorber nanostructures interactions with miR-155, a limit of detection (LOD) of 100×10^{-15} m and a sensitivity of 1.162% pm^{-1} was achieved. The reported LOD is 5000 and 100 times lower than that using DNA single strand as probes and low than that of the fluorescence detection method, respectively. The recent innovative work by Yao et al. is very promising. They applied SEIRA to biosensor analysis for the determination of SARS-CoV-2 viral genomic segments (Yao et al., 2021). In this experiment, the SEIRA substrates are functionalized with the single-stranded DNA, which binds to selected SARS-CoV-2 genomic sequences. By using a statistical method based on the principal component analysis, they identified key characteristic differences between infected and control samples. The employed SEIRA platform enables rapid detection of SARS-CoV-2 with a detection limit of 1 μM viral nucleic acids within less than 5 min with no amplification and 2.98 copies per μL (5 aM) within 30 min when combined with the recombinase polymerase amplification treatment.

The investigation of nucleic acids and phospholipids structure from tumor cells, sensitive and drug resistant using SEIRA technique (Dovbeshko, 2001; Chekhun, 2002), has paved the way for the possibility of using SEIRA for diagnostic criteria in cancer research. SEIRA, as a vibrational spectroscopy, makes it possible to enhance vibrational bands of nucleic acid structural peculiarities from tumor tissues and nucleic acid interactions with anticancer drugs (Dovbeshko, 2002). In a more recent work by Huang et al. (2021), SEIRA was employed as a real-time and label free characterization of live cells and their responses. They used SEIRA substrate as a label-free phenotypic assay, allowing multiple detection of cellular responses to external stimuli: changes in cell morphology, adhesion, lipid composition of the cellular membrane, as well as intracellular signaling. In addition, by considering a multivariate statistical method such as principal component analysis (PCA), they analyzed phenotypic response barcode. In term of single cell, more interesting studies by combining SEIRA and scanning near-field optical microscopy (s-SNOM) have been showed. O' Callahan et al. proposed a method to obtain fingerprint IR spectra of ferritin proteins using s-SNOM (O' Callahan, 2019). The ferritin was deposited onto the sample by spin-coating 10 μL of an aqueous 12 μM protein solution at 3000 rpm for 20 s proteins and mapped with s-SNOM to find the surface topography and reveal the molecular signature of the protein. A zeptomole detection limit of protein reported was because of an effective coupling between plasmonic nanostructure and the vibrational polarization of the ferritin protein. In another study, spherical beads of poly- (methyl methacrylate) (PMMA) with 30–70 nm diameter and cylindrical TSV with 18 nm diameter were investigated using s-SNOM (Brehm, 2006). The single virus nanoparticles were cast from suspension and dried on Au/silicon substrate and mapped with s-SNOM to find the surface topography, IR amplitude, and phase contrast. IR amplitude and phase contrast were visible even at very small-lit probe volumes (10–20) and topography was visible at ~16 nm of the height of the virus nanoparticle.

5. Conclusion and outlook

SEIRA spectroscopy integrated with optical platforms has witnessed rapid development as a suitable alternative analytical biosensor to their counterpart SERS. Since it was first reported, several theoretical EM models have been developed and experiments performed to better understand the mechanism of SEIRA spectroscopy. In addition, the continuous advancement in numerical simulation tools and nano/microfabrication technology have made it possible to fabricate highly accurate and low-cost SEIRA substrates. Owing to these efforts, the myriad of SEIRA-based sensing platforms is continuously emerging, especially in biomedical applications to improve field enhancement, high selectivity and incredible limit of detection. Although recent years have seen several pioneer research works on SEIRA that with intuitive insight into different optimization method to increase SEIRA performance, ranging from material selection, structural design and coupling between the metasurface and the vibrational molecules, however, their performance are still limited and also, large part of sensing application based on SEIRA techniques are lab-based. Therefore, challenges to overcome in order to further improve SEIRA sensing platform are: (1) Overcoming SEIRA substrates fabrication challenges by employing a low-cost, scalable and high-resolution technique will pave the way for more reliable SEIRA substrates and will eventually lead towards practical industrial applications; (2) optimization methods that will lead to better understandings the coupling effect between nanostructure and vibrational molecules will further improve SEIRA enhancement; (3) new experimental setups are crucial for improving the sensitivity and selectivity; (4) employing machine learning algorithms based on statistical multivariate will make SEIRA as a suitable biomedical sensing platform. By investigating these issues, the potential of SEIRA as an analytical biosensor is very promising and competitive.

References

Zhu, et al., 2011. Lithographically Fabricated Optical Antennas with Gaps Well Below 10 nm. Small 7 (13), 1761–1766.

Stuart, 2004. Infrared Spectroscopy: Fundamentals and Applications, John Wiley & Sons, Ltd ISBNs: 0-470-85427-8 (HB); 0-470-85428-6 (PB).

Osawa, 2006. Surface-enhanced Vibrational Spectroscopy, Handbook of Vibrational Spectroscopy. John Wiley & Sons, Ltd.

Fuglerud, et al., 2020. Surface-Enhanced Absorption Spectroscopy for Optical Fiber Sensing. Materials 13 (1), 34. https://doi.org/10.3390/ma13010034.

Galarreta, et al., 2013. Microfluidic channel with embedded SERS 2D platform for the aptamer detection of ochratoxin A. Anal. Bioanal. Chem. 405, 1613–1621.

Wallace, G.Q., et al., 2016. Superimposed Arrays of Nanoprisms for Multispectral Molecular Plasmonics. ACS Photonics 3 (9), 1723–1732.

Wang, et al., 2011. Surface plasmon resonance detection of small molecule using split aptamerfragments. Sensors and Actuators B: Chemical 156 (2), 893–898. Wang, et al., 2013. Use of mercaptophenyl bosonic acid functionalized gold nanoparticles in a sensitive and selective dynamic light scattering assay for glucose detection in serum. Analyst 138, 5146–5150.

Adato, et al., 2013. In-situ ultra-sensitive infrared absorption spectroscopy of biomolecule interactions in real time with plasmonic nanoantennas. Nat. Commun. 4 (2154), 1–10.

Janneh, M., et al., 2016. Bandwidth Optimization and Frequency Tuning of Plasmonic Functionalized Metasurfaces for Optical Sensing of Chemical and Biological Substances. Procedia Engineering, 30th EuroSensor Conference.

Feldheim, et al., 2002. Metal Nanoparticles. Synthesis, Characterization and Applications. Marcel Dekker, New York.

Hui, et al., 2021. Infrared Plasmonic Biosensor with Tetrahedral DNA Nanostructure as Carriers for Label-Free and Ultrasensitive Detection ofmiR-155. Adv. Sci. 2100583.

Cobley, et al., 2009. Shape-Controlled Synthesis of Silver Nanoparticles for Plasmonic and Sensing Applications. Plasmonics 4 (2), 171–179.

Coronado, et al., 2003. Surface plasmon broadening for arbitrary shape nanoparticles: a geometrical probability approach. J. Chem. Phys. 119 (7), 3926–3934.

Di Meo, et al., 2021. Advanced DNA Detection via Multispectral Plasmonic Metasurfaces. Frontiers in Bioengineering and Biotechnology 9. https://doi.org/10.3389/fbioe.2021.66612110.3389/fbioe.2021.666121.s001.

Meo, et al., 2019. Metasurface based on cross-shaped plasmonic nanoantennas as chemical sensor for surface-enhanced infrared absorption spectroscopy, Sensors & Actuators: B. Chemical 286, 600–607.

Michieli, et al., 2015. Optimal geometric parameters of ordered arrays of nanoprisms for enhanced sensitivity in localized plasmon based sensors. Biosensors and Bioelectronics 65, 346–353.

Neubrech, et al., 2008. Resonant Plasmonic and Vibrational Coupling in a Tailored Nanoantenna for Infrared Detection. PRL 101 (15). https://doi.org/10.1103/PhysRevLett.101.157403.

Neubrech, et al., 2017. Surface-Enhanced Infrared Spectroscopy Using Resonant Nanoantennas. Chem. Rev. 117 (7), 5110–5145.

Gomez, et al., 2012. Surface plasmon hybridization and exciton coupling. Phys. Rev. B 86, 035411.

Neuman, et al., 2015. On the Importance of Plasmonic Scattering for an Optimal Enhancement of Vibrational Absorption in SEIRA with Linear Metallic Antennas. J. Phys. Chem. C. 119 (47), 26652–26662.

Neumann, et al., 2015. Mapping the near fields of plasmonic nanoantennas by scattering-type scanning near-field optical microscopy. Laser Photon. Rev. 9, 637–649.

Dovbeshko, et al., 2002. Surface enhanced IR absorption of nucleic acids from tumor cells: FTIR reflectance study. Biopolymers 67 (6), 470–486. https://doi.org/10.1002/bip.10165.

Dregely, et al., 2012. Vibrational near field mapping of planar and buried three-dimensional plasmonic nanostructures. Nat. Com. 4, 2237.

Schnell, et al., 2009. controlling the near-field oscillations of loaded plasmonic nanoantennas. Nat. Photonics 3 (5), 287–291.

Seiça, et al., 2021. Study of Membrane Protein Monolayers Using Surface-Enhanced Infrared Absorption Spectroscopy (SEIRAS): Critical Dependence of Nanostructured Gold Surface Morphology. ACS Sens. 6 (8), 2875–2882.

Sönnichsen, et al., 2002. Drastic Reduction of Plasmon Damping in Gold Nanorods, PRL, 031–9007chip sensing. International Journal of Optomechatronics 15 (1), 97–119.

Lee, et al., 2015. Sub-10 nm near-field localization by plasmonic metal nanoaperture arrays with ultrashort light pulses. Scientific Reports 5 (1). https://doi.org/10.1038/srep17584.

López-Lorente, et al., 2016. towards label-free mid-infrared protein assays: in-situ formation of bare gold nanoparticles for surface enhanced infrared absorption spectroscopy of bovine serum albumin. MicrochimActa.

Adato, et al., 2013. Engineered Absorption Enhancement and Induced Transparency in Coupled Molecular and Plasmonic Resonator Systems. Nano Lett. 13, 2584–2591. https://doi.org/10.1021/nl400689q.

Osawa, et al., 1993. Surface-Enhanced Infrared Spectroscopy: The Origin of the Absorption Enhancement and Band Selection Rule in the Infrared Spectra of Molecules Adsorbed on Fine Metal Particles. Appl. Spectrosc. 47 (9), 1497–1502.

Osley, et al., 2013. Fano resonance resulting from a tunable interaction between molecular vibrational modes and a double continuum of plasmonic metamolecule. Phys. Rev. Lett. 110, 087402.

Di Meo et al., (2020), Pixeled metasurface for multi-wavelength detection of vitamin D, Nanophotonic, 20200103.

Torma et al., 2014. Strong coupling between surface plasmon polaritons and emitters, PACS numbers: 33.80.-b, 73.20.Mf, 42.50.Nn.

Krauth, et al., 1999. Asymmetric line shapes and surface enhanced infrared absorption of CO adsorbed on thin iron films on MgO (001). J. Chem. Phys. 110 (6), 3113–3117.

Kyo, et al., 2021. Nanoimprint lithography for high-throughput fabrication of metasurface, front. Optoelectron. 14.

Bagheri, et al., 2014. Large-Area Antenna Assisted SEIRA Substrates by Laser Interference Lithography. Adv. Opt. Mater. 2 (11), 1050–1056.

Bagheri, et al., 2015. Fabrication of Square-Centimeter Plasmonic Nanoantenna Arrays by Femtosecond Direct Laser Writing Lithography: Effects of Collective Excitations on SEIRA Enhancement. ACS Photonics 2 (6), 779–786.

Adato, et al., 2017. Engineering mid-infrared nanoantennas for surface enhanced infrared absorption spectroscopy. Materials Today 18, 8.

Aksu, et al., 2013. Plasmonically Enhanced Vibrational Bio Spectroscopy Using Low-Cost Infrared Antenna Arrays by Nanostencil Lithography. Adv. Opt. Mater. 1, 798–803.

Law, et al., 2013. All-Semiconductor Plasmonic Nanoantennas for Infrared Sensing. Nano Lett. 13, 4569–4574.

Nannan, et al., 2018. Infrared-Responsive Colloidal Silver Nanorods for Surface-Enhanced Infrared Absorption. Adv. Optical Mater. 1800436.

Domenici, et al., 2019. Ultrasound delivery of Surface Enhanced InfraRed Absorption active gold-nanoprobes into fibroblast cells: a biological study via Synchrotron-based InfraRed microanalysis at single cell level. Sci Rep 9 (1). https://doi.org/10.1038/s41598-019-48292-0.

Dong, et al., 2017. Nanogapped Au Antennas for Ultrasensitive Surface-Enhanced Infrared Absorption Spectroscopy. Nano Lett. 17 (9), 5768–5774.

Mbomson, et al., 2017. Asymmetric split H-shape nanoantennas for molecular sensing. Biomed. Opt. Express 8, 395.

Zhao, et al., 2014. Hole-mask colloidal nanolithography combined with tilted-angle-rotation evaporation: A versatile method for fabrication of low-cost and large-area complex plasmonic nanostructures and metamaterials, Beilstein. J. Nanotechnol. 5, 577–586.

Zhou, et al., 2021. Infrared metamaterial for surface-enhanced infrared absorption spectroscopy: pushing the frontier of ultrasensitive on-chip sensing. INT. J. OPTOMECHATRONICS 15 (1), 97–119.

Zhang, et al., 2001. Study of Surface-Enhanced Infrared Spectroscopy: Dependence of the Enhancement on Thickness of Metal Island Films and Structure of Chemisorbed Molecules. J. of Colloid and Interface Science 233 (1), 99–106.

Ashkarran, et al., 2013. Controlling the Geometry of Silver Nanostructures for Biological Applications, Physics Procedia 40. Complete 76–83.

Chong, et al., 2018. Surface-Enhanced Infrared Absorption: Pushing the Frontier for On-Chip Gas Sensing. ACS Sens. 3 (1), 230–238.

Cataldo, et al., 2012. Hole-Mask Colloidal Nano lithography for Large-Area Low-Cost Metamaterials and Antenna-Assisted Surface-Enhanced Infrared Absorption Substrates. ACSNANO 6 (1), 979–985.

Vogt, et al., 2015. Impact of the Plasmonic Near- and Far-Field Resonance-Energy Shift on the Enhancement of Infrared Vibrational Signals. Phys. Chem. Chem. Phys. 00, 1–3.

Wang, et al., 2019. Surface-Enhanced Infrared Absorption Spectroscopy Using Charge Transfer Plasmons. ACS Photonics 6 (5), 1272–1278.

Yao, et al., 2021. Rapid detection of SARS-CoV-2 viral nucleic acids based on surface enhanced infrared absorption spectroscopy. Nanoscale 13, 10133.

Yoo, et al., 2018. High-Contrast Infrared Absorption Spectroscopy via Mass-Produced Coaxial Zero-Mode Resonators with Sub-10-nm Gaps. Nano Lett. 18 (3), 1930–1936.

Herpin, et al., 2021. Infrared Metasurface Augmented by Deep Learning for Monitoring Dynamics between All Major Classes of Biomolecules. Adv. Mater. 2006054.

Luk'yanchuk, et al., 2010. The Fano resonance in plasmonic nanostructures and metamaterials. Nature Mater 9 (9), 707–715.

Marcellis, et al., 2017. Design Optimization of Plasmonic Metasurfaces for Mid-Infrared High-Sensitivity Chemical Sensing. Plasmonics 12, 293–298.

Hartstein, A., et al., 1980. Enhancement of the Infrared Absorption from Molecular Monolayers with Thin Metal Overlayers. Phys. Rev. Lett. 45 (3), 201–204.

Herpin, et al., 2018. Quantifying the Limits of Detection of Surface-Enhanced Infrared Spectroscopy with Grating Order-Coupled Nanogaps Antennas. ACS Photonics 5, 4117–4124.

Huck, et al., 2014. Surface-Enhanced Infrared Spectroscopy Using Nanometer-Sized Gaps. ACS Nano 8 (5), 4908–4914.

Huck, et al., 2019. Chemical Identification of Single Ultrafine Particles Using Surface-Enhanced Infrared Absorption. Phys. Rev. Applied 11, 014036.

Bjerke, et al., 2002. Surface-enhanced infrared absorption spectroscopy of p-nitrothiophenol on vapor-deposited platinum films (Society for Applied Spectroscopy. Appl Spectrosc 56 (10), 1275–1280.

Brehm, et al., 2006. Infrared Spectroscopic Mapping of Single Nanoparticles and Viruses at Nanoscale Resolution. Nano Lett. 6, 1307–1310. https://doi.org/10.1021/nl0610836.

Dovbeshko, et al., 2001. Surface enhanced infrared absorption of nucleic acids on gold substrate. Semicond. Phys. Quantum Electron. Optoelectron. 4, 202–206.

Cetin, et al., 2014. Accessible Nearfields by Nanoantennas on Nanopedestals for Ultrasensitive Vibrational Spectroscopy. Adv. Optical Mater. 2 (9), 866–872.

Chekhun, et al., 2002. The SEIRA spectroscopy data of nucleic acids and phospholipids from sensitive- and drug-resistant rat tumors. J. Exp. Clin. Cancer Res. 21, 599–607.

O' Callahan, et al., 2019. Ultrasensitive Tip- and Antenna-Enhanced Infrared Nanoscopy of Protein Complexes. J. Phys. Chem. C 123, 17505–17509.

Huang et al., 2021. Monitoring the effects of chemical stimuli on live cells with metasurface-enhanced infrared reflection spectroscopy.

Miniaturized lenses integrated on optical fibers: Towards a new milestone along the lab-on-fiber technology roadmap

Prakash Chandra Sahu, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, pksahoo88@gmail.com*

Srikanta Pradhan, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, s.pradhan91@gmail.com*

Satyajit Nayak, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, satyajit_nayak@gmail.com*

Subhendu Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srikant.p@yahoo.co.in*

ARTICLE INFO

Keywords:
Flat Lenses
Optical fibers
Lab-on-Fiber
Refractive lenses
Diffractive lenses
Resonant lenses
Metasurfaces

ABSTRACT

Optical fibers and lenses are key components in many optical systems. Lenses used to couple light into the fiber or to tailor the output beam coming out from the fiber itself typically involve bulk optical schemes and complex packaging that limit the compactness and flexibility of the overall system. In the last years, new technological solutions based on the lab-on-fiber technology have been proposed to directly integrate optical components on the fiber end-face with a monolithic approach. In this review, we provide a detailed state of the art of the miniaturized lenses integrated on the end-face of optical fibers reported so far. The proposed devices have been divided into three main categories (refractive, diffractive and resonant including metasurfaces) according to their working principle. A critical discussion on the different categories of lenses is given, also providing a comparison in terms of ease of design and fabrication, application flexibility, compactness, scientific and technological maturity, and time to market.

1. Introduction

Lenses and optical fibers are key components in optical systems and are used together in many situations, ranging from light focusing, collimation, coupling to a light source, beam tailoring, and also for imaging and trapping applications. However, the large majority of these applications rely on free-space coupling schemes, involving the presence of bulk discrete optical components. Although these well-established technologies allow reaching good performances, their bulk dimensions and complex packaging limit their compactness and flexibility. Overcoming these limitations, reducing the device footprint, is crucial to improve the performance of the final optical system.

In this framework, the scientific community is continuously aimed at finding new solutions to directly integrate optical components on the fiber end-face with a monolithic approach, exploiting the advantages of having an inherently light-coupled platform like the fiber tip. In fact, having the lens constantly connected to the fiber, without requiring any free-space optics, eliminates the need for complicated alignment procedures, while increasing the device stability. However, the fiber facet is a unique and unconventional platform due to its small cross-section and large aspect ratio, so that it represents a challenge from a manufacturing point of view, requiring to adapt established fabrication technologies, used for planar substrates, to the peculiar geometry of the fiber termination.

The first attempts to integrate lenses onto optical fibers, in compact ready-to-use devices, date back to the late seventies (Cohen & Schneider, 1974) and mostly consisted of directly modifying the fiber termination on a macroscopic scale, resulting in surface roughness with poor control over the final geometry (Hillerich & Guttmann, 1989). Since then, many other configurations have been proposed. Nowadays, considerable steps forward in the fabrication process have been possible thanks to the “Lab-on-Fiber” (LOF) technology (Cusano et al., 2015). The advent of the LOF technology has allowed the realization of a new class of fiber-based devices (Pisco & Cusano, 2020), by integrating on the fiber tip advanced photonic and plasmonic structures (Ricciardi et al., 2015; Scaravilli et al., 2018) and multi-functional materials (Giaquinto, 2021; Giaquinto et al., 2019). In the last years, many manufacturing techniques have been developed, including “bottom-up” and “top-down” methodologies (Vaiano et al., 2016). The progress in the fabrication techniques has paved the way to new applications, that were previously unattainable, enabling the integration of increasingly complex structures on fiber facets. In this perspective, it is also possible to integrate on the fiber tip a new class of optical components that have recently been gaining attention: flat lenses (Yu & Capasso, 2014). This term is used in opposition to optics with curved surfaces, indicating a broad range of planar lenses typically having a sub-wavelength thickness, capable of controlling the light at the nanoscale. A flat lens has the advantage of greater design flexibility, other than compact dimension compared to its bulk counterpart (Capasso, 2018). While the latter is based on refraction to focus light, with a phase profile modulated by a continuous variation of the optical thickness, flat optics restrict the modulation to its minimum, imparting the desired phase by appropriately designing the lens geometry, without any change of the thickness. It is important to highlight that the integration of flat lenses on the optical fiber has the potential to introduce a new significant milestone along the technological roadmap pertaining to LOF, towards the development of fiber lenses, with reduced footprint and new remarkable functionalities.

In this review, a classification of different fiber lenses has been carried out according to their working principle, identifying three main categories (refractive, diffractive and resonant fiber lenses), schematically represented in Fig. 1. Various types of fiber lenses can be distinguished depending on their shape, dimension, and operating mechanism, even if is not trivial to systematically catalog them. Integrated refractive lenses will be first discussed, providing a general overview of different types of fiber lenses which exploit the same working principle of conventional free-space optics, namely modulating the phase by a continuous variation of the optical thickness. To this category belong all the lenses on the fiber tip having micrometer dimensions, for this defined microlenses, and also the so-called lensed fibers, obtained by directly tailoring the fiber termination. Subsequently, the attention will be focused on “flat optics”, indicating planar lenses having dimensions in the order of the wavelength or with sub-wavelength thickness, integrated on the fiber facet. Particularly, in the third section different types of diffractive lenses fabricated on the fiber facet will be presented. In this broad class, we will take into account optical elements which base their main working principle on diffraction, including properly designed diffraction gratings, Fresnel diffractive lenses, Fresnel Zone Plates, Fresnel Phase Plate. The fourth section will instead be devoted to resonant lenses, such as metalenses, which exploit the resonant behavior of optical sub-wavelength scatters properly arranged to form arrays on the end-face of the optical fiber.

For each lens typology discussed, we will provide an overview of the main manufacturing processes used to implement these lenses on the fiber tip and the main applications reported so far, also comparing them in terms of performances. Specifically, for this comparison we will mainly take into account the following parameters: numerical aperture (NA) and coupling efficiency, for focusing and coupling applications, respectively. The numerical aperture, NA = n sinθ (where n is the medium refractive index and θ is the maximum half-angle of light), indicates the ability of a lens to focus or to collect light. The coupling efficiency (Edwards et al., 1993; Kataoka, 2010; Panda et al., 2018; Sakai & Kimura, 1980; Zhou et al., 2020) instead, may be calculated by integrating the two light fields to be coupled (the one of the source E_s and the one of the fiber E_f):

$$\eta = \frac{|\iint E_s E_f dxdy|^2}{\iint |E_s|^2 dxdy \iint |E_f|^2 dxdy}$$

Finally, in the last conclusive section, we will highlight the advantages and disadvantages of each category discussed, providing a comparison in terms of ease of design and fabrication, application flexibility, compactness, scientific and technological maturity, and time to market.

2. Refractive lenses

The first type of fiber lens analyzed in this review is also the first to have chronologically appeared, consisting of optical elements with micrometric dimensions whose geometry essentially resembles that of ‘bulk’ lenses, directly formed or attached on the optical fiber tip. These “microlenses” have some similarities to conventional bulk refractive lenses since are based on the same working principle of refraction and are designed according to geometrical optics but, thanks to their smaller size, they can be integrated on the end of an optical fiber. Although within the generic name of “microlens” a vast range of micrometric optical elements could be included, in this section, we will restrict our attention to the main fiber microlenses able to focus light by essentially relying on refraction. The microlens geometry and material are the most influential parameters and during the years various designs have been proposed for improving the performances. Specifically, to maximize the power entering the fiber, alignment with sub-micrometer accuracy is required and, given the inherent difficulty of precisely transferring the microlens on the fiber tip (Xiong & Xu, 2020), there has been a growing interest in the so-called lensed fibers, which are optical fibers with the tip shaped like a lens. In this case, the fabrication takes place directly on the fiber termination, using the same material of the optical fiber, allowing to eliminate optical interfaces between the fiber and the lens, thus reducing the losses. Thanks to their compact structure and reduced number of components, in addition to simplifying the alignment procedure, lensed fibers also allow to increase the packaging stability of the final system.

Fiber microlenses have been used for focusing or for collecting light however, the most common application consists in achieving efficient coupling between the optical fiber and a light source, generally a

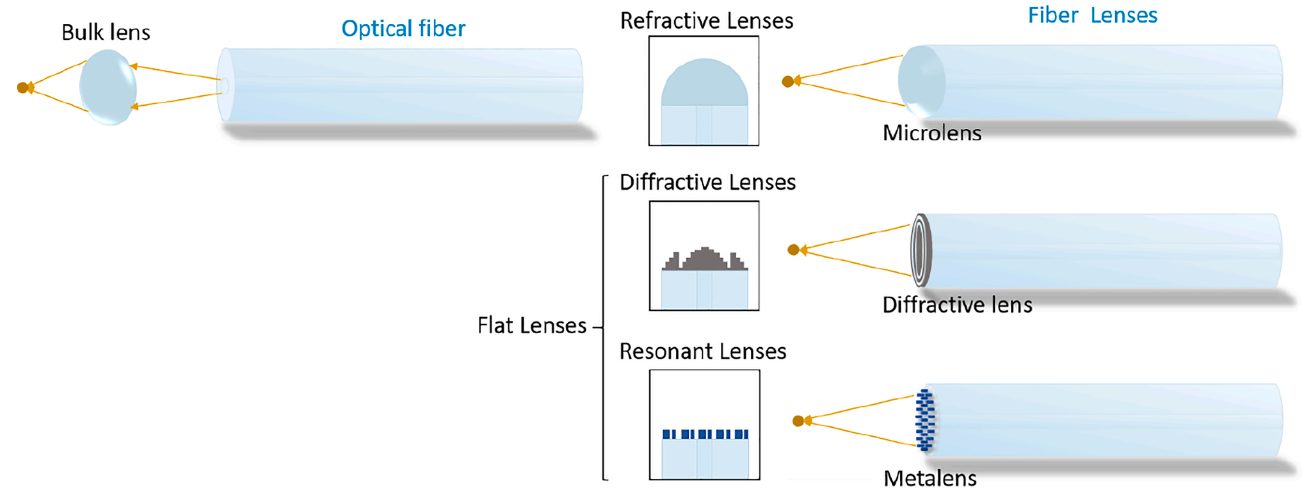


Fig. 1. Schematic representation of the different fiber lenses analyzed in this review for a common application such as light focusing, in comparison to a traditional free-space optical system based on a bulk lens (on the left). Fiber lenses, which are constituted by a lens directly integrated on the fiber facet, can be divided into refractive lenses, diffractive lenses, and resonant lenses. The latter two can be referred to as flat lenses for their compact dimensions.

semiconductor laser (Zhou et al., 2020). In order to enhance the coupling efficiency, different coupling scheme and microlens structures have been considered. Among the most common lens shapes there are hemispherical (Yamada et al., 1980), hyperbolic (Edwards et al., 1993), conical (Vitello & Eisenstein, 1982) and parabolic (Liu, 2008) ones. In theory, with hyperbolic microlenses should be possible to reach a 100% coupling efficiency (Mandal et al., 2018), and effectively a higher coupling efficiency of 87% has been reported in (Min Yang et al., 2010) with a tapered hyperbolic microlens, in respect to 62% achieved with a tapered hemispherical microlens. However, the use of hyperbolic microlens is limited by the expensive and difficult manufacturing process. The fabrication technique is indeed one of the main factors driving the development of fiber microlenses. Initial attempts of shaping the fiber facet were carried out by directly modifying the fiber tip by exploiting different techniques such as polishing, etching, high-temperature processes (for example arc discharge), or by depositing micro-particles. Although these fabrication methods are relatively simple, they did not allow for an accurate control of the final geometry. More specifically, the principal fabrication methods used for realizing microlenses can be gathered in physical methods and chemical ones, also with the possibility to combine both of them (Zhou et al., 2020). The physical method typically consists in directly removing the material through an external force in order to realize the desired shape; the chemical method, instead, "carves" the microlens through chemical corrosion, resulting in a more difficult control of the final shape. Plenty of different fabrication processes have been used to realize lensed fibers, including etching (Vitello & Eisenstein, 1982), photolithography (Cohen & Schneider, 1974), polishing (Lin, 2005), grinding (Lin et al., 2011), direct laser writing (DLW) Lin et al., 2014), and focused ion beam milling (FIB) milling (Melkonyan et al., 2019), just to name a few. Each of them has advantages and drawbacks, hence finding a manufacturing technique that combines low-cost, good performances and high-fabrication rate still remains an important objective to reach.

A common technique to fabricate lensed fibers consists in tapering the fiber down by acid etching and rounding its termination by melting. This simple method allows to automatically align the microlens with the fiber core, exploiting surface tension to obtain an approximately spherical shape, while the taper angle and melting degree define the curvature radius of the fiber tip and consequently the focused spot size. In Vitello and Eisenstein (1982) the authors exploited this manufacturing process to obtain a conical lens to match the modes of a single-mode fiber (SMF) and a laser diode, achieving low coupling losses of 3 dB. The etching method was also employed by (Ghafoori-Shiraz, 1988) to produce a conical microlens on an aluminium-coated fiber tip, obtaining a minimum coupling loss of 3 dB. To have a better control over the fabrication process, in various articles (Lee et al., 2004; Shiraishi et al., 2011; Tsai et al., 2008) the chemical method was used in combination with a physical one.

A similar coupling loss of 2.9 dB was achieved with a hemispherical microlens realized on the fiber facet with the electric arc discharge method, but in this case, alignment tolerances in the fiber axial and lateral directions for lens coupling resulted quite critical (Yamada et al., 1980). Even if a good reproducibility was reported in (Kuwahara et al., 1980) the tapered hemispherical fiber drawn by arc discharge allowed to obtain a coupling efficiency of only 35%. Other lensed fibers were realized by fusion splicer heating (Lay et al., 2004), by chemical etching (Kuchmizhak et al., 2014), by arc discharge polishing (Shiraishi et al., 2011), or by elastic polishing plate (Lin, 2005). A common purely physical approach to engineer lensed fibers is the grinding method. However, this fabrication method requires to constantly change and control the grinding speed and angle in order to optimize the surface curvature; furthermore it is not very promising for mass production due to a quite low efficiency (Yoda & Shiraishi, 2001). In (Lin et al., 2007) the grinding technique was used to realize an asymmetric elliptic-cone-shaped microlens, demonstrating an average coupling efficiency of 71%. A higher average coupling efficiency of 83% was measured by Liu et al. (2011), using a single-step grinding process to produce a double-variable-curvature microlens. Moreover, by exploiting the grinding and fusing techniques lensed fibers with a quadrangular pyramid shape (Huang et al., 2004) and a conical wedge (Yeh et al., 2005) were realized, obtaining in both cases a coupling efficiency of ~83%.

Cone-shaped microlenses with different radii of curvature were instead formed by means of the electrostatic pulling technique on a gradient-index fiber, by controlling the electric field intensity. An example has been reported by Wu et al., who demonstrated a coupling efficiency of 78% (Wu et al., 2011). The scanning electron microscopy (SEM) image of the realized prototype is shown in Fig. 2*a*.

A similar coupling efficiency of 72% was demonstrated with a lensed plastic fiber optic (Tseng et al., 2014) having a spherical shape obtained through electrostatic force. A high average coupling efficiency of about 80% was reported in Lin et al. (2017), where to fabricate a hyperbolic microlens on the fiber termination (Fig. 2*b*) was employed a three-step process combining grinding, spin-on-glass (SOG) coating, and electrostatic pulling. A low coupling variation of 0.116 ± 0.044% was achieved by appropriately tuning the radii of curvature.

Lee et al. also reported an hyperbolic microlens realized through a combination of etching and fusion methods (Lee et al., 2004), measuring a coupling efficiency with a laser diode of >82%, obtained thanks to an optimized mode matching, showing an improvement of 2 dB with respect to the performances previously obtained with a hemispherical microlens (Hillerich & Guttmann, 1989). In general, as stated before, a hyperbolic shape allows achieving better performances in comparison to those of hemispherical microlenses, improving the coupling efficiency of several decibels. In fact, most of the previously reported hemispherical microlenses do not overcome the value of about 55% (−2.5 dB) efficiency (Ghafoori-Shiraz, 1988; Ghafoori-shiraz & Asano, 1986). This is essentially because with hemispherical microlens it is difficult to obtain a small lens radius and a large aperture necessary for efficient coupling. Moreover, unwanted effects such as spherical aberration, mode-mismatch, Fresnel reflections, and fiber truncation cause high losses (Edwards et al., 1993). However, the shape of the microlens is not the only factor to take into consideration, since the performances also depend on the resolution of the fabrication technique. It is possible to obtain high-quality fiber microlenses with photolithographic techniques, however they are typically expensive and time-consuming (Wen et al., 2020). On the other hand, previously analyzed conventional lensed fibers fabricated by chemical etching, arc discharge, fusion, grinding, or polishing methods, are well established but generally not suitable for mass production (Yoda & Shiraishi, 2001).

An alternative manufacturing technique that has been shown to have good performance and speed is laser micromachining, which works by ablative removal of small portions of glass from the facet of the fiber optic. Besides, the lens curvature produced with this method appears to be more consistent with respect to those obtained by using traditional techniques, as reported in Presby et al. (1990). In this work, short intense laser pulses were used to melt the fiber tip in such a way to produce parabolic microlenses, controlling the radius through the proper choice of intensity and duration of the laser pulses. The results showed coupling losses of 1.5–3 dB, with a 2 dB improvement over lensed fibers fabricated with etching and melting techniques. A CO_2 laser was also employed to produce an hemispherical lenses on the tapered fiber, by melting its tip (Barnard & Lit, 1991) and to fabricate a polarization-maintaining fiber (Presby & Edwards, 1992), obtaining in the latter case a coupling efficiency up to 70% between the laser diode and the optical fiber. Always with a CO_2 laser an asymmetric hyperbolic microlens (Presby & Giles, 1993) was micromachined, achieving a coupling efficiency of 84% (−0.74 dB). Another example of hyperbolic microlens directly fabricated on the fiber tip by laser micromachining was proposed by the same authors (Edwards et al., 1993), improving the coupling efficiency to 90% (−0.45 dB), reducing the coupling loss to 0.22 dB due only to reflections. In (Dou et al., 2008) F_2-laser micromachining was used to create a microlens buried in the fiber surface in

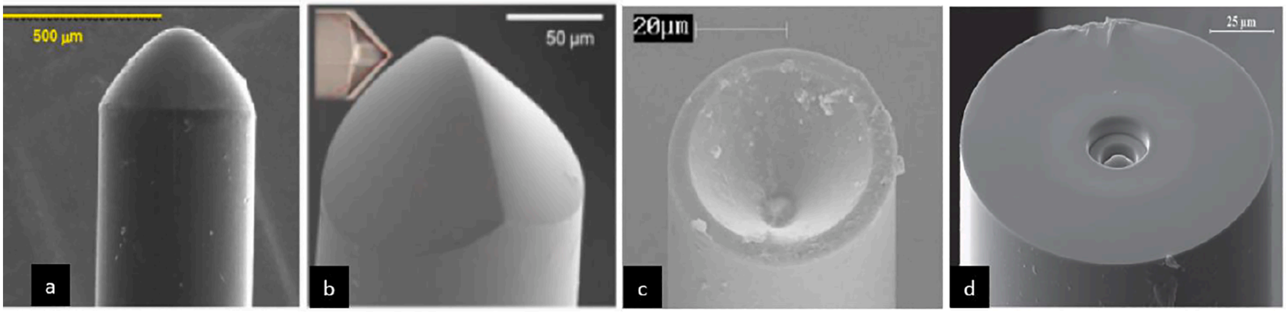


Fig. 2. SEM images of various refractive lenses: (a) aspherical lensed fiber manufactured with electro-static pulling (Wu et al., 2011) *Copyright © 2011 Optica Publishing Group*; (b) micro-hyperboloid lensed fiber realized with a three-step process (Lin et al., 2017) *Copyright © 2017 Optica Publishing Group*; (c) spherical polymer microlens produced through PDMS injection (Zaboub et al., 2016) *Copyright © 2016 Elsevier Optics Communications*; (d) embedded parabolic microlens fabricated by FIB milling (Melkonyan et al., 2019) *Copyright © 2019 Jphys Photonics*, Article licensed under a Creative Commons Attribution 4.0 International License.

order to focus the beam coming out from fiber termination. The microlens profile obtained, although appeared slightly rough due to the fabrication with an ablation spot, was approximately spherical. Despite the poor surface quality, with this type of lens was possible to produce in the proximity of the focal plane a nearly Gaussian beam shape, reaching a full width at half maximum (FWHM) of 2.0 µm, unlike conically shaped axicons.

In addition, a relatively simple and fast technique is represented by DLW, which allows fabricating multiple microlenses with high consistency and reproducibility. An example is represented by Lin et al. (2014), where the authors demonstrated a coupling efficiency of 53.5% at a long working distance of 16 µm. In another work (Gissibl et al., 2016b), 3D-DLW was used to construct a complex multi-lens device, consisting of various free-form lenses, all made out of the same material and having diameters of 120 µm. This compact optical system showed a large field of view of 80° and resolutions of 500 $lpmm^{-1}$. Specifically, this device is particularly suitable for imaging applications, since it also allows to compensate optical aberrations, such as field curvature, coma, astigmatism, distortion, by adding more refractive interfaces.

In the same year, Zaboub et al. reported on a work about micro-collimators (Zaboub et al., 2016), realizing spherical microlenses (shown in Fig. 2*c*) by injecting polydiethylsiloxane (PDMS) into a conical micro-cavity previously etched. The optimized values of waist (2.28 µm), working distance (19.27 µm), and curvature radius (10.08 µm) allow reaching a record-high maximum coupling efficiency up to 99.75% under specific circumstances.

As mentioned before, beyond the aim of improving the coupling between fiber optics and light sources, another typical application of microlenses is light focusing, which requires high NA (Kato et al., 2014). Indeed, some of the previously mentioned microlenses, used for enhancing the coupling efficiency, are characterized by a relatively small NA. On the other hand, in Kato et al. (2013) two different lensed fibers with high NAs and low losses were investigated for directly focusing the light coming out of the fiber into a sub-wavelength waist. In particular, the smallest FWHM reported was 0.62 λ, correlated to a NA of 0.85, also allowing direct coupling to a SMF. The spherically and hemispherically shaped tips were both realized by tapering down the fibers through a heat-and-pull process and then annealing the cut taper termination with a hydrogen–oxygen torch. The same authors, in another work (Kato et al., 2014), fabricated on an air-clad fiber an hemispherical microlens acting as a high-NA lens, in confocal microscopy. Interestingly, this lensed fiber was able to efficiently collect photons, reaching a coupling efficiency of around 25% other than showing a NA = 0.85 and a smallest waist FWHM of 0.61λ in this case as well. Another application was proposed by Rodrigues et al. (2018), who demonstrated an optical fiber tweezer for trapping. The polymeric microlens was formed on the facet of a modified optical fiber (a SMF terminated with a multi-mode fiber (MMF) segment) through a photo-polymerization process. With the so fabricated microlenses, various types of particles were trapped and distinguished.

In a recent article (Wen et al., 2020) a low-cost and simple method to produce microlenses, only requiring two steps (dripping and curing) was illustrated. In particular, exploiting the liquid surface tension, the authors realized hemispherical microlenses of different sizes, meant to focus the output beam while regulating the fiber divergence angle and reducing the beam focal spot size. Although the economic and flexible approach, the reduced control of the fabrication process prevents the microlens production on a mass scale.

On the other hand, FIB milling is gaining attention during the last years as a valid method to fabricate microlenses. As demonstrated by Li et al. this technique has the advantage of being compatible with non-conventional substrates, allowing to realize on the fiber tip complex structures with high efficiency, unlike more traditional manufacturing processes (Li et al., 2021). For example, in Melkonyan et al. (2017) a series of microlenses were fabricated on the fiber tip using this technique. Notably, a coupling efficiency slightly higher than that of a commercial tapered lensed fiber was achieved. This was possible thanks to an axicon lens carved in the fiber facet, converting the fiber-guided mode into a Bessel-type beam. In addition to enable efficient edge coupling, the embedded microlens, with an extended focus depth, ensured a larger alignment tolerance in the longitudinal direction. In one of their most recent works (Melkonyan et al., 2019), the same research group, fabricated another fiber lens engraved below the optical fiber surface using again FIB milling. Specifically, the embedded parabolic lens with a diameter of 15 µm and a height of 5 µm, shown in Fig. 2*d*, was engraved at a depth of 6.5 µm beneath the tip (a value equal to the focusing distance). Embedding the lens inside the fiber guarantees intrinsic longitudinal alignment and easier angular alignment too. This innovative design also increases compactness and efficiency, simplifying fiber-to-chip packaging while providing additional mechanical and environmental robustness. Directly bonding the fiber lens and the chip through physical contact is an unfeasible solution for commercial lensed fibers, which moreover present 1–1.5 dB lower coupling performance.

In Table 1 below some of the main articles analyzed in this section are reported, highlighting the principal characteristic of each fiber lens, such as the typology of optical fiber on which are realized, the lens shape and material, the manufacturing process used for the fabrication, and the type of application for which the fiber microlens is used.

3. Diffractive lenses

Although microlenses integrated on the fiber tip have represented a considerable step forward compared to their bulk free-space counterparts, in most cases the volume of these refractive elements is still significant, far from reaching an ultra-compact size. In addition, their applicability is limited by the restricted choice of microlenses shape and material, which largely influence the performances. Furthermore, a lens integrated onto the fiber end could be improved by further reducing the

Table 1
Comparison between the main parameters of selected refractive fiber lenses.

	Lens Characteristics				Declared Performances	
References	*Fiber type*	*Lens type*	*Material*	*Fabrication method*	*Application*	*Parameters of merit*
(Yamada et al., 1980)	SMF	Hemispherical microlens	fiber glass	Electric arc discharge	Light coupling	coupling efficiency: η = 49%
(Vitello and Eisenstein, 1982)	SMF	Conical microlens	fiber glass	Chemical etching	Light coupling	coupling efficiency: η = 51%
(Presby & Giles, 1993)	SMF	Asymmetric hyperbolic microlens	fiber glass	CO_2 laser	Light coupling	coupling efficiency: η = 84%
(Edwards et al., 1993)	SMF	Hyperbolic microlens	fiber glass	Laser micromachining	Light coupling	coupling efficiency: η = 90%
(Lee et al., 2004)	SMF	Hyperbolic microlens	fiber glass	Etching + fusion	Light coupling	coupling efficiency: η = 82%
(Lin et al., 2007)	SMF	Asymmetricelliptic-cone-shaped microlens	fiber glass	Grinding + arc heating	Light coupling	coupling efficiency: η = 85%
(Wu et al., 2011)	GIF	Aspherical cone-shaped microlens	polymer: SU-8	Electrostatic pulling	Light coupling	coupling efficiency: η = 78%
(Kato et al., 2013)	SMF	Sphericalmicrolens	fiber glass	Heat-and-pull +cutting + anneal	Light focusing	numerical aperture: NA = 0.85
(Kuchmizhak et al., 2014)	SMF	Hemispherical microaxicon	fiber glass	chemical etching	Light focusing	focal length: f~0.95 μm
(Lin et al., 2017)	SMF	Hyperboloidmicrolens	fiber glass	Mechanical grinding + SOG coating + electrostatic pulling	Light coupling	coupling efficiency: η = 84%
(Melkonyan et al., 2019)	SMF	Embedded Parabolic microlens	fiber glass	FIB milling	Light coupling	focal length: f = 6.5 μm spot size: 2.6 μm

thickness, especially in the perspective of realizing an ultra-compact packaging and improving the device overall efficiency. A method to obtain a thin planar lens on the fiber is to integrate on its facet a diffractive optical element (DOE), based on diffraction, instead of refraction. Diffractive lenses consist of multiple parts, with elements of dimensions in the order of the working wavelength, which operate by spatially arranging "zones" that impart a suitable phase in order to realize constructive interference of the transmitted light at the focus. By exploiting this strategy, it is sufficient to decrease the structure local period to obtain larger bending angles, without any thickness variations (Majumder et al., 2019). The design of diffractive Fresnel lenses (or kinoform diffractive lenses) is based on wave optics. This approach allows to obtain a considerable difference in terms of volume in comparison to their refractive counterparts (Gil et al., 2003).

During the last years, there has been an evolution of diffractive lenses leading to increasingly complicated structures. Different types of DOEs are reported for various applications that also go beyond the "simple" light focusing or coupling purpose. For example, diffractive lenses have been proposed to impart wavefront shaping of a light beam, as well as to redirect or split into different directions the propagating light (Kim et al., 2007).

Among the most common DOEs integrated on the fiber tip, it is also worth mentioning other sub-wavelength structures that could be considered as unconventional DOEs, such as Fresnel zone plates (FZP), Fresnel phase plates (FPP), as well as some periodic structures constituted by concentric rings or slits, acting as gratings. They can be differentiated between DOEs based on amplitude and phase structures. DOEs belonging to the first category are properly designed to essentially work by blocking part of the incident light; these amplitude elements are characterized by lower diffraction efficiency with respect to phase structures. In the second case, instead, DOEs can present binary, multilevel, or kinoform profiles to impart specific phase shifts to the light components. In particular, FZPs are amplitude structures in which opaque and transparent zones are alternated. In this way the light is concentrated onto a specific plane through concentric annular slits, suitably collocated to achieve constructive interference. In FPPs, instead, opaque zones are substituted by transparent π-phase steps, resulting in higher optical conversion efficiency due to the light no longer being blocked.

Binary phase elements are formed either by a single material but with elements of different thicknesses or by two materials with the same thickness, hence introducing two dissimilar phase shifts (Siemion, 2019). Blazed or multilevel diffractive lenses are instead proposed to enhance the focusing efficiency since they allow to approximate the ideal continuous phase distribution. In theory, using blazed diffractive lenses could be possible to achieve close to 100% efficiency (Hutley & Fleming, 1997). However, these optical elements suffer from chromatic aberrations, even if they can be reduced with multi-order diffractive lenses, implementable with any etchable material, not just polymers (Banerji et al., 2019). Nowadays, with fabrication techniques such as nano-imprint lithography (Koshelev et al., 2016), 3D printing (Asadollahbaik et al., 2019), electron beam lithography (EBL) (Sundaram & Wen, 2012), FIB milling (Rodrigues Ribeiro et al., 2017), is possible to engineer on the fiber tip more and more complex DOEs, with an unprecedented level of compactness and introducing new functionalities.

An example of a low-cost and reproducible manufacturing technique is nanoimprint lithography (NIL), used for example to realize a series of photonic structures on the fiber facet reported in Koshelev et al. (2016) for light manipulation. For instance, a Fresnel lens was realized on the fiber tip by means of UV NIL, using a material with a high refractive index, allowing efficient light focusing even in an immersion liquid, and obtaining a near-diffraction-limited focal spot with a diameter of 810 nm. In another article (Calafiore et al., 2016) the same authors reported the NIL-enabled realization of a 3D beam splitter for light manipulation, together with other 3D photonic structures directly onto the termination of the fiber. The same group has also provided examples of a high-refractive-index Fresnel lensed fiber, a vortex phase plate, and a beam shaper in Munechika et al. (2018).

On the other hand, FZPs and FPPs allow to further reduce the thickness, increasing the device compactness. Among the most common manufacturing processes to fabricate these structures, we find FIB milling, femtosecond laser micromachining, and EBL. For example in Sundaram and Wen (2012), micro FZPs, with focal lengths of about 3 μm, were implemented on the termination of multimode optical fibers by negative tone lift-off EBL. On the other hand, FIB milling was used to realize both a FZP (Fig. 3*c*) and a FPP (Fig. 3*d*) on mode expanded optical fibers for optical trapping (Rodrigues Ribeiro et al., 2017). The experimental results showed focal distances of ~ 5 μm and ~ 10 μm, respectively, while the maximum measured efficiencies, were about 38% for FZPs and 67% for FPPs, demonstrating higher losses in FZPs. In

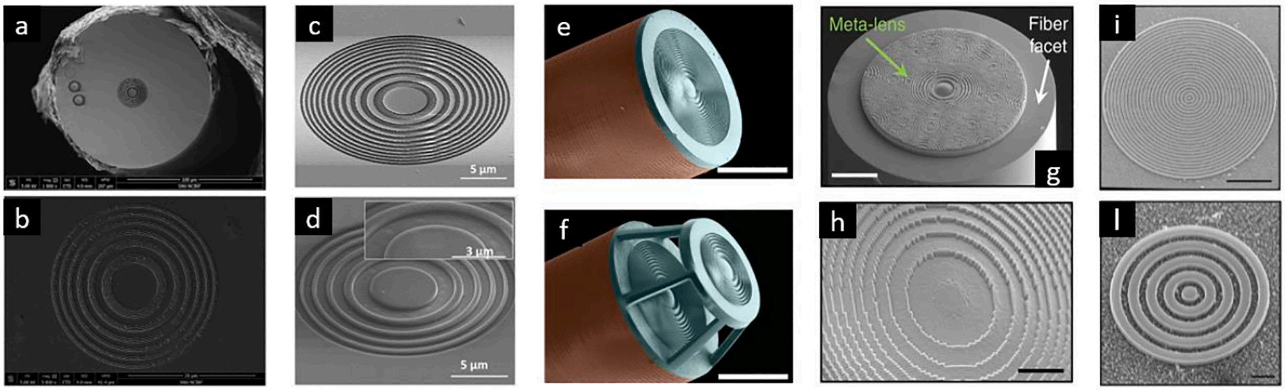


Fig. 3. SEM images of different kinds of diffractive lenses: (a-b) Metallic FZP (H. Kim et al., 2017) *Copyright © 2017 Optica Publishing Group*; (c) FZP and (d) FPP (Rodrigues Ribeiro et al., 2017) *Copyright © 2017 Scientific Reports,* Article licensed under a Creative Commons Attribution 4.0 International License; (e-f) 3D printed diffractive Fresnel lenses *adapted from* (Asadollahbaik et al., 2019) *Copyright © 2019 American Chemical Society* (g-h) ultrahigh NA *meta*-fibre (Pldschun et al., 2021) *Copyright © 2021 Light: Science & Application,* Article licensed under a Creative Commons Attribution 4.0 International License; (i-l) Nanostructure-Empowered efficient light-coupling *adapted from* (Yermakov et al., 2020). *Copyright © 2020 American Chemical Society.*

addition, a metallic FZP, whose SEM images are reported in Fig. 3*a-b*, was fabricated on the fiber tip to obtain super-variable light focusing. The tuning sensitivity of the proposed FZP was 20 times higher than the one obtained with a traditional spherical lens, offering the possibility to relocate the focal point by changing the incident wavelength. Nevertheless, the measured transmission efficiency was low (about 10% at the maximum) if compared to ordinary lenses (Kim et al., 2017). Furthermore, annular slits structures are characterized by high-order diffraction and short focal length, resulting in a restricted number of applications. An alternative to overcome these drawbacks was provided by Janeiro et al. (2016), with a photon sieve employed as a diffractive lens, where pinholes were distributed over the underlying FZP zones. In this way, it was possible to obtain higher resolution due to a higher NA and to the suppression of higher diffraction orders. The photon sieve was implemented via FIB milling on the fiber end, previously dip-coated with a conductive polymer (PEDOT:PSS) instead of using metallization. It was also demonstrated the feasibility of this device for light coupling applications, improving the coupling efficiency of 2.29 dB and achieving higher alignment tolerance in transversal and longitudinal directions. FIB milling was also used to fabricate a different kind of sub-wavelength focusing platform (Yuan et al., 2017) constituted by an achromatic super-oscillatory lens at the center of a photonic crystal fiber. This concentric ring nanostructure generated sub-diffraction hotspots but only 1–3% of the incident light was focused since the rest was mostly spread in halo rings surrounding it.

More recently, 3D nanoimprinting, based on Two-Photon Polymerization (2PP) is attracting the attention of the scientific community working in this field (Asadollahbaik et al., 2019; Hadibrata et al., 2021; Plidschun et al., 2021; Yu et al., 2020). This manufacturing technique enables printing complex structures with submicrometer resolution on any surfaces, including problematic substrates, such as the facet of the optical fiber. Specifically, this technique allows to digitalize arbitrary 3D nanostructures and, in comparison to other fabrication processes, is fast, cost-effective, and very versatile for a whole range of applications. Some examples of DOEs directly fabricated on the fiber tip by DLW have already been reported in literature. This approach has been used by Gissibl et al. (2016a) to realize on the fiber tip diffractive optical elements for spatial intensity beam tailoring, and in particular for shaping a doughnut or a top-hat intensity distribution. Since the feature size was below 100 nm, it was possible to produce an optical phase plate with a micrometric diameter (17.6 µm for the base and just 4.4 µm for the real diffractive element), to spatially redistribute the intensity distribution arising from the tip. With 3D printing was also implemented a spiral phase plate fiber for vortex beams generation (Ding et al., 2019), obtaining a 60.7% transmittance. With a similar approach, a vortex beam generator was produced on the fiber termination by Yu et al. (2020) over a composite fiber structure. The device was based on a kinoform spiral zone plate, derived from the superimposing of a spiral phase into a kinoform lens, and could efficiently convert the incident light into a single-focus vortex beam, avoiding unwanted order diffraction or focusing. In addition, the experimental results showed better performances compared to the ones of traditional fiber binary spiral zone plate, accomplishing 60% focusing efficiency and 86% vortex purity. Still, 3D-DLW based on 2PP was employed by Hadibrata et al. (2021) for an inverse-designed metalens on the fiber facet. The metalens was modeled as a circular grating and was able to convert parallel wavefronts into spherical ones, with a focal length of 8 µm at the operating wavelength.

The same fabrication method has also been used to realize diffractive lenses on the fiber end-face for trapping applications. For instance, in (Asadollahbaik et al. (2019) dual-fiber setups for optical trapping were demonstrated, having a trap stiffness 35–50 times higher than those achieved with conventional chemically-etched lensed fibers. This result was possible thanks to converging beams and NAs of up to 0.7 (Fig. 3*f*), enabling to produce a strong trapping efficiency in axial and transverse directions. The presented diffractive Fresnel lenses with their continuous design of the zone profiles, contrary to binary diffractive lenses, were able to suppress light in unwanted diffraction orders, allowing to approach 100% diffraction efficiencies (proportion of light at the focal position) even at low NA, while assuring a focal distance as large as 200 µm in an immersion medium. For similar trapping experiments, an even higher NA (nearly 0.9) was achieved by Schmidt et al. with their "*meta*-fibre" (Plidschun et al., 2021). The ultrathin lens (shown in Fig. 3*g*) was implemented on the end of a functionalized SMF (ending with a MMF segment for beam expansion). Even if the printing of a digitalized diffractive Fresnel lenses was already demonstrated in the previously mentioned work (Asadollahbaik et al., 2019), the higher NA achieved enabled successful particle trapping using only one platform for the setup. A kinoform phase distribution was obtained via the discretization of the phase profile of hyperbolic-type in 2π steps and, contrary to multilevel diffractive Fresnel lenses, the *meta*-fiber had an overall constant spatial resolution. Schmidt group has also contributed, in (Yermakov et al., 2020), at the integration on the fiber termination of a dielectric grating (images in Fig. 3*i–l*) with a high-index material (Si_3N_4), successfully reaching an in-coupling efficiency of 3 orders of magnitude higher than their previous plasmonic platform (Wang et al., 2019). This work will be discussed in more detail in the following section. The nanostructure, acting as a diffraction grating, was formed by dielectric concentric rings, implemented with EBL, providing polarization-insensitive in-coupling over a wide range of angles. In particular for angles $\theta > 60^{\circ}$, where conventional fibers cannot operate, were reported values of normalized in-coupling efficiencies of about 1% and 2% per TM and TE polarizations, respectively.

Some of the previously mentioned diffractive fiber lenses are compared in Table 2, showing the different features of each diffraction-based element.

4. Resonant lenses

If earlier works on "flat optics" have been mostly based on previously investigated diffractive lenses, recently the new technology of resonant lenses has emerged. In this category, metasurface-based lenses (also known as metalenses) are gaining considerable attention as a means to further reduce the overall thickness of the optical elements to sub-wavelength scale (Khorasaninejad & Capasso, 2017). These ultra-thin metasurfaces can be used to realize resonant lenses, which focus light by using a different working principle than their diffractive counterpart, allowing to completely control the full phase profile (range 0-2π). A metasurface-based lens consists of an array of subwavelength resonators, appropriately arranged in order to change the optical response of the interface. Meta-lenses exploit the resonant nature of these sub-wavelength optical scatterers, which introduce abrupt phase shifts, allowing to control amplitude, phase, and polarization of light (Lalanne & Chavel, 2017).

An initial classification of resonant lenses can be made considering the material of the metasurface building blocks, in particular, we distinguish between plasmonic and dielectric resonators. By suitably choosing their geometrical parameters (size, shape, orientation) and their spatial arrangement in precise patterns is possible to digitize and tailor the incoming wavefront at will. Different types of metalenses can be distinguished based on their phase mechanisms (Genevet et al., 2017). The first examples of metasurfaces to have chronologically appeared were plasmonic ones, consisting of metallic nanostructures, whereby the typically working principle relies on resonant electronic-electromagnetic oscillation known as localized plasmonic resonance. Anyway, to achieve wavefront control by introducing an arbitrary phase profile, it is necessary to completely cover the range from 0 to 2π, which can be achieved through one of the following strategies or a combination of them (Liang et al., 2019): localized resonances (e.g. plasmonic resonances and Mie resonances); extended resonances (e.g. high contrast gratings); and Pancharatnam-Berry phase in which the rotation of the scatterers controls the phase of a circularly polarized beam.

The metalens phase profile, for the case of normally incident light, is hyperbolic, guaranteeing a diffraction-limited spot (Yu & Capasso, 2014):

$$\varphi(r) = \frac{-2\pi}{\lambda}\left(\sqrt{r^2 + f^2} - f\right)$$

In the last years, different metalenses have been proposed using various types of optical resonators (metallic or dielectric antennas, apertures, slits, disks, holes, bars, etc.), and numerous articles have already been published showing a great interest in the topic. In this respect, we find metasurfaces used for shaping wavefronts at will obtaining light focusing (Paniagua-Domínguez et al., 2018) or polarization control (Desiatov et al., 2015), arbitrary reflection and refraction (Ni et al., 2012; Sun et al., 2012), holograms (Ni et al., 2013b), beam tailoring and deflection (Cheng et al., 2017) and so on. Since any planar ultrathin array of resonant scatterers at an interface, with a certain optical response could be seen as a metasurface, a comprehensive review of all the various types of metasurface-based lenses would be beyond the scope of this article (if interested in learning more on the topic we refer the reader to previous reviews (Capasso et al., 2017; Chen et al., 2020; Genevet et al., 2017; Hail et al., 2018; Lalanne & Chavel, 2017; Liang et al., 2019; Moon et al., 2020).

The first demonstration of a metasurface-based lens composed of subwavelength resonant elements directly realized on the fiber end-face has been presented by (Principe et al., 2017) with their "*meta*-tip". This article has represented a promising approach to achieve ultra-thin fiber lenses, based on the combination of metasurfaces and LOF technology.

However, despite their small footprint and planar geometry, it is not trivial to fabricate the metasurfaces subwavelength elements onto this small and unconventional platform, due to the already mentioned peculiar features of the fiber tip, as demonstrated by the limited number of fabricated devices reported until now. In particular, it is necessary to optimize the fabrication method employed for the specific case, with two different strategies possible: the metasurface can be either realized on a planar surface with conventional techniques and then transferred on the fiber end-face or it may be directly processed on the fiber termination via optimized manufacturing processes. Lithography is among the most common manufacturing method chosen to pattern metasurfaces on the fiber tip, since the nanostructures can be realized with a single lithographic step (Yu & Capasso, 2015). The first attempts of metasurfaces integration onto the optical fiber have mostly regarded plasmonic metalenses (Principe et al., 2017; Wang et al., 2019; Yang et al., 2019), while dielectric metasurfaces are generally more difficult to integrate on the fiber tip (Zhou et al., 2021). For example, the authors of the ground-breaking article (Yu et al., 2011), that revisited Snell's law at the interface between two uniform media by exploiting an array of plasmonic V-antennas, also successfully integrated a plasmonic metasurface on the fiber tip (Yu & Capasso, 2015). Such metasurface was

Table 2
Comparison between the main parameters of selected diffractive fiber lenses.

	Lens Characteristics				Declared Performances	
References	*Fiber type*	*Lens type*	*Material*	*Fabrication method*	*Application*	*Parameters of merit*
(Janeiro et al., 2016)	SMF	Phase photon sieve	fiber glass	etching + FIB milling	Light focusing	focal length: f = 0.5 μm spot size: 1.5 μm
(Koshelev et al., 2016)	SMF	Diffractive Fresnel plate	polymer	UV-NIL	Immersion application	focal spot diameter: d = 810 nm
(Kim et al., 2017)	MMF	Metallic FZP	silver	FIB milling	Light focusing	focal length: f = 20 μm
(Rodrigues Ribeiro et al., 2017)	SMF + MMF segment SMF + MMF segment	FZP FPP	platinum gold/ palladium	FIB milling FIB milling	Trapping Trapping	transmission efficiency: 38% focal length: f = 5 μm transmission efficiency: 67% focal length: f = 10 μm
(Asadollahbaik et al., 2019)	SMF	Diffractive Fresnel lens	polymer	DLW	Trapping	numerical aperture: NA = 0.7 (in water) focal length: f = 50 μm
(Yermakov et al., 2020)	SMF	Dielectric grating	Si_3N_4	EBL	Light coupling at large angles (θ > 60°)	in-fiber coupling efficiency: η = 14.2%
(Plidschun et al., 2021)	SMF + MMF segment	Diffractive *meta*-lens	polymer	DLW	Trapping	numerical aperture: NA = 0.88 (in water)

formed by an array of metallic nanoantennas fabricated by EBL and then transferred to the fiber termination using the decal transfer method.

Although not constituted by an array of subwavelength optical scatterers, there are some early examples of plasmonic structures which can be considered as resonant lenses, since their working principle is mainly based on plasmonic resonances. In this regard, Stief et al. directly fabricated a surface plasmonic (SP) lens on the gold-coated fiber tip by exploiting FIB milling (Stief et al., 2011). The SP lens, composed of three/four concentric annular slits, was designed for light superfocusing in both transverse directions, achieving far-field, sub-diffraction-limit sized focus through the excitation of surface plasmon polaritons. The obtained focal length was 1.2 μm with a transverse focal size of 0.31λ/NA and a calculated numerical aperture in water of 0.91. However the NA measured in air resulted too low for application in real case scenarios. Moreover, a similar fiber-based surface plasmonic lens was proposed by the same group for 3D optical trapping (Stief et al., 2013). A stable trapping, stronger than traditional optical tweezers, was demonstrated also using a reduced optical power with a consequent decrease of thermal effects caused by metal absorption and avoiding physical contact with the trapped objects.

The first real step toward the integration of metalenses on the fiber tip, to extend in an unprecedented way the (metasurface-enabled) light-manipulation capabilities of the fiber, is represented by the already mentioned work by Cusano group (Principe et al., 2017). The schematic of the metatip geometry, the illustration of its working principle, together with the SEM images of the fabricated prototypes are provided in Fig. 4*a–d*. This plasmonic metasurface was able to impart a constant phase-gradient along the chosen direction, resulting in a linear phase profile and dividing the transmitted beam into an ordinary component and an anomalous one that is subjected to the engineered steering. The design of the phase-gradient plasmonic metasurface was inspired by Babinet's principle, such as in a previous planar example (Ni et al., 2013a), presenting an inverted configuration, realized by patterning rectangular nanoholes of different sizes (half of which are rotated of 45°) on the gold-coated fiber tip, using FIB milling.

The same fabrication technique was recently employed by Yang et al. (2019) to create a plasmonic metalens on a large-mode-area photonic crystal fiber. This type of fiber has been chosen for the dimension of its core diameter, larger than that of a conventional single-mode fiber but maintaining single-mode properties. The in-fiber metalens (shown in Fig. 4*e–i*), with a phase gradient designed for light focusing in the telecom range, was realized by directly patterning a geometric phase (i.e. Berry phase) metasurface onto the previous deposited gold film. The measured values of focal lengths are 28 μm and 40 μm at $\lambda = 1550$ nm with numerical apertures of 0.37 and 0.23 other than a maximum operating efficiency of 16.4%, while the maximum enhanced optical intensity was as high as 234%.

Moreover, with metalenses is possible to address the problem of collecting light at large angles ($\theta > 30°$), a range not feasible with traditional optical fibers. Although the NAs achieved so far are still too low for realistic applications, in (Wang et al., 2019) the authors proposed a way to redirect light at large incident angles (up to 80°) by nanostructuring the fiber termination with plasmonic nanodots. EBL was used to produce this periodic array (Fig. 4*o–q*), following the same procedure reported in a previous article of the same authors to engineer nanotrimers on the fiber tip (Wang et al., 2018). Even if the resulting light-collection performances were higher than those of bare fibers, especially in regards to larger incident angles, the obtained coupling efficiency in the order of 10^{-4} to 10^{-5} was still too low for real case scenarios, because of weak diffraction and strong absorption in metals.

The intrinsic absorption losses in metal-based metasurfaces limit their performances, driving researchers to investigate metasurfaces formed by dielectric materials (Kamali et al., 2018). An example of a dielectric metasurface on the fiber tip has been provided by Zhou et al. (2021). In Fig. 4*l–n* are reported the SEM images and the schematic illustrating the operational principle of the fiber *meta*-tip designed for converting the light into a divergent vortex beam or a collimated beam for TE and TM polarization, respectively. This polarization-selective platform, composed of rectangular nanobricks, consisting of hydrogenated amorphous silicon (a-Si:H), was fabricated by a combination of EBL and plasma etching, then integrated on the fiber tip using a vision system. Furthermore, two fiber *meta*-tips so realized were used to form a

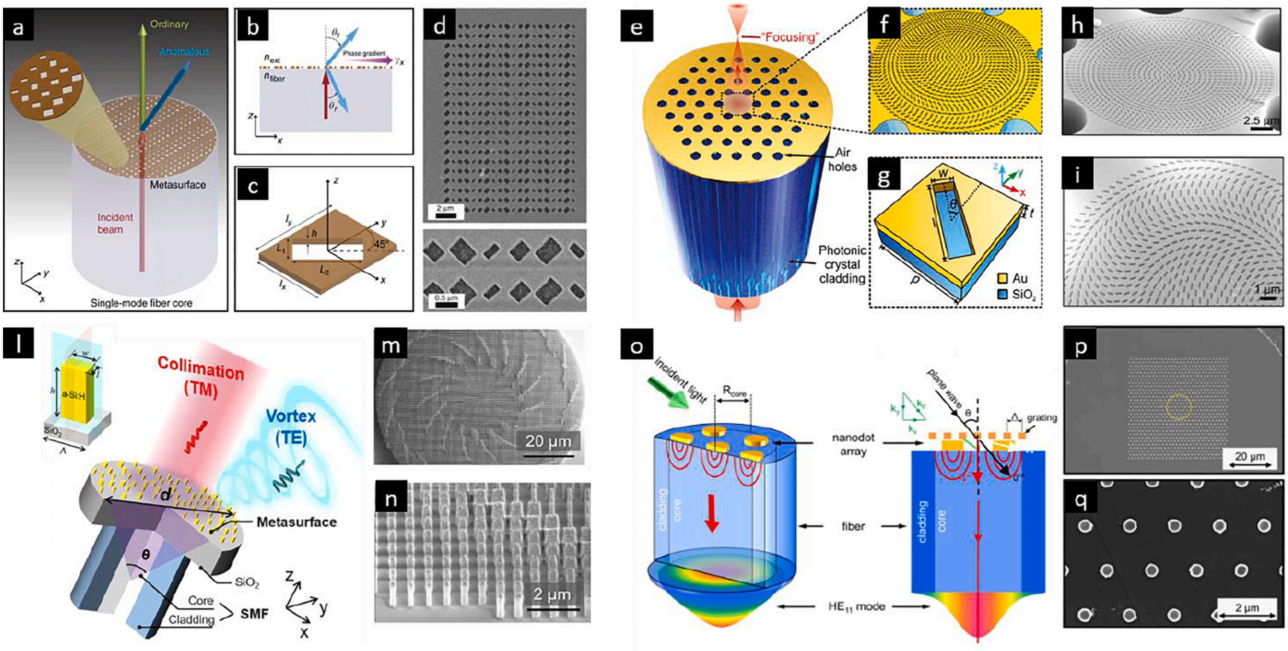


Fig. 4. Schematic illustrations (a,b,c,e,f,g,l,o) of the work principles and SEM images (d,h,i,m,n,p,q) pertaining to (a-d) optical fiber *meta*-tips (Principe et al., 2017) *Copyright © 2017 Light Science & Applications,* Article licensed under a Creative Commons Attribution 4.0 International License*;* (e-i) photonic crystal fiber metalens (Yang et al., 2019) *Copyright © 2019 De Gruyter,* Article licensed under a Creative Commons Attribution 4.0 International License*;* (l-n) all-dielectric fiber *meta*-tip (C. Zhou et al., 2021) *Copyright © 2021 Wiley Online Library;* (o-q) nanostructure-enhanced fiber adapted from (Wang et al., 2019) *Copyright © 2019 American Chemical Society.*

polarization-selective optical interconnect, obtaining a transmission efficiency of ≈57% for both polarizations.

The integration of all-dielectric metalens on the fiber tip can add to the optical fiber plenty of new functionalities with unprecedented performances, however their development is strictly connected to the evolution of the manufacturing processes. Until now, the preferred technique to create dielectric metasurfaces on the fiber tip has been EBL, showing feasibility for large-scale production, but the fabrication procedure should be still optimized.

The high interest in the subject is also underlined by the rising number of numerical articles being published in the last years. For example, a dielectric metalens made out of nanoscale rutile titanium dioxide (TiO2) pillars, acting as an immersion lens to enhance the collection efficiency, has been recently proposed (Zhang & Guo, 2020). Other interesting examples of numerical works concerning resonant lenses integrated on the facet of photonic crystal fibers are provided by M. Kim et al, with an in-fiber dielectric metalens constituted by silicon pillars, having a focal length of 30 μm and an 88% efficiency (Kim & Kim, 2020), and by Zhao group, which studied a lens formed by TiO2 nanorods of different radii, with a wavelength-depending focal length of 315–380 μm, designed for broadband and efficient focusing (Zhao et al., 2021). Moreover, in another recent study (Xie et al., 2021), a Photonic Crystal Fiber focusing metalens has been presented with the aim of increasing the numerical aperture of the fiber up to 0.65, reaching coupling efficiencies of nearly 90%.

These and other examples show the great potential represented by the integration of metasurface-based lenses on the fiber tip. In the following Table 3 the main features of the most significant articles on resonant lenses are synthesized, providing information about the optical fiber type, the typology and material of the metasurface-based lens, the chosen fabrication method, and the kind of application for which are employed.

5. Conclusions and outlook

In this review, we have reported on integrated lenses on optical fibers, discussing the actual performances and the recent development trends leading to increasingly efficient and miniaturized devices.

Indeed, the fiber facet is an inherently light coupled platform, hence the intrinsic advantage of placing optical elements directly on the tip of an optical fiber, allows for creating easy-to-use, plug and play and ultra-compact devices, that may supplement and/or replace bulky free-space optic systems, avoiding misalignment problems. Recently, it's established that small thickness, lightweight, and minimized absorption losses are desired features for a lens, that's why researchers have been paying increasing attention to the so-called "flat lenses" and to their integration on the fiber tip, towards increasingly ultra-compact packaging, with a small footprint.

In the previous sections, we have reported on numerous examples of refractive, diffractive, and resonant lenses; discussing the advantages and drawbacks of each category. Our study essentially reveals that the choice of the best suited fiber lens is strictly related to the specific application and to the requirements involved.

For a single application of interest (such as coupling, focusing, trapping, beam tailoring, etc.), the different features of each platform make it difficult to directly compare their performances. In any case, the above mentioned three types of fiber lenses can be compared in terms of their main characteristics from a qualitative point of view, as schematically shown in Fig. 5.

For the comparison each parameter was chosen in contrast to another one (Ease of Design vs Application Flexibility, Ease of Fabrication vs Compactness and Scientific & Technological Maturity vs Time to Market), highlighting the intrinsic trade-off among the different categories. Specifically, the figure shows that a simple design is opposed to the flexibility for different applications, while the fabrication complexity, together with cost and reproducibility, is strictly related to the dimension of the lens, with a greater complexity corresponding to a higher level of compactness.

Fig. 5 immediately shows that there is not a specific fiber lens that can outperform the others in every aspect.

For instance, refractive lenses are generally easier to design and less expensive to produce, with well-established fabrication techniques. On the other hand, their micrometric dimensions do not allow to achieve ultra-compact devices. Besides microlenses are less flexible, being suitable only for a limited number of applications, even if they represent a quite consolidated choice, as demonstrated by the large number of articles published on lenses for improving the coupling efficiency between the optical fiber and the light source.

Contrarily, diffractive and resonant lenses, which can be both referred to as flat lenses, have the advantage of reduced size and greater design flexibility compared to refractive lenses, but generally they are both more complex to design and fabricate. By the way, the greater complexity translates into a considerable versatility, allowing to introduce new optical functions (such as vortex generation and beam tailoring) unthinkable with standard microlenses.

For many years flat lenses have been realized employing diffractive optics technology, however in the last years, the new concept of "metasurfaces" has appeared. Indeed, with their arrival, the scientific community has questioned the advantages of these resonant lenses over diffractive ones. If diffractive lenses are generally more cost-effective and easier to design (Banerji et al., 2019; Engelberg and Levy, 2020b; Sensale-Rodriguez et al., 2019), on the other hand, metasurfaces offer more degrees of freedom, since with them is possible to model wavefront at will, creating optical elements with remarkable performances and allowing to structure light by shaping vector beams, with full control of polarization. Their subwavelength dimensions are extremely useful in integrated optics, but even though metalenses tend to be thinner than their diffractive counterparts, the overall thickness is still comparable (Engelberg and Levy, 2020a), showing only a partial advantage from a compactness point of view.

If flat lenses have undoubted advantages in terms of device compactness and performance, in comparison to their refractive counterpart, their maturity and their estimated time to the market is still far behind those of microlenses. Probably, the scientific and technological maturity is the most influential parameter, explaining the points of strength reached in the first two rows (easy of design and fabrication) by

Table 3
Comparison between the main parameters of selected resonant fiber lenses.

	Lens Characteristics				Declared Performances	
References	*Fiber type*	*Lens type*	*Material*	*Fabrication method*	*Application*	*Parameters of merit*
(Principe et al., 2017)	SMF	Plasmonic metasurface	gold	FIB milling	Beam steering	transmission efficiency: 12%
(Yang et al., 2019)	LMA-PCF	Plasmonic metalens	gold	EBL	Light focusing	numerical aperture: NA = 0.37 focal length: f = 28 μm
(Wang et al., 2019)	SMF	Plasmonic nanostructure	gold	EBL	Light collection at large angles (θ > 25°)	in-fiber coupling efficiency: η < 1%
(Zhou et al., 2021)	SMF	Dielectric metasurface	a-Si:H	EBL	Vortex generation & Beam collimation	Transmission efficiency: 57% focal length: f = 200 μm

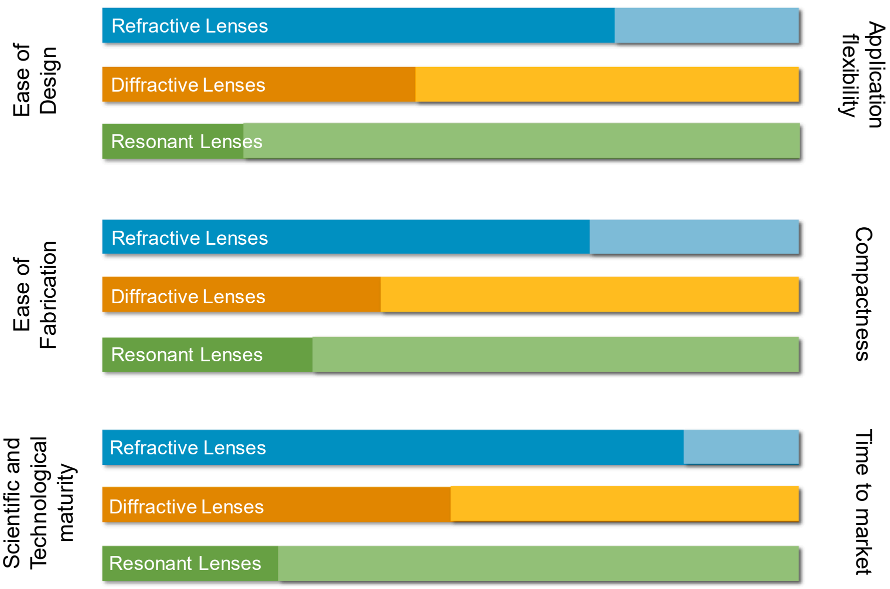


Fig. 5. Qualitative comparison among refractive, diffractive and resonant fiber lenses.

refractive lenses. As a matter of fact, microlenses are the 'oldest' type of fiber lenses, with the first articles on the subject dating to decades ago, representing a well-established technology even if with all the limitations in terms of flexibility and compactness, previously listed.

At the same time, the promising technology of flat lenses has the potential to set an important milestone along the technological roadmap pertaining to LOF may open the way to new remarkable applications, by overcoming the limitations of refractive lenses. In this respect, it is important to underline also the gap between diffractive and resonant lenses; in fact, as can be seen from Fig. 5, diffractive lenses are halfway between refractive and resonant lenses in quite all categories. Actually, metasurface-based lenses are less mature from both scientific and technological point of view, considering that the first articles on their integration on the fiber tip have been only recently published (Principe et al., 2017). Although resonant lenses show great potentialities (Chen & Capasso, 2021), they may not completely substitute diffractive lenses but rather expands on them (Luo, 2018). However, despite the great interest in the subject, there are still some limitations to their advancement, the primary reason being fabrication constraints, due to their complex geometry required. Even if a lot of effort has been put in terms of simplifying the metasurface design (Chen et al., 2020; Hadibrata et al., 2021; Lin et al., 2018), the (design and fabrication) complexity is still higher than the one required for the other fiber lenses. For this reason, further developments of resonant lenses could be achieved by improving the fabrication techniques, reducing the gap with diffractive lenses.

In this framework, the evolution of the manufacturing methods could be possible thanks to the LOF technology, due to its ability to realize subwavelength platforms on the fiber tip with high throughput. At the same time, the successful integration of such multifunctional components on the fiber end-face would also result in a major breakthrough in the LOF technology roadmap. In fact, new advances in fabrication techniques applied to optical fibers can pave the way to the integration of increasingly complex structures on the fiber facet, also achieving new unprecedented functionalities by directly integrating on the fiber tip light sources, detectors, and sensors, obtaining a new class of ultra-compact ready-to-use fiber-based platforms.

References

Asadollahbaik, A., Thiele, S., Weber, K., Kumar, A., Drozella, J., Sterl, F., Herkommer, A. M., Giessen, H., Fick, J., 2019. Highly efficient dual-fiber optical trapping with 3D printed diffractive fresnel lenses. ACS Photonics 7 (1), 88–97. https://doi.org/10.1021/acsphotonics.9b01024.

Banerji, S., Meem, M., Majumder, A., Vasquez, F.G., Sensale-Rodriguez, B., Menon, R., 2019. Imaging with flat optics: metalenses or diffractive lenses? Optica 6 (6), 805. https://doi.org/10.1364/optica.6.000805.

Barnard, C.W., Lit, J.W.Y., 1991. Single-mode fiber microlens with controllable spot size. Appl. Opt. 30 (15), 1958. https://doi.org/10.1364/AO.30.001958.

Calafiore, G., Koshelev, A., Allen, F.I., Dhuey, S., Sassolini, S., Wong, E., Lum, P., Munechika, K., Cabrini, S., 2016. Nanoimprint of a 3D structure on an optical fiber for light wavefront manipulation. Nanotechnology 27 (37). https://doi.org/10.1088/0957-4484/27/37/375301.

Capasso, F., 2018. The future and promise of flat optics: a personal perspective. Nanophotonics 7 (6), 953–957. https://doi.org/10.1515/NANOPH-2018-0004.

Capasso, F., Aieta, F., Khorasaninejad, M., Genevet, P., Devlin, R., 2017. Recent advances in planar optics: from plasmonic to dielectric metasurfaces. Optica 4 (1), 139. https://doi.org/10.1364/OPTICA.4.000139.

Chen, W.T., Capasso, F., 2021. Will flat optics appear in everyday life anytime soon? Appl. Phys. Lett. 118 (10), 100503. https://doi.org/10.1063/5.0039885.

Chen, W.T., Zhu, A.Y., Capasso, F., 2020. Flat optics with dispersion-engineered metasurfaces. Nat. Rev. Mater. 5 (8), 604–620. https://doi.org/10.1038/s41578-020-0203-3.

Cheng, J., Inampudi, S., Mosallaei, H., 2017. Optimization-based dielectric metasurfaces for angle-selective multifunctional beam deflection. Sci. Rep. 7 (1) https://doi.org/10.1038/s41598-017-12541-x.

Cohen, L.G., Schneider, M.V., 1974. Microlenses for coupling junction lasers to optical fibers. Appl. Opt. 13 (1), 89. https://doi.org/10.1364/AO.13.000089.

Cusano, A., Consales, M., Crescitelli, A., & Ricciardi, A. (Eds.). (2015). Lab-on-Fiber Technology. 56. https://doi.org/10.1007/978-3-319-06998-2.

Desiatov, B., Mazurski, N., Fainman, Y., Levy, U., 2015. Polarization selective beam shaping using nanoscale dielectric metasurfaces. Opt. Express 23 (17), 22611. https://doi.org/10.1364/OE.23.022611.

Ding, M., Chen, Y., Zhao, Y., Zhou, W., Koshelev, A., Munechika, K., Liu, Q., Feng, T., Wang, Y., Pan, Z., Shen, D., Griebner, U., Petrov, V., 2019. Propagation and orbital angular momentum of vortex beams generated from a spiral phase plate-fiber. Laser Phys. Lett. 16 (3), 035106. https://doi.org/10.1088/1612-202X/aafcb1.

Dou, J., Li, J., Herman, P.R., Aitchison, J.S., Fricke-Begemann, T., Ihlemann, J., Marowsky, G., 2008. Laser machining of micro-lenses on the end face of single-mode optical fibers. Appl. Phys. A Mater. Sci. Process. 91 (4), 591–594. https://doi.org/10.1007/s00339-008-4425-2.

Edwards, C.A., Presby, H.M., Dragone, C., 1993. Ideal microlenses for laser to fiber coupling. J. Lightwave Technol. 11 (2), 252–257. https://doi.org/10.1109/50.212535.

Engelberg, J., Levy, U., 2020a. The advantages of metalenses over diffractive lenses. Nat. Commun. 11 (1), 9–12. https://doi.org/10.1038/s41467-020-15972-9.

Engelberg, J., Levy, U., 2020b. The advantages of metalenses over diffractive lenses. Nat Commun 11 (1). https://doi.org/10.1038/s41467-020-15972-9.

Genevet, P., Capasso, F., Aieta, F., Khorasaninejad, M., Devlin, R., 2017. Recent advances in planar optics: from plasmonic to dielectric metasurfaces. Optica 4 (1), 139. https://doi.org/10.1364/optica.4.000139.

Ghafoori-Shiraz, H., 1988. Experimental investigation on coupling efficiency between semiconductor laser diodes and single-mode fibres by an etching technique. Opt. Quant. Electron. 20 (6), 493–500. https://doi.org/10.1007/BF00635750.

Ghafoori-shiraz, H., Asano, T., 1986. Microlens for coupling a semiconductor laser to a single-mode fiber. Opt. Lett. 11 (8), 537. https://doi.org/10.1364/OL.11.000537.

Giaquinto, M., 2021. (INVITED) Stimuli-responsive materials for smart Lab-on-Fiber optrodes. Results Opt. 2, 100051. https://doi.org/10.1016/j.rio.2020.100051.

Giaquinto, M., Aliberti, A., Micco, A., Gambino, F., Ruvo, M., Ricciardi, A., Cusano, A., 2019. Cavity-enhanced lab-on-fiber technology: toward advanced biosensors and nano-opto-mechanical active devices. ACS Photonics 6 (12), 3271–3280.

Gil, D., Menon, R., Smith, H.I., 2003. The case for diffractive optics in maskless lithography. J. Vacuum Sci. Technol. B: Microelectron. Nanometer Struct. Process. Measure. Phenomena 21 (6), 2810. https://doi.org/10.1116/1.1629288.

Gissibl, T., Schmid, M., Giessen, H., 2016a. Spatial beam intensity shaping using phase masks on single-mode optical fibers fabricated by femtosecond direct laser writing. Optica 3 (4), 448. https://doi.org/10.1364/OPTICA.3.000448.

Gissibl, T., Thiele, S., Herkommer, A., Giessen, H., 2016b. Two-photon direct laser writing of ultracompact multi-lens objectives. Nat. Photonics 10 (8), 554–560. https://doi.org/10.1038/nphoton.2016.121.

Hadibrata, W., Wei, H., Krishnaswamy, S., Aydin, K., 2021. Inverse design and 3D printing of a metalens on an optical fiber tip for direct laser lithography. Nano Lett. 21 (6), 2422–2428.

Hail, C.U., Poulikakos, D., Eghlidi, H., 2018. High-efficiency, extreme-numerical-aperture metasurfaces based on partial control of the phase of light. Adv. Opt. Mater. 6 (22), 1–8. https://doi.org/10.1002/adom.201800852.

Hillerich, B., Guttmann, J., 1989. Deterioration of taper lens performance due to taper asymmetry. J. Lightwave Technol. 7 (1), 99–104. https://doi.org/10.1109/50.17739.

Huang, S.Y., Yeh, S.M., Lu, Y.K., Lin, H.H., Cheng, W.H., 2004. A novel scheme of lensed fiber employing a quadrangular-pyramid-shaped fiber endface for coupling high-power lasers to single mode fibers. OSA Trends Opt. Photon. Ser. 95B (5), 533–535.

Hutley, M.C., Fleming, M.B., 1997. Blazed diffractive optics. Appl. Opt. 36 (20), 4635. https://doi.org/10.1364/AO.36.004635.

Janeiro, R., Flores, R., Dahal, P., Viegas, J., 2016. Fabrication of a phase photon sieve on an optical fiber tip by focused ion beam nanomachining for improved fiber to silicon photonics waveguide light coupling. Opt. Express 24 (11), 11611. https://doi.org/10.1364/oe.24.011611.

Kamali, S.M., Arbabi, E., Arbabi, A., Faraon, A., 2018. A review of dielectric optical metasurfaces for wavefront control. Nanophotonics 7 (6), 1041–1068. https://doi.org/10.1515/nanoph-2017-0129.

Kataoka, K., 2010. Estimation of coupling efficiency of optical fiber by far-field method. Opt. Rev. 17 (5), 476–480.

Kato, S., Chonan, S., & Aoki, T. (2013). *Micro-lensed single-mode optical fiber with high numerical aperture*. 3–4. http://arxiv.org/abs/1305.5937.

Kato, S., Chonan, S., Aoki, T., 2014. High-numerical-aperture microlensed tip on an air-clad optical fiber. Opt. Lett. 39 (4), 773. https://doi.org/10.1364/ol.39.000773.

Khorasaninejad, M., Capasso, F., 2017. Metalenses: versatile multifunctional photonic components. Science 358 (6367). https://doi.org/10.1126/science.aam8100.

Kim, H., Kim, J., An, H., Lee, Y., Lee, G., Na, J., Park, K., Lee, S., Lee, S.-Y., Lee, B., Jeong, Y., 2017. Metallic Fresnel zone plate implemented on an optical fiber facet for super-variable focusing of light. Opt. Express 25 (24), 30290. https://doi.org/10.1364/oe.25.030290.

Kim, J.K., Jung, Y., Lee, B.H., Oh, K., Chun, C., Kim, D., 2007. Optical phase-front inscription over optical fiber end for flexible control of beam propagation and beam pattern in free space. Opt. Fiber Technol. 13 (3), 240–245. https://doi.org/10.1016/J.YOFTE.2007.02.003.

Kim, M., Kim, S., 2020. High efficiency dielectric photonic crystal fiber metalens. Sci. Rep. 10 (1), 1–6. https://doi.org/10.1038/s41598-020-77821-5.

Koshelev, A., Calafiore, G., Piña-Hernandez, C., Allen, F.I., Dhuey, S., Sassolini, S., Wong, E., Lum, P., Munechika, K., Cabrini, S., 2016. High refractive index Fresnel lens on a fiber fabricated by nanoimprint lithography for immersion applications. Opt. Lett. 41 (15), 3423. https://doi.org/10.1364/OL.41.003423.

Kuchmizhak, A., Gurbatov, S., Nepomniaschii, A., Vitrik, O., Kulchin, Y., 2014. High-quality fiber microaxicons fabricated by a modified chemical etching method for laser focusing and generation of Bessel-like beams. Appl. Opt. 53 (5), 937. https://doi.org/10.1364/ao.53.000937.

Kuwahara, H., Sasaki, M., Tokoyo, N., 1980. Efficient coupling from semiconductor lasers into single-mode fibers with tapered hemispherical ends. Appl. Opt. 19 (15), 2578. https://doi.org/10.1364/AO.19.002578.

Lalanne, P., Chavel, P., 2017. Metalenses at visible wavelengths: past, present, perspectives. Laser Photonics Rev. 11 (3), 1600295. https://doi.org/10.1002/LPOR.201600295.

Lay, T.S., Yang, H.M., Lee, C.W., Cheng, W.H., 2004. Fiber grating laser: a performance study on coupling efficiency of fiber microlens and the Bragg reflectivity. Opt. Commun. 233 (1–3), 89–96. https://doi.org/10.1016/J.OPTCOM.2004.01.023.

Lee, C.-W., Yang, H.-M., Huang, S.-Y., Lay, T.-S., & Cheng, W.-H. (2004). High-Coupling Tapered Hyperbolic Fiber Microlens and Taper Asymmetry Effect. *Journal of Lightwave Technology, Vol. 22, Issue 5, Pp. 1395-*, *22*(5), 1395-. https://www.osapublishing.org/abstract.cfm?uri=jlt-22-5-1395.

Li, K., Du, J., Shen, W., Liu, J., He, Z., 2021. Improved optical coupling based on a concave cavity lens fabricated by optical fiber facet etching. Chinese Opt. Lett. 19 (5), 050602 https://doi.org/10.3788/col202119.050602.

Liang, H., Martins, A., Borges, B.-H.-V., Zhou, J., Martins, E.R., Li, J., Krauss, T.F., 2019. High performance metalenses: numerical aperture, aberrations, chromaticity, and trade-offs. Optica 6 (12), 1461. https://doi.org/10.1364/optica.6.001461.

Lin, C.-C., Yeh, S.-M., Cheng, W.-H., Tsai, Y.-C., Liu, Y.-D., Lu, Y.-K., 2007. Asymmetric elliptic-cone-shaped microlens for efficient coupling to high-power laser diodes. Opt. Express 15 (4), 1434. https://doi.org/10.1364/OE.15.001434.

Lin, C.-H., Lei, S.-C., Hsieh, W.-H., Tsai, Y.-C., Liu, C.-N., Cheng, W.-H., 2017. Micro-hyperboloid lensed fibers for efficient coupling from laser chips. Opt. Express 25 (20), 24480. https://doi.org/10.1364/oe.25.024480.

Lin, C.-H., Wu, C.-C., Kuo, S.-M., Tseng, Y.-D., 2011. Fabrication of asperical lensed optical fibers with an electro-static pulling of SU-8 photoresist. Opt. Express 19 (23), 22993. https://doi.org/10.1364/OE.19.022993.

Lin, F., Huang, H., Zou, H., Fu, J., Li, Q., Chen, S., Wu, X., 2014. Laser printed fiber microlens for fiber-diode coupling by direct laser writing. Appl. Opt. 53 (36), 8444. https://doi.org/10.1364/AO.53.008444.

Lin, S.I.E., 2005. A lensed fiber workstation based on the elastic polishing plate method. Precis. Eng. 29 (2), 146–150. https://doi.org/10.1016/J.PRECISIONENG.2004.05.008.

Lin, Z., Groever, B., Capasso, F., Rodriguez, A.W., Lončar, M., 2018. Topology-optimized multilayered metaoptics. *Phys. Rev. Appl.* 9 (4) https://doi.org/10.1103/PhysRevApplied.9.044030.

Liu, Y.D., Tsai, Y.C., Lu, Y.K., Wang, L.J., Hsieh, M.C., Yeh, S.M., Cheng, W.H., 2011. New scheme of double-variable-curvature microlens for efficient coupling high-power lasers to single-mode fibers. J. Lightwave Technol. 29 (6), 898–904. https://doi.org/10.1109/JLT.2010.2103394.

Liu, H., 2008. The approximate ABCD matrix for a parabolic lens of revolution and its application in calculating the coupling efficiency. Optik 119 (14), 666–670. https://doi.org/10.1016/J.IJLEO.2007.01.014.

Luo, X., 2018. Engineering optics 2.0: a revolution in optical materials, devices, and systems. ACS Photonics 5 (12), 4724–4738. https://doi.org/10.1021/ACSPHOTONICS.8B01036.

Majumder, A., Sensale-Rodriguez, B., Vasquez, F.G., Meem, M., Menon, R., Banerji, S., 2019. Imaging with flat optics: metalenses or diffractive lenses? Optica 6 (6), 805. https://doi.org/10.1364/OPTICA.6.000805.

Mandal, H., Maiti, S., Chiu, T.L., Gangopadhyay, S., 2018. Mismatch considerations in laser diode to single-mode circular core triangular index fiber excitation via upside down tapered hemispherical microlens on the fiber tip. Optik 168, 533–540. https://doi.org/10.1016/J.IJLEO.2018.04.115.

Melkonyan, H., Sloyan, K., Odeh, M., Almansouri, I., Chiesa, M., Dahlem, M.S., 2019. Embedded parabolic fiber lens for efficient fiber-to-waveguide coupling fabricated by focused ion beam. J. Phys. Photonics 1 (2), 025004. https://doi.org/10.1088/2515-7647/ab044f.

Melkonyan, H., Sloyan, K., Twayana, K., Moreira, P., Dahlem, M.S., 2017. Efficient fiber-to-waveguide edge coupling using an optical fiber axicon lens fabricated by focused ion beam. IEEE Photonics J. 9 (4), 1–9. https://doi.org/10.1109/JPHOT.2017.2710189.

Min Yang, H., Ting Chen, C., Ro, R., Chun Liang, T., 2010. Investigation of the efficient coupling between a highly elliptical Gaussian profile output from a laser diode and a single mode fiber using a hyperbolic-shaped microlens. Opt. Laser Technol. 42 (6), 918–926. https://doi.org/10.1016/J.OPTLASTEC.2010.01.009.

Moon, S.-W., Kim, Y., Yoon, G., Rho, J., 2020. Recent progress on ultrathin metalenses for flat optics. IScience 23 (12), 101877. https://doi.org/10.1016/j.isci.2020.101877.

Munechika, K., Koshelev, A., Calafiore, G., Pina-Hernandez, C., Allen, F.I., Dhuey, S., Sassolini, S., Wong, E., Lum, P., Cabrini, S. (2018). *Photonics on a fiber for wavefront manipulation. February*, 8. https://doi.org/10.1117/12.2287571.

Ni, X., Emani, N.K., Kildishev, A.V., Boltasseva, A., Shalaev, V.M., 2012. Broadband light bending with plasmonic nanoantennas. Science 335 (6067), 427.

Ni, X., Ishii, S., Kildishev, A.V., Shalaev, V.M., 2013a. Ultra-thin, planar, Babinet-inverted plasmonic metalenses. Light Sci. Appl. 2 (4), e72.

Ni, X., Kildishev, A.V., Shalaev, V.M., 2013b. Metasurface holograms for visible light. Nat. Commun. 4 (1) https://doi.org/10.1038/ncomms3807.

Panda, A., Sarkar, P., Palai, G., 2018. Studies on coupling of optical power in fiber to semiconductor waveguide at wavelength 1550 nm for photonics integrated circuits. Optik 157, 944–950. https://doi.org/10.1016/J.IJLEO.2017.11.119.

Paniagua-Domínguez, R., Yu, Y.F., Khaidarov, E., Choi, S., Leong, V., Bakker, R.M., Liang, X., Fu, Y.H., Valuckas, V., Krivitsky, L.A., Kuznetsov, A.I., 2018. A metalens with a near-unity numerical aperture. Nano Lett. 18 (3), 2124–2132.

Pisco, M., Cusano, A. (2020). Lab-On-Fiber Technology: A Roadmap toward Multifunctional Plug and Play Platforms. *Sensors 2020, 20*(17), 4705. https://doi.org/10.3390/S20174705.

Plidschun, M., Ren, H., Kim, J., Förster, R., Maier, S.A., Schmidt, M.A., 2021. Ultrahigh numerical aperture meta-fibre for flexible optical trapping. Light Sci. Appl. 10 (1) https://doi.org/10.1038/s41377-021-00491-z.

Presby, H.M., Benner, A.F., Edwards, C.A., 1990. Laser micromachining of efficient fiber microlenses. Appl. Opt. 29 (18), 2692. https://doi.org/10.1364/ao.29.002692.

Presby, H.M., Edwards, C.A., 1992. Efficient coupling of polarization-maintaining fiber to laser diodes. IEEE Photonics Technol. Lett. 4 (8), 897–899. https://doi.org/10.1109/68.149901.

Presby, H.M., Giles, C.R., 1993. Asymmetric Fiber microlenses for efficient coupling to elliptical laser beams. IEEE Photonics Technol. Lett. 5 (2), 184–186. https://doi.org/10.1109/68.195998.

Principe, M., Consales, M., Micco, A., Crescitelli, A., Castaldi, G., Esposito, E., La Ferrara, V., Cutolo, A., Galdi, V., Cusano, A., 2017. Optical fiber meta-tips. Light Sci. Appl. 6 (3), 1–10. https://doi.org/10.1038/lsa.2016.226.

Ricciardi, A., Crescitelli, A., Vaiano, P., Quero, G., Consales, M., Pisco, M., Esposito, E., Cusano, A., 2015. Lab-on-fiber technology: a new vision for chemical and biological sensing. Analyst 140 (24), 8068–8079. https://doi.org/10.1039/C5AN01241D.

Rodrigues Ribeiro, R.S., Dahal, P., Guerreiro, A., Jorge, P.A.S., Viegas, J., 2017. Fabrication of Fresnel plates on optical fibres by FIB milling for optical trapping, manipulation and detection of single cells. Sci. Rep. 7 (1), 1–14. https://doi.org/10.1038/s41598-017-04490-2.

Rodrigues, S.M., Paiva, J.S., Ribeiro, R.S.R., Soppera, O., Cunha, J.P.S., Jorge, P.A.S., 2018. Fabrication of multimode-single mode polymer fiber tweezers for single cell trapping and identification with improved performance. Sensors (Switzerland) 18 (9), 1–26. https://doi.org/10.3390/s18092746.

Sakai, J.I., Kimura, T., 1980. Design of a miniature lens for semiconductor laser to singlemode fiber coupling. IEEE J. Quantum Electron. 16 (10), 1059–1067. https://doi.org/10.1109/JQE.1980.1070359.

Scaravilli, M., Micco, A., Castaldi, G., Coppola, G., Gioffrè, M., Iodice, M., La Ferrara, V., Galdi, V., Cusano, A., 2018. Excitation of bloch surface waves on an optical fiber tip. Adv. Opt. Mater. 6 (19), 1800477. https://doi.org/10.1002/adom.201800477.

Sensale-Rodriguez, B., Meem, M., Menon, R., Banerji, S., Menon, R. (2019). Metalenses or diffractive lenses for imaging? *Imaging and Applied Optics 2019 (COSI, IS, MATH, PcAOP) (2019), Paper ITu4B.3*, *Part F157-ISA 2019*, ITu4B.3. https://doi.org/10.1364/ISA.2019.ITU4B.3.

Shiraishi, K., Yoda, H., Kogami, Y., Tsai, C.S., 2011. High focusing power lensed fibers employing graded-index fiber with eigen-beam diameter. IEEE Photonics Technol. Lett. 23 (19), 1376–1378. https://doi.org/10.1109/LPT.2011.2161278.

Siemion, A., 2019. Terahertz diffractive optics—Smart control over radiation. J. Infrared Milli. Terahz Waves 40 (5), 477–499.

Liu, Y., Xu, H., Stief, F., Zhitenev, N., Yu, M., 2011. Far-field superfocusing with an optical fiber based surface plasmonic lens made of nanoscale concentric annular slits. Opt. Express 19 (21), 20233. https://doi.org/10.1364/OE.19.020233.

Liu, Y., Stief, F., Yu, M., 2013. Subwavelength optical trapping with a fiber-based surface plasmonic lens. Opt. Lett. 38 (5), 721. https://doi.org/10.1364/OL.38.000721.

Sun, S., Yang, K.-Y., Wang, C.-M., Juan, T.-K., Chen, W.T., Liao, C.Y., He, Q., Xiao, S., Kung, W.-T., Guo, G.-Y., Zhou, L., Tsai, D.P., 2012. High-efficiency broadband anomalous reflection by gradient meta-surfaces. Nano Lett. 12 (12), 6223–6229. https://doi.org/10.1021/NL3032668.

Sundaram, V.M., Wen, S.-B., 2012. Fabrication of micro-optical devices at the end of a multimode optical fiber with negative tone lift-off EBL. J. Micromech. Microeng. 22 (12), 125016. https://doi.org/10.1088/0960-1317/22/12/125016.

Tsai, C.S., Yoda, H., Shiraishi, K., Muro, K., Kagaya, M., Kogami, Y., 2008. Single-mode fiber with a plano-convex silicon microlens for an integrated butt-coupling scheme. Appl. Opt. 47 (34), 6345. https://doi.org/10.1364/AO.47.006345.

Tseng, Y.T., Huang, J.B., Hung, T.Y., Jhou, H.Y., Kuo, C.L., 2014. Lensed plastic optical fiber employing hyperbolic end filled with high-index resin using electrostatic force. Precis. Eng. 38 (1), 183–189. https://doi.org/10.1016/J.PRECISIONENG.2013.09.003.

Vaiano, P., Carotenuto, B., Pisco, M., Ricciardi, A., Quero, G., Consales, M., Crescitelli, A., Esposito, E., Cusano, A., 2016. Lab on fiber technology for biological sensing applications. Laser Photonics Rev. 10 (6), 922–961. https://doi.org/10.1002/LPOR.201600111.

Eisenstein, G., Vitello, D., 1982. Chemically etched conical microlenses for coupling single-mode lasers into single-mode fibers. Appl. Opt. 21 (19), 3470. https://doi.org/10.1364/AO.21.003470.

Wang, N., Zeisberger, M., Hübner, U., Schmidt, M.A., 2018. Nanotrimer enhanced optical fiber tips implemented by electron beam lithography. Opt. Mater. Express 8 (8), 2246. https://doi.org/10.1364/ome.8.002246.

Wang, N., Zeisberger, M., Hübner, U., Schmidt, M.A., 2019. Boosting light collection efficiency of optical fibers using metallic nanostructures. ACS Photonics 6 (3), 691–698. https://doi.org/10.1021/acsphotonics.8b01560.

Wen, C., Hu, J., Xu, W., Shi, J., Zhen, S., Cao, Z., Yu, B., Xu, F., 2020. A simple and low-cost fabrication method of microlens on optical fiber end facet. Optik 214 (April), 164829. https://doi.org/10.1016/j.ijleo.2020.164829.

Wu, C.-C., Tseng, Y.-D., Kuo, S.-M., Lin, C.-H., 2011. Fabrication of asperical lensed optical fibers with an electro-static pulling of SU-8 photoresist. Opt. Express 19 (23), 22993. https://doi.org/10.1364/oe.19.022993.

Xie, Y., Zhang, J., Wang, S., Liu, D., 2021. High-efficiency broadband photonic crystal fiber metalens with a large numerical aperture. Opt. Commun. 481, 126524. https://doi.org/10.1016/j.optcom.2020.126524.

Xiong, Y., Xu, F., 2020. Multifunctional integration on optical fiber tips: challenges and opportunities. Adv. Photonics 2 (06), 1–24. https://doi.org/10.1117/1.ap.2.6.064001.

Yamada, J.I., Murakami, Y., Sakai, J.I., Kimura, T., 1980. Characteristics of a hemispherical microlens for coupling between a semiconductor laser and singlemode fiber. IEEE J. Quantum Electron. 16 (10), 1067–1072. https://doi.org/10.1109/JQE.1980.1070355.

Yang, J., Ghimire, I., Wu, P.C., Gurung, S., Arndt, C., Tsai, D.P., Lee, H.W.H., 2019. Photonic crystal fiber metalens. *Nanophotonics* 8 (3), 443–449. https://doi.org/10.1515/nanoph-2018-0204.

Yeh, S.M., Huang, S.Y., Cheng, W.H., 2005. A new scheme of conical-wedge-shaped fiber endface for coupling between high-power laser diodes and single-mode fibers. J. Lightwave Technol. 23 (4), 1781–1786. https://doi.org/10.1109/JLT.2005.844511.

Yermakov, O., Schneidewind, H., Hübner, U., Wieduwilt, T., Zeisberger, M., Bogdanov, A., Kivshar, Y., Schmidt, M.A., 2020. Nanostructure-empowered efficient coupling of light into optical fibers at extraordinarily large angles. ACS Photonics 7 (10), 2834–2841. https://doi.org/10.1021/acsphotonics.0c01078.

Yoda, H., Shiraishi, K., 2001. A new scheme of a lensed fiber employing a wedge-shaped graded-index fiber tip for the coupling between high-power laser diodes and single-mode fibers. J. Lightwave Technol. 19 (12), 1910–1917. https://doi.org/10.1109/50.971684.

Yu, J., Bai, Z., Zhu, G., Fu, C., Li, Y., Liu, S., Liao, C., Wang, Y., 2020. 3D nanoprinted kinoform spiral zone plates on fiber facets for high-efficiency focused vortex beam generation. Opt. Express 28 (25), 38127. https://doi.org/10.1364/OE.411209.

Yu, N., Capasso, F., 2014. Flat optics with designer metasurfaces. Nat. Mater. 13 (2), 139–150. https://doi.org/10.1038/nmat3839.

Yu, N., Capasso, F., 2015. Optical metasurfaces and prospect of their applications including fiber optics. J. Lightwave Technol. 33 (12), 2344–2358. https://doi.org/10.1109/JLT.2015.2404860.

Yu, N., Genevet, P., Kats, M.A., Aieta, F., Tetienne, J.P., Capasso, F., Gaburro, Z., 2011. Light propagation with phase discontinuities: generalized laws of reflection and refraction. Science 334 (6054), 333–337. https://doi.org/10.1126/SCIENCE.1210713.

Yuan, G.H., Rogers, E.TF., Zheludev, N.I., 2017. Achromatic super-oscillatory lenses with sub-wavelength focusing. Light Sci. Appl. 6 (9) e17036 e17036.

Zaboub, M., Guessoum, A., Demagh, N.E., Guermat, A., 2016. Fabrication of polymer microlenses on single mode optical fibers for light coupling. Opt. Commun. 366, 122–126. https://doi.org/10.1016/J.OPTCOM.2015.12.010.

Zhang, H., Guo, Z., 2020. Single mode-fiber scale based square solid immersion metalens for single quantum emitters. Opt. Mater. 105 (March), 2–8. https://doi.org/10.1016/j.optmat.2020.109850.

Zhao, Q., Qu, J., Peng, G.D., Yu, C., 2021. Endless single-mode photonics crystal fiber metalens for broadband and efficient focusing in near-infrared range. Micromachines 12 (2), 1–11. https://doi.org/10.3390/mi12020219.

Zhou, C., Lee, W.B., Gao, S., Li, H., Park, C.S., Choi, D.Y., Lee, S.S., 2021. All-dielectric fiber meta-tip enabling vortex generation and beam collimation for optical interconnect. Laser Photonics Rev. 15 (5), 1–9. https://doi.org/10.1002/lpor.202000581.

Zhou, H., Xu, H., Duan, J.-a., 2020. Review of the technology of a single mode fiber coupling to a laser diode. Opt. Fiber Technol. 55, 102097. https://doi.org/10.1016/j.yofte.2019.102097.

Surface tension coefficient of liquid sensor based on FBG

Soumya Datta Mohanty, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, soumya_datta.mohanty@gmail.com*

Satyajit Nayak, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, satyajit_nayak@gmail.com*

S. Sivasakthiselvan, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, s.sivasakthiselavan@live.com*

Ankita Panda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ankitapanda21@yahoo.co.in*

ARTICLE INFO

Keywords:
Surface tension coefficient of liquid
FBG
Pull-off method

ABSTRACT

A highly accuracy optical fiber sensor based on a fiber Bragg grating (FBG) is proposed and demonstrated for surface tension coefficient of liquid (STCL) measurement. In the experiment, pull-off method is applied to measure the STCL, and FBG is used as a high precision tension sensor. The performance is minutely studied in the surface tension coefficient of water measurement. The experimental results showed that the Bragg wavelength can be used to assess the STCL, and the average absolute percentage deviation of measurements from the correlation of the surface tension coefficient of water is just 0.4%. In order to show that the sensor has the versatility of other STCL measurement, surface tension coefficients of NaCl solution, and sucrose solution with different concentrations are experimentally measured. The results showed that the sensor presented higher accuracy. The STCL measurement system based on FBG has the advantages of simple structure, cost-effectiveness, and high accuracy. It can offer an alternative scheme for high accuracy STCL measurement in the fields such as surface physics, surface chemistry, and medicine.

1. Introduction

The surface tension coefficient of liquid (STCL) is a crucial property parameter of liquids. STCL affects the surface tension of liquid directly. In the recently years, studies on the surface tension of liquids have acquired more attention (Nayar et al., 2014; Pauletti et al., Matter (2021).; Navascues, 1979); but the STCL was less. Measurement and study the STCL, it can distinguish the liquid solution's concentration, viscosity, impurity content (Khaleduzzaman et al., 2013; Peng et al., 2021; Ghatee et al., 2010). Therefore, measurement the STCL is extremely important for surface physics, surface chemistry, medicine, agriculture, smelting, and chemical engineering (Hanks et al., 2020; Martínez-Calvo and Sevilla, 2020; Masuhara and Yuyama, 2021). The traditional methods for measuring the STCL include the pull-off method, capillary method, maximum bubble pressure method, hanging drop method, surface wave laser interferometry, Yang-Laplace relation method, etc. The pull-off method based on the Jolly balance is a classic method (Tang et al., 2016; Saepuzaman et al., 2018), which has the advantages of simple operation, low cost, and fast measurement speed. However, its experimental measurement device lead to the accuracy is not satisfactory (Tang et al., 2016). The capillary method is likewise a common method to assess the surface tension coefficient of liquid with higher accuracy (An, 2010). However, the capillary method is restricted to the kinds of liquid, such as the liquids with very high resistivity. The maximum bubble pressure method was also reported measuring the STCL (Zhao et al., 2007); but it depended on a complicated device and inconvenient operation.

Measurement the STCL, the optical fiber sensor is a good candidate compared to traditional methods due to the advantages of superior sensitivity, strong anti-interference ability, light-weight, fast response, and cost-efficient. Over the preceding research, the optical fiber sensor application has been reported widely (Márquez-Cruz and Hernández-Cordero, 2014; Tian et al., 2019; Luo et al., 2021), and its potential in determining STCL has attracted some attention. Such as the Fabry-Perot interferometer optical fiber sensor based on the hollow fiber has been applied to measure STCL at different temperatures (Zhu et al., 2015). Even though the sensors based on optical fiber tension showed a satisfactory result, the measurement accuracy still needs further improvement (Wang et al., 2020). Therefore, it encourages us to put forward a new method to further improve the measure accuracy of STCL.

In this article, a highly accurate STCL sensor based on FBG is proposed and experimentally demonstrated. Measuring characteristics of the sensor have been investigated using the water, NaCl solution, and sucrose solution. The experimental procedure and an uncertainty analysis are presented. The results showed that the average absolute percentage deviation of measurements from the correlation of the measured surface tension coefficient of water compared with the standard reference value is just 0.4%. Besides, for different concentration of NaCl solution and sucrose solution, this sensor has good linear response and acquired accurate results. The sensor gets the versatility of STCL measurement, and has higher measurement results than conventional methods.

2. Experimental techniques

The pull-off method is applicable to measure the STCL, and the FBG is used as a high precision tension sensor. The scheme of experiment is illustrated in Fig. 1. It comprises of a broadband light source (BBS, Golight, 1250 nm-1650 nm), optical spectrum analyzers (OSA, YOKOGAWA, AQ6370D), FBG, and lifting platform. The sensor used in the experiment is a commercial FBG with a central wavelength of 1561.42 nm, side mold rejection ratio is 17 dB, and reflectivity is 92.92%. A circulator is used to connect the FBG and sent a reflected signal from the FBG to the OSA. The wire ring connected with the FBG by a heat shrink tubing, and the bottom of the wire ring is initially completely immersed in the liquid in a glass dish. The lifting platform can be lifted up and down freely to make the wire ring pull-off the liquid. Optical connecting rods are used to support the fiber which is held the FBG.

The light from the BBS is transferred into the circulator and then injected into the FBG. The light with wavelength meeting Bragg condition is reflected and then recorded by the OSA with the resolution of 0.02 nm. The central wavelength of the FBG can be expressed as the following (Liu et al., 2018):

$$\lambda = 2n_{eff}\Lambda \tag{1}$$

where n_{eff} is the effective refractive index of the optical fiber core of FBG, and Λ is the grating period of FBG. It is known that the temperature and tension of the FBG environmental change, the grating period and effective refractive index of fiber core of the optical grating changes. As a result, the central wavelength of the reflected light changes correspondingly. The central wavelength of FBG change caused by the axial tension and temperature can be expressed as (Othonos, 1997):

$$\Delta\lambda = 2(\Lambda\frac{\partial n_{eff}}{\partial F} + n_{eff}\frac{\partial \Lambda}{\partial F})\Delta F + 2(\Lambda\frac{\partial n_{eff}}{\partial T} + n_{eff}\frac{\partial \Lambda}{\partial T})\Delta T \tag{2}$$

where ΔF and ΔT are the axial tension and temperature variation, respectively. In the experiment, only the tension acted on the FBG when the wire ring is pulled-off from the liquid and hanged in the air (ambient temperature keeps constant, 28 °C). The central wavelength of the FBG caused by the gravity force can be expressed as:

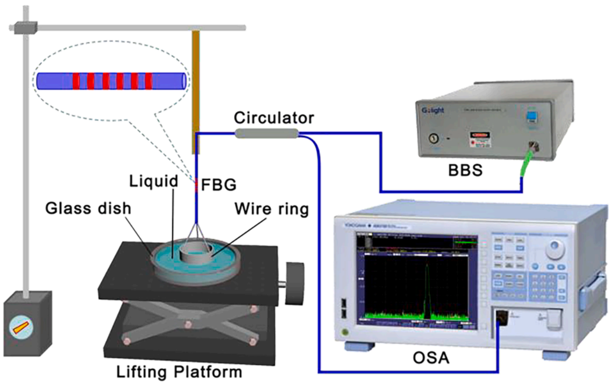


Fig. 1. Schematic diagram of experimental setup.

$$\lambda_1 = k{\cdot}mg \tag{3}$$

$$\lambda_2 = k{\cdot}(mg + f) \tag{4}$$

Where f is the surface tension of liquid, therefore, the surface tension of liquid can be expressed as:

$$f = \frac{\lambda_2 - \lambda_1}{k} \tag{5}$$

The moment of the wire ring is pulled-off from the surface of liquid, and the surface tension of liquid can also be expressed as:

$$f = \pi{\cdot}(D_1 + D_2)\alpha \tag{6}$$

where D_1 and D_2 are the inner diameter and outer diameter of the wire ring, respectively. The α is STCL. The relationship between the central wavelength of the FBG with the STCL can be expressed as:

$$\alpha = \frac{(\lambda_2 - \lambda_1)}{k\pi(D_1 + D_2)} \tag{7}$$

Therefore, monitor the shifted of spectral wavelength of the sensor, and then the surface tension coefficient of liquid can be calculated.

3. Results and discussion

A tensile test is firstly conducted to study the axial tension sensitivity of the FBG. The laboratory temperature is keep at 28 °C controlled by an air conditioning during the whole experimental process. Therefore, the effect of temperature on the shift of Bragg wavelength can be negligible. The effective refractive index of the fiber core and the grating period changed with the tension generated by added the weight, and the central wavelength of the FBG changed at the same time. The gravity force ranging from 0.1 N to 0.6 N in intervals of 0.1 N is applied in the experimental, and the measured spectra is shown in Fig. 2(a). The reflected spectrum of the FBG exhibits regular shifted with the gravity force increased. The central wavelength of the FBG under the tension was extracted for linear analysis, and the result as depicted in Fig. 2(b). It can be seen that the central wavelength of the FBG show a good linear response with the increase of tension. The equation of linear regression can be express as $y = 1561.02 + 1.30x$, and the fit has a coefficient of determination (R^2) value is up to 0.999. Based on the result above, tension sensitivity of the FBG is $k = 1.30$ nm/N.

Secondly, the surface tension coefficient of water is measured in the experiment. The inner diameter and outer diameter of the wire ring used in the experiment should be measured accurately. To obtain accurate measurement results, a vernier caliper (Deli, Type: DL91150) with a division value is 0.01 mm was used to measure the inner diameter and outer diameter of the wire ring. The results are listed in Table 1.

STCL is then measured, and the feasibility and repeatability of the method is also experimentally validated. In the experiment, the bottom of the wire ring is initially completely immersed in the water in the glass dish. The glass dish is transferred up and down by means of the lifting platform to make the wire ring pull-off the liquid surface. Before the wire ring is pulled-off from the water surface, the surface tension acted on the FBG increased with the lifting platform goes down. Under the tension, the central wavelength of the FBG increased gradually, and it reached the maximum value at the moment when the wire ring pulled-off from the liquid surface. The result is displayed in Fig. 3.

It can be seen from Fig. 3 that the surface tension of liquid increased with the lifting platform goes downs, and the central wavelength of the FBG increase gradually. The central wavelength of the FBG λ_2 reaches maximum value of 1561.091 nm at the moment that the wire ring is pulled-off from the liquid surface. After pulled-off from the liquid surface, the wire ring was hanged in the air and experienced several oscillations, so the spectrum showed jitter. The wavelength is only caused by the gravity of the wire ring after it ended vibration. Now, the λ_1

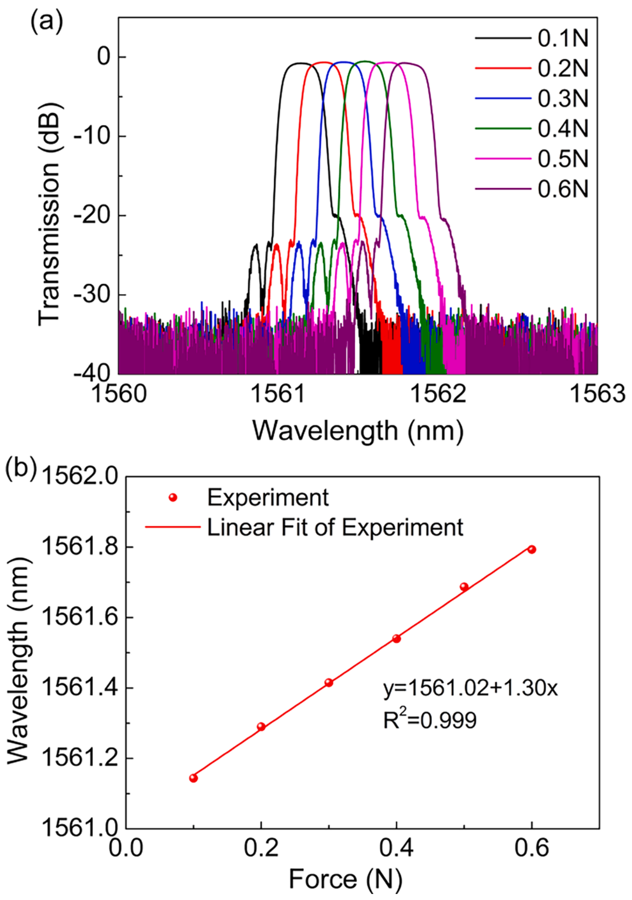


Fig. 2. (a) Spectral responses of the sensor obtained after applying a tensile force, (b) linear relationship between tensile force and the wavelength.

Table 1
Inner diameter and outer diameter of the wire ring.

Sequence (mm)	1	2	3	Average
Inner (D)	32.86	32.91	32.89	32.88
Outer (D)	34.97	35.03	34.99	34.99

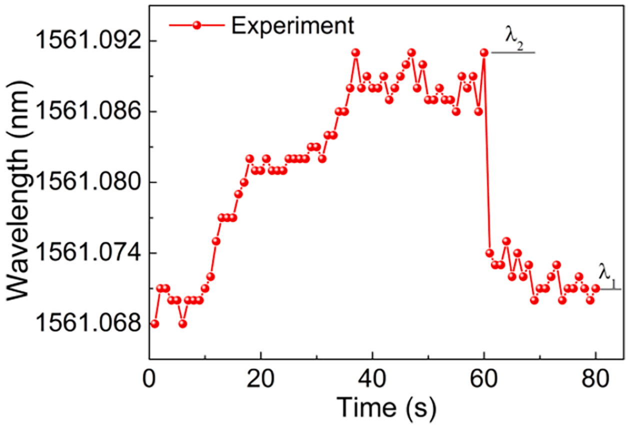


Fig. 3. The dynamic curve of the central wavelength of FBG with time.

reaches 1561.071 nm. The FBG wavelength shift is 0.02 nm, and substituting the relevant values into the formula (7), and the surface tension coefficient of water is 72.19 × 10^{-3}N/m. The experiment is carried out in the laboratory under the room temperature of 28 °C, and the reference value of the surface tension coefficient of water estimated by the literature is 71.90 × 10^{-3}N/m (Vargaftik et al., 1983), the absolute percentage deviation of measurements from the correlation is just 0.4%.

The experiment was repeated for many times to improve the accuracy of the result, it can be observed in Fig. 4(a). A total of 10 time measurements were performed, and the FBG wavelength shift floated at 0.02 nm. After statistical analysis, the average FBG wavelength shift is 0.02 nm. The relative error of the experimental results of each measurement of surface tension coefficient of water was calculated, and the results can be observed in Fig. 4(b). The percentage deviation of the sixth time measurement experiments is slightly larger, but most of the experimental results have smaller errors. After statistical analysis, the average of percentage deviation is only 0.4%.

To verify that the proposed sensor has the versatility of STCL measurement of other liquid, surface tension coefficients of NaCl solution and sucrose solution with different concentrations are experimentally measured under the same conditions separately. The test solutions were prepared by diluting original solutions of known concentration with the initial and final weight being measured using a CAAKEr FA2204 electronic mass balance that had a resolution of 0.0001 g and an accuracy of 0.0002 g. The concentration of NaCl solution range from 5% to 25% at intervals of 5%, and the measurement results are shown in Fig. 5. It can be seen that the wavelength shifts grow with the increase of NaCl solution concentration, and showed a good linear response, as showed in Fig. 5(a). The surface tension coefficient of NaCl solution at the corresponding concentration can be calculated with the wavelength shifted, and the result as showed in Fig. 5(b). It can be seen that the surface

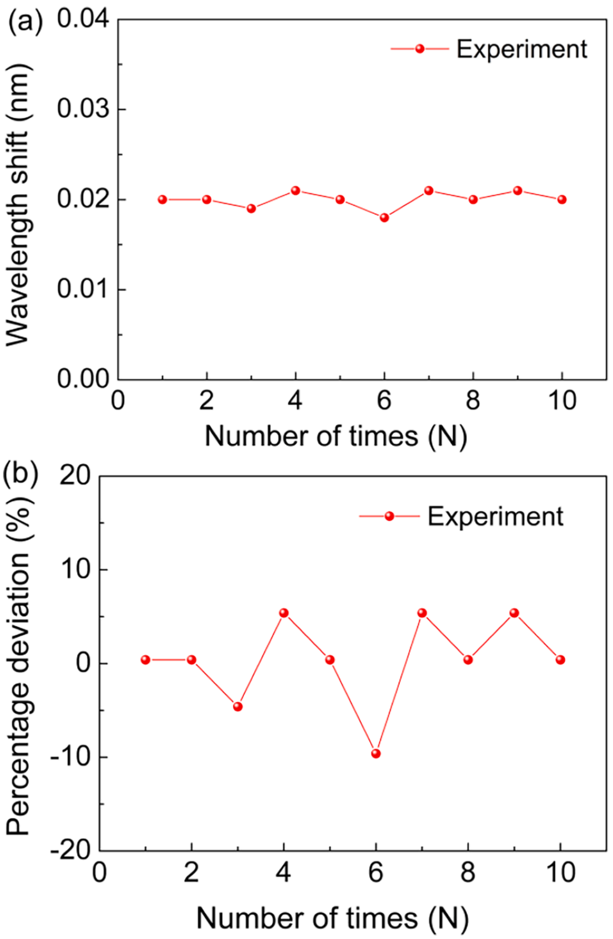


Fig. 4. (a) Repeatability of the pull-off experiment of ten times, (b) percentage deviation of the pull-off experiment of ten times.

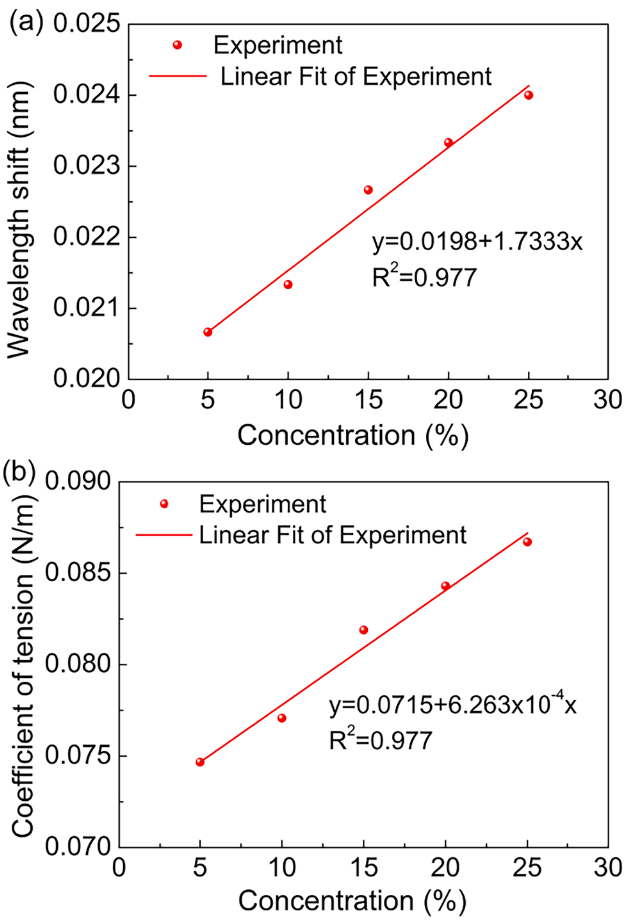


Fig. 5. (a) Linear relationship between the wavelength shifts with concentration of NaCl solution, (b) linear relationship between the coefficient of tension with concentration of NaCl solution.

tension coefficient increases with the NaCl solution concentration increase, and exhibits a linearly response. The equation of linear regression is $y = 0.0715 + 6.263 \times 10^{-4}x$, and the fit has a coefficient of determination (R^2) value is up to 0.977.

For NaCl solution, the linear regression is $y = 0.0715 + 6.263 \times 10^{-4}x$. If $x = 0$, it means that the concentration of the solution is zero, the solution can be regarded as pure water, and the surface tension coefficient is $71.5 \times 10^{-3}\ N/m$. Compared with the reference value of the STCL, the percentage deviation of measurements from the correlation is only 0.5%.

Similarly to the NaCl solution, surface tension coefficient of sucrose solution in different concentration was also used for experimental measurement, and the results are presented in Fig. 6. The results presented that the wavelength shift increased with the sucrose solution concentration increased ranging from 5% to 25% in intervals of 5%, and showed a good linear response, as showed in Fig. 6(a). The surface tension coefficient presented linear relationship with the corresponding concentration, as showed in Fig. 6(b). The equation of linear regression is $y = 0.069 + 4.817 \times 10^{-4}x$, and the fit has a coefficient of determination (R^2) value is up to 0.999. Similar to NaCl solution, for a sucrose solution, if $x = 0$, the percentage deviation of measurements from the correlation is only 2.7%. Based on the three examples above, the accuracy of the sensor is highly relative.

It is well known that the FBG is sensitive to temperature, so the effect of the temperature on the performance is needed to be considered (Hirayama and Sano, 2000). The temperature has a significant effect during the experiment, so it has to be kept constantly or compensated in the course of the experiment to avoid influence on the STCL test.

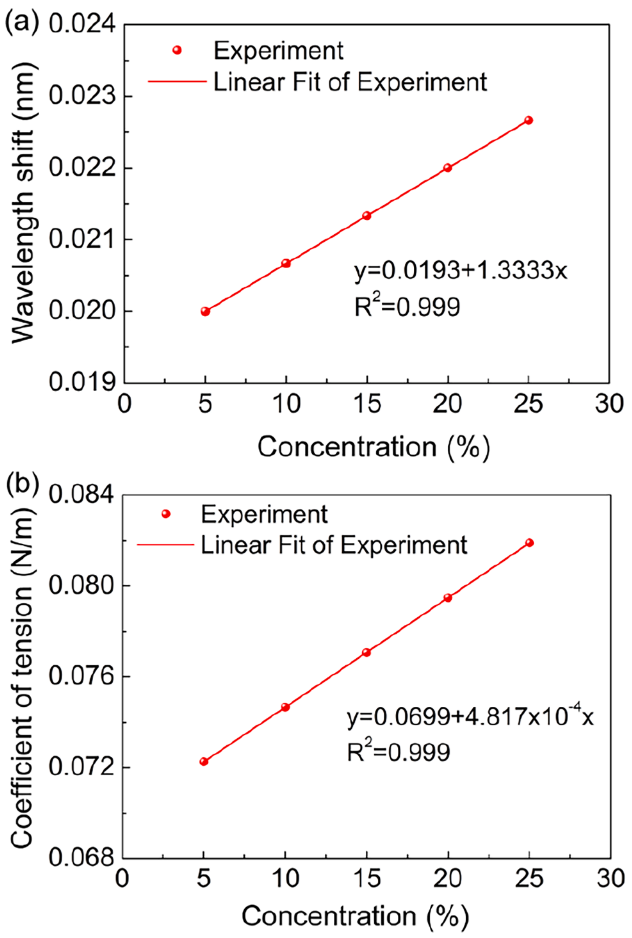


Fig. 6. (a) Linear relationship between the wavelength shifts with the concentration of the sucrose solution, (b) linear relationship between coefficient of tension with concentration of sucrose solution.

4. Conclusions

In this article, a STCL sensor based on a FBG has been experimental and demonstrated. The characteristics of STCL measurement of the sensor have been investigated, and a reference correlation for the STCL as a function of concentration was developed. The sensor has the versatility of STCL measurement, and the percentage deviation of measurements from the correlation of the surface tension coefficient of water is just 0.4%. The STCL measurement system based on FBG has the advantages of simple structure, repeatability, cost-effectiveness, and high accuracy. The proposed scheme with its excellent performances and has potential applications in the fields of surface physics, surface chemistry, and medicine.

Funding

This work was financially supported by the Natural Science Foundation of China (NSFC) under Grant No. 62165002 and Guangxi National key research and development program under Grant Nos. AB18221033, and Guangxi 1000 Young and Middle-aged College and University Backbone Teachers Cultivation Program, and Science Research Foundation Project of Guangxi Normal University of Science and Technology under Grant No. GXKS2020QN025, and Research Basic Ability Improvement Project of Young and Middle-aged Teachers in Guangxi Universities in 2021 under Grant No. 2021KY0856.

References

An, Y.k., 2010. Determination of the surface tension coefficient of liquid with the capillary tube probe method. College Phys. 29 (10), 37–40.

Ghatee, M.H., Zare, M., Zolghadr, A.R., et al., 2010. Temperature dependence of viscosity and relation with the surface tension of ionic liquids. Fluid Phase Equilib. 291 (2), 188–194.

Hanks, D.F., Lu, Z., Sircar, J., et al., 2020. High heat flux evaporation of low surface tension liquids from nanoporous membranes. ACS Appl. Mater. Interfaces 12 (6), 7232–7238.

Hirayama, N., Sano, Y., 2000. Fiber Bragg grating temperature sensor for practical use. ISA Trans. 39 (2), 169–173.

Khaleduzzaman, S.S., Mahbubul, I.M., Shahrul, I.M., et al., 2013. Effect of particle concentration, temperature and surfactant on surface tension of nanofluids. Int. Commun. Heat Mass Transfer 49, 110–114.

Liu, H.L., Zhu, Z.W., Zheng, Y., et al., 2018. Experimental study on an FBG strain sensor. Opt. Fiber Technol. 40, 144–151.

Luo, D.B., Qian, L.L., Wu, S.B., et al., 2021. Measurement of liquid surface tension and shape of a droplet on vertical plate by far field scattering technique. Opt. Commun. 482, 126578.

Márquez-Cruz, V.A., Hernández-Cordero, J.A., 2014. Fiber optic Fabry-Perot sensor for surface tension analysis. Opt. Express 22 (3), 3028–3038.

Martínez-Calvo, A., Sevilla, A., 2020. Universal thinning of liquid filaments under dominant surface forces. Phys. Rev. Lett. 125 (11), 114502.

Masuhara, H., Yuyama, K.I., 2021. Optical Force-Induced Chemistry at Solution Surfaces. Annu. Rev. Phys. Chem. 72, 565–589.

Navascues, G., 1979. Liquid surfaces: theory of surface tension. Rep. Prog. Phys. 42 (7), 1131.

Nayar, K., Panchanathan, D., McKinley, G., et al., 2014. Surface tension of seawater. J. Phys. Chem. Ref. Data 43 (4), 043103.

Othonos, A., 1997. Fiber bragg gratings. Rev. Sci. Instrum. 68 (12), 4309–4341.

M. Pauletti, V. Rybkin, M. Iannuzzi, 2021. Surface tension of liquids and binary mixtures from molecular dynamics simulations, J. Phys.: Condens. Matter.

Peng, M., Duignan, T.T., Nguyen, C.V., et al., 2021. From Surface Tension to Molecular Distribution: Modeling Surfactant Adsorption at the Air-Water Interface. Langmuir 37 (7), 2237–2255.

Saepuzaman, D., Nugraha, M.G., Dewi, R., et al., 2018. Using Jolly Balance Spring Method to Determine Pure Water Surface Tension Coefficient. Pertanika J. Sci. Technol. 26 (3), 1435–1441.

Tang, L., Liu, G.N., Qian, J., et al., 2016. Discussion on the measurement of the surface tension coefficient by the pull-off method. Eur. J. Phys. 37 (2), 025801.

Tian, Y., Xu, B., Chen, Y., et al., 2019. Liquid surface tension and refractive index sensor based on a side-hole fiber Bragg grating. IEEE Photonics Technol. Lett. 31 (12), 947–950.

Vargaftik, N., Volkov, B., Voljak, L., 1983. International tables of the surface tension of water. J. Phys. Chem. Ref. Data 12 (3), 817–820.

Wang, B.Y., Mao, X.Q., Liu, Y., et al., 2020. Measuring liquid surface tension coefficient based on optical fiber tension sensor. Phys. Exp. 40 (7), 16–18.

Zhao, H.W., Li, M.X., Bai, S.G., 2007. Research on the maximum bubble pressure method to measure the liquid surface tension coefficient. Phys. Exp. 27 (7), 36–38.

Zhu, Y.H., Kang, J., Sang, T., et al., 2015. Liquid surface tension coefficient measurement and temperature impact based on optical interference method. J. Optoelect. Laser 26 (1), 130–134.

Interferometric Fabry-Perot sensors for ultrasound detection on the tip of an optical fiber

Sudhansu Sekhar Khuntia, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sskhuntia88@gmail.com*

Abhishek Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, abhishekdas2256@gmail.com*

Manoranjan Sahoo, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, manoranjansahoo14@gmail.com*

Prajnadipta Sahoo, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, prajnadipta567@gmail.com*

ARTICLE INFO

Keywords:
Fabry-Perot
Acoustic detection
Immaging

ABSTRACT

To satisfy the requirements of the precision medicine, the demand for miniaturized online, real time and localized diagnostic tools is constantly increasing. Local acoustic inspection is assuming a large interest with particular reference to both oncology and neurology. With these considerations in mind, in this paper, we focus our attention on a comparative analysis of the best available solutions which can be exploited to realize a high frequency acoustic detector based on Fabry- Perot interferometry, located on the tip of an optical fiber.

1. Introduction

An increasing interest is being devoted to the concept of precision medicine which is becoming a golden rule to reduce any negative collateral effect in any therapy or diagnostic process (Hodson, 2016; König et al., 2017; Collins and Varmus, 2015). It can be declined into two main approaches. The first one is based on treatments designed according to the genomic characteristics of the patient (Badani et al., 2015; Aldape et al., 2015). The second one tends to avoid systemic drug delivery approach trying either to release minor drug quantities but with high precision or to perform highly localized surgery or therapeutic treatments (Garraway et al., 2013; Hsu et al., 2013; Alphandéry, 2020; Wiwatchaitawee et al., 2021). This is becoming almost a golden rule in any oncology or neurology treatment. The final objective is to cancel or at least minimize the negative side effects of many therapeutic approaches. An ideal approach demands for the adoption of minimally invasive strategies capable of selectively attacking cancer cells by reducing the collateral effects on the healthy environment (Vaiano et al., 2016). In order to get this goal, very precise online and real time diagnostic tools are required in order to drive the therapy in a very localized way independently of his specific nature (e.g., classical surgery, laparoscopy, particle beams, radiology, both laser and ultrasound treatments) (Bush et al., 2017). In many cases, this strategy requires for highly precise diagnostic tools to process the area of interest with a high spatial resolution very often far beyond the performance of the actual instrumentation. On this line of argument, the ultrasound echography assumes a particular relevance for the absence of negative collateral effects (Kuo et al., 2017). The classic echography indeed fails when high resolution is required in very small portions of the body like in the case of any treatment independently of its specific nature (Carotenuto et al., 2019). As long as it is well known that the resolution improves when the frequency increases but, on the other hand, the attenuation is proportional to the square of the frequency. Taking in mind that is not possible to increase the acoustic frequency without many negative collateral effects. The actual system typically works around a few tens of megahertz with a penetration depth at higher frequencies within few centimeters. This implies that a higher spatial resolution is needed to drive specific localized therapies in a more efficient/effective way (Ting et al., 2010). This goal can be achieved only if the acoustic source/detector is placed very close to the area to be investigated. A possible solution with a very poor invasiveness, is to employ of either a needle or of a catheter. This allows to say that the echograph must be integrated inside the needle. For the reason above, some technologies (e.g., Mach-Zehnder configuration and Michelson configuration (Teixeira et al., 2014) are not considered because of their huge dimensions. The best solution is offered by the technology of the optical fiber sensors (Poduval et al., 2017) taking advantage by the concept of lab of fiber (Vaiano et al., 2016), which is the natural extension of the concept of lab on chip on the tip of an optical fiber. The employment of optical fiber enables to achieve extremely small dimensions of the sensor (around tens of micrometers) thus allowing its insertion inside a needle. This configuration, with the sensor placed very close the under-test object, generates the possibility to get high resolution with low invasiveness. The actual technology allows to integrate high frequency acoustic sources and detectors on an optical fiber (Poduval et al., 2017). This possibility to obtain an all-in system brings the advantages to reduce the overall dimension, gain electromagnetic interference immunity, low invasiveness and to have a compact system. This paper focusses the attention on the detection solutions by performing a comparative analysis among a set of the most performant ones, that introduce a particular of innovation (e.g., new manufacturing method, lenses or different materials). The solutions are based on the Fabry-Perot interferometric (FP) mechanism that represents the best synergy between high resolution and small dimension, opening new doors to the detection mechanism, not only exclusively related to medical applications. The FP sensors merge a simple implementation and cost-effective, providing advantages and an alternative to overcome the limitations of other aforementioned technologies. This paper is organized as follows: In the Section 2 the mechanism of acoustic detection based on Fabry-Perot interferometry on the terminal end of an optical fiber is briefly described; in Section 3 some proposed solutions are reviewed. The Section 3 is divided in two main categories deal with how the sensor is made: Polymeric film configuration and Diaphragm configuration. In the Section 4 the conclusion is given.

2. Acoustic detection with the terminal tip of an optical fiber

The structure is mainly made up by two components. The first one deals with the generation of the acoustic waves and the second one with the reception of the reflected waves. To generate the acoustic waves with an optical fiber it is possible to exploit either a piezoelectric (Li et al., 2020) or the photoacoustic principle (Li and Wang, 2009), generating them via optical fiber (Noimark et al., 2018). As for the receiver part, it is possible to create a fiber optic sensor using the Fabry-Perot interferometric mechanism. The Fabry-Perot interferometer (FP) is an interferometer with two reflective surfaces (with a high reflection coefficient), which form a cavity. The light entering the cavity generates multiple reflections which result in the presence of peaks in the reflection spectrum of the signal (Morris et al., 2009). The resonant nanostructure can be designed in such a way to provide a spectral feature in the range of wavelengths of interest (Teixeira et al., 2014). A Fabry-Perot interferometer made of optical fiber can be constructed using a wall of the fiber itself and a highly reflective surface. A material can be interposed between the two surfaces, either air or polymers, for a cavity that expands or restrains following external stimuli (Jingcheng et al., 2019). Alternatively, two reflecting surfaces can be used, deposited directly on the optical fiber, which has the unique purpose of transmitting and collecting light (Ansari et al., 2018). The Fabry-Perot sensor can be divided into two main categories Polymeric film and Diaphragm (Ma et al., 2013). The first is based on the realization of a cavity consisting of reflecting walls inside which there is a polymeric layer, whose thickness variation induces a shift in the reflection spectrum (Cox and Beard, 2007). When an acoustic wave hits the sensor tip, it modulates the thickness of the polymer spacer. this thickness modulation produces an optical phase shift of the light reflected by the two mirrors, as a consequence there is a corresponding modulation of the reflected optical power. This modulation is expressed as a peak in the reflection spectrum. The first configuration can be observed in Fig. 1. In Fig. 2, there is shown the second configuration, that consists in forming a cavity, using two reflective layers spaced with air (Jingcheng et al., 2019). The operation mechanism is similar to the previous structure. As the acoustic wave touches the sensor tip, the diaphragm vibrates accordingly. The induced vibration results in a variation in length between the diaphragm and the facet of the fiber. In this way, a phase modulation of the reflected light is obtained. Typical materials involved are for example ~10 nm thick metal coatings (e.g., gold) or ~10 μm thick polymeric films (Sun et al., 2015) (e.g., Parylene C (Teixeira et al., 2014), UV- curable adhesive (Tan et al., 2014), epoxy (Wang et al., 2015). The cavity must be designed considering some fundamental parameters and their mutual relationship, such as the length of the cavity itself is inversely proportional to the operating frequency or how the acoustic phase sensitivity is proportional to cavity thickness (Jingcheng et al., 2019).

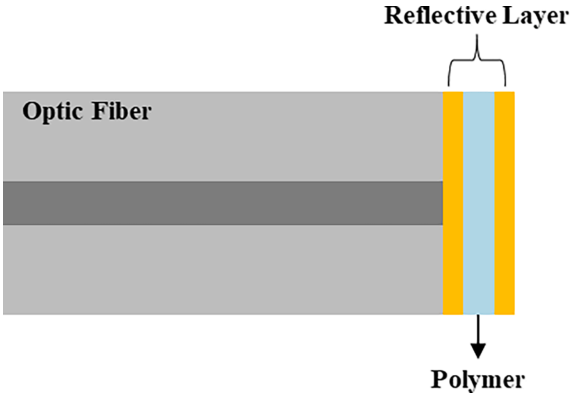


Fig. 1. Schematic illustration of a sensor based on Polymer configuration.

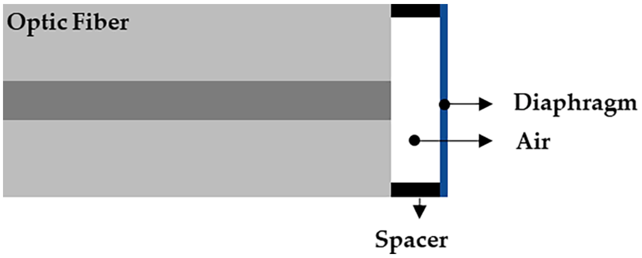


Fig. 2. Schematic illustration of a sensor based on diaphragm configuration.

3. Acoustic wave detection configuration

3.1. Sensor based on polymeric film

In this section four different sensors are described, as previously mentioned. These sensors are all based on a polymeric film between two reflecting surfaces but are distinguished mainly by the polymer used (consequently the manufacturing method) as well as by the dimensions. Starting from a classic polymeric film FP sensor, each sensor described introduces an innovative feature, resulting in a more compact and complex system.

Xu Guo et al. Proposed in 2020, a compact sensor consisting of both generator and receiver (Guo et al., 2020). The purpose of the sensor was to measure temperature variations (in vitro), correlating it with the time of flight of the detected acoustic waves. The main application was monitoring the temperature during the radiofrequency ablation to remove tumors. The generator was manufactured by depositing on a multimode fiber (400/425 μm) a photoacoustic material obtained by mixing Polydimethylsiloxane (PDMS) with gold salt. The absorption peak was detected at 530 nm with an acoustic band generated around 20 MHz. The receiver was made by depositing a 10 nm layer of gold on the tip of a single mode optical fiber. Before deposition, the fiber was treated and cut. A polymeric layer of PDMS was deposited on the first layer of gold, on top of which a second layer of gold (10 nm) was deposited, to form the cavity. When the light hits the sensor, the light is reflected by the two respective gold layers, generating a characteristic interference pattern. As the temperature increases, the polymer layer undergoes a variation in thickness, causing a variation in the optical path and consequently in the flight time. The overall dimensions of the receiver were only 125 μm, while those of the generator were 425 μm. The testing procedure include a room with pressure and temperature controlled. They verified that the forced temperature in the room matched with the temperature measured by the sensor. The match was performed by detecting the time of flight of the returned wavelength, managing to discretize the echoes with a difference of a few μs. Although the sensor has been tested exclusively for temperature monitoring, the small size and ease of manufacture open new doors to this type.

Yang et al. proposed a new innovative FP cavity (Yang et al., 2021). The structure finds his application in the medical field, given its small size and high sensitivity, it would be suitable for ultrasound measurements for image detection. The sensor was composed of a polymeric microcavity, deposited on a double cladding optical fiber (dimensions equal to 9 / 250 μm), as can be seen in Fig. 3. Following a theoretical study, the optimal length of the cavity and the relative radius of curvature has been identified, so that it has a very high reflectivity response. The cavity length exhibits a very precise range of 10–68 μm. The cavity was created and processed by means of two-photon three dimensions (3D) photolithography, using an innovative machine, namely the Photonic Professional GT, Nanoscribe, Germany. This machine can reproduce any structure, solving not only the processing problem, but also allowing the use of different photoresists with refractive indices and differentiated properties. This structure, in addition to the innovative aspect of using a latest generation machinery, guarantees precision and efficiency in terms of batch production. It leaves the opportunity to create structures free from typical manufacturing problem. The working optical wavelength was 1530 nm. The measured sensitivity of the FP sensor was found to be 13 Pa as a peak-to-peak value, with a bandwidth of 20 MHz. In terms of mV / kPa, the sensitivity measured under an acoustic pressure of 107 kPa and 302 mV of output voltage was equal to 2.82 mV / kPa. This device has several advantages including the ease of mass production, together with the small size, but compared with other sensors the dimension and the sensitivity are not so competitive.

Cholchester et al. had created an ultrasound probe with a diameter of<0.84 mm, which included receiver and transmitter both inserted in a stainless-steel tube (Colchester et al., 2015). It was developed for imaging of vascular tissues. The transmitter was made using an absorbent layer of carbon nanotubes (CNT), deposited on an optical fiber. The receiver based on the principle of a Fabry-Perot interferometer was made using a mono-modal optical fiber and a polymer-based cavity. Two highly reflective mirrors (98%) were used inside which a 20.1 μm thick epoxy layer was deposited. The mirrors and the epoxy layer were then covered with a 10 μm thick layer of Parylene- C, to protect the system. The transmitter and the receivers through a heat shrink tube were fixed to each other and finally inserted into a stainless-steel tube with an internal diameter of 0.84 mm to one millimeter from the end. The sensor was interrogated using a continuous wave laser in the wavelength range between 1500 and 1600 nm, while the acoustic transducer was modulated with a bandwidth equal to 20 MHz. Fig. 4 shows the experimental set up used to test the receiver and the transmitter, that was the system as a whole (ex vivo tests). An optical circulator was used to ensure that the interrogation light flows into the receiver, and at the same time the reflected light flows into a photodiode-transimpedance amplifier (custom to 50 MHz). The photodiode had two outputs, one operating at low frequencies (<50 kHz) and one at high frequencies (>500 kHz). The operating bandwidth of the ultrasonic sensor was 20 MHz, while the optical bandwidth was 1500–1600 nm. Due to the steel tube, artifacts were present in the acquired signals, removed following processing before reconstruction of the resulting image. The samples under test, consisting of a porcine aorta and a carotid artery (taken ex vivo), were placed on an ultrasound polyurethane attenuator. The sample vessels fixed with a cork ring were opened longitudinally, to allow internal imaging. The image acquisition time included a stabilization time of 100 ms, and 400 lateral passes with a 50 μm step. The lateral resolution was better as the depth was reduced, while the axial resolution increased with increasing depth. For depths equal to 3.5 mm, 1.5 mm and 10.5 mm, the lateral resolution varied respectively by 88 μm, 137 μm and 230 μm while the axial resolution at 2.5 mm was 58 μm and at other depths it was included in the range 58–79 μm, as shown in Fig. 5. For penetration thicknesses ranging from 1 to 10 mm, they could obtain a resolution of about 70 μm. This structure, despite the defects due to crosstalk, still manages to obtain an overall resolution of the order of a few microns, making it particularly innovative.

Guggenheim et al. in 2017 (Guggenheim et al., 2017), repurposed a sensor for ultrasound detection based on a polymeric plane-concave microresonator. The application was imaging acquisition, where the validation tests were performed in-vivo and ex-vivo. The microresonator consists of a micrometric cavity (100 μm in size) with a flat side resting on the optical fiber and a concave side. Both sides were made up of highly reflective dielectric mirrors (R > 98%), inside which there was a polymeric layer of similar shape. The tip of the single-mode optical fiber was first coated with a mirror dielectric substrate, then a drop of polymeric material was deposited. This material has the characteristics of being optically transparent and polymerizable by UV rays. Once stabilized, the drop was polymerized by UV rays, forming a smooth semi-spherical cap, on which the second layer of dielectric material was deposited, thus forming the cavity. The design of the cavity was extremely accurate, considering the radius of the concave shell such as to be perfectly matched to the diverging beam. In this way, the divergence and focus of the beam was corrected. 14 sensors with variable thickness (from 30 μm to 530 μm) have been designed and manufactured. To evaluate the acoustic performance, they characterized the sensor in terms of sensitivity, frequency response, directivity and practical ability to acquire biomedical images. As shown by the experimental data obtained, the sensitivity increases with increasing cavity thickness, at the same time by decreasing the thickness it was possible to obtain very broad bandwidths. Sensors that have a high sensitivity could obtain a penetration thickness in the order of centimeters if they had a cut-off frequency lower than 10 MHz. To show the actual imaging capacity they performed two different tests. One showed the micro vascularization of the individual capillaries of one ear of a mouse, and the other a swine aorta. In the first case an acquisition was made with pulsed photoacoustic methodology (with a 7 μm laser beam) on an area of about 8 mm × 8 mm. Using a gel for coupling with the optical fiber (set at 1.2 mm), it was possible to obtain high-contrast images and a wide field of view that demonstrated at the same time the omnidirectionality of the sensor. The second application example consisted in the acquisition of a three dimensions image, by raster scanning of two optical fibers. One fiber employed as laser ultrasound source and the other consisting of the microresonator. The three dimensions images were reconstructed from the sample returned echoes using the k-wave30 toolbox. The operating acoustic band oscillated from 0 to 70 MHz, with a wavelength for the

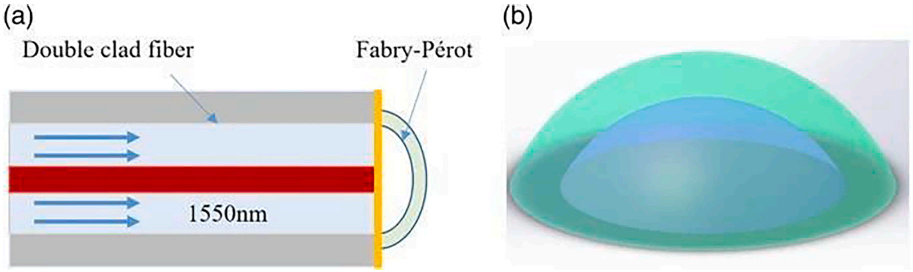


Fig. 3. (a) Schematic illustration of the sensor; (b) the microcavity (Yang et al., 2021).

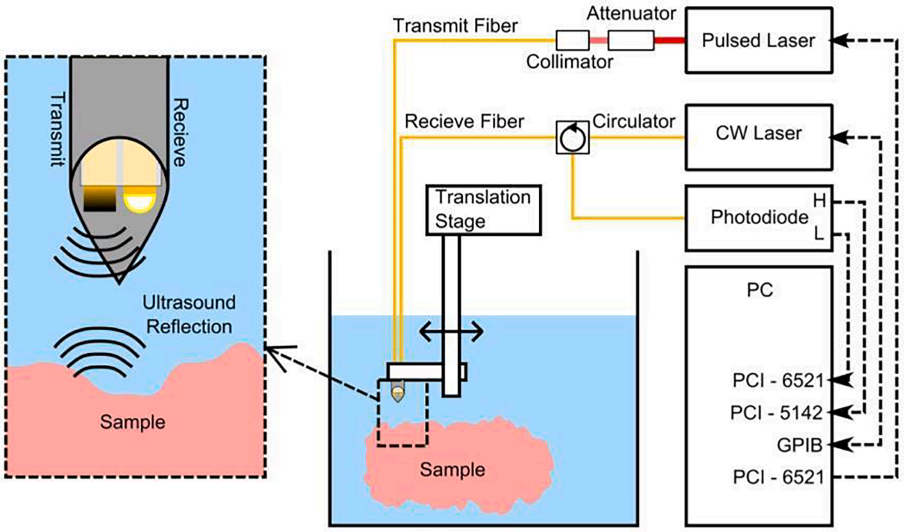


Fig. 4. Imaging system setup with an expanded section showing the ultrasound generation and detection fiber (based on epoxy film) within tubular metal housing (inner diameter 0.84 mm). (Colchester et al., 2015) Image © 2015 Optical Society ofAmerica.

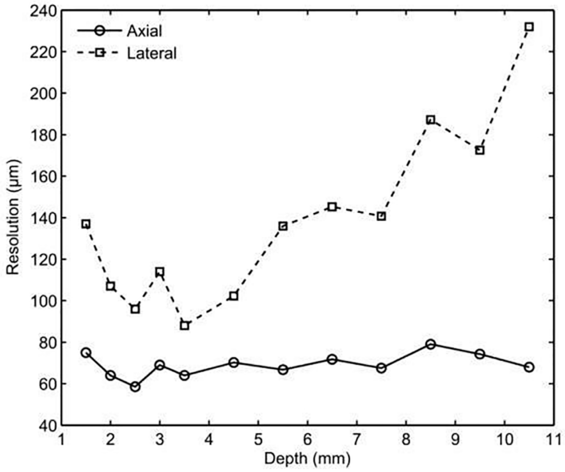


Fig. 5. Measured ultrasound imaging system resolution in the axial dimension (circles) and in the lateral dimension (squares) for different target depths, referring to the epoxy spacer sensor. (Colchester et al., 2015) Image © 2015 Optical Society of America.

optical sensor equal to 578 nm. The resulting spatial resolution obtained was approximately 125 µm. Both practical applications demonstrate how the sensor manages to be highly sensitive, omnidirectional, and flexible in design.

Table 1 shows the devices of the various groups, just described, in which the performances are compared. Colchester et al. exhibits an innovative sensor including both the generation and reception part (Colchester et al., 2015). This factor compensates for the large size (<0.84 mm) compared to other sensors (around 125 µm), while also exhibiting unrivaled spatial resolution (~70 µm). On the other hand, Guggenheim et al. propose a sensor with dimensions reduced to a minimum and a lower spatial resolution (~125 µm) than the previous group (~70 µm), but at the same time it is easy to manufacture (Guggenheim et al., 2017). The sensor, proposed by Guo et al. (Guo et al., 2020); used as a thermal probe, represents an interesting structure due to the simple manufacturing methods and small dimensions. The tests carried out (exclusively in terms of validation) do not allow a comparison in terms of performance. The sensor proposed by Yang et al. was fabricated using two-photon 3D photolithography (Photonic Professional GT, Nanoscribe, Germany) (Yang et al., 2021). The probe is also able to obtain advantages in terms of sensitivity, obtaining it equal to 13 Pa, with an overall dimension of 250 big than other. The dimensions are

Table 1
Comparison of the described devices based on polymeric films.

Title	Authors	Year	Type of sensor	Spatial resolution	Working acoustic frequency	Pressure sensitivity	Dimension
Optical Fiber Ultrasound Probe for Radiofrequency Ablation Temperature Monitoring: In-Vitro Results	Xu Guo et al. (Guo et al., 2020)	2020	Two gold film (10 µm) within PDMS on fiber tip.	n/a	20 MHz	n/a	125 µm
Ultrasonic signal detection based on Fabry– Perot cavity sensor	Wu Yang et al. (Yang et al., 2021)	2021	Double-cladding fiber with microcavity (30 µm) made with photoresist with Nanoscribe	n/a	20 MHz	13 Pa	250 µm
Broadband miniature optical ultrasound probe for high resolution vascular tissue imaging.	Colchester et al. (Colchester et al., 2015)	2015	Epoxy spacer (20.1 µm) between two dielectric mirrors (98%) on fiber tip coated with Parylene-C.	Axial resolution 64 µm and lateral resolution was 88 µm.	20 MHz	n/a	<0.84 mm
Ultra-sensitive planoconcave optical microresonator for 3 ultrasound sensing	Guggenheim et.al (Guggenheim et al., 2017)	2017	A solid planoconcave UV-curable liquid polymer microcavity formed between two highly reflective dielectric mirrors (R > 98%) on the top of an optical fiber.	~125 µm	0–70 MHz	n/a	125 µm

obviously smaller than the Colchester et al. sensor, but they include both the generator, the receiver and the needle (Colchester et al., 2015).

3.2. *Sensor based on diaphragm*

In the previous section, the sensors described were made up of a polymeric film. This section describes four sensors based on two reflective layers spaced with air (often referred to as diaphragm configuration). All the sensors described use a single mode fiber on the tip of which the diaphragm is manufactured, in a different way both in terms of geometry and materials.

Zhang et al. presented an Interferometric Sensor based on the diaphragm mechanism for ultrasound sensing and imaging (Zhang et al., 2017). The concept idea dealt with the creation of a cavity an epoxy resin film on the optical fiber tip. They first spliced a hollow core fiber (with a diameter of 100 μm) to a single mode fiber (SMF), with a diameter of 106.31 μm. The end of the hollow core fiber (HCF) was dipped into 353ND adhesive epoxy resin. The 353ND adhesive was a two-component epoxy resin with not only the characteristic of reliability and the resistance to high temperature environment up to 250 °C, but also a lower young's modulus (~1 GPa). In this component, it was mixed with the proportion of 10:1. After curing process, the epoxy resin created a diaphragm with the HCF. To protect the structure from possibly harmful external agents such as water or other fluids, a surrounding layer had been added to the overall component. The protective layer was made by inserting the structure, composed of the two fibers and the resin layer, in a thin glass tube (with inner diameter of 0.3 mm). A thin layer of polyethylene was deposited on the top of one of the ends. This final step was made with particular attention to keeping the diaphragm of 353ND resin and the polyethylene film well separated, to avoid mutual interferences. The Fabry- Perot cavity was made by two interfaces, that generate two reflections in spectrum, one was the interface SMF-Air and the second was 353ND resin-air. As a first step Zhang et al. decided to acquire the spectra of the samples before and after packaging, to evaluate possible influence of the protection. Then, they had evaluated the response in temperature, concluding that there was a slight reduction in the extinction ratio for temperatures around 250° C. The most interesting experiment was the testing of the device as ultrasound probe for imaging. The sideband filter technique has been used for its advantages including the low cost and the ability to reduce fluctuations in source output power. The test model consisted of a water container inside which a plexiglass rectangle was positioned on copper supports, to avoid overlapping of the return signals from the bottom of the two structures. In Fig. 7 it is possible to see the set up that was used. The rectangular Plexiglas block has 50 cm × 50 cm × 6 cm dimensions. To generate a square wave with a frequency of 300 Hz, a piezoelectric was used. The optical fiber was mounted together with the piezoelectric, suitably spaced 5 cm apart, on a scanner built above the box containing the water. A tunable laser was used (Santec, TSL-710), having as characteristics an output power of 20 mW and a line width of 100 kHz. The pulse ultrasonic excitation was at 300 kHz and the optical bandwidth was from 1480 to 1640 nm. The system formed by the piezoelectric and the sensor could scan a surface with a radius of 6 cm and a pitch of 1 cm. To perform a 2D surface scan, it took about 8 min. With this type of device, it was possible to obtain a longitudinal spatial resolution of approximately 5 mm. In conclusion, the structure has the qualities of being compact, stable in temperature, resistant to external agents and above all quite small (the internal diameter of the outer layer was 0.3 mm). The performance in terms of spatial resolution is disappointing, compared to other structures since exhibits a weak resolution of 5 mm. Despite the very low resolution, the component is inspiring from the point of view of the composition of the structure.

Zhang et al. proposed in 2018 a device consisting of a collimator and a Fabry-Perot interferometer, demonstrating experimentally how the combination of these two elements lead not only to fringe visibility enhanced but also to obtain a high resolution (Zhang et al., 2018). The first step for the realization of the sensor was the splicing of a single mode optical fiber and one with a graded index, which increases in the radial direction. This choice was based on the consideration that the outgoing light was collimated thanks to the focusing effect, minimizing divergence. The same group had shown (Wang et al., 2016) that the optimal choice for the length of the gradual index fiber was precisely that of 260 μm, which had an extremely low angle of divergence. Once the splicing was carried out, the two fibers were inserted into a plexiglass tube (internal and external diameter respectively of 0.4 mm and 1.6 mm) so that the graded index fiber protruded for its entire length from the end of the tube. This tube was used exclusively to make the fiber adhere better to the inside of an aluminum tube. At the end of the aluminum tube (internal diameter of 2 mm) a 30 nm film of PTFE (polytetrafluoroethylene) was glued using the T530 glue, to form a FP optical fiber interferometer. A diagram of the sensor was proposed in Fig. 6, where it is possible to see the structure as a whole and the total dimensions (around 2 mm). To test the capacities of the sensor, a piezoelectric device with a frequency of 300 kHz was used. In Fig. 7 it is possible to view the set up used, briefly described in the last one work. Specifically, the piezoelectric and the FP sensor moved simultaneously along the surface of the water for 5 cm with 1 mm pitch. The images reported in Fig. 8, show how the surfaces of the blocks can be well discriminated. The images were reconstructed starting from the calculation of the time of flight and subsequent processing and amplifications. Mainly finalized for seismic imaging, this application shows how to obtain a sensor with a total size of 2 mm and a resolution approximately equal to 3.7 mm.

Q. Rong Ruixiang et al. developed an innovative FP sensor based on optical fiber and the use of focusing lenses, in 2017 (Rong et al., 2017). The sensor was made mainly for imaging through ultrasonic waves in the field of seismic monitoring. The sensor was built starting from a single-mode optical fiber (SMF), inserted in the central hole of a ceramic tube at the end of which there was a gold foil (130 nm thick). The fiber and the gold layer were 100 μm apart from each other. This system was inserted into an additional plastic tube terminated by a focusing acoustic lens. The lens was made of Polymethyl methacrylate (PMMA) and spherical in shape, to focus the acoustic lenses directly on the gold foil. The total size of the sensor was mainly given by the size of the lenses. Three types of sensors with three different lenses were produced and tested: 9 mm, 16 mm and 23 mm. The optical working wavelengths adopted to test the sensors was 1480 to 1640 nm. Following a testing phase, it emerged that the sensor could reconstruct the actual structure under test, creating a 2-D image. The sensitivity in terms of pressure, measured, appeared to be in the range 0 ÷ 60 μm / psi (function of thickness gold and cavity length). This sensor represents a valid inspiration, considering the use of lenses, but at the same time the big size limits its use for medical applications.

Guo et al. proposed in 2019 the design, testing and fabrication of a sensor based on a Fabry-Perot fiber optic interferometer (Guo et al., 2019). The purpose of this sensor was to take pressure measurements. The overall sensor size was 125 μm. The FP interferometer was created by creating a silica diaphragm on the tip of the optical fiber. The manufacture of the sensor can be divided into three steps. The first was the manufacture of the silica diaphragm (1.2 μm thickness). The second step consists of the splicing of a single mode and a multimode fiber (the core made of germanium doped silica with a diameter of 105 μm, and the cladding made of pure silica with a diameter of 125 μm), followed by an etching phase of the multimode fiber to create the cavity (of 50 μm). The third step consisted in the thermal bonding of the diaphragm on the cavity. To carry out a component testing, the sensor was placed inside a controlled room for both temperature and pressure. The optical working wavelength was 1520 to 1570 nm. Tests revealed sensor sensitivity to static pressure of 12.4 nm / kPa. The high sensitivity and minimal dimensions make the device suitable for medical applications.

In Table 2 is possible to observe the previously described devices compared. Small dimensions and simplicity of manufacturing of the

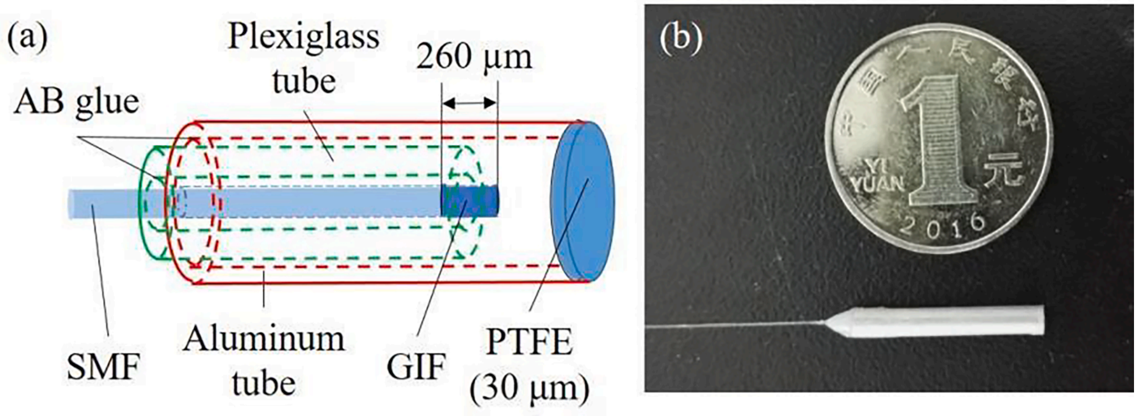


Fig. 6. (a) Sensor's structure; (b) dimension of the sensor. (Zhang et al., 2018) Image © 2018 Optical Society of America under theterms of the OSA.

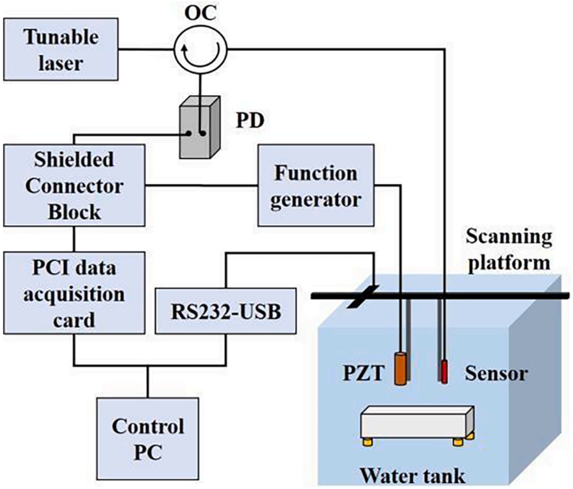


Fig. 7. Schematic illustration of the set up used to test the diaphragm sensor made with PTFEE. (Zhang et al., 2018) Image © 2018 Optical Society of America under theterms of the OSA.

sensor are offered by Zhang et al. (Zhang et al., 2017); who with their 0.3 mm probe manage to obtain a scarce 5 mm of resolution (Zhang et al., 2017). The same group in 2018 then re-proposed another sensor, made in a different way from the previous one, managing to reach an Inferred resolution up to 3.7 mm, at the expense of the dimensions that reach 1.6 mm (Zhang et al., 2018). The sensor proposed by Qiangzhou Rong Ruixiang (Rong et al., 2017), had the peculiarity of introduce the use of a lens for focusing the acoustic waves. Despite the excellent performance of the probe, the large dimensions of the device limit its use in medical imaging. The use of acoustic focusers is very interesting. Despite the complex manufacturing process, the structure, proposed by Guo et al., offers compact and small dimensions (125 µm), as well as a remarkable sensitivity to pressure in the order of 12 nm / kPa (Guo et al., 2019).

4. Conclusions and future trends

In this review, we analyzed the best solutions available for ultrasound detection by integrating a Fabry Perot interferometer on the tip of an optical fiber. There are two main categories to differentiate some operative structures: the polymeric one and the diaphragm based. The diaphragm configuration results to be the one which exhibits sensors with big overall dimensions (around millimeters). The resolution results remarkably low, not suitable for medical and biological imaging application. One of the series proposed by Xu Guo et al. exhibited a high sensitivity to pressure and petite dimensions. Despite that, the sensors had a complex fabrication (Guo et al., 2019). The polymeric configuration is the best structure in terms of performance and fabrication methods. The resulting resolution is of the order of few tens of

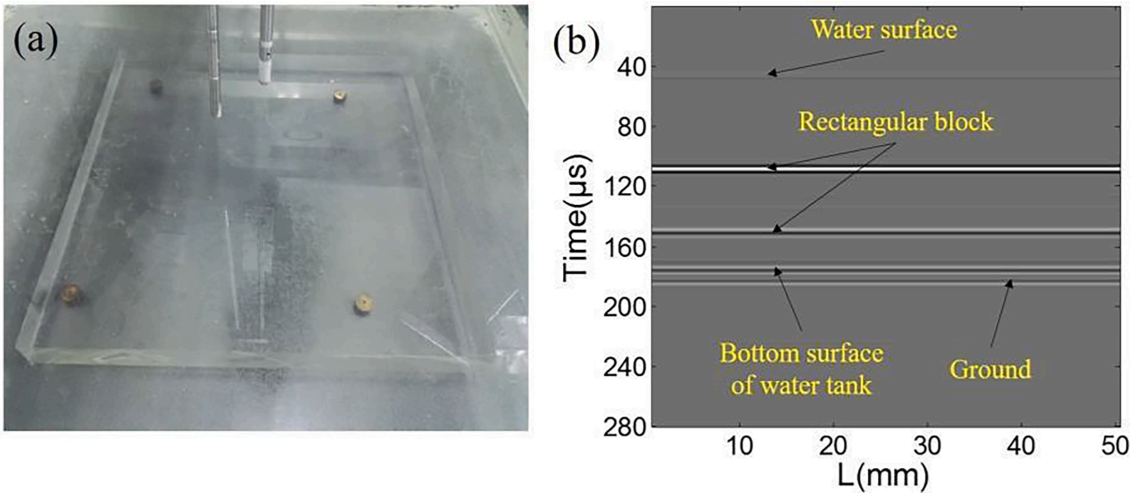


Fig. 8. (a) The water tank employed to test the PTFEE sensor. (b) The image resulting (Zhang et al., 2018). Image © 2018 Optical Society of America under the terms of the OSA.

Table 2
comparison of the described devices based on diaphragm.

Title	Authors	Year	Type of sensor	Spatial resolution	Working acoustic frequency	Pressure sensitivity	Dimension
An Optical Fiber Fabry–Perot Interferometric Sensor Based on Functionalized Diaphragm for Ultrasound Detection and Imaging	Zhang et al. (Zhang et al., 2017)	2017	SMF spliced to Hollow core fiber coated with 353ND (two component epoxy resin)	5 mm	300 kHz	n/a	~0.3 mm
Ultrasonic imaging of seismic physical models using a fringe visibility enhanced fiber-optic Fabry-Perot interferometric sensor	Zhang et al. (Zhang et al., 2018)	2018	Section of Graded index spliced to a SMF (for collimation). Around an aluminum tube (2 mm) with PTFE on the top.	Inferred up to 3.7 mm	300 kHz	n/a	~1.6 mm
Ultrasonic Sensitivity- Improved Fabry–Perot Interferometer Using Acoustic Focusing and Its Application for Noncontact Imaging	Ruixiang et al. (Rong et al., 2017)	2017	SMF in a ceramic tube at the end 130 nm of gold foil. Spaced with air, then there is an acoustic focus lens.	0 ÷ 60 μm/psi Function of thickness gold and cavity length	330 kHz	n/a	9 to23 mm
Highly Sensitive Miniature All- Silica Fiber Tip Fabry–Perot Pressure Sensor	Xu Guo et al. (Guo et al., 2019)	2019	SMF spliced to MMF (core: germanium doped silica) etched and on the top 1.2 μm silica diaphragm	n/a	n/a	12 nm/kPa	125 μm

micrometers. All the described solutions introduce an innovative detail, whether it is the simplicity of construction (Guo et al., 2020) or the use of tools for the enhancement of acoustic waves, as the employ of focusing lenses. Three sensors stand out among all for their characteristics. One introduces the possibility of using an innovative machine (Nanoscribe) that would allow both prototypes and mass production (Yang et al., 2021). The sensor proposed by Colchester et al. demonstrates that is possible to build an all-in system (generator part and receiver part in a needle) with high performance. This work offers an excellent trade-off between working frequency, probe size and spatial resolution (Colchester et al., 2015). In fact, it can obtain high resolutions between 64 and 88 μm, with a probe having a total size of<0.84 mm. The last one is the sensor proposed by Guggenheim et. al, that introduces the employment of UV-curable liquid polymer, describing an easy fabrication process (Guggenheim et al., 2017). In terms of the overall dimensions of the probe, the sensor manages to be of the order of 125 μm, considerably smaller than the others. All of three, present some limits that should be overcome. The best solution would be a configuration that include the strong points of each proposed solution.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

References

R. Hodson Precision medicine Nature 537 7619 2016 S49 S49 10.1038/537S49a.
König, I.R., Fuchs, O., Hansen, G., von Mutius, E., Kopp, M.V., 2017. What is precision medicine? Eur. Respirat. J. 50 (4), 1700391. https://doi.org/10.1183/13993003.00391-2017.
Collins, F.S., Varmus, H., 2015. A new initiative on precision medicine. New Engl. J. Med. 372 (9), 793–795. https://doi.org/10.1056/NEJMp1500523.
Badani, K.K., Thompson, D.J., Brown, G., Holmes, D., Kella, N., Albala, D., Singh, A., Buerki, C., Davicioni, E., Hornberger, J., 2015. Effect of a genomic classifier test on clinical practice decisions for patients with high-risk prostate cancer after surgery. BJU Int. 115 (3), 419–429. https://doi.org/10.1111/bju.12789.
Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., von Deimling, A., 2015. Glioblastoma: pathology, molecular mechanisms and markers. Acta Neuropathol. 129 (6), 829–848. https://doi.org/10.1007/s00401-015-1432-1.
Garraway, L.A., Verweij, J., Ballman, K.V., 2013. Precision oncology: an overview. J. Clin. Oncol. 31 (15), 1803–1805. https://doi.org/10.1200/JCO.2013.49.4799.
Hsu, W., Markey, M.K., Wang, M.D., 2013. Biomedical imaging informatics in the era of precision medicine: progress, challenges, and opportunities. J. Am. Med. Inform. Assoc. JAMIA 20 (6), 1010–1013. https://doi.org/10.1136/amiajnl-2013-002315.
Alphandéry, E., 2020. Nano-therapies for glioblastoma treatment. Nano-Therap. Glioblastoma Treat. Cancers 12 (1), 242. https://doi.org/10.3390/cancers12010242.
Wiwatchaitawee, K., Quarterman, J.C., Geary, S.M., Salem, A.K., 2021. Enhancement of therapies for glioblastoma (GBM) using nanoparticle-based delivery systems. AAPS Pharm. SciTech 22 (2), 71. https://doi.org/10.1208/s12249-021-01928-9.
Vaiano, P., Carotenuto, B., Pisco, M., Ricciardi, A., Quero, G., Consales, M., Crescitelli, A., Esposito, E., Cusano, A., 2016. Lab on Fiber Technology for biological sensing applications. Laser Photonics Rev. 10 (6), 922–961. https://doi.org/10.1002/lpor.201600111.
Bush, N.A., Chang, S.M., Berger, M.S., 2017. Current and future strategies for treatment of glioma. Neurosurg. Rev. 40 (1), 1–14. https://doi.org/10.1007/s10143-016-0709-8.
Kuo, W.-C., Kao, M.-C., Tsou, M.-Y., Ting, C.-K., Stieger, K., 2017. In vivo images of the epidural space with two- and three-dimensional optical coherence tomography in a porcine model. PLoS ONE 12 (2), e0172149. https://doi.org/10.1371/journal.pone.0172149.
Carotenuto, B., Ricciardi, A., Micco, A., Amorizzo, E., Mercieri, M., Cutolo, A., Cusano, A., 2019. Optical fiber technology enables smart needles for epidurals: an in-vivo swine study. Biomed. Opt. Express 10 (3), 1351. https://doi.org/10.1364/BOE.10.001351.
Ting, C.K., Tsou, M.Y., Chen, P.T., Chang, K.Y., Mandell, M.S., Chan, K.H., Chang, Y., 2010. A new technique to assist epidural needle placement: fiberoptic-guided insertion using two wavelengths. Anesthesiology 112 (5), 1128–1135. https://doi.org/10.1097/ALN.0b013e3181d3d958.
Teixeira, J.G.V., Leite, I.T., Silva, S., Frazão, O., 2014. Advanced fiber-optic acoustic sensors. Advanced fiber-optic acoustic sensors. Photon. Sens. 4 (3), 198–208. https://doi.org/10.1007/s13320-014-0148-5.
Poduval, R.K., Noimark, S., Colchester, R.J., Macdonald, T.J., Parkin, I.P., Desjardins, A. E., Papakonstantinou, I., 2017. Optical fiber ultrasound transmitter with electrospun carbon nanotube-polymer composite. Appl. Phys. Lett. 110 (22), 223701. https://doi.org/10.1063/1.4984838.
Li, Z., Li, J., Chen, H., 2020. Experimental research on excitation condition and performance of airflow-induced acoustic piezoelectric generator. Micromachines 11 (10), 913. https://doi.org/10.3390/mi11100913.
Li, C., Wang, L.V., 2009. Photoacoustic tomography and sensing in biomedicine. Phys. Med. Biol. 54 (19), R59–R97. https://doi.org/10.1088/0031-9155/54/19/R01.
Noimark, S., Colchester, R.J., Poduval, R.K., Maneas, E., Alles, E.J., Zhao, T., Zhang, E.Z., Ashworth, M., Tsolaki, E., Chester, A.H., Latif, N., Bertazzo, S., David, A.L., Ourselin, S., Beard, P.C., Parkin, I.P., Papakonstantinou, I., Desjardins, A.E., 2018. Polydimethylsiloxane composites for optical ultrasound generation and multimodality imaging. Adv. Funct. Mater. 28 (9), 1704919. https://doi.org/10.1002/adfm.201704919.
Morris, P., Hurrell, A., Shaw, A., Zhang, E., Beard, P., 2009. A fabry-perot fiber-optic ultrasonic hydrophone for the simultaneous measurement of temperature and acoustic pressure. J. Acoust. Soc. Am. 125 (6), 3611–3622. https://doi.org/10.1121/1.3117437.
Jingcheng, Z., Guo, X., Du, C., Cao, C., Wang, X., 2019. A fiber optic ultrasonic sensing system for high temperature monitoring using optically generated ultrasonic waves. Sensors 19 (2), 404. https://doi.org/10.3390/s19020404.
Ansari, R., Zhang, E.Z., Desjardins, A.E., Beard, P.C., 2018. All-optical forward- viewing photoacoustic probe for high-resolution 3D endoscopy. Light Sci Appl. 7, 1–9. https://doi.org/10.1038/s41377-018-0070-5.
Ma, J., Xuan, H., Ho, H.L., Jin, W., Yang, Y., Fan, S., 2013. Fiber-optic fabry– pérot acoustic sensor with multilayer graphene diaphragm. IEEE Photon. Technol. Lett. 25 (10), 932–935. https://doi.org/10.1109/LPT.2013.2256343.
Cox, B., Beard, P., 2007. The frequency-dependent directivity of a planar fabry-perot polymer film ultrasound sensor. IEEE Trans. Ultrason., Ferroelect., Freq. Contr. 54 (2), 394–404.
Sun, B., Wang, Y., Qu, J., Liao, C., Yin, G., He, J., Zhou, J., Tang, J., Liu, S., Li, Z., Liu, Y., 2015. Simultaneous measurement of pressure and temperature by employing Fabry-Perot interferometer based on pendant polymer droplet. Opt. Express 23 (3), 1906–1911. https://doi.org/10.1364/OE.23.001906.
Tan, X.L., Geng, Y.F., Li, X.J., Deng, Y.L., Yin, Z., Gao, R., 2014. UV- curable polymer micro hemisphere-based fiber-optic fabry-perot interferometer for simultaneous

measurement of refractive index and temperature. IEEE Photon. J. 6 (4), 1–8. https://doi.org/10.1109/JPHOT.2014.2332460.

Wang, J., Wu, S., Ren, W., 2015. Epoxy resin cap-based low-finesse fiber fabry-perot interferometer for temperature sensing. IEEE Sens. J. 15 (11), 6385–6389. https://doi.org/10.1109/JSEN.2015.2458895.

Guo, X.u., Zhou, J., Du, C., Wang, X., 2020. Optical fiber ultrasound probe for radiofrequency ablation temperature monitoring: in-vitro results. IEEE Photon. Technol. Lett. 32 (12), 689–692. https://doi.org/10.1109/LPT.2020.2991720.

Yang, W.u., Zhang, C., Zeng, J., Song, W., 2021. Correction to: ultrasonic signal detection based on Fabry-Perot cavity sensor. Vis. Comput. Ind. Biomed. Art 4 (1). https://doi.org/10.1186/s42492-021-00079-9.

Colchester, R.J., Zhang, E.Z., Mosse, C.A., Beard, P.C., Papakonstantinou, I., Desjardins, A.E., 2015. Broadband miniature optical ultrasound probe for high resolution vascular tissue imaging. Biomed. Opt. Express 6 (4), 1502–1511. https://doi.org/10.1364/BOE.6.001502.

Guggenheim, J.A., Li, J., Allen, T.J., Colchester, R.J., Noimark, S., Ogunlade, O., Parkin, I.P., Papakonstantinou, I., Desjardins, A.E., Zhang, E.Z., Beard, P.C., 2017. Ultrasensitive plano-concave optical microresonators for ultrasound sensing. Nat. Photon. 11 (11), 714–719. https://doi.org/10.1038/s41566-017-0027-x.

Zhang, W., Wang, R., Rong, Q., Qiao, X., Guo, T., Shao, Z., Li, J., Ma, W., 2017. An optical fiber Fabry-Perot interferometric sensor based on functionalized diaphragm for ultrasound detection and imaging. IEEE Photon. J. 9 (3), 1–8.

Zhang, W., Chen, F., Ma, W., Rong, Q., Qiao, X., Wang, R., 2018. Ultrasonic imaging of seismic physical models using a fringe visibility enhanced fiber-optic Fabry-Perot interferometric sensor. Opt. Express 26 (8), 11025. https://doi.org/10.1364/OE.26.011025.

Wang, R., Liu, Z., Qiao, X., 2016. Fringe visibility enhanced Fabry-Perot interferometer and its application as gas refractometer. Sens. Actuat. B 234, 498–502. https://doi.org/10.1016/j.snb.2016.05.028.

Rong, Q., Zhou, R., Hao, Y., Yin, X., Shao, Z., Gang, T., Qiao, X., 2017. Ultrasonic sensitivity-improved fabry–perot interferometer using acoustic focusing and its application for noncontact imaging. IEEE Photon. J. 9 (3), 1–11.

Guo, X.u., Zhou, J., Du, C., Wang, X., 2019. Highly sensitive miniature all- silica fiber tip fabry-perot pressure sensor. IEEE Photon. Technol. Lett. 31 (9), 689–692. https://doi.org/10.1109/LPT.2019.2904420.

Advances in fiber optic sensors for soil moisture monitoring: A review

Madhusudan Das, *Department of Electronics and Communication Engineering , Raajdhani Engineering College, Bhubaneswar, madhusudandas55@hotmail.com*

Somnath Mishra, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, somnathmishra@yahoo.co.in*

Prangya Paramita Padhi, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, prangya.p.padhi@gmail.com*

Supriya Nayak, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, supriyanayak443@gmail.com*

ARTICLE INFO

Keywords:
Fiber optics sensor
Soil water measurements
Distrubuite fiber optic sensors
Fiber Bragg Grating
Smart farming

ABSTRACT

Real time measurements of soil water content play a critical role in many fields of science as agronomy, geology, engineering, and hydrology. In the last years, agriculture has become one of the most important application fields of soil water sensors technology in order to optimize the irrigation process and to guarantee sustainable water resources management. This work provides a review on the latest emerging methodologies based on optical fiber sensing for soil moisture monitoring for agricultural and hydrological applications. In particular, the main studies referring to optical fiber sensors based on a variation of the refractive index of the external medium, sensors based on heated distributed temperature sensor (HDTS) and sensors based on Fiber Bragg Gratings (FBG) or Long Period Gratings (LPG) are here explored. Finally, some approaches based on the NIR absorbance spectroscopy are proposed for measuring the soil water content. Most of these approaches and technologies are still in a prototyping phase and only few of them are properly evaluated in situ real context.

1. Introduction

The growing population, which corresponds to an increase in food demand, together with the climatic changes and global warming, which reduce water availability, have a significant impact on agriculture, which is one of the main sources for food production and for economic growth (Yin et al., 2021).

Therefore, there is an increasing need to improve crops production and optimize the use of water resources. In this regard, smart agriculture (or precision farming) represents an attractive and efficient solution to guarantee an improvement in the management of traditional agricultural practices. The recent technology innovations introduced by environmental monitoring systems are playing a key role in the development of precision farming and have the potential to be an important driver of the sustainable expansion of agricultural systems.

Many sensor applications have a significant impact on all agricultural practices. For instance, soil moisture sensors support farmers' decisions for irrigation scheduling to prevent plants from drought stress and over-application of irrigation (Yin et al., 2021). Moreover, the use of artificial intelligence systems integrated with sensors (mainly soil moisture and temperature sensors) is becoming an innovative approach to increase agricultural productivity and optimize natural resources (Kayad et al., 2020).

Monitoring the soil conditions provides key information not only to improve resource utilization to maximize farming outputs and minimize environmental side effects but also to build site-specific databases of relationships between soil conditions and plant growth for smart and sustainable agriculture systems. Novel technologies for collecting soil information are in demand to build efficient smart or precision agriculture systems. (Kayad et al., 2020)

In this scenario, the global soil moisture sensor market size was estimated at USD 173.6 million in 2018 and is expected to grow at a compound annual growth rate (CAGR) of 14.0% from 2019 to 2025. (`Soil Moisture Sensor Market Size, 2019)

Increasing the use of these sensors by the agricultural sector to improve crops productivity and optimize water utilization is expected to drive market growth over the forecast period.

Moreover, the use appropriate soil water content sensors also help in avoiding the irrigation issues through constant monitoring, thus spurring the growth of the market. The introduction of new technologies that aid residential owners to monitor soil moisture conditions of potted plants, vegetable gardens, and lawns are also expected to drive the market. (Soil Moisture Sensor Market Size, 2019)

Several techniques were developed to determine the soil moisture content as attested by the prestigious reviews about the traditional sensors present in literature (Susha Lekshmi et al., 2014; Dobriyal et al., 2012). The typical approach widely used to determine the water content in soil is based on the measurement of the changes of soil properties related to water content, such as the dielectric permittivity, matrix potential, and mass (volumetric water content) of soils (Schmugge et al., 1980; Robinson et al., 2008; Susha Lekshmi et al., 2014; Dobriyal et al., 2012).

There are different types of soil water content sensors available in the market, including resistive, capacitive, dielectric, and tensiometric sensors (Schmugge et al., 1980). Nevertheless, standard sensors are not compatible with monitoring large areas, as the use of a high number of sensors requires a high number of cables.

On the other hand, fiber optic sensors, which are capable of providing continuous real-time measurements, are presently used in many fields of science to measure and monitor various physical and chemical processes. Since these sensors are small and can access a measurement field with little disturbance, they may be especially suitable for the measurement of small-scale processes in soil, such as water content.

The goal of this paper is to show last advances in the optical sensing devices developed for soil moisture monitoring. In Section 2, the basic theories about soil moisture estimation are presented. Section 3 consists of an overview of soil moisture measurements performed using fiber optic sensors. Finally, in section 4, the Near-infrared spectroscopy approach to detect soil water content is reported.

2. Soil water estimation: basic theories

Water content is a measurement of the amount of water in the soil by weight or volume and is defined as the water lost from the soil upon drying to constant mass at 105 °C (Soil Science Society of America, 2008). It is expressed in units of either mass of water per unit mass of dry soil (kg/kg), and it is defined as gravimetric soil moisture content w, estimated according the following equation (Schmugge et al., 1980):

$$w = \frac{M_{wet} - M_{dry}}{M_{dry}}$$

where M_{wet} is the mass of wet soil and M_{dry} is the mass of dry soil. The gravimetric method is a direct method and results the most reliable technique applied for soil moisture determination. It is used as the standard for the calibration of all other soil moisture determination techniques (Schmugge et al., 1980; Robinson et al., 2008; Dobriyal et al., 2012).

Soil water content can be also expressed in units of volume of water per unit bulk volume of soil (m^3/m^3) (Kirkham, 2014) and the amount of moisture present is expressed in percentage as volumetric soil moisture content θ which is defined as the ratio of the volume of moisture present in the soil (V_{water}) to the total volume of the soil (V_{soil}):

$$\theta = \frac{V_{water}}{V_{soil}}$$

Both these parameters are related as reported in the following equation:

$$\theta = w \times \rho_b$$

where ρ_b is the dry bulk density of the soil.

Moreover, several solutions based on indirect methods, which are not object of the present paper, have been also proposed in literature for the water content estimation in soil (e.g dielectric sensors (Dasberg, 1985; Dean et al., 1987; Nadler et al., 1991; Whalley et al., 2013; Xu et al., 2012a,b; Topp and Daniel, 1998; Arulanandan, 1991; Cardenas-Lailhacar and Dukes, 2010; Schwartz et al., 2003) and resistive sensors (Susha Lekshmi et al., 2014; Dobriyal et al., 2012)). Such methods foreseen the measurement of a parameter that depends on the changes of SWC through physically based or empirical relationships.

3. Fiber optic-based sensor

Fiber optic sensors-based (FOS) are used in many fields of science to measure and monitor various physical and chemical processes. Since these sensors can be relatively small and can access a measurement arena with little disturbance, they may be especially suitable for the measurement of small-scale processes in soil, such as water dynamics. Moreover, FOS are capable of providing continuous real-time monitoring, and thanks to the multiplexing capabilities, they are able to provide measurements over large areas. Several studies are available in the literature concerning the development of fiber optical-based sensors for soil moisture monitoring. In particular, the main studies refer to sensors based on a variation of the refractive index of the external medium (Alessi and Lyle, 1986), sensors based on heated distributed temperature sensor (HDTS) (Striegl and Loheide, 2012; Steele-Dunne et al., 2010), and sensors based on Fiber Bragg Gratings (FBG) (Yeo et al., 2008) or Long Period Gratings (LPG) (Berruti et al., 2014).

One of the first attempts to develop an optical fiber-based sensor for soil moisture measurements was performed by Alessi and Prunty in 1985; they investigated the possibility to determinate soil water content using optical fibers (Alessi and Lyle, 1986).

They experimentally verified that the light transmission through an optical fiber, surrounded by porous media, decreases as the water content in the media increases but they did not provide a sensor calibration.

In 1999, Garrido et Al. verified the possibility to use a fiber optic mini probe (FOMP) for in situ measurements of soil water content (Garrido et al., 1999). The working principle of the proposed system is based on the relationship between volumetric water content and the intensity of light reflected back from the soil and in particular, on the variation of the reflection capacity of the soil in presence of water. The FOMP system is composed by: a light source, a fiber optic mini-probes consisting of bifurcated fiber optic bundles with diameters of 3.4 and 2.5 mm at the common end and a photodetector. In this simple design, a constant beam of light is sent through the input leg to the common end, where it exits the fiber optic bundle, interacts with the soil volume immediately in front of the probe and is partially reflected back into the probe. The reflected light is carried through the output leg to a photodetector for quantification. Three soils have been used, with different soil texture, porosity and bulk density, to construct and test calibrations of the FOMPs to measure soil water content in the range (Dong et al., 2017) of VWC %. Experimentally results demonstrated the correlation ($r^2 > 0.98$) between the soil light reflectivity and the water content. In particular, it was found that wetted soil reflects less light than dry soil, as the water absorbs the incident light.

The FOMP system's major advantages are that it measures water content at very small scales (15–20 mm^3) with high temporal resolution and it is very sensitive to small changes in water content. In addition, the system can be multiplexed to give a spatial water content distribution.

Nevertheless, this approach exhibits some limitations. Firstly, the calibration requires extremely accurate knowledge of the water content directly in front of the probe. Moreover, the FOMP's performance decrease in presence of coarse textured soil. The sensor was found to be able only to measure the first layer of soil particles that are in direct contact whit the probe. For all of these reasons, the FOMP is not suitable for soil water content measurements in agricultural applications.

Recently, Haroon et al presented another approach for soil moisture detection, based on the capability of a fiber optic sensor to detect any changes in the intensity of light that travels inside the fiber due to a variation of the refractive index of the material surrounding the fiber itself. The light is typically confined inside the core of the fiber by the cladding. When the cladding is removed, the light can leak out from the fiber core. It happens when the material used to wrap the fiber is removed (Haroon, 2018).

To this aim, they realized a prototype based on fiber optic with the fiber core exposed so that the evanescent field is in contact with the surrounding. Once the buffer coating and cladding are removed, the exposed fiber optic is covered with plaster of Paris and protected by the plastic cover at the outer layer. Plaster of Paris absorbs moisture from the soil around it. In presence of water, the refractive index of the material that covers the fiber changes. The authors experimentally demonstrated that in the wet conditions the output signal of the sensor increases while in dry conditions the output signal value decreases. The feasibility proposed in (Haroon, 2018) shows the capability of the proposed fiber optic probe to detect water. Nevertheless, the study does not provide a soil test where the drying process is investigated.

3.1. Heated fiber optic distributed temperature sensors

Fiber Optic Distributed Temperature Sensors (DTS) uses a fibre optic cable to provide continuous temperature measurements with high spatial and thermal resolution (up to 1-m spatial resolution and 0.01 °C) over the distance up of 30 km

. A Schematic of distributed DTS is shown in Fig. 1.

Fibre optics for distributed temperature measurements has been in use since 1980s (e.g. Dakin et al., 1985), it is a flexible and powerful tool to monitor the hydrological systems (e.g. Johansson and Farhadiroushan, 1999; Selker et al., 2006).

As a matter of fact, several approaches to measure the soil water content are based on DTS technique through a relationship between soil thermal conductivity and soil moisture (Striegl and Loheide, 2012; Olmanson and Ochsner, 2006).

There are two categories of DTS method for soil water measurement, namely Actively Heated Fiber Optics (AHFO) and passive DTS. AHFO system is shown in Fig. 2, its working principle consists in the application of an electrically generated heat pulse to the fiber optic cable and the resulting temperature change (ΔT) during or after the heat pulse is related to the water content; ΔT is dependent on the surrounding soil water content because heat dissipates away from the cable more efficiently in wet conditions than in dry conditions (Striegl and Loheide, 2012; Sourbeer and Loheide, 2016). The temperature variation ΔT can be estimated by the following equation (Florides and Kalogirou, 2008):

$$\Delta T = -\frac{Q}{4\pi\lambda}\left[ln\left(\frac{4\pi t}{r^2}\right) - \gamma\right]$$

where Q is the energy applied per unit of time [J/ms], λ is the soil thermal conductivity, r is the distance from the heating source [m], t is the time after the heat pulse [s] and γ is the Euler's constant.

Therefore, AHFO has the potential to map soil moisture over large areas with high spatial and temporal resolutions.

Passive DTS, on the other hand, uses soil thermal responses to the net solar radiation to estimate soil water (Krzeminska et al., 2012). Unlike the AHFO method, the passive DTS method does not require an external source of energy making this approach useful in remote areas where power supply is limited (Steele-Dunne et al., 2010).

Steele-Dunne et al. (2010) demonstrated the feasibility of using the thermal response to the diurnal temperature cycle of buried fiber optic cables for distributed measurements of soil thermal properties and soil moisture content (Steele-Dunne et al., 2010). They first determined the thermal diffusivity from the observed cable temperatures and then, the inferred the soil moisture from the thermal diffusivity. The obtained results demonstrate that passive soil DTS alone can yield information on surface soil moisture.

Although this approach is advantageous because the low power requirements are ideal for remote study locations, its application remains challenging under conditions where the thermal response to the diurnal temperature cycle is not large enough to allow accurate estimation of soil moisture content (e.g., under dense vegetative canopy, at depths beyond the top few centimeters of the soil column, cloudy days, or other surface energy flux limited systems) (Steele-Dunne et al., 2010).

As to the AHFO, the soil moisture can be calculated in a reliable way from two numerical methods: the cumulative temperature (T_{cum}) and the maximum temperature (T_{max}) (Sayde et al., 2010; Sayde et al., 2014).

Sayde et al. (Sayde et al., 2010) provides a small-scale laboratory experiment where they demonstrated the feasibility of an approach for estimating soil water content (θ) using DTS technology in which a heat pulse is introduced from the cable itself and the resulting temperature response is monitored (Sayde et al., 2010).

They installed the optical fiber DTS cables into a soil column, and used a relatively short (2 min) and strong (20 W/m) heat pulse to heat the DTS cables.

The obtained thermal response of the soil to the heat pulse in the form of cumulative temperature increase, T_{cum}, over a certain period of time, T_{cum} is calculetd as follow:

$$T_{cum} = \int_{t_s}^{t_e} \Delta T(t)dT$$

where ΔT is the temperature increment with respect to a reference temperature, t_s and t_e are the time start and the end of the heat pulse, respectively.

The cumulative temperature variation is a function of the soil thermal properties and they related it to soil moisture with an empirical calibration equation demonstrating that T_{cum} is a good indicator of soil water content.

The obtained results provided in (Sayde et al., 2010) show that soil moisture can be measured using an active DTS approach in the range 0.05–0.41 m^3 m¯3 with an accuracy of 0.046 m^3 m^{-3} (Sayde et al., 2010).

A different approach for distributed soil water content measurements was proposed by Striegl and Loheide (2012). Here the authors presented the development and field deployment of sensing (DθS) system, where an optical fibre cable is buried into the soil at the constant depth of 20 cm and for 130 m length. In Fig. 3, is shown the field scale experiment.

They used longer (10 min) and lower power (3.07 W/m) heat pulses to estimate the soil moisture along a transect of a floodplain of the Upper East Branch Pecatonica River. Instead of using T_{cum}, the maximum temperature increase after heating (T_{max}) was used to estimate soil moisture (Striegl and Loheide, 2012).

This method allows to calculate T_{max} as an average of temperature data collected over a certain period of time once the temperature rise has plateaued:

$$T_{max} = \frac{1}{N}\sum_{t_e-\Delta t}^{t_e} \Delta T(t_i)$$

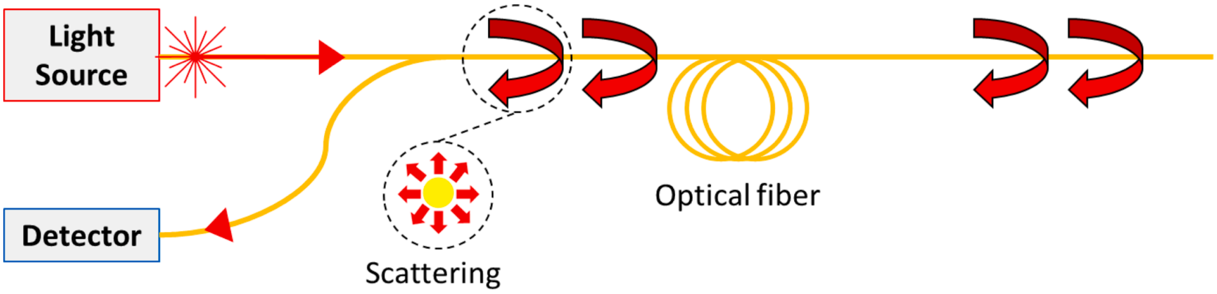


Fig. 1. Schematic of fiber optic distributed temperature sensor with light backscattering.

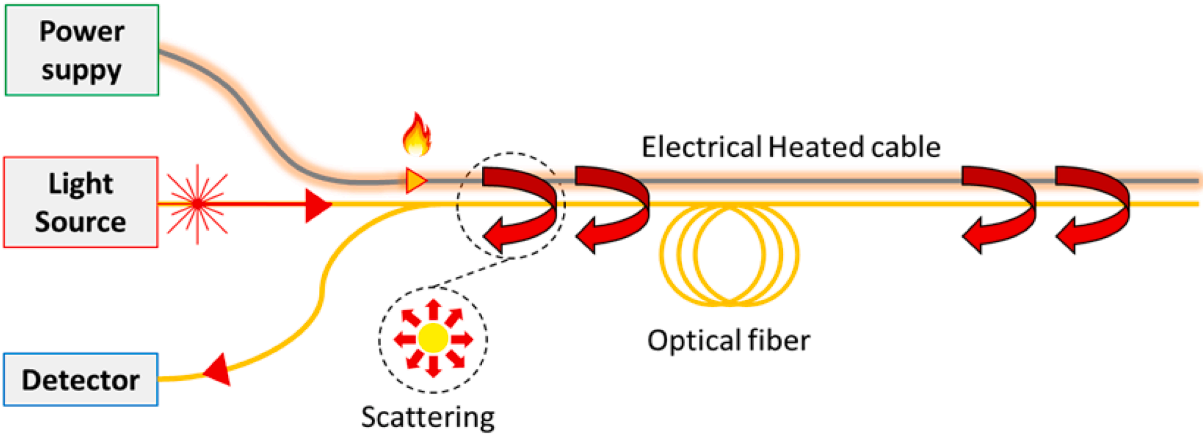


Fig. 2. Schematic of Heated fiber optic DTS with light backscattering.

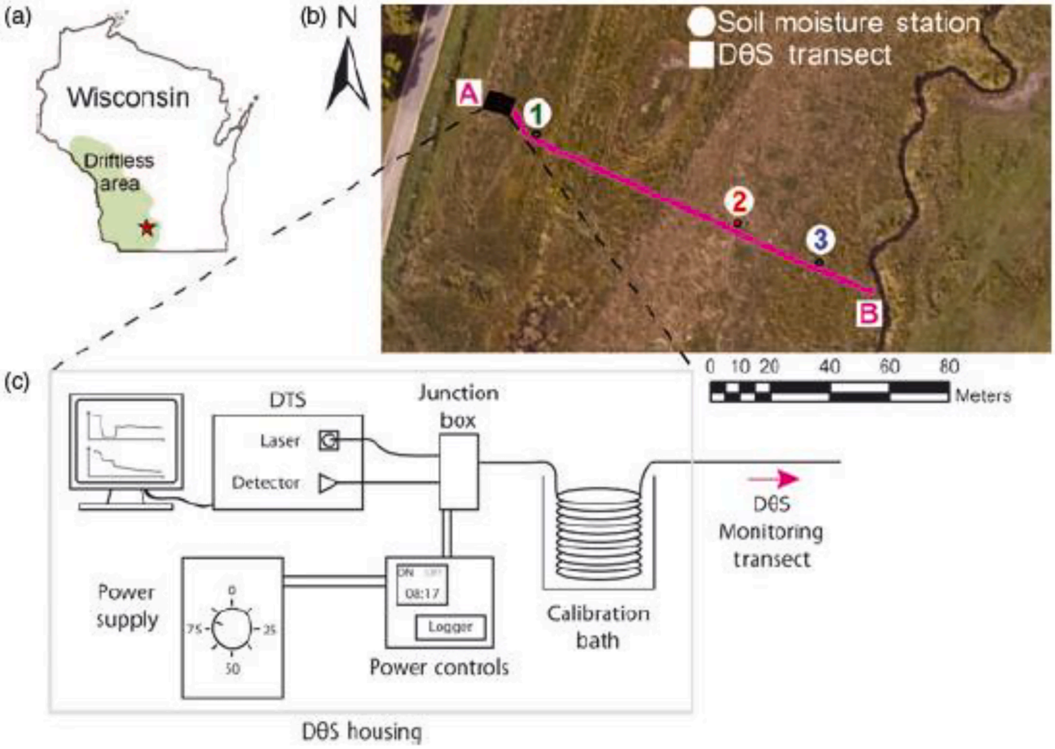


Fig. 3. (a) Study site located in the Driftless Area of southwest Wisconsin within the floodplain of the Upper East Branch Pecatonica River. (b) Aerial photograph shows the DθS monitoring transect (AB) spanning three distinct hydrologic regimes. Stations 1 to 3 are continuously logging dielectric soil moisture sensors. (c) Schematic illustrates the equipment located within the environmentally controlled DθS housing used to monitor θ at a 20 cm depth along the DθS monitoring transect. Copyright Ground Water.

where Δt is the length of the period (s) used for calculating T_{max} and N is the number of the measurements within this period. The thermal response ΔT provided by the $D\theta S$ is considered as the difference between two measurements and is defined form the following equation:

$$\Delta T = T_{heat} - T_0$$

where T_{heat} is the temperature measurement due to induced heating after 8 min and T_0, is the ambient temperature considered as average of 5 min of measurements before starting heating process.

T_{max} was than related to the θ measurements performed by co-located commercial dielectric sensors. The resulting curve (T_{max}-θ) empirically relate ΔT measurements on the DθS cable with the dielectric ones.

The slope of the T_{max}-θ relationship (DθS response curve) decreased with increasing θ, further suggesting decreased sensitivity at higher soil moisture contents, which is consistent with findings from previous studies (Sayde et al., 2010).

Moreover, the relationship DθS estimated θ vs. observed θ presents two regions, the strong estimator region for $\theta \leq 0.31$ and the week estimator region for $\theta > 0.31$; the first one is characterized by a RMSE = 0.0016 and the other one RMSE = 0.050 (Xu et al., 2012).

Recently, Dong et al presented a study where the performances of the heat pulse analysis methods based on the cumulative temperature (T_{cum}) and the maximum temperature (T_{max}) were evaluated for different heating strategies (Dong et al., 2017). In particular, the impact of heating strategy, i.e., difference in the duration and the input power of the applied heat pulses, on the performance of *Tcum* and *Tmax* was studied.

To do this, they analysed the two heat pulse processing methods using three different heating strategies across the full range of moisture values of sand.

As for the heating strategies used, they can be summarized as follow: the first heat pulse applied (H5) used a power input of 9.2 W/m with a duration of 5 min. This corresponds to the highest possible input energy without overheating the exposed DTS cables. In the second case (L5), the power was then lowered by half and the same heating duration (5 min) was used. Finally, a low power input (4.6 W/m) was combined with a longer duration (10 min), denoted as L10. A comparison of the characteristics of the three heating strategies is shown in Table 1.

The AHFO estimated soil moisture were validated against the dielectric soil moisture measurements provided by Decagon EC5 sensors using two metrics. The first is the root mean squared difference (RMSD) given by:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(\theta_{est,i} - \theta_{obs,i}\right)^2}$$

where θ_{est} and θ_{obs} represent the AHFO estimated and the EC5 observed soil moisture, respectively, and n is the number of the data points. Note that the AHFO method is also calibrated using the EC5

sensors.

A second metric used for validating the AHFO method is the coefficient of determination (R^2).

It represents the goodness of fit of the data to the regression curve:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the measured value and f_i is the corresponding estimated value.

A comparison of the R^2 for Heat Pulse Analysis Methods is reported in Table 2:

Results show that increased power input improves the accuracy of all two AHFO methods. If the available power input is limited, one of the temperatures based (i.e., *Tcum* and *Tmax*) methods should be employed instead. For the *Tcum* method, increasing the heat pulse duration can produce the similar improvement in sensitivity as an increase in power (a RMSD comparison for T_{cum} and *Tmax* VS Observed soil water content is reported in Table 3).

Finally, Gamage et al (2018) presented a feasibility study to measure soil water content at field scale using an in-situ calibration approach by means of AHFO technique (Vidana Gamage et al., 2018). They installed three fiber optic cables (147 m long) at three depths (0.05, 0.10, 0.20); each cable was connected to an electrical cable used for heating. As for the heating strategies, they used a heat pulse about 5 min and 7.28 W/m (q) to heat the DTS cables while for the data analysis they used the cumulative temperature (T_{cum}) method. T_{cum} was normalized by power intensity as $T_{cum\_N} = T_{cum}/q$ and using the SWC (provided by dielectric sensors Decagon 5TE) they developed a depth specific calibration relationships $T_{cum\_N} - \theta$.

The AHFO estimated SWC was validated using the 5TE SWC data. Root mean square error (RMSE) and coefficient of determination (R^2) were calculated to obtain the averaged predictive accuracy and goodness of fitness, respectively.

Results for specific depth calibration are summarized in the Table 4.

Overall, this study confirmed a great potential of the AHFO technique to measure soil water at high spatial resolutions (<1 m) and to monitor soil water of surface soil in a crop grown field over a cropping season with a satisfactory accuracy when compared with commercial soil water sensors (Whalley et al., 2013).

Table 1
The strength and the duration of the pulse in three heating strategies.

Strategy	L5	L10	H5
Strength (W/m)	4.6	4.6	9.2
Duration (min)	5	10	5

Table 2
R^2 comparison for T_{cum} and T_{max} VS Observed soil water content.

Strategy	T_{cum} (R^2)	T_{max} (R^2)
L5	0.66	0.66
L10	0.86	0.82
H5	0.95	0.96

Table 3
RMSD comparison for Tcum and Tmax VS Observed soil water content.

Strategy	T_{cum} (RMSD) [m^3/m^3]	T_{max} (RMSD) [m^3/m^3]
L5	0.061	0.059
L10	0.040	0.042
H5	0.020	0.018

Table 4
Comparison between observed soil water content estimated by 5TE soil moisture sensors and predicted soil water calculated by AHFO.

Depth [m]	RMSD [m^3/m^3]	R^2
0.05	0.028	0.87
0.10	0.037	0.46
0.20	0.037	0.86
Single calibration including all three depths	0.033	0.66

3.2. Fiber Bragg Gratings-Based sensors

Recently, the fiber Bragg grating (FBG) sensor has been widely used for monitoring of several parameters as temperature, strain, displacement, pressure, pH value, humidity, high magnetic field and acceleration, due to the advantages of small size, anti-electromagnetic interference, corrosion resistance and high sensitivity, for the possibility of multiplexing a large number of different sensors into the same optical fiber, reducing the multiple cabling used in traditional electronic sensing (Berruti et al., 2014).

Though the FBG based sensors represent an innovative solution for distributed monitoring, only a few studies are available in which the FBGs were applied in the agriculture field; in this section will be presented some attempts to use FBGs as soil water content sensors.

In 2002, Laylor et al. developed a first soil moisture sensor based on optical fiber Bragg Grating technology; the realized probe is made of two FBGs multiplexed on a single fiber (Laylor et al., 2002). One of the FBGs responds to temperature only and the other one responds to both temperature and relative humidity. They are installed in a stainless steel tube that includes a hydrophilic medium encased in a water permeable housing. The obtained results show that the limit of the moisture sensing appears to be between 2% and 4% of soil moisture (Laylor et al., 2002).

Since 2014, our research group was involved in the development of a first prototype of FBGs-based soil moisture and temperature sensor for irrigation optimization. We proposed the integration of a FBG-based thermo-hygrometer with a polymer micro-porous membrane for monitoring temperature and water content to perform measurements directly in soil. The thermo-hygrometer was realized using two FBGs, which are installed on the same fiber optic; one FBG is coated with a sensitive polyamide overlay for relative humidity (RH) measurement and the other one is bare for temperature measurement (and thermal compensation) ((Kronenberg et al., 2002; Yeo et al., 2005; Berruti et al., 2013)]. A schematic and the real prototype are shown in Fig. 4.

The micro-porous membrane is aimed at avoiding the direct contact of FBGs with water in the liquid phase, while allowing water in the

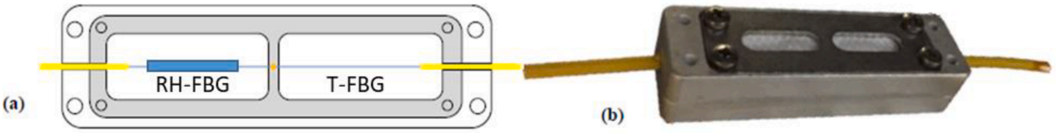


Fig. 4. (a) Sketch design of the protection package and thermos-hygrometer placing (b) real prototype.

gaseous phases to pass and interact with the FBG sensitive overlay. The FBG thermo-hygrometer was encapsulated in in a metallic package that is needed to guarantee a safety installation in soil. They experimentally demonstrated that the developed fiber optic thermo-hygrometer is able to provide soil water content continuous measurements (Leone et al., 2015).

In 2017, our group has experimentally demonstrated a novel solution for rainfall-induced landslide prevention applications consisting in an FBG-based soil VWC sensor, which overcomes the limitations of the first proposed solution due to the limited range of measurement and enables a continuous monitoring over large areas (Leone et al., 2017). Indeed, the optimized version of the probe was demonstrated to correctly monitor VWC values up to 37% when buried in the soil. The basic idea for enlarging the VWC measuring range of the soil moisture sensor relied on the integration of the above-mentioned thermo-hygrometers within a larger, customized functional package characterized by an increased exchange volume and thus able to promote a better distribution of the water molecules coming from the soil.

Nevertheless, due to the intrinsic thermal sensitivity of FBGs (~10 pm/°C), a long-term reliable temperature compensation is required in order to avoid errors in the correct VWC measurement.

Cao et al. in 2018 presented a different approach to estimate soil water content using an FBG-based Carbon Fiber Heated Sensor (CFHS).

The principle of CFHS is based on the correlation between the thermal responses of the heated sensor and the moisture content of the surrounding soil. When measurement starts, the CFHS is heated under a constant current, and the temperature variation of the CFHS is recorded by an FBG interrogator during heating. Based on a series of calibration tests, a temperature characteristic value (T_t) can then be used to calculate the soil moisture (θ) (Cao et al., 2015).

The CFHS is shown in Fig. 5; it consists of a carbon fiber rod with a diameter of 5 mm, FBG sensors multiplexed on the same optical fiber, a coating, two clips, a power supply, and two electric cables. The ends of the CFHS and electrodes of power supply are connected by electric cables. The carbon fiber rod is not only as a skeleton of the CFHS, but also conducts the electric current. Moreover, they applied a coating on the carbon rod in order to protect the optical fiber and to avoid power leakage.

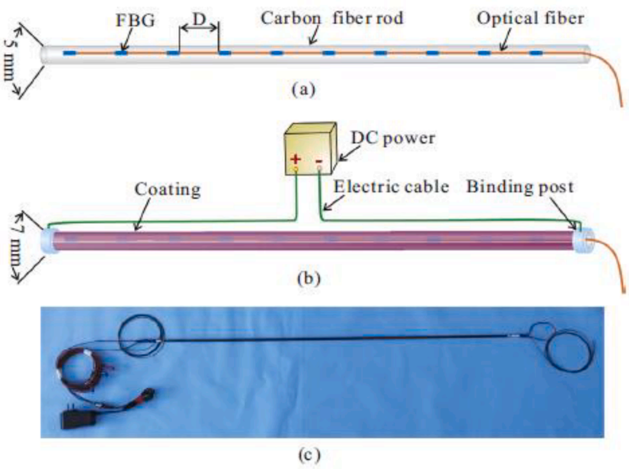


Fig. 5. Schematic of the CFHS, (a) internal structure of the CFHS, (b) external. Copyright Elsevier Engineering Geology.

They performed a characterization of the CFHS using different type of soil: Medium sand, Fine sand, Silt and clay. The calibrations test were performed filling a PVC tube with soil samples at different and predetermined VWC values (obtained with gravimetric methods) and burying the CFHS probe. The heating strategy consisted in a heat pulse of 5.4 W/m for 5 min.

They leaded a laboratory validation tests and the feasibility to perform soil water measurements in the range 0, 35% of VWC was demonstrated. However, the realized probe due to weak coating and feeble strength does not be applied to use at field scale.

For these reason, in 2018, Cao et al carried out some further improvements in order to make the probe able to quickly detect in situ soil moisture profiles of highways slopes and subgrades (Cao et al., 2018). In this work, they presented an FBG-based aluminium oxide tube packed sensor (ATPS) that overcome the limitation in strength of the above-mentioned CFHS for the in-situ moisture monitoring. Also, for the ATPS, the working principle is based on the relationship between thermal response of the characteristic temperature (T_t and soil moisture; once installed in soil, ATPS is heated by an electric cable in it and the heating strategy provides a power of 2.21 W/m for 5 min. The ATPS is reported in Fig. 6; as shown in figure, the new probe was realized by means of an aluminium oxide tube in which are inserted FBGs multiplexed on the same fiber optic cable and a U-shaped resistance wire. At the bottom of the tube a plastic cusp is fixed using epoxy glue. The more robust package was needed in order to perform in situ measurements.

They performed a laboratory characterization to obtain the calibration parameter and two laboratory experiments to verify the capability on a real field scale. Test results demonstrated that with the advantages

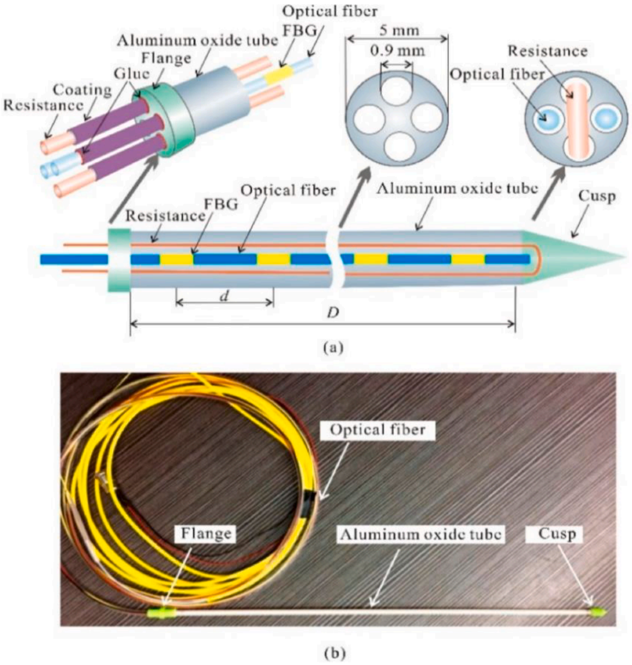


Fig. 6. Details of the ATPS: (a) frame of the ATPS, and (b) images of the ATPS. Copyright MDPI Sensors 2018.

in terms of strength and stiffness, the ATPS is able to perform fast monitoring of highway slope and soil moisture measurements; moreover, as for the estimation of soil water, ATPS shows batter performance with respect the CFHS in terms of R^2 and RMSE obtained in the calibration test, respectively 0.972 VS 0.902 and 0.018 m^3/m^3 VS 0.033 m^3/m^3. Nevertheless, the presented sensor is not suitable for monitoring over large-scale and does not allow cost-effective multiplexing due to the hardware complexity.

In 2017, Hallet et al. presented for the first time the use of an optical fibre long period grating (LPG) as soil water content sensor; they used the refractive index sensitivity of LPG to measure the soil moisture content (Hallett et al., 2017). The LPG is installed in a 3D printed package that is need to protect the fibre when the probe is buried into the soil. The package allows the soil water flow around the LPG.

They firstly demonstrated the sensitivity of the LPG when it is immersed in water and they registered 10 nm of wavelength shift in the transition air–water of the centre wavelength of the resonance bands.

Afterwards, the probe was tested in the range 10–40% VWC into two different soil (clay and sandy loam) and buried to a depth of 5 cm together with a conventional soil moisture sensor (Theta Probe SM200). In both soil types, the probe presented a correlation with the readings of the reference soil sensor even if the LPG wavelength variation results to be non-linear when compared with Theta Probe. Moreover, a time delay between the two sensors was registered (Hallett et al., 2017).

4. Near-Infrared reflectance technique for soil water content

Soil water content can be estimated using the near-infrared (NIR) reflectance technique by means of several absorption bands that soil moisture has in the NIR region (970, 1200, 1450 and 1970 nm) (Stenberg et al., 2010; Yoshio Kano and McClure, 2009). Moreover, such approach provides fast and non-destructive measurements and allows the detection of several soil attributes such as organic matter, minerals, pH and heavy metal content (Stenberg et al., 2010).

In 2006, Mouazen et al. demonstrated the possibility to measure soil water content by the numerical analysis of VIS-NIR soil spectra acquired by means of a mobile commercial spectrophotometer (Mouazen et al., 2006). A schematic is shown in Fig. 7:

Liang et al presented a similar concept where the prediction model of soil moisture content was developed to analyse the sensitive spectral band (wavelengths strongly affected by water absorption 1450, 1940) and insensitive spectral band (wavelength not affected by water absorption 1281). The over-mentioned approaches provide an estimation of surface soil moisture but they are not well suited for a direct in soil and continuous monitoring of soil water content (Liang et al., 2012).

In 2012, Yin et al presented a near-infrared (NIR) based optical sensor designed to perform in situ measurements of the reflectance of soil surface (Yin et al., 2021). The reflectance sensor used for soil surface moisture measurement is based on the reflectance of two light-emitting diodes (LEDs) for generating NIR light at two different wavelengths, in which one has a wavelength of 1940 nm and a strong water absorption band, whereas the other has a wavelength of 1800 nm and a weak water absorption band, as related to soil moisture reflectance. Moreover, it uses photodiodes for receiving reflected light signals. A sensor schematic is shown in Fig. 8:

They investigate the relationship between soil moisture and surface reflectance by studying four different soils; the soil samples water content covered a wide range from completely dry to saturation conditions. The results indicate a strong linear correlation between soil moisture and relative absorption depths for the different soils tested, and the reflectance models are dependent on soil type (Yin et al., 2013). The relative absorption depth (R) is used to estimate the soil water content is provided by the following equation [A near infrared reflectance soil moisture meter]:

$$R = 1 - \frac{R_b}{R_c}$$

where R is the relative absorption depth, R_b is the reflection intensity at 1940 nm and R_c is the reflection intensity at 1800 nm. The soil moisture is than correlated with the relative absorption depth using a linear equation. As results, a strong linear correlation has been found between soil volumetric water content and the relative absorption depth for all soils ($R^2 = 0.642$).

The results of this work show the feasibility to evaluate the surface soil moisture by measuring the soil reflectance but the proposed sensor is not properly suited for in filed agricultural applications where in-depth soil water estimation are required.

In 2015, Yin et al. proposed a small size, portable and low power consumption system as near-infrared surface soil moisture sensor. In this work, they used a quadratic function to describes the relationship between relative reflectance and moisture content obtaining a high correlation efficient ($R^2 = 0.941$) and low root mean square error (*RMSE*) (Yin et al., 2015).

Recently, Chen et al. presented an optical-based solution for a direct and continuous soil water monitoring in combination with a nanoporous ceramic plate (Chen et al., 2019). In particular, a miniaturized sensor consisting in a thin water reservoir included between a nanoporous ceramic disc (Al_2O_3) and a silicon diaphragm in combination with a miniature optical displacement detection unit is used to provide a long-term continuous measurement of soil water potential (Fig. 9).

When the sensor is buried into unsaturated soil water, a negative pressure inside the reservoir is established, thus inducing a diaphragm bending into the water reservoir itself. The displacement of the

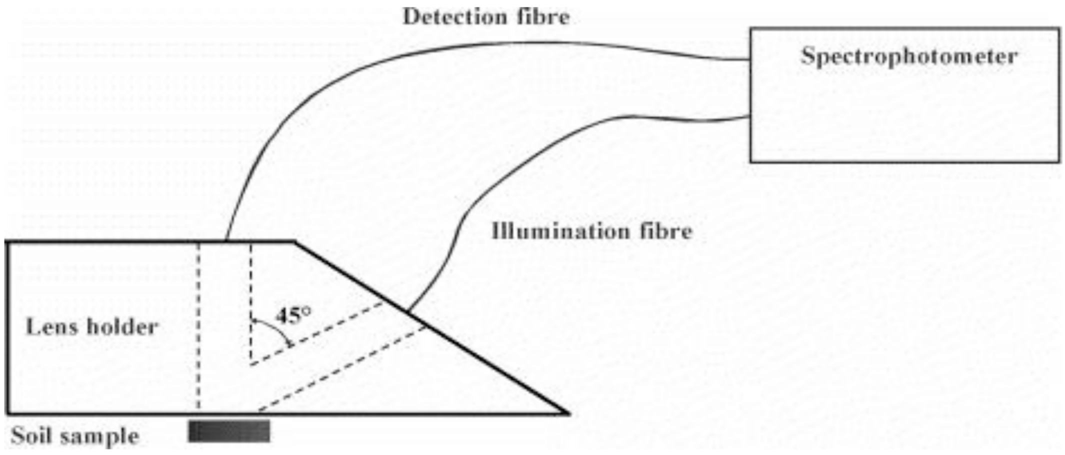


Fig. 7. Schematic illustration of experimental set up (from Mouanzen et al., 2005c).

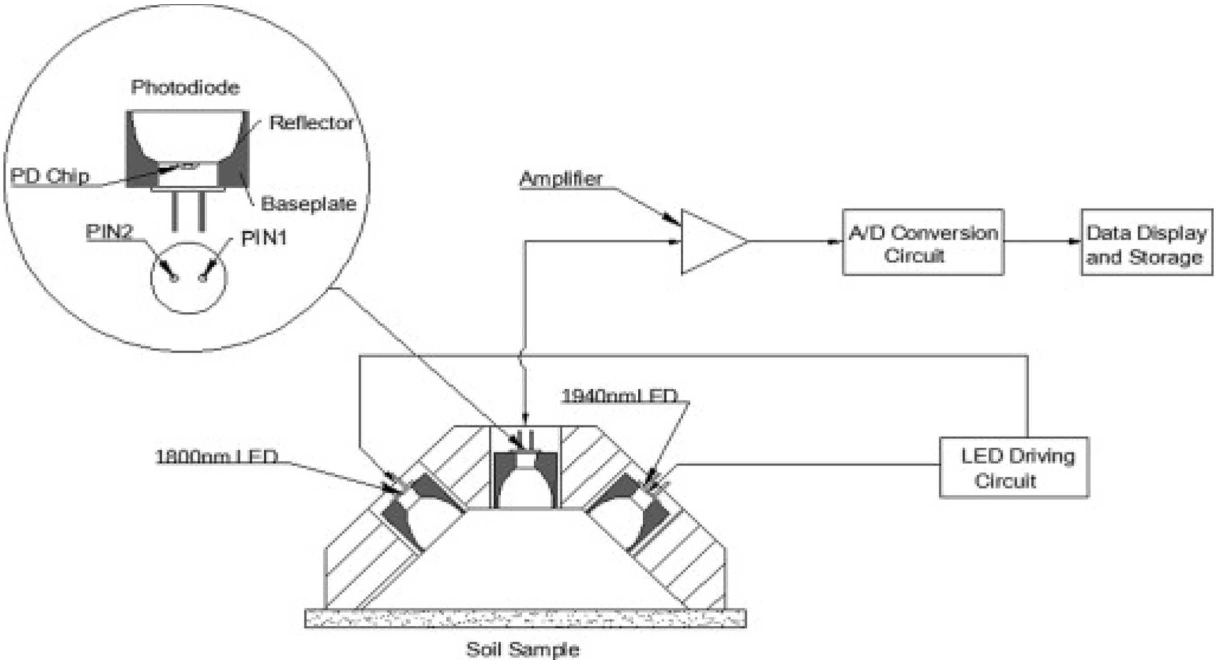


Fig. 8. NIR sensor case schematic. Copyright Elevier Computers and Electronics in Agricolture, 2013.

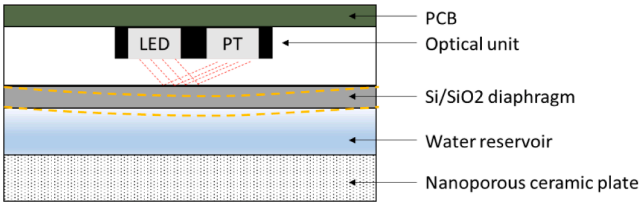


Fig. 9. Cross-sectional schematic representation of the soil water potential sensor.

diaphragm is measured by the optical displacement detector that is composed of integrated light source and photodetector.

Nevertheless, the presented sensor is not suitable for continuous monitoring over large-scale as it requires a periodic refilling of the water reservoir. Moreover, the device does not allow cost-effective multiplexing due to the hardware complexity.

In 2021, our research group presented a compact innovative optical and low-cost platform for soil water content measurement based on a properly combination of the optical property's variations of a nanoporous ceramic disc in connection with an engineered optical fiber with NIR-based detection techniques (Leone et al., 2022). It consists of a bifurcated cable Y-shaped that includes directly a couple of fibers in a single body. The two fibers are placed side-by-side in the common end, converging in a SMA905 connector, and breaks out into two legs, one for the connection to the light source and one for the detector. The common leg has to illuminates and collects light from the disc, as shown in Fig. 10 (a). Moreover, an appropriate structure made of some optical commercial components was used to couple the optical fiber with the nanoporous ceramic disc (Fig. 10 (b)). For a soil test, the prototype was placed in a protection PVC tube buried in a soil tank, as shown in Fig. 10 (c).

The obtained results show the capability of such prototype to measure soil VWC values up to 35%, with a sensitivity of about 2.3%/%VWC and resolution lower than 1%. The proposed platform is extremely flexible and suitable for continuous monitoring over large areas thus allowing to build extended sensors-base grid due to the optical fibers multiplexing capability.

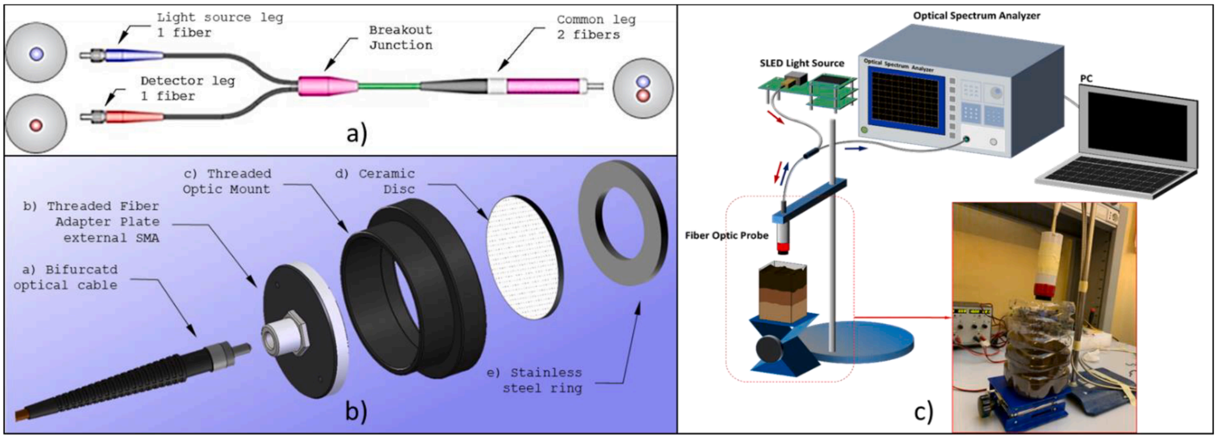


Fig. 10. (a) Design of bifurcated fibers; (b) Optical kit and integration sequence; (c) Experimental setup and Protection PVC package for validation in soil.(from Leone et al., 2022).

5. Conclusions

Climate change and the limited availability of water resources represent the most important challenges that technology must face providing efficient soil moisture monitoring solutions in both agriculture and hydrology fields.

The direction to follow within the development of innovative systems for smart agriculture and hydrological risk monitoring should be oriented towards the integration of different technologies. For example, an integrated solution could include soil sensors and aerial sensors, for visual inspections, supported by reliable Artificial Intelligence and Machine Learning algorithms, in order to provide decision-making support both for the development of customized irrigation strategies in the agricultural field and for the monitoring and early warning of hydrological risk situations. Moreover, the integration capabilities of such sensors in both operational context (e.g. installation in field) and the interconnection with smart systems have to be improved for ensuring real-time automated monitoring especially over large areas. Optical fiber sensors, in this sense, represent a technological platform able to perform soil moisture monitoring over a large-scale with extreme integration capabilities thanks to their small dimensions and reduced energy consumption.

In this paper, a review of the major advances in the development of optical fiber based devices for soil water content monitoring was presented. The working principle and the performances of such platforms were reported in order to provide a complete scenario of soil moisture sensors for smart farming applications and hydrological monitoring proposed in literature.

The emerging optical fiber based technology for the soil water content monitoring is potentially effective to overcome the limitations of the commercial conventional sensors. Indeed, the use of fiber optic sensors allows to implement non-invasive solutions for large-scale monitoring by exploiting the possibility of multiplexing, thus making them particularly suitable for applications in the agri-food field.

The state-of-art analysis reported in the present work shows that the majority of fiber optic devices for soil water content monitoring proposed in literature is still in the prototype phase; only some of them, such as the AHFO, CFHS and FBG-based devices, have been demonstrated for continuous and large scale real site test.

As for the reflectance techniques based on the absorption bands of the soil moisture in the NIR region, they provide fast and non-destructive measurements but the water estimation is limited to the soil surface, thus making them not well suited for a direct in soil and continuous monitoring for agricultural uses. On the contrary, the solution combining the ceramic disk with the optical fiber bifurcated probe based on NIR detection overcomes the over-mentioned limitations as it was found to be suitable for burial at any depth providing a continuous monitoring.

References

Berruti, G., Consales, M., Giordano, M., Sansone, L., Petagna, P., Buontempo, S., Breglio, G., Cusano, A., 2013. Radiation hard humidity sensors for high energy physics applications using polyimide-coated fiber Bragg gratings sensors. Sens. Actuators, B 177, 94–102. https://doi.org/10.1016/j.snb.2012.10.047.

Alessi, R.S., Lyle, P., 1986. Soil-water determination using fiber optics. Soil Sci. Soc. Am. J. 50 (4), 860–863.

Arulanandan, K., 1991. Dielectric method for prediction of porosity of saturated soil. J. Geotech. Eng. 117 (2), 319–330.

Berruti, G., Consales, M., Borriello, A., Giordano, M., Buontempo, S., Makovec, A., Breglio, G., Petagna, P., Cusano, A., 2014. A comparative study of radiation-tolerant fiber optic sensors for relative humidity monitoring in high-radiation environments JPHOT.2014.2357433.

Cao, D., Shi, B., Zhu, H., Wei, G., Chen, S.E., Yan, J., 2015. A distributed measurement method for in-situ soil moisture content by using carbon-fiber heated cable. J. Rock Mech. Geotech. Eng. 7 (6), 700–707. https://doi.org/10.1016/j.jrmge.2015.08.003.

Cao, D., Fang, H., Wang, F., Zhu, H., Sun, M., 2018. A fiber bragg-grating-based miniature sensor for the fast detection of soil moisture profiles in highway slopes and subgrades. Sensors (Switzerland) 18 (12), Dec. https://doi.org/10.3390/s18124431.

Cardenas-Lailhacar, B., Dukes, M.D., 2010. Precision of soil moisture sensor irrigation controllers under field conditions. Agric. Water Manag. 97 (5), 666–672. https://doi.org/10.1016/j.agwat.2009.12.009.

Chen, Y., Tian, Y., Wang, X., Dong, L., 2019. Miniaturized Soil Sensor for Continuous, In-Situ Monitoring of Soil Water Potential, 2019 20th International Conference on Solid-State Sensors, Actuators and Microsystems and Eurosensors XXXIII, TRANSDUCERS 2019 and EUROSENSORS XXXIII, no. June, pp. 2025–2028, doi: 10.1109/TRANSDUCERS.2019.8808562.

Dasberg, S., 1985. Time domain reflectometry field measurements of soil water content and electrical conductivity. Soil Sci. Soc. Am. J. 49 (2), 293–297.

Dean, T.J., Bell, J.P., Baty, A.J.B., 1987. Soil moisture measurement by an improved capacitance technique, Part I. Sensor design and performance. J. Hydrol. 93 (1–2), 67–78. https://doi.org/10.1016/0022-1694(87)90194-6.

Dobriyal, P., Qureshi, A., Badola, R., Hussain, S.A., 2012. A review of the methods available for estimating soil moisture and its implications for water resource management. J. Hydrol. 458-459, 110–117.

Dong, J., Agliata, R., Steele-Dunne, S., Hoes, O., Bogaard, T., Greco, R., van de Giesen, N., 2017. The impacts of heating strategy on soil moisture estimation using actively heated fiber optics. Sensors (Switzerland) 17 (9), 2102. https://doi.org/10.3390/s17092102.

Florides, G., Kalogirou, S., 2008. First in situ determination of the thermal performance of a U-pipe borehole heat exchanger, in Cyprus. Appl. Therm. Eng. 28 (2-3), 157–163.

Garrido, F., Ghodrati, M., Chendorain, M., 1999. Small-scale measurement of soil water content using a fiber optic sensor. Soil Sci. Soc. Am. J. 63 (6), 1505–1512.

Hallett, J.S. *et al.*, 2017. Soil moisture content measurement using optical fiber long period gratings, in 25th International Conference on Optical Fiber Sensors, vol. 10323, p. 103232J. doi: 10.1117/12.2263427.

Haroon, H., et al., 2018. Design and implementation of fibre optic sensor for soil moisture detection. J. Telecommun., Electr. Comp. Eng. 10 (2–5), 131–134.

Kayad, A., Paraforos, D.S., Marinello, F., Fountas, S., 2020. Latest advances in sensor applications in agriculture. Agriculture (Switzerland) 10 (8), 1–8. https://doi.org/10.3390/agriculture10080362.

Kirkham, M.B., 2014. Principles of Soil and Plant Water Relations. Academic Press.

Kronenberg, P., Rastogi, P.K., Giaccari, P., Limberger, H.G., 2002. Relative humidity sensor with optical fiber Bragg gratings. Opt. Lett. 27 (16), 1385. https://doi.org/10.1364/OL.27.001385.

Krzeminska, D.M., Steele-Dunne, S.C., Bogaard, T.A., Rutten, M.M., Sailhac, P., Geraud, Y., 2012. High-resolution temperature observations to monitor soil thermal properties as a proxy for soil moisture condition in clay-shale landslide. Hydrol. Process. 26 (14), 2143–2156. https://doi.org/10.1002/hyp.7980.

Laylor, M., Calvert, S., Taylor, T., Schulz, W., Lumsden, R., Udd, E., 2002. Fiber optic grating moisture and humidity sensors, 2002. [Online]. Available: http://www.blueroadresearch.com.

Leone, M., Consales, M., Passeggio, G., Buontempo, S., Zaraket, H., Youssef, A., Persiano, G.V., Cutolo, A., Cusano, A., 2022. Fiber optic soil water content sensor for precision farming. Opt. Laser Technol. 149, 107816. https://doi.org/10.1016/j.optlastec.2021.107816.

Leone, M., Consales, M., Laudati, A., Mennella, F., Cutolo, A., Cusano, A., 2015. Fiber optic thermo-hygrometers for soil moisture and temperature measurements: the SFORI project, 24th International Conference on Optical Fibre Sensors, vol. 9634, p. 96342P, doi: 10.1117/12.2194756.

Leone, M. *et al.*, 2017. Fiber optic thermo-hygrometers for soil moisture monitoring, Sensors (Switzerland), 17 (6), doi: 10.3390/s17061451.

Liang, X., Li, X., Lei, T., 2012. A new NIR technique for rapid determination of soil moisture content, 2012 International Conference on Systems and Informatics, ICSAI 2012, no. Icsai, pp. 16–20, 2012, doi: 10.1109/ICSAI.2012.6223513.

Mouazen, A.M., Karoui, R., De Baerdemaeker, J., Ramon, H., 2006. Characterization of soil water content using measured visible and near infrared spectra. Soil Sci. Soc. Am. J. 70 (4), 1295–1302. https://doi.org/10.2136/sssaj2005.0297.

Nadler, A., Dasberg, S., Lapid, I., 1991. Time domain reflectometry measurements of water content and electrical conductivity of layered soil columns. Soil Sci. Soc. Am. J. 55 (4), 938–943. https://doi.org/10.2136/sssaj1991.03615995005500040007x.

Olmanson, O.K., Ochsner, T.E., 2006. Comparing ambient temperature effects on heat pulse and time domain reflectometry soil water content measurements. Vadose Zone J. 5 (2), 751–756. https://doi.org/10.2136/vzj2005.0114.

Robinson, D.A., Campbell, C.S., Hopmans, J.W., Hornbuckle, B.K., Jones, S.B., Knight, R., Ogden, F., Selker, J., Wendroth, O., 2008. Soil moisture measurement for ecological and hydrological watershed-scale observatories: a review. Vadose Zone J. 7 (1), 358–389. https://doi.org/10.2136/vzj2007.0143.

Sayde, C., Gregory, C., Gil-Rodriguez, M., Tufillaro, N., Tyler, S., van de Giesen, N., English, M., Cuenca, R., Selker, J.S., 2010. Feasibility of soil moisture monitoring with heated fiber optics: rapid communication. Water Resour. Res. 46 (6) https://doi.org/10.1029/2009WR007846.

Sayde, C., Buelga, J.B., Rodriguez-Sinobas, L., El Khoury, L., English, M., van de Giesen, N., Selker, J.S., 2014. Mapping variability of soil water content and flux across 1–1000 m scales using the actively heated fiber optic method. Water Resour. Res. 50 (9), 7302–7317. https://doi.org/10.1002/2013WR014983.

Schmugge, T.J., Jackson, T.J., McKim, H.L., 1979. Survey of methods for soil moisture determination, 16 (6), 961–979.

Schwartz, R.C., Evett, S.R., Unger, P.W., 2003. Soil hydraulic properties of cropland compared with reestablished and native grassland. Geoderma 116 (1–2), 47–60. https://doi.org/10.1016/S0016-7061(03)00093-4.

Soil Moisture Sensor Market Size, Share & Trends Analysis Report By Sensors, By Connectivity (Wired, Wireless), By Application (Agriculture, Residential, Forestry, Research Studies), By Region, And Segment Forecasts, 2019 – 2025., Grand View Research, Nov 2019, https://www.grandviewresearch.com/industry-analysis/soil-moisture-sensors-market.

Sourbeer, J.J., Loheide, S.P., 2016. Obstacles to long-term soil moisture monitoring with heated distributed temperature sensingc. Hydrol. Process. 30 (7), 1017–1035. https://doi.org/10.1002/hyp.10615.

Steele-Dunne, S.C., Rutten, M.M., Krzeminska, D.M., Hausner, M., Tyler, S.W., Selker, J., Bogaard, T.A., van de Giesen, N.C., 2010. Feasibility of soil moisture estimation using passive distributed temperature sensing: passive soil DTS for soil moisture estimation. Water Resour. Res. 46 (3) https://doi.org/10.1029/2009WR008272.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Adv. Agron. 107 (C), 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.

Striegl, A.M., Loheide, S.P., 2012. Heated distributed temperature sensing for field scale soil moisture monitoring. Ground Water 50 (3), 340–347. https://doi.org/10.1111/j.1745-6584.2012.00928.x.

Susha Lekshmi, S.L., Singh, D.N., Shojaei Baghini, M., 2014. A critical review of soil moisture measurement. Measurement 54, 92–105.

Topp, G.C., Daniel, R.W., 1998. Time domain reflectometry: A seminal technique for measuring mass and energy in soil. Soil and Tillage Research 47 (1-2), 125–132.

Vidana Gamage, D.N., Biswas, A., Strachan, I.B., Adamchuk, V.I., 2018. Soil water measurement using actively heated fiber optics at field scale, Sensors (Switzerland), 18 (4), doi: 10.3390/s18041116.

Whalley, W.R., Ober, E.S., Jenkins, M., 2013. Measurement of the matric potential of soil water in the rhizosphere. J. Exp. Bot. 64 (13), 3951–3963. https://doi.org/10.1093/jxb/ert044.

Xu, J., Ma, X., Logsdon, S.D., Horton, R., 2012. Short, multineedle frequency domain reflectometry sensor suitable for measuring soil water content. Soil Sci. Soc. Am. J. 76 (6), 1929–1937. https://doi.org/10.2136/sssaj2011.0361.

Xu, J., Ma, X., Logsdon, S.D., Horton, R., 2012. Short, multineedle frequency domain reflectometry sensor suitable for measuring soil water content. Soil Sci. Soc. Am. J. 76 (6), 1929–1937. https://doi.org/10.2136/sssaj2011.0361.

Yeo, T.L., Sun, T., Grattan, K.T.V., Parry, D., Lade, R., Powell, B.D., 2005. Characterisation of a polymer-coated fibre Bragg grating sensor for relative humidity sensing. Sens. Actuators, B 110 (1), 148–156. https://doi.org/10.1016/j.snb.2005.01.033.

Yeo, T.L., Sun, T., Grattan, K.T.V., 2008. Fibre-optic sensor technologies for humidity and moisture measurement. Sens. Actuators, A 144 (2), 280–295.

Yin, H., Cao, Y., Marelli, B., Zeng, X., Mason, A.J., Cao, C., 2021. Soil sensors and plant wearables for smart and precision agriculture. Adv. Mater. 33 (20), 2007764. https://doi.org/10.1002/adma.202007764.

Yin, Z. *et al.*, 2015. Reflection Model for Soil Moisture Measurement Using Near-infrared Reflection Sensor, no. Ifeesm, pp. 854–860, doi: 10.2991/ifeesm-15.2015.158.

Yin, Z., Lei, T., Yan, Q., Chen, Z., Dong, Y., 2021. Licensed Content Publisher Elsevier Licensed Content Publication Computers and Electronics in Agriculture Licensed Content Title A near-infrared reflectance sensor for soil surface moisture measurement Licensed Content Author.

Yin, Z., Lei, T., Yan, Q., Chen, Z., Dong, Y., 2013. A near-infrared reflectance sensor for soil surface moisture measurement. Comput. Electron. Agric. 99, 101–107. https://doi.org/10.1016/j.compag.2013.08.029.

Yoshio Kano, R.W.S., McClure, W.F., A Near Infrared Reflectance Soil Moisture Meter, vol. 2, p. 2009, 2009.

High-resolution demodulation of interference envelope peak at arbitrary positions by warped discrete Fourier transform

Rudra Prasad Nanda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, rudraprasad858@gmail.com*

Smruti Ranjan Panda, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, sr_panda@outlook.com*

Swaha Pattnaik, *Department of Electronics and Communication Engineering , NM Institute of Engineering & Technology, Bhubaneswar, swaha.pattanayak@gmail.com*

Manoranjan Sahoo, *Department of Electronics and Communication Engineering , Capital Engineering College, Bhubaneswar, manoranjansahoo14@gmail.com*

ARTICLE INFO

Keywords:
Instrumentation, measurement, and metrology
Fourier optics and signal processing
Intreferometry
Metrology
Fringe analysis

ABSTRACT

We propose a novel approach to address the high-resolution problem for analyzing the peak position of an interference envelope in vertical-scanning wideband interferometry. Nonlinear envelope demodulation cannot be performed for arbitrary interference fringes because the envelope peak positions fluctuate depending on the acquisition statuses of the interference fringes. In the proposed method, first, the interference fringes are shifted until the desired results are obtained; this helps solve the discrepancy problem. Then, nonlinear envelope demodulation processing is applied to the selected shifted interference fringes. This paper describes the proposed procedure used in this study and its development using a shift selection scheme based on the envelope match characteristics of the shifted interference fringes. The proposed approach is introduced within the framework of super wideband light-based Michelson interferometry, and the experimental results demonstrate the superiority of the proposed algorithm over the conventional Fourier transform method for envelope peak resolution. The present technique is expected to be useful for the high-precision measurement of distances for not only scientific purposes but also industry requirements.

1. Introduction

An interferometer is an optical system used to analyze information obtained from interference fringes and produce the desired measurements (length/distance, surface shape, refractive index, etc.) (Steel, 1985; Langenbeck, 2014). Distance positioning using the envelope peak position of an interference fringe (Hariharan, 2003; Harding, 2013; de Groot and Leach, 2011) is a well-known method that is applied in full-field optical coherence microscopy (Dubois, 2016), surface metrology (Yoshizawa, 2017; Osten, 2018; Quinten and A, , 2020); and micro-manufacturing technologies (Qin, 2010; Fassi and Shipley, 2017), among others. Fourier transform techniques (Takeda et al., 1982; Pawłowski et al., 2006) are often used to demodulate the envelopes of fringes and are based on the theory of amplitude modulation/demodulation (Vijayachitra, 2013), which is well known in the communication field (Takeda, 1997). In this method, the Fourier transform is applied to the obtained interference fringes to express the deterministic signals and additive noise in the frequency domain. One of the main features of this method is that the signal can be separated from noise using a frequency-domain filter. The envelope can be reproduced from the fringe spectrum by removing the noise components. This method realizes envelope demodulation using the discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) pair. The DFT and IDFT are both linear processes; therefore, the envelope of the demodulated interference fringe has the same number of samples as the original interference fringe obtained from the interferometer, i.e., both have the same position resolutions.

Recently, many studies (Makur and Mitra, 2001; Marvasti, 2012; Jalali et al., 2014; Mahjoubfar et al., 2017; Phillips et al., 2017) have been inspired by the human visual and auditory systems, whose typical responses to input signals are nonlinear. To identify the envelope peak positions of interference fringes, we only need the data around these peaks because such locations have high interference fringe intensities. Hence, we need not scrutinize the envelopes in the low interference fringe intensity ranges. In other words, all envelopes need not be demodulated with the same resolutions at the peak locations. Once the range of the envelope peak position is located, we only need to demodulate the envelope within this range with higher density of sampling points. We note that the idea of peak detection in the frequency domain is also used for carrier estimation (Guo et al., 2007; Pandey et al., 2018) and the subsequent phase extraction (de Groot and Deck, 1995; Kemao, 2015) in optical fringe analysis. The proposed method could also find potential application in the frequency domain of fringe analysis.

Based on this idea, previous research (Wei et al., 2020) proposed a nonlinear envelope demodulation approach using the warped discrete Fourier transform (WDFT) (Makur and Mitra, 2001; Makur, 2008). This scenario can be understood as follows. Based on the symmetries of the DFT and IDFT, the IDFT–DFT process can be used to evaluate a signal envelope by time sampling at equal intervals. In other words, the signal frequency components at equidistant frequency sampling points obtained by IDFT processing are subjected to DFT processing to obtain the envelope of the time-domain signal sampled at equal intervals. The DFT can be used to evaluate the spectrum of a time signal sampled at equal intervals with equally spaced frequency sampling points. However, the WDFT can be used to evaluate the spectrum of a time signal sampled at equal intervals with non-equidistant frequency sampling points. Based on the analogical relationship between the DFT and WDFT, we replaced the DFT in the IDFT–DFT process with WDFT. The IDFT–WDFT process can, thus, be used to evaluate a signal envelope via non-equidistant time sampling of a time signal originally sampled at equal intervals. Specifically, the signal components obtained by IDFT at equally spaced frequency sampling points are subjected to WDFT to obtain the envelope of the time signal sampled at nonequal intervals. As noted above, the analogical relationship between DFT and WDFT as well as the symmetry of the Fourier transform between the time and frequency domains are required to understand the IDFT–WDFT method. The relationships between these processes are listed in Table 1.

There was a problem with the IDFT–WDFT approach for arbitrary interference fringes. For an N-point input time signal, the densest sampling arrangement range for the envelope position resolution obtained by IDFT–WDFT processing is near the N/2 and N/2 + 1 points. To achieve the adaptive high resolution of the envelope peak, we match the first and second highest points of the envelope (i.e., the first two envelope peak positions) with the two N/2 and N/2 + 1 points under the assumption that the interference fringe envelope peak exists between these points. Interference fringes with arbitrary peak positions do not satisfy this condition. In other words, to express the vicinity of the envelope peak position using IDFT–WDFT with the densest sampling of the envelope position resolution, fringe shifting is required. Therefore, we propose a fringe shifting method to satisfy the above requirements. We first shift the interference fringes and then measure its envelope peak position. Shifting is stopped when the envelope peak of the partially shifted interference fringes is at the center of the signal data length (i.e., N/2 and N/2 + 1 points). This means that we set the shifted envelope peak position to match the densest sampling range for position resolution. Thus, the discrepancy between the envelope peak position of the interference fringe and densest sampling range by IDFT–WDFT processing can be resolved.

2. Principle

To simplify the methodology, we use numerically calculated fringes to explain the proposed procedure. Fig. 1(a)–3(a) show interference fringes (black plus symbols) with different envelope peak positions obtained by numerical calculations. The data length of the calculated interference fringe data was N = 128 points. Fig. 1(c)–3(c) show the envelopes (blue circles) obtained by IDFT–WDFT of these interference fringes. For comparison, Fig. 1(b)–3(b) show the envelopes (blue circles) obtained by IDFT–DFT of the interference fringes. In this study, to enable easier visualization of the graphs, we have drawn curves (dashed or dotted lines) for the plot points. As shown in Fig. 2(c) and 3(c), the densest sampling range of the envelope for the IDFT–WDFT process is fixed at the center of the x axis. For these 128-point data signals, the densest sampling arrangement of the envelope position resolution for IDFT–WDFT is near points N/2 = 64 and N/2 + 1 = 65. However, the envelope peak position fluctuates depending on the acquisition status of the interference fringes. The problem with IDFT–WDFT processing for arbitrary interference fringes is that the densest sampling arrangement may not always match the envelope peak position. To achieve peak-adaptive high resolution for arbitrary interference fringes, we match the positions of the first and second strongest envelope strengths of the fringes to the densest sampling locations.

Table 1
Comparison of the different processing methods.

method	Input		output	
DFT	time domain	equal interval	frequency domain	equal interval
WDFT	time domain	equal interval	frequency domain	nonequal interval
IDFT–DFT	time domain	equal interval	time domain	equal interval
IDFT–WDFT	time domain	equal interval	time domain	nonequal interval

The difference between Fig. 2(a) and 3(a) is whether the peak value of the signal is on the right or left side of the x axis. When Fig. 3 is rotated with respect to the center of the x axis, the interference fringe obtained is the same as that in Fig. 2. Hence, the processing of the interference fringe in Fig. 2 is described below as an example.

This study proposes a method to shift the interference fringes before performing IDFT–WDFT and returning the envelope peak position obtained for the partially shifted interference fringes. We explain the method to match the envelope peak position with the densest sampling range by IDFT–WDFT using the numerically calculated interference fringes. Assume that the interference fringe to be processed is as shown in Fig. 4(a). DFT is performed first on this signal, as shown in Fig. 4(b). Filtering is then performed so that only the signal spectrum range is extracted, as shown in Fig. 4(c). Next, IDFT is performed on the spectrum in Fig. 4(c). Subsequently, two processes are then performed on the signal obtained after the IDFT. The first involves finding the absolute value of the IDFT signal and doubling it to obtain the envelope (blue circle), as shown in Fig. 4(d). The second involves finding the real part of the signal and doubling it to obtain the noise-removed interference fringes (red dashed line), as shown in Fig. 4(d).

Using the detected envelope, we obtain the positions of the points with the strongest envelope (position k) and the second strongest (at either k − 1 or k + 1). We assume that the position of the point with the second strongest envelope strength is k + 1. This means that our target matches k with N/2 (at the same time, k + 1 = N/2 + 1). The noise-removed fringes have the same lengths as the original fringes, i.e., N data points. We shift the interference fringes in unit time steps toward the center (i.e., left). This means that the first data point of the noise-removed interference fringe is the last data point of the shifted fringe. We then calculate the new envelope peak position using the same DFT and IDFT processes as those used for the shifted fringe. The continuation of the shifting process is decided by assessing the match between k and N/2.

We show an example of the interference fringe that satisfies the above conditions with a shift of 21 steps, as shown in Fig. 4(e). The 21-step shifted fringe is used for IDFT–WDFT processing. The filtered Fourier spectrum obtained by IDFT of this shifted fringe is shown in Fig. 4(f). The envelope (blue circles) of the filtered Fourier spectrum of the fringe shifted by DFT is shown in Fig. 5(a). We confirmed the match between the envelope peak position and center of the x axis. Fig. 5(b) shows the nonlinear processing results of the envelope expressed by the sampling points at non-equidistant intervals when applying the filtered Fourier spectrum (Fig. 4(f))) to WDFT. In Fig. 5(b), we confirm envelope peak adaptive high resolution using the IDFT–WDFT process. When the position of the point with the strongest envelope strength is k and that of the point with the second strongest envelope strength is k − 1, we need to find the match between k − 1 and N/2 (here, k = N/2 + 1).

The power spectrum of the signal can be calculated from the noise-

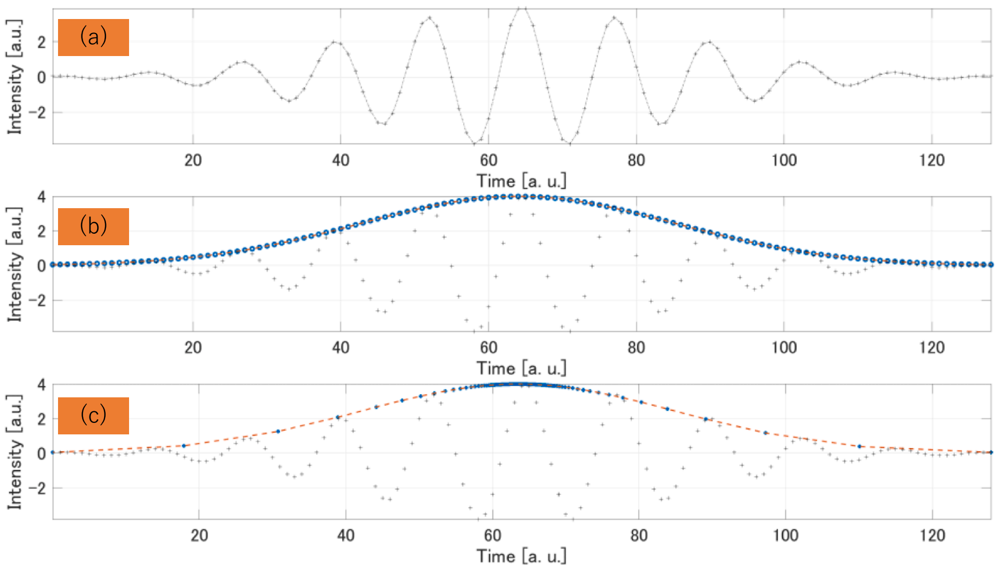


Fig. 1. (a) Numerically calculated interference fringe with the envelope peak at the center of the x axis; (b) envelope obtained by IDFT–DFT processing of the signal in (a); (c) envelope obtained by IDFT–WDFT processing of the signal in (a).

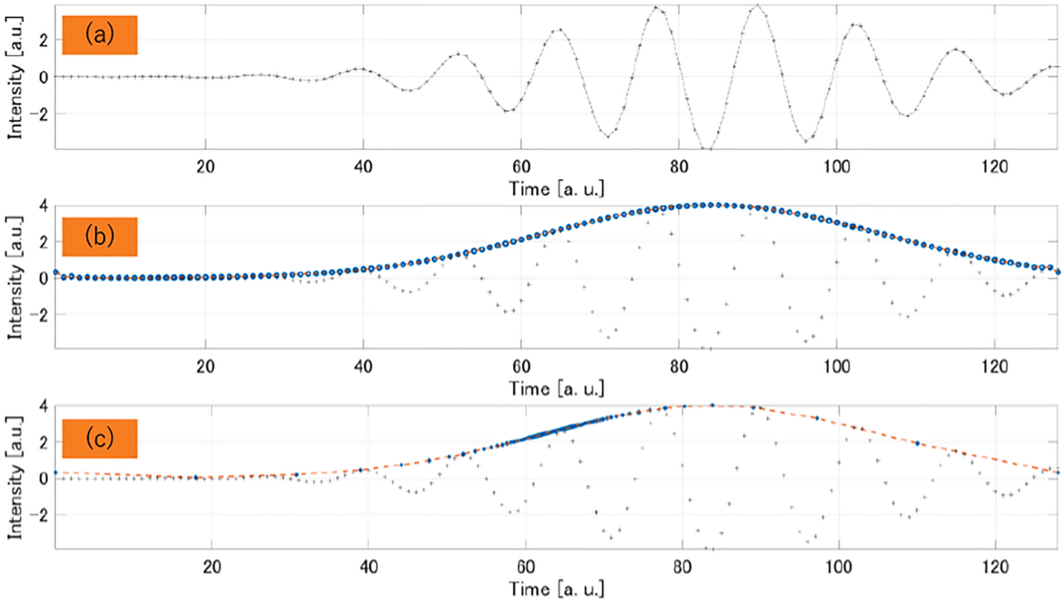


Fig. 2. (a) Numerically calculated interference fringe with the envelope peak to the right side of the x axis; (b) envelope obtained by IDFT–DFT processing of the signal in (a); (c) envelope obtained by IDFT–WDFT processing of the signal in (a).

free interference fringes. The noise power spectrum can be calculated from a noise-only signal. From both of these, the signal–noise ratio (SNR) can be calculated. By subtracting noise-free interference fringes from the interference fringes obtained by the DFT-IDFT process, a noise-only signal can be obtained. The power spectrum of noise can then be calculated. The power spectrum of the noise-free interference fringes can be used to calculate the signal-to-noise ratio after DFT–IDFT processing. In this study, as shown in Fig. 4(a), a noise-only signal was used.

As shown in Fig. 4(c), noise in the spectral region of the signal cannot be removed. The Fourier transform peak detection method is sensitive to noise in the spectral region of the signal.

3. Results and discussion

We conducted a super luminescent diode (SLD)-based vertical-scanning wideband interferometry experiment to demonstrate the application of the proposed method for high-resolution capability near the envelope peak in a fringe whose peak is not at the center of the x axis. The input light (ASLD-CWDM-3-B-FA, Amonics) is a super wideband source that provides high spectral density across the range of 1450–1650 nm. The fringes were recorded using a Michelson interferometer, which divides the light from the SLD laser into two beams traveling in different directions using a beam splitter placed in the optical path. Each beam is reflected by a mirror at the end of its optical path and returned to the original path to be recombined at the beam splitter. One of the mirrors was fixed on the optical table and served as a target mirror, while the other served as a reference mirror for scanning using a controlled scanning mechanism. The recorded interference fringes were processed by a digital oscilloscope and sent to a personal computer.

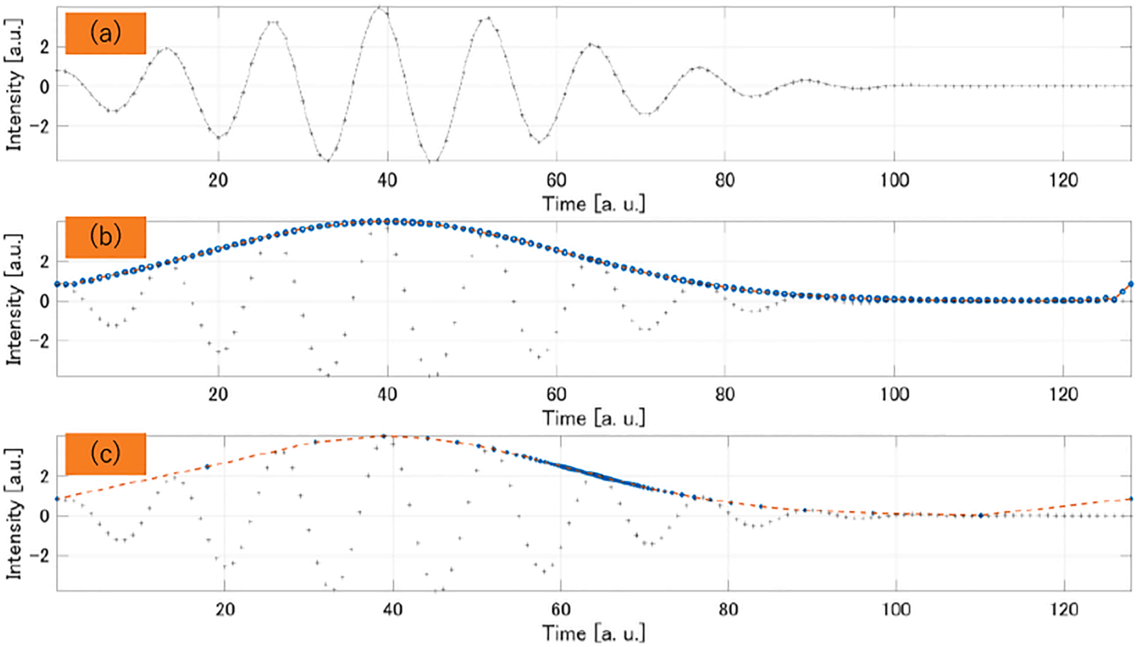


Fig. 3. (a) Numerically calculated interference fringe with the envelope peak to the left side of the x axis; (b) envelope obtained by IDFT–DFT processing of the signal in (a); (c) envelope obtained by IDFT–WDFT processing of the signal in (a).

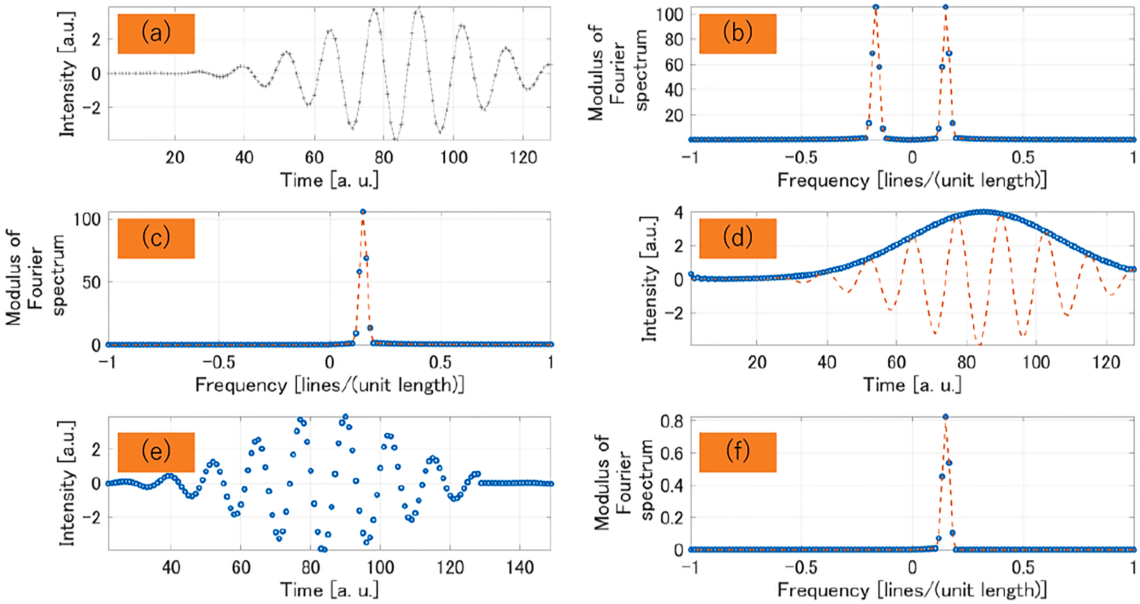


Fig. 4. (a) Numerically calculated interference fringe with the envelope peak on the left side of x axis; (b) Fourier spectrum obtained by DFT of the signal in (a); (c) filtered Fourier spectrum of (b); (d) envelope and noise-removed fringes obtained by DFT of (c), (e) fringe in (d) shifted by 21 steps; (f) filtered Fourier spectrum obtained by IDFT of (e).

In this experiment, the measurement target is a flat mirror. Furthermore, to suppress the influence of air fluctuations, a light shielding enclosure is used to cover the entire optical table. Therefore, the wavefront reflected by the mirror is not easily disturbed, and the interference fringes are less affected by speckle noise. If the interference fringes are affected by speckle noise due to the influence of air fluctuations, the interference fringes will be deformed. In this case, the envelope deviates from the form of the Gaussian function.

The proposed method was applied to multiple interference fringe signals recorded at different fixed positions of the object mirror. Because of constraints on the length of this paper, we only introduce one result here. The recorded fringes and results of peak position analysis using the proposed method are shown in Figs. 6–8. Shifting was performed using DFT–IDFT based match-selection scheme, as shown in Fig. 6(e). A sequence of noise-removed fringe spectra of the shifted fringe is shown in Fig. 6(f). The selected fringe spectrum was then converted to time domains using DFT and WDFT to produce the reconstructed envelopes of the interference fringes (blue circles) in Fig. 7(a) and (b). For comparison, the shifted interference fringe signal, which is the same as that shown in Fig. 6(e), is shown using black crosses.

This experiment was designed to demonstrate the principle. In this study, the warping control parameter *a* was set to 0.9. If *a* is less than 1 and greater than 0.5, the purpose of this study can be achieved. For example, if *a* is set to 0.99, a resolution of about 1/198 of that of the

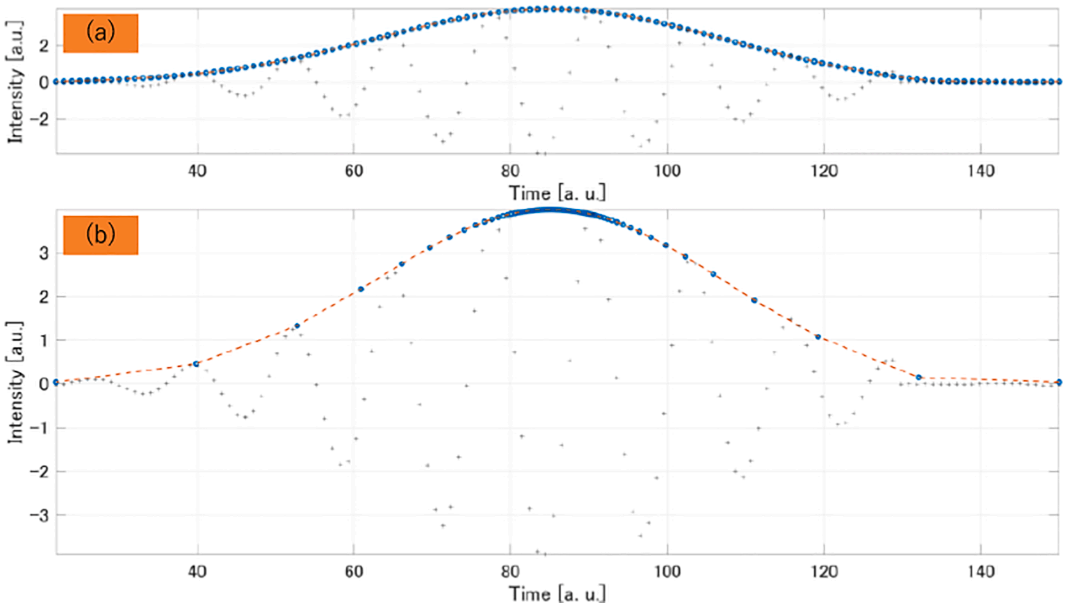


Fig. 5. Comparison of the initially set (black plus symbols) and calculated envelope (blue circles) using (a) IDFT–DFT and (b) IDFT–WDFT. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

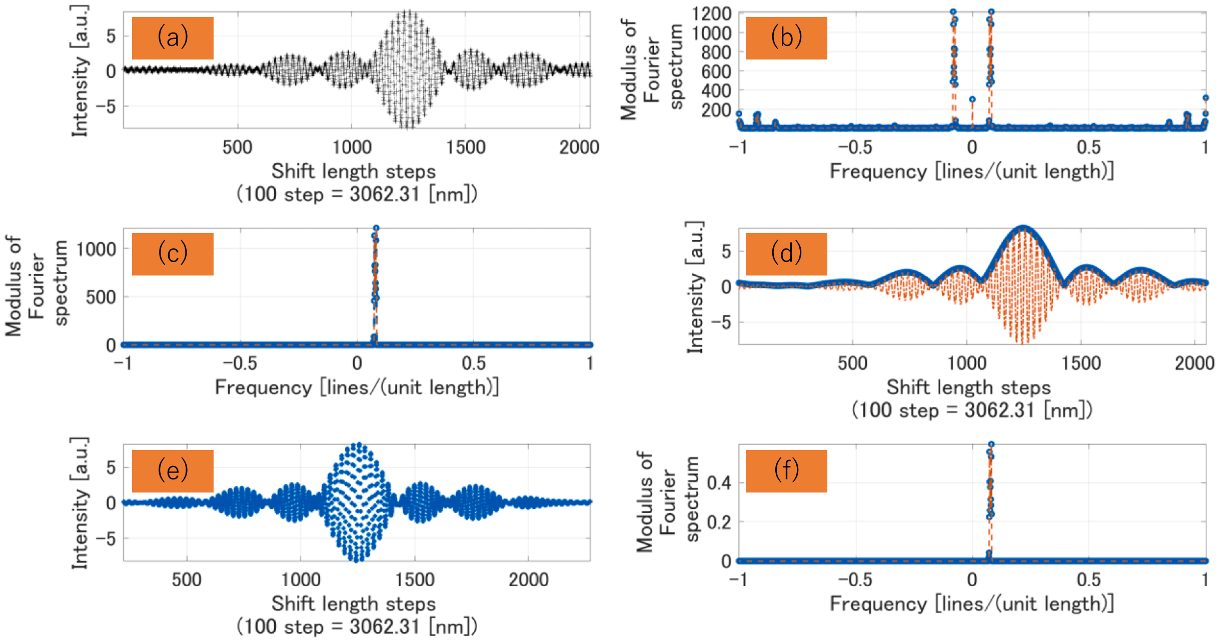


Fig. 6. (a) Interference fringe; (b) frequency spectrum components of (a); (c) filtered frequency spectrum components of (b); (d) noise-removed interference fringe (blue dashed line) and envelope obtained using the DFT–IDFT method (red solid line); (e) shifted fringe; (f) filtered frequency spectrum components obtained by IDFT of (e). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conventional method can be achieved, which was confirmed by experiments. The limiting factor for *a* seems to depend on the accuracy of the numerical calculation software. This point is currently under investigation.

Fig. 8 shows a magnified portion of the fringe having the maximum intensity. As the DFT–IDFT method is a linear method, the sampling points of the fringes and that of the envelope have a one-to-one relationship. The resolution of the IDFT–DFT method is about 30.62 nm (namely, one shift length step). By contrast, the IDFT–WDFT is a nonlinear method. The distances between adjacent envelope sampling data points are not in integer steps. As shown in the Fig. 8(b), the resolution of the IDFT–WDFT method is about 30.62/19 = 1.6 nm (namely, 1/19 of one shift length step). The position resolution around the envelope peak improved by a factor of almost 19 compared with that obtained using the linear conventional method.

Tables 2 and 3 list some selected experimental results for the fringe data used. Inspection of the data in the tables shows that a finer resolution of the peak position than that in the conventional method is obtained because the proposed method involves WDFT analysis. Note that the IDFT–WDFT process is a nonlinear demodulation; therefore, the sampling points obtained are different from those obtained by the conventional IDFT–DFT process. Therefore, the differences between the two data cannot be evaluated further.

The results of this investigation show that, the resolution of an

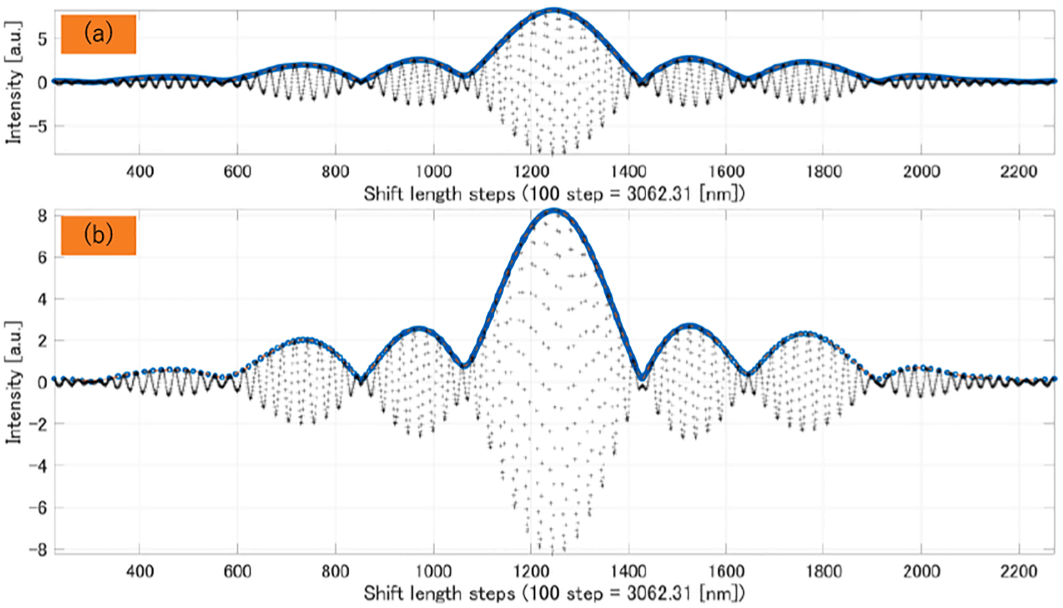


Fig. 7. (a) Obtained envelope (blue circles) of the filtered Fourier spectrum of shifted fringe using DFT and the shifted interference fringe used (black crosses); (b) envelope (blue circles) of the filtered Fourier spectrum of shifted fringe using WDFT and the shifted interference fringe used (black crosses). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

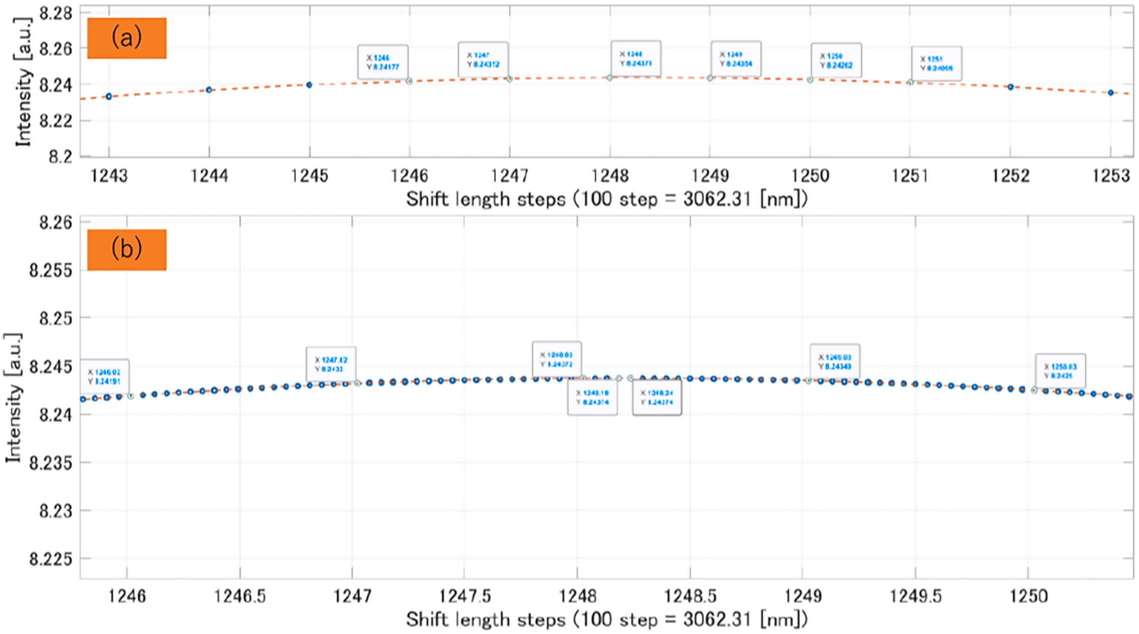


Fig. 8. Selected envelope ranges near the envelope peak positions using (a) IDFT–DFT and (b) IDFT–WDFT.

interferometry is also strongly dependent on the demodulation technology used. With an appropriate demodulation method, arbitrary interference fringes can be demodulated with a high-resolution envelope for position measurements. Finally, it is anticipated that the present technique will be a powerful metrological tool not only for surface profilometry and tomography but also for super-resolution metrology.

Table 2
Selected data values near the envelope peak position using the IDFT–DFT processing method.

X [steps]	1246	1247	1248	1249	1250
Intensity [a. u.]	8.24177	8.24312	8.24371	8.24354	8.24262

Table 3
Selected data values near the envelope peak position using the IDFT–WDFT processing method.

X [steps]	1246.02	1247.02	1248.03	1248.18	1248.24	1249.03	1250.03
Intensity [a. u.]	8.24191	8.2432	8.24372	8.24374	8.24374	8.24349	8.2425

4. Conclusions

This paper presents an envelope peak-adaptive high-resolution analysis method, where WDFT is used to analyze the peak position of an interference envelope produced by an SLD-based vertical-scanning wideband interferometer. When performing IDFT–WDFT for arbitrary interference fringes, the envelope peak positions fluctuate depending on the signal acquisition status. We shifted these interference fringes and then measured the envelope peak positions of the shifted fringes; the shifting process is stopped when the envelope peak position of the shifted signal matches the peak of the densest sampling range for position resolution using the IDFT–WDFT method. Moreover, experiments were performed to confirm that the proposed match-selection scheme yields robust results; the proposed method is effective at providing higher resolution estimates of the envelope peak positions. However, the accuracy of the numerical calculation software may be a limiting factor. This issue is currently under investigation. Therefore, our future research will focus on the optimization of the shifting process for the stepwise shifting method used in this study. We also plan to expand the vertical resolution of conventional white-light interferometry, which can be applied to optical tomography and profilometry.

References

Steel, W.H., 1985. Interferometry. Cambridge University Press.

Langenbeck, P., 2014. Interferometry for Precision Measurement. SPIE.

Hariharan, P., 2003. Optical Interferometry. Elsevier, pp. 1–8. https://doi.org/10.1016/B978-012311630-7/50002-2.

Harding, K., 2013. Handbook of Optical Dimensional Metrology. Taylor & Francis.

de Groot, P., 2011. Coherence scanning interferometry. In: Leach, R. (Ed.), Optical Measurement of Surface Topography. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 187–208.

Dubois, A., 2016. Handbook of Full-field Optical Coherence Microscopy: Technology and Applications. Jenny Stanford Publishing.

Yoshizawa, T., 2017. Handbook of Optical Metrology: Principles and Applications, second ed. CRC Press.

Osten, W., 2018. Optical Inspection of Microsystems. CRC Press.

Quinten, M., 2020. A Practical Guide to Surface Metrology. Springer International Publishing.

Qin, Y., 2010. Micromanufacturing Engineering and Technology. Elsevier Science.

Fassi, I., Shipley, D., 2017. Micro-Manufacturing Technologies and Their Applications: A Theoretical and Practical Guide. Springer International Publishing.

Takeda, M., Ina, H., Kobayashi, S., 1982. Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. J. Opt. Soc. Am. 72, 156–160.

Pawłowski, M.E., Sakano, Y., Miyamoto, Y., Takeda, M., 2006. Phase-crossing algorithm for white-light fringes analysis. Optics Communications 260, 68–72.

Vijayachitra, S., 2013. Communication Engineering. McGraw Hill Education (India).

Takeda, M., 1997. The philosophy of fringes: Analogies and dualities in fringe generation and analysis. In: Fringe '97 Automatic Processing of Fringe Patterns, Akademie Verlag Series in Optical Metrology (Akademie Verlag), pp. 17–26.

Makur, A., Mitra, S.K., 2001. Warped discrete-Fourier transform: Theory and applications. IEEE Trans. Circuits Syst. I Fund. Theory Appl. 48, 1086–1093.

Marvasti, F., 2012. Nonuniform Sampling: Theory and Practice. Springer, US.

Jalali, B., Chan, J., Asghari, M.H., 2014. Time-bandwidth engineering. Optica 1, 23–31.

Mahjoubfar, A., Churkin, D.V., Barland, S., Broderick, N., Turitsyn, S.K., Jalali, B., 2017. Time stretch and its applications. Nature Photonics 11 (6), 341–351.

Phillips, D.B., Sun, M.-J., Taylor, J.M., Edgar, M.P., Barnett, S.M., Gibson, G.M., Padgett, M.J., 2017. Adaptive foveated single-pixel imaging with dynamic supersampling. Science Advances 3 (4). https://doi.org/10.1126/sciadv.1601782.

Guo, H., Yang, Q., Chen, M., 2007. Local frequency estimation for the fringe pattern with a spatial carrier: Principle and applications. Appl. Optics 46, 1057–1065.

Pandey, N., Singh, M., Ghosh, A., Khare, K., 2018. Optical surface measurement using accurate carrier estimation in Fourier transform fringe analysis and phase unwrapping based upon transport of intensity equation. J. Optics (India) 47, 389–395.

de Groot, P., Deck, L., 1995. Surface Profiling by Analysis of White-light Interferograms in the Spatial Frequency Domain. J. Modern Optics 42 (2), 389–401.

Kemao, Q., 2015. Applications of windowed Fourier fringe analysis in optical measurement: A review. Opt. Las. Eng. 66, 67–73.

Wei, D., Nagata, Y., Aketagawa, M., 2020. Envelope peak-adaptive resolution enhancement of interference fringes using a nonlinear frequency-to-time mapping. Optics Commun. 472, 125870. https://doi.org/10.1016/j.optcom.2020.125870.

Makur, A., 2008. Computational Schemes for Warped DFT and Its Inverse. IEEE Trans. Circuits Syst. I: Regul. Papers 55 (9), 2686–2695.

GravCPA: Controller Placement Algorithm Based on Traffic Gravitation in SDN

Durga Prasanna Mohanty, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, dp_mohanty@yahoo.co.in*

Aruna Rout, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, aruna.rout@hotmail.com*

Snigadha Sagarika, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, snigadha216@gmail.com*

Prasanta Kumar Sahoo, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering &Technology,Bhubaneswar,prasantakumarsahoo@outlook.com*

Abstract

Software-defined network separates the control plane and the data plane, making the network more flexible. With the expansion of the network scale, one centralized controller cannot meet the latency needs of large-scale networks. Therefore, it is necessary to use multicontroller architecture, which has some problems with the controller placement. In this article, we take both the average latency and the worst latency between switch and controller into consideration and make a multi-objective optimization model. An improved label propagation algorithm based on traffic gravitation is proposed to solve the subdomain division problem, and a heuristic method is for subdomain controller placement. The simulation experiments show the effectiveness of the proposed algorithm and the time complexity guarantee for large-scale networks.

1. Introduction

Software-defined network (SDN) provides extremely flexible and customizable network services by separating the control plane and the data plane. However, with the expansion of network scale and data traffic, centralized architectures cannot fulfill the needs of efficiency, scalability, and availability [1]. In that case, there should be several distributed synchronous controllers in the SDN network. It has some problems that need to be solved, like how many controllers are needed in an SDN network, where should they locate, and what is the mapping of the controllers and the switches. Heller et al. [2] proposed the controller placement problem (CPP) for the first time and established a mathematical model for it, and the experiments showed that adding controllers for most networks would reduce the latency, but the benefits would decrease with the increase of controllers.

Most research regards CPP as a clustering problem, in which case the switches are regarded as the nodes with community attributes. However, CPP and traditional community detection [3] are not exactly the same. The nodes in the latter have strong connections with neighbors but weak connections with distant nodes, which is different from the forwarding devices in the network. Although the two network nodes are far apart (cannot share a controller), a large amount of data traffic can be exchanged between them. Therefore, in addition to considering the nature of the network topology itself, it is also necessary to increase the consideration of traffic demands.

The CPP has many different optimization objectives [4], such as network responsiveness, fault tolerance, resilience, and QoS. We focus on the responsiveness of the control plane, including the average and the worst latency between switch and controller, and solve the CPP in two steps. The first step is to solve the problem of controller number and mapping, by dividing the network topology into multiple fully connected subdomains through an improved label propagation algorithm (LPA) that uses the abstract gravitation between nodes as a value function. In the second step, we find the best controller placement position in each subdomain, using the gravitational force of nodes to the controller and a heuristic algorithm based on open searching. Our main contributions are as follows:

(i) A model that considers both the average latency and the worst latency is designed for the CPP, and a two-step heuristic algorithm is proposed to get the approximate optimal solution

(ii) We define the traffic gravitation using the traffic demands of the network and propose a subdomain division algorithm based on the LPA to determine the number of controllers and the mapping

(iii) By defining the space gravitation between nodes and controllers, a controller placement algorithm is proposed to determine the specific location of the controller in the subdomain

(iv) Experiments show that the proposed placement algorithm is effective and has less time complexity compared with some excellent algorithms

The rest of this article is organized as follows: Section 2 surveys the related work, showing the present research methods and their insufficiency. Section 3 optimizes the model of CPP with both average latency and worst latency and shows a brief of LPA. A two-step heuristic algorithm for the CPP is proposed in Section 4. Section 5 shows comparative simulations between the proposed algorithm and others. Finally, there are the future work in Section 6 and the conclusion in Section 7.

2. Related Work

There has been some research on the CPP, and some progress has also been made. The most common method is to optimize the proposed mathematical model through (integer) linear programming (ILP). Yao et al. [5] consider the controller's capacity for the first time, which relies on the k-center algorithm but adds capacity constraints. Finally, the problem is solved by integer linear programming. Simulation experiments show that the strategy can effectively reduce the number of controllers and the load of the busiest controller. Sallahi et al. [6] take the deployment cost as the optimization goal to determine the optimal number and the placement of controllers. The research uses linear programming to build the model, which is solved by a linear solver. Although this method is effective, it is very time-consuming and only suitable for static calculations in small networks. He et al. [7] select the forwarding devices in the network topology as an alternative location and complete the construction of a linear optimization model for the end-to-end data flow establishment time. They convert the problem into a mixed-integer linear optimization problem and adopt a linear optimizer, Gurobi [8], to solve it. It can be seen that the linear programming method has the advantages of strong versatility and optimization. But no matter what the optimization goal is, it requires a lot of prior knowledge, and its computational complexity is too high to be accepted in the network solution for dynamical adjusting in time.

The heuristic algorithm based on modularity is a common community detection method, which is also suitable for solving many CPP problems. Fan et al. [9] establish an optimization model considering both control latency and reliability, which uses an improved Louvain algorithm [10] based on modularity to calculate subdomains. Then, they use particle swarm optimization (PSO) to further calculate the placement of the controller in the subdomain. In order to solve the problem of resolution limit and subdomain disconnection caused by the Louvain algorithm, Traag et al. [11] propose a Leiden algorithm, which is based on the subdomain calculated by the Louvain algorithm and performs an internal subdivision. It combines the nodes with a certain probability and forms further-subdivided subdomains. Experiments show that the Leiden algorithm has more efficient subdomain partitioning capabilities while avoiding the problem of resolution limit and subdomain disconnection, and it is still applicable in the weighted graph. Chen et al. [12] adopt the same research steps. First, the Louvain algorithm based on modularity is used to calculate the subdomain, and then, the nodes with the smallest average and the smallest worst latency are found in the subdomains as the placement of the controller. A modularity-based heuristic algorithm is a common community detection algorithm, which has good results in community discovery. However, the CPP is not exactly the same as ordinary community detection, because of the traffic demands between network nodes. So when this type of algorithm is applied to the real network topologies, the performance is slightly unsatisfactory.

Another common method of community detection is the label propagation algorithm (LPA) [13], which is also widely used in CPP problems. CLPA is an algorithm for controller load balancing and network stability proposed by Liu et al. [14]. The network stability is abstracted from the reciprocal of the number of hops from the forwarding device to the controller. CLPA divides the network into different subdomains by using LPA and then uses the k-median algorithm [15] to calculate the placement of the controller in each subdomain. The simulation experiments show that its computational complexity and delay performance are higher than LPA.

Table 1 is a brief overview of the current and the proposed algorithms, where the SC-avg and SC-wst, respectively, represent the average and the worst latency between the switch and the controller, and MILP represents mixed-integer linear programming. And the remaining CPP research involves many optimization methods and goals, such as NSGA-II [16] based on the genetic algorithm and effective Pareto algorithm [17] based on the Nash bargaining model. Most CPP research sets the constant (or infrequently changing) attribute value of the nodes as the optimization goals, but the combination of the immutability of network topology and the variability of traffic demands is actually one of the difficulties of CPP [18]. For this reason, this article converts the characteristics of traffic demands into the attributes of network nodes as much as possible and proposes a controller placement algorithm that focuses on the influence of traffic, so that the final strategy is more in line with the real network topology.

Table 1: An overview of current controller placement approaches and the proposed approach.

Author	Objective(s)	Method	Tool(s)	Evaluation
Yao et al. [5]	The SC-wst	LP	Capacitated K-center	Fewer number of controllers, but is complex for large-size networks
Sallahi et al. [6]	Network cost	ILP	Linear solver	High complexity, only for the small size networks
He et al. [7]	Flow setting time	MILP	Gurobi [8]	Moderate complexity
Fan et al. [9]	Control latency and reliability	Modularity	Louvain and PSO [10]	Low complexity, but has subdomain disconnection problem
Traag et al. [11]	Control latency	Modularity	Leiden	Low complexity, fix the problem of the Louvain algorithm
Chen et al. [12]	The SC-avg or SC-wst	Modularity	Louvain, traverse search	Moderate complexity, well-done for subdomain division
Liu et al. [14]	Load balancing and stability	LPA	K-median	Much low complexity, but is bad for subdomain division with traffic
The proposed	The SC-avg and SC-wst	LPA	Heuristic algorithm	Much low complexity and is good for both SC-avg and SC-wst

3. Optimization Model

3.1. Problem Description. The SDN infrastructure is shown in Figure 1, and CPP is to calculate the optimal number, location, and mapping of controllers in the control plane when the data plane and related network parameters are known. As shown in Figure 1, the entire topology deploys two controllers, located at SW2 and SW4, respectively. The control domain of CTL A is \{SW1, SW2, SW3\}, and the control domain of CTL B is \{SW4, SW5\}.

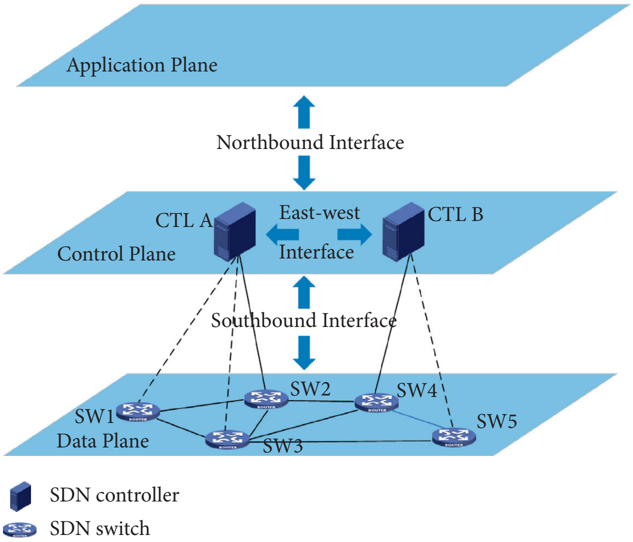


Figure 1: The infrastructure of SDN.

3.2. Model and Symbol. This article uses the models of graph theory to abstract the optimization model. For a given network topology, the data layer is $G = (V, E)$, where $V = \{v_i\}_n$ is the set of network nodes, $|V| = n$ is the total number of nodes, and $E = \{e_{ij}\}_{n*n}$ is the set of network links. When $e_{ij} = 0$, there is no direct physical link between nodes v_i and v_j. The control layer is $C = \{v_i\}_k$, where k is the total number of controllers.

For a simple description, it is necessary to describe the mapping of switches and controllers using a specific data structure. First of all, the entire network topology is divided into several subdomains according to the control domain, presented as a dictionary structure subDomain = $\{l_i\colon \{v_0, v_1 \ldots v_{n_i}\}\}, i \in [0, k)$, where l_i is the label (unique identifier) of i subdomain and n_i is the number of nodes in i subdomain. We have

(1) Each switch only has one controller, that is,

$$\cup_{i=0}^{k-1} s_i = V, \tag{1}$$

$$s_i \cap s_j = \phi, \quad \forall i, j \in [0, k) \text{ and } i \neq j, \tag{2}$$

where s_i = subDomain$[i]$ is the node set of i subdomain. Equation (1) means that all nodes are divided into subdomains, and equation (2) means that the subdomains do not overlap each other.

(2) Each subdomain only has one controller, and the controllers of all subdomains are different from each other. In this article, unless otherwise specified, the label of the subdomain and the subscript of the controller are considered to have a one-to-one correspondence, shown in the following equation:

$$s_i \leftrightarrow l_i \leftrightarrow c_i. \tag{3}$$

Then, we use traffic matrix $M = \{m_{ij}\}_{n*n}$ to describe the end-to-end traffic demands between nodes in network G, where m_{ij} is the total traffic that node v_i needs to send to node v_j in the initial state. Define $T_{fwd} = \{tf_i\}_n$ and $T_{snd} = \{ts_{ij}\}_{n*n}$, where tf_i is the total traffic actually forwarded by the node v_i and ts_{ij} is the total traffic actually required to be sent from v_i to v_j, shown as in the following equations:

$$tf_i = \sum_{m_{sd} \in M} m_{sd} \cdot \delta_{sd}(i), \quad \forall i \in [0, n), \tag{4}$$

$$ts_{ij} = \sum_{m_{sj} \in M} m_{sj} \cdot \delta_{sj}(i), \quad \forall i, j \in [0, n), \tag{5}$$

$$\delta_{sd}(i) = \begin{cases} 1, & \text{if } v_i \in \text{Path}[v_s][v_d], \\ 0, & \text{else,} \end{cases} \tag{6}$$

where $\text{Path}[v_s][v_d]$ is the routing path from v_s to v_d.

Finally, summarize the optimization model of controller placement. We focus on the response latency between switches and controllers, in which the average latency (represented by SC-avg) can reflect the basic performance of propagation in the SDN network and the worst latency (represented by SC-wst) can reflect the performance under strict constraints [19]. The optimization problem for optimizing these two goals can be obtained as

$$\min \alpha L_{avg} + \beta L_{wst}, \tag{7}$$

$$L_{\text{avg}} = \frac{1}{n} \sum_{v \in V} d(v, c_i) | v \in s_i, \tag{8}$$

$$L_{wst} = \max_{v \in V} d(v, c_i) | v \in s_i, \tag{9}$$

$$\text{s.t. } \alpha + \beta = 1, \tag{10}$$

$$\begin{aligned} \text{len}(\text{subDomain}) = k \\ \text{satisfy equation } (1)(2)(3), \end{aligned} \tag{11}$$

where $d(v, c_i) | v \in s_i$ is the latency from the node v to the controller c_i of its subdomain. Equations (8) and (9) are the calculation formulas for the average and the worst latency. Equation (10) is the weight constraint, and equation (11) is the controller number constraint.

3.3. Brief of LPA. Since the subdomain division algorithm is based on LPA, in order to ensure the continuity of the article, it is necessary to briefly introduce LPA. LPA is originally applied to solve the problem of community detection. Because of its simple idea and approximately linear time complexity; meanwhile, the result does not depend on the initial solution, and it is widely used in various fields. The main process is

(1) Initialize the network; each node is regarded as an independent community and marked with a globally unique label

(2) Randomly select some nodes and update the node label to the label with the largest value of the evaluation function among its neighbors

(3) Iterate the second step until there are no more changes to the label

When LPA converges, nodes with the same label belong to the same community. The limitations of LPA are dichotomous oscillation, unstable results of multiple calculations, and resolution limitations [11]. These problems are solved in Section 4.1 using some restrictions.

4. Solution of the Problem

For the optimization goal of Section 3.2, a controller placement algorithm is proposed, which aims at minimizing the average latency and the worst latency between switches and controllers as much as possible in a large-scale network. As analyzed above, CPP needs to be divided into two parts, which are the subdomain division of the network and the placement of the controller in each subdomain. Therefore, algorithms are proposed to solve these two subproblems based on the node's traffic gravitation and space gravitation, respectively. The flow diagram of the whole algorithm is shown in Figure 2, and the details are introduced in Sections 4.1 and 4.2.

4.1. Subdomain Division Algorithm. In community detection, there are some indicators to measure the importance of nodes, such as degree centrality [20], betweenness centrality [21], and LeaderRank value [22]. However, these indicators cannot fully measure the importance of nodes in the field of communication networks, in which one of the main differences is the traffic demands. We use T_{fwd}, the total traffic forwarded by nodes, to present the importance of nodes. Furthermore, inspired by the betweenness centrality, the out-degree $d_{\text{out}}(v)$ is selected to be the penalty factor. Combining the results of several rounds of simulation experiments, the importance of node I_v is defined in the following equation:

$$I_v = \frac{tf_v}{\sqrt{d_{\text{out}}(v)}}, \quad \forall v \in V. \tag{12}$$

Influenced by traffic demand, we believe that there is a mutual attraction relationship between nodes. The more traffic that needs to be transmitted between two nodes, the greater the possibility of belonging to the same subdomain. Therefore, through analogy with the concept of gravity, the traffic T_{snd} actually transmitted between nodes is defined as the quality of the nodes, and the length hops_{ij} of the routing path between two nodes is defined as the distance. Therefore, the traffic gravitation $F_{traf}(v_i, v_j)$ between nodes is defined in the following equation:

$$F_{traf}(v_i, v_j) = \frac{ts_{ij}}{\text{hops}_{ij}^2}, \quad \forall v_i, v_j \in V. \tag{13}$$

Section 3.3 has briefly introduced LPA, which is improved in this section to get better performance. In the label propagation process of LPA, the update selection of nodes is random, which may increase the convergence time exponentially in the worst case. So we consider sacrificing a certain amount of randomness to greatly shorten the convergence time. When initializing the network topology, the traffic matrix is used to calculate the importance of each node, and the initial update queue is formed in descending order. In each round of updates, the node whose label

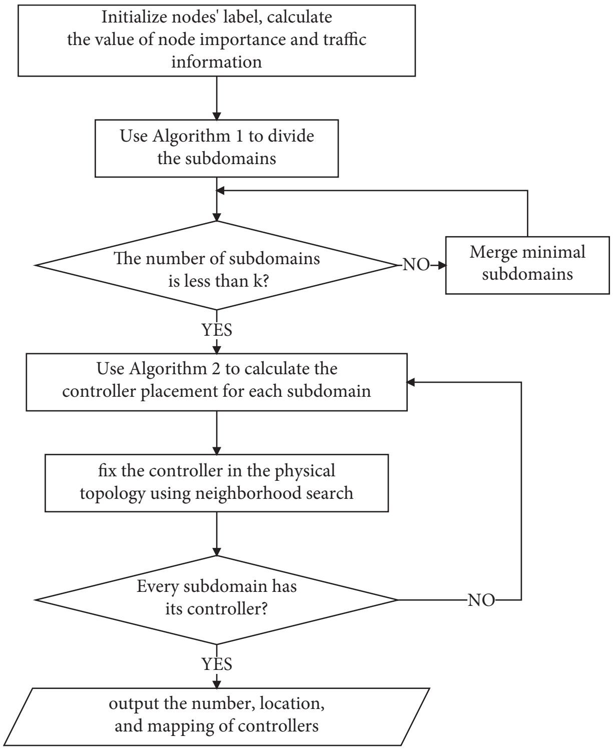

Initialize nodes' label, calculate the value of node importance and traffic information
Use Algorithm 1 to divide the subdomains
The number of subdomains is less than k?
NO
Merge minimal subdomains
YES
Use Algorithm 2 to calculate the controller placement for each subdomain
fix the controller in the physical topology using neighborhood search
Every subdomain has its controller?
NO
YES
output the number, location, and mapping of controllers

Figure 2: The flow diagram of the placement algorithm.

changes and its neighbors join the update queue again according to the node's importance, to wait for the next update round. This node selection strategy can effectively reduce the number of nodes that need to be processed during each update round in a large-scale network, thereby greatly reducing the algorithm convergence time. The complete process of the subdomain division algorithm (gravSDA) is described in detail through Algorithm 1, and Table 2 shows the meaning of the symbols used in gravSDA.

Algorithm 1, gravSDA, improves on LPA, in which the initialization phase completes the initialization of the node label l_i, the calculation of T_{fwd} and T_{snd}, and the initialization of the update queue Q_{update}. The entire iterative process of label propagation is completed by lines 1 to 11. Lines 3 to 10 are the core of gravSDA, which uses the traffic gravitation among neighboring nodes to update the label of the node; lines 7 and 10 indicate that the update queue is updated to the union of the nodes with the label change and its neighbors, and it is sorted in descending order according to I_v. In the iterative process, the maximum number of iterations maxIter is added to prevent the dichotomous oscillation that may occur. It should be noted that if the number of subdomains is larger than k in line 15, we will combine the smallest subdomain to its neighbor until there are just k subdomains. The flow diagram of gravSDA is shown in Figure 3.

GravSDA uses the network traffic matrix and the traffic gravitation between nodes as the standard for label update, so it is more in line with the requirements of subdomain division in the communication network topology compared with other graph theories and clustering algorithms. Its effectiveness for the real network is experimentally verified in Section 5.3. Furthermore, gravSDA improves the order and range of label updating on the basis of LPA, which greatly reduces the algorithm convergence time, and it is verified in Section 5.4.

4.2. Controller Placement Algorithm in Subdomain. With the results of Section 4.1, it can further study the placement of controllers in each subdomain. Since there is no strict correspondence between the data flow and the control flow [23], the traffic gravitation model is not suitable for controllers. Thus, we assume the following:

(1) All new data flows trigger packet-in message from the switch to the controller, and the total amount of transmitted data is equal

(2) The soft and hard timeout time of all flow tables is constant

(3) Although in-band communication is used between the switch and the controller, sufficient bandwidth is reserved under any circumstances to transmit control messages

Under the above assumptions, the latency of control messages is only determined by the length of the routing path, and the importance of the node depends on its out-in degree in a subdomain. Since the synchronization latency between the controllers needs to be considered, when the subdomain s_i is analyzed separately, the remaining subdomains should be regarded as virtual nodes to analyze the impact on s_i.

Inspired by the idea of force balance, CPP can be analogous to a scene where one particle is balanced by force and is fixed in a force field. First, select an initial position for the controller c_i, which can be any position in the space area. And then, it is affected by the gravitational force $\overrightarrow{F_g}$ of both nodes in the domain and virtual nodes outside the domain. The force $\overrightarrow{F_g}$ of the node to the controller is similar to the node's demand for the controller. When the controller is far away, the node "eagerly" wants the controller to be closer, so the calculation formula should be more similar to Hooke's law. We propose the calculation formula in the following equation:

$$\overrightarrow{F_g}(v_i, c, e) = \begin{cases} 0, & \text{if } \operatorname{dis}(v_i, c) < e, \\ \omega \cdot (\operatorname{dis}(v_i, c) - e), & \text{else}, \end{cases} \tag{14}$$

where e is the original length, that is, when the distance between nodes and controllers is less than e, the gravitational force is ignored, $\operatorname{dis}(v_i, c)$ is the Euclidean distance between node v_i and controller c, and ω is the coefficient of elasticity.

Finally, some special processing is added to ensure the time complexity and effectiveness of the controller placement algorithm (gravCPA). When selecting the initial position, choose the center of the smallest covering circle that contains all nodes in a subdomain, which can be obtained with the Elzinga–Hearn algorithm [24] in $O(n)$ time complexity. The controller may oscillate in the space force field, so every time in the iteration, the penalty factor φ is used to reduce the moving step length of the controller to ensure

```
Input: $G(V,E), M$, maxIter, $k$
Output: subDomain = $\{l_i: \{v_0, v_1 \ldots v_{n_i}\}\}$
Initialize: subDomain = $\{v_i$: set for $v_i$ in $V\}$
  $T_{fwd}, T_{snd}$ calculate values using equations (4) and (5)
  $Q_{\text{update}}$ = sorted ($V$, key = lambda $x$ $I_v(x)$)
Process
while is change and iterTimes < maxIter do
    iterTimes+ = 1, $Q_{\text{new}} = \phi$
    for node $v_i$ in $Q_{\text{update}}$ do
        $l_{\text{new}}$ = findMaxNeighbors$(G, T_{snd}, v_i)$
        if $l_i \neq l_{\text{new}}$ then
            update $l_i$ and is change
            $Q_{\text{new}}$.add$(v_i \cup G$.neighbours$(v_i))$
        end if
    end for
    $Q_{\text{update}}$ = sorted ($Q_{\text{new}}$, key = lambda $x$ $I_v(x)$)
end while
for node $v_i$ in $V$ do
    subDomain$[l_i]$.add$(v_i)$
end for
return subDomain
Function 1 findMaxNeighbors $(G, T_{sn\,d}, v_i)$ $\longrightarrow$ label:
res = $[-1, G$.Nodes$(-1)]$
for node $v_j$ in $G$.neighbours$(v_i)$ do
    res = $[F_{\text{traf}}(v_i, v_j), v_j]$if $F_{\text{traf}}(v_i, v_j) >$ res$[0]$
end for
return res[1].label if res[0] $\neq -1$ else $v_i$.label
```

Algorithm 1: Subdomain Division Algorithm (gravSDA).

Table 2: Symbols of Algorithm 1.

Symbol	Description
$Q_{\text{update}}/Q_{\text{new}}$	Node queue to be updated in this round of iteration/ next round
isChange	Flag indicated whether the label has changed in this iteration
G.nodes(i)	Construct an instance of the node numbered i in the network G

final convergence. Due to the open search, the final controller may place outside the network topology. In order to make gravCPA more general, it needs to search again in its surroundings to find an alternative placement in the topology. It is described in Algorithm 2 in detail, and Table 3 is a description of the symbols used in gravCPA.

The idea of Algorithm 2 is based on the space gravitation $\overrightarrow{F}_g$ of the node to the controller, which is firstly finding the optimal controller placement through open search and finally fixing the placement in the network topology. The open search process is lines 2 to 8, in which the resultant force in intradomain $\overrightarrow{F}_{\text{int}}$ is calculated by equation (14) and the resultant force extra-domain $\overrightarrow{F}_{\text{ext}}$ only considers the unit force formed by the relative positions. Each iteration penalizes the step length to ensure that the algorithm eventually converges. The surrounding search process is the 9th to 13th line, in which the placement is found through traversal in some subdomain nodes with suitable distance tolerance. The flow diagram of Algorithm 2 is shown in Figure 4.

5. Simulation

The performance and convergence time of gravCPA are simulated and evaluated on the x86 platform in this section. The operating system of the simulation experiment is Ubuntu 16.04.3 LTS, the processor model is Intel(R) Xeon(R) CPU E5-2609 0 @ 2.40 GHz, and the physical memory is 16 GB. And in this section, we define the sum of SC-avg latency and SC-wst latency as the integrated latency for a simple description.

5.1. Influence of the Number of Controllers. Except for a few algorithms that depend on the initial conditions, other CPP algorithms can obtain the optimal number of controllers when running. In fact, the number of controllers cannot be increased without any upper limit. Based on this consideration, we limit the number of controllers and compare the proposed gravCPA with CTR [7], LDN [11], and CLPA [14]. The metrics are SC-avg latency and SC-wst latency in the network. Particularly, when the limited number is greater than the calculated number, the latter is selected.

We use Python 3.8.0 and NetworkX components for simulation and use a randomly generated LFR benchmark network that is fully connected. After 200 times repeated experiments, the average results are shown in Figure 5. The parameters in this experiment are shown in Table 4, where PLD means power law distribution.

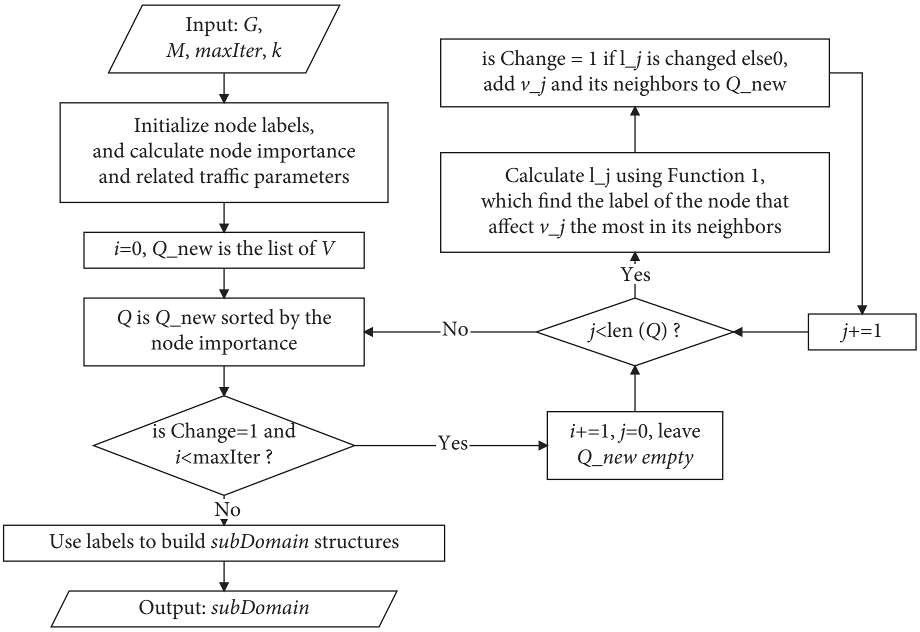

Input: G,
M, maxIter, k
Initialize node labels,
and calculate node importance
and related traffic parameters
i=0, Q_new is the list of V
Q is Q_new sorted by the
node importance
is Change=1 and
i<maxIter ?
No
Use labels to build subDomain structures
Output: subDomain
Yes
i+=1, j=0, leave
Q_new empty
No
j<len (Q) ?
Yes
Calculate l_j using Function 1,
which find the label of the node that
affect v_j the most in its neighbors
is Change = 1 if l_j is changed else0,
add v_j and its neighbors to Q_new
j+=1

Figure 3: The flow diagram of Algorithm 1.

```
Input: $G(V, E)$, subDomain, $\varphi, \varepsilon, \delta$
Output: dictionary structure: $ctls = \{l_i\colon v\}$
Initialize: $ctls = \{l_i\colon \text{EH}(s_i)\ \text{for}\ l_i, s_i\ \text{in subDomain}\}$
Process
for $s_i$ in subDomain do
  $\text{loc}_i = ctls[l_i]$
  while STEP $> \varepsilon$ and $\text{loc}_i$ change do
    $\overrightarrow{F}_{\text{int}} = \text{sum}(\overrightarrow{F_g}(v_j, \text{loc}_i, e)\text{for}\ v_j\ \text{in}\ s_i)$
    $\overrightarrow{F}_{\text{ext}} = \text{sum}\left( \begin{array}{c} (v_j.x - \text{loc}_i.x, v_j.y - \text{loc}_i.y)/\sqrt{v_j.x - \text{loc}_i.x^2 + v_j.y - \text{loc}_i.y^2} \\ \text{for}\ v_j\ \text{in}\ V_{\text{ext}}(s_i) \end{array} \right)$
    $\text{loc}_i \leftarrow \text{loc}_i + (\overrightarrow{F}_{\text{int}} + \overrightarrow{F}_{\text{ext}}) \cdot \text{STEP}$
    $\text{STEP} * = \varphi$
  end while
  $v_{\text{res}} \leftarrow$ find the closest node to $\text{loc}_i$
   $D = \text{dis}(v_{\text{res}}, \text{loc}_i)$
   for $v_j$ in $\backslash\{v$ for $v$ in $s_i$ if $\text{dis}(v, \text{loc}_i) < D + \delta\backslash\}$ do
     $v_{\text{res}} = v_j$ if calOptFunc $(v_j) <$ calOptFunc $(v_{\text{res}})$
   end for
   $ctls[l_i] = v_{\text{res}}$
end for
return ctls
Function 2 calOptFunc $(v) \longrightarrow$ int:
res$\leftarrow$calculate the value of equation (7)
return res
```

Algorithm 2: Controller Placement Algorithm (gravCPA).

Figure 5 shows the response latency curve of the network control plane when the number of controllers changes under gravCPA and three comparative algorithms. Figure 5(a) shows the change of SC-avg latency, and it can be seen that gravCPA has better performance on SC-avg than other CPP algorithms. When the number of controllers $k = 5$, gravCPA's SC-avg is 233.86 μs, which is 21.1% lower than CLPA's 296.52 μs and 32.6% lower than CTR's 347.1 μs; when $1 \leq k \leq 6$, the SC-avg of gravCPA is lower than LDN, but it is a little worse when $k > 6$. Combined with Figure 6, it can be seen that 95% of k is less than 8.63 and 80% is less than 7.03 for gravCPA,

Table 3: Symbols of Algorithm 2.

Symbol	Description
φ	Factor for STEP penalty
ε	Iteration accuracy for STEP
δ	Factor for distance tolerance
EH (s_i)	Function that calculates the minimum covering circle's center of s_i using Elzinga–Hearn [24]
STEP	Step length of controller movement
$\overrightarrow{F}_{\text{int}}/\overrightarrow{F}_{\text{ext}}$	Resultant force of intra-/extra-domain nodes
$V_{\text{est}}(s_i)$	Set of subdomain abstract nodes except s_i

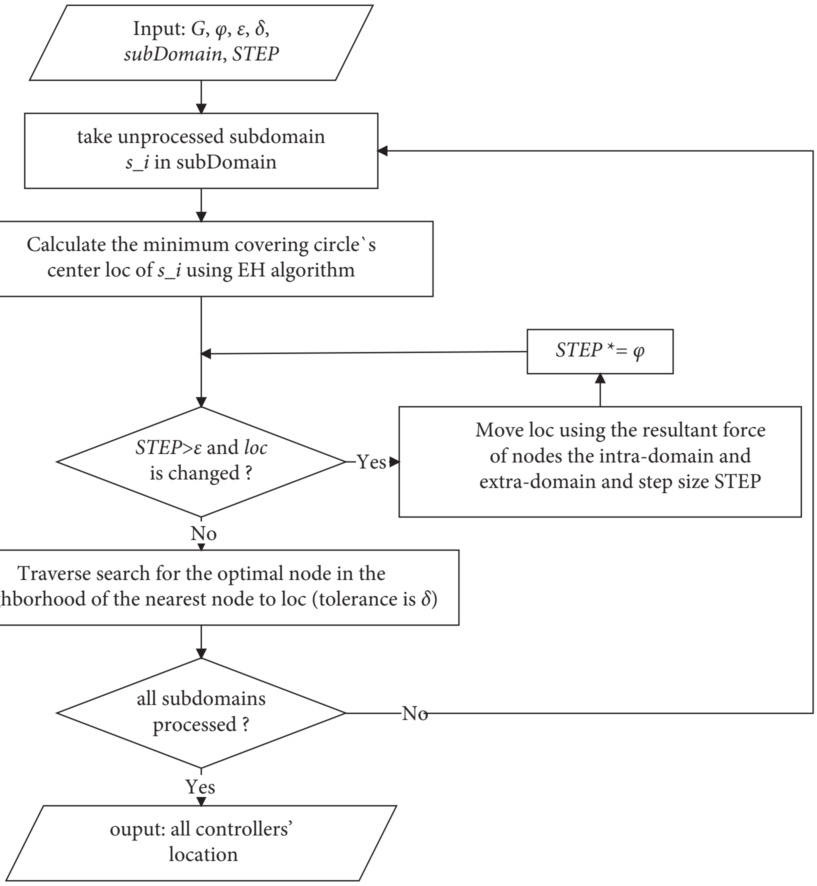

Input: G, φ, ε, δ, subDomain, STEP
take unprocessed subdomain s_i in subDomain
Calculate the minimum covering circle`s center loc of s_i using EH algorithm
STEP *= φ
STEP>ε and loc is changed ?
Yes
Move loc using the resultant force of nodes the intra-domain and extra-domain and step size STEP
No
Traverse search for the optimal node in the neighborhood of the nearest node to loc (tolerance is δ)
all subdomains processed ?
No
Yes
ouput: all controllers' location

Figure 4: The flow diagram of Algorithm 2.

while these two values for LDN are 10.25 and 9.24, respectively. In another word, the controller's number of gravCPA is not more than 7 in most cases. When k increases in the gravCPA experiment, the number of controllers does not increase actually so that the SC-avg latency does not decrease. In summary, gravCPA can obtain a better average latency with a smaller number of controllers.

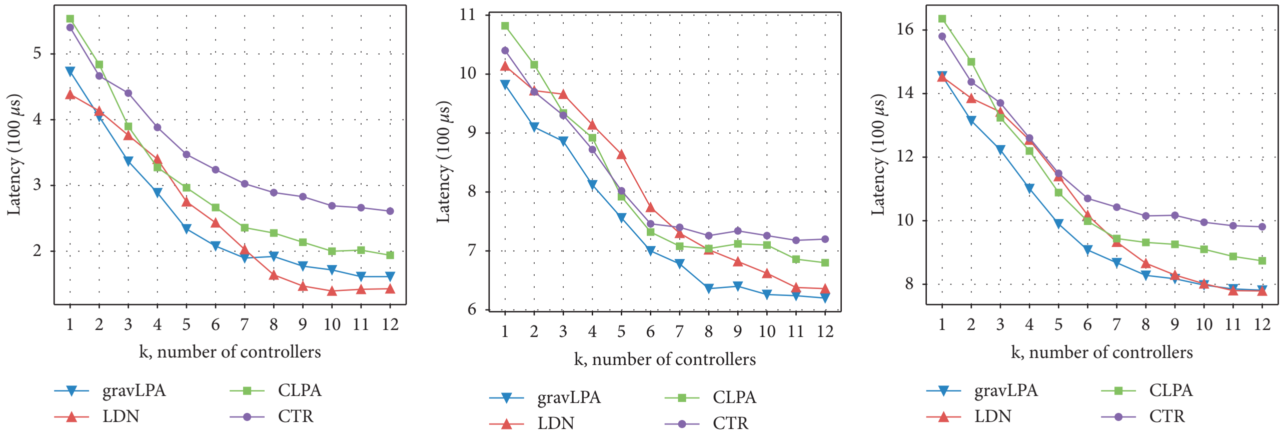
Figures 5(b) and 5(c) show the curve of the worst latency and the integrated latency, respectively. It can be seen that these two parameters of gravCPA are better than other algorithms. Furthermore, based on the numerical error of 4.56%, gravCPA converges when $k = 7$, and at this time, the integrated latency 867.62 μs decreases by 7.0%, 8.1%, and 16.8%, respectively, compared with LDN's 932.68 μs, CLPA's 943.82 μs, and CTR's 1042.6 μs.

5.2. Cumulative Distribution Function of Latency. Although each CPP algorithm wants to make an optimal placement for every network as much as possible, its performance varies with the network parameters and situation. We use a randomly generated network topology with full connection to verify the versatility, where the number of nodes is $n \in [12, 200]$ and the number of links is $m \in [n-1, n*(n-1)/2]$. The cumulative distribution function of the integrated latency of gravCPA is compared with CTR, LDN, and CLPA.

We also use Python 3.8.0 and NetworkX components for simulation and use Seaborn components to generate the cumulative distribution function curve from the results of 200 repeated experiments. The results are shown in Figure 7.

Figure 7 shows the cumulative distribution function of the integrated latency of four CPP algorithms in a

]



(a) (b) (c)

FIGURE 5: Response latency curve of controllers' number k.

TABLE 4: Simulation parameters settings.

Parameter	Value	Description
n	100	Number of nodes
$d_{\min}$	2	Minimum degree of nodes
τ_1	3	PLD index for the degree
τ_2	1.5	PLD index for the community size
μ	0.1	Fraction of intercommunity edges
$\alpha, \varphi, \varepsilon, \delta$	0.5, 0.99, 10^{-6}, 3 units	Parameters of gravCPA

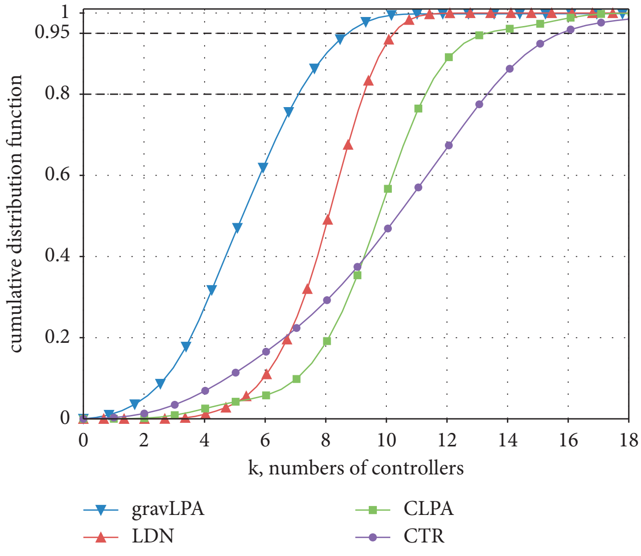


FIGURE 6: The cumulative distribution function curve of controllers' number without k-limitation.

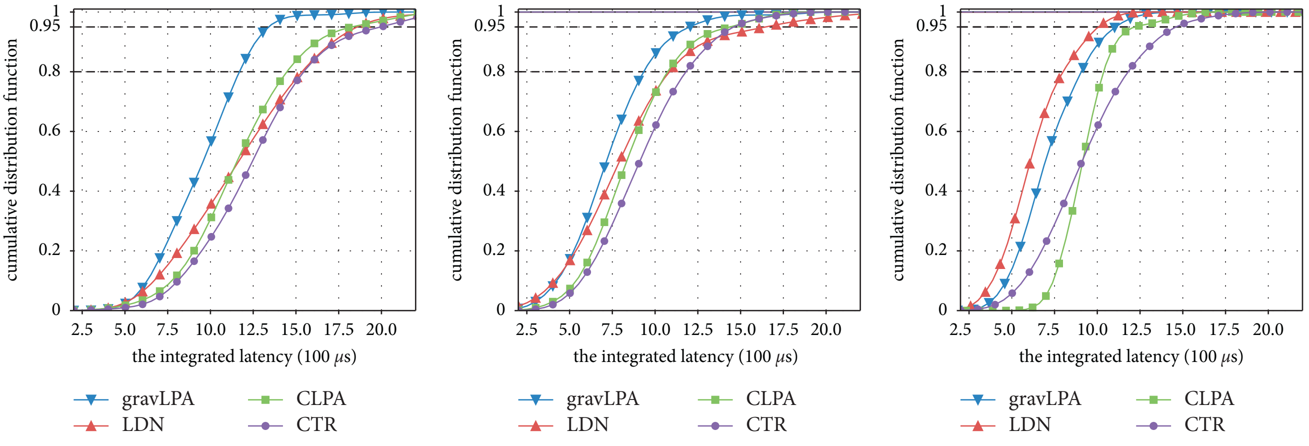
randomly generated network. When $k = 3$, as shown in Figure 7(a), the integrated latency performance of gravCPA is significantly better than the other three algorithms. Compared with CLPA that has the second best performance, when the confidence levels are 95% and 80%, the integrated latency of gravCPA decreases by 26.2% (from 1789.11 μs to 1320.58 μs) and 18.8% (from 1420.98 μs to 1153.25 μs), respectively.

Comparing Figures 7(b) and 7(c), it can be seen that when $k > 6$, the integrated delay of gravCPA no longer decreases with the increase of k. Relatively, the other three algorithms have different improvements and the performance of LDN is improved the most. However, when $k = 12$, the integrated latency of gravCPA (1084.46 μs) is just slightly larger than LDN (972.9 μs), which is about 111.4% of LDN.

In summary, it is concluded that gravCPA can still achieve better performance of integrated latency with a small number of controllers when faced with the randomly generated network.

5.3. Performance in Real Networks. Sections 5.1 and 5.2 show the latency performance of the four algorithms in the LFR benchmark network and the randomly generated network. Since the traffic demands between nodes in the real network are not exactly the same, it is more complicated for research. In this section, we use the real network topologies and traffic demands [25] to compare the latency performance of each algorithm.

We use the real topologies in Table 5 simulated by Python 3.8.0 and NetworkX components. Assume that the first data packet of each flow triggers the packet-in message, and the transmission latency of the control message is the sum of each hop latency from the node to the subdomain controller, where the hop latency is equal to the reciprocal of one-tenth of its bandwidth. The experimental results are shown in Figure 8.



(a)

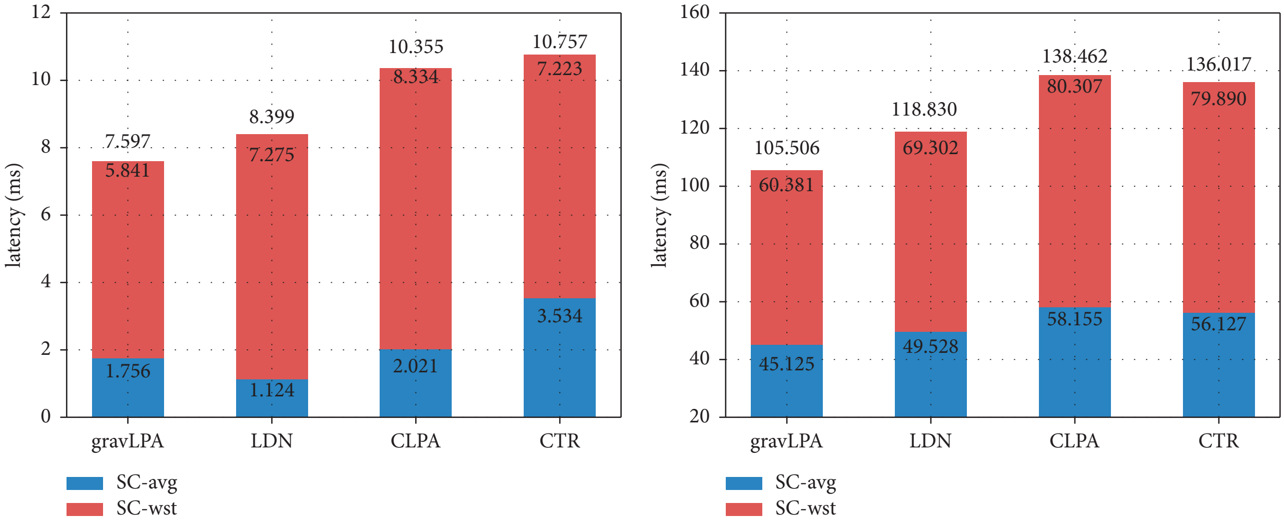
(b)

(c)

Figure 7: The cumulative distribution function curve of the integrated latency.

Table 5: Topologies in the experiment.

Topology	Nodes	Links	Demand pairs
Abilene	12	15	132
Cost266	37	57	1332



(a)

(b)

Figure 8: The latency of CPP algorithms in real networks.

Figure 8(a) shows the simulation result of the Abilene network with 12 nodes, which is representative of the small-scale network. It can be seen that although the SC-avg latency of gravCPA is slightly larger than LDN, the remaining parameters are all better than the other three algorithms. The integrated latency of gravCPA is only 7.597 ms, compared with 8.399 ms of LDN, 10.355 ms of CLPA, and 10.757 ms of CTR, which are reduced by 9.54%, 26.63%, and 39.37%, respectively. In a word, the gravCPA based on traffic gravitation has a stronger optimization performance when considering the real traffic demands.

Figure 8(b) shows the simulation result of the Cost266 network with 37 nodes, which is a medium-scale network. Compared with Alibene, the traffic demands in large- and medium-scale networks are more complex, and better performance cannot be obtained without considering the impact of traffic. When gravCPA restricts the number of controllers in Cost266, the integrated latency is 60.381 ms. Compared with LDN's 69.302 ms, CLPA's 80.307 ms, and

Table 6: Convergence time (ms) of four algorithms.

Nodes number	10	50	100	500	1,000	5,000	10,000
gravCPA	0	1.137	3.794	17.784	35.453	929.28	2070.218
LDN	0	1.099	5.191	15.399	52.202	1170.729	4725.17
CLPA	0	2.592	14.956	71.583	194.544	3616.096	TLE
CTR	0	1.027	16.74	202.514	9152.028	TLE	TLE

CTR's 79.89 ms, it decreases by 12.87%, 24.81%, and 24.42%, respectively.

5.4. Convergence Time. Although this article does not focus on the dynamic placement, time complexity will seriously affect the universality of the algorithm faced with large-scale network topologies. Thus, we increase the scale of the network to compare the convergence time of each algorithm. The results are shown in Table 6, which cannot be shown in a line or bar figure because of the serious nonlinear growth of the value. It should be noticed that the convergence time includes the time of subdomain division and controller placement, and the TLE in Table 6 means the code is the time limit (5 minutes) exceeded.

Although there may be errors in code optimization, the experimental results are in line with expectations. GravCPA and LDN are based on LPA and Louvain algorithms, respectively, which improve the iterative operations and have good time complexity. However, CLPA does not consider calculation acceleration, while CTR is an algorithm based on linear programming. It is difficult for the two to complete the calculation in a short time with a larger network.

6. Future Work

6.1. Flow Types. In traffic engineering, data flow is usually divided into two types, elephant flow (bulk data transfer), and mice flow (short-lived data exchange) [26]. These two types have completely different effects on the control messages between the switch and the controller. The former has huge data transmission but only exchanges little control messages, while the latter is totally opposite. In this article, the placement strategy is only based on the size of the traffic demands, the performance will decrease when facing extreme conditions. We consider completing the controller deployment strategy based on the predicted flow type in the next step.

6.2. Stability. We have noticed that both LPA and gravitation models have problems with oscillations and non-unique results. Although we have added a lot of assumptions and constraints, there is still a problem that the placement strategy results are not unique in specific or complex scenarios. And this part will become the direction of follow-up research.

7. Conclusion

In this article, we focus on the CPP in the SDN multi-controller architecture. Different from many research, we optimize the average latency and the worst latency using the data plane traffic demands. For the subdomain division problem, the traffic gravitation is defined and an improved LPA is designed accordingly. On the other hand, for the subdomain controller placement, we use the open search for the first and then use the traversal search in the surrounding of the first step's result to decide which placement is located in the topology. The comparative experiment proves the effectiveness of the proposed algorithm, which can achieve lower average latency and worst latency with a smaller number of controllers, and it also has a certain time complexity guarantee.

References

[1] O. Blial, M. Ben Mamoun, and R. Benaini, "An overview on sdn architectures with multiple controllers," *Journal of Computer Networks and Communications*, vol. 2016, 8 pages, Article ID 9396525, 2016.

[2] B. Heller, R. Sherwood, and N. McKeown, "The controller placement problem," *ACM SIGCOMM-Computer Communication Review*, vol. 42, no. 4, pp. 473–478, 2012.

[3] B. S Khan and A. N. Muaz, "Network community detection: a review and visual survey," 2017, https://arxiv.org/abs/1708.00977.

[4] T. Das, V. Sridharan, and M. Gurusamy, "A survey on controller placement in SDN," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 472–503, 2019.

[5] G. Yao, J. Bi, Y. Li, and L. Guo, "On the capacitated controller placement problem in software defined networks," *IEEE Communications Letters*, vol. 18, no. 8, pp. 1339–1342, 2014.

[6] A. Sallahi and M. St-Hilaire, "Optimal model for the controller placement problem in software defined networks," *IEEE Communications Letters*, vol. 19, no. 1, pp. 30–33, 2014.

[7] H. Mu, A. Basta, A. Blenk, and W. Kellerer, "Modeling flow setup time for controller placement in SDN: evaluation for dynamic flows," in *Proceedings of the 2017 IEEE International Conference on Communications (ICC)*, pp. 1–7, IEEE, 2017.

[8] G. Gurobi, "Optimizer reference manual," 2015, http://www.gurobi.com.

[9] Z. Fan, J. Yao, X. Yang, Z. Wang, and X. Wan, "A multi-controller placement strategy based on delay and reliability optimization in sdn," in *Proceedings of the 2019 28th Wireless and Optical Communications Conference (WOCC)*, pp. 1–5, IEEE, Beijing, China, May 2019.

[10] D. B. Vincent, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10, Article ID P10008, 2008.

[11] V. A Traag, L Waltman, and N. J Van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, pp. 5233–5312, 2019.

[12] W. Chen, C. Chen, X. Jiang, and L. Liu, "Multi-controller placement towards sdn based on louvain heuristic algorithm," *IEEE Access*, vol. 6, pp. 49486–49497, 2018.

[13] U. N Raghavan, R Albert, and S Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, no. 3, Article ID 036106, 2007.

[14] B. Liu, B. Wang, and X. Xi, "Heuristics for sdn controller deployment using community detection algorithm," in *Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 253–258, IEEE, Beijing, August 2016.

[15] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit, "Local search heuristics for k-median and facility location problems," *SIAM Journal on Computing*, vol. 33, no. 3, pp. 544–562, 2004.

[16] F. Bannour, S. Souihi, and A. Mellouk, "Scalability and reliability aware sdn controller placement strategies," in *Proceedings of the 2017 13th International Conference on Network and Service Management (CNSM)*, pp. 1–4, IEEE, Tokyo, Japan, November 2017.

[17] A. Ksentini, M. Bagaa, T. Taleb, and I. Balasingham, "On using bargaining game for optimal placement of sdn controllers," in *Proceedings of the 2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, Kuala Lumpur, Malaysia, May 2016.

[18] M. Priyadarsini and P. Bera, "Software defined networking architecture, traffic management, security, and placement: a survey," *Computer Networks*, vol. 192, Article ID 108047, 2021.

[19] J. Lu, Z. Zhang, T. Hu, P. Yi, and J. Lan, "A survey of controller placement problem in software-defined networking," *IEEE Access*, vol. 7, pp. 24290–24307, 2019.

[20] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[21] L. C Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[22] L Lü, Y. C Zhang, C. H Yeung, and T Zhou, "Leaders in social networks, the delicious case," *PLoS One*, vol. 6, no. 6, Article ID e21202, 2011.

[23] Q. Yang, H. Deng, and L. Wang, "A cache strategy based on control of traffic load in the information-centric networking," *Journal of Network New Media*, vol. 10, no. 05, pp. 17–22, 2021.

[24] E. Jack and D. W Hearn, "Geometrical solutions for some minimax location problems," *Transportation Science*, vol. 6, no. 4, pp. 379–394, 1972.

[25] Snd lib, "A library of test instances," 2006, http://sndlib.zib.de/home.action.

[26] W. Wang, Y. Sun, K. Zheng, M. Kaalifar Ali, D. Li, and Z. Li, "Freeway: adaptively isolating the elephant and mice flows on different transmission paths," in *Proceedings of the 2014 IEEE 22nd International Conference on Network Protocols*, pp. 362–367, IEEE, Raleigh, NC, USA, October 2014.

Multirobot Adaptive Task Allocation of Intelligent Warehouse Based on Evolutionary Strategy

Debendra Kumar Sahoo, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, dksahoo312@gmail.com*

Pravas Behera, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, p_behera91@outlook.com*

Manoj Mohanta, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, manoj.mohanta62@outlook.com*

Kamal Samal, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, kamal_samal21@gmail.com*

Abstract

To solve the dynamic and real-time problem of multirobot task allocation in intelligent warehouse system under parts-to-picker mode, this paper presents a combined solution based on adaptive task pool strategy and Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) algorithm. In the first stage of the solution, a variable task pool is used to store dynamically added tasks, which can dynamically divide continuous and large-scale task allocation problems into small-scale subproblems to solve them to meet dynamic requirements. And an adaptive control strategy is used to automatically adjust the total number of tasks in the task pool to achieve a trade-off among throughput, energy consumption, and waiting time, which has better adaptability than manually adjusting the size of the task pool. In the second stage of the solution, when the task pool is full, tasks in the task pool will be assigned to robots. For the task allocation problem, this paper regards it as an optimization problem and uses the CMA-ES algorithm to find the optimal task assignment solution for all the robots. By comparing with fixed threshold method under 56 different task pool sizes, the experimental results show that the throughput can be close to reaching the optimal level, and the average distance traveled by robots to handle each unit is lower using adaptive threshold method; so, adaptive task pool solution has better adaptability and can find the optimal task pool size by itself. This method can satisfy the dynamic and real-time requirements and can be effectively applied to the intelligent warehouse system.

1. Introduction

In recent years, the orders of various e-commerce platforms have soared, and the scale of distribution centers has become increasingly large, which has brought great challenges to the traditional logistics industry [1]. In the traditional warehouse, 60% to 70% of the workers' time is spent on picking up goods [2], and the efficiency is extremely low. Therefore, more and more automatic machines and equipment have been applied in the field of warehouse [2]. Many companies have started to adopt a new kind of parts-to-picker intelligent warehouse system, such as Kiva system [3]. In the system as shown in Figure 1, robots transport the shelves from storage areas to workstations, and workers need to wait at the stations. When the shelves reach the workstations, they take the needed goods from the shelves or store bundles into the shelves. It has been proved that this kind of the intelligent warehouse system greatly saves labor cost and improves the efficiency of warehouse operation [4].

Cooperative control of multiple mobile robots is the key to realize intelligent warehousing. In a warehouse as shown in Figure 1, there are often numerous tasks such as replenishment and picking, as well as numerous robots to perform these tasks. In addition, the costs of different robots to perform a task are also different. Therefore, the efficiency of the warehouse is determined by selecting suitable robots to perform specific tasks. This is a typical multirobot task allocation (MRTA) problem [5]. With the operation of the warehouse, tasks and the warehouse environment will constantly change. How to find a better task allocation scheme for pick-task and replenishment-task assignment in such a highly dynamic environment [3, 4] is the focus of this paper.

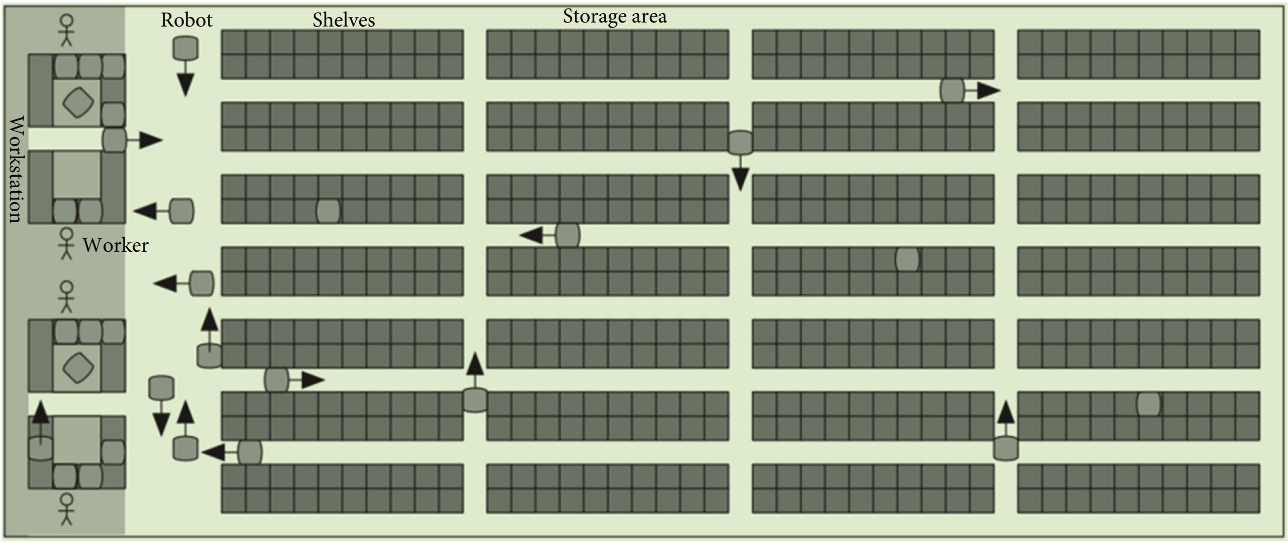


Figure 1: Parts-to-picker intelligent warehouse system from ref. [4].

MRTA is one of the most challenging problems in the multirobot system [6]. Market-based methods are the most studied methods at present, such as the single-task auction algorithm proposed in ref. [7]. In order to solve the problem that the single-task auction algorithm is difficult to get the optimal solution, a combined auction algorithm which considers the correlation between tasks was proposed in ref. [8]. When the number of robots and tasks is small, MRTA can be regarded as a zero-one integer linear programming problem and solved by simplex method, branch and bound method, Hungarian algorithm [9], etc. For example, the Hungarian algorithm was adopted in ref. [10] to solve the role assignment problem in robot soccer game. There are also some thresholding based methods such as ALLIANCE [11] and Broadcast of Local Eligibility (BLE) [12], which have good real-time, fault tolerance, and robustness, but usually only local optimal solution can be obtained. For large-scale problems, the heuristic algorithm can effectively reduce solution space and improve search efficiency. For example, in ref. [13], the heuristic algorithm was adopted to solve the task assignment problem in multi-core processor. Evolutionary algorithms are mature global optimization methods with high robustness and wide applicability, which can effectively deal with complex problems that are difficult to be solved by traditional optimization algorithms. Various evolutionary algorithms such as genetic algorithm and simulated annealing algorithm have been widely used in MRTA problem. In ref. [14], the genetic algorithm was used to solve the time-extended multirobot task allocation problem in the case of disaster. A hybrid genetic and ant colony algorithm was proposed in ref. [15] to improve the solving accuracy of the genetic algorithm. In ref. [16], the genetic algorithm was used to solve MRTA problem in the intelligent warehouse. Ref. [17] designed an improved quantum evolutionary algorithm based on the niche coevolution strategy and enhanced particle swarm optimization (IPOQEA) to solve the airport gate allocation problem. In ref. [18], an improved quantum-inspired cooperative coevolution algorithm with multistrategy is used to solve the knapsack problem and the actual airport gate allocation problem. Refs. [17–20] use the cooperative coevolution framework to divide the complex optimization problem into several subproblems, and these subproblems were solved by independent searching in order to improve the solution efficiency. Similarly, the situation where the number of tasks is variable in an intelligent warehouse can be studied using the idea of divide-and-conquer in Refs. [17–20].

Therefore, we use a task pool to store dynamically added tasks and propose an adaptive control strategy to automatically adjust the task pool size according to the current environment. When the task pool is full, the tasks in the pool will be assigned to the robots. Then, the task allocation problem is regarded as an optimization problem and solved by the CMA-ES algorithm [21].

2. Problem Formulation

The intelligent warehouse system consists of many movable shelves and robots as well as some workstations. The robots transport the needed shelves from the storage area to the workstations, and the workers can complete the replenishment and picking without moving. A typical intelligent warehouse layout (a screenshot from the open source software RAWSim-O [22]) is shown in Figure 2. In the figure, the four squares on the left represent the replenishment station, and the replenished bundles are temporarily stored here waiting for shelves. The four squares on the right represent picking stations. After receiving orders, the system will use a special algorithm to assign orders to different stations. There will be an upper limit on the number of orders in the stations [23]. The squares in the middle area are the shelves, in which the goods in the warehouse are stored. Shelves can be lifted and moved by robots. The circles in the figure are robots. A robot can carry a shelf to move. When a robot does not carry a shelf, it can move freely under the shelf.

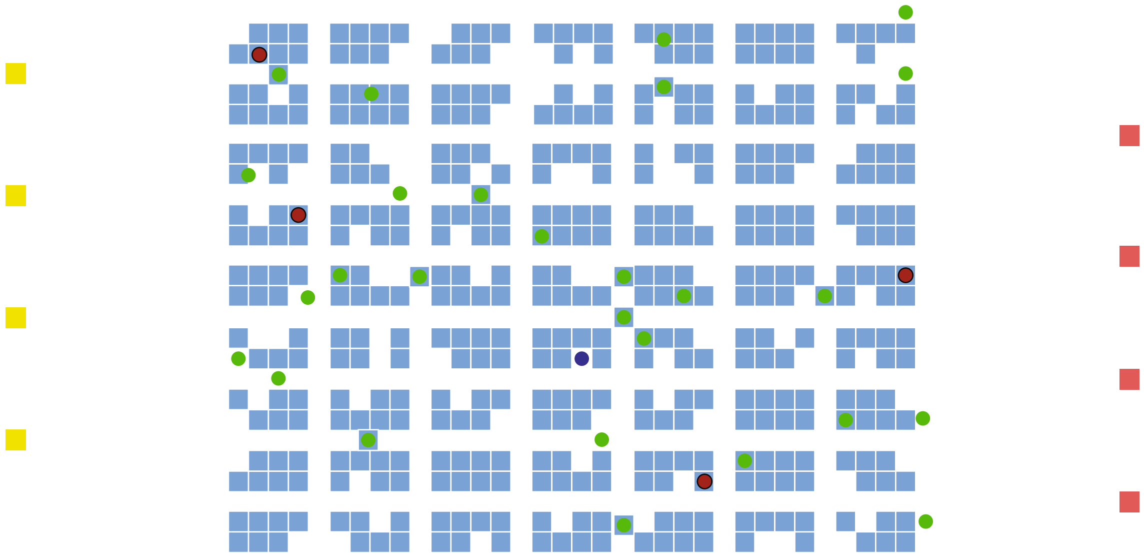


FIGURE 2: A typical intelligent warehouse layout from ref. [22].

In order to facilitate problem analysis, we make the following assumptions:

(1) Robots are all isomorphic and travel at exactly the same speed. They can only move forward, backward, left, and right.

(2) The time for a robot to lift a shelf and stay at a workstation is very short, which can be ignored.

(3) Every robot carries the required shelf and travels from the position of the shelf to the designated station and then carries the shelf back to its original location.

The shelf selection algorithm will select shelves for each workstation according to requirements. The selected shelves need to be transported from the shelf storage area to the appropriate station for picking up or replenishing goods, and then they are transported back to the original position, which is the task of the robots. If a robot is not assigned a task, it will move to a special resting area for rest. How to reasonably assign tasks to robots is the problem to be studied in this paper.

Referring to ref. [16], suppose that there are m tasks (refers to all tasks from the beginning to the end of the warehouse operation) and n robots in the warehouse, the set of tasks is $T = \{t_1, t_2, t_3, \cdots, t_m\}$, and the set of robots is $R = \{r_1, r_2, r_3 \cdots, r_n\}$. The set of tasks assigned to robot r_i is T_i, which is a subset of T. $T_1 \cup T_2 \cup T_3 \cup \cdots \cup T_n = T$ and $T_1 \cap T_2 \cap T_3 \cap \cdots \cap T_n = \varnothing$. Let $T_i = \{t_{i1}, t_{i2}, t_{i3}, \cdots, t_{ik}\}$ and T_i is ordered, and then the sequence of tasks to be completed by the robot r_i is $t_{i1} \longrightarrow t_{i2} \longrightarrow t_{i3} \longrightarrow \cdots \longrightarrow t_{ik}$. The cost of robot r to complete its task sequence can be expressed as

$$C(r_i) = I(r_i, t_{i1}) + \sum_{h=1}^{k} S(t_h) + \sum_{h=1}^{k-1} R(t_h, t_{h+1}), \qquad (1)$$

where $C(r_i)$ represents the cost of the robot r_i to complete all tasks. Since all robots travel at the same speed, the cost can be expressed as the distance traveled by the robot. The robot can only move forward, backward, left, and right; so, the distance traveled between the two points can be expressed as Manhattan distance.

$I(r_i, t_{i1})$ represents the cost for the robot to get from the initial position to the position of required shelf for the first task t_{i1}. Let the initial coordinate of the robot be (x_r, y_r) and the coordinate of the required shelf for the first task be (x_{t1}, y_{t1}), and then

$$I(r_i, t_{i1}) = |x_r - x_{t1}| + |y_r - y_{t1}|. \qquad (2)$$

$S(t_h)$ represents the cost for the robot to complete task t_h , which is only related to task t_h itself. It can be represented by the distance that after the robot carries the required shelf, it travels from the position of the required shelf for the task to the designated station and then returns to the shelf's original position from the station. Let the coordinate of required shelf for task t_h be (x_p, y_p) and the coordinate of target station be (x_s, y_s), and then

$$S(t_h) = \left(|x_p - x_s| + \left|y_p - y_s\right|\right) * 2. \qquad (3)$$

$R(t_h, t_{h+1})$ represents the cost for the robot to reach the starting position of the next task t_{h+1} after completing task t_h. Since the robot needs to transport the shelf back to the original position after completing task t_h, it can be directly represented by the Manhattan distance from the position of required shelf for task t_h to the position of required shelf for task t_{h+1}. Let the coordinate of required shelf for task t_h be (x_{p1}, y_{p1}) and the coordinate of required shelf for task t_{h+1} be (x_{p2}, y_{p2}), and then

$$R(t_h, t_{h+1}) = \left|x_{p1} - x_{p2}\right| + \left|y_{p1} - y_{p2}\right|. \qquad (4)$$

In order to make the overall allocation scheme as optimal as possible, we consider the following two optimization objectives:

(1) The maximum time taken by all robots to complete all tasks (C_{time})

(2) The mean distance traveled by all robots (C_{distance})

where

$$C_{\text{time}} = \max_i C(r_i), \qquad C_{\text{distance}} = \frac{\sum_{i=1}^{n} C(r_i)}{n}. \qquad (5)$$

C_{time} describes the efficiency of the robots to complete tasks. The smaller C_{time} is, the less time the robots take to complete all tasks, and the higher the efficiency is. C_{distance} describes the power consumption of the multirobot system. The smaller C_{distance} is, the shorter the total travel distance of all robots is, and the lower the power consumption is. The goal of the method studied in this paper is to reasonably assign all tasks in the system to all robots so that these two values can be as small as possible.

3. Method

3.1. Architecture. With the entry of new orders, new tasks are constantly generated and must be completed as soon as possible; so, the warehouse system is a highly dynamic and real-time system. In such a highly dynamic system, it is difficult to find the global optimal solution; so, the problem is divided into many subproblems. Specifically, we created a task pool P. When a new task is generated, it is immediately added to P. When the number of tasks in the task pool P reaches the threshold value (automatic adjustment of the threshold will be described in Section 3.3), the CMA-ES method in Section 3.2 is used to allocate the tasks in the task pool to robots. The robots insert the new task sequence allocated into the rear of the previous unfinished task sequence, and then the task pool is emptied. The robots execute tasks according to their own task sequence, and the executed tasks are deleted from the sequence. As the new tasks are generated again, the tasks are added to P again. Loop until the warehouse stops running. In Figure 3, the specific steps are as follows:

Step 1. Initialize the task pool size and set the task pool P to be empty. For all robots, initialize task sequence T_i of every robot r_i.

Step 2. The threshold of the task pool size is automatically adjusted using adaptive control strategy in Section 3.3.

Step 3. New tasks are constantly added to P. Jump to step 4 when the number of tasks in the task pool reaches the threshold.

Step 4. The tasks in the task pool are assigned to the robots using the CMA-ES method in Section 3.2, and for all robots, the new task sequence assigned to robot r_i is inserted at the end of the current task sequence T_i.

Step 5. Clear the task pool P and jump to step 2.

The above solution in Figure 3 is executed by the central controller, and the robot only needs to execute the tasks according to the assigned task sequence. The parallel operation of the two parts enables the robots to be busy all the time, which saves time and meets the requirement of real-time storage system.

3.2. CMA-ES Algorithm. As mentioned in Section 3.1, tasks are assigned to robots when the number of tasks in the task pool reaches the threshold. This problem is regarded as an optimization problem in a static environment. This is a NP-hard problem, and the CMA-ES algorithm is used to find the optimal solution. The successful application in many fields [24–26] proves that the CMA-ES algorithm is a good search algorithm.

3.2.1. Representation of Solutions. Referring to ref. [27], for the task allocation problem with m tasks and n robots, a candidate to represent a task assignment scheme is $X = [x_1, x_2, x_3 \cdots x_m]$. X contains m real numbers, and for each real number x_i, it satisfies $1 \le x_i < n + 1, i = 1, 2, 3, \cdots, m$, where x_i means task i is performed by robot $\text{Int}(x_i)$, and $\text{Int}(x_i)$ means the integer of real number x_i. If $\text{Int}(x_i) = \text{Int}(x_j), i \neq j$, this means that the task x_i and x_j are both assigned to the same robot, and the task represented by the smaller number between x_i and x_j is executed first. If $x_i = x_j$, the execution order of these two tasks is determined randomly.

For example, there are 8 tasks (represented by numbers 1, 2, 3,..., 8) and 3 robots (represented by numbers 1, 2, 3), and an individual [1.7, 3.8, 2.2, 1.3, 2.8, 1.5, 3.3, 3.7] is generated. Then, the task sequence assigned to robot 1 is $4 \longrightarrow 6 \longrightarrow 1$. The task sequence assigned to robot 2 is $3 \longrightarrow 5$. The task sequence assigned to robot 3 is $7 \longrightarrow 8 \longrightarrow 2$.

3.2.2. Fitness Function. Fitness function is used to evaluate candidates. For the CMA-ES algorithm, individuals with lower fitness value are more excellent. In Section 2, two optimization goals are proposed for the whole system: one is the time C_{time} for the robots to complete all tasks; the second is the mean driving distance C_{distance} of all robots. Each planning can be regarded as a subproblem of the whole. For each subproblem, in order to achieve the optimal overall performance, these two goals are still considered; so, fitness function f is calculated through the following equation [16]:

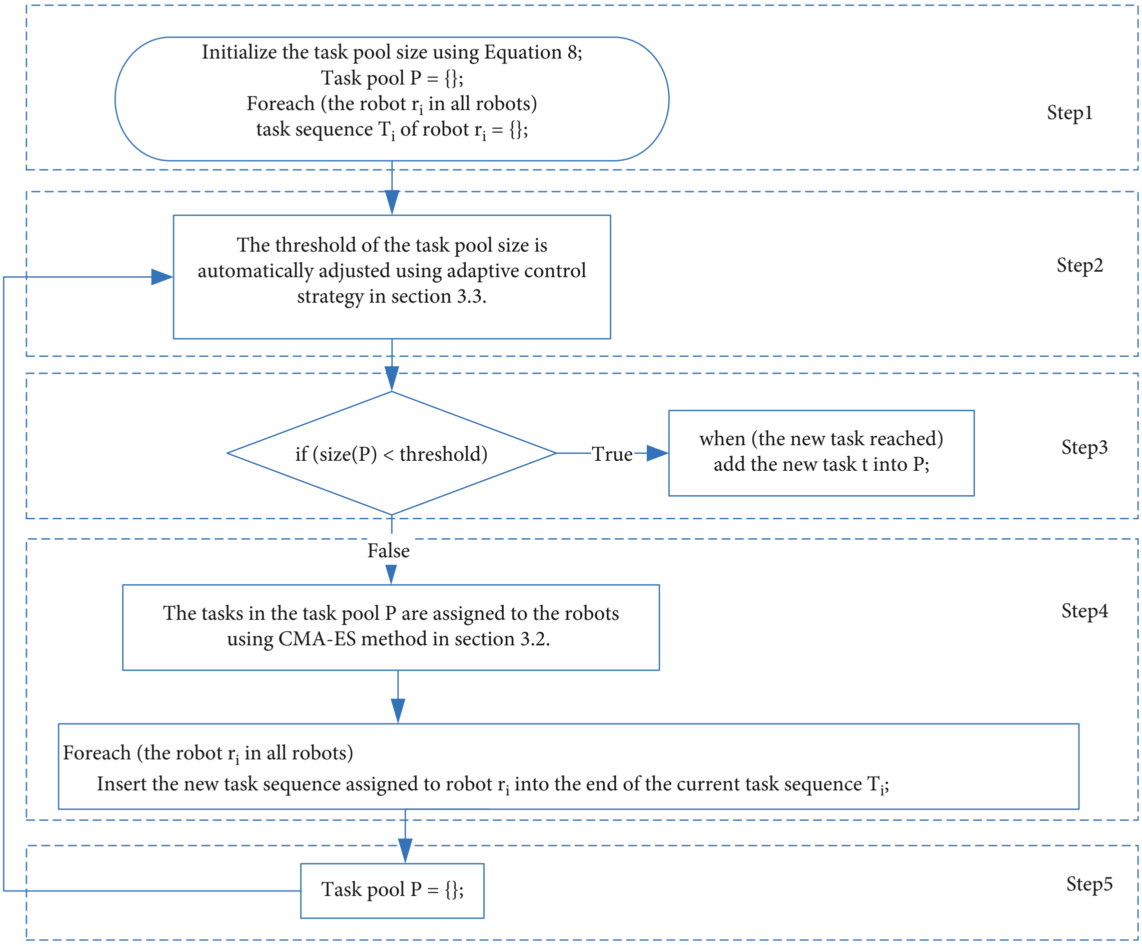

Initialize the task pool size using Equation 8;
Task pool P = {};
Foreach (the robot r_i in all robots)
task sequence T_i of robot r_i = {};
Step1
The threshold of the task pool size is automatically adjusted using adaptive control strategy in section 3.3.
Step2
if (size(P) < threshold)
True
when (the new task reached) add the new task t into P;
Step3
False
The tasks in the task pool P are assigned to the robots using CMA-ES method in section 3.2.
Step4
Foreach (the robot r_i in all robots)
Insert the new task sequence assigned to robot r_i into the end of the current task sequence T_i;
Task pool P = {};
Step5

Figure 3: The flow chart of the combined solution based on adaptive task pool strategy and CMA-ES.

$$
\begin{gathered}
f = \alpha C'_{\text{time}} + (1-\alpha) C'_{\text{distance}}, \quad 0 \le \alpha \le 1, \\
C'_{\text{time}} = \max_{i} C'(r_i), \\
C'_{\text{distance}} = \frac{\sum_{i=1}^{n} C'(r_i)}{\mathrm{n}},
\end{gathered} \tag{6}
$$

where α is a constant that can be adjusted according to the actual demand. If more attention is paid to the completion time of a single order, α can be increased. If more attention is paid to the energy consumption of all robots, α can be reduced. $C'(r_i)$ is the cost of robot r_i to execute the tasks in the current task sequence first and then execute the tasks according to the candidate. C'_{time} is the maximum time taken by the robots. C'_{distance} is the mean distance traveled by all robots. In the current moment, there may be unfinished tasks in the task sequence. The robot must first complete these tasks before performing the tasks assigned at the current moment. Therefore, for $C'(r_i)$, we divide it into two parts to calculate:

$$
C'(r_i) = C'_1(r_i) + C'_2(r_i), \tag{7}
$$

where $C'_1(r_i)$ is the cost for the robot to complete the tasks in the current task sequence, and $C'_2(r_i)$ is the cost for the robot to execute the tasks according to the candidate. $C'_1(r_i)$ and $C'_2(r_i)$ are represented by the distance traveled by the robot and calculated using the method described in Equation (1).

With this fitness function, we try to find the optimal solution at that moment in each optimization and try to approximate the global optimal solution by this method.

3.3. Automatic Adjustment of Task Pool. When the number of tasks in the task pool reaches the threshold, the tasks in the task pool will be assigned to the robots. The threshold plays a decisive role in the efficiency of assignment. The larger the threshold is, the more tasks will be involved in the optimization, and then the more the planned scheme will be close to the global optimal solution. If an optimization contains all the tasks in the system, the optimal solution found by the optimization will be the optimal solution of the whole system. But orders in the warehouse are added dynamically over time, so tasks are also generated dynamically. As the threshold increases, the time required for the task pool to be filled will also increase, and this situation will occur: the robot has finished all the tasks assigned to it, but the number of tasks in the task pool has not reached the threshold; so, the next optimization cannot start, and the robot can only wait. This leads to a waste of time and cannot meet the real-time of the warehouse system. Moreover,

```
Input: lastAdjustTime, currentTime, lastTasksCompleted, tasksCompleted, oldThreshold, lastAction
Output: newThreshold, lastAction
1: if $currentTime - lastAdjustTime > I$ then
2:   if $tasksCompleted = 0$ then
3:     $newThreshold \longleftarrow oldThreshold/2$
4:     $lastAction \longleftarrow -1$
5:   else if $tasksCompleted - lastTaskCompleted \geq 0$ then
6:     $newThreshold \longleftarrow newThreshold + lastAction$
7:   else
8:     $newThreshold \longleftarrow newThreshold - lastAction$
9:     $lastAction \longleftarrow -lastAction$
10: else
11:   $newThreshold \longleftarrow oldThreshold$
12: return newThreshold, lastAction
```

Algorithm 1: Adaptive control strategy.

because each workstation has an order capacity limit, there is also an upper limit on the total number of tasks in the system, and if the task pool size exceeds this upper limit, the number of tasks in the task pool will never reach the threshold, and the system will be stagnant. Therefore, it is very important to set a threshold of appropriate size.

Obviously, for different warehouses, the threshold should be set differently depending on the actual situation. Even for the same warehouse, the number of robots may be adjusted, and the rate of order generation may vary at different times; so, it is not appropriate to set the threshold to a fixed value. Therefore, we design an adaptive control strategy to dynamically adjust the task pool, as shown in Algorithm 1.

First, the setting of the initial threshold is important, which determines the speed of finding the optimal threshold. We believe that the size of the initial threshold should be related to the number of robots and the upper limit number of tasks in the warehouse. The upper limit number of tasks in the warehouse is related to the number of workstations and the capacity of each workstation. So, we propose the following heuristic formula to calculate the initial threshold:

$$\text{initialThreshold} = \frac{(\gamma * \text{stations} + \text{robots})}{2}, \qquad (8)$$

where γ is a constant representing the average number of tasks per workstation in unit time, which is set according to the actual situation. stations is the number of stations, and robots is the number of robots. We set a time interval I (It is a constant that can be set according to actual requirements), and every I seconds, the threshold is adjusted (line 1). lastAction is used to record the last adjustment. We counted the total number of tasks completed by the robot from the last adjusted moment to the current moment, and the total number of tasks completed from the penultimate adjusted moment to the last adjusted moment, expressed by tasksCompleted and lastTasksCompleted, respectively. If taskCompleted is 0, indicating that the threshold has been set so high that the number of tasks has not reached the threshold, then simply cut the threshold in half and set lastAction to −1 (line 2, line 3, and line 4). If tasksCompleted is greater than or equal to lastTasksCompleted, it indicates that the last adjustment has had a positive effect on the system, and the same adjustment will be performed (line 5 and line 6). If tasksCompleted is less than lastTasksCompleted, it indicates that the last adjustment had a negative effect on the system, and the reverse adjustment will be performed (line 7 and line 8). In addition, lastAction will be reversed (line 9).

4. Experiments

We used RAWSim-O [22], an open source framework developed by Merschformann et al., as the experimental platform. RAWSim-O is a simulation framework that simulates the operation of an intelligent warehouse system and allows us to test our own methods.

We used the warehouse layout shown in Figure 2. In the warehouse layout, there are 32 robots and 550 shelves. The storage positions of the shelves are at the middle area of the layout. And there are four replenishment stations on the left and four picking stations on the right. To simplify the problem, we set the duration of a robot staying at a workstation to a very small value of 0.1.

For the assessment of performance we take the sum of SKUs (stock keeping unit) in both item bundles stored at the replenishment stations and orders picked at the picking stations as handled units. This represents the throughput of the warehouse, and the higher the better. We also look at the average distance traveled by robots to handle each unit. This can represent the power consumption of the multirobot system.

In order to test the impact of task pool threshold size on the allocation effect, we did 56 experiments, each experiment corresponding to different pool sizes. Each experiment was simulated for 24 hours with 10 repetitions.

Under different task pool sizes, the number of units handled by robots is shown in the blue solid line in Figure 4, and the average distance traveled by robots to handle each unit is shown in the blue solid line in Figure 5. The comparison results among different fixed threshold on handled units and travel distance per unit are shown in Table 1. The maximum number of handled units is 207583 when the fixed threshold is set to 18. The minimum number of travel

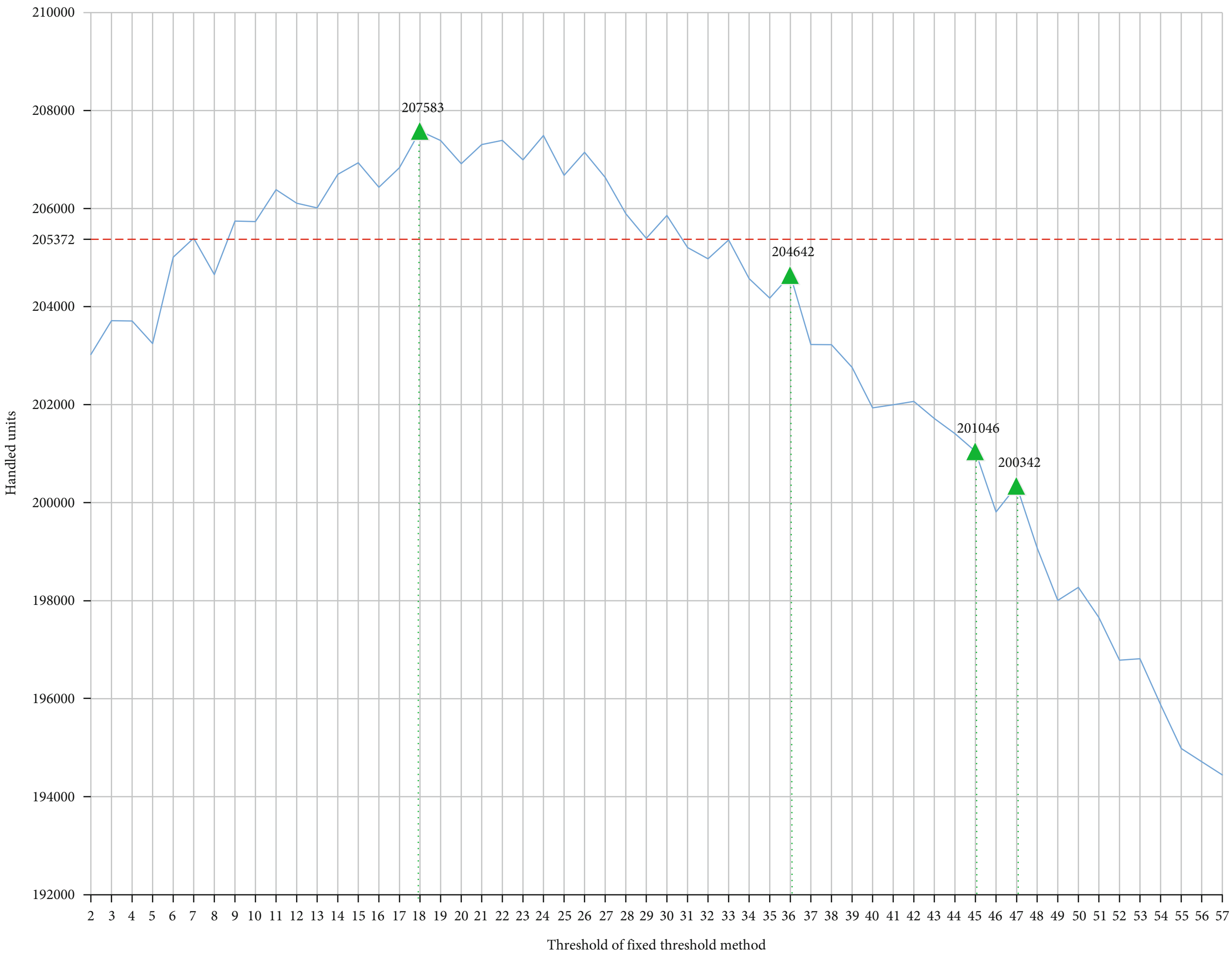


FIGURE 4: Comparison between adaptive threshold method and fixed threshold method on handled units. The red dotted line is the adaptive threshold method, and the blue solid line is the fixed threshold method.

distance per unit is 10.73 when the fixed threshold is set to 36, 45, or 47. According to Figures 4 and 5 and Table 1, it is not good to set the threshold too large or too small, which is consistent with our conjecture. If the threshold is set too small, the solution will be too far away from the global optimal solution; therefore, the number of handled units is small, and the travel distance per unit is large. If the threshold is set too large, the solution will be closer to the global optimal solution; so, the travel distance per unit is small, but the robot will have a long waiting time; therefore, the number of handled units will be small.

To sum up, a bad threshold can be very inefficient; so, setting the threshold manually is very risky. Therefore, a method of automatically adjusting threshold is necessary. We used the adaptive control strategy proposed by ourselves to conduct the experiment again, and all conditions were identical except the threshold. According to the workstation capacity, γ in Equation (8) was set to 4; so, the initial threshold was calculated as 32. The results are shown in Table 1. We compared the results with the fixed threshold approach, as shown in Figures 4 and 5. The red dotted line is the adaptive threshold method, and the blue solid line is the fixed threshold method. Compared with fixed threshold 18, the adaptive threshold method gets worse result in handled units but better result in travel distance per unit. Compared with fixed threshold 36, 45, and 47, the adaptive threshold method gets better result in handled units but worse result in travel distance per unit. Taken together, it can be seen from the two figures that the adaptive threshold method can be close to reaching the level when the threshold is set to the optimal in both indexes. The experimental results show that the proposed adaptive control strategy has good application effect.

5. Conclusion

In order to solve the dynamic and real-time problem of multirobot task allocation in the intelligent warehouse system, a combined solution based on adaptive task pool strategy and CMA-ES algorithm is proposed in the paper. In the early

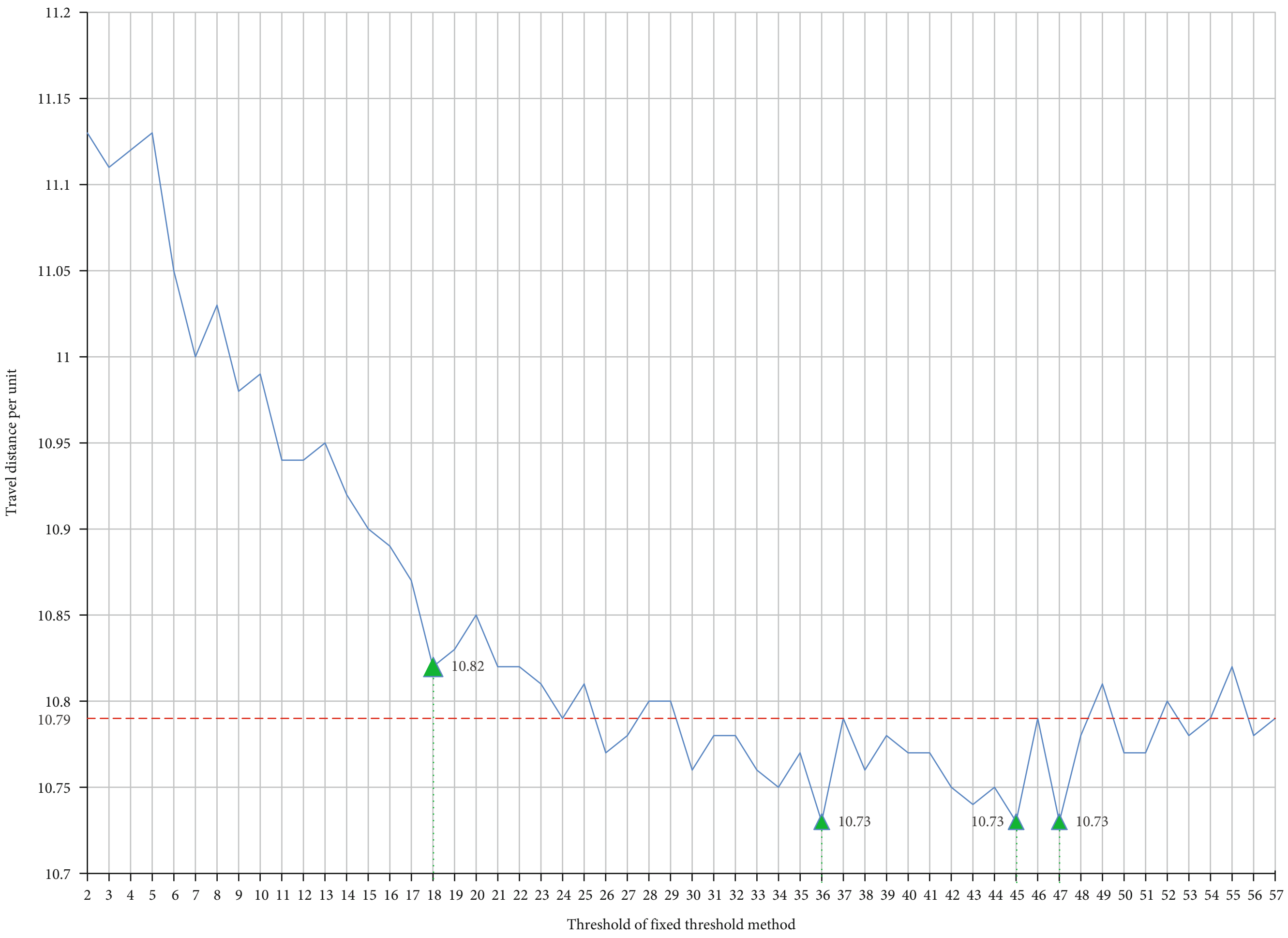


Figure 5: Comparison between adaptive threshold method and fixed threshold method on travel distance per unit. The red dotted line is the adaptive threshold method, and the blue solid line is the fixed threshold method.

Table 1: Comparison between adaptive threshold method and fixed threshold method on handled units and travel distance per unit.

Method (initial threshold)	Handled units	Travel distance per unit
Fixed threshold (18)	207583	10.82
Fixed threshold (36)	204642	10.73
Fixed threshold (45)	201046	10.73
Fixed threshold (47)	200342	10.73
Adaptive threshold (32)	205372	10.79

stage of the solution, the divide-to-conquer idea is used to design a variable task pool that is used to store dynamically added tasks. The variable task pool is designed to dynamically divide continuous and large-scale task allocation problems into small-scale subproblems to solve them to meet dynamic requirements. And an adaptive control strategy is used to automatically adjust the threshold of the task pool size in real time to achieve a trade-off among throughtput, energy consumption, and waiting time, which has better adaptability than manually adjusting the size of the task pool. In the later stage of the solution, when the task pool is full, tasks in the task pool will be assigned to robots using the CMA-ES algorithm to find the optimal task assignment solution for all the robots according to the fitness function including the maximum time and the mean travel distance required by all robots to complete all the tasks. By comparing with fixed threshold method under 56 different task pool sizes, the experimental results show that the handled units can be close to reaching the optimal level, and the average travel distance per unit is lower using adaptive threshold method; so, adaptive threshold solution indeed has better adaptability. This method can satisfy the dynamic and real-time requirements and can be effectively applied to the intelligent warehouse system.

However, because of the complexity and dynamics of the warehouse environment, it may not be accurate to measure the cost by Manhattan distance. Therefore, how to introduce accurate robot motion model to evaluate the cost will be the next work. Furthermore, the relationships among handled units, travel distance per unit, the maximum time taken by all robots to complete all tasks, and the mean distance traveled by all robots need further study. In addition, the effect of communication quality on allocation is not taken into account and will be deeply studied.

References

[1] M. Zhou and M. Y. Wang, "Analysis on the development of e-commerce logistics service industry and countermeasures," *Computer and Information Technology*, vol. 20, no. 6, pp. 10–12, 2012.

[2] S. X. Zou, "The present and future of warehouse robot," *Logistics Engineering and Management*, vol. 35, no. 6, pp. 171-172, 2013.

[3] J. J. Enright and P. R. Wurman, "Optimization and coordinated autonomy in mobile fulfillment systems," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 33–38, San Francisco, California, 2011.

[4] P. R. Wurman, R. D'Andrea, and M. Mountz, "Coordinating hundreds of cooperative, autonomous vehicles in warehouses," *AI Magazine*, vol. 29, no. 1, p. 9, 2008.

[5] B. P. Gerkey and M. J. Matarić, "A formal analysis and taxonomy of task allocation in multi-robot systems," *International Journal of Robotics Research*, vol. 23, no. 9, pp. 939–954, 2004.

[6] A. Khamis, A. Hussein, and A. Elmogy, "Multi-robot task allocation: a review of the state-of-the-art," *Eds. Cham: Springer International Publishing*, vol. 604, pp. 31–51, 2015.

[7] B. P. Gerkey and M. J. Matarić, "Sold!: auction methods for multirobot coordination," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 5, pp. 758–768, 2002.

[8] M. Berhault, H. Huang, P. Keskinocak et al., "Robot Exploration with Combinatorial Auctions," *In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1957–1962, 2003.

[9] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[10] P. MacAlpine, E. Price, and P. Stone, "SCRAM: scalable collision-avoiding role assignment with minimal-makespan for formational positioning," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2096–2102, Austin, Texas, USA, 2015.

[11] L. E. Parker, "ALLIANCE: an architecture for fault tolerant multirobot cooperation," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 2, pp. 220–240, 1998.

[12] B. B. Werger and M. J. Mataric, "Broadcast of local eligibility: behavior-based control for strongly cooperative robot teams," in *Proceedings of the 4th International Conference on Autonomous Agents*, pp. 21-22, Barcelona, Spain, 2000.

[13] Y. Liu, X. Zhang, H. Li, and D. Qian, "Allocating tasks in multi-core processor based parallel system," in *2007 IFIP International Conference on Network and Parallel Computing Workshops*, pp. 748–753, Liaoning, China, 2007.

[14] E. G. Jones, M. B. Dias, and A. Stentz, "Time-extended multi-robot coordination for domains with intra-path constraints," *Autonomous Robots*, vol. 30, no. 1, pp. 41–56, 2011.

[15] J. Zhang and Y. Q. Cao, "Research on dynamic task allocation for MAS based on hybrid genetic and ant colony algorithm," *Computer Science*, vol. 38, no. S1, pp. 268–270, 2011.

[16] J. J. Dou, C. L. Chen, and P. Yang, "Genetic scheduling and reinforcement learning in multirobot systems for intelligent warehouses," *Mathematical Problems in Engineering*, vol. 2015, 10 pages, 2015.

[17] W. Deng, J. Xu, H. Zhao, and Y. Song, "A novel gate resource allocation method using improved PSO-based QEA," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1737–1745, 2022.

[18] X. Cai, H. Zhao, S. Shang et al., "An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application," *Expert Systems with Applications*, vol. 171, article 114629, 2021.

[19] W. Deng, J. J. Xu, X. Z. Gao, and H. M. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 3, pp. 1578–1587, 2022.

[20] W. Deng, S. Shang, X. Cai et al., "Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization," *Knowledge-Based Systems*, vol. 224, article 107080, 2021.

[21] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)," *Evolutionary Computation*, vol. 11, no. 1, pp. 1–18, 2003.

[22] M. Merschformann, L. Xie, and H. Li, "RAWSim-O: a simulation framework for robotic mobile fulfillment systems," *Logistics Research*, vol. 11, no. 8, pp. 1–11, 2018.

[23] L. Xie, N. Thieme, R. Krenzler, and H. Y. Li, *Efficient Order Picking Methods in Robotic Mobile Fulfillment Systems*, 2019, https://arxiv.org/abs/1902.03092.

[24] F. Stulp and O. Sigaud, "Path integral policy improvement with covariance matrix adaptation," in *29th International Conference on Machine Learning*, Edinburgh, Scotland, 2012.

[25] T. Geijtenbeek, M. Van De Panne, and A. F. Van Der Stappen, "Flexible muscle-based locomotion for bipedal creatures," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 1–11, 2013.

[26] P. MacAlpine and P. Stone, "Overlapping layered learning," *Artificial Intelligence*, vol. 254, pp. 21–43, 2018.

[27] H. R. Zhou, W. S. Tang, and H. L. Wang, "Optimization of multiple traveling salesman problem based on differential evolution algorithm," *Systems Engineering Theory & Practice*, vol. 30, no. 8, pp. 1471–1476, 2010.

Multirobot Adaptive Task Allocation of Intelligent Warehouse Based on Evolutionary Strategy

Subhasish Mohanty, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, subhasish.mohanty@hotmail.com*

Sudipt Mandal, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, sudiptmandal@hotmail.com*

Bikash Ranjan Dash, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, br_dash214@gmail.com*

Srikanta Pradhan, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, s.pradhan91@gmail.com*

Abstract

To solve the dynamic and real-time problem of multirobot task allocation in intelligent warehouse system under parts-to-picker mode, this paper presents a combined solution based on adaptive task pool strategy and Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) algorithm. In the first stage of the solution, a variable task pool is used to store dynamically added tasks, which can dynamically divide continuous and large-scale task allocation problems into small-scale subproblems to solve them to meet dynamic requirements. And an adaptive control strategy is used to automatically adjust the total number of tasks in the task pool to achieve a trade-off among throughput, energy consumption, and waiting time, which has better adaptability than manually adjusting the size of the task pool. In the second stage of the solution, when the task pool is full, tasks in the task pool will be assigned to robots. For the task allocation problem, this paper regards it as an optimization problem and uses the CMA-ES algorithm to find the optimal task assignment solution for all the robots. By comparing with fixed threshold method under 56 different task pool sizes, the experimental results show that the throughput can be close to reaching the optimal level, and the average distance traveled by robots to handle each unit is lower using adaptive threshold method; so, adaptive task pool solution has better adaptability and can find the optimal task pool size by itself. This method can satisfy the dynamic and real-time requirements and can be effectively applied to the intelligent warehouse system.

1. Introduction

In recent years, the orders of various e-commerce platforms have soared, and the scale of distribution centers has become increasingly large, which has brought great challenges to the traditional logistics industry [1]. In the traditional warehouse, 60% to 70% of the workers' time is spent on picking up goods [2], and the efficiency is extremely low. Therefore, more and more automatic machines and equipment have been applied in the field of warehouse [2]. Many companies have started to adopt a new kind of parts-to-picker intelligent warehouse system, such as Kiva system [3]. In the system as shown in Figure 1, robots transport the shelves from storage areas to workstations, and workers need to wait at the stations. When the shelves reach the workstations, they take the needed goods from the shelves or store bundles into the shelves. It has been proved that this kind of the intelligent warehouse system greatly saves labor cost and improves the efficiency of warehouse operation [4].

Cooperative control of multiple mobile robots is the key to realize intelligent warehousing. In a warehouse as shown in Figure 1, there are often numerous tasks such as replenishment and picking, as well as numerous robots to perform these tasks. In addition, the costs of different robots to perform a task are also different. Therefore, the efficiency of the warehouse is determined by selecting suitable robots to perform specific tasks. This is a typical multirobot task allocation (MRTA) problem [5]. With the operation of the warehouse, tasks and the warehouse environment will constantly change. How to find a better task allocation scheme for pick-task and replenishment-task assignment in such a highly dynamic environment [3, 4] is the focus of this paper.

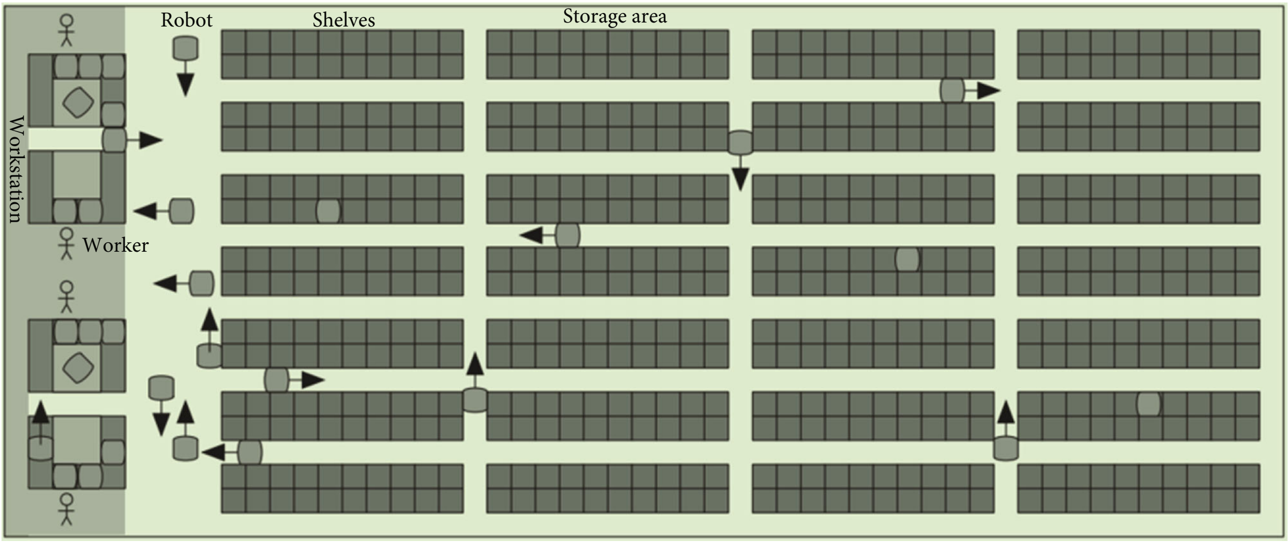


Figure 1: Parts-to-picker intelligent warehouse system from ref. [4].

MRTA is one of the most challenging problems in the multirobot system [6]. Market-based methods are the most studied methods at present, such as the single-task auction algorithm proposed in ref. [7]. In order to solve the problem that the single-task auction algorithm is difficult to get the optimal solution, a combined auction algorithm which considers the correlation between tasks was proposed in ref. [8]. When the number of robots and tasks is small, MRTA can be regarded as a zero-one integer linear programming problem and solved by simplex method, branch and bound method, Hungarian algorithm [9], etc. For example, the Hungarian algorithm was adopted in ref. [10] to solve the role assignment problem in robot soccer game. There are also some thresholding based methods such as ALLIANCE [11] and Broadcast of Local Eligibility (BLE) [12], which have good real-time, fault tolerance, and robustness, but usually only local optimal solution can be obtained. For large-scale problems, the heuristic algorithm can effectively reduce solution space and improve search efficiency. For example, in ref. [13], the heuristic algorithm was adopted to solve the task assignment problem in multi-core processor. Evolutionary algorithms are mature global optimization methods with high robustness and wide applicability, which can effectively deal with complex problems that are difficult to be solved by traditional optimization algorithms. Various evolutionary algorithms such as genetic algorithm and simulated annealing algorithm have been widely used in MRTA problem. In ref. [14], the genetic algorithm was used to solve the time-extended multirobot task allocation problem in the case of disaster. A hybrid genetic and ant colony algorithm was proposed in ref. [15] to improve the solving accuracy of the genetic algorithm. In ref. [16], the genetic algorithm was used to solve MRTA problem in the intelligent warehouse. Ref. [17] designed an improved quantum evolutionary algorithm based on the niche coevolution strategy and enhanced particle swarm optimization (IPOQEA) to solve the airport gate allocation problem. In ref. [18], an improved quantum-inspired cooperative coevolution algorithm with multistrategy is used to solve the knapsack problem and the actual airport gate allocation problem. Refs. [17–20] use the cooperative coevolution framework to divide the complex optimization problem into several subproblems, and these subproblems were solved by independent searching in order to improve the solution efficiency. Similarly, the situation where the number of tasks is variable in an intelligent warehouse can be studied using the idea of divide-and-conquer in Refs. [17–20].

Therefore, we use a task pool to store dynamically added tasks and propose an adaptive control strategy to automatically adjust the task pool size according to the current environment. When the task pool is full, the tasks in the pool will be assigned to the robots. Then, the task allocation problem is regarded as an optimization problem and solved by the CMA-ES algorithm [21].

2. Problem Formulation

The intelligent warehouse system consists of many movable shelves and robots as well as some workstations. The robots transport the needed shelves from the storage area to the workstations, and the workers can complete the replenishment and picking without moving. A typical intelligent warehouse layout (a screenshot from the open source software RAWSim-O [22]) is shown in Figure 2. In the figure, the four squares on the left represent the replenishment station, and the replenished bundles are temporarily stored here waiting for shelves. The four squares on the right represent picking stations. After receiving orders, the system will use a special algorithm to assign orders to different stations. There will be an upper limit on the number of orders in the stations [23]. The squares in the middle area are the shelves, in which the goods in the warehouse are stored. Shelves can be lifted and moved by robots. The circles in the figure are robots. A robot can carry a shelf to move. When a robot does not carry a shelf, it can move freely under the shelf.

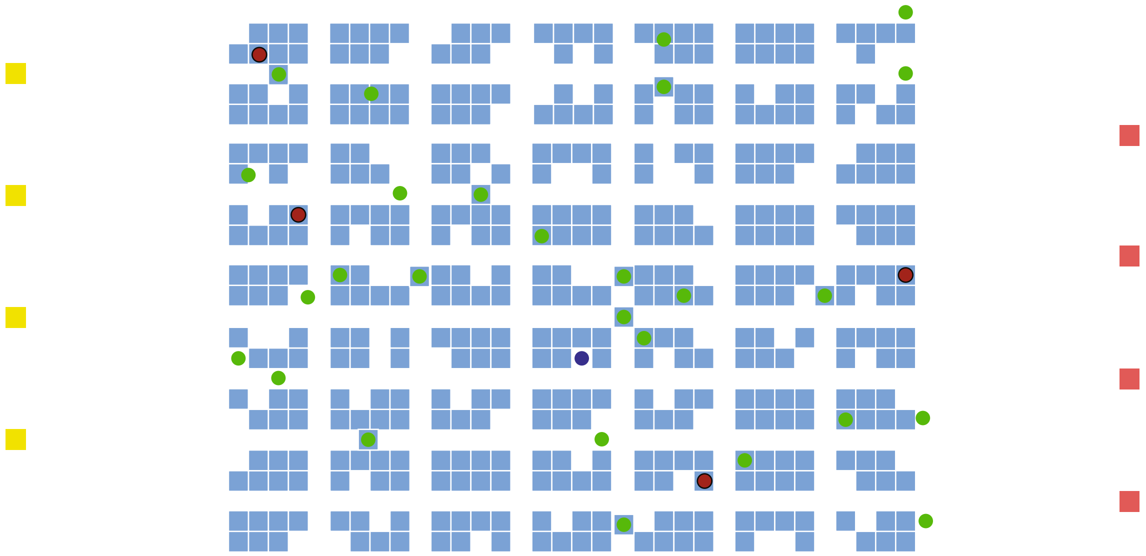


Figure 2: A typical intelligent warehouse layout from ref. [22].

In order to facilitate problem analysis, we make the following assumptions:

(1) Robots are all isomorphic and travel at exactly the same speed. They can only move forward, backward, left, and right.

(2) The time for a robot to lift a shelf and stay at a workstation is very short, which can be ignored.

(3) Every robot carries the required shelf and travels from the position of the shelf to the designated station and then carries the shelf back to its original location.

The shelf selection algorithm will select shelves for each workstation according to requirements. The selected shelves need to be transported from the shelf storage area to the appropriate station for picking up or replenishing goods, and then they are transported back to the original position, which is the task of the robots. If a robot is not assigned a task, it will move to a special resting area for rest. How to reasonably assign tasks to robots is the problem to be studied in this paper.

Referring to ref. [16], suppose that there are m tasks (refers to all tasks from the beginning to the end of the warehouse operation) and n robots in the warehouse, the set of tasks is $T = \{t_1, t_2, t_3, \cdots, t_m\}$, and the set of robots is $R = \{r_1, r_2, r_3 \cdots, r_n\}$. The set of tasks assigned to robot r_i is T_i, which is a subset of T. $T_1 \cup T_2 \cup T_3 \cup \cdots \cup T_n = T$ and $T_1 \cap T_2 \cap T_3 \cap \cdots \cap T_n = \varnothing$. Let $T_i = \{t_{i1}, t_{i2}, t_{i3}, \cdots, t_{ik}\}$ and T_i is ordered, and then the sequence of tasks to be completed by the robot r_i is $t_{i1} \longrightarrow t_{i2} \longrightarrow t_{i3} \longrightarrow \cdots \longrightarrow t_{ik}$. The cost of robot r to complete its task sequence can be expressed as

$$C(r_i) = I(r_i, t_{i1}) + \sum_{h=1}^{k} S(t_h) + \sum_{h=1}^{k-1} R(t_h, t_{h+1}), \qquad (1)$$

where $C(r_i)$ represents the cost of the robot r_i to complete all tasks. Since all robots travel at the same speed, the cost can be expressed as the distance traveled by the robot. The robot can only move forward, backward, left, and right; so, the distance traveled between the two points can be expressed as Manhattan distance.

$I(r_i, t_{i1})$ represents the cost for the robot to get from the initial position to the position of required shelf for the first task t_{i1}. Let the initial coordinate of the robot be (x_r, y_r) and the coordinate of the required shelf for the first task be (x_{t1}, y_{t1}), and then

$$I(r_i, t_{i1}) = |x_r - x_{t1}| + |y_r - y_{t1}|. \qquad (2)$$

$S(t_h)$ represents the cost for the robot to complete task t_h , which is only related to task t_h itself. It can be represented by the distance that after the robot carries the required shelf, it travels from the position of the required shelf for the task to the designated station and then returns to the shelf's original position from the station. Let the coordinate of required shelf for task t_h be (x_p, y_p) and the coordinate of target station be (x_s, y_s), and then

$$S(t_h) = \left(\left|x_p - x_s\right| + \left|y_p - y_s\right|\right) * 2. \qquad (3)$$

$R(t_h, t_{h+1})$ represents the cost for the robot to reach the starting position of the next task t_{h+1} after completing task t_h. Since the robot needs to transport the shelf back to the original position after completing task t_h, it can be directly represented by the Manhattan distance from the position of required shelf for task t_h to the position of required shelf for task t_{h+1}. Let the coordinate of required shelf for task t_h be (x_{p1}, y_{p1}) and the coordinate of required shelf for task t_{h+1} be (x_{p2}, y_{p2}), and then

$$R(t_h, t_{h+1}) = \left|x_{p1} - x_{p2}\right| + \left|y_{p1} - y_{p2}\right|. \tag{4}$$

In order to make the overall allocation scheme as optimal as possible, we consider the following two optimization objectives:

(1) The maximum time taken by all robots to complete all tasks (C_{time})

(2) The mean distance traveled by all robots (C_{distance})

where

$$C_{\text{time}} = \max_i C(r_i), \qquad C_{\text{distance}} = \frac{\sum_{i=1}^{n} C(r_i)}{n}. \tag{5}$$

C_{time} describes the efficiency of the robots to complete tasks. The smaller C_{time} is, the less time the robots take to complete all tasks, and the higher the efficiency is. C_{distance} describes the power consumption of the multirobot system. The smaller C_{distance} is, the shorter the total travel distance of all robots is, and the lower the power consumption is. The goal of the method studied in this paper is to reasonably assign all tasks in the system to all robots so that these two values can be as small as possible.

3. Method

3.1. Architecture. With the entry of new orders, new tasks are constantly generated and must be completed as soon as possible; so, the warehouse system is a highly dynamic and real-time system. In such a highly dynamic system, it is difficult to find the global optimal solution; so, the problem is divided into many subproblems. Specifically, we created a task pool P. When a new task is generated, it is immediately added to P. When the number of tasks in the task pool P reaches the threshold value (automatic adjustment of the threshold will be described in Section 3.3), the CMA-ES method in Section 3.2 is used to allocate the tasks in the task pool to robots. The robots insert the new task sequence allocated into the rear of the previous unfinished task sequence, and then the task pool is emptied. The robots execute tasks according to their own task sequence, and the executed tasks are deleted from the sequence. As the new tasks are generated again, the tasks are added to P again. Loop until the warehouse stops running. In Figure 3, the specific steps are as follows:

Step 1. Initialize the task pool size and set the task pool P to be empty. For all robots, initialize task sequence T_i of every robot r_i.

Step 2. The threshold of the task pool size is automatically adjusted using adaptive control strategy in Section 3.3.

Step 3. New tasks are constantly added to P. Jump to step 4 when the number of tasks in the task pool reaches the threshold.

Step 4. The tasks in the task pool are assigned to the robots using the CMA-ES method in Section 3.2, and for all robots, the new task sequence assigned to robot r_i is inserted at the end of the current task sequence T_i.

Step 5. Clear the task pool P and jump to step 2.

The above solution in Figure 3 is executed by the central controller, and the robot only needs to execute the tasks according to the assigned task sequence. The parallel operation of the two parts enables the robots to be busy all the time, which saves time and meets the requirement of real-time storage system.

3.2. CMA-ES Algorithm. As mentioned in Section 3.1, tasks are assigned to robots when the number of tasks in the task pool reaches the threshold. This problem is regarded as an optimization problem in a static environment. This is a NP-hard problem, and the CMA-ES algorithm is used to find the optimal solution. The successful application in many fields [24–26] proves that the CMA-ES algorithm is a good search algorithm.

3.2.1. Representation of Solutions. Referring to ref. [27], for the task allocation problem with m tasks and n robots, a candidate to represent a task assignment scheme is $X = [x_1, x_2, x_3 \cdots x_m]$. X contains m real numbers, and for each real number x_i, it satisfies $1 \le x_i < n + 1, i = 1, 2, 3, \cdots, m$, where x_i means task i is performed by robot $\text{Int}(x_i)$, and $\text{Int}(x_i)$ means the integer of real number x_i. If $\text{Int}(x_i) = \text{Int}(x_j), i \neq j$, this means that the task x_i and x_j are both assigned to the same robot, and the task represented by the smaller number between x_i and x_j is executed first. If $x_i = x_j$, the execution order of these two tasks is determined randomly.

For example, there are 8 tasks (represented by numbers 1, 2, 3,..., 8) and 3 robots (represented by numbers 1, 2, 3), and an individual [1.7, 3.8, 2.2, 1.3, 2.8, 1.5, 3.3, 3.7] is generated. Then, the task sequence assigned to robot 1 is $4 \longrightarrow 6 \longrightarrow 1$. The task sequence assigned to robot 2 is $3 \longrightarrow 5$. The task sequence assigned to robot 3 is $7 \longrightarrow 8 \longrightarrow 2$.

3.2.2. Fitness Function. Fitness function is used to evaluate candidates. For the CMA-ES algorithm, individuals with lower fitness value are more excellent. In Section 2, two optimization goals are proposed for the whole system: one is the time C_{time} for the robots to complete all tasks; the second is the mean driving distance C_{distance} of all robots. Each planning can be regarded as a subproblem of the whole. For each subproblem, in order to achieve the optimal overall performance, these two goals are still considered; so, fitness function f is calculated through the following equation [16]:

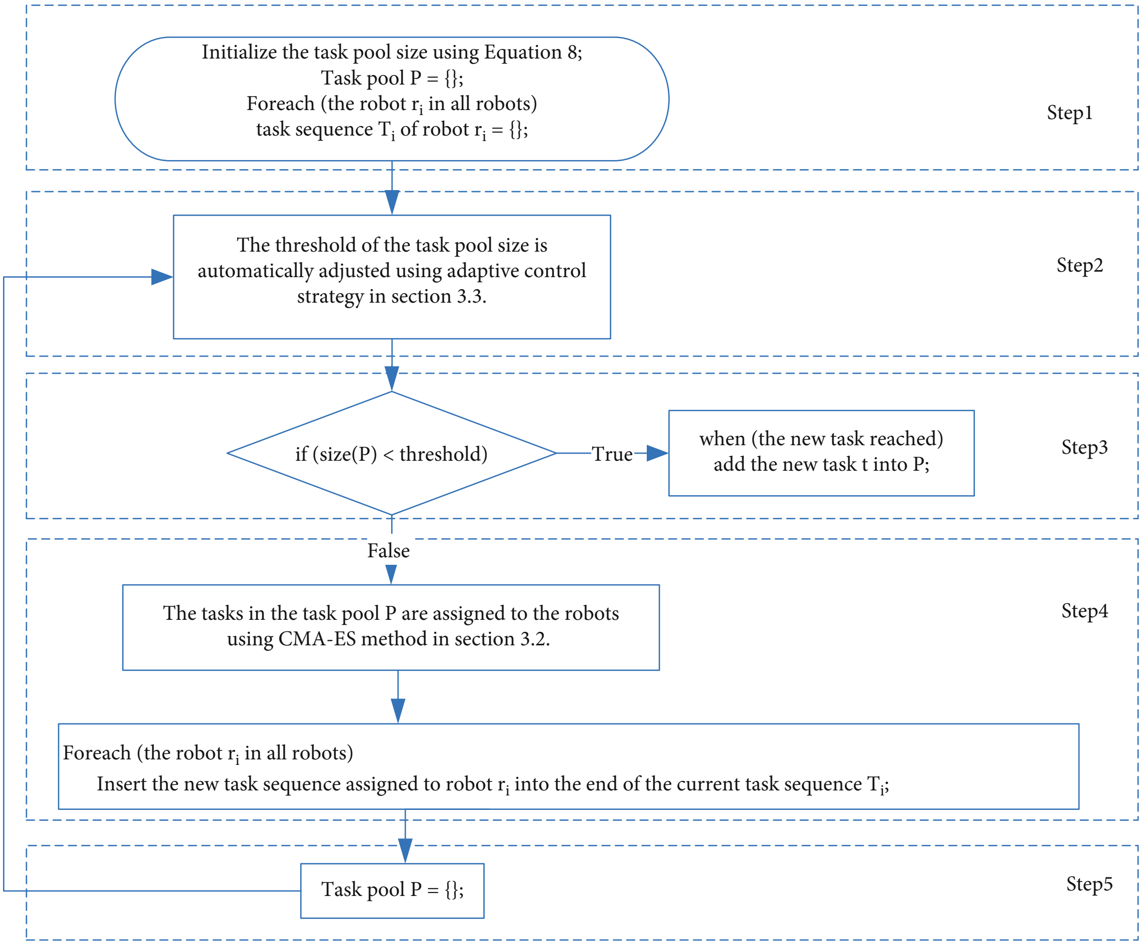

Initialize the task pool size using Equation 8;
Task pool P = {};
Foreach (the robot r_i in all robots)
task sequence T_i of robot r_i = {};
Step1
The threshold of the task pool size is automatically adjusted using adaptive control strategy in section 3.3.
Step2
if (size(P) < threshold)
True
when (the new task reached) add the new task t into P;
Step3
False
The tasks in the task pool P are assigned to the robots using CMA-ES method in section 3.2.
Step4
Foreach (the robot r_i in all robots)
Insert the new task sequence assigned to robot r_i into the end of the current task sequence T_i;
Task pool P = {};
Step5

Figure 3: The flow chart of the combined solution based on adaptive task pool strategy and CMA-ES.

$$f = \alpha C'_{\text{time}} + (1-\alpha) C'_{\text{distance}}, \; 0 \le \alpha \le 1,$$
$$C'_{\text{time}} = \max_i C'(r_i), \qquad (6)$$
$$C'_{\text{distance}} = \frac{\sum_{i=1}^{n} C'(r_i)}{\text{n}},$$

where α is a constant that can be adjusted according to the actual demand. If more attention is paid to the completion time of a single order, α can be increased. If more attention is paid to the energy consumption of all robots, α can be reduced. $C'(r_i)$ is the cost of robot r_i to execute the tasks in the current task sequence first and then execute the tasks according to the candidate. C'_{time} is the maximum time taken by the robots. C'_{distance} is the mean distance traveled by all robots. In the current moment, there may be unfinished tasks in the task sequence. The robot must first complete these tasks before performing the tasks assigned at the current moment. Therefore, for $C'(r_i)$, we divide it into two parts to calculate:

$$C'(r_i) = C'_1(r_i) + C'_2(r_i), \qquad (7)$$

where $C'_1(r_i)$ is the cost for the robot to complete the tasks in the current task sequence, and $C'_2(r_i)$ is the cost for the robot to execute the tasks according to the candidate. $C'_1(r_i)$ and $C'_2(r_i)$ are represented by the distance traveled by the robot and calculated using the method described in Equation (1).

With this fitness function, we try to find the optimal solution at that moment in each optimization and try to approximate the global optimal solution by this method.

3.3. Automatic Adjustment of Task Pool. When the number of tasks in the task pool reaches the threshold, the tasks in the task pool will be assigned to the robots. The threshold plays a decisive role in the efficiency of assignment. The larger the threshold is, the more tasks will be involved in the optimization, and then the more the planned scheme will be close to the global optimal solution. If an optimization contains all the tasks in the system, the optimal solution found by the optimization will be the optimal solution of the whole system. But orders in the warehouse are added dynamically over time, so tasks are also generated dynamically. As the threshold increases, the time required for the task pool to be filled will also increase, and this situation will occur: the robot has finished all the tasks assigned to it, but the number of tasks in the task pool has not reached the threshold; so, the next optimization cannot start, and the robot can only wait. This leads to a waste of time and cannot meet the real-time of the warehouse system. Moreover,

```
Input: lastAdjustTime, currentTime, lastTasksCompleted, tasksCompleted, oldThreshold, lastAction
Output: newThreshold, lastAction
1: if $currentTime - lastAdjustTime > I$ then
2:   if $tasksCompleted = 0$ then
3:     $newThreshold \longleftarrow oldThreshold/2$
4:     $lastAction \longleftarrow -1$
5:   else if $tasksCompleted - lastTaskCompleted \geq 0$ then
6:     $newThreshold \longleftarrow newThreshold + lastAction$
7:   else
8:     $newThreshold \longleftarrow newThreshold - lastAction$
9:     $lastAction \longleftarrow -lastAction$
10: else
11:   $newThreshold \longleftarrow oldThreshold$
12: return newThreshold, lastAction
```

Algorithm 1: Adaptive control strategy.

because each workstation has an order capacity limit, there is also an upper limit on the total number of tasks in the system, and if the task pool size exceeds this upper limit, the number of tasks in the task pool will never reach the threshold, and the system will be stagnant. Therefore, it is very important to set a threshold of appropriate size.

Obviously, for different warehouses, the threshold should be set differently depending on the actual situation. Even for the same warehouse, the number of robots may be adjusted, and the rate of order generation may vary at different times; so, it is not appropriate to set the threshold to a fixed value. Therefore, we design an adaptive control strategy to dynamically adjust the task pool, as shown in Algorithm 1.

First, the setting of the initial threshold is important, which determines the speed of finding the optimal threshold. We believe that the size of the initial threshold should be related to the number of robots and the upper limit number of tasks in the warehouse. The upper limit number of tasks in the warehouse is related to the number of workstations and the capacity of each workstation. So, we propose the following heuristic formula to calculate the initial threshold:

$$\text{initialThreshold} = \frac{(\gamma * \text{stations} + \text{robots})}{2}, \qquad (8)$$

where γ is a constant representing the average number of tasks per workstation in unit time, which is set according to the actual situation. stations is the number of stations, and robots is the number of robots. We set a time interval I (It is a constant that can be set according to actual requirements), and every I seconds, the threshold is adjusted (line 1). lastAction is used to record the last adjustment. We counted the total number of tasks completed by the robot from the last adjusted moment to the current moment, and the total number of tasks completed from the penultimate adjusted moment to the last adjusted moment, expressed by tasksCompleted and lastTasksCompleted, respectively. If taskCompleted is 0, indicating that the threshold has been set so high that the number of tasks has not reached the threshold, then simply cut the threshold in half and set lastAction to −1 (line 2, line 3, and line 4). If tasksCompleted is greater than or equal to lastTasksCompleted, it indicates that the last adjustment has had a positive effect on the system, and the same adjustment will be performed (line 5 and line 6). If tasksCompleted is less than lastTasksCompleted, it indicates that the last adjustment had a negative effect on the system, and the reverse adjustment will be performed (line 7 and line 8). In addition, lastAction will be reversed (line 9).

4. Experiments

We used RAWSim-O [22], an open source framework developed by Merschformann et al., as the experimental platform. RAWSim-O is a simulation framework that simulates the operation of an intelligent warehouse system and allows us to test our own methods.

We used the warehouse layout shown in Figure 2. In the warehouse layout, there are 32 robots and 550 shelves. The storage positions of the shelves are at the middle area of the layout. And there are four replenishment stations on the left and four picking stations on the right. To simplify the problem, we set the duration of a robot staying at a workstation to a very small value of 0.1.

For the assessment of performance we take the sum of SKUs (stock keeping unit) in both item bundles stored at the replenishment stations and orders picked at the picking stations as handled units. This represents the throughput of the warehouse, and the higher the better. We also look at the average distance traveled by robots to handle each unit. This can represent the power consumption of the multirobot system.

In order to test the impact of task pool threshold size on the allocation effect, we did 56 experiments, each experiment corresponding to different pool sizes. Each experiment was simulated for 24 hours with 10 repetitions.

Under different task pool sizes, the number of units handled by robots is shown in the blue solid line in Figure 4, and the average distance traveled by robots to handle each unit is shown in the blue solid line in Figure 5. The comparison results among different fixed threshold on handled units and travel distance per unit are shown in Table 1. The maximum number of handled units is 207583 when the fixed threshold is set to 18. The minimum number of travel

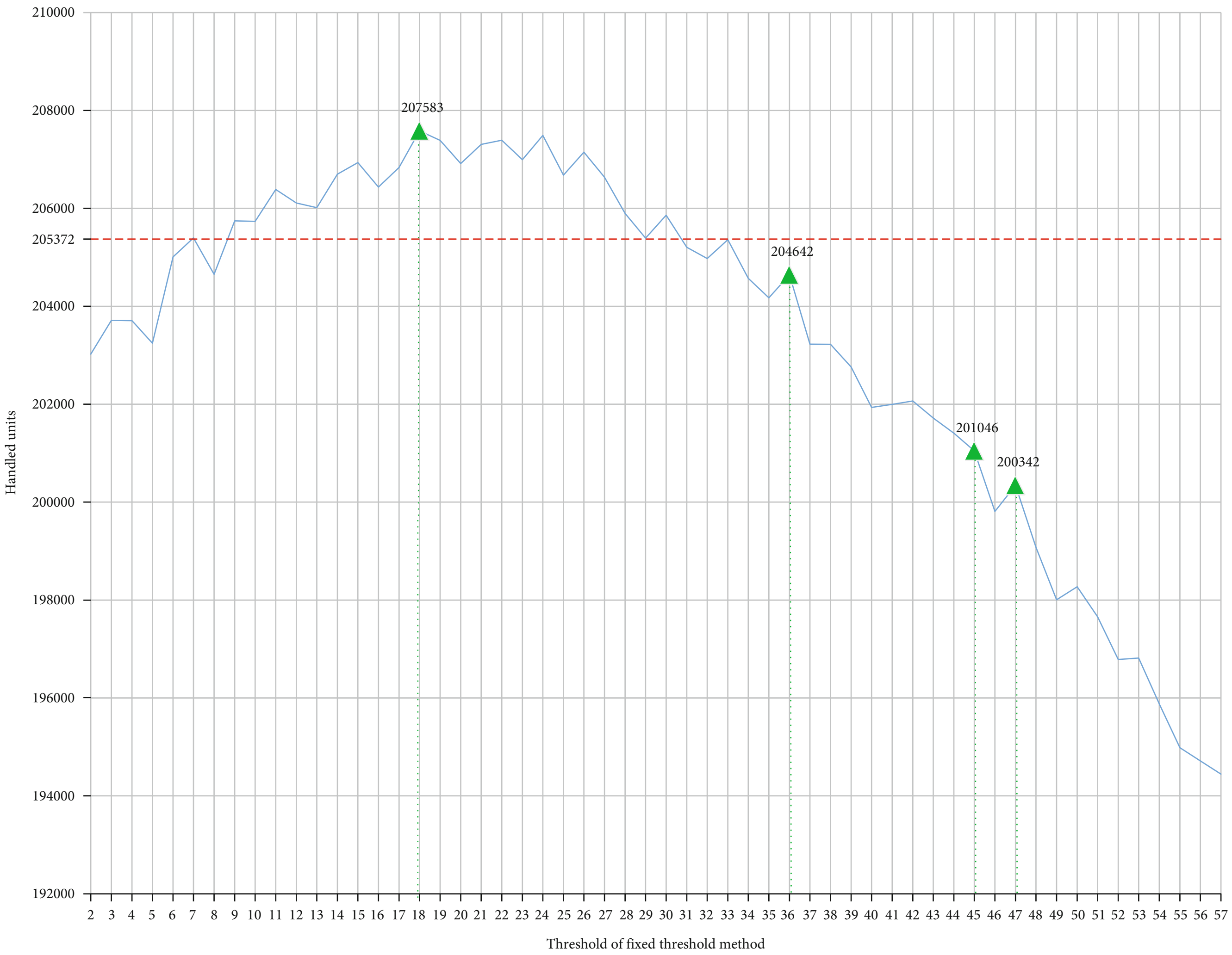


Figure 4: Comparison between adaptive threshold method and fixed threshold method on handled units. The red dotted line is the adaptive threshold method, and the blue solid line is the fixed threshold method.

distance per unit is 10.73 when the fixed threshold is set to 36, 45, or 47. According to Figures 4 and 5 and Table 1, it is not good to set the threshold too large or too small, which is consistent with our conjecture. If the threshold is set too small, the solution will be too far away from the global optimal solution; therefore, the number of handled units is small, and the travel distance per unit is large. If the threshold is set too large, the solution will be closer to the global optimal solution; so, the travel distance per unit is small, but the robot will have a long waiting time; therefore, the number of handled units will be small.

To sum up, a bad threshold can be very inefficient; so, setting the threshold manually is very risky. Therefore, a method of automatically adjusting threshold is necessary. We used the adaptive control strategy proposed by ourselves to conduct the experiment again, and all conditions were identical except the threshold. According to the workstation capacity, γ in Equation (8) was set to 4; so, the initial threshold was calculated as 32. The results are shown in Table 1. We compared the results with the fixed threshold approach, as shown in Figures 4 and 5. The red dotted line is the adaptive threshold method, and the blue solid line is the fixed threshold method. Compared with fixed threshold 18, the adaptive threshold method gets worse result in handled units but better result in travel distance per unit. Compared with fixed threshold 36, 45, and 47, the adaptive threshold method gets better result in handled units but worse result in travel distance per unit. Taken together, it can be seen from the two figures that the adaptive threshold method can be close to reaching the level when the threshold is set to the optimal in both indexes. The experimental results show that the proposed adaptive control strategy has good application effect.

5. Conclusion

In order to solve the dynamic and real-time problem of multirobot task allocation in the intelligent warehouse system, a combined solution based on adaptive task pool strategy and CMA-ES algorithm is proposed in the paper. In the early

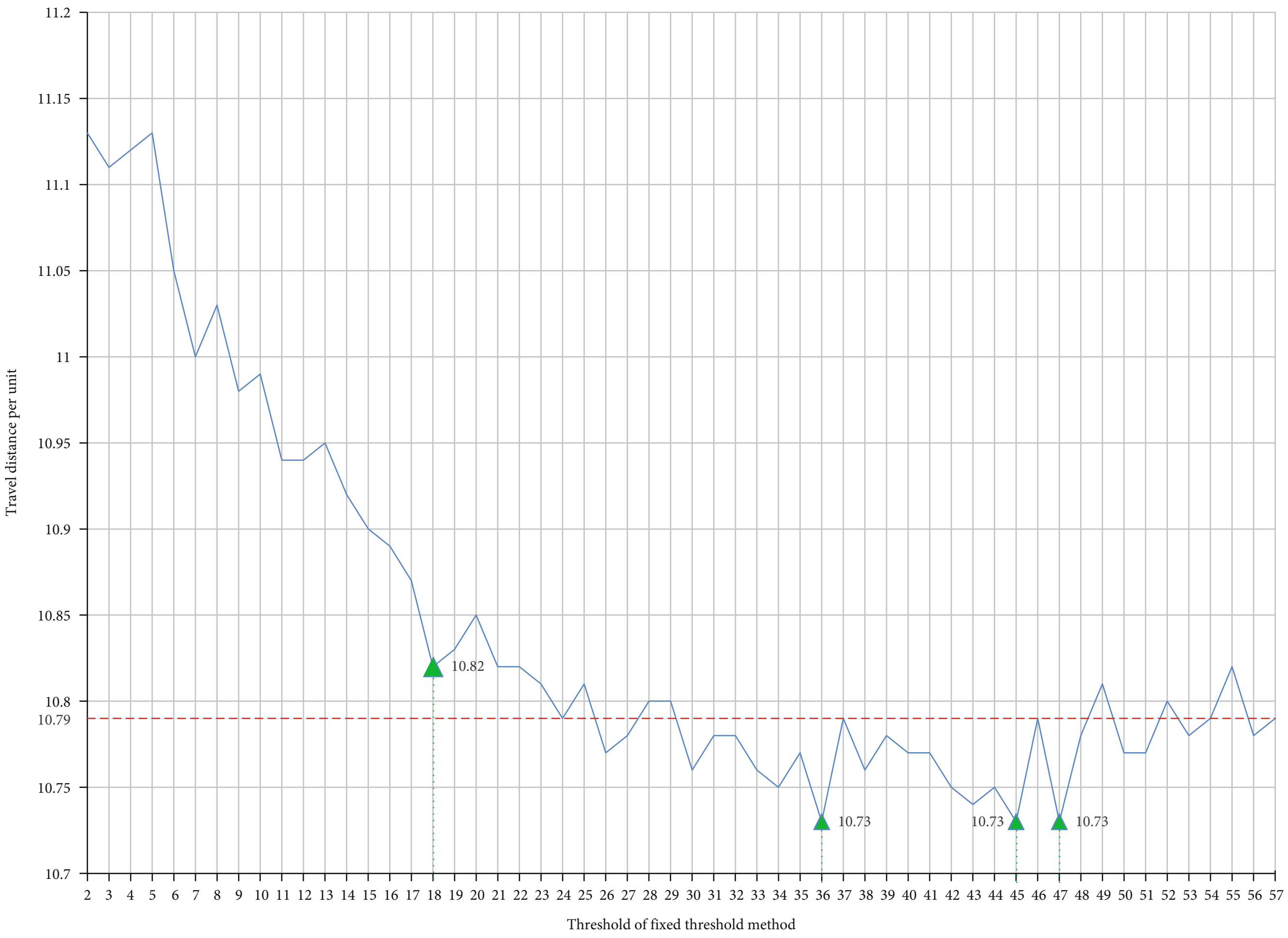


Figure 5: Comparison between adaptive threshold method and fixed threshold method on travel distance per unit. The red dotted line is the adaptive threshold method, and the blue solid line is the fixed threshold method.

Table 1: Comparison between adaptive threshold method and fixed threshold method on handled units and travel distance per unit.

Method (initial threshold)	Handled units	Travel distance per unit
Fixed threshold (18)	207583	10.82
Fixed threshold (36)	204642	10.73
Fixed threshold (45)	201046	10.73
Fixed threshold (47)	200342	10.73
Adaptive threshold (32)	205372	10.79

stage of the solution, the divide-to-conquer idea is used to design a variable task pool that is used to store dynamically added tasks. The variable task pool is designed to dynamically divide continuous and large-scale task allocation problems into small-scale subproblems to solve them to meet dynamic requirements. And an adaptive control strategy is used to automatically adjust the threshold of the task pool size in real time to achieve a trade-off among throughtput, energy consumption, and waiting time, which has better adaptability than manually adjusting the size of the task pool. In the later stage of the solution, when the task pool is full, tasks in the task pool will be assigned to robots using the CMA-ES algorithm to find the optimal task assignment solution for all the robots according to the fitness function including the maximum time and the mean travel distance required by all robots to complete all the tasks. By comparing with fixed threshold method under 56 different task pool sizes, the experimental results show that the handled units can be close to reaching the optimal level, and the average travel distance per unit is lower using adaptive threshold method; so, adaptive threshold solution indeed has better adaptability. This method can satisfy the dynamic and real-time requirements and can be effectively applied to the intelligent warehouse system.

However, because of the complexity and dynamics of the warehouse environment, it may not be accurate to measure the cost by Manhattan distance. Therefore, how to introduce accurate robot motion model to evaluate the cost will be the next work. Furthermore, the relationships among handled units, travel distance per unit, the maximum time taken by all robots to complete all tasks, and the mean distance traveled by all robots need further study. In addition, the effect of communication quality on allocation is not taken into account and will be deeply studied.

References

[1] M. Zhou and M. Y. Wang, "Analysis on the development of e-commerce logistics service industry and countermeasures," *Computer and Information Technology*, vol. 20, no. 6, pp. 10–12, 2012.

[2] S. X. Zou, "The present and future of warehouse robot," *Logistics Engineering and Management*, vol. 35, no. 6, pp. 171-172, 2013.

[3] J. J. Enright and P. R. Wurman, "Optimization and coordinated autonomy in mobile fulfillment systems," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 33–38, San Francisco, California, 2011.

[4] P. R. Wurman, R. D'Andrea, and M. Mountz, "Coordinating hundreds of cooperative, autonomous vehicles in warehouses," *AI Magazine*, vol. 29, no. 1, p. 9, 2008.

[5] B. P. Gerkey and M. J. Matarić, "A formal analysis and taxonomy of task allocation in multi-robot systems," *International Journal of Robotics Research*, vol. 23, no. 9, pp. 939–954, 2004.

[6] A. Khamis, A. Hussein, and A. Elmogy, "Multi-robot task allocation: a review of the state-of-the-art," *Eds. Cham: Springer International Publishing*, vol. 604, pp. 31–51, 2015.

[7] B. P. Gerkey and M. J. Matarić, "Sold!: auction methods for multirobot coordination," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 5, pp. 758–768, 2002.

[8] M. Berhault, H. Huang, P. Keskinocak et al., "Robot Exploration with Combinatorial Auctions," *In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1957–1962, 2003.

[9] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[10] P. MacAlpine, E. Price, and P. Stone, "SCRAM: scalable collision-avoiding role assignment with minimal-makespan for formational positioning," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2096–2102, Austin, Texas, USA, 2015.

[11] L. E. Parker, "ALLIANCE: an architecture for fault tolerant multirobot cooperation," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 2, pp. 220–240, 1998.

[12] B. B. Werger and M. J. Mataric, "Broadcast of local eligibility: behavior-based control for strongly cooperative robot teams," in *Proceedings of the 4th International Conference on Autonomous Agents*, pp. 21-22, Barcelona, Spain, 2000.

[13] Y. Liu, X. Zhang, H. Li, and D. Qian, "Allocating tasks in multi-core processor based parallel system," in *2007 IFIP International Conference on Network and Parallel Computing Workshops*, pp. 748–753, Liaoning, China, 2007.

[14] E. G. Jones, M. B. Dias, and A. Stentz, "Time-extended multi-robot coordination for domains with intra-path constraints," *Autonomous Robots*, vol. 30, no. 1, pp. 41–56, 2011.

[15] J. Zhang and Y. Q. Cao, "Research on dynamic task allocation for MAS based on hybrid genetic and ant colony algorithm," *Computer Science*, vol. 38, no. S1, pp. 268–270, 2011.

[16] J. J. Dou, C. L. Chen, and P. Yang, "Genetic scheduling and reinforcement learning in multirobot systems for intelligent warehouses," *Mathematical Problems in Engineering*, vol. 2015, 10 pages, 2015.

[17] W. Deng, J. Xu, H. Zhao, and Y. Song, "A novel gate resource allocation method using improved PSO-based QEA," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1737–1745, 2022.

[18] X. Cai, H. Zhao, S. Shang et al., "An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application," *Expert Systems with Applications*, vol. 171, article 114629, 2021.

[19] W. Deng, J. J. Xu, X. Z. Gao, and H. M. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 3, pp. 1578–1587, 2022.

[20] W. Deng, S. Shang, X. Cai et al., "Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization," *Knowledge-Based Systems*, vol. 224, article 107080, 2021.

[21] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)," *Evolutionary Computation*, vol. 11, no. 1, pp. 1–18, 2003.

[22] M. Merschformann, L. Xie, and H. Li, "RAWSim-O: a simulation framework for robotic mobile fulfillment systems," *Logistics Research*, vol. 11, no. 8, pp. 1–11, 2018.

[23] L. Xie, N. Thieme, R. Krenzler, and H. Y. Li, *Efficient Order Picking Methods in Robotic Mobile Fulfillment Systems*, 2019, https://arxiv.org/abs/1902.03092.

[24] F. Stulp and O. Sigaud, "Path integral policy improvement with covariance matrix adaptation," in *29th International Conference on Machine Learning*, Edinburgh, Scotland, 2012.

[25] T. Geijtenbeek, M. Van De Panne, and A. F. Van Der Stappen, "Flexible muscle-based locomotion for bipedal creatures," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 1–11, 2013.

[26] P. MacAlpine and P. Stone, "Overlapping layered learning," *Artificial Intelligence*, vol. 254, pp. 21–43, 2018.

[27] H. R. Zhou, W. S. Tang, and H. L. Wang, "Optimization of multiple traveling salesman problem based on differential evolution algorithm," *Systems Engineering Theory & Practice*, vol. 30, no. 8, pp. 1471–1476, 2010.

An Efficient Method for Diagnosing Brain Tumors Based on MRI Images Using Deep Convolutional Neural Networks

Swarna Manjari Samal, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, swarnamanjari88@gmail.com*

Binayini Pradhan, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, binayini.pradhan@gmail.com*

Chinmaya Ranjan Pradhan, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, cr.pradhan23@gmail.com*

Satyajit Nayak, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, satyajit_nayak@gmail.com*

Abstract

This paper proposes a system to effectively identify brain tumors on MRI images using artificial intelligence algorithms and ADAS optimization function. This system is developed with the aim of assisting doctors in diagnosing one of the most dangerous diseases for humans. The data used in the study is patient image data collected from Bach Mai Hospital, Vietnam. The proposed approach includes two main steps. First, we propose the normalization method for brain MRI images to remove unnecessary components without affecting their information content. In the next step, Deep Convolutional Neural Networks are used and then we propose to apply ADAS optimization function to build predictive models based on that normalized dataset. From there, the results will be compared to choose the most optimal method. Those results of the evaluated algorithms through the coefficient F1-score are greater than 94% and the highest value is 97.65%.

1. Introduction

The brain is a particularly important organ, the control center of the central nervous system, coordinating the activities of all organs and parts in the human body. The brain has a complex structure and is protected and covered by the skull, a very hard bone box. However, a rigid skull may help protect the brain parenchyma from minor trauma but does not prevent the development of lesions and abnormal structures within the brain. One of the brain diseases of primary concern in medicine is brain tumors. A brain tumor is a condition in which abnormal cells grow in the brain. Brain tumors are divided into two types: benign brain tumors and malignant brain tumors (called cancer) [1]. Whether it is a benign brain tumor or a malignant brain tumor, it affects brain cells, causing brain damage and being even life-threatening. There are about 120 different types of brain tumors, most of which are tumors in the brain tissue, in addition to tumors in the meninges, pituitary gland, cranial nerves. Any form of brain tumor can be dangerous for the patient. Tumors in brain tissue or benign brain tumors often progress slowly; the symptoms of brain tumors in this case will also appear slower and more insidious. In contrast, if the brain tumor grows rapidly, the patient will feel the symptoms more pronounced in both frequency and extent. With current medical capabilities, early detection of abnormal structures in the patient's brain can improve the likelihood of successful treatment and limit the sequelae of tumors to the brain in general and the patient's health in particular.

The detection of brain tumors today is mostly based on the ability of doctors to distinguish abnormalities on MRI images which is a type of high-quality image in the field of imaging [2]. This is a process that requires a lot of experience and concentration to detect and classify brain diseases and brain tumors. From brain MRI, it is possible to diagnose and recognize many different types of brain tumors and offer appropriate treatment methods [3]. However, the increasing number of patients with the large number of images obtained has become a major challenge in the field of diagnostic imaging, a field that requires rapid and accurate evaluation of results by doctor. Artificial intelligence technology will help classify diseases from MRI images quickly and bring high accuracy in disease diagnosis. The classification of diseases based on MRI images has not been too difficult with high accuracy due to the introduction of GPUs (Graphics Processing Unit) and image processing based on artificial intelligence (AI).

This research focuses on the application of image preprocessing techniques and the development of algorithms using convolutional neural network (CNN) models, which are advanced deep learning models such as DenseNet201 [4], ResNet152V2 [5], MobileNetV3 [6], and VGG19 [7]. At the same time, the research also focuses on developing and applying the ADAS optimization algorithm to improve the accuracy in classifying normal people and brain tumor patients. The dataset in this work includes 1307 brain MRI images in JPEG format that are manually classified by specialists into 2 categories: normal human brain MRI images and brain MRI images of people with brain tumor disease. The comparison of all experimental results will evaluate the effectiveness of each model.

The article is organized as follows. Section 2 presents the previously conducted MRI brain tumor classification studies. Section 3 provides an overview of brain MRI images and the CNN algorithm models used. Section 4 presents the experimental results and gives evaluation for each algorithm. Conclusions and future work are outlined in Section 5.

2. Related Work

Several technical methods related to brain MRI images classification since 2017 based on different classification models are summarized in Table 1. They are divided into two basic methods: using CNN network architecture and not using CNN network architecture. In [10], the authors divided brain MRI images into two categories: normal images and images with abnormal signs. They used GLCM to get the features of the MRI images; then a probabilistic neural network (PNN) was used to classify the MRI brain images of people as normal or abnormal. As a result, they obtained a classification model with an accuracy of 95%. In [14], Ullah et al. proposed a scheme to classify the brain MRI images of normal people and patients using equilibrium histograms, discrete wavelet transforms, and Feedforward Artificial Neural Networks. Recently, deep learning method has been widely used for the classification of brain tumors on MRI images [8, 9]. The deep learning method does not need to manually extract the features of the images; it combines the extraction and classification stages in the self-learning process. The deep learning method requires a dataset where normalized processing of the MRI images is sometimes required, and then salient features are identified during machine learning [13].

Convolution Neural Network (CNN), one of the well-known deep learning techniques for image data, can be used as a feature extraction tool from which to capture related features to perform data classification task. Feature maps in the initial and higher layers of the CNN model extract low-level features and specific features of high-level content, respectively. Feature maps in the earlier layer construct simple structural information, such as shapes, textures, and edges, while the higher layers combine these low-level features into constructing (encoding) expressions performance, integrating local and global information.

Various researchers have proposed to use CNN to classify brain tumors based on brain MRI image datasets [11, 21, 22]. Deepak and Ameer [12] used pretrained GoogLeNet to extract features from brain MRI images with CNN network architecture to classify three types of brain tumors and obtained up to 98% accuracy. Çinar and Yildirim, [15] modified the ResNet50 network based on the pretrained CNN network architecture by removing the last 5 layers and adding 8 new layers, and that method achieved 97.2% accuracy. Saxena et al. [17] used InceptionV3, ResNet50, and VGG16 network architectures with legacy methods to classify brain tumor data. In this study, ResNet50 model obtained the highest accuracy rate with 95%. Díaz-Pernas et al. [18] presented a CNN network architecture for automatic brain tumor segmentation such as glioma, meningioma, and pituitary tumor. They evaluated their proposed model using the T1-weighted contrast-enhanced MRI dataset and obtained an accuracy of 97.3%.

Siddiaue et al. [16] proposed a model based on modified vgg-16 network architecture for brain tumor images classification which achieved an accuracy of 96% and an F1-score of 97%. Abd El Kader et al. [19] developed a differential deep-CNN-based model to classify MRI images with and without tumors. In fact, this model was still based on the basic CNN architecture but obtained an accuracy of 99.25% and an F1-score of 95.23%. In [20], the authors successfully deployed transfer learning for some variant architectures of CNN to apply to the classification of MRI images with and without brain tumors, in which MobileNetV2 had an accuracy of 92% and F1-score of 92%; InceptionV3 had an accuracy of 91% and an F1-score of 90.98%; VGG19 had an accuracy of 88% and F1-score of 88.18%.

In summary, as observed from the above studies, the accuracy obtained by using deep learning with CNN network architecture to classify brain MRI is significantly higher than that of the old traditional techniques. However, deep learning models require a large amount of data to train in order to perform better than traditional machine learning techniques.

3. Materials and Methods

3.1. Brain Tumor MRI Images

3.1.1. Content Contained in MRI Images. The commonly used standard for MRI images today is DICOM, an acronym for Digital Imaging and Communications in Medicine Standards [23]. This is an industry standard system developed to meet the needs of manufacturers and users in connecting, storing, exchanging, and printing medical images.

As for the DICOM image format standard, in addition to the image files, it also includes header files as in Figure 1.

Table 1: Summary of studies on brain tumor classification.

Author	Classification method	Objective	Dataset	Feature extraction method	Accuracy
Khawaldeh et al. [8]	CNN	Classification of brain MRI into normal and abnormal	587 MR images	CNN	91.16%
Paul et al. [9]	Fully connected and CNN	Brain tumor classification of MR brain image	3064 MR images	CNN	91.43%
Varuna Shree and Kumar [10]	Probabilistic neural network (PNN)	Classification of brain MRI into normal and abnormal	650 MR images	Gray level cooccurrence matrix (GLCM)	95%
Hemanth et al. [11]	CNN	Classification into normal and abnormal	220 MR images	CNN	94.5%
Deepak and Ameer [12]	Deep transfer learning	Classification of glioma, meningioma, and pituitary tumors	3064 MR images	Google Net	98%
Das et al. [13]	CNN	Brain tumor classification	3064 MR images	CNN	94.39%
Ullah et al. [14]	Feedforward neural network	Classification of brain MRI into normal and abnormal	71 MR images	DWT	95.8%
Çinar and Yildirim [15]	CNN models	Brain tumor detection and classification	253 MR images	CNN	97.2%
Siddiaue et al. [16]	Proposed DCNN model	Brain tumor classification	253 MR images	CNN	96%
Saxena et al. [17]	CNN networks with transfer learning	Binary classification of brain tumor normal and abnormal	253 MR images	CNN	95%
Díaz-Pernas et al. [18]	Multipathway CNN	Brain tumor classification	3064 MR images	CNN	97.3%
Abd El kader et al. [19]	Proposed differential deep-CNN	Brain tumor classification	25000 MR images	CNN	99.25%
Tazin et al. [20]	CNN architectures	Brain tumor classification	2513 MR images	CNN	Up to 92%

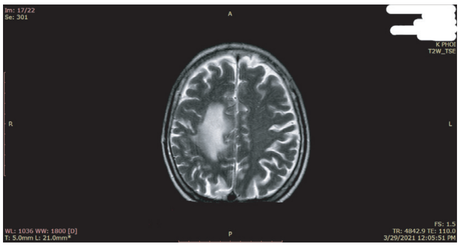


Figure 1: Actual DICOM image (patient information has been removed).

Although stored in different files, when displayed, the header information is displayed along with the MRI image information via a "DICOM browser." Data in MRI images include demographic information, patient information, parameters acquired for imaging studies, image size, and image matrix size. The patient's information displayed includes patient's first and last name, gender, age, date of birth, and place where the MRI scan was performed.

3.1.2. The Role of MRI Images in the Diagnosis of Brain Tumors. Magnetic resonance imaging of the brain [24] can very clearly detect and describe abnormalities in the brain parenchyma in general such as vascular tumors, arterial occlusion, and invasion of the venous sinuses as well as the relationship between tumor and surrounding structures. There are three basic image formats of MRI images: T1W, T2W, and T2 Flair. They are used in specific cases depending on the situation of the disease.

T1W imaging is mainly used to identify necrotic tumors, hemorrhage in tumors, or cysts. For example, with MRI images in meningiomas, on T1-weighted images, most meningioma shows no difference in signal intensity compared with cortical gray matter.

For image in the T2W phase, the received signal has been changed completely; it is a fairly homogeneous gain signal block. Imaging is also helpful in evaluating hemorrhages and cysts. In particular, the role of the T2W phase is very useful in reflecting the homogeneity of benign soft tumors or meningiomas.

For Fluid-Attenuated Inversion Recovery (T2-FLAIR), this type of phase image is very useful to evaluate the consequences and effects of edema. Although this finding is not specific for meningiomas in particular, it is very meaningful in the diagnosis as well as the long-term prognosis for the patient.

Overall, the sensitivity and specificity of MRI are very high in the diagnosis of meningiomas. MRI has been shown to be superior in tumor delineation by its relationship to surrounding structures.

3.2. Model Architectures

3.2.1. Supervised Learning. Supervised learning [25] is an algorithm that predicts the output (outcome) of a new data (new input) based on previously known (input, outcome) pairs. This data pair is also known as (data, label). Supervised learning is the most popular group of machine learning algorithms.

Mathematically, supervised learning consists of a set of input variables $X = \{x_1, x_2, \ldots, x_N\}$ and a corresponding set of labels $Y = \{y_1, y_2, \ldots, y_N\}$, where x_i and y_i are vector. The data pairs $(x_i, y_i) \in X \times Y$ are called the training dataset. From this training dataset, we need to create a function that maps each element from the set X to a corresponding (approximate) element of the set Y as

$$y_i \approx f(x_i), \quad \forall i = 1, 2, \ldots, N. \tag{1}$$

The goal is to approximate the function f very well so that when we have a new data x we can compute its corresponding label:

$$y = f(x). \tag{2}$$

A problem is called classification if the labels of the input data are divided into a finite number of groups.

3.2.2. Convolutional Neural Network Architectures. Convolutional Neural Network (CNN) [31] is one of the most popular and most influential deep learning models in the computer vision community. CNN is used in many problems such as image recognition and video analysis or for problems in the field of natural language processing and solves most of these problems well.

CNN includes a set of basic layers such as convolution layer, nonlinear layer, pooling layer, and fully connected layer. These layers are linked together in a certain order. Basically, an image will be passed through the convolution layer and nonlinear layer first; then the calculated values will be passed through the pooling layer to reduce the number of operations while preserving the characteristics of the data. The convolution layer, nonlinear layer, and pooling layer can appear one or more times in the CNN network. Finally, the data is passed through fully connected network and soft-max to calculate the probability of object classification.

Table 2 summarizes some typical CNN network architectures since 2012. To evaluate and compare network structures, two parameters are used, Top 1 Accuracy and Top 5 Accuracy. In the case of Top 1 Accuracy, the correct model's prediction must be the model that predicts the class with the highest probability. In the case of Top 5 Accuracy, the correct model's prediction is the model that correctly predicts one of the 5 classes with the highest probability.

In this study, four different network architectures are used: DenseNet201 [4], ResNet152V2 [5], MobileNetV3 [6], and VGG19 [7]. All the above four network architectures are developments and upgrades based on the basic network architecture CNN, one of the advanced deep learning models for image classification that has been verified with high accuracy on image sets, ImageNet [32]. These CNN variant network architectures are widely used in image recognition and classification problems. All four network architectures have a structure consisting of two basic layers, the feature extraction layer and the classifier layer. In this research, the input to the network architecture is a 256 × 256 brain MRI image containing information with or without brain tumors. The feature extraction layer has the role of extracting features of brain MRI images such as white matter, gray matter, cerebrospinal fluid, cerebral cortex, and brain tumor. Then, the classification layer is responsible for synthesizing the features of brain MRI images, giving specific features of images with tumors and images without tumors to serve the classification process.

3.3. Optimal Algorithms. The optimization algorithm is the basis for building a neural network model with the aim of "learning" the features (or patterns) of the input data, from which it is possible to find a suitable pair of weights and biases to optimize the model. But the question is how to "learn?" Specifically, how the weights and biases are found, not just randomly taking the weights and biases values for a finite number of times and hoping after some steps a solution can be found. Therefore, it is necessary to find an algorithm to improve weights and biases step by step, and that is why optimizer algorithms were created.

Some of the factors commonly used to evaluate an optimizer algorithm are as follows:

Fast convergence (Training Process)

High generalization (can recognize previously untrained patterns)

High accuracy

The popular optimization algorithms are listed in Figure 2. These algorithms are GD-Gradient Descent [33], SGD-Stochastic Gradient Descent [34], ADAS-Adaptive Scheduling of Stochastic Gradients [35], AdaGrad-Adaptive Subgradient Methods for Online Learning and Stochastic Optimization [36], Momentum [37], RMSProp [33], and ADAM-Adaptive Moment Estimation [38].

Among the above algorithms, the Optimal Algorithms belonging to the adaptive family usually have fast convergence speed. Meanwhile, algorithms belonging to the SGD family often have high generalization. However, this study only focuses on the development and application of ADAM and ADAS algorithms.

3.3.1. ADAM Algorithm: A Method for Stochastic Optimization. ADAM is a combination of Momentum and RMSProp. One of the key components of ADAM is exponential weighted moving averages (also known as leak averages) that estimate both the momentum and the second-order moment of the gradient. Specifically, it uses state variables as follows:

$$\begin{aligned} v &\leftarrow \beta_1 v_t - 1 + (1 - \beta_1) g_t, \\ s &\leftarrow \beta_2 s_t - 1 + (1 - \beta_2) g_t^2, \end{aligned} \tag{3}$$

where v is the first moment vector, s is the second moment vector, β_1 and β_2 are the jump parameters at the initial and the second points in ADAM's algorithm, t is the time for the correction steps, and g is the gradient. Here β_1 and β_2 are nonnegative weight parameters. Popular choices for them are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. This means that the variance estimate moves much slower than the momentum term.

Table 2: Some CNN architectures and ImageNet benchmark (image classification).

CNN architectures	Author	Top 1 accuracy (%)	Top 5 accuracy (%)
AlexNet	Krizhevsky et al. [26]	63.3	84.6
VGG16	Simonyan and Zisserman [7]	74.4	91.9
VGG19		74.5	92.0
GoogLeNet (InceptionV1)	Szegedy et al. [27]	69.8	89.5
ResNet50	He et al. [5]	77.1	93.3
ResNet152		78.6	94.3
DenseNet201	Huang et al. [4]	77.4	93.7
MobileNet224	Howard et al. [6]	70.6	89.5
MobileNetV3		79.0	94.5
ResNeXt101	Xie et al. [28]	80.9	95.6
EfficientNet-L2	Tan and Le [29]	90.2	98.8
RegNet-Y	Radosavovic et al. [30]	79.9	95.0

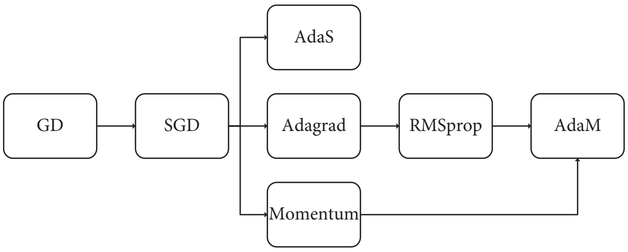


Figure 2: The development of optimization algorithms.

Note that if initializing the values $v_0 = s_0 = 0$, the algorithm will have a significant initial bias towards smaller values. This problem can be solved using $\sum_{i=0}^{t}(1-\beta_t)/(1-\beta)$ to normalize the terms. Similarly, state variables are normalized as follows:

$$v_t' = \frac{v_t}{1-\beta_1^t},$$
$$v_t' = \frac{s_t}{1-\beta_2^t}. \tag{4}$$

From the appropriate estimates, the updated equations can be established. First, the gradient value will be adjusted, similar to that in RMSProp [33] to get

$$g_t' = \frac{\eta s_t}{\sqrt{s_t'} + \varepsilon}, \tag{5}$$

where ε is a constant and it is chosen to be $\varepsilon = 10^{-6}$ to balance arithmetic stability and reliability, and η is the learning rate. From there, the update step is defined as follows:

$$x_t \leftarrow x_t - 1 - g_t'. \tag{6}$$

When looking at the design of ADAM, the inspiration of the algorithm is clear. Momentum and range are clearly represented in the state variables. Moreover, based on RMSProp it is easy to see that the combination of both terms is quite simple. Finally, the learning rate η allows us to control the update step length to solve convergence problems.

3.3.2. ADAS Algorithm: Adaptive Scheduling of Stochastic Gradients. ADAS [35] is an optimization algorithm belonging to the family of Stochastic Gradient Descent (SGD) algorithms. The updated rules for ADAS are established using SGD with momentum as follows:

$$\begin{aligned}
&\eta(t,l) \leftarrow \beta.\eta(t-1,l) + \zeta \cdot [\overline{G}(t,l) - \overline{G}(t-1,l)],\\
&v_l^k \overline{G} \alpha \cdot v_l^{k-1} - \eta(t,l) \cdot g_l^k,\\
&\theta_l^k \overline{G} \theta_l^{k-1} + v_l^k,
\end{aligned} \tag{7}$$

where η is the learning rate, t is the time for the correction steps, β is ADAS gain factor, ζ is the knowledge gain hyperparameter, k is the current minibatch, t is the current epoch iteration, l is the convolution block index, $\overline{G}(\cdot)$ is the average knowledge gain obtained from both mode-3 and mode-4 decompositions, v is the velocity term, and θ is the learnable parameter. The learning rate is calculated relative to the rate of change of knowledge acquired after the training epochs. The learning rate $\eta(t,l)$ is then further updated by an exponential moving average called the gain factor, with the hyperparameter β, to accumulate the history of the knowledge gained over the series epochs. In fact, β controls the trade-off between convergence rate and training accuracy of ADAS.

ADAS is an adaptive optimization tool for scheduling the learning rate in the training of a CNN network. ADAS exhibits a much faster convergence speed than other optimization algorithms. ADAS demonstrated generalization characteristics (low test loss) on par with SGD-based optimizers, improving on the poor generalization characteristics of adaptive optimizers. In addition to optimization, ADAS introduces new polling metrics for CNN layer removal (quality metrics).

3.4. Accuracy and F1-Score. The classification problem in this study is a binary classification problem, in which one class is an MRI image with a brain tumor and the other is an MRI image without a brain tumor. This study considers the image class with brain tumor to be positive and the remaining image class without brain tumor to be negative. The parameters True Positive (TP), False Positive (FP), True

Negative (TN), and False Negative (FN) are described as in Table 3.

In this paper, the parameters used to evaluate the effectiveness of the model are accuracy, precision, recall, and F1-score [39]. When building a classification model, the ratio of correctly predicted cases to the total number of cases is always considered. That ratio is called accuracy. Precision is the answer to the question: how many true positives are there out of the total number of positive diagnoses? Recall measures the rate of correctly predicting positive cases across all samples in the positive group. F1-score is the harmonic mean between precision and recall. Therefore, in situations where the precision and recall are too different, the F1-score will balance both values and help us to make an objective assessment. Accuracy, precision, recall, and F1-score are defined as the following equations:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{total sample}} \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{8}$$

4. Experiments and Results

This study will compare the results of the network architectures DenseNet201, ResNet152V2, MobileNetV3, and VGG19 in the cases before and after data normalization with the ADAM optimization function. Then, the study will specifically compare the performance of the above algorithms with the ADAM and ADAS optimization functions on the same normalized dataset.

4.1. Collecting and Normalizing Data

4.1.1. Collecting Data. In this study, the dataset is a set of MRI brain tumors of 123 patients with brain tumors at Bach Mai Hospital, Hanoi, Vietnam, of all ages. Initially, the MRI image was in DICOM format; to remove the information in the patient's DICOM image and convert the image format for machine learning, the DICOM format was converted to the JPEG image format. The size of the converted images is 256 × 256 pixels.

The image used during training is a T2 pulse sequence image as in Figure 3. Signal intensity with T2 phase correlates very well with not only homogeneity but also tissue profile. Specifically, with low-intensity signals, the tumor has a fibrous and stiffer character than the normal parenchyma. For example, the tumor is fibroblastic in nature, while the more intense sections show a softer characteristic such as a vascular tumor. Therefore, the image of the T2 pulse sequence is considered a pulse sequence that best assesses whether the patient has a brain tumor or not.

Table 3: Confusion matrix.

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

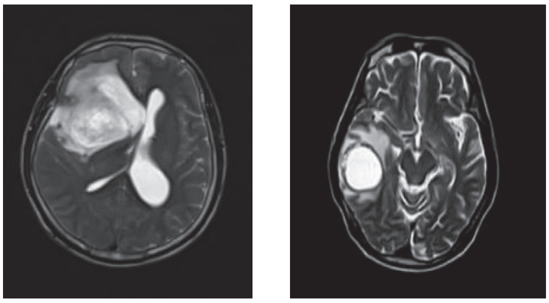


Figure 3: Image of the T2 pulse sequence showing the patient's brain tumor.

With the above 123 patients with brain tumor pathology and 100 healthy persons, 1307 images of T2 pulse sequence were selected, of which 647 images showed brain tumors and 660 images did not show brain tumors. The images are all brought to a size of 256 × 256 pixels to serve the training and testing process of the algorithms.

4.1.2. Normalizing Data

(1) Minimizing Image Redundancy. In the raw MRI image data, it is easy to see that there is a rather large black border, but that is the air in the optical field of the machine, so it does not carry information about the skull to be examined. Therefore, it is really necessary to remove the black out-of-the-edge image from the MRI image without affecting the image information content.

The skull on an MRI is usually surrounded by a bright white border, the outer layer of fat around the skull. Meanwhile, the MRI image is a grayscale image (one-dimensional); the range of values of each element in the image matrix representing the brightness of the pixel is in the range [0, 255]. In order to maximize the black border on the image, the easy method implemented by this study is to find the first pixel with a nonzero value in the directions from left to right, from right to left, from top to bottom, and from bottom to top as shown in Figure 4. After determining the coordinates of those pixels, remove the outer edges. The normalization of images by cutting out the parts that do not make sense in image classification aims to increase the accuracy of the training process and reduce the training time of the algorithm.

(2) Normalizing Image Size. The normalization of the image size helps to improve the accuracy and efficiency of the algorithm. In this study, the image size is 256 × 256. This is the right image size for AI algorithms and ensures MRI

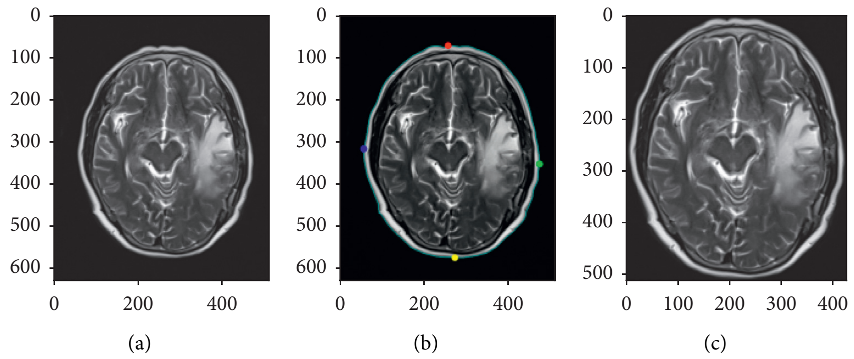


Figure 4: Removal of redundant areas of MRI images.

image quality after resizing. Choosing a smaller size will make it difficult for AI algorithms to detect small differences between pixels, affecting the accuracy of the algorithm. If the image size is larger, it will affect the quality of the MRI image after resizing/reducing image quality, negatively affecting the accuracy and performance of the algorithm.

Normalization of data is processed by image data files corresponding to each type of patient's MRI image and by using Python programming. The normalized data removes the nonsignificant parts of the image classification, which increases the accuracy of the model training process and reduces the training time of the algorithm.

4.2. Image Classification Process

Step 1. Preparing the training dataset and feature extraction.

This step is considered an important step in machine learning problems because it is the input for learning to find the model of the problem. We must know how to select the good features, remove the bad features of the data or the noisy components, and estimate how many dimensions of the data are good or in other words how many features to select. If the number of dimensions is too large, making it difficult to calculate, it is necessary to reduce the number of dimensions of the data while maintaining the accuracy of the data (reduce dimension).

In this step, the dataset to test on the model is needed to be prepared. Usually, cross-validation will be used to divide the dataset into two parts, one for training (training dataset) and the other for testing purposes on the model (testing dataset). There are two ways commonly used in cross-validation: splitting and k – folding. For the above algorithms, during the training process, the data is divided according to the ratio 6 : 2 : 2, in which 60% of the data is for training and 20% is for the training validation process (validation). And the remaining 20% is for the process of retesting the model after training.

With the dataset consisting of 1307 images (T2-Images) as mentioned above, the image set has been divided according to the ratio 6 : 2 : 2 to serve the training, validation, and testing processes. Specifically, the number of images used includes 813 images for training, of which 414 images do not show brain tumors and 399 images show brain tumors; 239 images for validation, including 121 images showing brain tumors and 118 images not showing brain tumors; 255 images for the test process, including 130 images showing brain tumors and 125 images not showing brain tumors.

Step 2. Classifier model.

The purpose of the training model is to find a function $f(x)$ from which to label the data. This step is often called learning or training.

$$f(x) = y, \tag{9}$$

where x is the feature or input of the data and y is the class label or output.

The classification model used here is the above supervised learning algorithms DenseNet201, ResNet152V2, MobileNetV3, and VGG19.

Step 3. Checking data with model to make prediction.

After finding the classification model in Step 2, in this step, new data will be added to test on the classification model.

Step 4. Evaluating the classification model and selecting the best model.

In the final step, the model will be evaluated by assessing the error level of the testing data and the training data through the found model. If the system results are not as expected, the parameters (turning parameters) of the learning algorithms must be changed to find a better model as well as to test and reevaluate the classification model. From there, it is possible to choose the best classification model for the problem. All steps mentioned above can be described as in Figure 5.

4.3. Evaluation of Experiment Results

4.3.1. Evaluating the Effectiveness of Applying Data Normalization

(1) Results of Training Process. In order to appraise the effectiveness of data normalization, this work evaluates the convergence (accuracy) of network architectures in

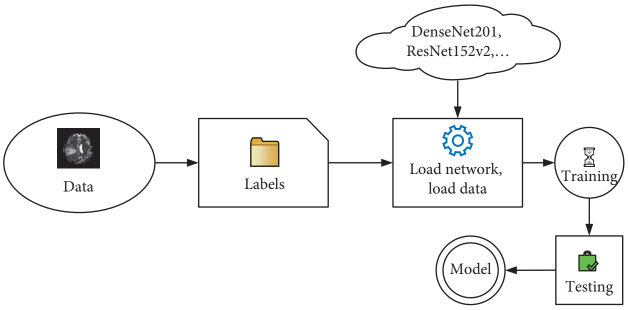


Figure 5: The process of creating a model of the algorithm.

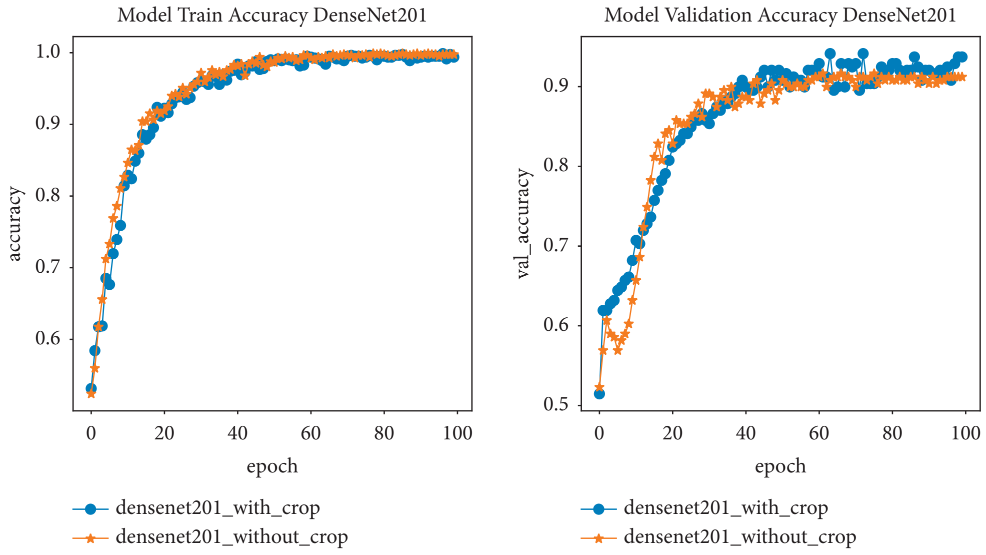


Figure 6: Accuracy of DenseNet201 network architecture before and after normalization.

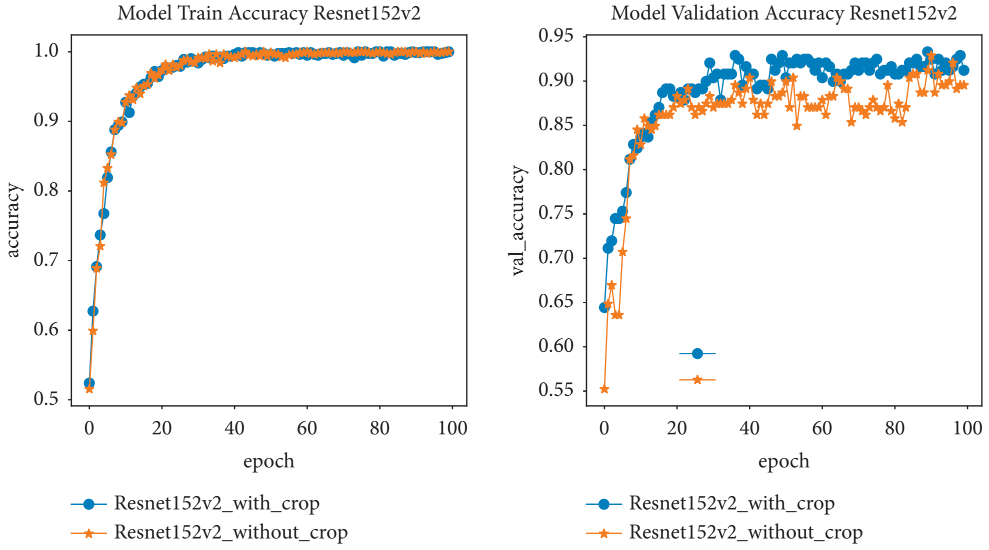


Figure 7: Accuracy of ResNet152V2 network architecture before and after normalization.

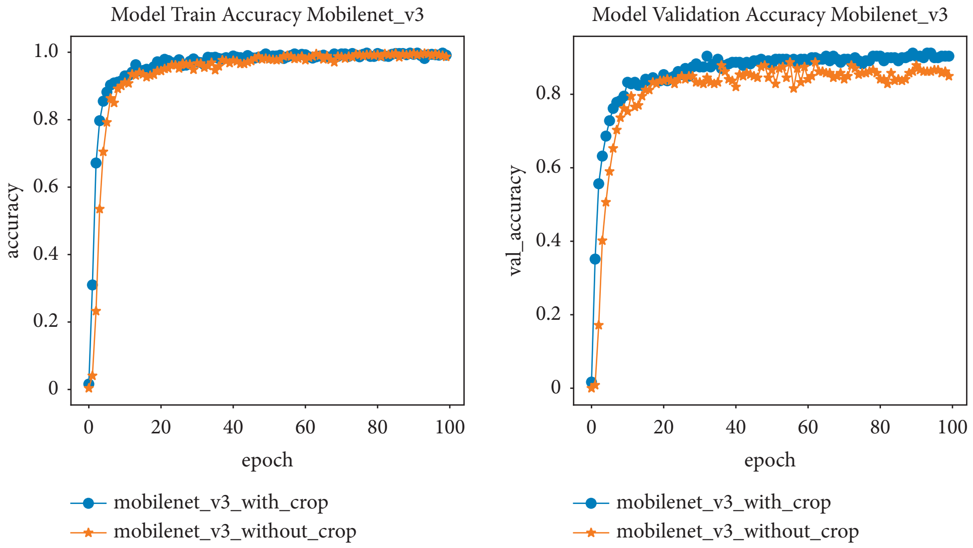


FIGURE 8: Accuracy of MobileNetV3 network architecture before and after normalization.

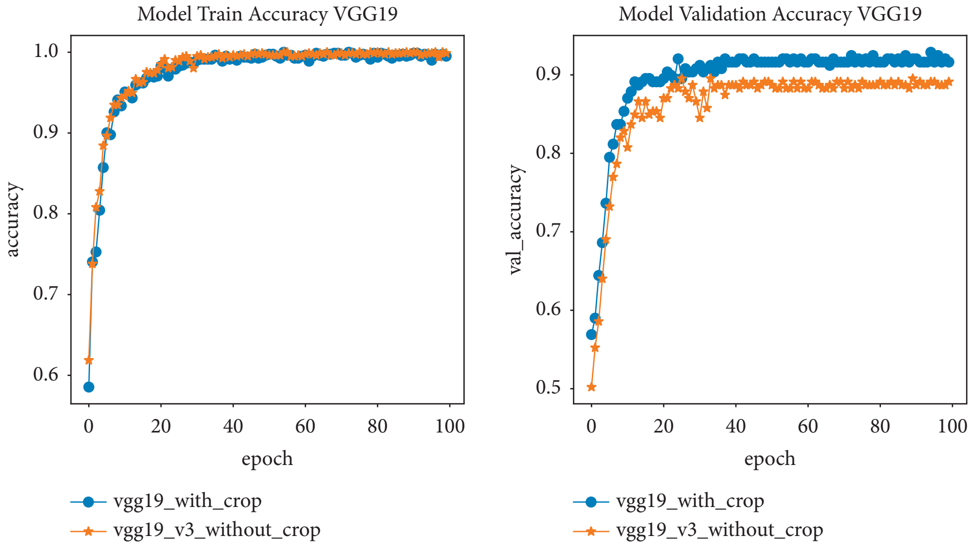


FIGURE 9: Accuracy of VGG19 network architecture before and after normalization.

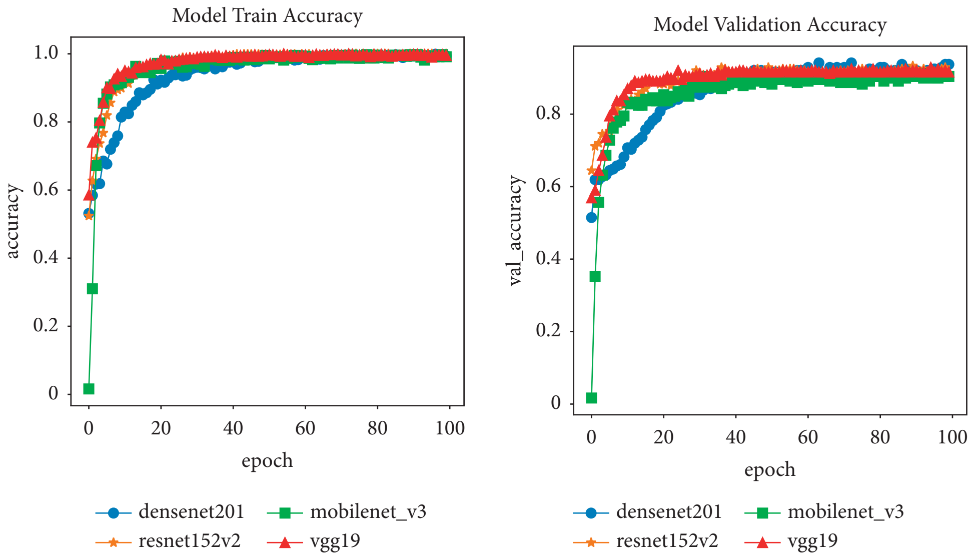


FIGURE 10: Accuracy of network architectures after data normalization.

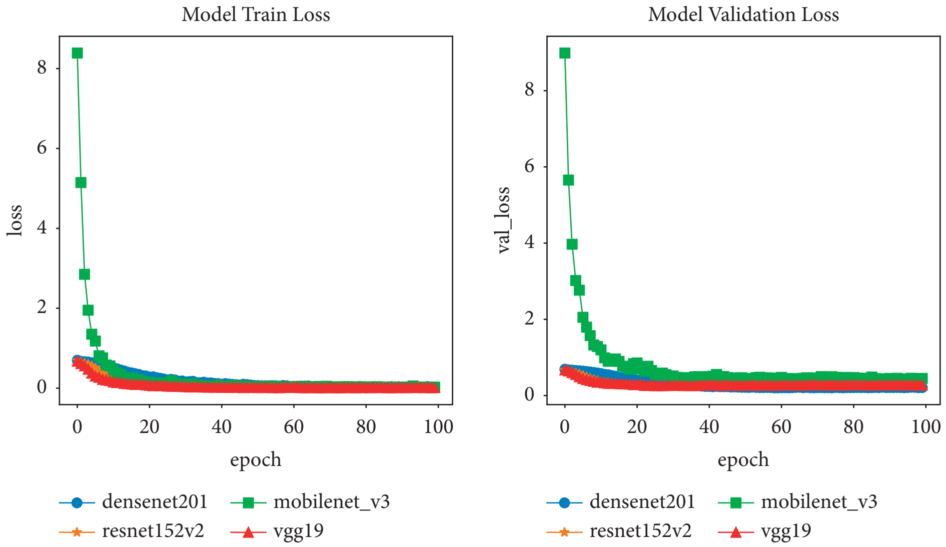


Figure 11: Loss of network architectures after data normalization.

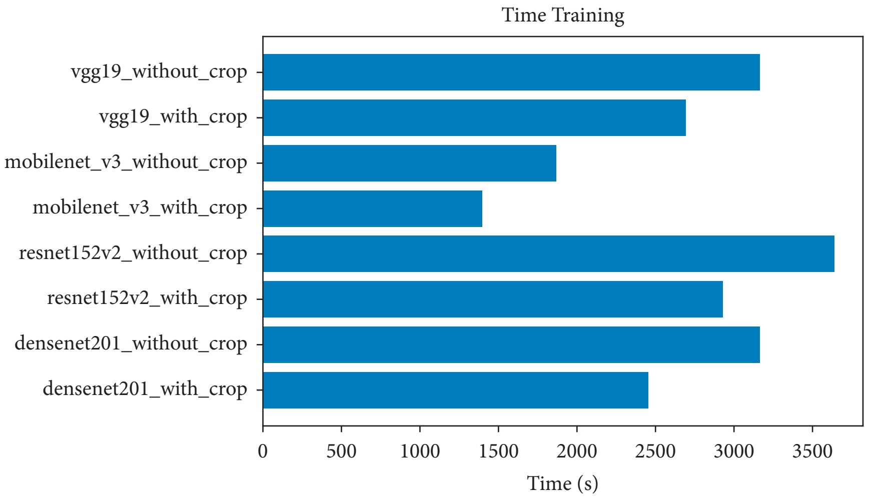


Figure 12: Training time over 100 epochs of network architectures before and after normalization.

Table 4: Table of comparison results between algorithms before and after data normalization.

CNN architectures	Training accuracy		Validation accuracy		Time (s)	
	Before data normalization	After data normalization	Before data normalization	After data normalization	Before data normalization	After data normalization
DenseNet201	99.87%	99.88%	91.63%	94.14%	3165	455
ResNet152V2	99.95%	99.95%	92.86%	93.31%	3639	927
MobileNetV3	99.66%	99.70%	88.70%	91.21%	1869	1398
VGG19	99.91%	99.91%	89.54%	92.88%	3165	2691

classification through 100 epochs. The network architectures used here are DenseNet201, ResNet152V2, VGG19, and MobileNetV3. The optimal algorithm used in the training processes in this section is the ADAM optimization algorithm.

From Figures 6–9, it is easy to see that the training accuracies before and after normalization of the network architectures are almost the same. The specific results of the DenseNet201, ResNet152V2, MobileNetV3, and VGG19 network architectures are 99.88%, 99.95%, 99.7%, and

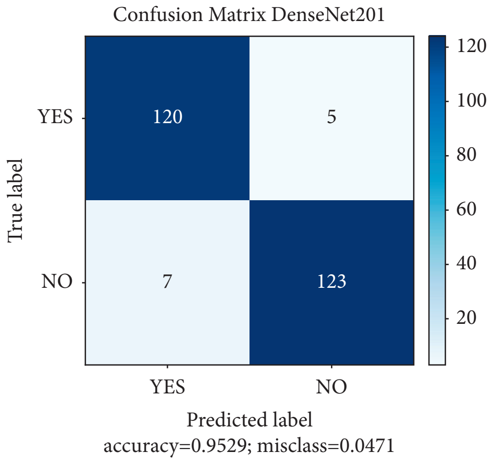


FIGURE 13: DenseNet201 network architecture accuracy based on F1-score.

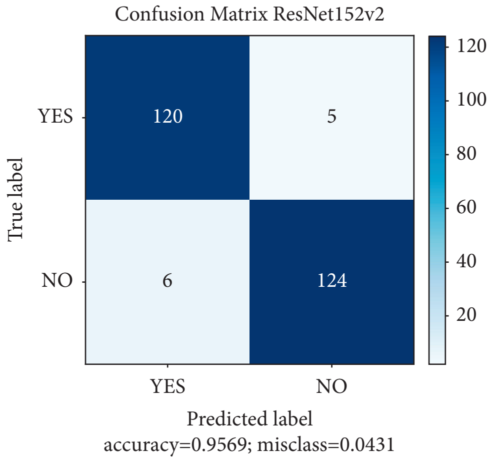


FIGURE 14: ResNet152V2 network architecture accuracy based on F1-score.

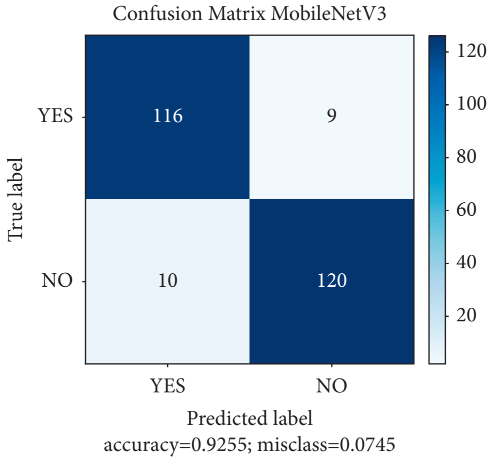


FIGURE 15: MobileNetV3 network architecture accuracy based on F1-score.

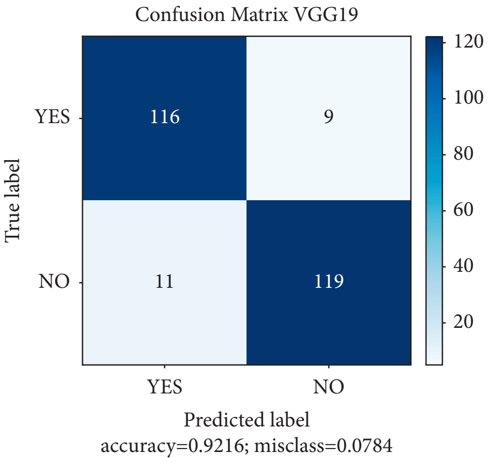


FIGURE 16: VGG19 network architecture accuracy based on F1-score.

TABLE 5: Accuracy of network architectures based on F1-score.

CNN architectures	DenseNet201	ResNet152V2
F1-score	95.29%	95.69%
CNN architectures	MobileNetV3	VGG19
F1-score	92.55%	92.16%

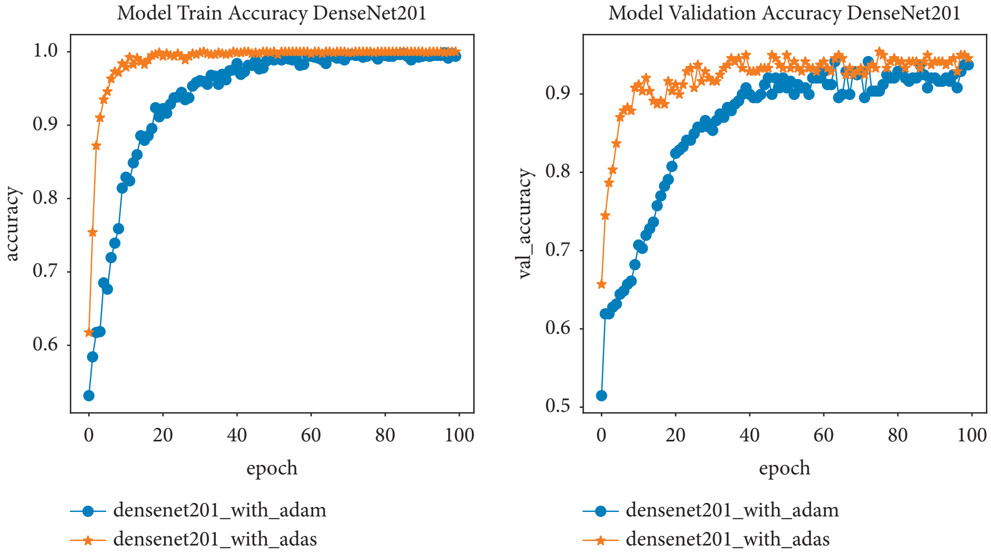


FIGURE 17: Accuracy of DenseNet201 network architectures using ADAM and ADAS optimization algorithms.

99.91%, respectively. However, the results of the validation step of the algorithms showed a marked increase in the accuracy when comparing before and after data normalization. Specifically, the accuracy of the validation process for the DenseNet201 network architecture after normalization is 94.14%, higher than before normalization with an accuracy of 91.63%. The validation result of ResNet152V2 network architecture after normalization has an accuracy of 93.31%, slightly better than before normalization with an accuracy of 92.86%. And this result of ResNet152V2 network architecture after normalization has higher stability than before normalization as presented in Figure 7; Figure 8 indicates that the validation process of MobileNetV3 network architecture after normalization has higher accuracy than before normalization with accuracy of 91.21% and 88.70%, respectively. The validation results of the VGG19 network architecture are similar to those of the three algorithms above with an accuracy of 92.88% after normalization compared to 89.54% before normalization as shown in Figure 9. And it can be seen that all network architectures have convergence with 90% accuracy after only 40 epochs when they use normalized data.

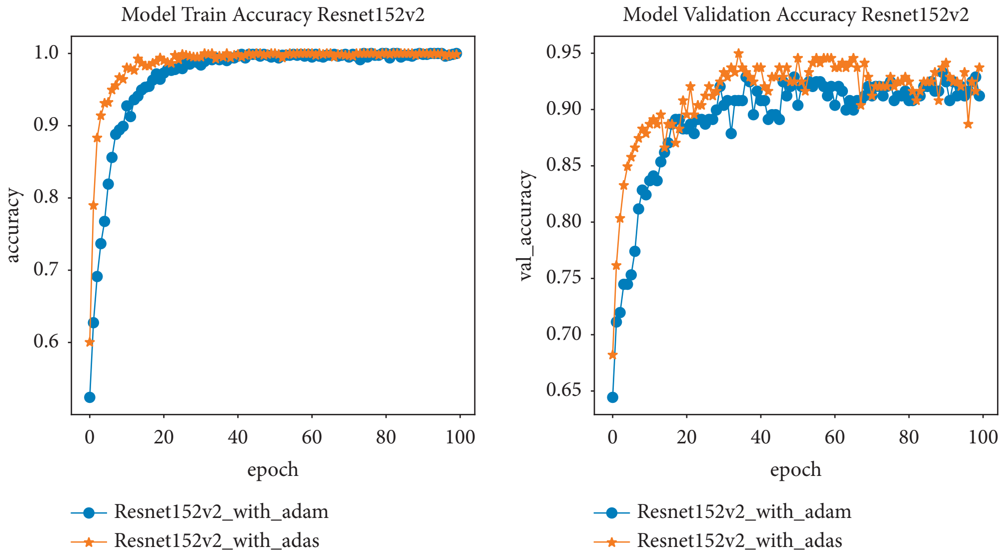


FIGURE 18: Accuracy of ResNet152V2 network architectures using ADAM and ADAS optimization algorithms.

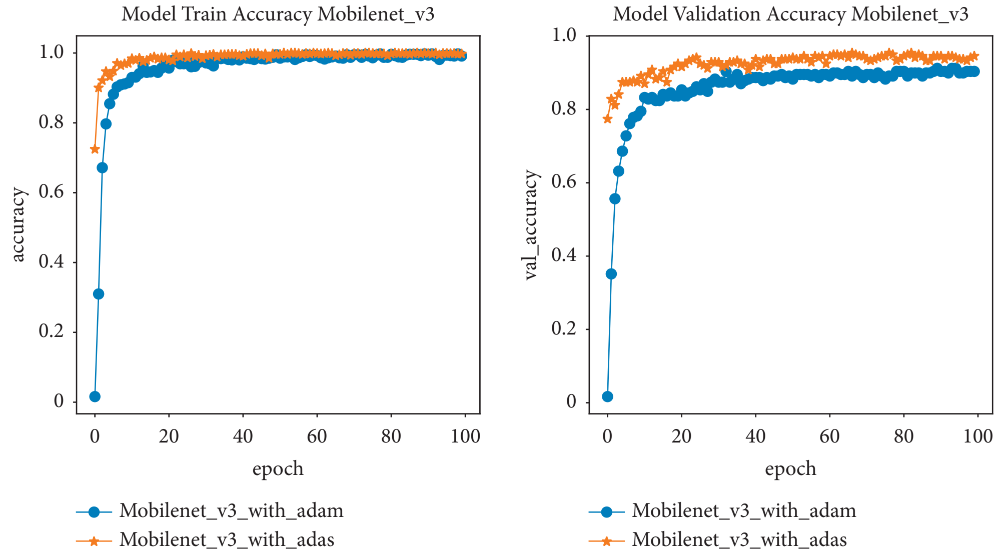


FIGURE 19: Accuracy of MobileNetV3 network architectures using ADAM and ADAS optimization algorithms.

In this paper, in order to be consistent with the collected brain MRI image data, with the ADAM optimal algorithm and the training process with steadily increasing accuracy, where the loss (loss) decreases the most, this study used different learning rates for each network architecture. Specifically, with the network architectures DenseNet201, ResNet152V2, MobileNetV3, and VGG, the initial learning coefficients are $\eta_0 = \{3e-6; 3e-6; 2e-5; 3e-6\}$, respectively. And the results of using these learning coefficients have shown the stability of the training process to avoid overfitting and are shown in Figures 10 and 11.

In practice, it is not always the case that the longer the model training process, the lower the loss function. When it reaches a certain number of epochs, the loss function value will reach saturation; it can no longer decrease and may even increase again. That is overfitting phenomenon. To prevent this phenomenon and free up computational resources, the training process should be stopped right at that saturation point. In this study, as shown in Figure 11, it can be seen that the values of the loss function for all architectures reach saturation when the number of epochs is 100.

When comparing the efficiency of processing speed on the same resource, based on Figure 12, it can be seen that all 4 network architectures give a shorter training time with normalized image data than the training time with denormalized image data. Comparison results between algorithms with datasets before and after normalization are shown in Table 4. Clearly, the results showed that the benefits of

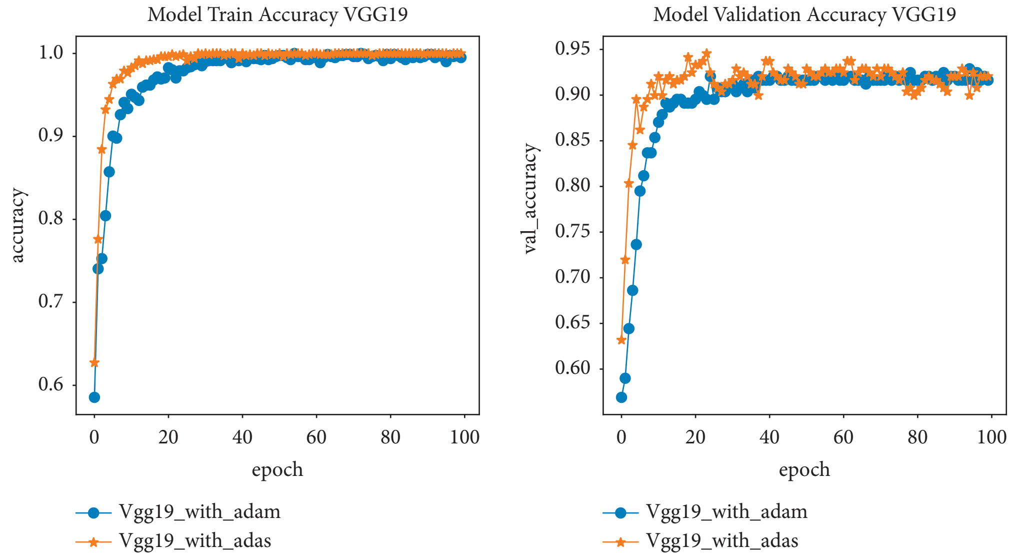


FIGURE 20: Accuracy of VGG19 network architectures using ADAM and ADAS optimization algorithms.

normalizing the image data make the network architectures capable of classifying brain tumors with higher accuracy and shorter training time.

(2) Evaluating the Accuracy of Network Architectures Based on F1-Score. After performing the training process, models of the respective network architectures were generated. In this part, they will be tested for testing on the test dataset. This dataset includes 255 images of which 130 images are showing brain tumors and 125 images are not showing brain tumors.

The results illustrated in Figures 13–16 and the summary data in Table 5 show that all algorithms have an accuracy greater than 92% when based on the F1-score, in which ResNet152V2 network architecture has the highest results. This is expected to be implemented in practice.

4.3.2. Comparing the Accuracy of Models Using ADAM and ADAS Optimal Function

(1) Results of Training Process. In this section, the accuracy of the classification network architectures will be evaluated and compared using the ADAM and ADAS optimization functions. The network architectures will execute the training, validation, and testing processes on the same normalized database with the same computational resources.

Similar to the ADAM optimization algorithm, in order to fit the brain MRI image data, it is suitable for the ADAS optimization algorithm and the training process has a steady increase in accuracy and the most uniform decrease in loss. Each network architecture uses its own learning coefficient. In this study, the learning coefficients of DenseNet201, ResNet152V2, MobileNetV3, and VGG network architectures are $\eta_0 = \{7e - 3; 5e - 3; 4e - 3; 1e - 2\}$, respectively.

With the above input data, the experimental results of network architectures with ADAM and ADAS optimal functions are shown in Figures 17–20, respectively. These results show that the training accuracy of the network architectures using the ADAM and ADAS optimization algorithms are almost the same with the obtained values being greater than 99%. However, for the results of the validation process of the network architectures, the accuracy when implementing the ADAS optimization algorithm has improved significantly in comparison with when using the ADAM optimal algorithm. Specifically, the accuracy of the validation process for the DenseNet201 network architecture using the ADAS optimization algorithm is 95.39% compared to 94.14% when using the ADAM optimization algorithm. And to achieve accuracy, with the ADAM optimal algorithm, the DenseNet201 network needs 40 epochs while with the ADAS optimization algorithm it only needs 10 epochs. The training validation process of ResNet152V2 network architecture using ADAS and ADAM optimization algorithms has the accuracy of 94.47% and 93.31%, respectively. To achieve 90% accuracy, ResNet152V2 network needs 30 epochs when using ADAM optimal algorithm while with ADAS optimal algorithm it only needs 20 epochs. For the MobileNetV3 network architecture, the accuracy of the validation process when using the ADAS optimization function is 95.39% compared to 91.21% when using the ADAM optimization algorithm. The convergence speed for using the ADAS optimization function is also much higher than using the ADAM function, specifically to achieve 90%

TABLE 6: Comparison results between network architectures using ADAM and ADAS optimization algorithms.

CNN architectures	Training accuracy		Validation accuracy		Times (s)	
	ADAM	ADAS	ADAM	ADAS	ADAM	DAS
DenseNet201	99.87%	99.88%	94.14%	95.39%	2455 s	2340 s
ResNet152V2	99.95%	99.97%	93.31%	94.47%	2927 s	3161 s
MobileNetV3	99.66%	99.71%	91.21%	95.39%	1398 s	1234 s
VGG19	99.91%	99.9%	92.88%	94.56%	2691 s	2447 s

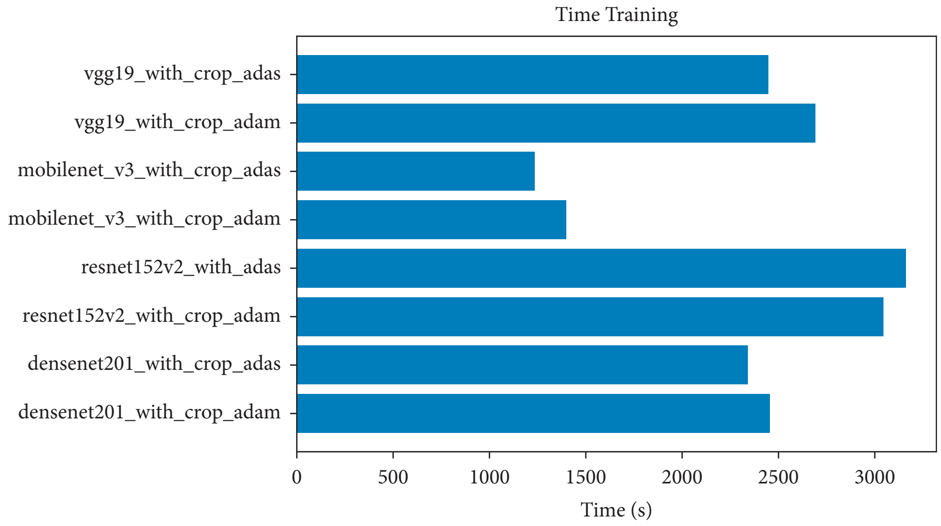


FIGURE 21: Training time over 100 epochs of network architectures with ADAM and ADAS optimization functions.

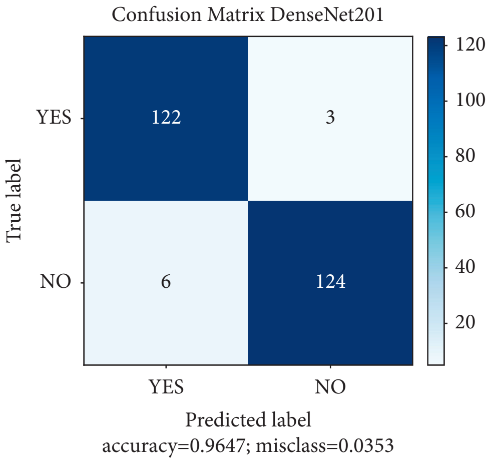


FIGURE 22: Accuracy evaluation matrix of DenseNet201 network architecture using ADAS optimization algorithm via F1-score.

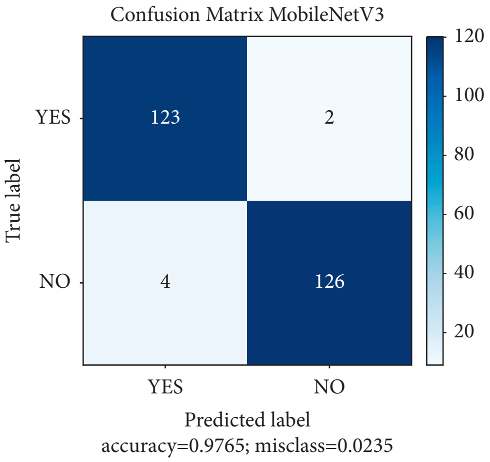


FIGURE 24: Accuracy evaluation matrix of MobileNetV3 network architecture using ADAS optimization algorithm via F1-score.

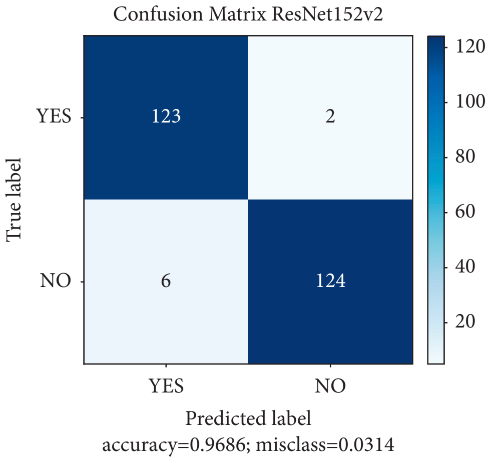


FIGURE 23: Accuracy evaluation matrix of ResNet152V2 network architecture using ADAS optimization algorithm via F1-score.

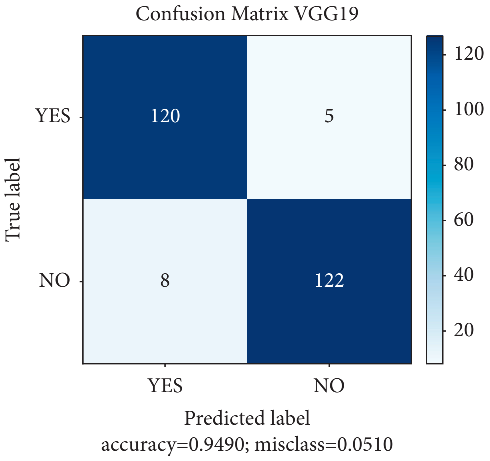


FIGURE 25: Accuracy evaluation matrix of VGG19 network architecture using ADAS optimization algorithm via F1-score.

Table 7: Accuracy of network architectures based on F1-score using ADAS optimization function.

CNN architectures	DenseNet201	ResNet152V2
F1-score	96.47%	96.86%
CNN architectures	MobileNetV3	VGG19
F1-score	97.65%	94.90%

Table 8: Accuracy and F1-score comparison with other previous studies.

Paper	Algorithms	Accuracy	F1-score
Díaz-Pernas et al. 2021 [18]	Multipathway CNN	99.4%	97.3%
Siddiaue et al. 2021 [16]	Proposed DCNN model	96%	97%
Abd El Kader et al. 2021 [19]	Proposed differential deep-CNN	99.25%	95.23%
Tazin et al. 2021 [20]	CNN architectures	Up to 92%	Up to 92%
This paper	**CNN architectures**	**Up to 99.97%**	**Up to 97.5%**

accuracy, with the ADAM MobileNetV3 function requiring 40 epochs while only 10 epochs are required when using the ADAS function. The VGG19 architecture also has the same results as the above architectures with the accuracy of 94.56% and 92.88%, respectively, with the ADAS and ADAM functions. And also with 90% accuracy, the number of VGG19 network architecture epochs needs to be 25 and 11 when using the ADAM and ADAS functions, respectively.

The performance comparison between ADAS and ADAM algorithms is summarized in Table 6. According to this table as well as the above analysis, it is easy to see that the ADAS optimization algorithm has increased the accuracy of the training process; the convergence in the training process also occurs faster. Figure 21 shows the comparison of training time when using 2 optimization functions with the same normalized dataset. Obviously, the model training time when using the ADAS function in most network architectures is faster. Only for ResNet152V2 architecture, the training time with the use of the ADAS function is slightly longer than with the use of the ADAM function. This can also be one of the problems that need to be studied in the future.

(2) Evaluation of F1-Score of Network Architectures Using ADAS Optimization Algorithm. Performing evaluation through F1-score similar to ADAM's algorithm, according to Figures 22–25, the accuracy evaluation through F1-score of network architectures using ADAS optimization function is established as shown in Table 7.

Obviously, when comparing the synthetic results presented in Tables 5 and 7, it is easy to see that the ADAS optimization algorithm has significantly increased the accuracy of the aforementioned models, in which the MobileNetV3 network model gives the highest accuracy of 97.65%. Combined with the results analyzed above, for the problem of brain tumor identification on MRI-T2 images, the ADAS optimization algorithm has significantly improved the accuracy of the training, validation, and testing processes of all the models surveyed in this work as well as shortening the training time of those models compared to the ADAM algorithm.

4.3.3. Comparison of Results. The performance of the proposed system in our study will be compared with the most recently published studies mentioned above. The results of that comparison are shown in Table 8. Based on this table, it is easy to see that the proposed system gave better results in both accuracy and F1-score than other studies with the same subjects. Obviously, although using the same variants of the DCNNs family, the data normalization and the ADAS optimization function helped to significantly improve the performance of the proposed system compared to those other systems.

5. Conclusion

This article has focused on deploying the application of artificial intelligence algorithms in classifying brain tumor patients and normal people using human brain MRI images. The dataset used is MRI images of Vietnamese people, including 123 patients and 100 healthy people. The four algorithms that are experimentally compared in the study are DenseNet201, ResNet152V2, MobileNetV3, and VGG19. The experimental results in the study have shown that the normalization of the initial data processing is very important when it has significantly increased the accuracy in classifying and detecting patients as well as reducing the training time of those models. On the other hand, the paper has also shown the efficiency of the ADAS optimization function compared with the very popular ADAM optimization function. In particular, the ADAS algorithm has advantages in comparison with the ADAM function in improving accuracy as well as reducing model training time. Of the four algorithms mentioned above, the MobileNetV3 algorithm is the most efficient. This can be considered as the foundation for implementing the above system in practice. However, the system also has the disadvantage that the dataset is still small. In the future, besides collecting more data to increase the accuracy of the system, the research will also develop methods to specifically classify those tumor types according to their tumor characteristics (benign or malignant) or by type of disease.

References

[1] J. C. Buckner, P. D. Brown, B. P. O'Neill, F. B. Meyer, C. J. Wetmore, and J. H. Uhm, "Central nervous system tumors," *Mayo Clinic Proceedings*, vol. 82, no. 10, 2007.

[2] A. Lashkari, "A neural network based method for brain abnormality detection in MR images using Gabor wavelets," *International journal of computer Applications*, vol. 4, no. 7, pp. 9–15, 2010.

[3] M. Vargo, "Brain tumor rehabilitation," *American Journal of Physical Medicine & Rehabilitation*, vol. 90, no. 5, pp. S50–S62, 2011.

[4] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, November 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, December 2016.

[6] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[8] S. Khawaldeh, U. Pervaiz, A. Rafiq, and R. S. Alkhawaldeh, "Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks," *Applied Sciences*, vol. 8, no. 1, p. 27, 2018.

[9] J. S. Paul, A. J. Plassard, B. A. Landman, and F. Daniel, "Deep learning for brain tumor classification," in *Proceedings of the Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10137, International Society for Optics and Photonics, Bellingham, WA, USA, March 2017.

[10] N. Varuna Shree and T. N. R. Kumar, "Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network," *Brain informatics*, vol. 5, no. 1, pp. 23–30, 2018.

[11] D. J. Hemanth, J. Anitha, A. Naaji, O. Geman, D. E. Popescu, and L. Hoang Son, "A modified deep convolutional neural network for abnormal brain image classification," *IEEE Access*, vol. 7, pp. 4275–4283, 2019.

[12] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Computers in Biology and Medicine*, vol. 111, Article ID 103345, 2019.

[13] S. Das, O. F. M. R. R. Aranya, and N. N. Labiba, "Brain tumor classification using convolutional neural network," in *Proceedings of the International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), IEEE, Dhaka, Bangladesh,*, December 2019.

[14] Z. Ullah, M. U. Farooq, S.-H. Lee, and D. An, "A hybrid image enhancement based brain MRI images classification technique," *Medical Hypotheses*, vol. 143, Article ID 109922, 2020.

[15] A. Çinar and M. Yildirim, "Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture," *Medical Hypotheses*, vol. 139, Article ID 109684, 2020.

[16] M. A. B. Siddiaue, S. Sakib, M. M. R. Khan, A. K. Tanzeem, M. Chowdhury, and N. Yasmin, "Deep convolutional neural networks model-based brain tumor detection in brain MRI images," in *Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Palladam, India, November 2020.

[17] P. Saxena, A. Maheshwari, and S. Maheshwari, "Predictive modeling of brain tumor: a Deep learning approach," *Innovations in Computational Intelligence and Computer Vision Advances in Intelligent Systems and Computing*, Springer, Berlin, Germany, pp. 275–285, 2021.

[18] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodrígue, and D. González-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," *Healthcare*, vol. 9, no. 2, 2021.

[19] I. Abd El Kader, G. Xu, Z. Shuai, S. Saminu, I. Javaid, and I. Salim Ahmad, "Differential deep convolutional neural network model for brain tumor classification," *Brain Sciences*, vol. 11, no. 3, p. 352, 2021.

[20] T. Tazin, S. Sarker, P. Gupta et al., "A robust and novel approach for brain tumor classification using convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 2392395, 11 pages, 2021.

[21] J. Seetha and S. S. Raja, "Brain tumor classification using convolutional neural networks," *Biomedical and Pharmacology Journal*, vol. 11, no. 3, pp. 1457–1461, 2018.

[22] N. M. Balasooriya and R. D. Nawarathna, "A sophisticated convolutional neural network model for brain tumor classification," in *Proceedings of the IEEE International Conference on Industrial and Information Systems (ICIIS), IEEE, Peradeniya, Sri Lanka*, February 2017.

[23] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *European Radiology*, vol. 12, no. 4, pp. 920–927, 2002.

[24] A. Haase, "Snapshot flash mri. applications to t1, t2, and chemical-shift imaging," *Magnetic Resonance in Medicine*, vol. 13, no. 1, pp. 77–89, 1990.

[25] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[27] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, October 2015.

[28] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[29] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, PMLR, Long Beach, CA, USA, May 2019.

[30] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 2020.

[31] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, https://arxiv.org/abs/1511.08458.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA*, June 2009.

[33] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, https://arxiv.org/abs/1609.04747.

[34] L. Bottou, "Stochastic gradient descent tricks," *Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, pp. 421–436, 2012.

[35] M. S. Hosseini and K. N. Plataniotis, "Adas: adaptive scheduling of stochastic gradients," 2020, https://arxiv.org/abs/2006.06587.

[36] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, p. 7, 2011.

[37] N. Jegadeesh and S. Titman, "Momentum," *Annual Review of Financial Economics*, vol. 3, no. 1, pp. 493–509, 2011.

[38] D. P. Kingma and Ba. Jimmy, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[39] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proceedings of the European Conference on Information Retrieval*, Santiago de Compostela, Spain, March 2005.

An Efficient Method for Diagnosing Brain Tumors Based on MRI Images Using Deep Convolutional Neural Networks

Soumya Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, soumyamohanty614@gmail.com*

Mahendra Kumar Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, mk_sahoo34@hotmail.com*

Sunita Baral, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, sunita.baral95@gmail.com*

Ajit Kumar Panda, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, akpanda555@hotmail.com*

Abstract:

This paper proposes a system to effectively identify brain tumors on MRI images using artificial intelligence algorithms and ADAS optimization function. This system is developed with the aim of assisting doctors in diagnosing one of the most dangerous diseases for humans. The data used in the study is patient image data collected from Bach Mai Hospital, Vietnam. The proposed approach includes two main steps. First, we propose the normalization method for brain MRI images to remove unnecessary components without affecting their information content. In the next step, Deep Convolutional Neural Networks are used and then we propose to apply ADAS optimization function to build predictive models based on that normalized dataset. From there, the results will be compared to choose the most optimal method. Those results of the evaluated algorithms through the coefficient F1-score are greater than 94% and the highest value is 97.65%.

1. Introduction

The brain is a particularly important organ, the control center of the central nervous system, coordinating the activities of all organs and parts in the human body. The brain has a complex structure and is protected and covered by the skull, a very hard bone box. However, a rigid skull may help protect the brain parenchyma from minor trauma but does not prevent the development of lesions and abnormal structures within the brain. One of the brain diseases of primary concern in medicine is brain tumors. A brain tumor is a condition in which abnormal cells grow in the brain. Brain tumors are divided into two types: benign brain tumors and malignant brain tumors (called cancer) [1]. Whether it is a benign brain tumor or a malignant brain tumor, it affects brain cells, causing brain damage and being even life-threatening. There are about 120 different types of brain tumors, most of which are tumors in the brain tissue, in addition to tumors in the meninges, pituitary gland, cranial nerves. Any form of brain tumor can be dangerous for the patient. Tumors in brain tissue or benign brain tumors often progress slowly; the symptoms of brain tumors in this case will also appear slower and more insidious. In contrast, if the brain tumor grows rapidly, the patient will feel the symptoms more pronounced in both frequency and extent. With current medical capabilities, early detection of abnormal structures in the patient's brain can improve the likelihood of successful treatment and limit the sequelae of tumors to the brain in general and the patient's health in particular.

The detection of brain tumors today is mostly based on the ability of doctors to distinguish abnormalities on MRI images which is a type of high-quality image in the field of imaging [2]. This is a process that requires a lot of experience and concentration to detect and classify brain diseases and brain tumors. From brain MRI, it is possible to diagnose and recognize many different types of brain tumors and offer appropriate treatment methods [3]. However, the increasing number of patients with the large number of images obtained has become a major challenge in the field of diagnostic imaging, a field that requires rapid and accurate evaluation of results by doctor. Artificial intelligence technology will help classify diseases from MRI images quickly and bring high accuracy in disease diagnosis. The classification of diseases based on MRI images has not been too difficult with high accuracy due to the introduction of GPUs (Graphics Processing Unit) and image processing based on artificial intelligence (AI).

This research focuses on the application of image preprocessing techniques and the development of algorithms using convolutional neural network (CNN) models, which are advanced deep learning models such as DenseNet201 [4], ResNet152V2 [5], MobileNetV3 [6], and VGG19 [7]. At the same time, the research also focuses on developing and applying the ADAS optimization algorithm to improve the accuracy in classifying normal people and brain tumor patients. The dataset in this work includes 1307 brain MRI images in JPEG format that are manually classified by specialists into 2 categories: normal human brain MRI images and brain MRI images of people with brain tumor disease. The comparison of all experimental results will evaluate the effectiveness of each model.

The article is organized as follows. Section 2 presents the previously conducted MRI brain tumor classification studies. Section 3 provides an overview of brain MRI images and the CNN algorithm models used. Section 4 presents the experimental results and gives evaluation for each algorithm. Conclusions and future work are outlined in Section 5.

2. Related Work

Several technical methods related to brain MRI images classification since 2017 based on different classification models are summarized in Table 1. They are divided into two basic methods: using CNN network architecture and not using CNN network architecture. In [10], the authors divided brain MRI images into two categories: normal images and images with abnormal signs. They used GLCM to get the features of the MRI images; then a probabilistic neural network (PNN) was used to classify the MRI brain images of people as normal or abnormal. As a result, they obtained a classification model with an accuracy of 95%. In [14], Ullah et al. proposed a scheme to classify the brain MRI images of normal people and patients using equilibrium histograms, discrete wavelet transforms, and Feedforward Artificial Neural Networks. Recently, deep learning method has been widely used for the classification of brain tumors on MRI images [8, 9]. The deep learning method does not need to manually extract the features of the images; it combines the extraction and classification stages in the self-learning process. The deep learning method requires a dataset where normalized processing of the MRI images is sometimes required, and then salient features are identified during machine learning [13].

Convolution Neural Network (CNN), one of the well-known deep learning techniques for image data, can be used as a feature extraction tool from which to capture related features to perform data classification task. Feature maps in the initial and higher layers of the CNN model extract low-level features and specific features of high-level content, respectively. Feature maps in the earlier layer construct simple structural information, such as shapes, textures, and edges, while the higher layers combine these low-level features into constructing (encoding) expressions performance, integrating local and global information.

Various researchers have proposed to use CNN to classify brain tumors based on brain MRI image datasets [11, 21, 22]. Deepak and Ameer [12] used pretrained GoogLeNet to extract features from brain MRI images with CNN network architecture to classify three types of brain tumors and obtained up to 98% accuracy. Çinar and Yildirim, [15] modified the ResNet50 network based on the pretrained CNN network architecture by removing the last 5 layers and adding 8 new layers, and that method achieved 97.2% accuracy. Saxena et al. [17] used InceptionV3, ResNet50, and VGG16 network architectures with legacy methods to classify brain tumor data. In this study, ResNet50 model obtained the highest accuracy rate with 95%. Díaz-Pernas et al. [18] presented a CNN network architecture for automatic brain tumor segmentation such as glioma, meningioma, and pituitary tumor. They evaluated their proposed model using the T1-weighted contrast-enhanced MRI dataset and obtained an accuracy of 97.3%.

Siddiaue et al. [16] proposed a model based on modified vgg-16 network architecture for brain tumor images classification which achieved an accuracy of 96% and an F1-score of 97%. Abd El Kader et al. [19] developed a differential deep-CNN-based model to classify MRI images with and without tumors. In fact, this model was still based on the basic CNN architecture but obtained an accuracy of 99.25% and an F1-score of 95.23%. In [20], the authors successfully deployed transfer learning for some variant architectures of CNN to apply to the classification of MRI images with and without brain tumors, in which MobileNetV2 had an accuracy of 92% and F1-score of 92%; InceptionV3 had an accuracy of 91% and an F1-score of 90.98%; VGG19 had an accuracy of 88% and F1-score of 88.18%.

In summary, as observed from the above studies, the accuracy obtained by using deep learning with CNN network architecture to classify brain MRI is significantly higher than that of the old traditional techniques. However, deep learning models require a large amount of data to train in order to perform better than traditional machine learning techniques.

3. Materials and Methods

3.1. Brain Tumor MRI Images

3.1.1. Content Contained in MRI Images. The commonly used standard for MRI images today is DICOM, an acronym for Digital Imaging and Communications in Medicine Standards [23]. This is an industry standard system developed to meet the needs of manufacturers and users in connecting, storing, exchanging, and printing medical images.

As for the DICOM image format standard, in addition to the image files, it also includes header files as in Figure 1.

Table 1: Summary of studies on brain tumor classification.

Author	Classification method	Objective	Dataset	Feature extraction method	Accuracy
Khawaldeh et al. [8]	CNN	Classification of brain MRI into normal and abnormal	587 MR images	CNN	91.16%
Paul et al. [9]	Fully connected and CNN	Brain tumor classification of MR brain image	3064 MR images	CNN	91.43%
Varuna Shree and Kumar [10]	Probabilistic neural network (PNN)	Classification of brain MRI into normal and abnormal	650 MR images	Gray level cooccurrence matrix (GLCM)	95%
Hemanth et al. [11]	CNN	Classification into normal and abnormal	220 MR images	CNN	94.5%
Deepak and Ameer [12]	Deep transfer learning	Classification of glioma, meningioma, and pituitary tumors	3064 MR images	Google Net	98%
Das et al. [13]	CNN	Brain tumor classification	3064 MR images	CNN	94.39%
Ullah et al. [14]	Feedforward neural network	Classification of brain MRI into normal and abnormal	71 MR images	DWT	95.8%
Çinar and Yildirim [15]	CNN models	Brain tumor detection and classification	253 MR images	CNN	97.2%
Siddiaue et al. [16]	Proposed DCNN model	Brain tumor classification	253 MR images	CNN	96%
Saxena et al. [17]	CNN networks with transfer learning	Binary classification of brain tumor normal and abnormal	253 MR images	CNN	95%
Díaz-Pernas et al. [18]	Multipathway CNN	Brain tumor classification	3064 MR images	CNN	97.3%
Abd El kader et al. [19]	Proposed differential deep-CNN	Brain tumor classification	25000 MR images	CNN	99.25%
Tazin et al. [20]	CNN architectures	Brain tumor classification	2513 MR images	CNN	Up to 92%

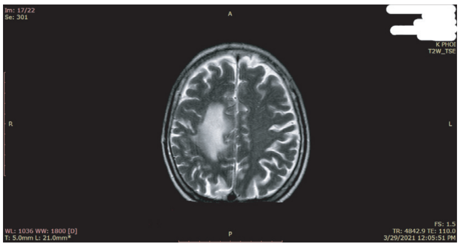


Figure 1: Actual DICOM image (patient information has been removed).

Although stored in different files, when displayed, the header information is displayed along with the MRI image information via a “DICOM browser.” Data in MRI images include demographic information, patient information, parameters acquired for imaging studies, image size, and image matrix size. The patient’s information displayed includes patient’s first and last name, gender, age, date of birth, and place where the MRI scan was performed.

3.1.2. The Role of MRI Images in the Diagnosis of Brain Tumors. Magnetic resonance imaging of the brain [24] can very clearly detect and describe abnormalities in the brain parenchyma in general such as vascular tumors, arterial occlusion, and invasion of the venous sinuses as well as the relationship between tumor and surrounding structures. There are three basic image formats of MRI images: T1W, T2W, and T2 Flair. They are used in specific cases depending on the situation of the disease.

T1W imaging is mainly used to identify necrotic tumors, hemorrhage in tumors, or cysts. For example, with MRI images in meningiomas, on T1-weighted images, most meningioma shows no difference in signal intensity compared with cortical gray matter.

For image in the T2W phase, the received signal has been changed completely; it is a fairly homogeneous gain signal block. Imaging is also helpful in evaluating hemorrhages and cysts. In particular, the role of the T2W phase is very useful in reflecting the homogeneity of benign soft tumors or meningiomas.

For Fluid-Attenuated Inversion Recovery (T2-FLAIR), this type of phase image is very useful to evaluate the consequences and effects of edema. Although this finding is not specific for meningiomas in particular, it is very meaningful in the diagnosis as well as the long-term prognosis for the patient.

Overall, the sensitivity and specificity of MRI are very high in the diagnosis of meningiomas. MRI has been shown to be superior in tumor delineation by its relationship to surrounding structures.

3.2. Model Architectures

3.2.1. Supervised Learning. Supervised learning [25] is an algorithm that predicts the output (outcome) of a new data (new input) based on previously known (input, outcome) pairs. This data pair is also known as (data, label). Supervised learning is the most popular group of machine learning algorithms.

Mathematically, supervised learning consists of a set of input variables $X = \{x_1, x_2, \ldots, x_N\}$ and a corresponding set of labels $Y = \{y_1, y_2, \ldots, y_N\}$, where x_i and y_i are vector. The data pairs $(x_i, y_i) \in X \times Y$ are called the training dataset. From this training dataset, we need to create a function that maps each element from the set X to a corresponding (approximate) element of the set Y as

$$y_i \approx f(x_i), \quad \forall i = 1, 2, \ldots, N. \tag{1}$$

The goal is to approximate the function f very well so that when we have a new data x we can compute its corresponding label:

$$y = f(x). \tag{2}$$

A problem is called classification if the labels of the input data are divided into a finite number of groups.

3.2.2. Convolutional Neural Network Architectures. Convolutional Neural Network (CNN) [31] is one of the most popular and most influential deep learning models in the computer vision community. CNN is used in many problems such as image recognition and video analysis or for problems in the field of natural language processing and solves most of these problems well.

CNN includes a set of basic layers such as convolution layer, nonlinear layer, pooling layer, and fully connected layer. These layers are linked together in a certain order. Basically, an image will be passed through the convolution layer and nonlinear layer first; then the calculated values will be passed through the pooling layer to reduce the number of operations while preserving the characteristics of the data. The convolution layer, nonlinear layer, and pooling layer can appear one or more times in the CNN network. Finally, the data is passed through fully connected network and soft-max to calculate the probability of object classification.

Table 2 summarizes some typical CNN network architectures since 2012. To evaluate and compare network structures, two parameters are used, Top 1 Accuracy and Top 5 Accuracy. In the case of Top 1 Accuracy, the correct model's prediction must be the model that predicts the class with the highest probability. In the case of Top 5 Accuracy, the correct model's prediction is the model that correctly predicts one of the 5 classes with the highest probability.

In this study, four different network architectures are used: DenseNet201 [4], ResNet152V2 [5], MobileNetV3 [6], and VGG19 [7]. All the above four network architectures are developments and upgrades based on the basic network architecture CNN, one of the advanced deep learning models for image classification that has been verified with high accuracy on image sets, ImageNet [32]. These CNN variant network architectures are widely used in image recognition and classification problems. All four network architectures have a structure consisting of two basic layers, the feature extraction layer and the classifier layer. In this research, the input to the network architecture is a 256 × 256 brain MRI image containing information with or without brain tumors. The feature extraction layer has the role of extracting features of brain MRI images such as white matter, gray matter, cerebrospinal fluid, cerebral cortex, and brain tumor. Then, the classification layer is responsible for synthesizing the features of brain MRI images, giving specific features of images with tumors and images without tumors to serve the classification process.

3.3. Optimal Algorithms. The optimization algorithm is the basis for building a neural network model with the aim of "learning" the features (or patterns) of the input data, from which it is possible to find a suitable pair of weights and biases to optimize the model. But the question is how to "learn?" Specifically, how the weights and biases are found, not just randomly taking the weights and biases values for a finite number of times and hoping after some steps a solution can be found. Therefore, it is necessary to find an algorithm to improve weights and biases step by step, and that is why optimizer algorithms were created.

Some of the factors commonly used to evaluate an optimizer algorithm are as follows:

Fast convergence (Training Process)

High generalization (can recognize previously untrained patterns)

High accuracy

The popular optimization algorithms are listed in Figure 2. These algorithms are GD-Gradient Descent [33], SGD-Stochastic Gradient Descent [34], ADAS-Adaptive Scheduling of Stochastic Gradients [35], AdaGrad-Adaptive Subgradient Methods for Online Learning and Stochastic Optimization [36], Momentum [37], RMSProp [33], and ADAM-Adaptive Moment Estimation [38].

Among the above algorithms, the Optimal Algorithms belonging to the adaptive family usually have fast convergence speed. Meanwhile, algorithms belonging to the SGD family often have high generalization. However, this study only focuses on the development and application of ADAM and ADAS algorithms.

3.3.1. ADAM Algorithm: A Method for Stochastic Optimization. ADAM is a combination of Momentum and RMSProp. One of the key components of ADAM is exponential weighted moving averages (also known as leak averages) that estimate both the momentum and the second-order moment of the gradient. Specifically, it uses state variables as follows:

$$\begin{aligned} v &\leftarrow \beta_1 v_t - 1 + (1 - \beta_1) g_t, \\ s &\leftarrow \beta_2 s_t - 1 + (1 - \beta_2) g_t^2, \end{aligned} \tag{3}$$

where v is the first moment vector, s is the second moment vector, β_1 and β_2 are the jump parameters at the initial and the second points in ADAM's algorithm, t is the time for the correction steps, and g is the gradient. Here β_1 and β_2 are nonnegative weight parameters. Popular choices for them are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. This means that the variance estimate moves much slower than the momentum term.

Table 2: Some CNN architectures and ImageNet benchmark (image classification).

CNN architectures	Author	Top 1 accuracy (%)	Top 5 accuracy (%)
AlexNet	Krizhevsky et al. [26]	63.3	84.6
VGG16	Simonyan and Zisserman [7]	74.4	91.9
VGG19		74.5	92.0
GoogLeNet (InceptionV1)	Szegedy et al. [27]	69.8	89.5
ResNet50	He et al. [5]	77.1	93.3
ResNet152		78.6	94.3
DenseNet201	Huang et al. [4]	77.4	93.7
MobileNet224	Howard et al. [6]	70.6	89.5
MobileNetV3		79.0	94.5
ResNeXt101	Xie et al. [28]	80.9	95.6
EfficientNet-L2	Tan and Le [29]	90.2	98.8
RegNet-Y	Radosavovic et al. [30]	79.9	95.0

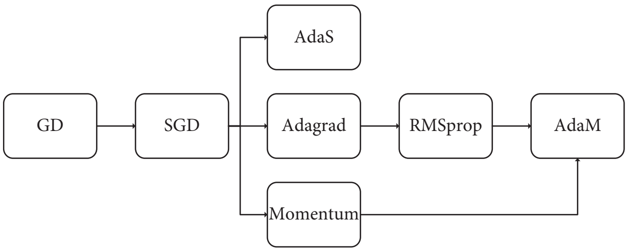


Figure 2: The development of optimization algorithms.

Note that if initializing the values $v_0 = s_0 = 0$, the algorithm will have a significant initial bias towards smaller values. This problem can be solved using $\sum_{i=0}^{t}(1-\beta_t)/(1-\beta)$ to normalize the terms. Similarly, state variables are normalized as follows:

$$\begin{aligned} v_t' &= \frac{v_t}{1-\beta_1^t}, \\ v_t' &= \frac{s_t}{1-\beta_2^t}. \end{aligned} \tag{4}$$

From the appropriate estimates, the updated equations can be established. First, the gradient value will be adjusted, similar to that in RMSProp [33] to get

$$g_t' = \frac{\eta s_t}{\sqrt{s_t'} + \varepsilon}, \tag{5}$$

where ε is a constant and it is chosen to be $\varepsilon = 10^{-6}$ to balance arithmetic stability and reliability, and η is the learning rate. From there, the update step is defined as follows:

$$x_t \leftarrow x_t - 1 - g_t'. \tag{6}$$

When looking at the design of ADAM, the inspiration of the algorithm is clear. Momentum and range are clearly represented in the state variables. Moreover, based on RMSProp it is easy to see that the combination of both terms is quite simple. Finally, the learning rate η allows us to control the update step length to solve convergence problems.

3.3.2. ADAS Algorithm: Adaptive Scheduling of Stochastic Gradients. ADAS [35] is an optimization algorithm belonging to the family of Stochastic Gradient Descent (SGD) algorithms. The updated rules for ADAS are established using SGD with momentum as follows:

$$\begin{aligned} &\eta(t,l) \leftarrow \beta.\eta(t-1,l) + \zeta \cdot [\overline{G}(t,l) - \overline{G}(t-1,l)], \\ &v_l^k \overline{G} \alpha \cdot v_l^{k-1} - \eta(t,l) \cdot g_l^k, \\ &\theta_l^k \overline{G} \theta_l^{k-1} + v_l^k, \end{aligned} \tag{7}$$

where η is the learning rate, t is the time for the correction steps, β is ADAS gain factor, ζ is the knowledge gain hyperparameter, k is the current minibatch, t is the current epoch iteration, l is the convolution block index, $\overline{G}(\cdot)$ is the average knowledge gain obtained from both mode-3 and mode-4 decompositions, v is the velocity term, and θ is the learnable parameter. The learning rate is calculated relative to the rate of change of knowledge acquired after the training epochs. The learning rate $\eta(t,l)$ is then further updated by an exponential moving average called the gain factor, with the hyperparameter β, to accumulate the history of the knowledge gained over the series epochs. In fact, β controls the trade-off between convergence rate and training accuracy of ADAS.

ADAS is an adaptive optimization tool for scheduling the learning rate in the training of a CNN network. ADAS exhibits a much faster convergence speed than other optimization algorithms. ADAS demonstrated generalization characteristics (low test loss) on par with SGD-based optimizers, improving on the poor generalization characteristics of adaptive optimizers. In addition to optimization, ADAS introduces new polling metrics for CNN layer removal (quality metrics).

3.4. Accuracy and F1-Score. The classification problem in this study is a binary classification problem, in which one class is an MRI image with a brain tumor and the other is an MRI image without a brain tumor. This study considers the image class with brain tumor to be positive and the remaining image class without brain tumor to be negative. The parameters True Positive (TP), False Positive (FP), True

Negative (TN), and False Negative (FN) are described as in Table 3.

In this paper, the parameters used to evaluate the effectiveness of the model are accuracy, precision, recall, and F1-score [39]. When building a classification model, the ratio of correctly predicted cases to the total number of cases is always considered. That ratio is called accuracy. Precision is the answer to the question: how many true positives are there out of the total number of positive diagnoses? Recall measures the rate of correctly predicting positive cases across all samples in the positive group. F1-score is the harmonic mean between precision and recall. Therefore, in situations where the precision and recall are too different, the F1-score will balance both values and help us to make an objective assessment. Accuracy, precision, recall, and F1-score are defined as the following equations:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{total sample}} \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{8}$$

4. Experiments and Results

This study will compare the results of the network architectures DenseNet201, ResNet152V2, MobileNetV3, and VGG19 in the cases before and after data normalization with the ADAM optimization function. Then, the study will specifically compare the performance of the above algorithms with the ADAM and ADAS optimization functions on the same normalized dataset.

4.1. Collecting and Normalizing Data

4.1.1. Collecting Data. In this study, the dataset is a set of MRI brain tumors of 123 patients with brain tumors at Bach Mai Hospital, Hanoi, Vietnam, of all ages. Initially, the MRI image was in DICOM format; to remove the information in the patient's DICOM image and convert the image format for machine learning, the DICOM format was converted to the JPEG image format. The size of the converted images is 256 × 256 pixels.

The image used during training is a T2 pulse sequence image as in Figure 3. Signal intensity with T2 phase correlates very well with not only homogeneity but also tissue profile. Specifically, with low-intensity signals, the tumor has a fibrous and stiffer character than the normal parenchyma. For example, the tumor is fibroblastic in nature, while the more intense sections show a softer characteristic such as a vascular tumor. Therefore, the image of the T2 pulse sequence is considered a pulse sequence that best assesses whether the patient has a brain tumor or not.

Table 3: Confusion matrix.

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

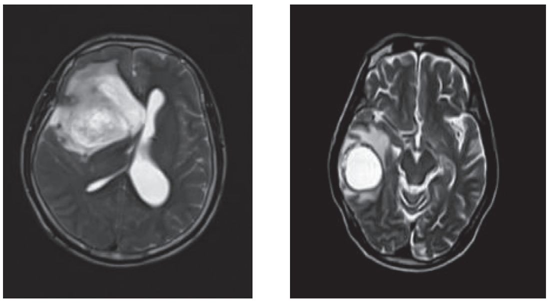


Figure 3: Image of the T2 pulse sequence showing the patient's brain tumor.

With the above 123 patients with brain tumor pathology and 100 healthy persons, 1307 images of T2 pulse sequence were selected, of which 647 images showed brain tumors and 660 images did not show brain tumors. The images are all brought to a size of 256 × 256 pixels to serve the training and testing process of the algorithms.

4.1.2. Normalizing Data

(1) Minimizing Image Redundancy. In the raw MRI image data, it is easy to see that there is a rather large black border, but that is the air in the optical field of the machine, so it does not carry information about the skull to be examined. Therefore, it is really necessary to remove the black out-of-the-edge image from the MRI image without affecting the image information content.

The skull on an MRI is usually surrounded by a bright white border, the outer layer of fat around the skull. Meanwhile, the MRI image is a grayscale image (one-dimensional); the range of values of each element in the image matrix representing the brightness of the pixel is in the range [0, 255]. In order to maximize the black border on the image, the easy method implemented by this study is to find the first pixel with a nonzero value in the directions from left to right, from right to left, from top to bottom, and from bottom to top as shown in Figure 4. After determining the coordinates of those pixels, remove the outer edges. The normalization of images by cutting out the parts that do not make sense in image classification aims to increase the accuracy of the training process and reduce the training time of the algorithm.

(2) Normalizing Image Size. The normalization of the image size helps to improve the accuracy and efficiency of the algorithm. In this study, the image size is 256 × 256. This is the right image size for AI algorithms and ensures MRI

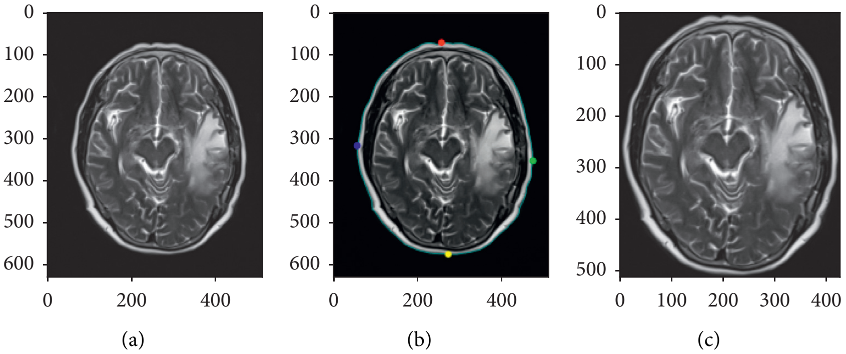


FIGURE 4: Removal of redundant areas of MRI images.

image quality after resizing. Choosing a smaller size will make it difficult for AI algorithms to detect small differences between pixels, affecting the accuracy of the algorithm. If the image size is larger, it will affect the quality of the MRI image after resizing/reducing image quality, negatively affecting the accuracy and performance of the algorithm.

Normalization of data is processed by image data files corresponding to each type of patient's MRI image and by using Python programming. The normalized data removes the nonsignificant parts of the image classification, which increases the accuracy of the model training process and reduces the training time of the algorithm.

4.2. Image Classification Process

Step 1. Preparing the training dataset and feature extraction.

This step is considered an important step in machine learning problems because it is the input for learning to find the model of the problem. We must know how to select the good features, remove the bad features of the data or the noisy components, and estimate how many dimensions of the data are good or in other words how many features to select. If the number of dimensions is too large, making it difficult to calculate, it is necessary to reduce the number of dimensions of the data while maintaining the accuracy of the data (reduce dimension).

In this step, the dataset to test on the model is needed to be prepared. Usually, cross-validation will be used to divide the dataset into two parts, one for training (training dataset) and the other for testing purposes on the model (testing dataset). There are two ways commonly used in cross-validation: splitting and k – folding. For the above algorithms, during the training process, the data is divided according to the ratio 6 : 2 : 2, in which 60% of the data is for training and 20% is for the training validation process (validation). And the remaining 20% is for the process of retesting the model after training.

With the dataset consisting of 1307 images (T2-Images) as mentioned above, the image set has been divided according to the ratio 6 : 2 : 2 to serve the training, validation, and testing processes. Specifically, the number of images used includes 813 images for training, of which 414 images do not show brain tumors and 399 images show brain tumors; 239 images for validation, including 121 images showing brain tumors and 118 images not showing brain tumors; 255 images for the test process, including 130 images showing brain tumors and 125 images not showing brain tumors.

Step 2. Classifier model.

The purpose of the training model is to find a function $f(x)$ from which to label the data. This step is often called learning or training.

$$f(x) = y, \tag{9}$$

where x is the feature or input of the data and y is the class label or output.

The classification model used here is the above supervised learning algorithms DenseNet201, ResNet152V2, MobileNetV3, and VGG19.

Step 3. Checking data with model to make prediction.

After finding the classification model in Step 2, in this step, new data will be added to test on the classification model.

Step 4. Evaluating the classification model and selecting the best model.

In the final step, the model will be evaluated by assessing the error level of the testing data and the training data through the found model. If the system results are not as expected, the parameters (turning parameters) of the learning algorithms must be changed to find a better model as well as to test and reevaluate the classification model. From there, it is possible to choose the best classification model for the problem. All steps mentioned above can be described as in Figure 5.

4.3. Evaluation of Experiment Results

4.3.1. Evaluating the Effectiveness of Applying Data Normalization

(1) Results of Training Process. In order to appraise the effectiveness of data normalization, this work evaluates the convergence (accuracy) of network architectures in

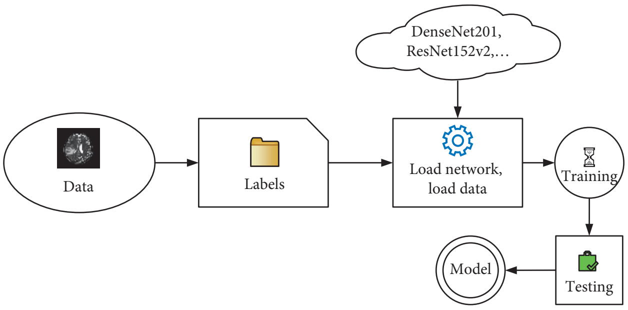
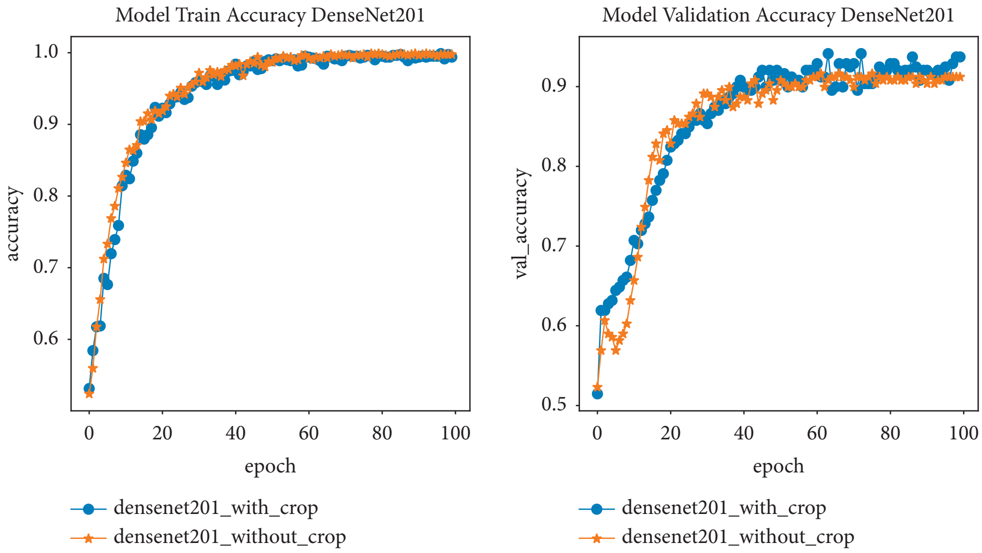
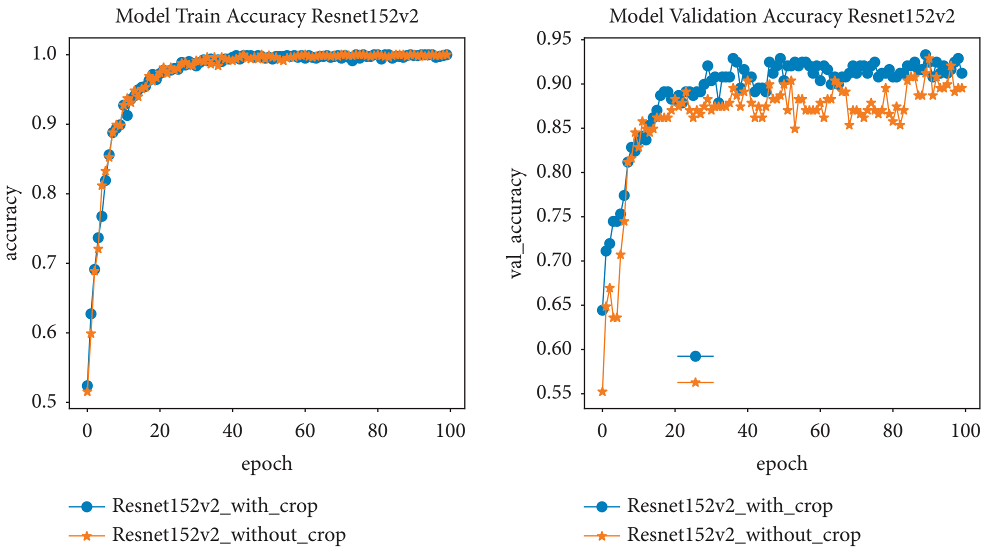


Figure 5: The process of creating a model of the algorithm.


Model Train Accuracy DenseNet201

Model Validation Accuracy DenseNet201

Figure 6: Accuracy of DenseNet201 network architecture before and after normalization.


Model Train Accuracy Resnet152v2

Model Validation Accuracy Resnet152v2

Figure 7: Accuracy of ResNet152V2 network architecture before and after normalization.

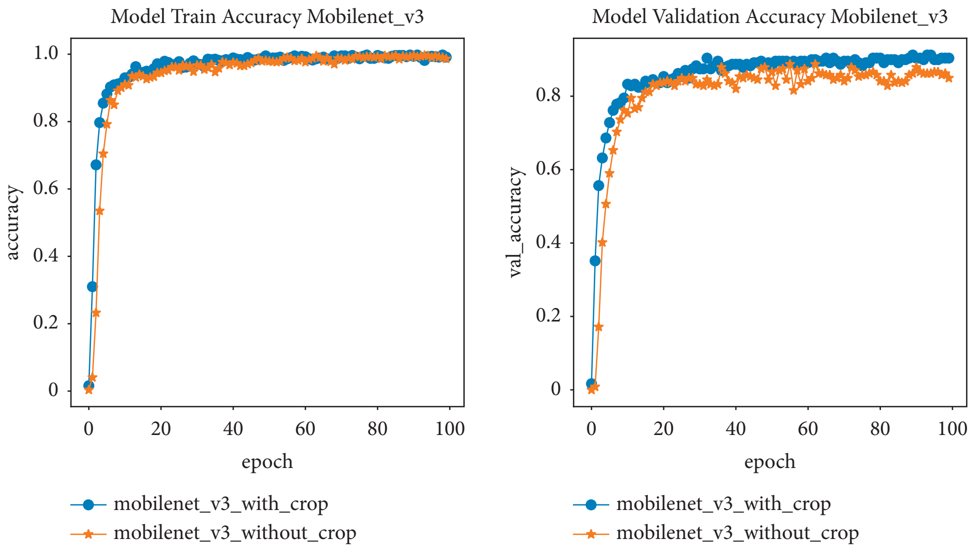


FIGURE 8: Accuracy of MobileNetV3 network architecture before and after normalization.

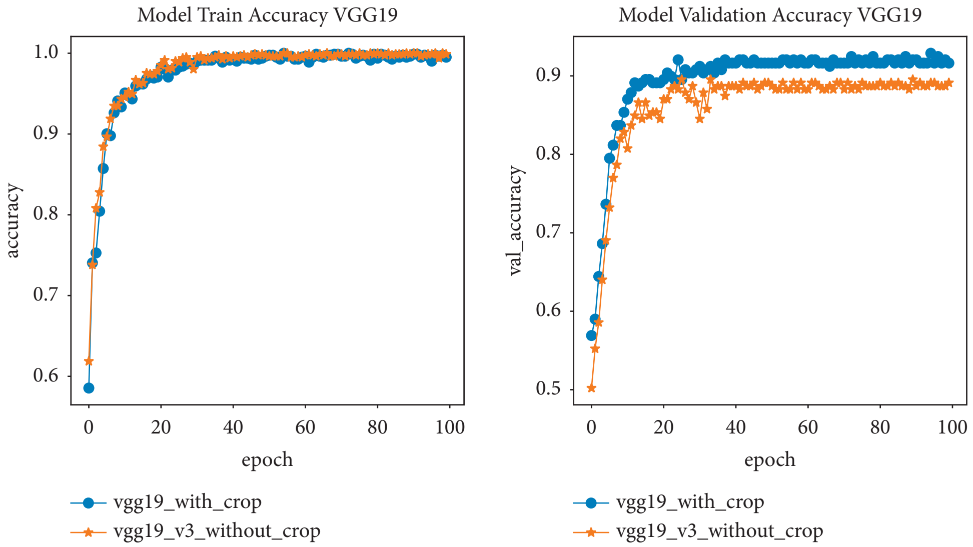


FIGURE 9: Accuracy of VGG19 network architecture before and after normalization.

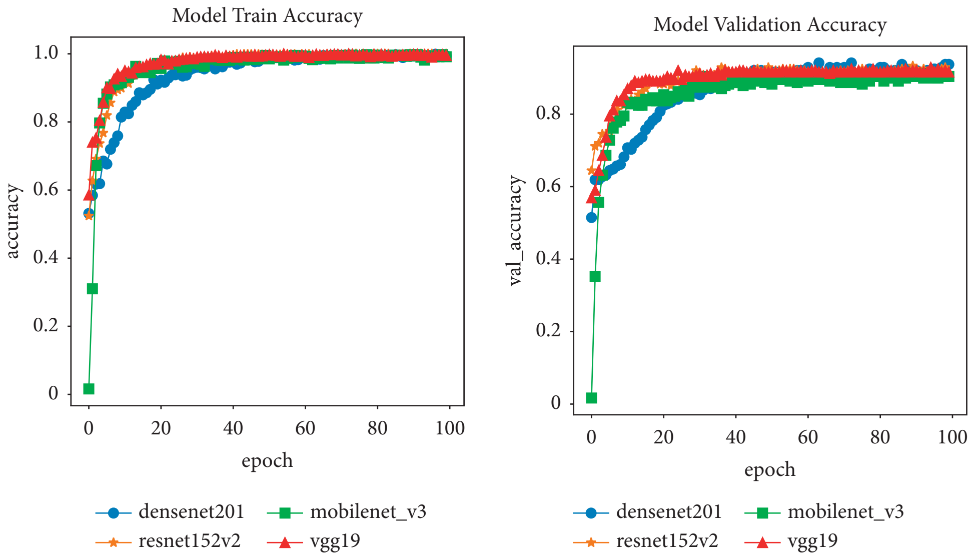


FIGURE 10: Accuracy of network architectures after data normalization.

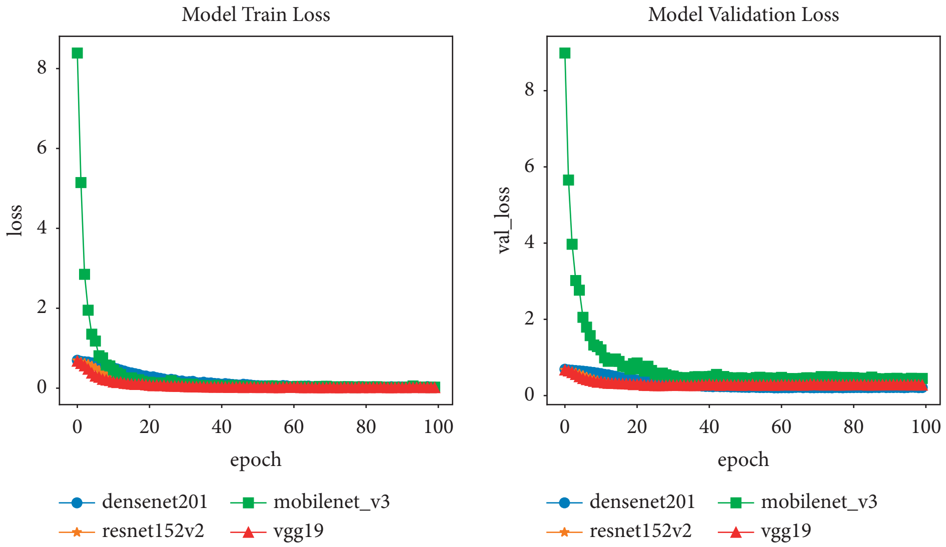


FIGURE 11: Loss of network architectures after data normalization.

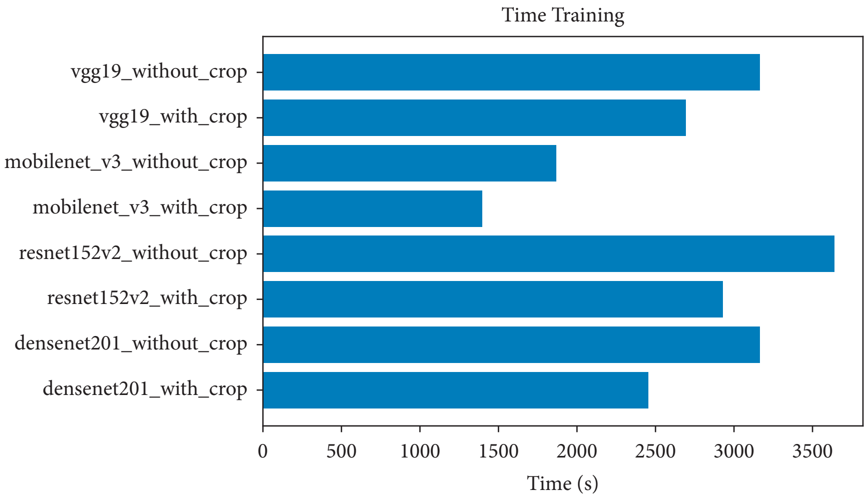


FIGURE 12: Training time over 100 epochs of network architectures before and after normalization.

TABLE 4: Table of comparison results between algorithms before and after data normalization.

CNN architectures	Training accuracy		Validation accuracy		Time (s)	
	Before data normalization	After data normalization	Before data normalization	After data normalization	Before data normalization	After data normalization
DenseNet201	99.87%	99.88%	91.63%	94.14%	3165	455
ResNet152V2	99.95%	99.95%	92.86%	93.31%	3639	927
MobileNetV3	99.66%	99.70%	88.70%	91.21%	1869	1398
VGG19	99.91%	99.91%	89.54%	92.88%	3165	2691

classification through 100 epochs. The network architectures used here are DenseNet201, ResNet152V2, VGG19, and MobileNetV3. The optimal algorithm used in the training processes in this section is the ADAM optimization algorithm.

From Figures 6–9, it is easy to see that the training accuracies before and after normalization of the network architectures are almost the same. The specific results of the DenseNet201, ResNet152V2, MobileNetV3, and VGG19 network architectures are 99.88%, 99.95%, 99.7%, and

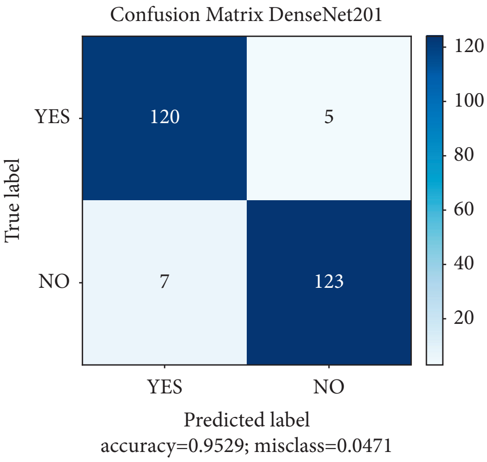


FIGURE 13: DenseNet201 network architecture accuracy based on F1-score.

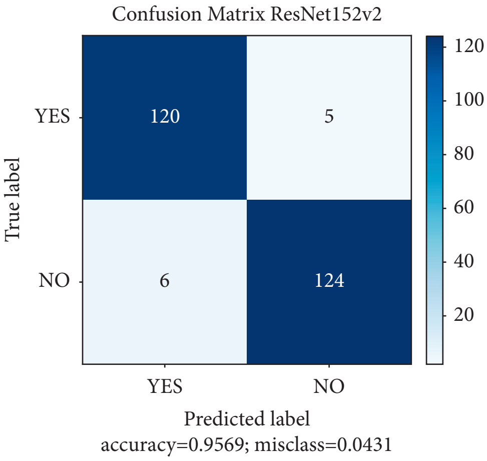


FIGURE 14: ResNet152V2 network architecture accuracy based on F1-score.

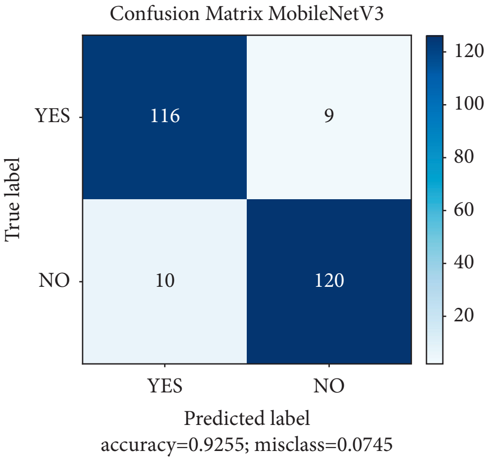


FIGURE 15: MobileNetV3 network architecture accuracy based on F1-score.

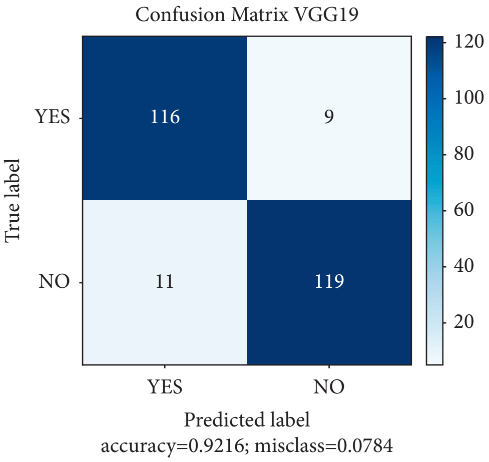


FIGURE 16: VGG19 network architecture accuracy based on F1-score.

TABLE 5: Accuracy of network architectures based on F1-score.

CNN architectures	DenseNet201	ResNet152V2
F1-score	95.29%	95.69%
CNN architectures	MobileNetV3	VGG19
F1-score	92.55%	92.16%

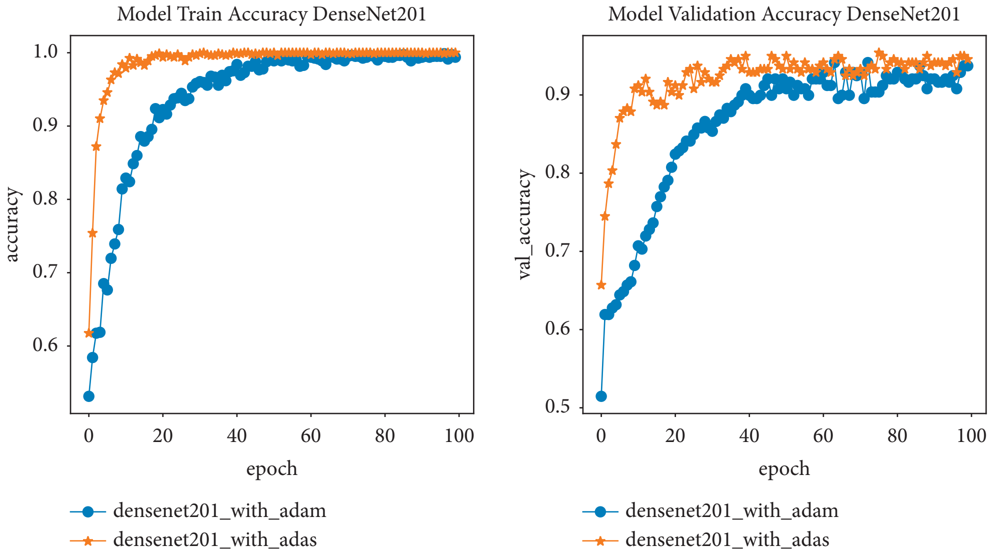


FIGURE 17: Accuracy of DenseNet201 network architectures using ADAM and ADAS optimization algorithms.

99.91%, respectively. However, the results of the validation step of the algorithms showed a marked increase in the accuracy when comparing before and after data normalization. Specifically, the accuracy of the validation process for the DenseNet201 network architecture after normalization is 94.14%, higher than before normalization with an accuracy of 91.63%. The validation result of ResNet152V2 network architecture after normalization has an accuracy of 93.31%, slightly better than before normalization with an accuracy of 92.86%. And this result of ResNet152V2 network architecture after normalization has higher stability than before normalization as presented in Figure 7; Figure 8 indicates that the validation process of MobileNetV3 network architecture after normalization has higher accuracy than before normalization with accuracy of 91.21% and 88.70%, respectively. The validation results of the VGG19 network architecture are similar to those of the three algorithms above with an accuracy of 92.88% after normalization compared to 89.54% before normalization as shown in Figure 9. And it can be seen that all network architectures have convergence with 90% accuracy after only 40 epochs when they use normalized data.

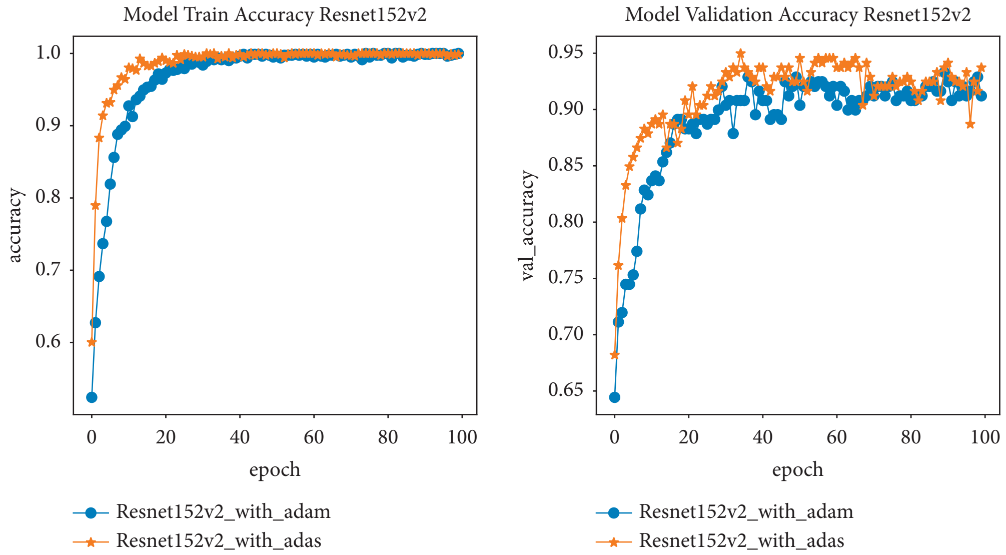


Figure 18: Accuracy of ResNet152V2 network architectures using ADAM and ADAS optimization algorithms.

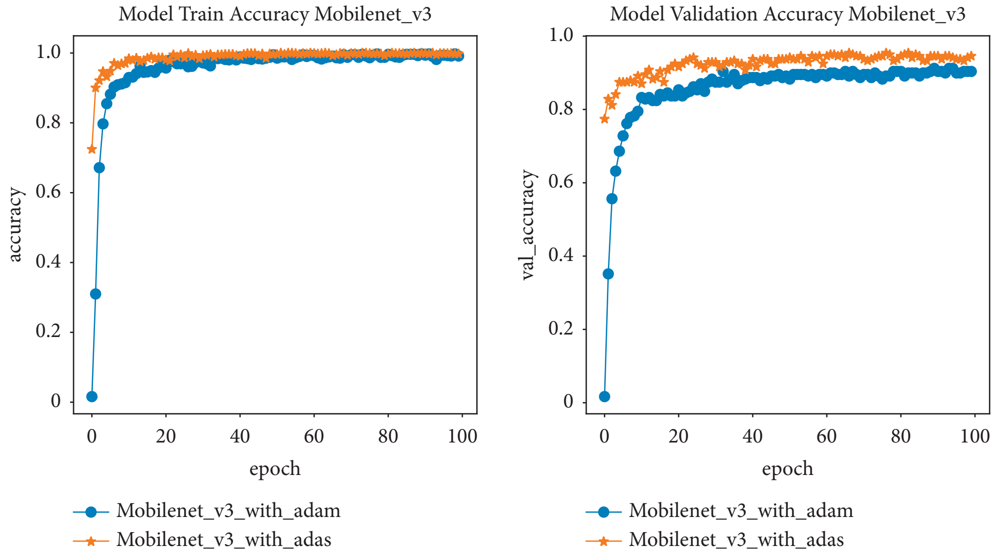


Figure 19: Accuracy of MobileNetV3 network architectures using ADAM and ADAS optimization algorithms.

In this paper, in order to be consistent with the collected brain MRI image data, with the ADAM optimal algorithm and the training process with steadily increasing accuracy, where the loss (loss) decreases the most, this study used different learning rates for each network architecture. Specifically, with the network architectures DenseNet201, ResNet152V2, MobileNetV3, and VGG, the initial learning coefficients are $\eta_0 = \{3e-6; 3e-6; 2e-5; 3e-6\}$, respectively. And the results of using these learning coefficients have shown the stability of the training process to avoid overfitting and are shown in Figures 10 and 11.

In practice, it is not always the case that the longer the model training process, the lower the loss function. When it reaches a certain number of epochs, the loss function value will reach saturation; it can no longer decrease and may even increase again. That is overfitting phenomenon. To prevent this phenomenon and free up computational resources, the training process should be stopped right at that saturation point. In this study, as shown in Figure 11, it can be seen that the values of the loss function for all architectures reach saturation when the number of epochs is 100.

When comparing the efficiency of processing speed on the same resource, based on Figure 12, it can be seen that all 4 network architectures give a shorter training time with normalized image data than the training time with denormalized image data. Comparison results between algorithms with datasets before and after normalization are shown in Table 4. Clearly, the results showed that the benefits of

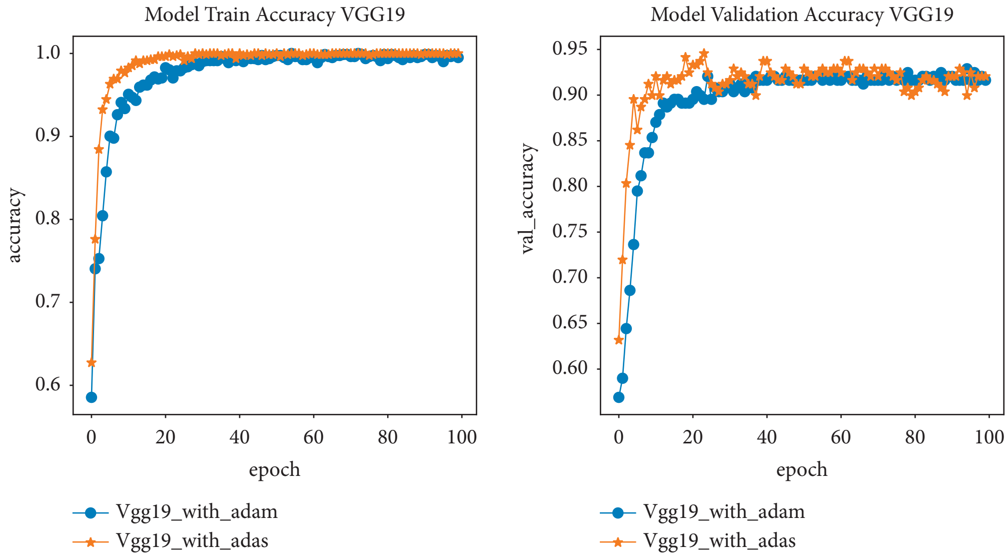


Figure 20: Accuracy of VGG19 network architectures using ADAM and ADAS optimization algorithms.

normalizing the image data make the network architectures capable of classifying brain tumors with higher accuracy and shorter training time.

(2) Evaluating the Accuracy of Network Architectures Based on F1-Score. After performing the training process, models of the respective network architectures were generated. In this part, they will be tested for testing on the test dataset. This dataset includes 255 images of which 130 images are showing brain tumors and 125 images are not showing brain tumors.

The results illustrated in Figures 13–16 and the summary data in Table 5 show that all algorithms have an accuracy greater than 92% when based on the F1-score, in which ResNet152V2 network architecture has the highest results. This is expected to be implemented in practice.

4.3.2. Comparing the Accuracy of Models Using ADAM and ADAS Optimal Function

(1) Results of Training Process. In this section, the accuracy of the classification network architectures will be evaluated and compared using the ADAM and ADAS optimization functions. The network architectures will execute the training, validation, and testing processes on the same normalized database with the same computational resources.

Similar to the ADAM optimization algorithm, in order to fit the brain MRI image data, it is suitable for the ADAS optimization algorithm and the training process has a steady increase in accuracy and the most uniform decrease in loss. Each network architecture uses its own learning coefficient. In this study, the learning coefficients of DenseNet201, ResNet152V2, MobileNetV3, and VGG network architectures are $\eta_0 = \{7e - 3; 5e - 3; 4e - 3; 1e - 2\}$, respectively.

Table 6: Comparison results between network architectures using ADAM and ADAS optimization algorithms.

CNN architectures	Training accuracy		Validation accuracy		Times (s)	
	ADAM	ADAS	ADAM	ADAS	ADAM	DAS
DenseNet201	99.87%	99.88%	94.14%	95.39%	2455 s	2340 s
ResNet152V2	99.95%	99.97%	93.31%	94.47%	2927 s	3161 s
MobileNetV3	99.66%	99.71%	91.21%	95.39%	1398 s	1234 s
VGG19	99.91%	99.9%	92.88%	94.56%	2691 s	2447 s

With the above input data, the experimental results of network architectures with ADAM and ADAS optimal functions are shown in Figures 17–20, respectively. These results show that the training accuracy of the network architectures using the ADAM and ADAS optimization algorithms are almost the same with the obtained values being greater than 99%. However, for the results of the validation process of the network architectures, the accuracy when implementing the ADAS optimization algorithm has improved significantly in comparison with when using the ADAM optimal algorithm. Specifically, the accuracy of the validation process for the DenseNet201 network architecture using the ADAS optimization algorithm is 95.39% compared to 94.14% when using the ADAM optimization algorithm. And to achieve accuracy, with the ADAM optimal algorithm, the DenseNet201 network needs 40 epochs while with the ADAS optimization algorithm it only needs 10 epochs. The training validation process of ResNet152V2 network architecture using ADAS and ADAM optimization algorithms has the accuracy of 94.47% and 93.31%, respectively. To achieve 90% accuracy, ResNet152V2 network needs 30 epochs when using ADAM optimal algorithm while with ADAS optimal algorithm it only needs 20 epochs. For the MobileNetV3 network architecture, the accuracy of the validation process when using the ADAS optimization function is 95.39% compared to 91.21% when using the ADAM optimization algorithm. The convergence speed for using the ADAS optimization function is also much higher than using the ADAM function, specifically to achieve 90%

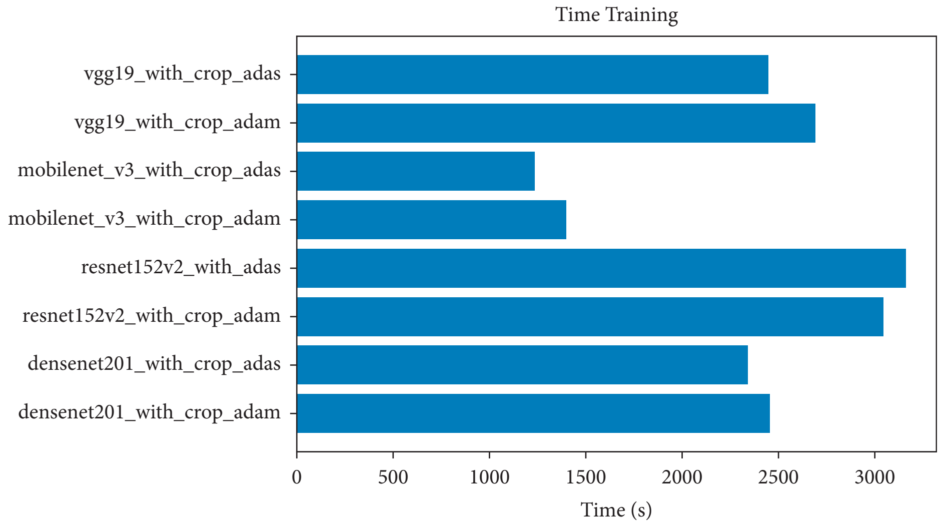


Figure 21: Training time over 100 epochs of network architectures with ADAM and ADAS optimization functions.

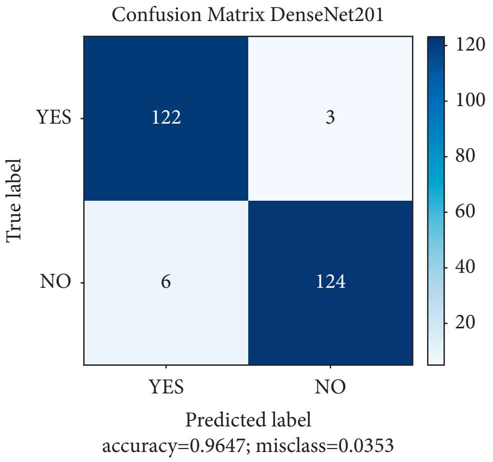


Figure 22: Accuracy evaluation matrix of DenseNet201 network architecture using ADAS optimization algorithm via F1-score.

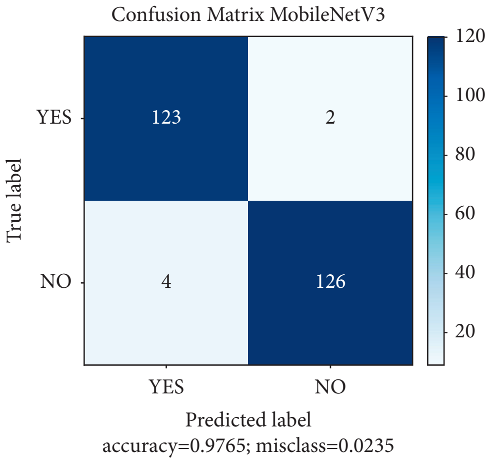


Figure 24: Accuracy evaluation matrix of MobileNetV3 network architecture using ADAS optimization algorithm via F1-score.

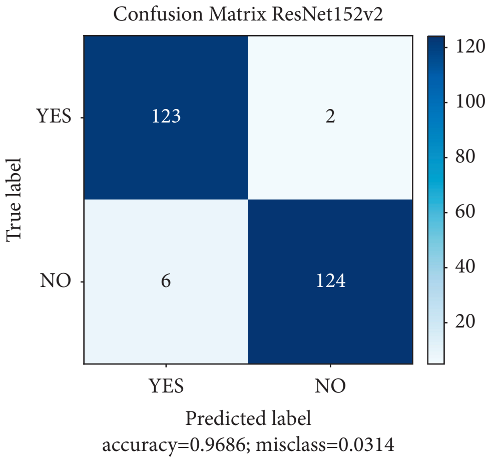


Figure 23: Accuracy evaluation matrix of ResNet152V2 network architecture using ADAS optimization algorithm via F1-score.

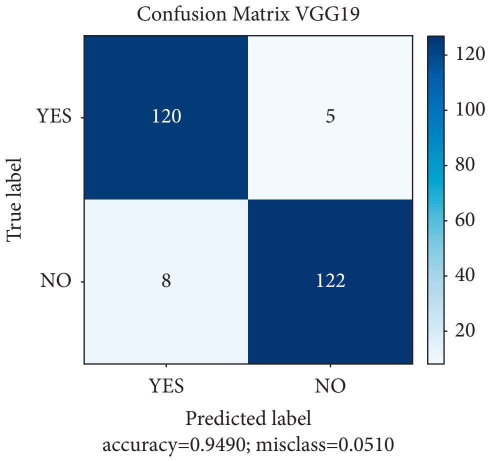


Figure 25: Accuracy evaluation matrix of VGG19 network architecture using ADAS optimization algorithm via F1-score.

Table 7: Accuracy of network architectures based on F1-score using ADAS optimization function.

CNN architectures	DenseNet201	ResNet152V2
F1-score	96.47%	96.86%
CNN architectures	MobileNetV3	VGG19
F1-score	97.65%	94.90%

Table 8: Accuracy and F1-score comparison with other previous studies.

Paper	Algorithms	Accuracy	F1-score
Díaz-Pernas et al. 2021 [18]	Multipathway CNN	99.4%	97.3%
Siddiaue et al. 2021 [16]	Proposed DCNN model	96%	97%
Abd El Kader et al. 2021 [19]	Proposed differential deep-CNN	99.25%	95.23%
Tazin et al. 2021 [20]	CNN architectures	Up to 92%	Up to 92%
This paper	**CNN architectures**	**Up to 99.97%**	**Up to 97.5%**

accuracy, with the ADAM MobileNetV3 function requiring 40 epochs while only 10 epochs are required when using the ADAS function. The VGG19 architecture also has the same results as the above architectures with the accuracy of 94.56% and 92.88%, respectively, with the ADAS and ADAM functions. And also with 90% accuracy, the number of VGG19 network architecture epochs needs to be 25 and 11 when using the ADAM and ADAS functions, respectively.

The performance comparison between ADAS and ADAM algorithms is summarized in Table 6. According to this table as well as the above analysis, it is easy to see that the ADAS optimization algorithm has increased the accuracy of the training process; the convergence in the training process also occurs faster. Figure 21 shows the comparison of training time when using 2 optimization functions with the same normalized dataset. Obviously, the model training time when using the ADAS function in most network architectures is faster. Only for ResNet152V2 architecture, the training time with the use of the ADAS function is slightly longer than with the use of the ADAM function. This can also be one of the problems that need to be studied in the future.

(2) Evaluation of F1-Score of Network Architectures Using ADAS Optimization Algorithm. Performing evaluation through F1-score similar to ADAM's algorithm, according to Figures 22–25, the accuracy evaluation through F1-score of network architectures using ADAS optimization function is established as shown in Table 7.

Obviously, when comparing the synthetic results presented in Tables 5 and 7, it is easy to see that the ADAS optimization algorithm has significantly increased the accuracy of the aforementioned models, in which the MobileNetV3 network model gives the highest accuracy of 97.65%. Combined with the results analyzed above, for the problem of brain tumor identification on MRI-T2 images, the ADAS optimization algorithm has significantly improved the accuracy of the training, validation, and testing processes of all the models surveyed in this work as well as shortening the training time of those models compared to the ADAM algorithm.

4.3.3. Comparison of Results. The performance of the proposed system in our study will be compared with the most recently published studies mentioned above. The results of that comparison are shown in Table 8. Based on this table, it is easy to see that the proposed system gave better results in both accuracy and F1-score than other studies with the same subjects. Obviously, although using the same variants of the DCNNs family, the data normalization and the ADAS optimization function helped to significantly improve the performance of the proposed system compared to those other systems.

5. Conclusion

This article has focused on deploying the application of artificial intelligence algorithms in classifying brain tumor patients and normal people using human brain MRI images. The dataset used is MRI images of Vietnamese people, including 123 patients and 100 healthy people. The four algorithms that are experimentally compared in the study are DenseNet201, ResNet152V2, MobileNetV3, and VGG19. The experimental results in the study have shown that the normalization of the initial data processing is very important when it has significantly increased the accuracy in classifying and detecting patients as well as reducing the training time of those models. On the other hand, the paper has also shown the efficiency of the ADAS optimization function compared with the very popular ADAM optimization function. In particular, the ADAS algorithm has advantages in comparison with the ADAM function in improving accuracy as well as reducing model training time. Of the four algorithms mentioned above, the MobileNetV3 algorithm is the most efficient. This can be considered as the foundation for implementing the above system in practice. However, the system also has the disadvantage that the dataset is still small. In the future, besides collecting more data to increase the accuracy of the system, the research will also develop methods to specifically classify those tumor types according to their tumor characteristics (benign or malignant) or by type of disease.

References

[1] J. C. Buckner, P. D. Brown, B. P. O'Neill, F. B. Meyer, C. J. Wetmore, and J. H. Uhm, "Central nervous system tumors," *Mayo Clinic Proceedings*, vol. 82, no. 10, 2007.

[2] A. Lashkari, "A neural network based method for brain abnormality detection in MR images using Gabor wavelets," *International journal of computer Applications*, vol. 4, no. 7, pp. 9–15, 2010.

[3] M. Vargo, "Brain tumor rehabilitation," *American Journal of Physical Medicine & Rehabilitation*, vol. 90, no. 5, pp. S50–S62, 2011.

[4] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, November 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, December 2016.

[6] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[8] S. Khawaldeh, U. Pervaiz, A. Rafiq, and R. S. Alkhawaldeh, "Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks," *Applied Sciences*, vol. 8, no. 1, p. 27, 2018.

[9] J. S. Paul, A. J. Plassard, B. A. Landman, and F. Daniel, "Deep learning for brain tumor classification," in *Proceedings of the Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10137, International Society for Optics and Photonics, Bellingham, WA, USA, March 2017.

[10] N. Varuna Shree and T. N. R. Kumar, "Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network," *Brain informatics*, vol. 5, no. 1, pp. 23–30, 2018.

[11] D. J. Hemanth, J. Anitha, A. Naaji, O. Geman, D. E. Popescu, and L. Hoang Son, "A modified deep convolutional neural network for abnormal brain image classification," *IEEE Access*, vol. 7, pp. 4275–4283, 2019.

[12] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Computers in Biology and Medicine*, vol. 111, Article ID 103345, 2019.

[13] S. Das, O. F. M. R. R. Aranya, and N. N. Labiba, "Brain tumor classification using convolutional neural network," in *Proceedings of the International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), IEEE, Dhaka, Bangladesh,*, December 2019.

[14] Z. Ullah, M. U. Farooq, S.-H. Lee, and D. An, "A hybrid image enhancement based brain MRI images classification technique," *Medical Hypotheses*, vol. 143, Article ID 109922, 2020.

[15] A. Çinar and M. Yildirim, "Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture," *Medical Hypotheses*, vol. 139, Article ID 109684, 2020.

[16] M. A. B. Siddiaue, S. Sakib, M. M. R. Khan, A. K. Tanzeem, M. Chowdhury, and N. Yasmin, "Deep convolutional neural networks model-based brain tumor detection in brain MRI images," in *Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Palladam, India, November 2020.

[17] P. Saxena, A. Maheshwari, and S. Maheshwari, "Predictive modeling of brain tumor: a Deep learning approach," *Innovations in Computational Intelligence and Computer Vision Advances in Intelligent Systems and Computing*, Springer, Berlin, Germany, pp. 275–285, 2021.

[18] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodrígue, and D. González-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," *Healthcare*, vol. 9, no. 2, 2021.

[19] I. Abd El Kader, G. Xu, Z. Shuai, S. Saminu, I. Javaid, and I. Salim Ahmad, "Differential deep convolutional neural network model for brain tumor classification," *Brain Sciences*, vol. 11, no. 3, p. 352, 2021.

[20] T. Tazin, S. Sarker, P. Gupta et al., "A robust and novel approach for brain tumor classification using convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 2392395, 11 pages, 2021.

[21] J. Seetha and S. S. Raja, "Brain tumor classification using convolutional neural networks," *Biomedical and Pharmacology Journal*, vol. 11, no. 3, pp. 1457–1461, 2018.

[22] N. M. Balasooriya and R. D. Nawarathna, "A sophisticated convolutional neural network model for brain tumor classification," in *Proceedings of the IEEE International Conference on Industrial and Information Systems (ICIIS), IEEE, Peradeniya, Sri Lanka*, February 2017.

[23] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *European Radiology*, vol. 12, no. 4, pp. 920–927, 2002.

[24] A. Haase, "Snapshot flash mri. applications to t1, t2, and chemical-shift imaging," *Magnetic Resonance in Medicine*, vol. 13, no. 1, pp. 77–89, 1990.

[25] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[27] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, October 2015.

[28] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[29] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, PMLR, Long Beach, CA, USA, May 2019.

[30] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 2020.

[31] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, https://arxiv.org/abs/1511.08458.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA*, June 2009.

[33] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, https://arxiv.org/abs/1609.04747.

[34] L. Bottou, "Stochastic gradient descent tricks," *Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, pp. 421–436, 2012.

[35] M. S. Hosseini and K. N. Plataniotis, "Adas: adaptive scheduling of stochastic gradients," 2020, https://arxiv.org/abs/2006.06587.

[36] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, p. 7, 2011.

[37] N. Jegadeesh and S. Titman, "Momentum," *Annual Review of Financial Economics*, vol. 3, no. 1, pp. 493–509, 2011.

[38] D. P. Kingma and Ba. Jimmy, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[39] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proceedings of the European Conference on Information Retrieval*, Santiago de Compostela, Spain, March 2005.

Optimal Wireless Information and Power Transfer Using Deep Q-Network

Debasish Mishra, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, debasishmishra1@gmail.com*

Pinaki Prasanna, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, pinakiprasanna91@gmail.com*

Laxminarayan Mishra, *Department of Electrical and Electronics Engineering, Capital Engineering College, Bhubaneswar, laxminarayan.s@gmail.com*

Prajnadipta Sahoo, *Department of Electrical and Electronics Engineering, NM Institute of Engineering & Technology, Bhubaneswar, prajnadipta567@gmail.com*

Abstract

In this paper, a multiantenna wireless transmitter communicates with an information receiver while radiating RF energy to surrounding energy harvesters. The channel between the transceivers is known to the transmitter, but the channels between the transmitter and the energy harvesters are unknown to the transmitter. By designing its transmit covariance matrix, the transmitter fully charges the energy buffers of all energy harvesters in the shortest amount of time while maintaining the target information rate toward the receiver. At the beginning of each time slot, the transmitter determines the particular beam pattern to transmit with. Throughout the whole charging process, the transmitter does not estimate the energy harvesting channel vectors. Due to the high complexity of the system, we propose a novel deep Q-network algorithm to determine the optimal transmission strategy for complex systems. Simulation results show that deep Q-network is superior to the existing algorithms in terms of the time consumption to fulfill the wireless charging process.

1. Introduction

For a wireless transceiver pair with multiple antennas, optimizing the transmit covariance matrix can achieve high data-rate communication over the multiple-input multiple-output (MIMO) channel. Meanwhile, the radiated radio frequency (RF) energy can be acquired by the nearby RF energy harvesters to charge the electronic devices [1].

The problem of simultaneous wireless information and power transfer (SWIPT) has been widely discussed in recent years. SWIPT systems are divided into two categories: (1) the receiver splits the received signals for information decoding and energy harvesting [2, 3]; (2) separated and dedicated information decoders (ID) and RF energy harvesters (EH) exist in the systems [4]. For the second type of the system, different transmission strategies have ever been proposed to achieve good performance points in the rate-energy region [1, 2, 5]. For the multiple RF energy harvesters, which are in the vicinity of the wireless transmitter, the covariance matrix at the transmitter is designed to either maximize the net energy harvesting rate or fairly distribute the radiated RF energy at the harvesters [6, 7]. The achievable information rate of the wireless transmitter-receiver pair is beyond a minimum requirement for reliable communication. Most of the existing works assume the channel state information (CSI) is completely known. Given the complete CSI, the transmitter designs the transmit covariance matrix to achieve the maximum information rate while satisfying the RF energy harvesting requirement [4, 8].

However, in practice, it is difficult for the transmitter to obtain the channel state information to the nearby RF energy harvesters because the scattering distribution of the hardware-limited energy harvesters makes the channel estimation at the RF energy harvesters challenging [9, 10]. The analytic center cutting plane method (ACCPM) was proposed for the transmitter to approximate the channel information with a few bits of feedback from the RF energy receiver iteratively [10]. Since this method is implemented by solving a convex optimization problem, the algorithm leads to high computational complexity. To reduce complexity, channel estimation based on Kalman filtering was proposed [11]. Nevertheless, the disadvantage of this approach is the slow convergence rate. In order to effectively deal with the CSI acquisition problem, in our paper, we will use the deep learning algorithm to solve the optimization problem in the SWIPT system only with partial channel information. The partial CSI is easy to acquire, which is already enough to achieve superior system performance using the deep Q-network. To the best of our knowledge, we are the first one to use the deep Q-network to optimize the SWIPT system performance and validate its superiority.

In our model, the transmitter intends to fully charge all surrounded energy harvesters' energy buffers in the shortest time while maintaining a target information rate toward the receiver. The communication link is defined as a strong line-of-sight (LOS) transmission, which is supposed to be invariant, but the energy harvesting channel conditions vary over time. Due to current hardware limitations, we assume that the estimation of the energy harvesting channel vectors is not able to be implemented under the fast varying channel conditions. As a result, the wireless charging problem can be modeled as a high complexity discrete-time stochastic control process with unknown system dynamics [12]. In [13], a similar problem has been explored. A multiarmed bandit algorithm is used to determine the optimal transmission strategy. In our paper, we apply a deep Q-network to solve the optimization problem and the simulation results demonstrate that the deep Q-network algorithm outperforms the multiarmed bandit algorithm. Historically, deep Q-network has a strongly proven record of attaining mastery over complex games with a very large number of system states, and unknown state transition probabilities [12]. More recently, a deep Q-network has been applied to deal with complex communication problems and has been shown to achieve good performance [14–16]. For this reason, we found deep Q-network fitting for our model. In our model, we consider the accumulated energy of the energy harvesters as the system states, while we define the action as the transmit power allocation. At the beginning of each time slot, each energy harvester sends feedback about the accumulated energy level to the wireless transmitter, and the transmitter collects all the information in order to generate the system state and inputs it into a well-trained deep Q-network. The deep Q-network outputs the Q values corresponding to all possible actions. The action with the maximum Q value is selected as the beam pattern to be used for the transmission during the current time slot.

Based on the traditional deep Q-network, the double deep Q-network and dueling deep Q-network algorithms are applied in order to reduce the observed overestimations [17] and improve the learning efficiency [18]. Henceforth, we apply dueling double deep Q-network to solve the varying channel multiple energy harvester wireless charging problem.

The novelties of this paper are summarized as follows:

(i) The simultaneous wireless information and power transfer problem is formulated as a Markov decision process (MDP) in an unknown varying channel condition for the first time.

(ii) The deep Q-network algorithm is applied to solve the proposed optimization problem for the first time. We demonstrate that, compared to the other existing algorithms, deep Q-network shows the superiority in efficient and stable wireless power transfer.

(iii) Multiple experimental scenarios are explored. By varying the number of transmission antennas and the number of energy harvesters in the system, the performance of both the deep Q-network and the other algorithms is compared and analyzed.

(iv) The evaluation for the algorithms is based on the real experimental data, which validate the effectiveness of the proposed deep Q-network in real-time simultaneous wireless information and power transfer systems.

The rest of the paper is organized as follows. In Section 2, we describe the simultaneous wireless information and power transfer system model. In Section 3, we model the optimization problem as a Markov decision process and present a deep Q-network algorithm to determine the optimal transmission strategy. In Section 4, we present our simulation results for different experimental environments. Section 5 concludes the paper.

2. System Model

As shown in Figure 1, an information transmitter communicates with its receiver while perceived by K nearby RF energy harvesters [8]. Both the transmitter and the receiver are equipped with M antennas, while each RF energy harvester is equipped with one receive antenna. The baseband received signal at the receiver can be represented as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{M\times M}$ denotes the normalized baseband equivalent channel from the information transmitter to its receiver, $\mathbf{x} \in \mathbb{C}^{M\times 1}$ represents the transmitted signal, and $\mathbf{z} \in \mathbb{C}^{M\times 1}$ is the zero-mean circularly symmetric complex Gaussian noise with $\mathbf{z} \sim \mathscr{CN}(\mathbf{0}, \rho^2\mathbf{I})$.

The transmit covariance matrix is denoted by $\mathbf{Q}$, i.e., $\mathbf{Q} = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$. The covariance matrix is Hermitian positive semidefinite, i.e., $\mathbf{Q} \succcurlyeq 0$. The transmit power is restricted by the transmitter's power constraint P, i.e., $\mathrm{Tr}(\mathbf{Q}) \le P$. For the information transmission, we assume that a Gaussian codebook with infinitely many code words is used for the symbols and the expectation of the transmit covariance matrix is taken over the entire codebook. Therefore, $\mathbf{x}$ is the zero-mean circularly symmetric complex Gaussian with $\mathbf{x} \sim \mathscr{CN}(\mathbf{0}, \mathbf{Q})$. With transmitter precoding and receiver filtering, the capacity of the MIMO channel is the sum of the capacities of the parallel noninterfering single-input single-output (SISO) channels (eigenmodes of channel $\mathbf{H}$) [19]. We

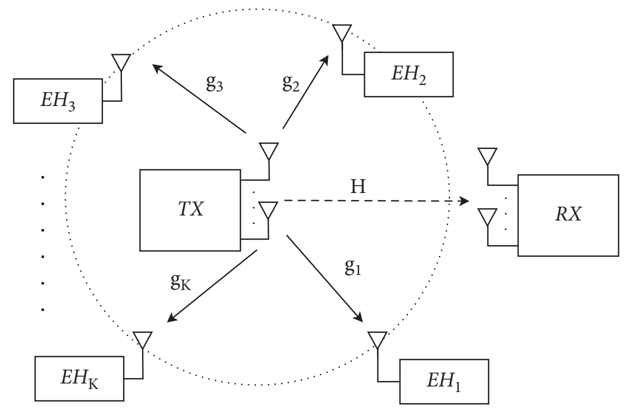


Figure 1: Wireless information transmitter and receiver surrounded by multiple RF energy harvesters.

convert the MIMO channel to M eigenchannels for information and energy transfer [20, 21]. A singular value decomposition (SVD) on $\mathbf{H}$ gives $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^H$, where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_M)$ contains the M singular values of $\mathbf{H}$. Since the MIMO channel is decomposed into M parallel SISO channels, the information rate can be given by

$$r = \sum_{m=1}^{M} \log\left(1 + \rho^{-2}|\sigma_m|^2 \widehat{q}_m\right), \tag{2}$$

where $\{\widehat{q}_m\}$ are the diagonal elements of $\widehat{\mathbf{Q}}$ with $\widehat{\mathbf{Q}} = \mathbf{V}^H\mathbf{Q}\mathbf{V}$.

The RF energy harvester received power specifies the harvested energy normalized by the baseband symbol period and scaled by the energy conversion efficiency. The received power at the ith energy harvester is

$$p_i = \mathbf{g}_i^H \mathbf{Q} \mathbf{g}_i, \tag{3}$$

where $\mathbf{g}_i \in \mathbb{C}^{M\times 1}$ is the channel vector from the transmitter to the ith energy harvester. With MIMO channel decomposition, the received power at energy harvester i is denoted as

$$p_i = \sum_{m=1}^{M} |\widehat{g}_{im}|^2 \widehat{q}_m, \tag{4}$$

where $\{\widehat{g}_{im}\}$ are the elements of vector $\widehat{\mathbf{g}}_i$ with $\widehat{\mathbf{g}}_i = \mathbf{V}^H\mathbf{g}_i$.

We define the simplified channel vector from the transmitter to the ith RF energy harvester as

$$\mathbf{c}_i = \left[|\widehat{g}_{i1}|^2, |\widehat{g}_{i2}|^2, \ldots, |\widehat{g}_{iM}|^2\right]^T, \tag{5}$$

for each $i \in \mathscr{K} = \{1, 2, \ldots, K\}$. The simplified channel vector contains no phase information. The K simplified channel vectors compose matrix $\mathbf{C} \in \mathbb{R}^{M\times K}$ as

$$\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K]. \tag{6}$$

In what follows, we assume that time is slotted, each time slot as a duration T, and that each energy harvester is equipped with an energy buffer of size $B_i \in [0, B_{\max}]$, $i \in \mathscr{K}$. Without loss of generality, we assume that, at $t = 0$, all harvesters' buffers are empty, which corresponds to system state $s_0 = [0, 0, \ldots, 0]$. At a generic time slot t, the transmitter transmits with one of the designed beam patterns. Each harvester i can harvest the specific amount of power p_i, and its energy buffer values increase to $B_i^{t+1} = B_i^t + p_i T$. Therefore, each state of the system includes the accumulated harvested energy information of all K harvesters, i.e.,

$$s_t = \left[B_1^t, B_2^t, \ldots, B_K^t\right], \tag{7}$$

where B_i^t denotes the ith energy harvester's accumulated energy up to time slot t.

Once all harvesters are fully charged, we assume that the system arrives at a final goal state denoted as $s_G = [B_{\max}, B_{\max}, \ldots, B_{\max}]$. We note that the energy buffer level $B_{\max}$ also accounts for situations in which $B_i > B_{\max}$.

3. Problem Formulation for Time-Varying Channel Conditions

In this section, we suppose that the communication link is characterized by strong LOS transmission, which results in an invariant channel matrix $\mathbf{H}$, while the energy harvesting channel vector $\mathbf{g}$ varies over time slots. We model the wireless charging problem as a Markov decision process (MDP) and show how to solve the optimization problem using reinforcement learning (RL). When the number of system states is very large, we apply a deep Q-network algorithm to acquire the optimal strategy at each particular system state.

3.1. Problem Formulation. In order to model our optimization problem as a RL problem, we define the beam pattern chosen in a particular time slot t as the action $\mathbf{a}^t$. The set of possible actions $\mathscr{A}$ is determined by equally generating L different beam patterns with power allocation vector $\widehat{\mathbf{q}} = [\widehat{q}_1, \ldots, \widehat{q}_m]$ that satisfies the power and information rate constraints, i.e., $\sum_{i=1}^{M} \widehat{q}_m = P$, $\sum_{i=1}^{M} \log(1 + \rho^{-2}|\sigma_m|^2 \widehat{q}_m) \geq R$. Each beam pattern corresponds to a particular power level p_i, which depends not only on the action $\mathbf{a}^t$ but also on the channel condition experienced by the harvester during time slot t.

Given the above, the simultaneous wireless information and energy transfer problem for a time-varying channel can be formulated as minimizing the time-consumption n to fully charge all the energy harvesters while maintaining the information rate between the information transceivers:

$$\mathscr{P}_1: \quad \begin{aligned} \underset{\{\mathbf{a}^t\}}{\text{minimize}} \quad & n \\ \text{subject to} \quad & a_m^t \geq 0 \\ & \sum_{m=1}^{M} a_m^t \leq P \\ & \sum_{m=1}^{M} \log\left(1 + \rho^{-2}|\sigma_m|^2 a_m^t\right) \geq R \\ & \sum_{t=1}^{n} \sum_{m=1}^{M} |\widehat{g}_{im}^t|^2 a_m^t T \geq B_{\max}, \quad \forall i \in \mathscr{K} \end{aligned} \quad . \tag{8}$$

In general, the action selected at each time slot will be different to adapt to the current channel conditions and current energy buffer state of the harvesters. Therefore, the evolution of our system can be described by a Markov chain, where the generic state s is identified by the current buffer levels of the harvester, i.e., $s = \{B_1, B_2, \ldots, B_K\}$. The set of all states is denoted by $\mathcal{S}$. Among all states, we are interested in the state in which all harvesters' buffer is empty, namely, $s_0 = \{0, \ldots, 0\}$, and the state s_G in which all the harvesters are fully charged, i.e., $s_G = \{B_{\max}, \ldots, B_{\max}\}$. If we suppose that we know all the channel coefficients at each time slot, problem $\mathcal{P}_1$ can be seen as a stochastic shortest path (SSP) problem from state s_0 to state s_G. At each time slot, the system is in a generic state s, the transmitter selects a beam pattern or action $\mathbf{a} \in \mathcal{A}$, and the system moves to a new state s'. The dynamics of the system is captured by transition probabilities $p_{s,s'}(\mathbf{a})$, $s, s' \in \mathcal{S}$, and $\mathbf{a} \in \mathcal{A}$, describing the probability that the harvesters' energy buffers reach the levels in s' after a transmission with beam pattern $\mathbf{a}$. We note that the goal state s_G is absorbing, i.e., $P_{s_G,s_G}(\mathbf{a}) = 1$, $\forall \mathbf{a} \in \mathcal{A}$.

Each transition also has an associated reward, $w(s, \mathbf{a}, s')$, that denotes the reward when the current state is $s \in \mathcal{S}$, action $\mathbf{a} \in \mathcal{A}$ is selected, and the system moves to state $s' \in \mathcal{S}$. Since we aim at reaching s_G in the fewest transmission time slots, we consider that the action entails a positive reward related to the difference between the current energy buffer level and the full energy buffer level of all harvesters. When the system reaches state s_G, we set the reward as 0. In this way, the system not only tries to fully charge all harvesters in the shortest time but will also uniformly charge all the harvesters. In detail, we define the reward function as

$$w(s, \mathbf{a}, s') = -\lambda \left(\mathrm{KB}_{\max} - \sum_{i=1}^{K} \min(s_i', B_{\max}) \right), \tag{9}$$

where

$$\lambda s_i' = \lambda s_i + \lambda \sum_{m=1}^{M} \left|\widehat{g}_{im}\right|^2 a_m T, \tag{10}$$

and λ denotes the unit price of the harvested energy.

It is noted that different reward functions can also be selected. As an example, it is also possible to set a constant negative reward (e.g., a unitary cost) for each transmission that the system does not reach the goal state and a big positive reward only for the states and actions that bring the system to the goal state s_G. This can be expressed as follows:

$$w(s, \mathbf{a}, s') = \begin{cases} +\infty, & s' = s_G, \\ -1, & \text{otherwise}. \end{cases} \tag{11}$$

We note that the reward formulation of equation (11) is actually equivalent to minimizing the number of time slots required to reach state s_G starting from state s_0.

Using the above formulation, the optimization problem $\mathcal{P} = (\mathcal{S}, \mathcal{A}, p, w, s_0, s_G)$ can then be seen as a stochastic shortest path search from state s_0 to state s_G on the Markov chain with states $\mathcal{S}$ and probabilities $\{p_{s,s'}(\mathbf{a})\}$, actions $\mathbf{a} \in \mathcal{A}$, and rewards $w(s, \mathbf{a}, s')$. Our objective is to find, for each possible state $s \in \mathcal{S}$, an optimal action $\mathbf{a}^*(s)$ so that the system will reach the goal state following the path with maximum average reward. A generic policy can be written as $\pi = \{\mathbf{a}(s)\colon s \in \mathcal{S}\}$.

Different techniques can be applied to solve problem $\mathcal{P}_1$, as it represents a particular class of MDPs. In this paper, however, we assume that the channel conditions at each time slot are unknown, which corresponds to not knowing the transition probabilities $\{p_{s,s'}(\mathbf{a})\}$. Therefore, in the next section, we describe how to solve the above problem using reinforcement learning.

3.2. *Optimal Power Allocation with Reinforcement Learning.*

Reinforcement learning is suitable for solving optimization problems in which the system dynamics follow a particular transition probability function, however, the probabilities $\{p_{s,s'}(\mathbf{a})\}$ are unknown. In what follows, we first show how to apply the Q-learning algorithm [22] to solve the optimization problem and then show how we can combine the reinforcement learning approach with a neural network to approximate the system model in case of large states and action sets, using deep Q-network [12].

3.2.1. *Q-Learning Method.*

If the number of system states is small, we can depend on the traditional Q-learning method to find the optimal strategy at each system state, as defined in the previous section.

To this end, we define the cost function of action $\mathbf{a}$ on system state s as $\{p_{s,s'}(\mathbf{a})\}$, with $s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$. The algorithm initializes with $Q(s, \mathbf{a}) = 0$ and then updates the Q values using the following equation:

$$Q(s, \mathbf{a}) = (1 - \alpha(s, \mathbf{a}))Q(s, \mathbf{a}) + \alpha(s, \mathbf{a})\left[w(s, \mathbf{a}, s') + \gamma f(s', \mathbf{a})\right], \tag{12}$$

where

$$f(s', \mathbf{a}) = \min_{\mathbf{a} \in \mathcal{A}} Q(s', \mathbf{a}), \tag{13}$$

and $\alpha(s', \mathbf{a})$ denotes the learning rate. In each time slot, only one Q value is updated, and hence, all the other Q values remain the same.

At the beginning of the learning iterations, since the Q-table does not have enough information to choose the best action at each system state, the algorithm randomly explores new actions. Hence, we first define threshold $\varepsilon_c \in [0.5, 1]$, and we then randomly generate a probability $p \in [0, 1]$. In the case that $p \geq \varepsilon_c$, we choose the action $\mathbf{a}$ as

$$\mathbf{a} = \max_{\mathbf{a} \in \mathcal{A}} Q(s, \mathbf{a}). \tag{14}$$

On the contrary, if $p < \varepsilon_c$, we randomly select one action from the action set $\mathcal{A}$.

When Q^* converges, the optimal strategy at each state is determined as

$$\pi^* (s) = \arg\max_{\mathbf{a}\in\mathscr{A}} Q^* (s, \mathbf{a}), \tag{15}$$

which corresponds to finding the optimal beam pattern for each system state during the charging process.

3.2.2. Deep Q-Network. When considering a complex system with multiple harvesters, large energy buffers, and time-varying channel conditions, the number of system states dramatically increases. In order to learn the optimal transmit strategy at each system state, the Q-learning algorithm described before requires a Q-table with a large number of elements, making it very difficult for all the values in the Q-table to converge. Therefore, in what follows, we describe how to apply the deep Q-network (DQN) approach to find the optimal transmission policy.

The main idea of DQN is to train a neural network to find the Q function of a particular system state and action combination. When the system is in state s, and action $\mathbf{a}$ is selected, the Q function is denoted as $Q(s, \mathbf{a}, \theta)$. θ denotes the parameters of the Q-network. The purpose of training the neural network is to make

$$Q(s, \mathbf{a}, \theta) \approx Q^* (s, \mathbf{a}). \tag{16}$$

According to the DQN algorithm [17], two neural networks are used to solve the problem: the evaluation network and the target network, which are denoted as eval_net and target_net, respectively. Both the eval_net and the target_net are set up with several hidden layers. The input of the eval_net and the target_net are denoted as s and s', which describe the current system state s and the next system state s', respectively. The output of eval_net and target_net are denoted as $Q_e (s, \mathbf{a}, \theta)$ and $Q_t (s, \mathbf{a}, \theta)$, respectively. The evaluation network is continuously trained to update the value of θ; however, the target network only copies the weight parameters from the evaluation network intermittently (i.e., $\theta' = \theta$). In each neural network learning epoch, the loss function is defined as

$$\text{loss}(\theta) = E\left[\left(y - Q_e (s, \mathbf{a}, \theta)\right)^2 \right]. \tag{17}$$

where y represents the real Q value and is calculated as

$$y = w(s, \mathbf{a}, s') + \varepsilon \max_{a' \in\mathscr{A}}, Q_t (s', \mathbf{a}', \theta'), \tag{18}$$

where ε is the learning rate. As the loss function updates, the values are backpropagated to the neural network to update the weight of the eval_net.

In order to better train the neural network, we apply the experience reply method to remove the correlation between different training data. Each experience consists of the current system state s, the action $\mathbf{a}$, the next system state s', and the corresponding reward $w(s, \mathbf{a}, s')$. The experience is denoted by the set $ep = \{s, \mathbf{a}, w(s, \mathbf{a}, s'), s'\}$. The algorithm records D experiences, and randomly select D_s (with $D_s < D$) experiences from D for training. After the training is finished, target_net clones all the weight parameters from the eval_net (i.e., $\theta' = \theta$).

The algorithm used for the DQN training process is presented in Algorithm 1. In the algorithm, we define in each training iteration, we generate D usable experiences ep and select D_s of all for training the eval_net. In total, we suppose there are U training iterations. We consider that, for both the eval_net and the target_net, there are N_l layers in the neural network. In the learning process, we use $\mathbf{C}$ to denote all energy harvesters' channel condition in a particular time slot.

3.2.3. Dueling Double Deep Q-Network. Since more harvesters and time-varying channel conditions incur more system states, even if we utilize the original DQN, it is hard to study the transmit rules for the transmitter. Therefore, we can apply dueling double DQN in order to deal with the overestimating problem during the training process and improve the learning efficiency of the neural network. Doubling DQN is a technique that strengthens the traditional DQN algorithm by preventing overestimating to happen [17]. In traditional DQN, as shown in equation (18), we utilize the target_net to predict the maximum Q value of the next state. However, the target_net is not updated at every training episode, which may lead to an increase in the training error and therefore complicate the training process. In doubling DQN, we utilize both the target_net and the eval_net to predict the Q value. The eval_net is used to determine the optimal action to be taken for the system state s' as follows:

$$y = w(s, \mathbf{a}, s') + \varepsilon \max_{a' \in\mathscr{A}} Q_e \left(s', \arg\max_{\mathbf{a}\in\mathscr{A}} Q(s', \mathbf{a}, \theta), \theta' \right). \tag{19}$$

It can be shown that, following this approach, the training error considerably decreases [17].

In traditional DQN, the neural network only has the Q value as the output. In order to speed up the convergence, we apply dueling DQN by setting up two output streams from the neural network. The first stream is represented by the output value $V(\mathbf{s}, \theta, \beta)$ results of the neural network, which represents the Q value of each system state. The second stream is called advantage output $A(s', \mathbf{a}, \theta, \alpha)$ and describes the advantage of applying each particular action to the current system state [18]. α and β are parameters that relate the two streams and the neural network output, which is denoted as

$$Q(s, \mathbf{a}, \theta, \alpha, \beta) = V(s, \theta, \beta) + \left(A(s', \mathbf{a}, \theta, \alpha) - \frac{1}{|A|} \sum_{a'} A(s', \mathbf{a}, \theta, \alpha) \right). \tag{20}$$

Dueling DQN can efficiently eliminate the extra training freedom, which speeds up the training [18].

4. Simulation Results

We simulate a MIMO wireless communication system with nearby RF energy harvesters. The wireless transmitter has at

```
(1) Randomly generate the weight parameter $\theta$ for the eval_net. The target_net clones the weight parameters $\theta' = \theta$. $u = 1$. $s = s_0$.
    $\mathbf{C} = \mathbf{C}_t$. $t = 1$. $D = d = 1$.
(2) At the beginning of the time slot, randomly generate a probability $p \in [0, 1]$.
    IF $D > 200$ and $p \geq \varepsilon_{ch}$:
    we choose the action $\mathbf{a}$ as $\mathbf{a} = \max_{\mathbf{a} \in \mathscr{A}} Q(s, \mathbf{a})$
    ELSEIF $p < \varepsilon_{ch}$:
    Randomly choose the action from action set $\mathscr{A}$.
    The transmitter transmits with the selected beam pattern.
(3) Throughout the whole time slot, the RF energy is accumulated in the harvesters' energy buffer, as $s_i' = s_i + \sum_{m=1}^{M} |\widehat{g}_{im}^{t}|^2 a_m T, \forall i \in \mathscr{K}$.
    At the end of each time slot, each harvester feedbacks the energy level to the transmitter and the system state is updated to $s'$.
(4) $ep(d) = \{s, \mathbf{a}, w(s, \mathbf{a}, s'), s'\}$. $d = d + 1$. If $D$ reaches the maximum of experience pool, $D$ remains constant, $d = 1$, otherwise, $D = d$.
    $s = s'$. $t = t + 1$. $\mathbf{C} = \mathbf{C}_t$.
(5) After experience pool accumulates enough data, from $D$ experiences, randomly select $D_s$ experiences to train the neural network
    eval_net. Backpropagation method is applied to minimize the loss function loss$(\theta)$. Clone the weight parameters from eval_net to
    target_net after several time intervals.
(6) IF $s' = s_G$:
    $s = s_0$. $t = 1$. $\mathbf{C} = \mathbf{C}_t$. $u = u + 1$. If $u = U$, algorithm terminates; otherwise, go back to step 2.
    IF $s' \neq s_G$:
    go to step 3.
```

Algorithm 1: Deep Q-network algorithm training process.

most $M = 4$ antennas. The 4×4 communication MIMO channel matrix $\mathbf{H}$ is measured by two Wireless Open-Access Research Platform (WARP) v3 boards. Both WARP boards are mounted with the FMC-RF-2X245 dual-radio module, which is operated in 5.805 GHz frequency band. The Xilinx Virtex-6 FPGA operates as the central processing system and the WARPLab is used for rapid physical layer prototyping which is compiled by MATLAB [23]. We deploy two transceivers as line-of-sight transmission. The maximum transmitted power is $P = 12\text{W}$. $\rho^2 = -70\,\text{dBm}$. The information rate requirement R is 53 bps/Hz. The average channel gain from the transmitter to the energy harvester is -30 dB. The energy conversion efficiency is 0.1. The duration of one time slot is defined as $T = 100\,\text{ms}$.

DQN is trained to solve for the optimal transmit strategies for each system state. The simulation parameters used for DQN are presented in Table 1.

As described in Section 3.2, the exploration rate ε_c determines the probability that the network selects an action randomly or follows the values of the Q-table. Initially, we set $\varepsilon_c = 1$ because the experience pool has to accumulate reasonable amount of data to train the neural network. $\varepsilon_c = 1$ decreases with 0.001 at each training interval and finally stops at $\varepsilon_{ch} = 0.1$, since the experience pool has collected enough training data.

Refer to [24]. The dueling double DQN is used in our paper, which is shown in Figure 2. The software environment for simulation is TensorFlow 0.12.1 with Python 3.6 in Jupyter Notebook 5.6.0.

For the energy harvesters' channel, to show an example of the performance achievable by the proposed algorithm, we consider the Rician channel fading model [25]. We suppose within each time slot t, the channel is invariant and varies in different time slots [26]. At the end of each time slot, the energy harvester feedbacks the current energy level back to the transmitter. For the Rician fading channel model, the total gain of the signal is denoted as $\mathbf{g} = \mathbf{g}^s + \mathbf{g}^d$, where $\mathbf{g}^s$ is the invariant LOS component and $\mathbf{g}^d$ denotes a zero-mean Gaussian diffuse component. The channel between the transmit antenna m and the energy harvester i can be denoted as $g_{im} = g_{im}^s + g_{im}^d$. The magnitude of the faded envelope can be modeled using the Rice factor K^r such that $K_{im}^r = \rho_{im}^2 / 2\sigma_{im}^2$, where ρ_{im}^2 denotes the average power of the main LOS component between the transmit antenna m and energy harvester i and σ_{im}^2 denotes the variance of the scatter component. We can derive the magnitude of the main LOS component as $|g_{im}^s| = \sqrt{2K_{im}^r}\sigma_{im}$ since $1/2E[(|g_{im}^d|)^2] = \sigma_{im}^2$. The mean and the variance of g_{im} are denoted as $\mu_{g_{im}} = g_{im}^s$ and $\sigma_{g_{im}}^2 = \sigma_{im}^2$, respectively. In polar coordinates, $g_{im} = r_{im} e^{j\theta_{im}}$.

First, we explore the optimal deep Q-network structure under fading channels. We suppose the number of antennas is $M = 3$ and the number of energy harvesters is $K = 2$. The channel between each antenna of the transmitter and each harvester is individually Rician distributed. The action set $\mathscr{A}$ contains 13 actions satisfying the information rate requirement: $[2, 2, 8]^T$, $[2, 4, 6]^T$, $[2, 6, 4]^T$, $[2, 8, 2]^T$, $[4, 2, 6]^T$, $[4, 4, 4]^T$, $[4, 6, 2]^T$, $[4, 8, 0]^T$, $[6, 2, 4]^T$, $[6, 4, 2]^T$, $[6, 6, 0]^T$, $[8, 2, 2]^T$, and $[8, 4, 0]^T$.

The LOS amplitude components of all channel links are defined as $r_{im} = 0.5$, with $i = 1, 2$ and $m = 1, 2, 3$. The LOS phase components of all channel links are defined as $\theta_{11} = \pi/4$, $\theta_{12} = \pi/2$, $\theta_{13} = -\pi/4$, $\theta_{21} = -\pi/2$, $\theta_{22} = 0$, and $\theta_{23} = 3\pi/4$. The standard deviation of the g_{im} amplitude and phase is denoted as σ_{im} and $1/\sqrt{2K_{im}^r}$, respectively. We suppose $\sigma_{im} = 0.05, \forall i, m$. Hence, $1/\sqrt{2K_{im}^r} = (r_{im}/\sigma_{im})^{-1} = 0.1, \forall i, m$.

Using the fading channel model above, in Figure 3, we show how the structure of the neural network together with the learning rate can affect the performance of the DQN, for a fixed number of training episodes (i.e., 40000). The performance of DQN is measured by the average number of

Table 1: DQN simulation parameters.

Dueling Deep Q-network	Value
Number of hidden layers (N_L)	4
Number of nodes of each hidden layer	100
Learning rate (ε)	≤0.1
Mini-batch size	10
Learning frequency	5
Training starting step	200
Experience pool	≥20000
Initial exploration rate (ε_c)	1
Final exploration rate (ε_c)	0.1
Exploration interval	0.001
target_net weight replacement interval	≥100
Discount factor	0.9
Training episodes	≥40000

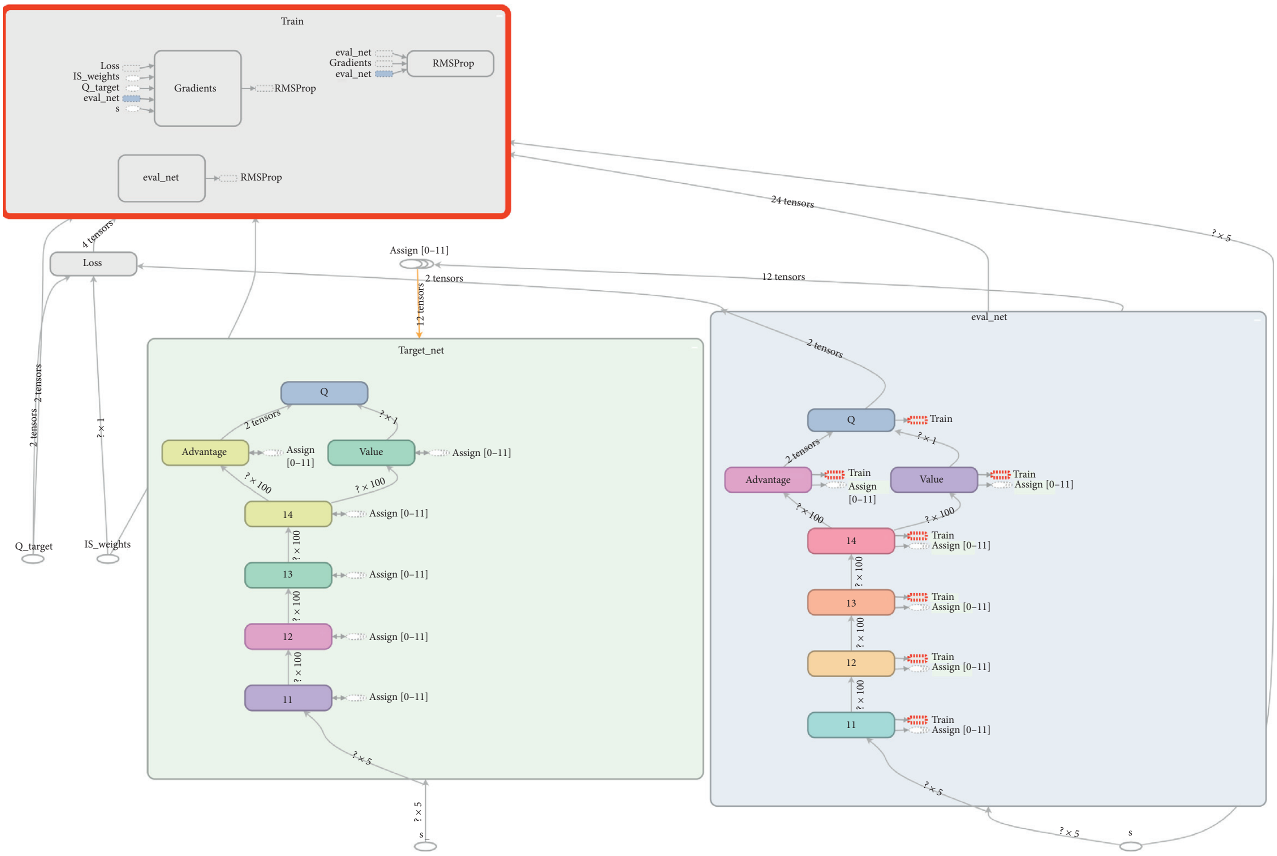


Figure 2: Dueling double deep Q-network structure.

time slots required to fully charge two harvesters. The average time-consumption is obtained over 1000 testing data. Figure 3 shows that if the deep Q-network has multiple hidden layers, a smaller learning rate is necessary to achieve better performance. When the learning rate is 0.1, the DQN with 4 hidden layers performs worse than a neural network with 2 or 3 hidden layers. On the other side, when the learning rate decreases, we can see that the neural network with 4 hidden layers and a learning rate of 0.00005 achieves the best overall performance. We do not see a monotonic decrease in the average number of time slots due to the stochastic nature of the channel that causes some fluctuations in the DQN optimization. After an initial improvement, decreasing the learning rate results in a slight increase in the average number of charging steps for all three neural network structures. This is due to the fixed number of training episodes. As a result, for all the simulations presented in this section, we consider a DQN algorithm using a 4 hidden layer deep neural network, with 100 nodes in each layer and a learning rate of 0.00005.

In Figure 4, we can observe that the size of the experience pool also affects the performance of DQN (40000 training episodes). To eliminate the correlation between the training data, we select part of the experience pool for training. In our

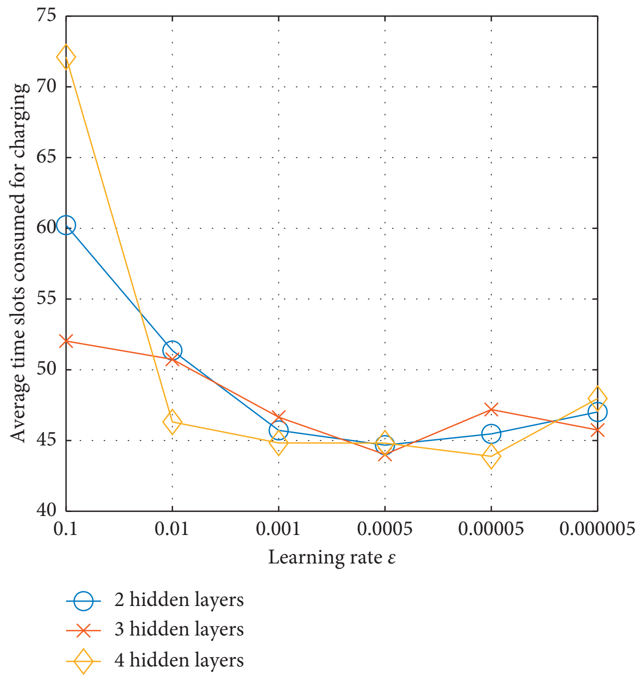


Figure 3: Deep Q-network performance on different learning rates and number of hidden layers for the neural network.

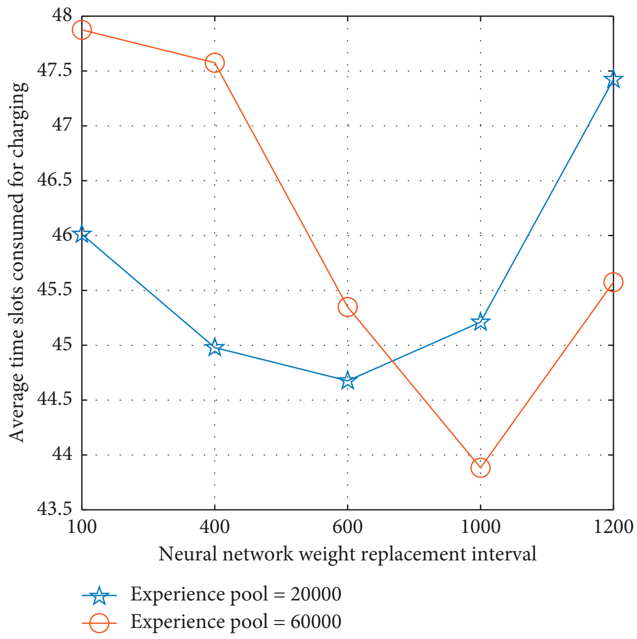


Figure 4: Deep Q-network performance for different values of neural network replacement iteration interval and experience pool.

simulation, this parameter, called mini-batch, is set to 10. Larger experience pool contains more training data; hence, selecting the mini-batch from it for training can eliminate the correlation between the training data. However, we need to balance the size of the experience pool and the target_net weight replacement interval. If the experience pool is large but the replacement iteration interval is small, even if we address the correlation problem between the training data, the neural network does not have enough training episodes to reduce the training error before the weight of the target_net is replaced. From Figure 4, we can observe that a large number of replacement iteration intervals may not be the best choice too. Therefore, we determine that, for our problem, DQN achieves the best performance when the size of the experience pool and the neural network replacement iteration interval are 60000 and 1000, respectively.

Figure 5 shows the impact of the reward function (see Section 3.1) on the DQN performance. In this figure, we consider the following reward functions: Reward$_1$: $w(s, \mathbf{a}, s') = 0$ if $s' = s_G$ and $w(s, \mathbf{a}, s') = -\lambda(\mathrm{K}B_{\max} - \sum_{i=1}^{K} \min(s'_i, B_{\max}))$ otherwise; Reward$_2$: $w(s, \mathbf{a}, s') = 10$ if $s' = s_G$ and $w(s, \mathbf{a}, s') = -1$ otherwise; Reward$_3$: $w(s, \mathbf{a}, s') = 1$ if $s' = s_G$ and $w(s, \mathbf{a}, s') = -1$ otherwise. Here, $K = 2$ and $\lambda = 0.25$. All three reward functions are designed to minimize the number of time slots required to fully charge all the harvesters. However, from Figure 5, we can observe that the best performance can be obtained using Reward$_1$. In this case, the energy level accumulated by each harvester increases uniformly, which results in the DQN to converge faster to the optimal policy. Both Reward$_2$ and Reward$_3$, instead, do not penalize states that unevenly charge the harvesters and therefore require more iterations to converge to the optimal solution (not shown in the figure) due to the large number of system states to explore. Therefore, in the following simulations, we use the reward function Reward$_1$ in both Figures 5 and 6, we average 40000 training steps every 100 steps in order to better show the convergence of the algorithm.

Figure 6 shows that when each energy harvester in the system is equipped with a larger energy buffer, the number of system states increases, and therefore, DQN requires more training period to converge to the steady transmit strategy for each system state. We can observe that when $B_{\max} = 1.6\,\mathrm{mJ}$, the system only needs less than 5000 training episodes to converge to the optimal strategy, and when $B_{\max} = 3.2\,\mathrm{mJ}$, the system needs around 12000 training episodes to converge to the optimal policy. However, for $B_{\max} = 4.8\,\mathrm{mJ}$, the system needs as many as 20000 training episodes to converge to the optimal strategy.

In the following simulations, we explore the impact of the channel model on optimization problem $\mathscr{P}_1$. For the Rician fading channel model, we consider $K^r_{im} \geq 10$ and to be the same for all i, m. In this way, we can approximate the Rician distribution as a Gaussian distribution. We fix $r_{im} = 0.5$, $\forall i, m$, but allow the standard deviation of both the amplitude and the phase of the channel to change to evaluate the performance on the system under different channel conditions. Since $r_{im} = 0.5$ and $\sqrt{2K^r_{im}} = r_{im}/\sigma_{im}$, $1/\sqrt{2K^r_{im}} = 2\sigma_{im}$. We define $\sigma_{im} \leq 0.1$ to guarantee $K^r_{im} \geq 10$.

In Figure 7, we express the standard deviation $\sigma_{\mathrm{amp}} = \sigma_{im}$, $\forall i, m$ of the phase and amplitude of the channel, and we compare the performance attained by the optimal policy with the performance of different other algorithms. The multiarmed bandit (MAB) algorithm is also implemented to compare with the DQN. In MAB, each bandit arm represents a particular transmission pattern. The upper confidence bound (UCB) algorithm [27] is implemented to

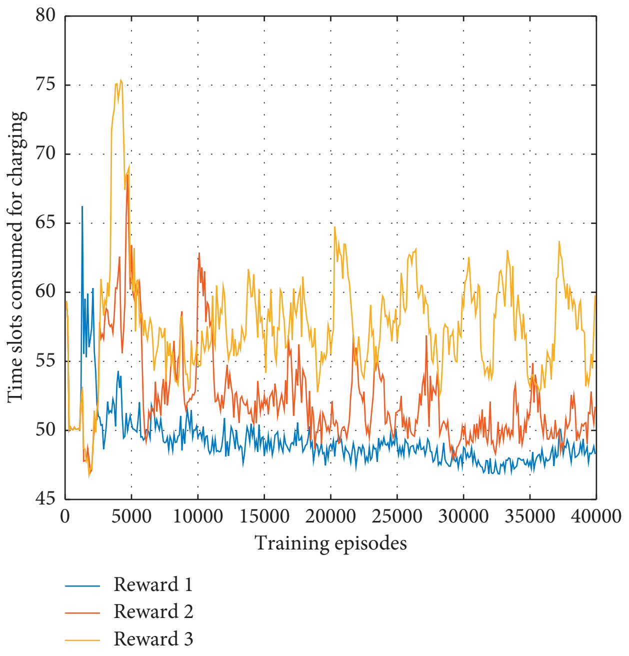


Figure 5: The deep Q-network performance for different reward functions.

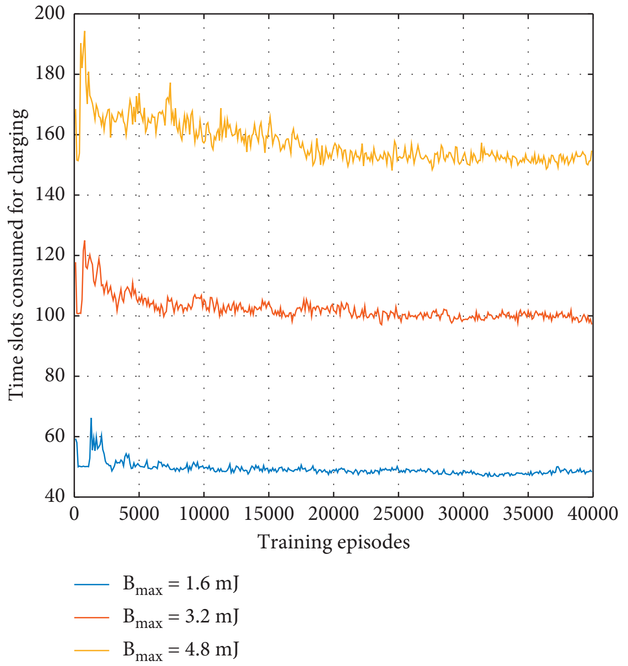


Figure 6: The deep Q-network performance for different energy buffer size B_{max}.

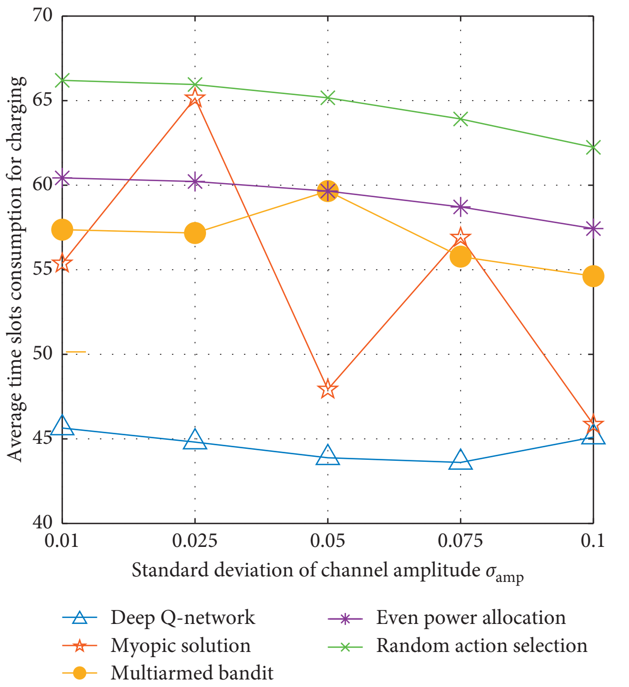


Figure 7: The comparison of average time consumption between DQN and other algorithms (myopic solution, multiarmed bandit, even power allocation, and random action selection) in the Rician fading channel model. The number of transmit antennas is $M = 3$. The number of energy harvesters is $N = 2$.

maximize the reward $w(s, \mathbf{a}, s')$ and determine the optimal action. Once the action is selected from the action space $\mathcal{A}$, it will be used for transmission continuously. The myopic algorithm is another machine learning algorithm that can be compared with DQN. Myopic solution has the same structure as the DQN; however, the reward discount is defined as $\gamma = 0$. As a result, the optimal strategy is determined only according to the current observation instead of considering the future consequence. Myopic solution has been widely used to solve the complex optimization in wireless communication problem and achieve good system performance [28]. Besides two machine learning algorithms, another two heuristic algorithms are also used for system performance comparison. For even power allocation, the transmit power P is evenly allocated on parallel channel for transmission. The random action selection is also applied for performance comparison. The random action selection has the worst performance while DQN performs best. Compared to the optimal existing algorithm multiarmed bandit algorithm, the DQN can consume 20% fewer time slots to complete charging. In some channel conditions, the myopic solution can achieve a similar performance as the DQN. However, the myopic solution cannot perform stably. For example, as the standard deviation of the channel amplitude is $\sigma_{amp} = 0.025$, DQN can outperform myopic solution by 45%. The instability can be explained as the myopic solution makes the decision only on the current system state and the current reward, which does not consider the future consequence. Hence, the training effects cannot be guaranteed. Overall, the DQN has superiority in both the charging time consumption and performing stability corresponding to different channel conditions.

To better explain the performance of the optimal policy, in Figure 8, we plot the action selected by DQN at a particular system state when $\sigma_{amp} = 0.05$. When $\sigma_{amp} = 0.05$,

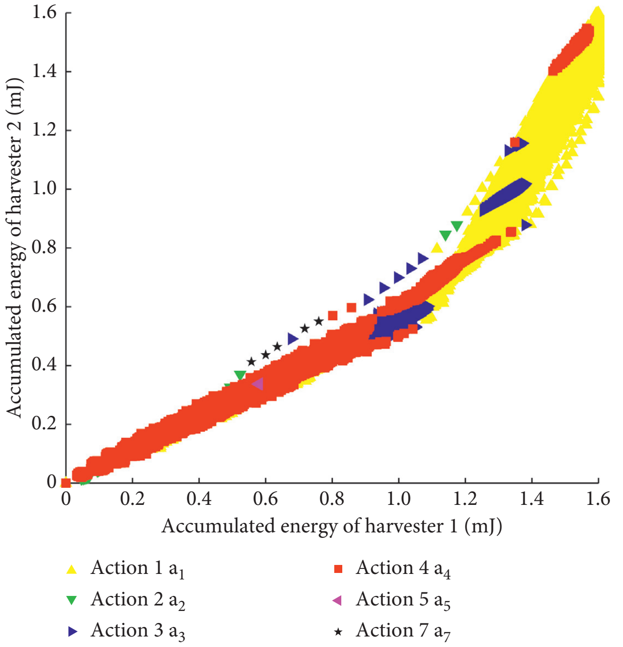


Figure 8: The action selection process of two harvesters scenario when $\sigma_{\text{amp}} = 0.05$.

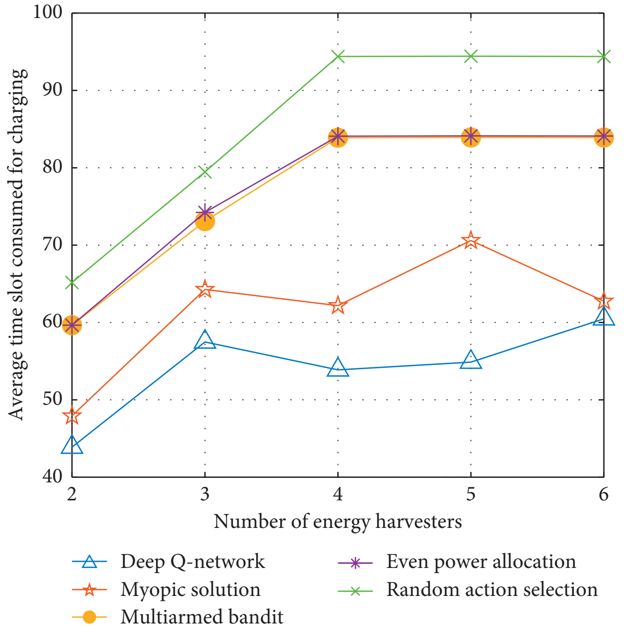


Figure 9: The comparison of average time consumption between DQN and other algorithms (myopic solution, multiarmed bandit, even power allocation, and random action selection) when the number of energy harvesters is $N = 2, 3, 4, 5, 6$. The number of transmit antennas is $M = 3$. The standard deviation of the channel amplitude is $\sigma_{\text{amp}} = 0.05$.

the optimal action selected by multiarmed bandit is the third action $\mathbf{a}_3 = [2, 6, 4]^T$, which can finish charging both harvesters in around 60 time slots. Meanwhile, the optimal policy determined by DQN can finish charging in around 43 time slots. To this end, Figure 8 shows that the charging process can actually be divided into two parts: before harvester 1 accumulates 1.2 mJ energy and harvester 2 accumulates 0.8 mJ energy, mostly action 4 $\mathbf{a}_4 = [2, 8, 2]^T$ is selected. After that, mostly action 1 $\mathbf{a}_1 = [2, 2, 8]^T$ is selected. As defined above, both the amplitude and the phase of the channel are Gaussian distributed with zero standard deviation, $\widehat{\mathbf{g}}_1^0 = [0.05, 0.59, 0.11]^T$ and $\widehat{\mathbf{g}}_2^0 = [0.04, 0.19, 0.51]^T$. So when both the amplitude and the phase of the channel change, the simplified channel state information will be distributed around $\widehat{\mathbf{g}}_1^0$ and $\widehat{\mathbf{g}}_2^0$. As a result, it can be shown that a policy that selects either action 1 or action 4 with different probabilities can have better performance than the policy that only selects action 3. Henceforth, the DQN can consume 40% fewer time slots to fully charge two energy harvesters.

In Figure 9, the performance of the DQN is compared with the other four algorithms by varying the number of energy harvesters in the system. In general, as the number of energy harvesters increases, all four algorithms consume more time slots to complete the wireless charging process. Compared to the random action selection, DQN can consume at least 58% less time slots to complete the charging. The performance of the multiarmed bandit and the even power allocation is very similar, which can be explained as the optimal action determined by the multiarmed bandit algorithm is close to the even power allocation strategy. Compared with two fixed action selection strategies (multiarmed bandit and even power allocation), DQN can reduce the time consumption by up to 72% (when the number of energy harvesters is $N = 3$). The myopic solution is still not the optimal strategy. From the figure, we can observe that the myopic solution outperforms two fixed action selection algorithms. Even though in some conditions ($N = 6$), the performance difference between DQN and myopic solution is very small, the myopic solution consumes more than 15% of the time slot than DQN in average. Overall, the DQN is the optimal algorithm which consumes fewest time slots to fully charge all the energy harvesters regardless of the number of energy harvesters.

In Figure 10, the number of transmit antennas is increased from $M = 3$ to $M = 4$. The number of energy harvesters varies from $N = 2$ to $N = 6$. Though the number of antennas increases, the channel conditions between the transmitter and the energy harvesters become more complicated; DQN still outperforms all the other four algorithms. Compared with myopic solution, multiarmed bandit, even action selection, and random action selection, DQN can consume up to 27%, 54%, 55%, and 76% fewer time slots to fulfill the wireless charging, respectively. As the number of energy harvesters increases, the superiority of the DQN becomes more obvious compared to two fixed action selection algorithms, which can be explained as it is more inefficient to select one fixed action to deal with a more complicated varying channel environment. Even though in some conditions, the performance of the myopic solution and DQN is similar, the myopic solution is not stable in dealing with different energy harvesters conditions. The

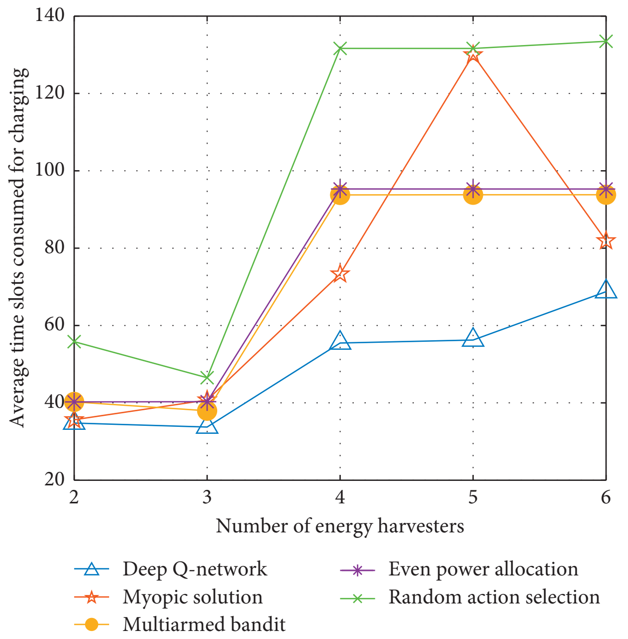


Figure 10: The comparison of average time consumption between DQN and other algorithms (myopic solution, multiarmed bandit, even power allocation, and random action selection) when the number of energy harvesters is $N = 2, 3, 4, 5, 6$. The n umber of transmit antennas is $M = 4$. The standard deviation of the channel amplitude is $\sigma_{\text{amp}} = 0.05$.

results from both Figures 9 and 1 0 demonstrate the superiority of the DQN in optimizing the time consumption for wireless power transfer.

5. Conclusions

In this paper, we design the optimal wireless power transfer system for multiple RF energy harvesters. Deep learning methods are used to enable the wireless transmitter to fully charge the energy buffers o f a ll e nergy h arvesters i n the shortest time while meeting the information rate requirement of the communication system.

As the channel conditions between the transmitter and the energy harvesters are time-varying and unknown, we model the problem as a Markov decision process. Due to the large number of system states in the model and the difficulty of training, we adapt a deep Q-network approach to find the best transmit strategy for each system state. In the simulation section, multiple experimental environments are explored. The measured real-time data are used to run the simulation. Deep Q-network is compared with the other four existing algorithms. The s imulation r esults v alidate t hat t he deep Q-network is superior to all the other algorithms in terms of the time consumption for fulfilling wireless power transfer.

References

[1] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 1989–2001, 2013.

[2] S. Lee, L. Liu, and R. Zhang, "Collaborative wireless energy and information transfer in interference channel," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 545–557, 2015.

[3] Z. Zong, H. Feng, F. R. Yu, N. Zhao, T. Yang, and B. Hu, "Optimal transceiver design for SWIPT in K-User MIMO interference channels," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 430–445, 2016.

[4] Y. Xing, Y. Qian, and L. Dong, "Deep learning for optimized wireless transmission to multiple rf energy harvesters," in *Proceedings of the 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Chicago, IL, USA, August 2018.

[5] J. Park and B. Clerckx, "Joint wireless information and energy transfer in a two-user MIMO interference channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4210–4221, 2013.

[6] W. Wu, X. Zhang, S. Wang, and B. Wang, "Max-min fair wireless energy transfer for multiple-input multiple-output wiretap channels," *IET Communications*, vol. 10, no. 7, pp. 739–744, 2016.

[7] A. Thudugalage, S. Atapattu, and J. Evans, "Beamformer design for wireless energy transfer with fairness," in *Proceedings of the 2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, May 2016.

[8] Y. Xing and L. Dong, "Passive radio-frequency energy harvesting through wireless information transmission," in *Proceedings of the 2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 73–80, Ottawa, ON, Canada, June 2017.

[9] P.-V. Mekikis, A. Antonopoulos, E. Kartsakli, A. S. Lalos, L. Alonso, and C. Verikoukis, "Information exchange in randomly deployed dense wsns with wireless energy harvesting capabilities," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 3008–3018, 2016.

[10] J. Xu and R. Zhang, "A general design framework for mimo wireless energy transfer with limited feedback," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2475–2488, 2016.

[11] K. W. Choi, D. I. Kim, and M. Y. Chung, "Received power-based channel estimation for energy beamforming in multiple-antenna RF energy transfer system," *IEEE Transactions on Signal Processing*, vol. 65, no. 6, pp. 1461–1476, 2017.

[12] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Playing atari with deep reinforcement learning," 2013, https://arxiv.org/abs/1312.5602.

[13] Y. Xing, Y. Qian, and L. Dong, "A multi-armed bandit approach to wireless information and power transfer," *IEEE Communications Letters*, vol. 24, no. 4, pp. 886–889, 2020.

[14] Y. He, Z. Zhang, F. R. Yu et al., "Deep reinforcement learning-based optimization for cache-enabled opportunistic interference alignment wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 433–510 445, 2017.

[15] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," 2016, https://arxiv.org/abs/1605.06676.

[16] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud rans," in *Proceedings of the 2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, Paris, France, May 2017.

[17] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," 2016, https://arxiv.org/abs/1509.06461.

[18] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," 2015, https://arxiv.org/abs/1511.06581.

[19] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*, Cambridge University Press, Cambridge, UK, 2007.

[20] S. Timotheou, I. Krikidis, S. Karachontzitis, and K. Berberidis, "Spatial domain simultaneous information and power transfer for mimo channels," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4115–4128, 2015.

[21] D. Mishra and G. C. Alexandropoulos, "Jointly optimal spatial channel assignment and power allocation for mimo swipt systems," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 214–217, 2018.

[22] A. G. Barto, S. J. Bradtke, and S. P. Singh, "Learning to act using real-time dynamic programming," *Artificial intelligence*, vol. 72, no. 1-2, pp. 81–138, 1995.

[23] L. Dong and Y. Liu, "Parallel sub-channel transmission for cognitive radios with multiple antennas," *Wireless Personal Communications*, vol. 79, no. 3, pp. 2069–2087, 2014.

[24] Y. Xing, H. Pan, B. Xu, T. Zhao, C. Tapparello, and Y. Qian, "Multiuser data dissemination in OFDMA system based on deep q-network," in *Proceedings of the IEEE IEMTRONIC (International IOT, Electronics and Mechatronics Conference)*, Toronto, Canada, April 2021.

[25] J. Cavers, *Mobile Channel Characteristics*, Springer Science & Business Media, Berlin, Germany, 2006.

[26] T.-Q. Wu and H.-C. Yang, "On the performance of overlaid wireless sensor transmission with rf energy harvesting," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 8, pp. 1693–1705, 2015.

[27] A. Slivkins, "Introduction to multi-armed bandits," 2019, https://arxiv.org/abs/1904.07272.

[28] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 2, pp. 257–265, 2018.

Fuzzy-Swarm Intelligence-Based Short-Term Load Forecasting Model as a Solution to Power Quality Issues Existing in Microgrid System

Sunil Kumar Mahapatro, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, skmahapatro78@outlook.com*

Prativa Barik, *Department of Electrical and Electronics Engineering, Raajdhani Engineering College, Bhubaneswar, p.barik213@gmail.com*

Srichandan Subhrajit Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, ss.sahoo93@outlook.com*

J. Uday Bhaskar, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, j.udaybhaskar1@gmail.com*

Abstract

Load demand is highly stochastic and uncertain. This is because it was highly influenced by a number of variables like load type, weather conditions, time of day, the seasonality factor, economic constraints, and other randomness effects. The loads are categorized as holiday loads (national and religious), weekdays, and weekend days. The nonlinearity and uncertain characteristics of electrical load in a microgrid are one of the major sources of power quality problems in a microgrid system, and they can be handled using an accurate load forecast model. The fuzzy load prediction model can effectively handle these nonlinearity and uncertainty characteristics to have an accurate load forecast, but the main challenge with this model is its inability to accommodate a large volume of historical load and weather information when the membership function of the input and output fuzzy variables and the number of the fuzzy rules are tremendous. The swarm intelligence load forecast model based on particle swarm optimization algorithms can improve the limitations of the fuzzy system and increase its forecasting performance. The parameters of time, temperature, historical load, and error correction factors are considered as the Fuzzy and Fuzzy-PSO model input variables, while the forecasted industrial load is the only output variable. The Gaussian membership function is considered for both the input and output fuzzy variables. A 3-year historical hourly load data of an Ethiopian industrial system is used to train and validate both prediction models. The mean absolute percentage error (MAPE) is used to evaluate the performance of these prediction models. The Fuzzy-PSO load prediction model shows results that have superior performance to the fuzzy-alone load prediction results.

1. Introduction

Load forecasting is a critical component of building an energy management system in a microgrid. It is categorized as very-short-term load forecast (VSTLF), short-term load forecast (STLF), medium-term load forecast (MTLF), and long-term load forecast (LTLF), which are based on the forecasting horizon and the application of predicting the load. The very-short-term load forecasting predicts the load in minutes to the hourly interval. Short-term load forecasting (STLF) predicts one-day up to one-week hourly loads, whereas medium-term load forecasting predicts loads from one-month to one-year time horizon, and long-term load forecasting is the prediction of loads that are more than one-year period [1–3]. The purpose of demand forecasting is to schedule energy generation, assess the security of the power system, and also schedule electricity prices. In order to reduce the capacity and investment cost of the energy storage system, we should maintain a permanent balance between generation and consumption. A disparity between supply and demand will lead to economic losses for the utility. Undergeneration will compromise the reliability and security of the grid, resulting in power outages. These power interruptions will necessitate compensation of the customers by the power suppliers, thus leading to reduced profits or even losses. On the other hand, overgeneration will result in losses due to high generation costs. Forecast information also influences the customers' decisions on energy management strategies like load shedding to smooth the load curve and limit peak loads. Utilities also place electricity price bids based on the future values of the demand and its corresponding prices [1, 3]. In this research, the STLF strategy was implemented in order to accurately balance generation and demand. It also helps us to maintain the power quality and stability of the distribution system with the minimum capacity of an energy storage system. Various STLF methods have been previously developed. Some of these methods include regressions, neural networks, similar day approach, time series method, fuzzy logic, and hybrid forecasting methods. The regression technique [4–11] is one of the most widely used statistical techniques usually employed to model the relationship between load consumption and other factors such as weather, day type, and customer class, but this technique needs a large volume of historical data to develop a mathematical load relationship.

Kumar Singla et al. [12] developed a short-term load prediction model using the Mamdani approach fuzzy logic toolbox and obtained a significant prediction error improvement between +3.67% and −3.75%. According to Gohil and Gupta [13], the fuzzy load prediction model provides a significant accuracy level for predicting holiday and working day loads as compared to conventional approaches. A forecast for special days (holidays) is carried out using an artificial neural network (ANN) and a fuzzy inference method [14]. The simulation result shows the ANN model improves the accuracy level of the fuzzy. Cevick and Cuncas [15] used the fuzzy load prediction approach to forecast future holiday loads. The study provides a significant accuracy level of MAPE of between 2.03% and 11.29%.

Gao et al. [16, 17] developed a short-term load forecasting method based on a least squares support vector machine (LS-SVM) combined with fuzzy control and bacterial colony chemotaxis optimization algorithm. In [16], the methodology is based on the prediction of the peak and valley loads and determining the prediction coefficients using the fuzzy rule tuning approach for the prediction of similar day future loads, whereas in [17] the bacterial colony chemotaxis optimization algorithm is used to determine hyperparameters of LS-SVM. Both methods improve the prediction accuracy. Short-term load forecasting using an artificial neural network (ANN) technique is conducted by [18, 19]. In [18], the weekday and weekend day loads are separately treated, and the training of neural network has been done separately for weekdays and weekend days. The neural network toolbox with 20 neurons has been used for forecasting an island load, and the prediction accuracy is promising. According to [19], short-term load forecasting is done to optimally estimate the load flow in a certain power system network. The result shows an accurate load forecast that helps with optimal generation planning and load flow studies.

Yang et al. [20–22] developed an improved Wang–Mendel fuzzy model based on the PSO algorithm to improve the learning capability and forecasting accuracy of the fuzzy system. The PSO algorithm helps to optimize the fuzzy rules of the WM fuzzy model. The model yields a very good forecasting accuracy of MAPE of 2.57%. The fuzzy load prediction model for short-term load forecasting lacks self-learning and tuning capability for stochastic and nonlinear load variations. The fuzzy rule base and the fuzzy membership function are the two variables tuned by using PSO. The PSO considers the mean-square error as the optimization objective function from the fuzzy membership function tuning approach. The simulation result demonstrated a significant improvement in forecasting accuracy.

Swarm intelligence is an area of artificial intelligence based on the collective and decentralized behavior of individuals that interact with each other and with the environment. PSO is a stochastic evolutionary algorithm based on swarm intelligence that searches for the solution to optimization problems in a specific search space and is able to predict the social behavior of individuals according to defined objectives [23]. A deep learning model for day-ahead load forecasting based on expert knowledge is discussed in [24, 25]. In these methods, the peak load is forecasted using nonlinear historical load data, temperature data, and economic metric data over a similar time horizon. In [25, 26], a hybrid short-term load forecasting approach using a fuzzy logic control system is developed. The short-term load forecasting approach in [26] and a short-term load forecasting model of a nuclear charging station based on PSO-SVM are proposed. The PSO is used to optimize the parameters of the support vector machine (SVM) for optimal charging and discharging operations of a nuclear charging station. The load forecast result based on the normalized root-mean-square error (NRMS) as a fitness measuring tool is used as a prewarning signal of the charging station. The Fuzzy-PSO load prediction model was also discussed in [27]. In [27], the load prediction model only accounts for the weekday and weekend industrial load, and the result obtained is at a high level of prediction accuracy. This paper is an extension of the paper in [27], which accounts for holiday loads and adds a new fuzzy input parameter called the error correction factor that further helps to tune the fuzzy system and also improves the forecasting accuracy.

The rest of the paper is organized as follows. Section 2 discusses the analysis of industrial raw data. Section 3 presents the problem formulation of the Fuzzy-PSO load prediction model. Section 4 is about results and Section 5 is about the discussion. Section 6 contains the conclusion, and the references and lists of abbreviations are at the end of the paper.

2. Industrial Load Data Analysis

Two data sets are required for an effective forecast model of the industrial load. These data sets are the training data set and the model validation (testing) data set. The training and testing load data and weather information are collected from an industrial load that is fed from a 15 kV Kaliti substation in Addis Ababa, Ethiopia. The 24-hour load data is collected from 2017 to 2020 G.C and the holiday, weekday (Monday to Friday), and weekend day (Saturday and Sunday) annual average hourly load data is identified and analyzed for the

Fuzzy-PSO load prediction model design. The d ata from 2017 to 2018 is used to train the Fuzzy-PSO model, whereas the 2019/20 load data is used to validate the model in all the load categories mentioned earlier. The M ATLAB programming and MATLAB/Simulink working environments are used to model both the Fuzzy and Fuzzy-PSO load prediction models. Both the training and validation industrial load data sets are prepared in a 24-hour data format based on the raw industrial load data that was collected from the case study area.

$$P_i = \frac{1}{T}\sum_{t=1}^{T} L_t, \tag{1}$$

where P_i is the average load in the ith hour, which is computed using the average value of the available load data size in a similar time frame. T is the total load data size in days, and it is 365 for the validation data set, but 730 for the training load data set. L_t is the industrial load in the ith hour of the tth day.

2.1. Holiday Load Profile. In the study area, the categories and number of holidays in each year are identified. In Ethiopia, five national holidays, five Christian holidays, and three Muslim holidays are found in a calendar year. After analysis using (1), the training and testing load profile curve from 2017 to 2020 is presented in Figures 1–6. Since each holiday has different load characteristics, the load prediction model is separately treated.

2.2. Weekday Load Profile. For the training and testing of the Fuzzy-PSO load prediction model, the annual average weekday load data from 2017 to 2020 G.C is computed using (1) and presented in Figure 7. The load data in 2017 and 2018 is used to train the Fuzzy-PSO load prediction model, whereas the annual average load data in the 2019/20 calendar year is used to validate the load prediction model for this load category.

2.3. Weekend Load Profile. For the training and testing of the Fuzzy-PSO prediction model, the annual average weekend load data from 2017 to 2020 G.C is presented in Figure 8. The load data in 2017 and 2018 is used to train the Fuzzy-PSO load prediction model, whereas the annual average load data in the 2019/20 calendar year is used to validate the model for the weekend load.

2.4. Overall System Load Profile. The total system load in Figure 9 is used to forecast the overall industrial system load in order to model the microgrid system which is part of ongoing research. To consider the inconsistency of load variation in various events mentioned, a scaling factor (S_F) based on the peak load and average system load ratio is considered for both training and testing data sets. The scaling factor is computed using the peak and average industrial load values of both the training and testing load data set values. It can be expressed as

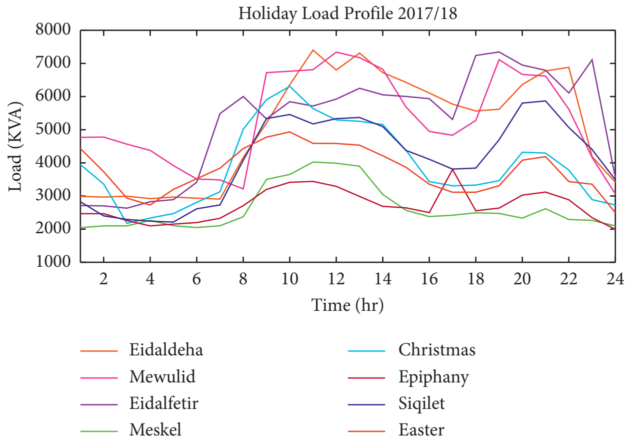


FIGURE 1: Hourly load profile of religious holidays in 2017/18.

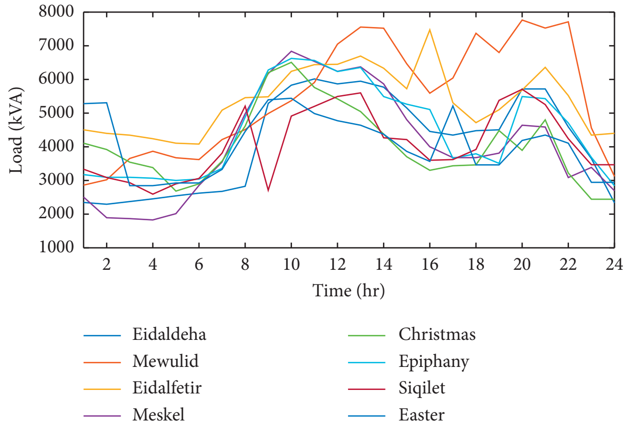


FIGURE 2: Hourly load profile of religious holidays in 2018/19.

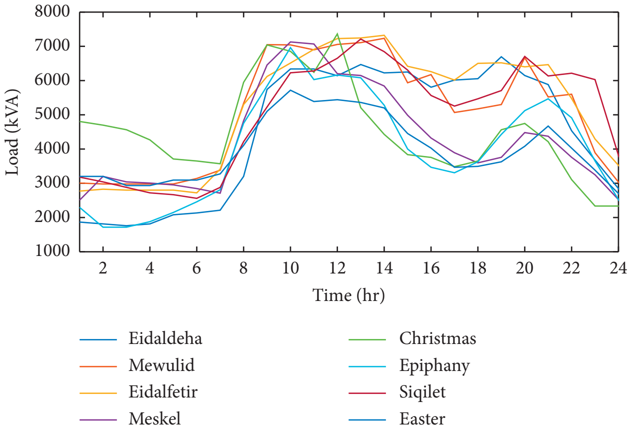


FIGURE 3: Hourly load profile of religious holidays in 2019/20.

$$S_F = \frac{L_{\text{peak}}}{L_{\text{mean}}}, \tag{2}$$

where S_F is the scaling factor, L_{peak} is the peak demand, and L_{mean} is the mean demand of the corresponding training and testing data set values.

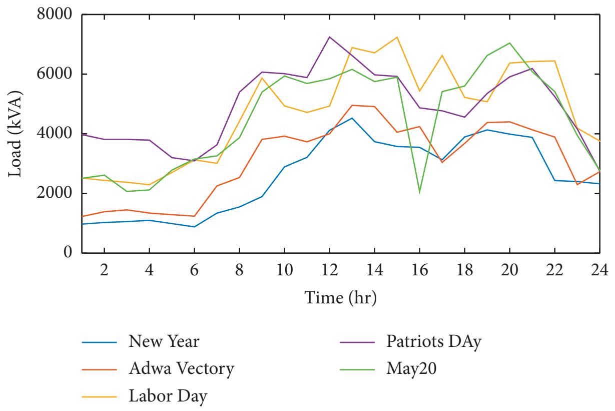


Figure 4: Hourly load profile of national holidays in 2017/18.

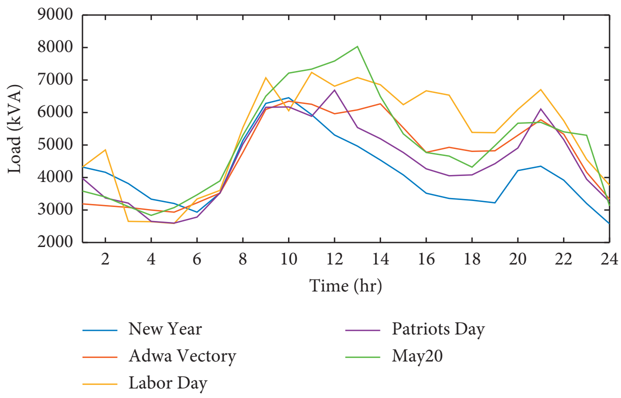


Figure 5: Hourly load profile of national holidays in 2018/19.

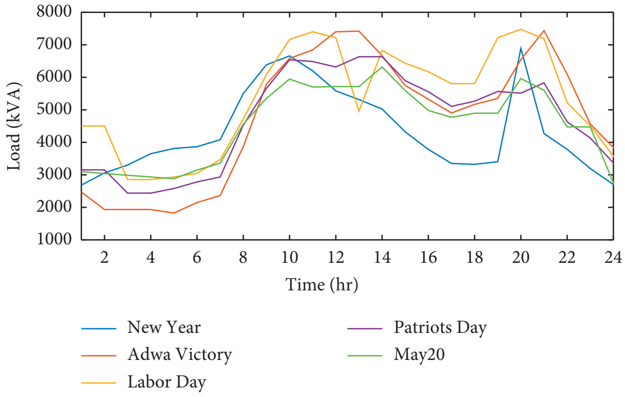


Figure 6: Hourly load profile of national holidays in 2019/20.

The Fuzzy-PSO load prediction model proposed in this paper can be scaled up by increasing the volume of the training and validation dataset. In addition to that, it also expanded by considering the different critical fuzzy input parameters that have a direct impact on load prediction result.

3. Problem Formulation of the Fuzzy-PSO Load Prediction Model

In this research, a new approach of industrial load forecasting to model a microgrid system has been developed based on a raw industrial load data from the field. The main contributions of this work are as follows:

(i) Introducing the PSO algorithm in the fuzzy load forecast model based on the training and validation data set correlation. An automatic learning and training input data-based generation of the fuzzy rules and optimal parameters of the Gaussian fuzzy membership functions.

(ii) A new fuzzy input variable called the error correction factor (ECF) has been introduced to further enhance the performance of the prediction model and.

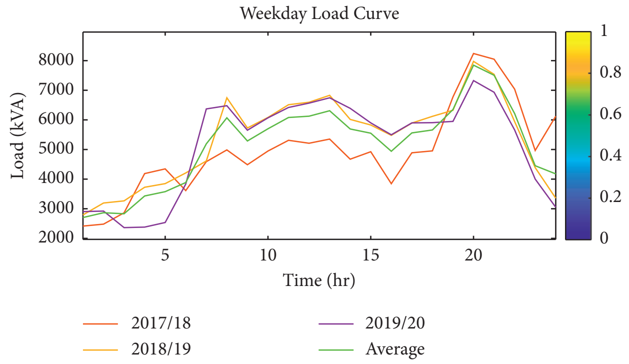


Figure 7: Weekdays hourly average load profile in 2017–2020.

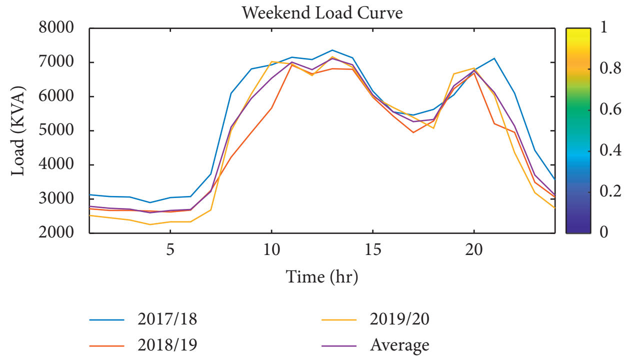


Figure 8: Weekend days hourly average load profile in 2017–2020.

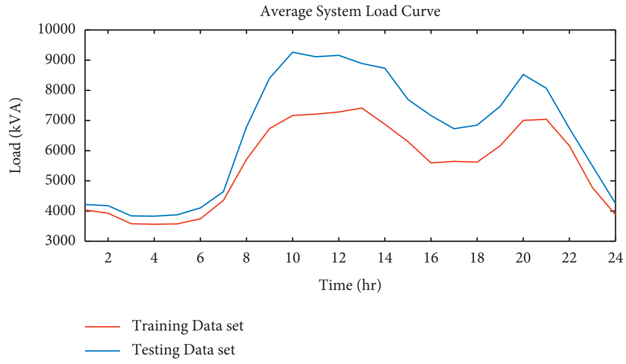


Figure 9: All events' hourly average training and testing load data set of the system.

(iii) The ECF fuzzy variable is computed from the Simulink fuzzy load prediction model based on equation (3) and it is considered as a time series fuzzy input variable.

The ECF is a time series normalized error of the validation load and forecasted load data values. The deviation of the forecast value from the validation data set gives us a pre-indicator to understand the forecasting direction and model adjustment of the Fuzzy-PSO industrial load prediction model.

3.1. Fuzzy Logic System. The load prediction model is based on four fuzzy input variables and a single fuzzy output variable whose Gaussian membership function and all the rules are later optimized using the particle swarm optimization algorithm. The fuzzy is trained using PSO based on the available input and output training data set correlation. The fuzzy input variables are temperature, error correction factor, historical load, and time of the day. The forecasted load is the only output fuzzy variable. Both the input and output fuzzy variables have Gaussian membership function as shown in Figure 10 and it has two fundamental parameters: the mean (c) and the standard deviation (b). The error correction factor introduced in this research has highly improved the performance of the prediction model, and it can be computed based on the available data values using the following mathematical relation, which was later modeled in Simulink.

$$\mathrm{ECF} = \frac{F_A - F_F}{F_A}, \tag{3}$$

where F_A is the actual load data set and F_F is the forecasted load data set. The ECF data is automatically calculated in the fuzzy load prediction model using Simulink and a 24-hour ECF time series data is obtained in the process. The ECF data computed in the fuzzy prediction model is later used to train the fuzzy model using PSO.

The fuzzy inference engine incorporates all possible rules that help to map the output from the fuzzy inputs based on the training dataset. Let A_i^k be the ith fuzzy input variable (x_i) of k's membership function and B^t is the t's membership function of the output variable (y_i); then the fuzzy rule is generated as follows:

If x_1 is A_1^k and x_2 is A_2^k and .... and x_n is A_n^k, then y_i is B^t.

The fuzzifier in the fuzzy load prediction model is formulated as follows [28]:

$$\mu_{A_{l}}(x_i') = \begin{cases} 1, & \text{if } x_i' = x_i, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The defuzzification of the load prediction model, based on the center-average defuzzifiers, is also formulated as follows [28]:

$$y = \frac{\sum_{l=1}^{T} y_c^l \omega^l}{\sum_{l=1}^{T} \omega^l}. \tag{5}$$

3.2. Particle Swarm Optimization. PSO is basically the intelligence of bird flocking or fish schooling and it was first introduced by Kennedy and Eberhart in 1995 [29] and later the original PSO is modified by Shi and Eberhart in 1999 [30] in order to improve the convergence rate and accuracy level by considering the inertia weight factor, ω.

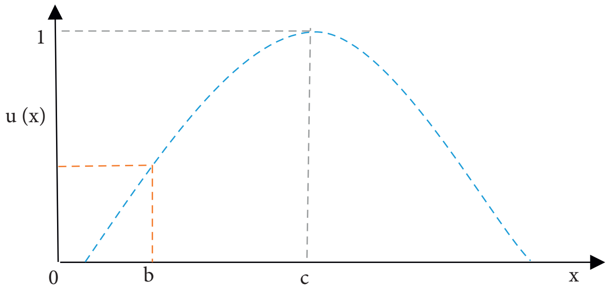


FIGURE 10: The Gaussian fuzzy membership function.

The algorithm is given as follows [27]:

(i) PSO parameters are being initialized [ω, c_1, c_2, r_1, r_2].

(ii) Random generation of the initial solution in the swarm's search space. The velocity and position of each particle are randomly initialized and evaluated according to the fitness function of the initial solution in order to determine the personal best and global best particle in the population.

(iii) Update the velocity and position of each particle using the following formula.

$$\begin{aligned} V_{id}^{t+1} &= \omega . V_{id}^{t} + c_1 r_1 \left(P_{id}^{t} - x_{id}^{t}\right) + c_2 r_2 \left(G_d^t - x_{id}^{t}\right), \\ x_{id}^{t+1} &= x_{id}^{t} + V_{id}^{t+1}. \end{aligned} \tag{6}$$

(iv) Update the personal best and global best position of the particle at each iteration.

$$\begin{aligned} P_{id}^{t+1} &= x_{id}^{t+1} \text{ if } f\left[X_k(t+1)\right] \le f\left[P_k^b(t)\right], \\ G_k^b(t+1) &= X_k(t+1)\, if\ f\left[X_k(t+1)\right] \le f\left[G_k^b(t)\right]. \end{aligned} \tag{7}$$

(v) Go to step (iii) when the termination criterion is not satisfied; otherwise, terminate the algorithm.

f is the fitness function; in this research, it is the mean absolute percentage error (MAPE), t is the current iteration, c_1 and c_2 are the acceleration coefficients, r_1 and r_2 are evenly distributed random number in the range 0 to 1, and $\mathrm{P}_{\mathrm{id}}^{\mathrm{t}}$ and G_d^t are the personal best and global best positions of the particles in the population.

3.3. Premature Convergence Problem of Fuzzy-PSO Algorithm. Early convergence to the local optimum point is one of the problems that rarely happen in a particle swarm optimization algorithm. The control of the inertia weight (ω), the cognitive, and social acceleration coefficients will improve the performance of PSO algorithm in various application. Zhang et al. [31] proposed a mutation strategy to avoid the premature convergence to local optimum of PSO. In this paper, in order to avoid the premature convergence of the

Fuzzy-PSO load prediction model, a method by Ratnaweera et al. [32] was used. This method needs less computational time, is easy to implement, and integrates to the basic PSO algorithm. According to [32], a linearly increasing social acceleration factor and a linearly decreasing cognitive acceleration factor are of help to boost the global searching capability while also balancing the local searching of the particles in a multidimensional particles search space. The inertia weight (ω), cognitive acceleration factor (c_1), and social acceleration factors (c_2) are dynamically varying and computed using equations (8)–(10) and later incorporated to the PSO algorithm in MATLAB programming. At every iteration, these PSO parameters are adjusted to minimize and avoid the premature convergence problem of the Fuzzy-PSO load prediction model.

$$\omega(t) = \omega_{\max} - \frac{(\omega_{\max} - \omega_{\min}) \times t}{I_t}, \tag{8}$$

$$c_1(t) = c_{1\max} + \frac{(c_{1\min} - c_{1\max}) \times t}{I_t}, \tag{9}$$

$$c_2(t) = c_{2\max} + \frac{(c_{2\max} - c_{2\min}) \times t}{I_t}. \tag{10}$$

Here, t is the current iteration, $I_t = 300$ is the total number of iterations, $\omega_{\max} = 0.99$ is the maximum inertia weight, $\omega_{\min} = 0.1$ is the minimum inertia weight, $c_{1\min} = 0.25$ is the minimum value of cognitive acceleration factor, $c_{1\max} = 2.5$ is the maximum value of cognitive acceleration factor, $c_{2\min} = 0.25$ is the minimum value of social acceleration factor, and $c_{2\max} = 2.5$ is the maximum value of social acceleration factor.

3.4. Encoding of Fuzzy Variables Using the PSO Algorithm. The PSO algorithm is used to tune the fuzzy membership function of an industrial load prediction model using the Mamdani fuzzy inference system approach. Figure 11 demonstrates the tuning process of the Fuzzy-PSO load prediction model to improve the performance of the fuzzy logic system. There are two parameters (standard deviation b and mean c) which represent each fuzzy membership function of the "n" input variables and the "q" fuzzy output variable of the fuzzy system. If each input variable has an "m" fuzzy membership function and each fuzzy output variable has a "t" fuzzy membership function, then the total parameters of the input (X) and output (Y) fuzzy variables are represented in equations (11) and (12). The parameters (b_i and c_i) of every fuzzy membership function in each fuzzy variable (input and output) can be encoded with the PSO particles using Eq. 13–Eq. 20.

$$X = \sum_{i=1}^{n} 2 \times m_i, \tag{11}$$

$$Y = \sum_{i=1}^{q} 2 \times t_i, \tag{12}$$

$$X = \begin{bmatrix} x^b \\ x^c \end{bmatrix}, \tag{13}$$

$$Y = \begin{bmatrix} y^b \\ y^c \end{bmatrix}, \tag{14}$$

$$x^b = \begin{bmatrix} x^b_{11} & x^b_{12} \cdots x^b_{1m} \\ x^b_{21} & x^b_{22} \cdots x^b_{2m} \\ . & . \; . \; . \\ . & . \quad . \quad . \\ x^b_{n1} & x^b_{n2} \cdots x^b_{nm} \end{bmatrix}, \tag{15}$$

$$x^c = \begin{bmatrix} x^c_{11} & x^c_{12} \cdots x^c_{1m} \\ x^c_{21} & x^c_{22} \cdots x^c_{2m} \\ . & . \; . \; . \\ . & . \quad . \quad . \\ x^c_{n1} & x^c_{n2} \cdots x^c_{nm} \end{bmatrix}, \tag{16}$$

$$y^b = \begin{bmatrix} y^b_{11} & y^b_{12} \cdots y^b_{1t} \\ y^b_{21} & y^b_{22} \cdots y^b_{2t} \\ . & . \; . \; . \\ . & . \quad . \quad . \\ y^b_{q1} & y^b_{q2} \cdots y^b_{qt} \end{bmatrix}, \tag{17}$$

$$y^c = \begin{bmatrix} y^c_{11} & y^c_{12} \cdots y^c_{1m} \\ y^c_{21} & x^c_{22} \cdots y^c_{2m} \\ . & . \; . \; . \\ . & . \quad . \quad . \\ y^c_{q1} & y^c_{q2} \cdots y^c_{qm} \end{bmatrix}, \tag{18}$$

$$x^b_{11} = [\, p_1 \;\; p_2 \;\; \cdots \;\; p_s \,], \tag{19}$$

$$y^b_{11} = [\, p_1 \;\; p_2 \;\; \cdots \;\; p_s \,], \tag{20}$$

where X is a $2 \times n \times m$ parameter of the input fuzzy variable, Y is a $2 \times q \times t$ parameter of the fuzzy output variables, x^c_{ij} is the mean parameter of the input variables' ijth membership function, x^b_{ij} is the standard deviation parameter of the input variables' ijth membership function, b_i and c_i are the fuzzy Gaussian membership function parameters for each input and output membership function.

A summary of the above discussion is presented in Table 1. A total of 42 unknown fuzzy parameters needs to be optimized using the PSO algorithm. If the swarm size is 100, then all the 42 parameters are encoded with 100 PSO particles, and in every iteration of running the simulation of the PSO algorithm, 42 fuzzy membership function parameters are computed, and the values are updated after evaluating the fitness of the cost function. Therefore, the parameters of both the input and output fuzzy variables are encoded with the particles in the swarm. For example,

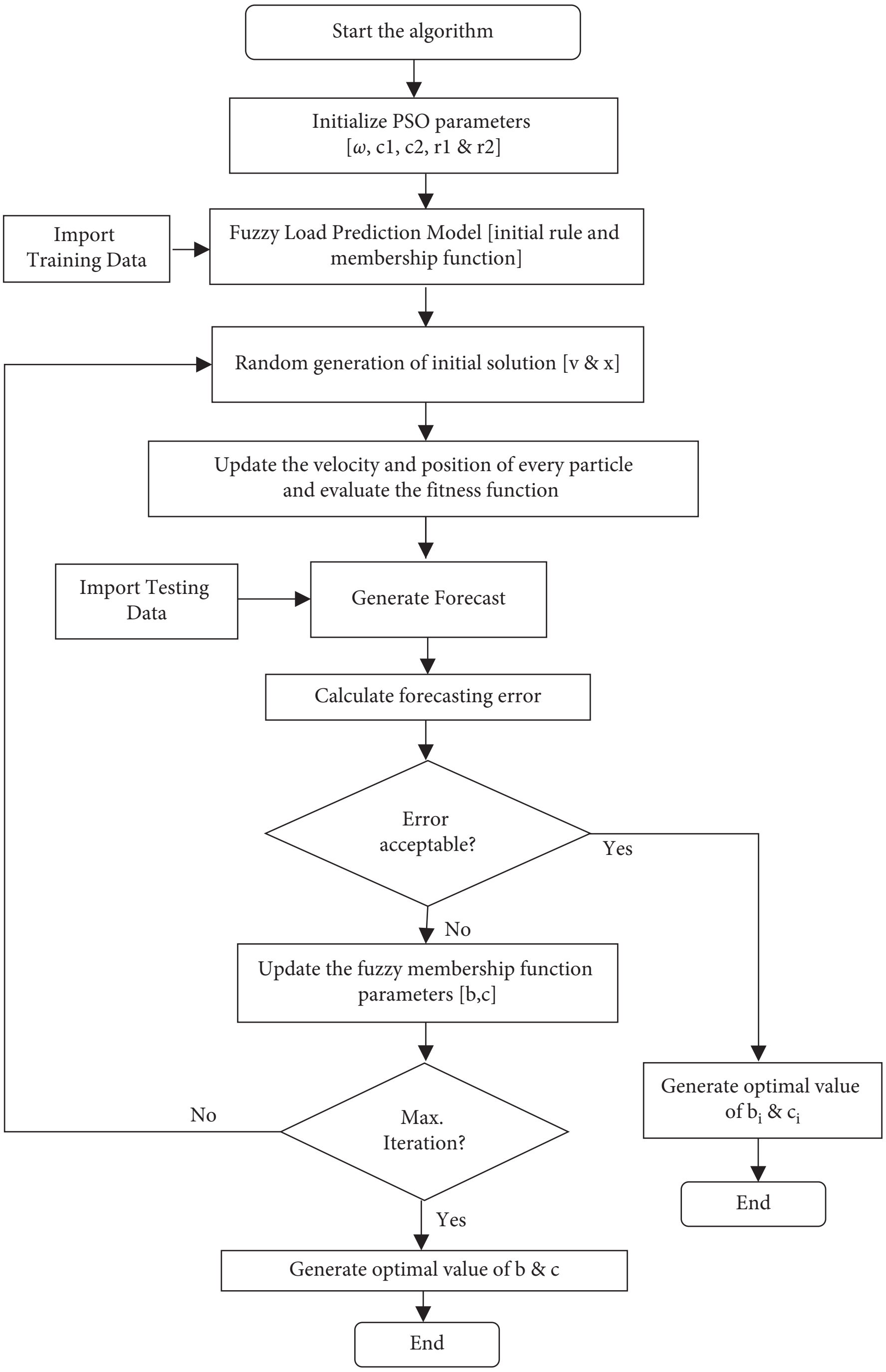

Start the algorithm
Initialize PSO parameters [ω, c1, c2, r1 & r2]
Import Training Data
Fuzzy Load Prediction Model [initial rule and membership function]
Random generation of initial solution [v & x]
Update the velocity and position of every particle and evaluate the fitness function
Import Testing Data
Generate Forecast
Calculate forecasting error
Error acceptable?
Yes
No
Update the fuzzy membership function parameters [b,c]
Max. Iteration?
No
Yes
Generate optimal value of b & c
End
Generate optimal value of b_i & c_i
End

Figure 11: Fuzzy-PSO tuning process for load prediction.

Table 1: The fuzzy parameter encoding using the particles of PSO.

Parameters	Parameters of the fuzzy membership function										Dimensional search space of particles
	b_i^1	c_i^1	b_i^2	c_i^2	b_i^3	c_i^3	b_i^4	c_i^4	b_i^5	c_i^5	
Time	x_{b11}	x_{c11}	x_{b12}	x_{c12}	x_{b13}	x_{c13}	x_{b14}	x_{c14}	x_{b15}	x_{c15}	$2 \times m_1 = 2 \times 5 = 10$
Temperature	x_{b21}	x_{c21}	x_{b22}	x_{c22}	x_{b23}	x_{c23}					$2 \times m_2 = 2 \times 3 = 6$
ECF	x_{b31}	x_{c31}	x_{b32}	x_{c32}	x_{b33}	x_{c33}					$2 \times m_3 = 2 \times 3 = 6$
Historical load	x_{b41}	x_{c41}	x_{b42}	x_{c42}	x_{b43}	x_{c43}	x_{b44}	x_{c44}	x_{b45}	x_{c45}	$2 \times m_4 = 2 \times 5 = 10$
Forecasted load	y_{b1}	y_{c1}	y_{b2}	y_{c2}	y_{b3}	y_{c3}	y_{b4}	y_{c4}	y_{b5}	y_{c5}	$2 \times m_t = 2 \times 5 = 10$
Total fuzzy system parameters encoding to run the PSO simulation											42

$x_{b11} = p_1 + p_2 + p_3 + \cdots + p_s$, where p is the individual particle in the swarm of population "s."

3.5. Fitness Function of the Optimization Problem. Based on the definition of fuzzy parameters in the previous expressions, the following fuzzy prediction model fitness measurement method given in (21) is used. The fitness function of the Mamdani fuzzy membership function is the mean absolute percentage error (MAPE) that measures the performance of the model and is evaluated as follows [27]:

$$\text{mape} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{y_i}. \tag{21}$$

The fitness function of the optimization problem is formulated as follows:

$$f_i\left(b_i, c_i\right) = \text{minimiz}\left(\text{mape}\right), \quad 0 \le b_i \le c_i. \tag{22}$$

Here y_i is the testing data set, $\hat{y}_i$ is the forecasted load data sets, and n is the forecasting time horizon.

4. Simulation Results

Both the training and testing data set values of the fuzzy input and output variables are presented on a 24-hour basis and it was discussed in Section 1. Therefore, the forecast is a short-term industrial load prediction of various events that exist in a calendar year with a 24-hour forecasting horizon.

4.1. Holiday Load Forecast. Holidays are special events that exist once a year in a calendar year, and such types of loads have distinctive characteristics such that they should be treated separately for the purpose of load prediction accuracy. These loads are special loads that exist during national holidays, Christian holidays, and Muslim holidays. The Christian holidays include Christmas, Siqilet, Easter, Epiphany, and Meskel. The Muslim holidays include Mewulid, Eidaldeha, and Eidalfetir, whereas in the national holidays, Adwa victory, New Year, Labor Day, and Patriots' Day are considered. The Fuzzy-PSO prediction of this special event provides a high level of prediction accuracy. The Fuzzy-PSO prediction results of all the holiday load scenarios are presented in Figures 12–23 and discussed in Section 5.

4.2. Weekday and Weekend Day Load Prediction. The weekday loads are the loads operated during normal periods where both national and religious holidays are omitted when they fall on Monday through Friday. The loads in this period should be treated separately because they have a different operational characteristic than holiday loads (Figure 24). On the other hand, the weekend day loads are the loads on Saturday and Sunday. Some private institutions consider Saturday as a working day, but still, the weekend has its own distinctive load characteristics, as shown in Figure 25.

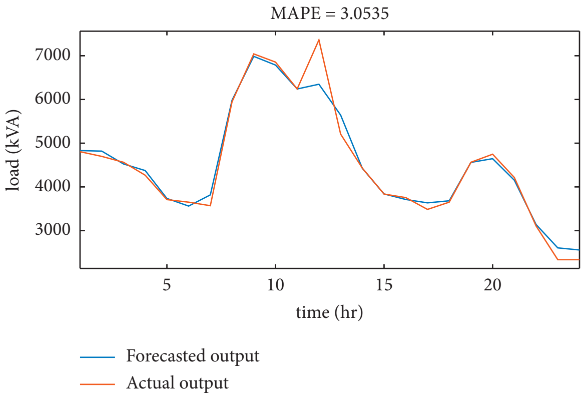


FIGURE 12: The Fuzzy-PSO prediction of Christmas load.

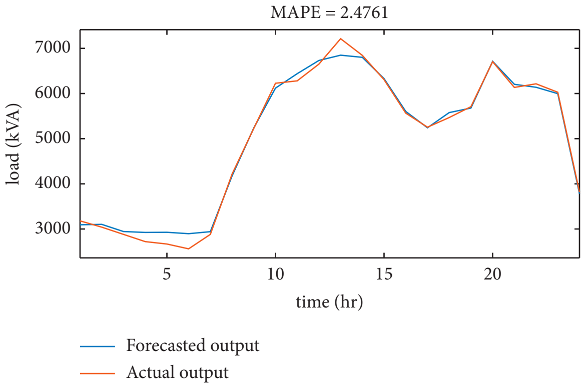


FIGURE 13: The Fuzzy-PSO prediction of Siqilet load.

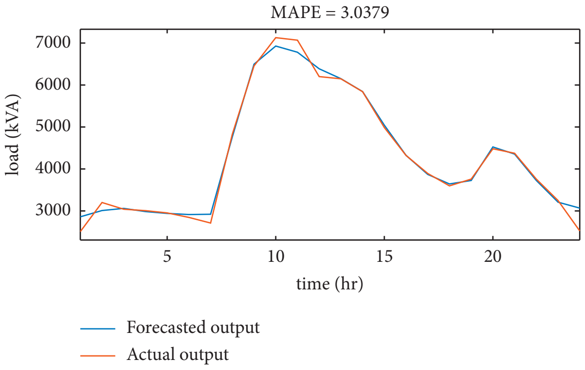


FIGURE 14: The Fuzzy-PSO prediction of Meskel load.

4.3. The Overall System Load Prediction. The total system load is the average load considering all the events mentioned earlier and it was scaled using a constant scaling factor to account its vulnerability due to the distinctive nature of various load events. The prediction result is used to model the microgrid system. The scaling factor is computed from the data sets and it is 1.3 for training data and 1.4 for testing data sets using equation (2).

The true correlation between the predicted total system load and the corresponding fuzzy input variables for the load

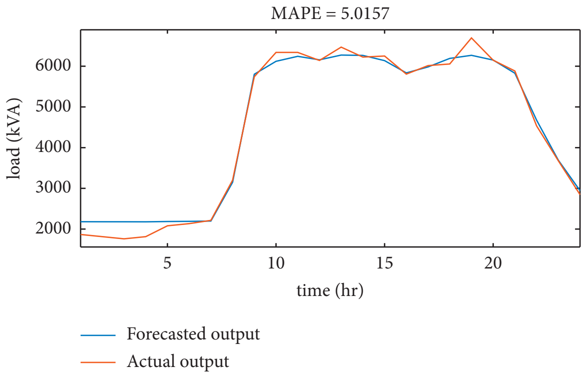


FIGURE 15: The Fuzzy-PSO prediction of Eidaldeha load.

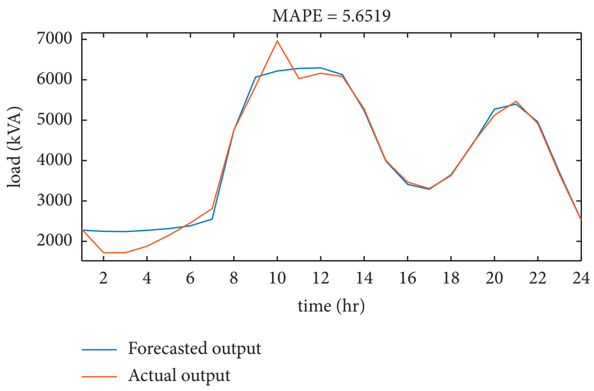


FIGURE 16: The Fuzzy-PSO prediction of Epiphany load.

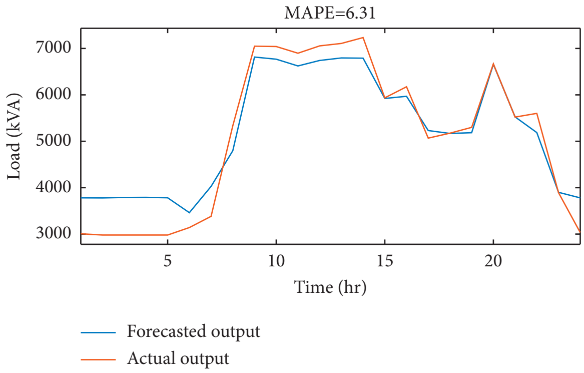


FIGURE 17: The Fuzzy-PSO prediction of Mewulid load.

prediction model based on the available data set is presented in Figures 26–29.

The optimized values of the fuzzy membership function and fuzzy rules are presented in Tables 2 to 6 for the total system load presented in Figure 30, and Table 7 presents the summary of the load prediction performance result.

5. Discussion

The computational cost of the Fuzzy-PSO load prediction model is measured based on the convergence rate of the

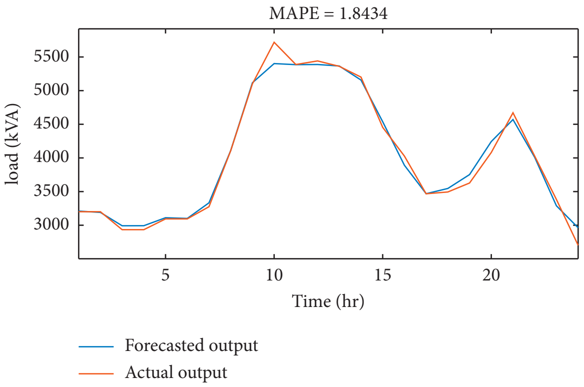


FIGURE 18: The Fuzzy-PSO prediction of Easter load.

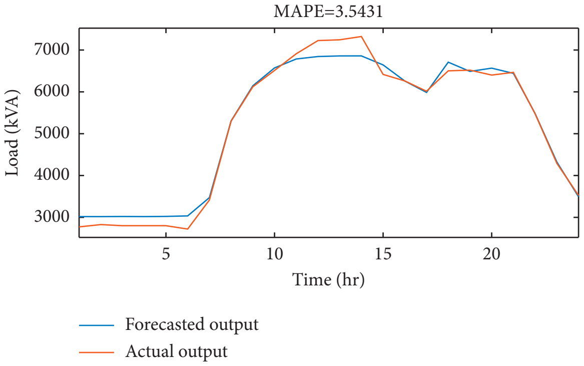


FIGURE 19: The Fuzzy-PSO prediction of Eidalfetir load.

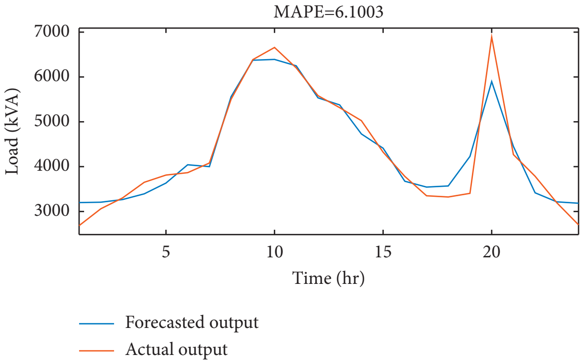


FIGURE 20: The Fuzzy-PSO prediction of New Year load.

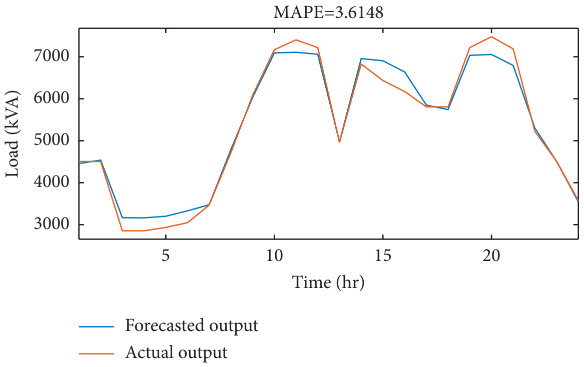


FIGURE 21: The Fuzzy-PSO prediction of Labor Day load.

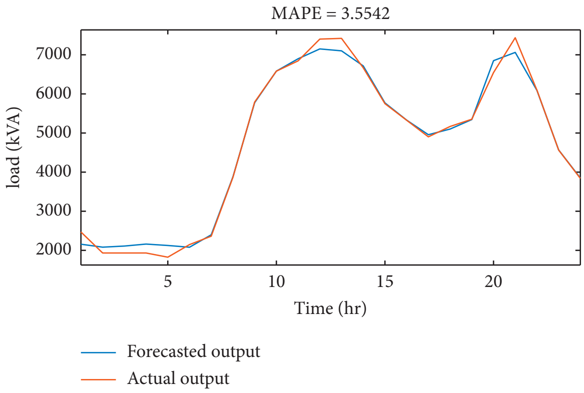


Figure 22: The Fuzzy-PSO prediction of Adwa load.

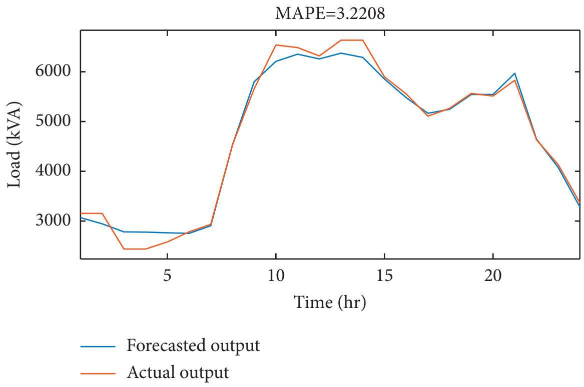


Figure 23: The Fuzzy-PSO prediction of Patriots' Day load.

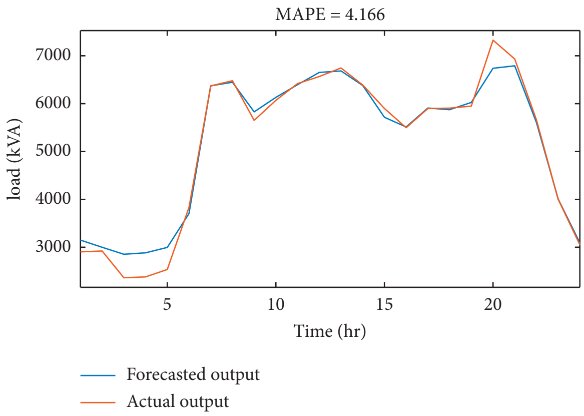


Figure 24: The Fuzzy-PSO prediction of weekday load.

algorithm, the computational time required to execute the prediction model, and the performance of the fitness function. The computational time depends on the number of iterations, the number of particles considered in the swarm, the number of fuzzy input variables, the number of fuzzy membership functions in each fuzzy input and output variable, the size of the input and output training dataset, and the performance of the hardware device (computer) used to simulate the model. For the Fuzzy-PSO load prediction model, 4 fuzzy input variables, 100 PSO particles, 300 iterations, and 17,520-hour training dataset values are

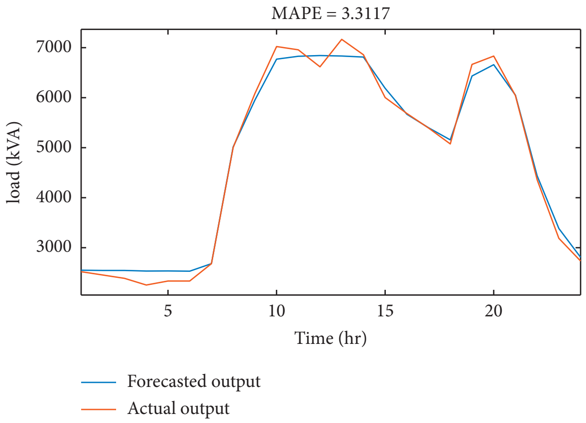


Figure 25: The Fuzzy-PSO prediction of weekend day load.

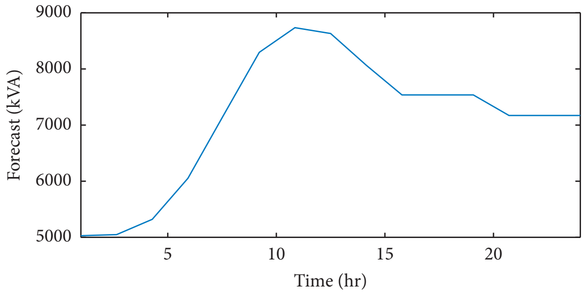


Figure 26: Correlation between forecast load and time for total system load.

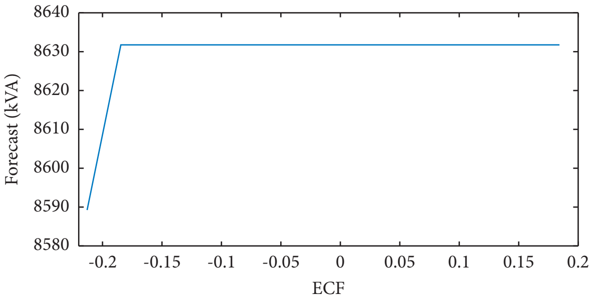


Figure 27: Correlation between forecast load and ECF for total system load.

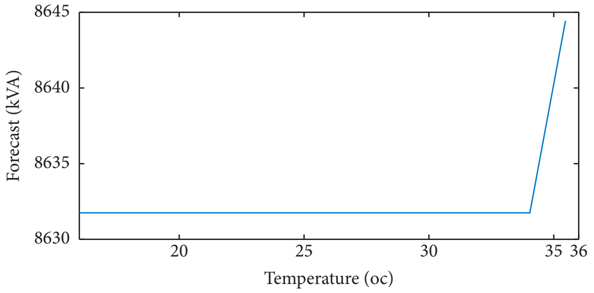


Figure 28: Correlation between forecast load and temperature for total system load.

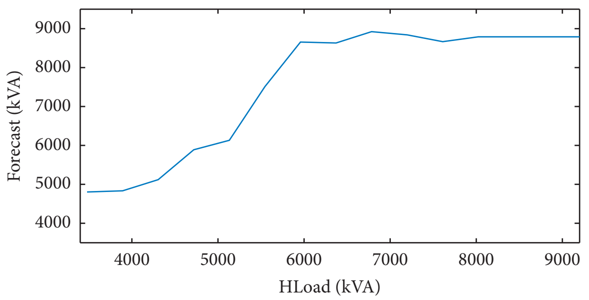


Figure 29: Correlation between forecast load and historical load for total system load.

Table 2: Fuzzy-PSO optimized result of temperature for microgrid model.

Membership function	Initial fuzzy parameters		PSO optimized parameters	
	b_i	c_i	b_i	c_i
Low	8.07	5.5	8.07	5.5
Normal	4.23	20	4.23	21.93
High	8.07	34.6	26.09	35.19

Table 3: Fuzzy-PSO optimized result of time for microgrid model.

Membership function	Fuzzy parameters		PSO optimized parameters	
	b_i	c_i	b_i	c_i
Night	2.803	1.7	2.396	1.628
Forenoon	2.803	8	2.302	9.115
Noon	2.803	12	2.081	17.591
Afternoon	2.803	16	1.640	16.176
Evening	2.803	22.3	2.396	20.995

Table 4: Fuzzy-PSO optimized result of ECF for microgrid model.

Membership function	Fuzzy parameters		PSO optimized parameters	
	b_i	c_i	b_i	c_i
Negative	0.6	−1	0.557	−1
Null	0.075	0	0.1134	0.0471
Positive	0.6	1	0.5062	1

Table 5: Fuzzy-PSO optimized result of historical load for microgrid model.

Membership function	Fuzzy parameters		PSO optimized parameters	
	b_i	c_i	b_i	c_i
Very low	270	2500	270	2503
Low	246	3636	609.9	5809
Average	246	4750	246	4930
High	246	5875	246	5697
Very high	270	7000	246	6875

simulated using an 8 GB RAM, core i7 Dell Latitude 7400 computer. The simulation takes several hours to simulate the Fuzzy-PSO load prediction model because of the complexity of the model, the large number of fuzzy membership functions, and the large number of iterations considered. The simulation converges to the solution very quickly after 15 iterations based on the calculation of the mean-absolute percentage error as a performance measurement index. The mean absolute percentage error is considered as a cost function to measure the performance of the Fuzzy-PSO load prediction model. The high level of prediction accuracy is recorded during the Easter holiday, and its industrial load prediction result is presented in Figure 18. The Fuzzy-PSO prediction model during Easter forecasts the load with a prediction accuracy of 1.84% (MAPE), which is significantly very high as compared with the fuzzy prediction model result alone of 3.54%. The Fuzzy-PSO load prediction model for the weekday yields a MAPE of 4.17%, which is a very high accuracy level compared to the fuzzy-alone load prediction model of 13.89% MAPE. The Fuzzy-PSO load prediction model for the weekend load has a prediction accuracy of 3.3% MAPE, whereas the overall industrial load Fuzzy-PSO load prediction result has a prediction accuracy of 3.62% MAPE. In general, the Fuzzy-PSO load prediction model has a fast convergence rate and very high prediction accuracy level and needs a supercomputer to improve the computational time of the prediction algorithm.

All the fuzzy inputs and fuzzy output membership functions are encoded using all 100 PSO particles. Each membership function is represented using two basic fuzzy parameters: the standard deviation (b) and mean (c), so the values of b and c are optimized using the training input and output dataset correlation. Similarly, the initial fuzzy rules are updated and optimized using the data correlation between the training input and output dataset. The Fuzzy-PSO algorithm executes the correlation between the fuzzy input variables (time, ECF, temperature, and historical load) and the fuzzy output variable (forecasted load) based on the training data set and training performance of the PSO. The correlation between time and forecasted load is presented in Figure 26, and from the correlation graph, we can understand that the time of the day can affect the load forecast result. During the early morning and after-night periods, the industrial load is low, while it is at its maximum during the daytime between 9 am and 15 am. The correlation between forecasted load and ECF demonstrates that at an ECF margin of ±20%, the change in the forecasted load is negligible, but at lower (negative big) and higher (positive big) values of the ECF, it has a great impact on the forecasted load. The incremental or decremental magnitude of the forecasted load depends on the sign and magnitude of the ECF values. The impact of temperature and historical load on the forecast of industrial load is very clear. They have a direct relationship with the forecasted load based on the characteristics of the temperature and historical load in the training dataset. Table 7 demonstrates the forecasting accuracy of the Fuzzy and Fuzzy-PSO load prediction models

Table 6: Fuzzy-PSO optimized result of overall forecasted load for microgrid model.

Membership function	Fuzzy parameters		PSO optimized parameters		Initial possible rules	PSO generated rules	Optimized rules
	b_i	c_i	b_i	c_i			
Very low	400	2000	400	20162	225	129	115
Low	400	4000	400	3930			
Average	400	6000	419	5759			
High	400	8000	400	9263			
Very high	400	10000	400	10000			

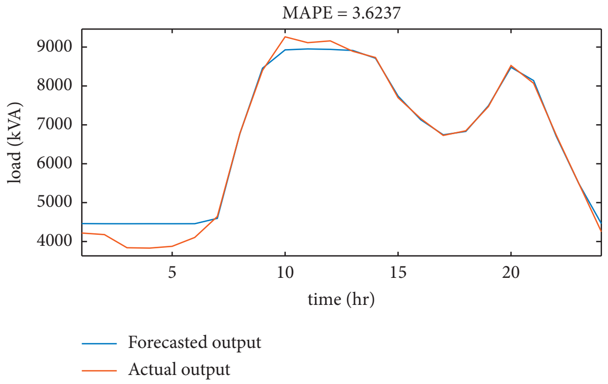


Figure 30: The Fuzzy-PSO prediction of the total system load for microgrid model.

Table 7: Performance evaluation of fuzzy-alone and Fuzzy-PSO load prediction model based on MAPE.

No.	Event	Methods	
		Fuzzy	Fuzzy-PSO
1	Easter	3.54	1.84
2	Epiphany	7.18	5.65
3	Siqilet	9.05	2.48
4	Christmas	11.59	3.05
5	Meskel	10.56	3.04
6	Eidalfetir	5.78	3.54
7	Mewulid	9.34	6.31
8	Eidaldeha	6.61	5.02
9	New Year	9.56	6.10
10	Adwa victory	11.09	3.55
11	Patriots' day	9.34	3.22
12	May_20	5.37	3.45
13	Labor Day	8.65	3.61
14	Weekday	13.89	4.17
15	Weekend day	13.9	3.31
16	System load	8.85	3.62

in terms of the mean absolute percentage error of the forecasted dataset and the validation dataset measured in percentage (%).

6. Conclusion

A microgrid is a viable option to supply reliable, efficient, and affordable power to an industrial customer. This load prediction work is ongoing research that forms an integral part of microgrid design. The particle swarm optimization algorithm is an intelligent technique implemented to optimize the fuzzy load prediction approach. The model provides an accurate load prediction result whose absolute-mean percentage error is less than 5% in most of the scenarios that were discussed in this paper. The Fuzzy-PSO load prediction model manipulates large volumes of fuzzy input and output data for both the training and testing datasets in order to generate fuzzy rules and fuzzy membership functions used for the load prediction model. The termination criteria of the PSO algorithm matter the accuracy level of the load prediction model. The number of iterations or the acceptable tolerance of error considered in the Fuzzy-PSO load prediction model also has a significant impact on the simulation result. The load prediction model during the New Year and Mewulid holidays is simulated based on a maximum fuzzy rule generation iteration of 50 and an overall fuzzy system optimization maximum iteration number of 250, which yields a very low level of prediction accuracy as compared with the other scenarios. Introducing a new fuzzy input variable (i.e., error correction factor) also improves the load prediction model based on the Fuzzy-PSO prediction approach.

Abbreviations

ANN:	Artificial neural network
E.C:	Ethiopian calendar
ECF:	Error correction factor
Eq:	Equation
Fig:	Figure
G.C:	Gregorian calendar
LS:	Least squares
LTLF:	Long-term load forecast
MAPE:	Mean absolute percentage error
MTLF:	Medium term load forecast
NRMS:	Normalized root-mean-square error
PSO:	Particle swarm optimization
STLF:	Short-term load forecast
SVM:	Support vector machine
VSTLF:	Very-short-term load forecast.

References

[1] E. A. Feinberg and D. Genethliou, *Applied Mathematics for Restructured Electric Power Systems: Load Forecasting*, Springer Link, Berlin, Germnay, 2005.

[2] S. J. Kiartzis, A. G. Bakirtzis, J. B. Theocharis, and G. Tsagas, “A fuzzy expert system for peak load forecasting. Application to the Greek power system,” in *Proceedings of the 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings. MeleCon 2000 (Cat. No.00CH37099)*, vol. 3, pp. 1097–1100, Lemesos, Cyprus, May 2000.

[3] A. A. Sallam and O. P. Malik, *Electric Distribution System*, Load Forecasting, Beijing, China, 2nd edition, 2019.

[4] K. Goswami, A. Ganguly, and A. K. Sil, “Day ahead forecasting and peak load management using multivariate auto regression technique,” in *Proceedings of the 2018 IEEE Applied Signal Processing Conference (ASPCON)*, pp. 279–282, Kolkata, India, December 2018.

[5] P. R. J. Campbell and K. Adamson, “Methodologies for load forecasting,” in *Proceedings of the 2006 3rd International IEEE Conference Intelligent Systems*, pp. 800–806, London, UK, September 2006.

[6] M. Lekshmi and K. N. A. Subramanya, “Short-term load forecasting of 400kV grid substation using R-tool and study of influence of ambient temperature on the forecasted load,” in *Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pp. 1–5, Gangtok, India, February 2019.

[7] K. Zhang, X. Feng, X. Tian, Z. Hu, and N. Guo, “Partial Least Squares regression load forecasting model based on the combination of grey Verhulst and equal-dimension and new-information model,” in *Proceedings of the 2020 7th International Forum on Electrical Engineering and Automation, IFEEA*, pp. 915–919, Hefei, China, September 2020.

[8] A. D. Papalexopoulos and T. C. Hesterberg, “A regression-based approach to short-term system load forecasting,” *IEEE Transactions on Power Systems*, vol. 5, no. 4, pp. 1535–1547, 1990.

[9] T. Haida and S. Muto, “Regression based peak load forecasting using a transformation technique,” *IEEE Transactions on Power Systems*, vol. 9, no. 4, pp. 1788–1794, 1994.

[10] W. Charytoniuk, M. S. Chen, and P. Van Olinda, “Non-parametric regression based short-term load forecasting,” *IEEE Transactions on Power Systems*, vol. 13, no. 3, pp. 725–730, 1998.

[11] L. Hu, L. Zhang, T. Wang, and K. Li, “Short-term load forecasting based on support vector regression considering cooling load in summer,” in *Proceedings of the 2020 Chinese Control and Decision Conference (CCDC)*, pp. 5495–5498, Hefei, China, August 2020.

[12] M. Kumar Singla, S. Hans, M. Kumar, and S. S. Hans, “Load forecasting using fuzzy logic tool box,” *Solar Energy*, vol. 3, no. 8, 2018, https://www.researchgate.net/publication/331114596 [Online]. Available:.

[13] P. Gohil and M. Gupta, “Short term load forecasting using fuzzy logic,” in *Proceedings of the National Conference on Recent Trends In Electrical and Electronics & Communication Engineering (RTEECE-2014)*, pp. 127–130, Vadodara, India, January 2014.

[14] K. H. Kwang-Ho Kim, H.-S. Hyoung-Sun Youn, and Y.-C. Yong-Cheol Kang, “Short-term load forecasting for special days in anomalous load conditions using neural networks and fuzzy inference method,” *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 559–565, 2000.

[15] H. H. Cevick and M. Cuncas, “A fuzzy logic based short term load forecast for the holidays,” *International Journal of Machine Learning and Computing*, vol. 6, no. 1, pp. 57–61, 2016.

[16] R. Gao, L. Liyuan Zhang, and X. Liu, “Short-term load forecasting based on least square support vector machine combined with fuzzy control,” in *Proceedings of the 10th World Congress on Intelligent Control and Automation*, pp. 1048–1051, Beijing, China, July 2012.

[17] Z.-b. Shi, Y. Li, and T. Yu, "Short-term load forecasting based on LS-SVM optimized by bacterial colony chemotaxis algorithm," in *Proceedings of the 2009 International Conference on Information and Multimedia Technology, ICIMT*, pp. 306–309, Jeju, Korea (South), December 2009.

[18] S. Singh, S. Hussain, and M. A. Bazaz, "Short term load forecasting using artificial neural network," in *Proceedings of the 2017 Fourth International Conference on Image Information Processing (ICIIP)*, pp. 1–5, Shimla, India, December 2017.

[19] N. Gautam, A. S. Mayal, V. S. Ram, and A. Priya, "Short term load forecasting of urban loads based on artificial neural network," in *Proceedings of the 2019 2nd International Conference on Power and Embedded Drive Control (ICPEDC)*, pp. 46–51, Chennai, India, August 2019.

[20] X. Yang, J. Yuan, J. Yuan, and H. Mao, "An improved WM method based on PSO for electric load forecasting," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8036–8041, 2010.

[21] C. O. Adika and L. Wang, "Short term energy consumption prediction using bio-inspired fuzzy systems," in *Proceedings of the 2012 North American Power Symposium (NAPS)*, pp. 1–6, Champaign, IL, USA, September 2012.

[22] P. Mukhopadhyay, G. Mitra, S. Banerjee, and G. Mukherjee, "Electricity load forecasting using fuzzy logic: short term load forecasting factoring weather parameter," in *Proceedings of the 2017 7th International Conference on Power Systems (ICPS)*, pp. 812–819, Pune, India, December 2017.

[23] X.-S. Yang and M. Karamanoglu, "Swarm intelligence and bio-inspired computation," p. 420, Elsevier, London, UK, 2013.

[24] X. Zhou, S. Yang, and S. Sun, "A Deep Learning model for day-ahead load forecasting taking advantage of expert knowledge," in *Proceedings of the 2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*, pp. 1–5, Wuhan, China, May 2021.

[25] B.-S. Kwon, R.-J. Park, and K.-B. Song, "Weekly peak load forecasting for 104 Weeks using deep learning algorithm," in *Proceedings of the 2019 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pp. 1–4, Macao, China, December 2019.

[26] X. Liao, X. Kang, M. Li, and N. Cao, "Short term load forecasting and early warning of charging station based on PSO-SVM," in *Proceedings of the 2019 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS*, pp. 305–308, March. 2019.

[27] D. M. Teferra, L. M. H. Ngoo, and G. N. Nyakoe, "A fuzzy based prediction of an industrial load in microgrid system using particle swarm optimization algorithm," in *Proceedings of the 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6, Mauritius, Mauritius, October 2021.

[28] G. C. Mouzouris and J. M. Mendel, "Nonsingleton fuzzy logic systems: theory and application," *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 1, pp. 56–71, 1997.

[29] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948, Perth, Australia, December 1995.

[30] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proceedings of the 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, pp. 69–73, Anchorage, AK, USA, May 1998.

[31] Y. Zhang, T. Li, G. Na, G. Li, and Y. Li, "Optimized extreme learning machine for power system transient stability prediction using s," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–8, 2015.

[32] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 240–255, 2004.

Optimization of a Two-Layer 3D Coil Structure with Uniform Magnetic Field

Ajanta Priyadarshinee, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ajantapriyadarshinee10@gmail.com*

Rajib Lochan Barik, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, rajib_barik543@gmail.com*

Subhendu Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srikant.p@yahoo.co.in*

Anil Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, anil_sahoo342@gmail.com*

Abstract

Conventional magnetically coupled resonant wireless power transfer systems are faced with resonant frequency splitting phenomena and impedance mismatch when a receiving coil is placed at misaligned position. These problems can be avoided by using uniform magnetic field distribution at receiving plane. In this paper, a novel 3D transmitting coil structure with improved uniform magnetic field distribution is proposed based on a developed optimization method. The goal is to maximize the average magnetic field strength and uniform magnetic field section of the receiving plane. Hence, figures of merit (FoM_1 and FoM_2) are introduced and defined as product of average magnetic field strength and length or surface along which uniform magnetic field is generated, respectively. The validity of the optimization method is verified through laboratory measurements performed on the fabricated coils driven by signal generator at operating frequency of 150 kHz. Depending on the allowed ripple value and predefined coil proportions, the proposed transmitting coil structure gives the uniform magnetic field distribution across 50% to 90% of the receiving plane.

1. Introduction

Numerous wireless power transfer (WPT) systems operate through nonuniform magnetic field strength distribution at receiving plane. Magnetic field nonuniformity in magnetically coupled resonant (MCR) WPT system causes resonant frequency splitting phenomena and impedance mismatching when receiving coil (RX-coil) is not properly aligned with transmitting coil (TX-coil) [1, 2]. Namely, the efficiency of WPT system is significantly impaired due to frequency splitting phenomena or impedance mismatching [1, 3]. Consequently, to achieve efficient energy transfer along with greater degree of freedom, in terms of RX-coil position, such WPT systems require frequency tracking [3] and automatic impedance matching [3, 4]. In order to create uniform magnetic field strength distribution at a receiving plane, various TX-coil designs are developed [1, 5–21].

In this paper, a novel 3D structure of TX-coil is proposed to achieve uniform magnetic field distribution at the given receiving plane (114 × 28 cm). Presence of uniform magnetic field strength distribution at the receiving plane will provide stable resonant frequency regardless of RX-coil position due to steady value of magnetic coupling factor k between TX- and RX-coil. Hence, uniform magnetic field strength distribution at a receiving plane results in a WPT system that does not require frequency tracking and automatic impedance matching [8]. Uniform magnetic field strength distribution also provides uniform and simultaneous power delivery to multiple RX-coils, i.e., loads. Furthermore, uniform magnetic field strength distribution at a receiving plane ensures nondegraded wireless power transfer efficiency regardless of RX-coil position within receiving plane [1].

The proposed TX-coil structure consists of two layers and is based on optimization method which is validated by simulation and measurement results. Since the proposed TX-coil structure is characterized by folded sides, it is more suitable for large sized WPT systems, for instance, installation under office desk. Computer simulations of TX-coils with different winding arrangement (coil layers relative spacing), but with the same outer dimensions (114 × 28 cm), were run, and magnetic field distribution at the given receiving plane which is above the TX-coil was observed. According to the simulation results, experimental TX-coils are made out of Litz wire instead of PCB realization which is not appropriate for large receiving plane. Both computer simulation and measurement results verified that uniform magnetic field distribution at considerable surface of the receiving plane is produced by the novel 3D TX-coil structure. With the approximately 100 W of output power at receiving side of a WPT system, application of such TX-coil enables simultaneous wireless charging of monitors, smartphones, laptops, etc.

2. Model Explanation

To maintain stable transfer efficiency and power delivery to freely moving RX-coil(s), TX-coil in WPT system should generate uniform magnetic field [18, 21]. Such field distribution is difficult to achieve with "standard" flat wound coils. When designing a 3D coil structure, a fast method for magnetic field evaluation is a significant advantage. A common approach for magnetic field evaluation is using dedicated software (e.g., Ansys Maxwell, FEMM, MATLAB, etc.) which executes a numerical analysis. Such approach is very time consuming, both in preparation (model preprocessing) and in simulation itself.

If the 3D TX-coil structure can be represented as a number of linear sections, the magnetic field simulations can be significantly simplified. Uniformity of magnetic field of the TX-coil is evaluated with respect to the plane of interest. In MCR-WPT, where multiple loads can be wirelessly powered, a flat surface (e.g., office desk) should have uniform magnetic field distribution. Therefore, magnetic field evaluations are done with respect to this plane of interest, i.e., referent plane. Under these conditions (coil represented by linear sections and a defined referent plane), a fast magnetic field simulation model can be developed. Since a coil is defined as a piecewise linear structure, a magnetic field in any given point can be calculated as a vector sum of magnetic fields generated by each linear segment of 3D coil structure. Each linear segment is defined by two points in 3D space, $P1$ and $P2$, as shown in Figure 1. When a current flows from $P1$ to $P2$, the linear segment $P1$-$P2$ generates a magnetic field. Figure 1 shows magnetic field generated by one linear segment ($P1$-$P2$) in a single point ($P0$) on referent plane. Magnitude of the magnetic field strength at point $P0$ can be calculated using (1).

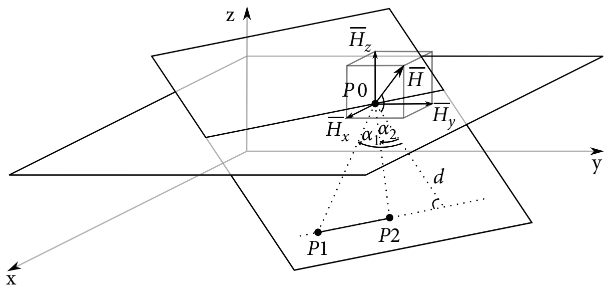


FIGURE 1: Magnetic field in the point $P0$ generated by the linear segment $P1$-$P2$.

$$|\overline{H}| = \frac{I}{4\pi d}\left(\sin\alpha_1 - \sin\alpha_2\right), \tag{1}$$

The shortest distance between $P0$ and the line which passes through $P1$ and $P2$ equals

$$d = \frac{|\overline{P1P2} \times \overline{P0P1}|}{|\overline{P1P2}|}. \tag{2}$$

The angles $\alpha 1$ and $\alpha 2$ are calculated as follows:

$$\begin{aligned}\alpha_1 &= 90^\circ - \arccos\left(\frac{P1P2 \cdot P0P1}{|P1P2| \cdot |P0P1|}\right),\\ \alpha_2 &= 90^\circ - \arccos\left(\frac{P1P2 \cdot P0P2}{|P1P2| \cdot |P0P2|}\right).\end{aligned} \tag{3}$$

For coil current I in linear segment flowing from P1 to P2, the direction of the magnetic field strength $\overline{H}$ at point $P0$ is perpendicular to the plane defined by points $P0$, $P1$, and $P2$ (Figure 1). Equation (1) gives only the magnitude of the magnetic field strength, but the direction of magnetic field strength can be calculated by

$$\widehat{H} = \frac{\overline{P1P2} \times \overline{P0P1}}{|\overline{P1P2} \times \overline{P0P1}|}. \tag{4}$$

With respect to the referent plane, only the part of the magnetic field directed in the z-axis ($\overline{H}z$ in Figure 1) is of interest. The proposed magnetic field modelling method is implemented in MATLAB. Coil is defined as piecewise linear structure using the array, where each row in array defines one segment of the coil. The first three values define the position of point $P1$, the next three values define the position of point $P2$, and the last value defines the current from $P1$ to $P2$.

$$\text{coil} = \begin{array}{lllllll} [P1x & P1y & P1z & P2x & P2y & P2z & I12 \\ P2x & P2y & P2z & P3x & P3y & P3z & I23 \\ \vdots & & & & & & \\ P(n-1)x & P(n-1)y & P(n-1)z & Pnx & Pny & Pnz & I(n-1)n] \end{array}. \tag{5}$$

For a coil structure defined as an array of linear current segments, a magnetic field in each point of a referent plane can be calculated. Figure 2 shows the simulated magnetic field for a single-layer rectangular coil (114 × 28 cm) at referent plane placed 30 mm above the coil.

The magnetic field has a bowl-like shape with pronounced spikes at the coil corners. For application in position tolerant (in terms of RX-coil) WPT system, such magnetic field shape is not suitable. The rectangular shape of the coil is one of the least favorable candidates for uniform magnetic field (with triangular coil shape being the worst), but it is the shape widely used in WPT systems. This is the main motivation for the development of the rectangular 3D coil structures which can generate uniform magnetic field.

In this paper, we focused on a two-layer 3D coil structure, due to simpler fabrication compared to a multilayer coil with respect to required precision of manufacture.

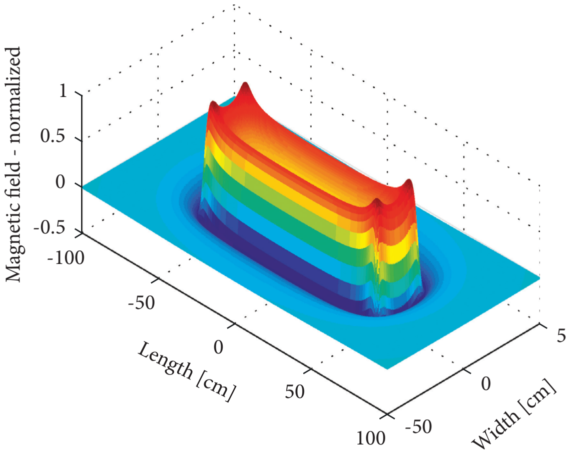


Figure 2: Simulated magnetic field distribution generated by single-layer rectangular coil (114 × 28 cm).

2.1. Optimization of Coil Structure. Figure 3 shows the generalized structure of two-layer 3D coil. Coil layers can be distinguished by their color; the first layer is shown in black and the second layer in red. Second coil layer is placed at a depth D_2 relative to the first coil layer, and it has narrower width W_2 compared to width of the first coil layer W. 3D coil structure has a length L with both coil layers folded to a depth D at the coil ends. Grey surface represents the referent plane placed at the distance h from the coil.

Coil structure optimization is a two-step procedure. In step one, the goal is to ensure a uniform field distribution across a width of the coil, i.e., over cross-section of the referent plane named cs_1 in Figure 3. In step two, the goal is to ensure a uniform field distribution across a length of the coil, i.e., over cross-section of the referent plane named cs_2 in Figure 3.

Figure 4 shows the proposed optimization framework. The maximal allowed ripple value of magnetic field (r) and first coil layer dimensions (width W and length L) are considered as inputs. The optimization outputs are optimal transfer distance (h), second coil layer variables (D_2, W_2), and end fold depth D, for which the magnetic field strength and uniform surface size are maximized.

Extensive analysis is conducted, resulting in mathematical model for rectangular coil optimization. Conducted analysis and mathematical model are explained in the next two sections.

2.2. Optimization of Second Coil Layer Variables. To ensure a uniform magnetic field distribution over cross-section cs_1 (width of the coil), a position of the second coil layer must be optimized. Perfectly uniform magnetic field distribution cannot be achieved, so one parameter that must be taken into account is the maximal ripple of the magnetic field. The second parameter is the distance h of the referent plane from the coil. For different distances h, different values of the second coil layer variables (W_2 and D_2) obtain "most" uniform field distribution. Similarly, for different allowed ripple values of a magnetic field, different values of the second coil layer variables (W_2 and D_2) result in different shortest distance of the referent plane h.

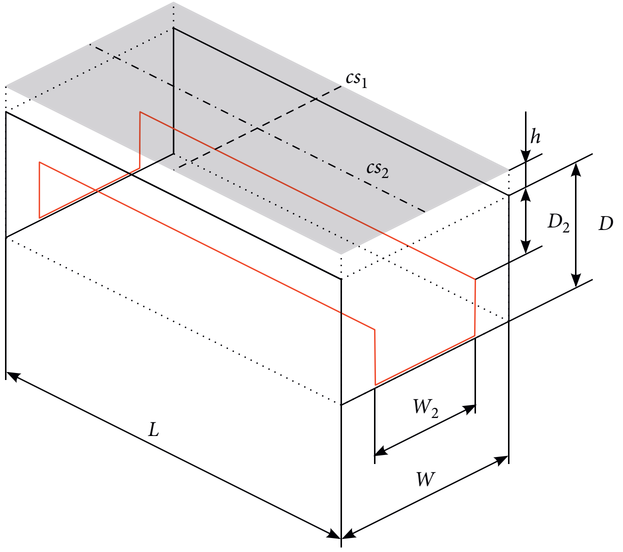


Figure 3: Two-layer 3D coil structure.

Optimization deals with the following problem: finding values of the second coil layer variables to generate the most uniform field possible at a shortest distance between TX-coil and a referent plane. However, such optimization problem misses some important aspects. Namely, the short distance generally results in high magnetic field strength. Certain values of the second coil layer variables that ensure uniform field at short distances do that at the cost of a lower magnetic field strength. This would decrease the overall performance of WPT system. The second important aspect is the percentage of the cross-section over which the uniform field is achieved. It is quite easy to get uniform field at short distance with high field strength, but only over a small fraction of total coil width.

Accordingly, the first part of the optimization process is finding values of the second coil layer variables to get

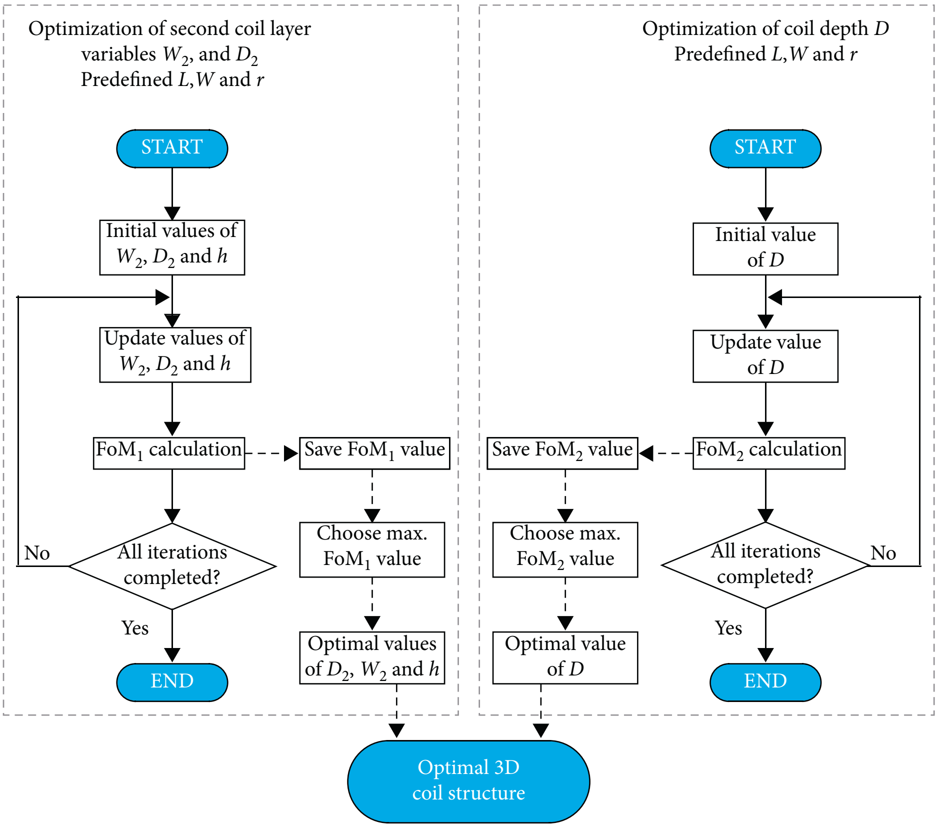

Optimization of second coil layer variables W_2, and D_2
Predefined L, W and r
START
Initial values of W_2, D_2 and h
Update values of W_2, D_2 and h
FoM_1 calculation
Save FoM_1 value
Choose max. FoM_1 value
All iterations completed?
No
Yes
END
Optimal values of D_2, W_2 and h
Optimization of coil depth D
Predefined L, W and r
START
Initial value of D
Update value of D
FoM_2 calculation
Save FoM_2 value
Choose max. FoM_2 value
All iterations completed?
No
Yes
END
Optimal value of D
Optimal 3D coil structure

Figure 4: Optimization framework.

uniform field on most of the cross-section, but also with a highest magnetic field strength value. Therefore, as a figure of merit (FoM_1), we propose the product of average magnetic flux density (calculated only for part of the cross-section where magnetic field strength is uniform with respect to defined ripple value) and percentage of cross-section on which the uniform field is achieved.

The first step in second coil variables' optimization is the FoM_1 analysis. This is a 4-dimensional problem. FoM_1 is affected by the referent plane distance h (first dimension), maximal allowed ripple of the uniform field (second dimension), and variables (W_2, D_2) of second coil layer (third and fourth dimension, respectively). Since the referent plane distance h must be observed with respect to coil width W, their ratio (h/W) will be used in FoM_1 analysis. We assume the same current magnitude and direction in both coil layers.

The four parameters are varied in the following ranges: h/W ratio ranges from 1% to 20% (20 steps), ripple value ranges from 1% to 20% (6 steps), width of the second coil layer W_2 ranges from 0 to the coil width W (50 steps), and depth of the second coil layer D_2 ranges from 0 to half of the coil width $W/2$ (50 steps).

The following methodology was used: for fixed values of h/W ratio and ripple, magnetic field strength was calculated for each possible value of the second coil layer variables. Namely, magnetic field strength for each value of D_2 and W_2 was calculated over the cs_1 cross-section of the coil, and magnetic field shape is analyzed. In this step of the coil optimization, the coil length L is set to be at least 10 times larger than the coil width W, and the ends of the coil are not folded down to the depth D. Figure 5 shows results of the analysis of a magnetic field shape for four different positions of second coil layer. Position of the second coil layer is completely determined by the values of the second coil layer variables, D_2 and W_2.

Ripple of the magnetic field is used as input parameter in the optimization process. It is used to determine the size of the uniform section of magnetic field. Uniform section is defined as a section of the coil where the magnetic field deviation does not exceed the defined ripple value. The analysis of the magnetic field begins at the center of the cross-section and moves to the sides. The magnetic field strength at the center is used as an initial average field strength value. The two adjacent magnetic field strengths are compared to the initial value, and if they do not deviate from initial value by more than the defined ripple, they are considered to be in the uniform section of the magnetic field. New average field strength of uniform section is then calculated. The next two adjacent field values are then evaluated using the same methodology. At one point, the field values that deviate by more than the defined ripple value will be reached. These field values and all the remaining ones are not a part of the uniform magnetic field section. This can be seen in Figures 5(a) and 5(b). The section of the magnetic field shape that is considered uniform (with defined ripple) corresponds to the top section of the square waveform (Figures 5(a) and 5(b)).

In Figures 5(c) and 5(d), different shapes of the magnetic field that correspond to different positions of the second coil

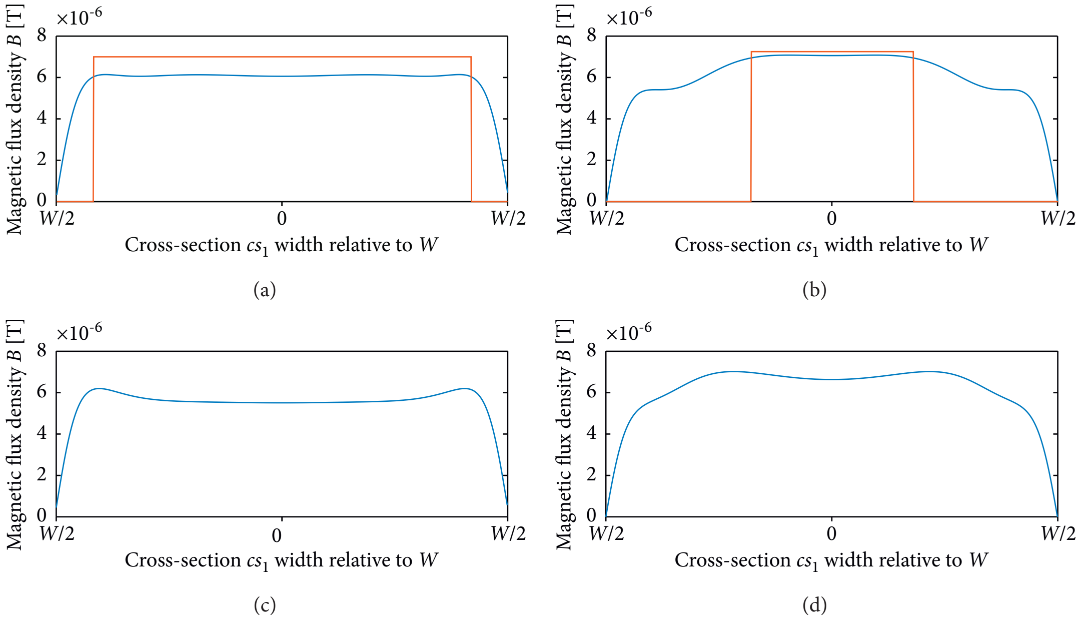


FIGURE 5: Examples of magnetic field shape analysis: (a) ripple = 1%, $h/W = 0.08$, $D_2/W = 0.18$, $W_2/W = 0.7$; (b) ripple = 1%, $h/W = 0.08$, $D_2/W = 0.12$, $W_2/W = 0.6$; (c) ripple = 1%, $h/W = 0.08$, $D_2/W = 0.24$, $W_2/W = 0.7$; (d) ripple = 1%, $h/W = 0.08$, $D_2/W = 0.1$, $W_2/W = 0.7$.

layer are shown. Such magnetic field shapes are not considered uniform because of the two peaks at the sides. While the center part of the magnetic field has uniform field distribution (Figure 5(c)), the overall magnetic field shape is considered nonuniform. The parts of the coil cross-section where the magnetic field strength exceeds the maximal allowed field strength (average strength + ripple value) can potentially be harmful for receiver circuits in WPT system, which are designed for uniform magnetic field. For that reason, Figures 5(c) and 5(d) have no square waveform representing the uniform section of the magnetic field.

To summarize, there are two possible outcomes of magnetic field shape analysis. The uniform section of the magnetic field is identified, or the magnetic field is considered nonuniform. For magnetic field shapes that have a uniform section, the FoM_1 is calculated:

$$\mathrm{FoM}_1 = B_{\mathrm{avg}} \cdot \frac{\text{uniform section width}}{\text{coil width}}, \tag{6}$$

where B_{avg} is an average value of magnetic flux density calculated only for the uniform section of magnetic field.

The same methodology is used to evaluate the magnetic field shapes for each possible position of the second coil layer. As a result, we get FoM_1 values for each position of the second coil layer. Figure 6 shows the FoM_1 values for five different h/W ratios with fixed ripple value.

For low h/W ratio value, majority of positions of second coil layer result in nonuniform field (dark blue areas). With higher h/W ratio values, more positions of second coil layer generate uniform magnetic field. For each fixed combination of h/W ratio and ripple value, there is an optimal position of the second coil layer which results in maximal FoM_1 value. This maximal FoM_1 value is represented as one dot in Figure 7. Figure 7 shows maximal FoM_1 values for all evaluated combinations of h/W ratio and ripple.

It can be seen that, for a higher ripple value, a higher FoM_1 can be achieved. For each ripple value, the maximal FoM_1 value is achieved at different h/W ratio (different distance between the referent plane and the coil). For higher values of h/W ratio, large uniform section can be easily achieved, but the average magnetic field strength is lower. For lower h/W ratio values, we have higher average magnetic field values, but narrower uniform sections. The maximal FoM_1 is achieved at optimal distance from the coil (optimal h/W ratio) with large uniform section and high average magnetic field strength. These optimal h/W ratios are shown with blue markers in Figure 8.

For a given ripple value and known coil width W, optimal distance h of the referent plane can be calculated as

$$h = W\,[0.0191 - 0.013 \cdot \ln(r)]. \tag{7}$$

Once the distance of the referent plane is selected, the position of second coil layer can be determined. Figure 9 shows the positions of second coil layer that achieve maximal FoM_1, for different ripple values with referent plane at optimal distance.

For the lowest evaluated ripple values (0.5%–2%), the optimal width of the second coil layer, W_2, is 70% of the first coil layer width, W. For larger ripple values, the second coil layer width increases. The depth of second coil layer, D_2, shows direct correlation with the ripple value (Figure 10).

For a given ripple value and known coil width W, with referent plane placed at optimal distance, the depth of second coil layer can be calculated as

$$D_2 = W\,[0.041 - 0.03 \cdot \ln(r)]. \tag{8}$$

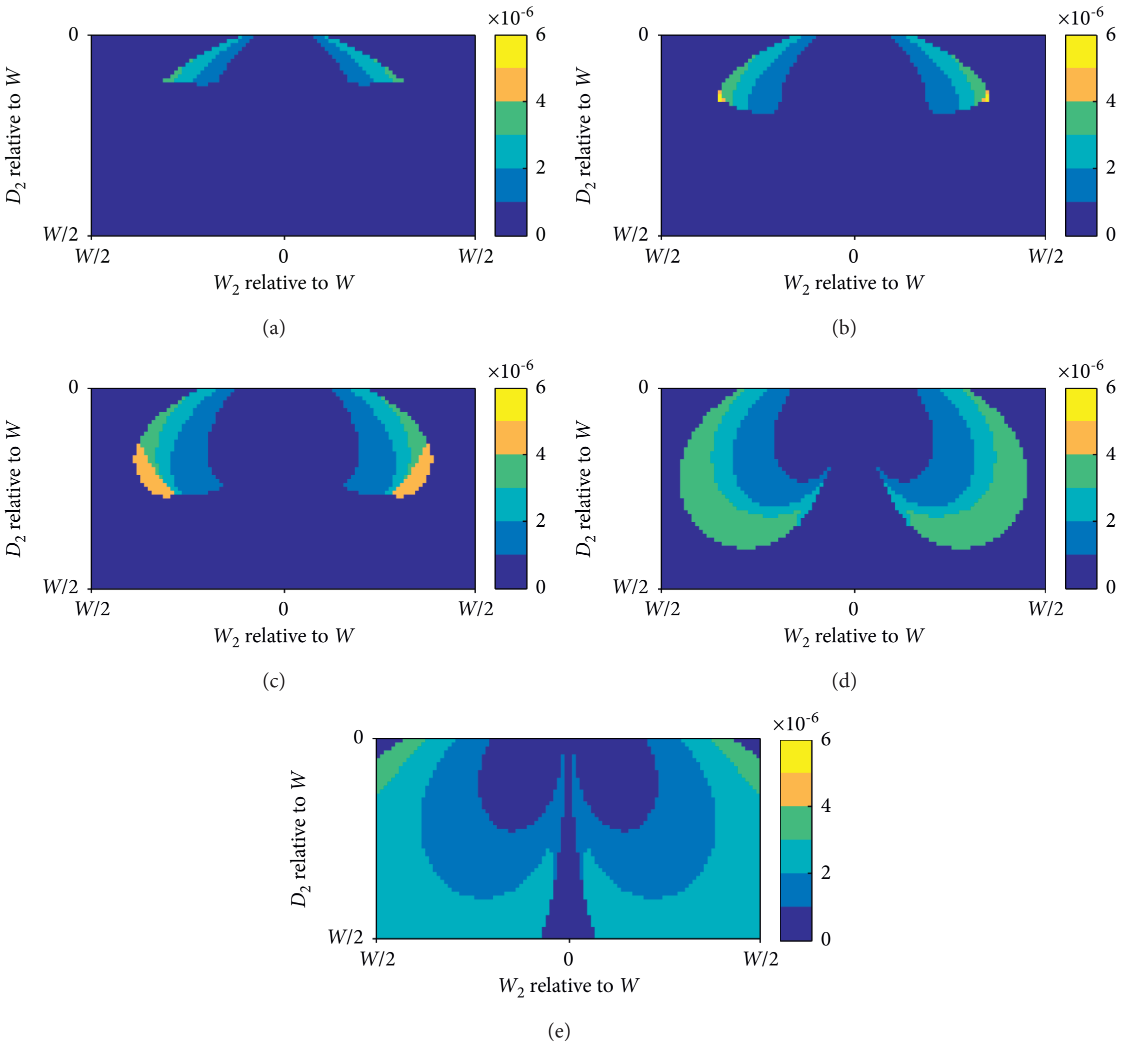


Figure 6: FoM_1 values for different combinations of h/W ratio and ripple value of 2%: (a) ripple = 2%, $h/W = 0.05$; (b) ripple = 2%, $h/W = 0.07$; (c) ripple = 2%, $h/W = 0.10$; (d) ripple = 2%, $h/W = 0.14$; (e) ripple = 2%, $h/W = 0.20$.

Figure 11 shows resulting magnetic field with optimized cs_1 cross-section, for 0.5% ripple.

2.3. Optimization of Coil Depth D. To achieve uniform magnetic field across cs_2 (Figure 3), the narrower sides of the coil structure are folded downwards to the depth D. The goal is to get magnetic field shape from cross-section cs_1 across the length of the coil L. If the coil ends are not folded down, at the ends of the coil there is a significant increase in the magnetic field strength (Figure 11).

The second part of the optimization process is determining the depth, D, to which the sides have to be folded down. The methodology to achieve this is the same as in the first part of the optimization process, but with one significant difference. The magnetic field shape throughout the coil length should be consistent, meaning that the magnetic field strength should not deviate across cs_2 cross-section. If the same ripple value as in cs_1 optimization would be allowed during cs_2 optimization, the result would have unwanted field increase at the coil ends, as shown in Figure 12(a). Thus, surface enclosed by W and L of the coil has a minor part of uniform magnetic field, as shown in Figure 13(a). In this stage of optimization, modified figure of merit is adopted:

$$\mathrm{FoM}_2 = B_{\mathrm{avg}} \cdot \frac{\text{uniform section surface}}{\text{coil surface}}. \qquad (9)$$

It is not practical to try to optimize cs_2 cross-section for 0% ripple, but using a ripple value 10 times lower than that used during cs_1 optimization gives good enough result (Figure 12(b)).

Due to high D/W ratio (0.43), the uniform section of the magnetic field at the coil end is quite irregular (Figure 13(b)). Such shape of the uniform field was obtained by optimizing cs_1 for 5% ripple and cs_2 for 0.5% ripple.

Alternative approach is to first optimize cs_1 for 0.5% ripple and then optimize cs_2 for 5% ripple. The resulting magnetic field is given in Figure 12(c). Uniform surface is 0.14% larger, and FoM_2 value is 10% lower, but the shape of

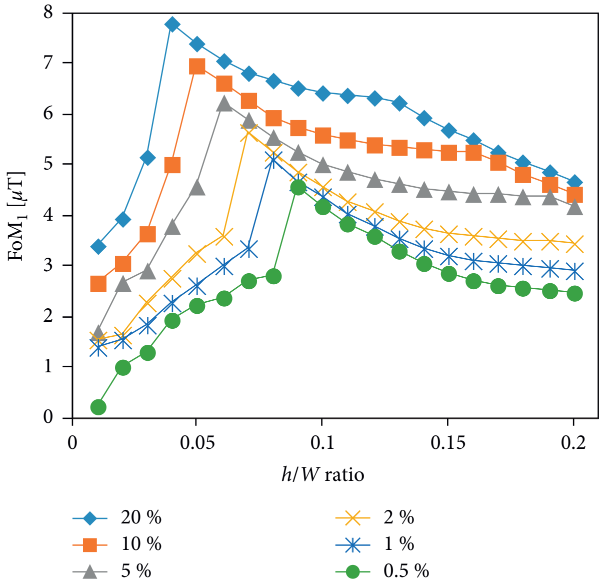


FIGURE 7: FoM$_1$ analysis results.

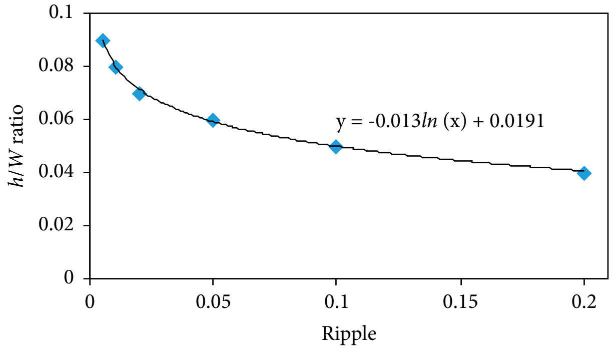


FIGURE 8: Optimal h/W ratios for different ripple values: simulated results are denoted by blue markers, and fitted mathematical model is represented by solid line.

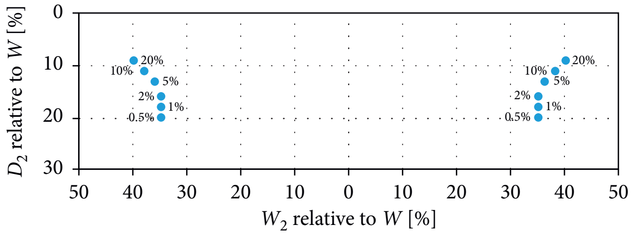


FIGURE 9: Positions of second coil layer that achieve maximal FoM$_1$, for different ripple values with referent plane placed at optimal distance.

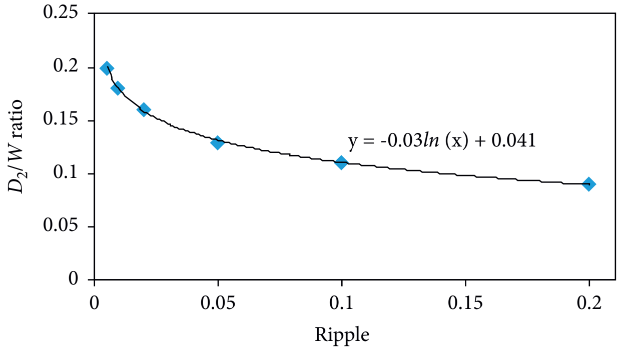


FIGURE 10: Optimal depth of second coil layer $(D)_2$ for different ripple values: simulated results are denoted by blue markers, and fitted mathematical model is represented by solid line.

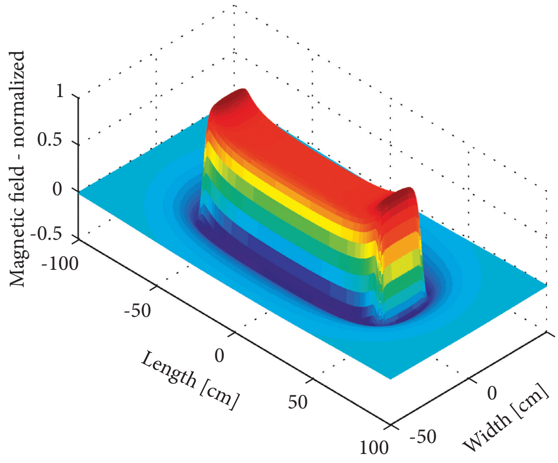


FIGURE 11: Simulated magnetic field distribution generated by two-layer coil with optimized cs_1.

the surface with uniform magnetic field follows the rectangular shape significantly better, as shown in Figure 13(c).

There is a trade-off between FoM$_2$ value and the shape of the surface with uniform magnetic field. The best results are obtained when both the cs_1 and the cs_2 optimizations are done for ripple value equal to one-half of the desired ripple. Figure 12(d) shows the magnetic field when cs_1 and cs_2 are optimized for 2.5% ripple. The uniform surface (Figure 13(d)) is drawn for ripple value of 5%. It results in the highest FoM$_2$ value and largest uniform surface. This approach is chosen as optimal.

The depth of the end fold D is a function of the ripple value (r) and the length to width ratio (L/W) of the coil. Figure 14 shows the end fold depths for different ripple values and L/W ratios. The fold depth increases for coils with larger L/W ratio and with lower allowed ripple value.

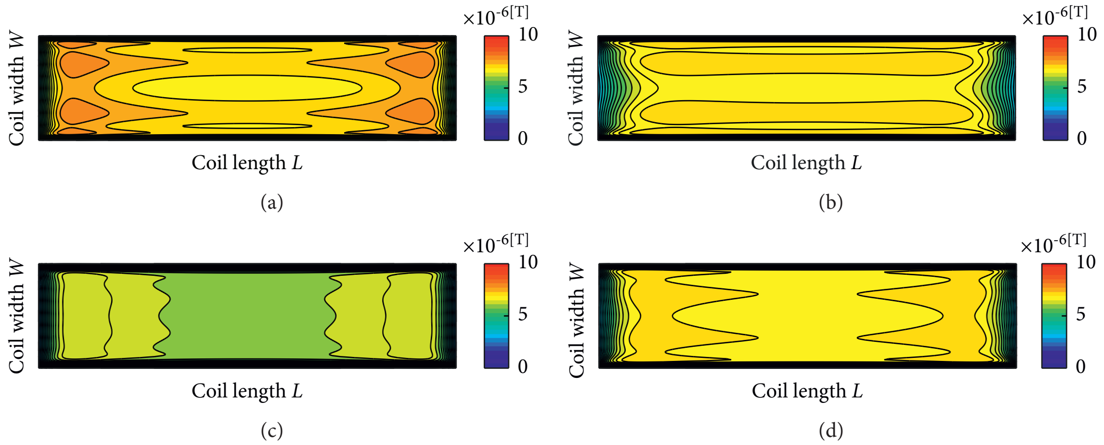
Results of cs_2 optimization are given as markers, and solid lines represent mathematical model, (9), which can be used to calculate required folding depth.

$$D = W \cdot \frac{1}{20.4 \cdot \sqrt{r}} \left(1 - e^{-3.25\sqrt{r}\ (L/W)}\right). \qquad (10)$$

The proposed optimization method is developed for rectangular coil where the percentage of uniform surface increases with higher L/W ratio of the coil (Figure 15).

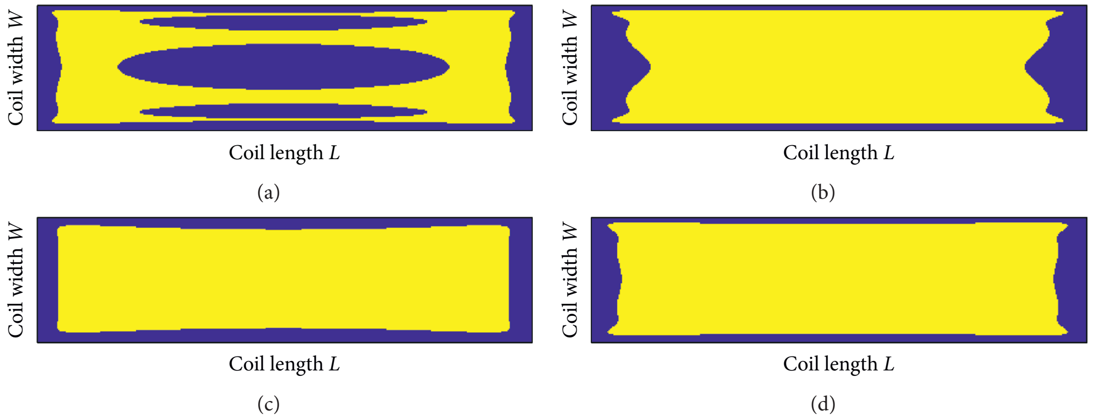
3. Measurements

To evaluate the developed optimization method, magnetic field measurements are performed on fabricated coils



(a) (b) (c) (d)

Figure 12: Magnetic field shape (top view) for different ripple values: (a) cs_1 5%, cs_2 5% ($D/W=0.19$); (b) cs_1 5%, cs_2 0.5% ($D/W=0.43$); (c) cs_1 0.5%, cs_2 5% ($D/W=0.19$); (d) cs_1 2.5%, cs_2 2.5% ($D/W=0.26$).



(a) (b) (c) (d)

Figure 13: Uniform field section (top view) for different combinations of cs_1 and cs_2 ripple values: (a) uniform magnetic field section (49.81%) for magnetic field shape from Figure 12(a), $\mathrm{FoM}_2=0.0015$ T; (b) uniform magnetic field section (75.78%) for magnetic field shape from Figure 12(b), $\mathrm{FoM}_2=0.0021$ T; (c) uniform magnetic field section (75.92%) for magnetic field shape from Figure 12(c), $\mathrm{FoM}_2=0.0019$ T; (d) uniform magnetic field section (79.85%) for magnetic field shape from Figure 12(d), $\mathrm{FoM}_2=0.0021$ T.

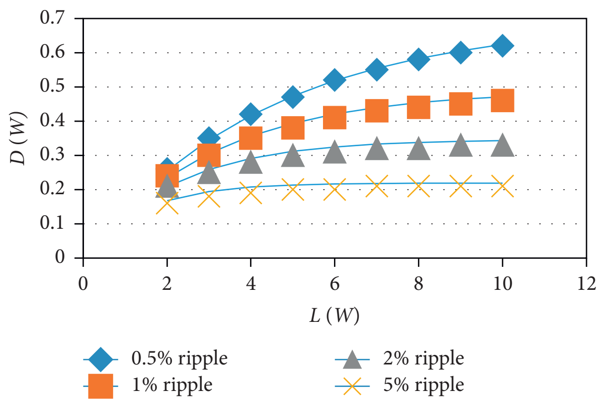


Figure 14: Required end fold depth (D) for different ripple values and L/W ratios.

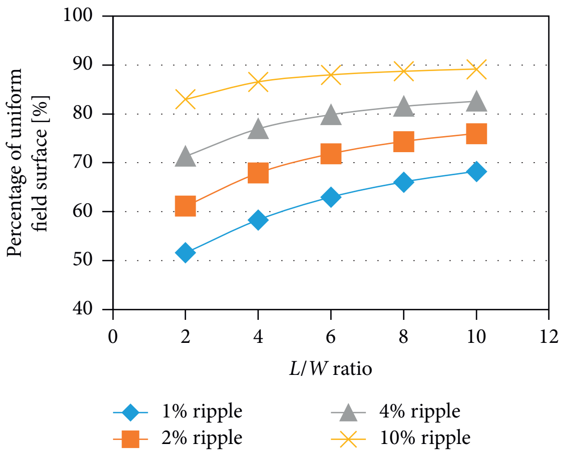


Figure 15: Percentage of uniform field surface for different L/W ratios.

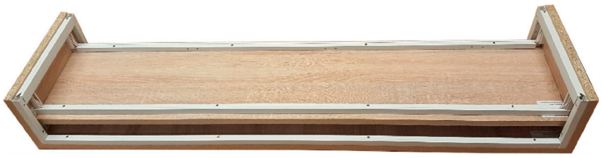


Figure 16: Fabricated proposed TX-coil, bottom view.

Figure 17: Fabricated conventional TX-coil, bottom view.

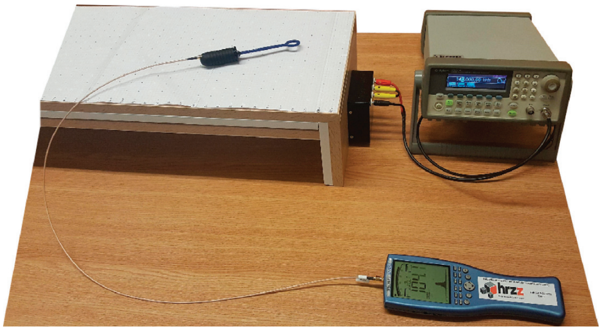


Figure 18: Measuring setup.

(proposed and conventional). The first coil layer size is set as $L = 114$ cm, $W = 28$ cm. The position of second coil layer is calculated using proposed optimization method for ripple value of 1%. Optimizations for cross-sections cs_1 and cs_2 are done for 0.5% ripple (one-half of the desired ripple). Optimal height of the referent plane equals (7): $h = 2.46$ cm. Calculated variables of the second coil layer position are $W_2 = 19.6$ cm, $D_2 = 5.6$ cm. Calculated end fold depth is $D = 11.8$ cm. The fabricated proposed TX-coil prototype is shown in Figure 16. Actual placement of the coil layers differs a bit from the calculated values, mainly due to the limited precision of fabrication ($W_2 = 20$ cm, $D_2 = 6$ cm, $D = 12$ cm). The wire that forms coil layers is placed in the white wire casing. Moreover, proposed TX-coil prototype is constructed using plywood framework to hold wire casings and ensure flat surface above in order to carry out magnetic field measurements in the same plane. Each coil layer consists of 4 windings of Litz wire, resulting in inductance of 103 μH.

Conventional, single-layer rectangular coil (114 × 28 cm) is also fabricated using Litz wire placed in the white wire casings installed at one side of the flat plywood board (Figure 17). Coil is made up of 6 windings, which results in inductance of 116 μH.

The magnetic field strength of these coils was measured at a temperature of 26°C and humidity of 50%. Field strength measurements are performed in the near field zone with measuring equipment consisting of a Spectran NF-5035 spectral analyzer, PBS-H3 probe (25 mm magnetic field test with 50 Ohms SMB m socket), and SMA cable. In performed experiments, fabricated TX-coils were energized by Agilent 33250A signal generator at operating frequency of 150 kHz.

Magnetic field strength of both fabricated TX-coils is measured at 900 points in a receiving plane 30 mm above the coil. Because it is measured by a PBS-H3 probe connected to the SMA input of a Spectran NF-5035 instrument, the analyzer provides highly sensitive measurement of an external alternating field up to 0.2 V max. Thus, the spectrum analyzer returns voltage values that are proportional to the magnetic field strength values. Hold mode is selected to measure the field strength values. Measuring setup is shown in Figure 18, and measurement results are given in Figure 19, along with the simulation results.

Simulation results of the proposed folded 3D coil structure are shown in Figures 19(a) and 19(b). Measurements of magnetic field strength distribution at receiving plane for both fabricated TX-coils (Figures 19(c) and 19(d)) confirmed that optimal and proposed folded 3D TX-coil structure provides significantly larger uniform magnetic field strength section in comparison to conventional TX-coil structure. From measured magnetic field shape (Figures 19(c) and 19(d)), it can be seen that one side of the coil has slightly higher field values than the opposite side. This is due to limited precision of wire windings placement in the wire casing. Aside from that, measurement and simulation results match to a high degree.

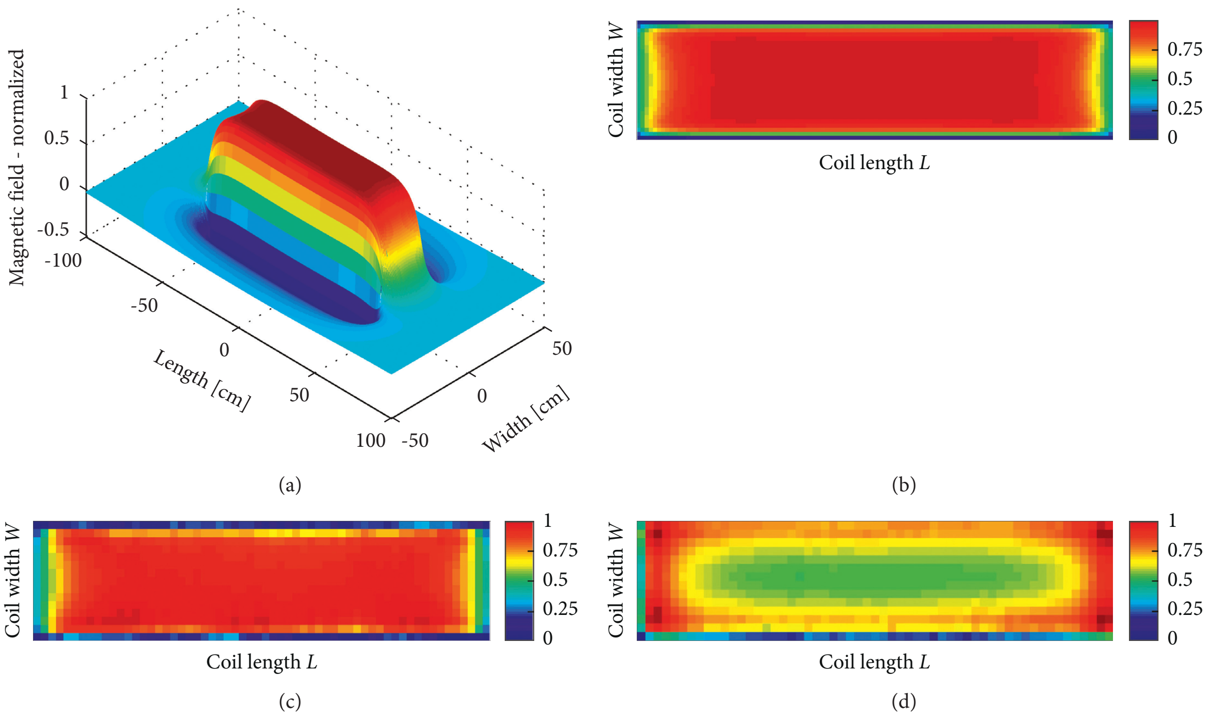


FIGURE 19: Simulation and measurement results: (a) simulated magnetic field distribution of optimized folded 3D coil structure; (b) normalized simulation results, top view; (c) normalized measurement results, top view (optimized folded 3D coil structure); (d) normalized measurement results, top view (conventional coil structure).

TABLE 1: Comparison of various optimized TX-coils' characteristics.

Ref.	Structure	TX-coil (cm)	r (%)	h (mm)	Uniform section (%)
[1]	3D rectangular	$22 \times 18 \times 2$	20	0.5	~32
[6]	Planar square	20×20	20	150	~48
[10]	Planar square	20×20	20	1	~36
[18]	Planar square	20×20	20	50	~51.8
[20]	Planar square	80×80	9.6	10	~42.3
This paper	3D rectangular	$114 \times 28 \times 12$	10	30	~55.3

Measured characteristics of different TX-coil designs which generate uniform magnetic field strength distribution at receiving plane are listed in Table 1. In comparison with other TX-coil characteristics from Table 1, the proposed coil generates improved magnetic field strength distribution with respect to uniform section of receiving plane. Further development of the proposed TX-coil will be aimed at reducing coil depth, D, which will result in lower profile of TX-coil.

4. Conclusion

A novel 3D TX-coil structure with improved uniform magnetic field distribution is proposed. Optimization methodology is presented and optimizing method is developed for rectangular 3D coil structure. Optimization results in coil structure that has maximized average magnetic field strength and uniform surface section. Computer simulations and experimental verification of the optimization method are successfully carried out. Both TX-coils, conventional and proposed, are fabricated to carry out magnetic field measurements which confirmed that proposed coil, unlike conventional coil, generates uniform magnetic field strength distribution at receiving plane. Furthermore, measured results show larger uniform section in comparison to other optimized TX-coil structures. Depending on the maximal ripple and proportions of the coil, optimization method results in 3D coil structure that generates improved uniform magnetic field strength distribution across 50% to 90% of the receiving plane.

References

[1] W.-S. Lee, H. L. Lee, K.-S. Oh, and J.-W. Yu, "Uniform magnetic field distribution of a spatially structured resonant coil for wireless power transfer," *Applied Physics Letters*, vol. 100, no. 21, 2012.

[2] D. Vinko, I. Biondić, and D. Bilandžija, "Impact of receiver power and coupling coefficient on resonant frequency in wireless power transfer system," in *Proceedings of the 2019 International Symposium ELMAR*, pp. 207–210, Zadar, Croatia, September 2019.

[3] J. Park, Y. Tak, Y. Kim, Y. Kim, and S. Nam, "Investigation of adaptive matching methods for near-field wireless power transfer," *IEEE Transactions on Antennas and Propagation*, vol. 59, no. 5, pp. 1769–1773, 2011.

[4] Q. Wang, W. Che, M. Dionigi, F. Mastri, M. Mongiardo, and G. Monti, "Gains maximization via impedance matching networks for wireless power transfer," *Progress In Electromagnetics Research*, vol. 164, pp. 135–153, 2019.

[5] Q. Wang, W. Che, M. Mongiardo, and G. Monti, "Wireless power transfer system with high misalignment tolerance for bio-medical implants," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3023–3027, 2020.

[6] S. Wang, Z. Hu, C. Rong, C. Lu, J. Chen, and M. Liu, "Planar multiple-antiparallel square transmitter for position-insensitive wireless power transfer," *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 2, pp. 188–192, 2018.

[7] C. Qiu, K. T. Chau, C. Liu, T. W. Ching, and Z. Zhang, "Modular inductive power transmission system for high misalignment electric vehicle application," *Journal of Applied Physics*, vol. 117, p. 17B528, 2015.

[8] F. Jolani, Y. Yu, and Z. Chen, "A planar positioning-free magnetically-coupled resonant wireless power transfer," in *Proceedings of the 2015 IEEE Wireless Power Transfer Conference (WPTC)*, pp. 1–3, Boulder, CO, USA, May 2015.

[9] X. Liu and S. Y. Hui, "Optimal design of a hybrid winding structure for planar contactless battery charging platform," *IEEE Transactions on Power Electronics*, vol. 23, no. 1, pp. 455–463, 2008.

[10] J. J. Casanova, Z. N. Low, J. Lin, and R. Tseng, "Transmitting coil achieving uniform magnetic field distribution for planar wireless power transfer system," in *Proceedings of the 2009 IEEE Radio and Wireless Symposium*, pp. 530–533, San Diego, CA, USA, January 2009.

[11] D. Yinliang, S. Yuanmao, and G. Yougang, "Design of coil structure achieving uniform magnetic field distribution for wireless charging platform," in *Proceedings of the 2011 4th International Conference on Power Electronics Systems and Applications*, pp. 1–5, Hong Kong, China, June 2011.

[12] L. Shen, W. Tang, H. Xiang, and W. Zhuang, "Uniform magnetic field by changing the current distribution on the planar coil for displacement-insensitive wireless power transfer/near field communication," *Microwave and Optical Technology Letters*, vol. 57, no. 2, pp. 424–427, 2014.

[13] Y. Zhang, T. Lu, Z. Zhao, F. He, K. Chen, and L. Yuan, "Quasi-uniform magnetic field generated by multiple transmitters of magnetically-coupled resonant wireless power transfer," in *Proceedings of the 2015 18th International Conference on Electrical Machines and Systems (ICEMS)*, pp. 1030–1034, Pattaya, Thailand, October 2015.

[14] T.-D. Yeo, D.-H. Kim, S. C. Chae, S.-T. Khang, and J.-W. Yu, "Design of free-positioning wireless power charging system for aaa rechargeable battery," in *Proceedings of the 2016 46th European Microwave Conference (EuMC)*, pp. 759–762, London, UK, October 2016.

[15] E. Elkhouly and S. Yang, "Transmitter coil design for resonant wireless power transfer," in *Proceedings of the 2016 IEEE PELS Workshop on Emerging Technologies: Wireless Power Transfer (WoW)*, pp. 1–5, Knoxville, TN, USA, October 2016.

[16] S. C. Tang, N. J. McDannold, and M. Vaninetti, "A wireless batteryless implantable radiofrequency lesioning device powered by intermediate-range segmented coil transmitter," in *Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1966–1969, Jeju Island, South Korea, July 2017.

[17] M. R. Basar, M. Y. Ahmad, F. Ibrahim, and J. Cho, "Resonant inductive power transfer system for freely moving capsule endoscope with highly uniform magnetic field," in *Proceedings of the 2016 IEEE Industrial Electronics and Applications Conference (IEACon)*, pp. 393–397, Kota Kinabalu, Malaysia, November 2016.

[18] T.-H. Kim, G.-H. Yun, W. Y. Lee, and J.-G. Yook, "Asymmetric coil structures for highly efficient wireless power transfer systems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 7, pp. 3443–3451, 2018.

[19] H. Wang, L. Deng, H. Luo, S. Huang, and C. Liao, "Omnidirectional wireless power transfer system with multiple receivers and a single wire wound spiral transmitter," *Progress In Electromagnetics Research C*, vol. 94, no. 7, pp. 189–202, 2019.

[20] Y. Zhang, L. Wang, Y. Guo, and Y. Zhang, "Optimisation of planar rectangular coil achieving uniform magnetic field distribution for EV wireless charging based on genetic algorithm," *IET Power Electronics*, vol. 12, pp. 2706–2712, 2019.

[21] Q. Xu, Q. Hu, H. Wang, Z.-H. Mao, and M. Sun, "Optimal design of planar spiral coil for uniform magnetic field to wirelessly power position-free targets," *IEEE Transactions on Magnetics*, vol. 57, no. 2, pp. 1–9, 2021.

Distributed Finite-Time Consensus Control of Second-Order Multiagent Systems Subject to Communication Time Delay

Dillip Kumar Nayak, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, dknayak225@yahoo.co.in*

Subhendu Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, subhendu.s@yahoo.co.in*

Alekha Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, alekha.sahoo241@gmail.com*

Pratik Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, pratikmohanty92@hotmail.com*

Abstract

This paper studies the consensus problem of a second-order multiagent system (MAS) with fixed communication delay under the structure of leaderless and leader-following systems. By using graph theory and finite-time control scheme, a distributed control protocol is proposed for each agent to reach consensus in a finite tim e. In practical application, the tim e delay of states is unavoidable, and for this, the consensus method is supposed to be extended to solve the time-delay problem. Thus, a finite-time consensus protocol with communication time delay is proposed in this paper. Compared with the general consensus method, the reliability and convergence speed of the system are increased by using the finite-tim e control. In addition, the protocol is distributed, and all agents have only local interactions. Finally, the effectiveness of the proposed protocol is verified by two num erical simulations.

1. Introduction

In recent years, the consensus control [1–4] of the multiagent system has attracted great attention, which has been widely applied in the fields of UAV formation [5], intelligent transportation system [6, 7], satellite attitude calibration [8, 9], aerospace, and other fields. The development of the multiagent system comes from people's research on the overall consensus behavior of biological clusters in nature, such as ants moving, birds flying in clusters, and so on. Consensus problem means that the final state information of each agent in a multiagent system converges to a common value. It is a very meaningful work to design effective control protocols and study how agents can reach agreement within a limited time through cooperation and accomplish tasks that cannot be accomplished by a single agent.

The concept of the multiagent system was first proposed by Minsky [10] in 1986 and later attracted widespread attention. A single entity with the ability of information interaction can be regarded as a single agent. In 2003, Jadbabaie [11] et al. made a specific theoretical analysis of the clustering phenomenon of the model by using the knowledge of graph theory and pointed out the conditions for the system to achieve consensus. Subsequently, Fax and Murray [12] proposed a theoretical research framework for the consensus problem and designed a consensus control algorithm based on the first-order multiagent system model. Zou et al. [13] studied the distributed consensus tracking problem for heterogeneous switched nonlinear multiagent systems with actuator faults and arbitrary switching signals.

The finite-time consensus (FTC) [14, 15] control method has attracted great attention of many researchers in recent years. Long Wang and Feng Xiao [16] analyzed the conditions for the continuous-time multiagent system to achieve consensus based on the finite-time stability theory. In [17], a control strategy based on a velocity-free distributed observer was designed to address the finite-time position consensus tracking control problem for heterogeneous leader-follower multiple AUV systems. In [18], the problem of practical FTC for the second-order MASs with unknown nonlinear dynamics has been studied, in which dynamics are extended to switched nonlinear systems. Li et al. [19] studied the finite-time cooperative tracking problem of a class of heterogeneous mixed-order multiagent systems. Zou et al. [20] developed protocols with only state transmissions, by which the consensus can be reached in fixed time.

In practice, due to the communication environment and the communication equipment itself, communication delay inevitably exists in the information transmission between multiple moving agents. Therefore, it is of practical significance to study the consensus of the multiagent system with communication delay [21–23]. Li et al. [24] studied the consensus of leadership following for a first-order multiagent system with noise disturbance and communication delay. Hou et al. [25] studied the consensus problem of a class of general second-order multiagent systems with communication delay by using the relationship between the roots of the characteristic equation and delay parameters. In [22], a distributed algorithm for a second-order multiagent system with nonuniform time delay was proposed to make all agents reach consensus.

Algorithm convergence speed is an important index to evaluate the algorithm performance, especially in some practical complex large-scale systems, which requires the system to be able to converge in a limited time. Considering the possible communication delays in real applications, it is significant to study the time-delay problem with the finite-time control.

Inspired by the above works, we complete the following works. A finite-time consensus problem involved with communication delay is considered in this paper. The communication graph is adopted as a directed connected graph with a fixed topology. Each agent is modelled by a second-order integrator with unknown nonlinear dynamics. In order to verify the effectiveness of the proposed control protocol, leaderless and leader-following systems are selected, respectively, in the numerical simulation. It is proved that the FTC can be achieved by given protocols. The contributions of this paper are summarized as follows:

(1) We use the sign function feedback control method based on partial state information to achieve a finite-time consensus for the MAS with communication time-delay problem when the time delay is under the upper limit. Meanwhile, the protocol is distributed and developed for each agent, and only partial information interaction is required.

(2) Two different structures of leaderless and leader-following multiagent systems have been discussed in this paper. Simulation results indicate that the finite-time consensus can be reached by the given protocols. Compared with the general consensus method which converges asymptotically, the convergence speed of the system is improved effectively by using the finite-time method proposed in this paper.

The outline of this work is as follows. Some preliminary knowledge is given in Section 2. Consensus control protocols proposed for leaderless and leader-following systems to solve the finite-time consensus problem with communication time delay and the proof are given in Section 3. Two examples of numerical simulation showing the effectiveness of the proposed protocol are given in Section 4. The summary of this paper and future research is given in Section 5.

2. Problem Formulation

2.1. Graph Theory. Graph theory analysis is an important tool to deal with distributed control problems, which can easily describe the information transmission relationship of a MAS.

In the multiagent system, for simple analysis, each single agent can be regarded as a node, and information can be exchanged between nodes. Then, a graph $G(A) = (V, E, A)$ can be used to describe the interaction between agents; a system of n intelligent agents can be described as a nonempty vertex set of $V(G) = \{v_1, v_2, \ldots, v_n\}$; $E \in V \times V$ represents a collection of nonempty directed edges; $A = [a_{ij}]_{n\times n} \in R^{n\times n}$ is the weighted adjacency matrix; and a_{ij} represents the weight of an edge in the topology starting at v_j and ending at v_i. Moreover, $a_{ij} > 0$ if $(v_j, v_i) \in E$; otherwise, $a_{ij} = 0$. In addition, $a_{ii} = 0$. For example, when the weight is set to 1, $a_{ij} = 1$. The weighted adjacency matrix $B = di\ ag\{b_1, b_2, \ldots, b_n\}$ is defined to describe the information interaction relationship between the leaders and the followers. If the followers can receive the information from the leader, then $b_{ij} > 0$, and when there is no information transmission, $b_{ij} = 0$.

Define the degree matrix $D = di\ ag\{d_1, d_2, \cdots, d_n\}$, where d_i represents the out degree of agent i, which is equal to the sum of every row of elements in the weighted adjacent matrix A. Define the Laplacian matrix $L = D - A$, which is the matrix representation of a topological graph.

2.2. Matrix Theory. Let matrix $A \in R^{m\times n}$ and matrix $B \in R^{p\times q}$; then, the Kronecker product of matrices $A \otimes B$ is a matrix with dimension of $mp \times nq$.

The Kronecker product is described by

$$A \otimes B = \begin{pmatrix} a_{11}B & \ldots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}. \tag{1}$$

Properties of the Kronecker product are given as follows:

(1) $(A \otimes B)(C \otimes D) = (AC) \otimes (B\ D)$.

(2) $(A \otimes B) + (A \otimes C) = A \otimes (B + C)$.

(3) $(AB)^{-1} = A^{-1} \otimes B^{-1}$.

(4) $(A \otimes B)^T = A^T \otimes B^T$.

2.3. Notations. Throughout this paper, the following notations are used. $\mathbf{1}_{\text{n}}$ represents an n-dimensional row vector with an element of 1; I_n represents the identity matrix of the dimensionality of $n \times n$; $\mathbf{0}_n$ denotes the zero matrix of appropriate dimension; $|\cdot|$ is the modulus of a real number; and $\|\cdot\|$ is the norm of a vector.

3. System Model and Protocol Design

3.1. Problem Description. Consider a class of second-order leaderless m ultiagent system s consisting of n agents. The dynam ics of ith agent can be described as

$$\begin{cases} \dot{x}_i(t) = v_i(t) \\ \dot{v}_i(t) = f(x_i(t), v_i(t)) + u_i(t) \end{cases}, \tag{2}$$

where $i = 1, 2, \cdots, n$, $x_i(t) \in R^n$ is the position state, $v_i(t) \in R^n$ is the velocity state of the ith agent at t, $u_i(t) \in R^n$ represents the input for agent i, and $f(x_i(t), v_i(t))$ is the unknown continuous nonlinear function with uncertainties.

Finite-time consensus is to be achieved when the following equations hold. As time t approaches T^*, under any given bounded initial state, it satisfies

$$\lim_{t \longrightarrow T^*} \left\| x_i(t) - x_j(t) \right\| = 0, \tag{3}$$

$$\lim_{t \longrightarrow T^*} \left\| v_i(t) - v_j(t) \right\| = 0, \tag{4}$$

where T^* is finite time.

If we design an effective consensus control protocol to make equations (3) and (4) hold, it can be said that the leader-following multiagent system of the second-order integrator model above can achieve consensus within a finite-time T^*.

With zero time delay, the general consensus control protocol is given as

$$u_i(t) = -\alpha\left(\sum_{j=1}^{n} L_{ij} x_j(t)\right) - \beta\left(\sum_{j=1}^{n} L_{ij} v_j(t)\right). \tag{5}$$

Lemma 1 (see [24]). The control protocol (5) makes the MAS with (5) achieve consensus if and only if G has a directed spanning tree and the following inequalities hold:

$$\alpha > 0, \beta > 0, \frac{\beta^2}{\alpha} > \max_{i=1,2,\cdots n} \frac{\mathrm{Im}^2(\lambda_i)}{\mathrm{Re}(\lambda_i)|\lambda_i|^2}. \tag{6}$$

Choose the appropriate Lyapunov function candidate, and the consensus control protocol of (5) can make the multiagent system described in (2) achieve consensus asymptotically. This was proved in [24].

The following discussion considers the case of communication time delay. Under the structure of leaderless and leader-following systems, effective finite-time consensus protocols were designed for MAS with communication time delay.

3.2. Finite-Time Consensus Protocol with Communication Delay for Leaderless MAS. In this section, consider the existence of communication delay in the leaderless system, and the following finite-time control protocol is proposed to solve the consensus problem with time delay.

For the ith agent in the multiagent system, we consider the following control protocol:

$$u_i(t) = -\alpha\left[\sum_{j=1}^{n} L_{ij} x_j(t-\tau) + sig\left(\sum_{j=1}^{n} L_{ij} x_j(t-\tau)\right)^k\right] - \beta\left[\sum_{j=1}^{n} L_{ij} v_j(t-\tau) + sig\left(\sum_{j=1}^{n} L_{ij} v_j(t-\tau)\right)^k\right], \tag{7}$$

where τ represents the time delay of information transfer from agent i to agent j.

We assume that $\tau_{ij} = \tau_{ji} = \tau$, that is, the communication delay between two agents is considered the same.

In (7),

$$\sum_{j=1}^{n} L_{ij} x_j = \sum_{j=1}^{n} a_{ij}\left(x_i - x_j\right). \tag{8}$$

Equation (8) represents the relative state information between agent i and its neighbor agent j. In equation (7), $sig(x)^k = |x|^k \mathrm{sign}(x)$, where $\mathrm{sign}(x)$ is a symbolic function. One has $\alpha > 0$, $\beta > 0$.

We select the state of agent 1 to facilitate the calculation and evaluation of the consensus error. Then, we define the position error and velocity error as follows:

$$e_{xi}(t) = x_i(t) - 1_{n-1} x_1(t), \tag{9}$$

$$e_{vi}(t) = v_i(t) - 1_{n-1} v_1(t), \tag{10}$$

where $i = 2, 3, \ldots, n$. So, it is easy to know that $e_{xi} \in R^{n-1}$, $e_{vi} \in R^{n-1}$.

$E = [e_{xi}, e_{vi}]^T$ represents the neighborhood error matrix.

Substituting the proposed control protocol (7) into system equation (2), the following error matrix equation can be obtained.

$$\dot{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{0}_{n-1} & \mathbf{I}_{n-1} \\ \mathbf{0}_{n-1} & \mathbf{0}_{n-1} \end{bmatrix} \mathbf{E}(t) + \begin{bmatrix} \mathbf{0}_{n-1} & \mathbf{0}_{n-1} \\ -\alpha\hat{\mathbf{L}} & -\beta\hat{\mathbf{L}} \end{bmatrix} \mathbf{E}(t-\tau)$$

$$+\begin{bmatrix} \mathbf{0}_{n-1} & \mathbf{0}_{n-1} \\ -\alpha\mathbf{I}_{n-1} & -\beta\mathbf{I}_{n-1} \end{bmatrix}\mathrm{sig}\,(\mathbf{E}\overline{\widehat{L}}(t-\tau))^{k} + \begin{bmatrix} \mathbf{0}_{(n-1)\times 1} \\ \mathbf{e_f} \end{bmatrix}, \tag{11}$$

where $e_f = f_i(t) - f_1(t)$, $\widehat{L} = L_{22} + 1_{n-1}A_{1n}^T$, $L_{22} = \begin{bmatrix} d_2 & -a_{23} & \cdots & -a_{2n} \\ -a_{32} & d_3 & \cdots & -a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n2} & -a_{n3} & \cdots & d_n \end{bmatrix}$, $A_{1n} = \begin{bmatrix} a_{12} \\ a_{13} \\ \vdots \\ a_{1n} \end{bmatrix}$, and $\overline{\widehat{L}} = \begin{bmatrix} \widehat{L} & 0_{n-1} \\ 0_{n-1} & \widehat{L} \end{bmatrix}$.

Define $A_{n1} = \begin{bmatrix} a_{21} & a_{31} & \cdots & a_{n1} \end{bmatrix}^T$; then, the Laplacian matrix $L = \begin{bmatrix} d_1 & -A_{1n}^T \\ -A_{1n} & L_{22} \end{bmatrix}$.

Equation (11) can be written as

$$\dot{E}(t) = (I_{n-1} \otimes B_1)E(t) + \left(\widehat{L} \otimes B_2\right)E(t-\tau) + (I_{n-1} \otimes B_2)sig\,(\overline{\widehat{L}}E(t-\tau))^k + \mathrm{B}_3, \tag{12}$$

where $B_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B_2 = \begin{bmatrix} 0 & 0 \\ -\alpha & -\beta \end{bmatrix}$, and $B_3 = \begin{bmatrix} 0_{(n-1)\times 1} \\ e_f \end{bmatrix}$.

Apparently, considering the consensus of system (2) translates into studying the convergence of error matrix equation (12).

Lemma 2. *Consider the system*

$$\begin{aligned} \dot{x} &= f(x), \\ f(0) &= 0, \\ x &\in R^n, \end{aligned} \tag{13}$$

and there exist a positive definite continuous function $V(x)$: $U \longrightarrow R$, real numbers $c_1, c_2 > 0$ and $\alpha \in (0,1)$, and an open neighborhood $U_0 \subset U$ of the origin such that $\dot{V}(x) + c_1V(x) + c_2V^{\alpha}(x) \leq 0$, $x = U_0/\{0\}$. Then, $V(x)$ will approach 0 in a finite-time T^*.

Lamma 3 (see [24]). The Laplacian matrix L has exactly one zero eigenvalue and all the other eigenvalues have positive parts if and only if the directed network has a directed spanning tree.

Assumption 1 In the leaderless multiagent system, assume that the communication topology in the MAS described by equation (2) is represented by a directed graph G and the network topology contains a spanning tree.

Assumption 2 (see [26]). In the leader-following multiagent system, assume that the communication topology in the MAS described by (2) and (24) contains a directed spanning tree in the digraph $\overline{G} = G \cup \{0\}$ with the leader $\{0\}$ as the root node.

Assumption 3 (see [22]). There exist two nonnegative constants k_1 and k_2, such that

$$\left\| f(x_1(t), x_2(t)) - f(y_1(t), y_2(t)) \right\| \leq k_1 \left\| x_1(t) - y_1(t) \right\| + k_2 \left\| x_2(t) - y_2(t) \right\|, \tag{14}$$

for any $x_1(t)$, $x_2(t)$, $y_1(t)$, $y_2(t) \in R^n$.

3.3. Consensus Analysis

Theorem 1. *Consider the leaderless system with equation (2). Suppose G has a directed spanning tree and that control protocol (7) can make the MAS achieve finite-time consensus in the time-delay case.*

Proof. The Lyapunov function is constructed as follows:

$$V(t) = \frac{1}{2}E^T(t)\mathrm{Q}\mathrm{E}(t), \tag{15}$$

where $\mathrm{Q} = \begin{bmatrix} 2\alpha\beta\widehat{L} & \alpha I_{n-1} \\ \alpha I_{n-1} & \beta I_{n-1} \end{bmatrix} \in R^{2(n-1)}$.

When $\beta > 0$, $\lambda_{\min}(\widehat{L}) > \alpha/2\beta^2$, one has $Q > 0$, so that $V(t)$ is positive definite.

Computing the derivative of time for $V(t)$, one has

$$\dot{V}(t) = E^T(t)\mathrm{Q}\dot{E}(t). \tag{16}$$

Substituting (12) into (16) yields

$$\begin{aligned} \dot{V}(t) &= E^T(t)Q\left(I_{n-1} \otimes B_1\right)E(t) + E^T(t)Q\left(\widehat{L} \otimes B_2\right)E(t-\tau) \\ &\quad + E^T(t)Q\left(I_{n-1} \otimes B_2\right)sig\,(\mathrm{E}\overline{\widehat{L}}\,(t-\tau))^k + E^T(t)\mathrm{Q}\mathrm{B}_3 \\ &= E^T(t)P_1E(t) + E^T(t)P_2E(t-\tau) \\ &\quad + E^T(t)P_3sig\,(\overline{\widehat{L}}\,E(t-\tau))^k + E^T(t)P_4, \end{aligned} \tag{17}$$

where $P_1 = \begin{bmatrix} 0_{n-1} & 2\alpha\beta\widehat{L} \\ 0_{n-1} & \alpha I_{n-1} \end{bmatrix}$, $P_2 = \begin{bmatrix} -\alpha^2\widehat{L} & -\alpha\beta\widehat{L} \\ -\alpha\beta\widehat{L} & -\beta^2\widehat{L} \end{bmatrix}$, $P_3 = \begin{bmatrix} -\alpha^2 I_{n-1} & -\alpha\beta I_{n-1} \\ -\alpha\beta I_{n-1} & -\beta^2 I_{n-1} \end{bmatrix}$, and $P_4 = \begin{bmatrix} \alpha e_f \\ \beta e_f \end{bmatrix}$.

Then,

$$\begin{aligned}\dot{V}(t) &= \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}^T \begin{bmatrix} P_1 & P_2 \\ 0_{2(n-1)} & 0_{2(n-1)} \end{bmatrix} \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}, \\ &+ \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}^T \begin{bmatrix} P_3 & 0_{2(n-1)} \\ 0_{2(n-1)} & 0_{2(n-1)} \end{bmatrix} sig\left(\begin{bmatrix} 0_{2(n-1)} & 0_{2(n-1)} \\ 0_{2(n-1)} & \overline{\widehat{L}} \end{bmatrix} \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix} \right)^k, \\ &+ \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}^T \begin{bmatrix} \alpha e_f + \beta e_f \\ 0_{2(n-1)\times 1} \end{bmatrix} = -z^T M_1 z - z^T M_2 sig(z)^k - z^T M_3, \end{aligned} \tag{18}$$

where $z = \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}$, $M_1 = \begin{bmatrix} -P_1 & -P_2 \\ 0_{2(n-1)} & 0_{2(n-1)} \end{bmatrix}$, $M_2 = \begin{bmatrix} -P_3 \overline{\widehat{L}}^k & 0_{2(n-1)} \\ 0_{2(n-1)} & 0_{2(n-1)} \end{bmatrix}$, and $M_3 = \begin{bmatrix} -(\alpha e_f + \beta e_f) \\ 0_{2(n-1)\times 1} \end{bmatrix}$.

From (18), one has

$$\dot{V}(t) \leq -\lambda_{\min}(M_1)\|z\|^2 - \lambda_{\min}(M_2)\|z\|^{k+1} - \|M_3\|\|z\| = -m_1\|z\|^2 - m_2\|z\|^{k+1} - \|M_3\|\|z\|, \tag{19}$$

where $m_1 = \lambda_{\min}(M_1)$ and $m_2 = \lambda_{\min}(M_2)$.

Also, it can be known from Assumption 3 that

$$\begin{aligned}\|M_3\| &= |\alpha+\beta|\|e_f\| = |\alpha+\beta|\|f(x_i(t), v_i(t)) - f(x_1(t), v_1(t))\|, \\ &\leq |\alpha+\beta|\left(k_1\|x_i(t) - x_1(t)\| + k_2\|v_i(t) - v_1(t)\|\right), \\ &= |\alpha+\beta|\left(k_1\|e_{xi}\| + k_2\|e_{vi}\|\right), \\ &= \begin{bmatrix} (k_1|\alpha+\beta| + k_2|\alpha+\beta|)1_{1\times 2(n-1)} & 0_{1\times 2(n-1)} \end{bmatrix} \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}, \\ &\leq |\alpha+\beta|\sqrt{k_1^2 + k_2^2}\|z\|. \end{aligned} \tag{20}$$

Then, substituting (20) into (19), one has

$$\dot{V}(t) \leq -\left(m_1 + |\alpha+\beta|\sqrt{k_1^2+k_2^2}\right)\|z\|^2 - m_2\|z\|^{k+1}. \tag{21}$$

According to equation (15),

$$V(t) = \frac{1}{2}\begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix}^T \begin{bmatrix} Q & 0_{2(n-1)} \\ 0_{2(n-1)} & 0_{2(n-1)} \end{bmatrix} \begin{bmatrix} E(t) \\ E(t-\tau) \end{bmatrix} \leq \lambda_{\max}(\overline{Q})\|z\|^2 = m_0\|z\|^2, \tag{22}$$

where $\overline{Q} = \begin{bmatrix} 1/2Q & 0_{2(n-1)} \\ 0_{2(n-1)} & 0_{2(n-1)} \end{bmatrix}$, $m_0 = \lambda_{\max}(\overline{Q}) = \beta + 2\alpha\beta\lambda_{\max}(\widehat{L})/4+ \sqrt{(\beta/2 + \alpha\beta\lambda_{\max}(\widehat{L}))^2 - 2\alpha\beta^2\lambda_{\max}(\widehat{L}) + \alpha^2}/2$, and $\lambda_{\max}(\widehat{L})$ is the minimum eigenvalue of the matrix $\widehat{L}$.

Substituting (22) into (21) yields

$$\dot{V}(t) \leq -\frac{\left(m_1 + |\alpha+\beta|\sqrt{k_1^2+k_2^2}\right)}{m_0}V(t) - \frac{m_2}{m_0^{1+k/2}}V^{1+k/2}(t). \tag{23}$$

Thus, according to Lemma 2, the system converges in a finite time. This completes the proof. □

3.4. Finite-Time Consensus Protocol with Communication Delay for Leader-Following MAS. In the previous section, we analyzed the consensus problem for leaderless system, and the final consensus state converges to some function related to the agent's initial state. In some practical applications, all agents are desired to converge to a specified value eventually. Then, we add a leader to the system to solve the problem of consensus tracking in this section.

The dynamics of the leader agent are described by

$$\begin{cases} \dot{x}_0(t) = v_0(t) \\ \dot{v}_0(t) = f\left(x_0(t), v_0(t)\right) \end{cases}, \tag{24}$$

and the followers are described as in equation (2).

In this case, for the leader-following system, the distributed finite-time consensus control protocol is proposed:

$$\begin{aligned} u_i(t) = {} & -\alpha\left[\sum_{j=1}^{n} L_{ij}x_j(t-\tau) + sig\left(\sum_{j=1}^{n} L_{ij}x_j(t-\tau)\right)^k\right] - \beta\left[\sum_{j=1}^{n} L_{ij}v_j(t-\tau) + sig\left(\sum_{j=1}^{n} L_{ij}v_j(t-\tau)\right)^k\right] \\ & - b_i\left[x_i(t-\tau) - x_0(t-\tau) + v_i(t-\tau) - v_0(t-\tau)\right], \end{aligned} \tag{25}$$

where b_i is the diagonal element in the adjacency matrix $B = [b_1, b_2, \ldots, b_n]^T$.

Theorem 2. *Consider the leader-following system with equations (2) and (24), and control protocol (25) can make the MAS achieve finite-time consensus in the time-delay case.*

Proof. Define the position error and velocity error as follows:

$$\begin{aligned} \delta_{xi}(t) &= x_i(t) - 1_n x_0(t), \\ \delta_{vi}(t) &= v_i(t) - 1_n v_0(t), \end{aligned} \tag{26}$$

where $i = 1, 2, \ldots, n$, $e_{xi} \in R^n$, $e_{vi} \in R^n$.

Defining $\delta = [\delta_{xi}, \delta_{vi}]^T$ and taking the derivative of δ, one has

$$\begin{aligned} \dot{\delta}(t) = {} & \begin{bmatrix} 0_n & I_n \\ 0_n & 0_n \end{bmatrix}\delta(t) + \begin{bmatrix} 0_n & 0_n \\ -\alpha L - B & -\beta L - B \end{bmatrix}\delta(t-\tau) \\ & + \begin{bmatrix} 0_n & 0_n \\ -\alpha I_n & -\beta I_n \end{bmatrix} sig\left(\overline{L}\delta(t-\tau)\right)^k + \begin{bmatrix} 0_{n\times 1} \\ \delta_f \end{bmatrix}, \end{aligned} \tag{27}$$

where $\overline{L} = \begin{bmatrix} L & 0_n \\ 0_n & L \end{bmatrix}$ and $\delta_f = f_i(t) - f_0(t)$.

Equation (27) can be written as

$$\dot{\delta}(t) = \left(I_n \otimes B_1\right)\delta(t) + \left(L \otimes \mathrm{B}_2 + \mathrm{B} \otimes B_0\right)\delta(t-\tau) + \left(I_n \otimes \mathrm{B}_2\right) sig\left(\overline{L}\delta(t-\tau)\right)^k + \mathrm{B}_3, \tag{28}$$

where $B_0 = \begin{bmatrix} 0 & 0 \\ -1 & -1 \end{bmatrix}$, $B_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B_2 = \begin{bmatrix} 0 & 0 \\ -\alpha & -\beta \end{bmatrix}$, and $B_3 = \begin{bmatrix} 0_{n\times 1} \\ \delta_f \end{bmatrix}$, which are the same as in (12).

Note that (12) and (28) have the same structure, and the rest is similar to the proof of Theorem 1 described in Section 3.3; we can also select the same Lyapunov function to prove that the system can converge in a finite time. Thus, it is omitted here. This completes the proof. □

4. Numerical Simulation

Example 1. Consider a second-order leaderless MAS described by equation (2), consisting of $n = 6$ follower agents. The communication topology is shown in Figure 1, whose weights are taken as follows.

From the graph theory knowledge, we know that

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \end{bmatrix}, \\ \mathbf{L} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -3 & -1 & 4 & 0 \\ 0 & 0 & 0 & -2 & 0 & 2 \end{bmatrix}. \end{aligned} \tag{29}$$

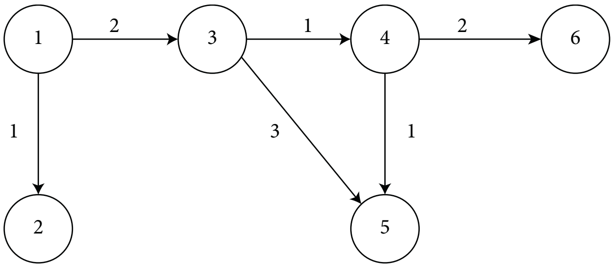


Figure 1: Topology of the graph G.

The initial states of each agent in the multiagent system are taken as follows: position $x(0) = [\,2\ \ 1.5\ \ 5\ \ -1.5\ \ 1\ \ 0.1\,]^T$ and velocity $v(0) = [\,4\ \ 2\ \ 0.1\ \ -2.1\ \ 1\ \ -0.5\,]^T$.

The nonlinear functions are selected as follows: $f(t) = -\sin(x) - 0.25\ \sin(v) + 1.5\ \cos(2.5t)$.

Choose the simulation duration as $T = 20$ s. The values of the parameters in the control protocol are taken as follows: $\alpha = 1.5$, $\beta = 1.5$, $k = 0.8$, and $\tau = 0.06$.

Figure 2 shows the simulation results when finite-time control protocol (7) is adopted.

It can be observed that the position and the velocity of the agents to reach a consensus takes about 9 s, which indicates that the consensus protocol is effective and the finite-time consensus for the time-delay case is achieved.

The state error of the leaderless system is defined as in (9) and (10).

Figure 3 shows the convergence of the state errors. It can be seen that the position and the velocity error converge to zero in a finite time.

Within the same context, general consensus protocol with communication time delay based on (5) could be designed as

$$u_i(t) = -\alpha\left(\sum_{j=1}^{n} L_{ij}x_i(t-\tau)\right) - \beta\left(\sum_{j=1}^{n} L_{ij}v_i(t-\tau)\right). \tag{30}$$

The consensus control protocol of (30) can make the leaderless multiagent system described in (2) achieve consensus asymptotically for the time-delay case. This was proved in [17].

Figure 4 shows the position and velocity state error when consensus protocol (30) is adopted.

It can be observed that the state of the agents can achieve consensus around 12 s, which is about 3 s slower than adopted protocol (7) proposed in Section 3.2.

Compared with the simulation result in Figure 3, it can be seen that the finite-time control protocol designed in this paper can accelerate the convergence speed of the system.

Define the MAE of the agents over the time interval $[0, N]$.

$$\begin{aligned} MAE\_x_i &= \frac{1}{N}\sum_{k=1}^{N}\left|x_i - x_0\right|, \\ MAE\_v_i &= \frac{1}{N}\sum_{k=1}^{N}\left|v_i - v_0\right|, \end{aligned} \tag{31}$$

where N represents the running time of the system in the simulation. For example, when the simulation duration is taken as T and the time interval is taken as T_s, then we have $N = T/T_s + 1$.

By calculating the mean absolute error (MAE) of the state of the agents, the performance of the algorithm with different consensus protocols (7) and (30) can be evaluated numerically.

According to Table 1, it can be intuitively seen that when finite-time protocol (7) is adopted, the MAE of the multiagent system is smaller, and the state error can converge to zero in a faster time.

Example 2. Consider a leader-following MAS described by equations (2) and (24). The communication topology is shown in Figure 5.

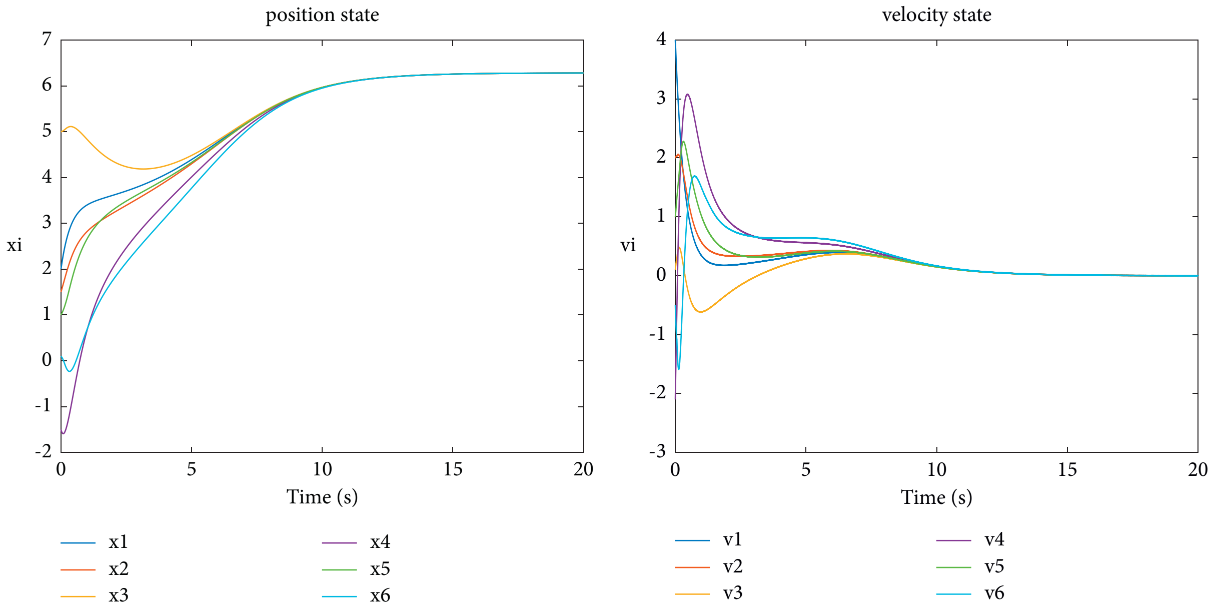
Choose the initial states of each agent in the multiagent system as follows: $x_1(0) = [\,0.2\ \ 0.4\ \ 0.5\,]^T$, $x_2(0) = [\,1.5\ \ 2\ \ 1.2\,]^T$, $x_3(0) = [\,0.5\ \ 0.1\ \ 0.3\,]^T$, $x_0(0) = [\,1\ \ 1\ \ 1\,]^T$; $v_1(0) = [\,-0.2\ \ 0.4\ \ 1\,]^T$, $v_2(0) = [\,1\ \ 0.2\ \ 0.8\,]^T$, $v_3(0) = [\,0.4\ \ 0.5\ \ 1.2\,]^T$, and $v_0(0) = [\,0.5\ \ 1\ \ 1\,]^T$. The state of each agent has a three-order component.

The nonlinear functions are selected as

$$f(t, x_i, v_i) = 0.5 * \begin{pmatrix} 9v_{i2}(t) - \dfrac{18}{7}v_{i1}(t) + \dfrac{24}{7}\left(\left|v_{i1}(t) + 1\right| - \left|v_{i1}(t) - 1\right|\right) \\ v_{i1}(t) - v_{i2}(t) + v_{i3}(t) \\ -\dfrac{100}{7}v_{i2}(t) \end{pmatrix}. \tag{32}$$

Similar to Example 1, choose the simulation duration as $T = 20$ s, and the values of the parameters are taken as follows: $\alpha = 1.5$, $\beta = 1.5$, $k = 0.8$, and $\tau = 0.06$.

Figure 6(a) shows the position of the agents when finite-time control protocol (25) is adopted. Figure 6(b) shows the convergence of the position error. The position error converges to zero in a finite time.



(a)

(b)

Figure 2: The state of each agent. (a) Position state of each agent. (b) Velocity state of each agent.

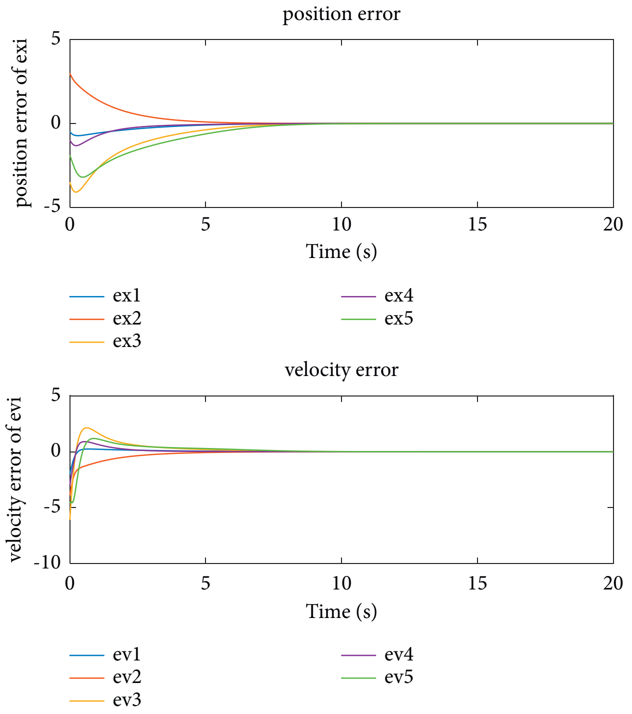


Figure 3: Position error and velocity error.

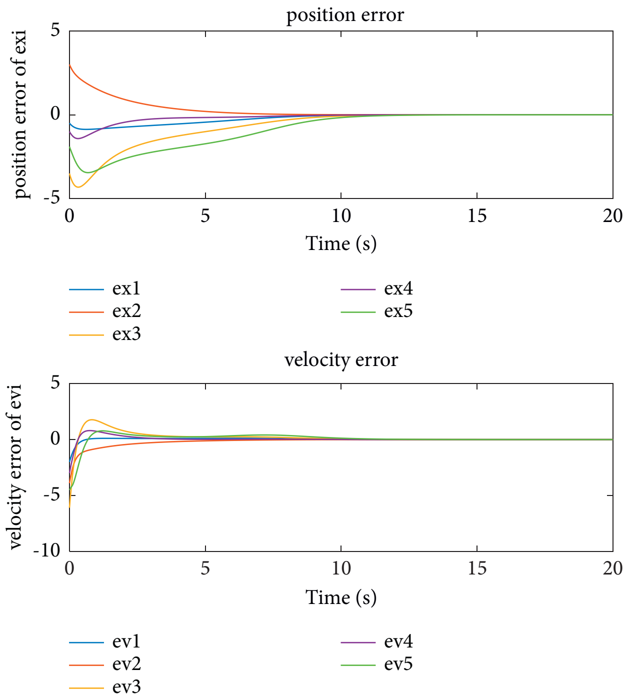


FIGURE 4: Position and velocity state error.

TABLE 1: The mean absolute error of the two methods.

Methods	MAE_x_i	MAE_v_i
With finite-time protocol (7)	0.9127	0.4984
With general protocol (30)	1.6477	0.5828

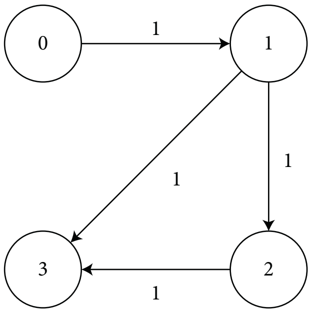
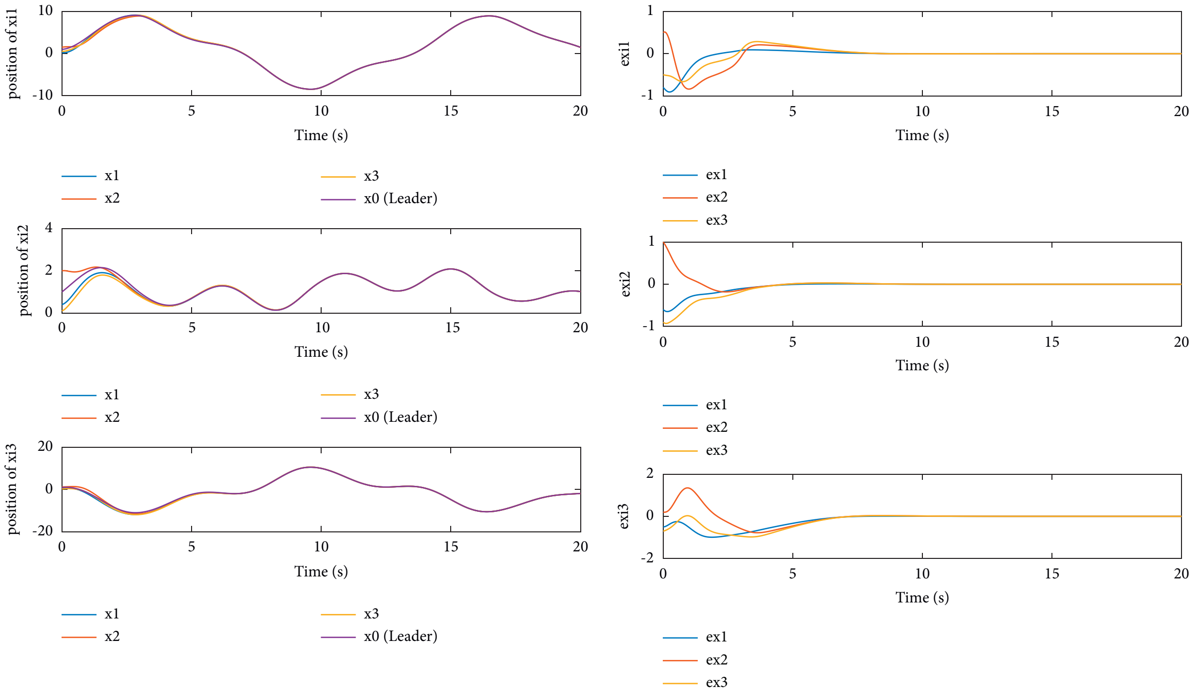


FIGURE 5: Topology of the graph G.

It can be observed that the position of the agents to reach a consensus takes about 8 s, which indicates that the consensus protocol is effective for the leader-following system with time delay.

Under the structure of the leader-following system, general consensus protocol with communication time delay based on (5) could be designed as

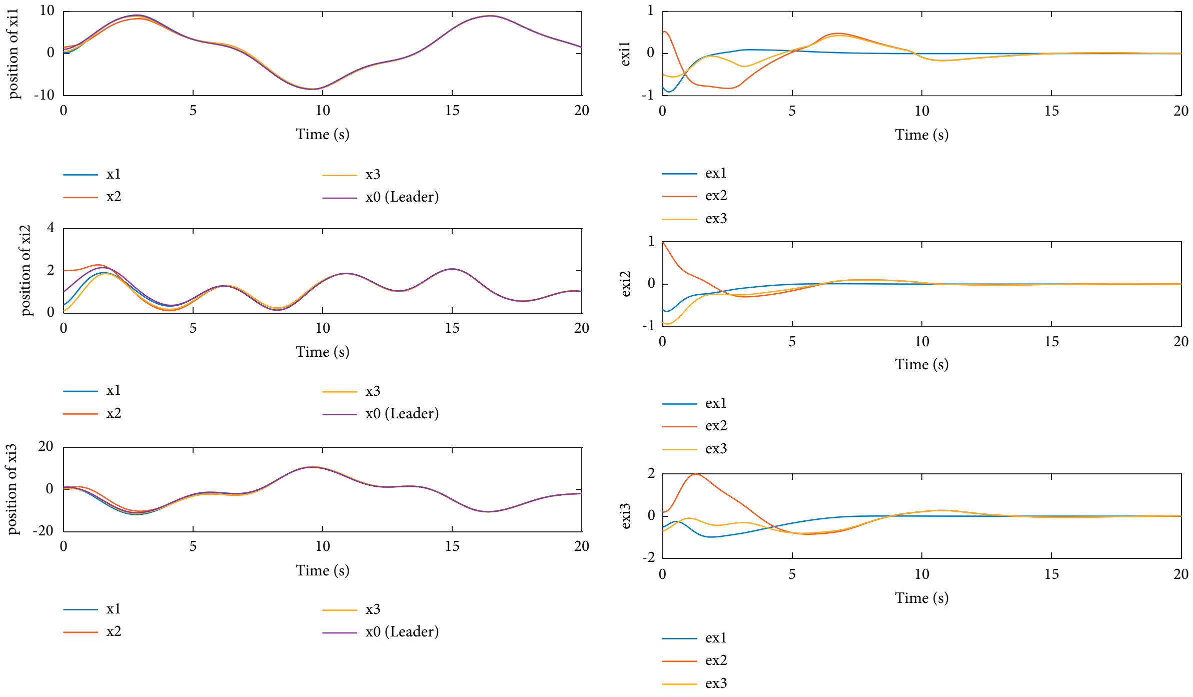
$$u_i(t) = -\alpha\left[\sum_{j=1}^{n} L_{ij} x_j(t-\tau)\right] - \beta\left[\sum_{j=1}^{n} L_{ij} v_j(t-\tau)\right] - b_i\left[x_i(t-\tau) - x_0(t-\tau) + v_i(t-\tau) - v_0(t-\tau)\right]. \qquad (33)$$



(a)

(b)

FIGURE 6: Position state and position error. (a) Position state. (b) Position error.



(a)

(b)

FIGURE 7: Position state and position error. (a) Position state. (b) Position error.

Table 2: The mean absolute error of the two methods.

Methods	MAE					
	x_{i1}	x_{i2}	x_{i3}	v_{i1}	v_{i2}	v_{i3}
With finite-time protocol (25)	0.2345	0.1249	0.4507	0.2176	0.0792	0.3968
With general protocol (33)	0.3755	0.1897	0.5314	0.2434	0.1099	0.4289

Figures 7(a) and 7(b) show the position state and error when consensus protocol (33) is adopted. Compared with Figures 6(a) and 6(b), note that the state of the agents can achieve consensus around 15 s, which is about 7 s slower than adopted protocol (25).

The simulation results of the agent here only give the position state and position error, and the velocity is similar, which is not given here.

We calculate the mean absolute error (MAE) in this case.

Table 2 shows the MAE of the three-order components of the agent's position state and velocity state. When finite-time protocol (25) is adopted, the MAE is smaller and the state error can converge to zero in a faster time.

5. Conclusions

In this paper, the consensus problem with fixed time delay is studied, and finite-time control protocols for the time-delay case are proposed. Two different structures of leaderless and leader-following multiagent systems have been discussed in this paper. Simulation result shows that the position and speed states of all agents can converge to the same in a fast time. Compared with the general consensus protocol, the proposed method can accelerate the convergence speed of the system and make the system more reliable. However, there are still many problems that remain to be solved, such as how to calculate the maximum delay limit of consensus. In addition, the fixed time delay is considered in this paper, and the future research on time-varying delay also presents greater challenges.

References

[1] Z. Li, W. Ren, X. Liu, and L. Xie, "Distributed consensus of linear multi-agent systems with adaptive dynamic protocols," *Automatica*, vol. 49, no. 7, pp. 1986–1995, 2013.

[2] H. Li, X. Liao, and G. Chen, "Leader-following finite-time consensus in second-order multi-agent networks with nonlinear dynamics," *International Journal of Control, Automation and Systems*, vol. 11, no. 2, pp. 422–426, 2013.

[3] F. Sun and Z.-H. Guan, "Finite-time consensus for leader-following second-order multi-agent system," *International Journal of Systems Science*, vol. 44, no. 4, pp. 727–738, 2013.

[4] Y. Wu, J. Hu, L. Xiang, Q. Liang, and K. Shi, "Finite-time output regulation of linear heterogeneous multi-agent systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2021, In press.

[5] J. Wang, Z. Zhou, C. Wang, and Z. Ding, "Cascade structure predictive observer design for consensus control with applications to UAVs formation flying," *Automatica*, vol. 121, 2020.

[6] H. Zhang, X. Liu, H. Ji, Z. Hou, and L. Fan, "Multi-agent-based data-driven distributed adaptive cooperative control in urban traffic signal timing," *Energies*, vol. 12, no. 7, p. 1402, 2019.

[7] D. R. Aleko and D. Soufiene, "An efficient adaptive traffic light control system for urban road traffic congestion reduction in smart cities," *Information*, vol. 11, no. 2, p. 119, 2020.

[8] M. Kiani, S. H. Pourtakdoust, and A. A. Sheikhy, "Consistent calibration of magnetometers for nonlinear attitude determination," *Measurement*, vol. 73, pp. 180–190, 2015.

[9] R. Haghighi and C. K. Pang, "Robust concurrent attitude-position control of a swarm of underactuated Nanosatellites," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 1, pp. 77–88, 2018.

[10] M. L. Minsky, *The Society of Mind*, pp. 371–396, Simon & Schuster, New York, NY, USA, 1988.

[11] A. Jadbabaie, J. Jie Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[12] J. A. Fax and R. M. Murray, "Information flow and cooperative control of vehicle formations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1465–1476, 2004.

[13] W. Zou, C. K. Ahn, and Z. Xiang, "Fuzzy-approximation-based distributed fault-tolerant consensus for heterogeneous switched nonlinear multiagent systems," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 10, pp. 2916–2925, 2021.

[14] P. Li and J. Hu, "An ADMM based distributed finite-time algorithm for economic dispatch problems," *IEEE Access*, vol. 6, pp. 30969–30976, 2018.

[15] H. Du, G. Wen, G. Chen, J. Cao, and F. E. Alsaadi, "A distributed finite-time consensus algorithm for higher-order leaderless and leader-following multiagent systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1625–1634, 2017.

[16] L. Long Wang and F. Feng Xiao, "Finite-time consensus problems for networks of dynamic agents," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 950–955, 2010.

[17] B. Chen, J. Hu, Y. Zhao, and B. Kumar Ghosh, "Finite-time velocity-free observer-based consensus tracking for

heterogeneous uncertain AUVs via adaptive sliding mode control," *Ocean Engineering*, vol. 237, Article ID 109565, 2021.

[18] W. Zou, "Finite-time consensus of second-order switched nonlinear multi-agent systems," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 5, pp. 1757–1762, 2019.

[19] X. Li, P. Shi, Y. Wang, and S. Wang, "Cooperative tracking control of heterogeneous mixed-order multiagent systems with higher-order nonlinear dynamics," *IEEE Transactions on Cybernetics*, 2020, In press.

[20] W. Zou, K. Qian, and Z. Xiang, "Fixed-time consensus for a class of heterogeneous nonlinear multiagent systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 7, pp. 1279–1283, 2019.

[21] J. Hu, "On robust consensus of multi-agent systems with communication delays," *Kybernetika*, vol. 45, no. 5, pp. 768–784, 2009.

[22] Y. Huang, Y. Li, and W. Hu, "Distributed rotating formation control of second-order leader-following multi-agent systems with nonuniform delays," *Journal of the Franklin Institute*, vol. 356, no. 5, pp. 3090–3101, 2019.

[23] S. Li, X. Peng, Y. Tang, and Y. Shi, "Finite-time synchronization of time-delayed neural networks with unknown parameters via adaptive control," *Neurocomputing*, vol. 308, no. 9, pp. 65–74, 2018.

[24] W. Li, Z. Chen, and Z. Liu, "Leader-following formation control for second-order multiagent systems with time-varying delay and nonlinear dynamics," *Nonlinear Dynamics*, vol. 72, no. 4, pp. 803–812, 2013.

[25] W. Hou, M. Fu, H. Zhang, and Z. Wu, "Consensus conditions for general second-order multi-agent systems with communication delay," *Automatica*, vol. 75, no. 6, pp. 293–298, 2017.

[26] Y. Wang, J. Cao, H. Wang, and F. E. Alsaadi, "Event-triggered consensus of multi-agent systems with nonlinear dynamics and communication delay," *Physica A: Statistical Mechanics and Its Applications*, vol. 522, pp. 147–157, 2019.

Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm

Sanam Devi, *Department of Electrical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sanamdevi226@yahoo.co.in*

Achyutananda Panda, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, achutanada.panda23@gmail.com*

Sandip Kar Mazumdar, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, sk_mazumdar1@gmail.com*

Prajnadipta Sahoo, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, prajnadipta.sahoo@yahoo.co.in*

Abstract

Electricity can be provided to small-scale communities like commercial areas and villages through microgrid, one of the small-scale, advanced, and independent electricity systems out of the grid. Microgrid is an appropriate choice for specific purposes reducing emission and generation cost and increasing efficiency, reliability, and the utilization of renewable energy sources. The main objective of this paper is to elucidate the combined economic emission dispatch CEED problem in the microgrid to attain optimal generation cost. A combined cost optimization approach is examined to minimize operational cost and emission levels while satisfying the load demand of the microgrid. With this background, the authors proposed a novel improved mayfly algorithm incorporating Levy flight to resolve the combined economic emission dispatch problem encountered in microgrids. The islanded mode microgrid test system considered in this study comprises thermal power, solar-powered, and wind power generating units. The simulation results were considered for 24 hours with varying power demands. The minimization of total cost and emission is attained for four different scenarios. Optimization results obtained for all scenarios using IMA giv e a comparatively better reduction in system cost than MA and other optimization algorithms considered revealing the efficacy of IMA taken for comparison with the same data. The proposed IMA algorithm can solv e the CEED problem in a grid-connected microgrid.

1. Introduction

Microgrid is one of the advanced small-scale centralized electricity systems and it usually contains energy storage resources, Distributed Generation (DG) units, and loads. Microgrids are generally designed and installed nearby energy consumers in a confined community [1]. But there has been a drastic increase in recent years regarding installing renewable energy sources in microgrids, owing to environmental advantages and low cost compared to its counterparts [2]. Microgrid meets different load requirements in residential, commercial, agriculture, and industrial sectors [3]. Microgrids can function under two different modes: grid-connected and islanded modes. In the former mode, the microgrid and main grid are connected, while in the latter the microgrid is isolated from the main grid during an emergency outbreak. It continues its power delivery functions to local loads as usual [4]. There are loads of advantages present in using microgrid reduction of carbon emission and generation cost, thanks to renewable energy sources, and increased reliability and power quality, etc. [5].

Optimal operations and effective planning of electric power generation systems are the two most crucial elements in electric industries. Controlling and operating power systems, cost-efficient load dispatch (Economic Load Dispatch, i.e., ELD) related problems are too much concerning to address [6]. Power system optimization problems that employ ELD are useful in identifying the most suitable, cheap, and seamless operations with the regulation of outputs produced by different power generation units that meet the load demands. ELD has a primary aim to mitigate the overall cost incurred upon power generation without compromising or producing any constraints [7]. ELD determines active power output generated by different power generation systems to attain the objective functions and simultaneously overcome many problems [8]. It is important to develop novel power management algorithms to translate the microgrid as a viable and happening choice compared to traditional power systems [9]. The mitigation of both power generation costs and the emission of environmental pollutants is the sole aim of utility operators. For these goals, two contradictory objectives must be considered. Combined economic emission dispatch (CEED) is utilized to mitigate emission levels from all generating units and costs incurred by the operating units [9].

CEED problem has been discussed earlier by different authors who proposed several optimization techniques to overcome the issue [10]. In [10], the researchers proposed a balanced trade-off method to resolve the ECED (Environment Constrained Economic Dispatch) problem. The study conducted a first-of-its-kind comparative analysis of three methods: Fractional Programming (FP), ECED, and price penalty factor (PPF) to overcome the CEED problem. Chimp Optimization Algorithm (ChOA) was proposed to address the optimal design of microgrid which comprises PV panels, wind turbines, and battery storage systems [11]. Optimization results had been compared with Improved Grey Wolf Optimizer (IGWO) and Grey Wolf Optimizer (GWO). Optimal sizing of photovoltaic cell and solar water heater by considering environmental parameters and fuel-saving was carried out in [12]. Energy and economic analysis of solar energy-based cogeneration system for a building in Saveh City were studied. The researchers simulated the model in a 3-Unit dynamic test system incorporating renewable energy sources. In the study conducted by Alamoush [13], Bernstein-search differential evolution (BSDE) algorithm was proposed to resolve the Dynamic Combined Heat and Power Economic emission Dispatch (DCHPED) generation problem in microgrid comprised of renewable energy sources, fixed nondeferrable and deferrable loads, fossil-fuel combined heat and power units, and thermal energy storage devices. In [14], whale optimization algorithm (WOA) has been applied to carry out the combined economic emission dispatch problem. The simulation was performed by considering four different load sharing scenarios among the distributed energy resources. The researchers also conducted ANOVA and Wilcoxon signed-rank tests to validate the supreme characteristic of WOA. In [15], Stochastic Fractal Search (SFS) algorithm was applied to resolve multiobjective economic emission dispatch problems that arise in combined heat and power (CHP) generation. This study was conducted in large microgrids in which solar-powered generating units, wind power units, and fossil-fuel-powered generating units were installed.

Collective Neuro Dynamic Optimization (CNO) method was proposed in the earlier study. In this study, the authors combined a heuristic approach and projection neural network (PNN) to optimize the scheduling of an electrical microgrid containing ten thermal generators and mitigate the costs incurred upon emission and generation [16]. A mixed-integer nonlinear programming formulation was proposed in [17] to dispatch the distributed generators cohesively. Further, the model was also aimed at fulfilling the water demands and the building's thermal energy requirements in a standalone water energy microgrid.

In [18], Giza Pyramids Construction (GPC) was proposed to implement the optimal design of an isolated microgrid. Net present cost (NPC), Levelized Cost of Energy (LCOE), loss of power supply probability, and availability index were considered objective functions. Modified adaptive accelerated particle swarm optimization (MAACPSO) algorithm was proposed to investigate the grid-tied PV systems reliability [19]. This study focused on the probability analysis and reliability assessment of the components of grid-tied PV systems through IEEE 24 bus integrated with PV system with four different case studies. A first-of-its-kind Sequential Optimization Strategy (SOS) was formulated in the study conducted earlier to allocate active and reactive power to Dispatchable Distributed Generator (DDG) units in an optimal manner. These DDG units are installed in a droop-controlled islanded AC microgrid [20]. The research proposed improved Quantum Particle Swarm Optimization (QPSO) earlier [21] to address the Short-Term Economic Environmental Dispatch (EED) problem in a microgrid. Recently, a multiobjective seeker optimization algorithm has been proposed to analyze the influence of charging and discharging behaviour of electric vehicles and demand side response resources on the economic functioning of PV-connected microgrid systems. The model considered three objectives: power fluctuation between microgrid and main grid, comprehensive operating cost of the microgrid, and utilization rate of photovoltaic energy [22].

In [23], the researchers proposed a mathematical optimization approach to achieve an optimal operation upon economic dispatch in DC microgrid. To allot the schedules for unit commitment and achieve economic dispatch in microgrid, the study conducted earlier [24] proposed an enhanced real-coded genetic algorithm in the enhanced mixed-integer linear programming (MILP) based method. The authors proposed a stochastic model in [25] to manage CHP-based microgrids optimally. The study took economic, reliability, and environmental aspects into consideration. Nonconvex and nonlinear stochastic problems are used to resolve complexity. The Exchange Market Algorithm (EMA) was proposed in a study. This study considered three contradictory objectives through a weighted sum approach to resolving the multiobjective problem as a single-objective problem.

A multiobjective optimal dispatch model was proposed in [26] for a grid-connected microgrid. In this study, the authors considered reducing environmental protection costs and the generation cost of the microgrid with the incorporation of an enhanced PSO algorithm. In [27], a genetic algorithm is applied to optimize the focal area of the parabolic trough concentration photovoltaic/thermal system. The objective function was considered a combination of electrical efficiency and thermal efficiency. A parabolic trough concentrating photovoltaic thermal (CPVT) was utilized to afford the energy required for a residential building [28]. CPVT was used as a source of heat and cooling and electrical energy for the building and simulation was carried out using TRNSYS software. Samy et al. addressed the problem of power outages in distant districts by taking advantage of the available renewable energy resources in the contiguous environment [29]. Hybrid Firefly and Harmony Search optimization technique (HFA/HS) was implemented to improve the net present cost of the proposed hybrid system which comprises photovoltaic (PV), wind turbine (WT), and fuel cell (FC). MINLP, a novel optimization model, was proposed in [30] to resolve the economic dispatch problem of a microgrid which contains numerous units of wind turbines (WTs), heat-only units, traditional power generators, photovoltaic (PV) systems, CHP units, and battery storage systems under certain uncertainties. In the study conducted earlier [31], MBGSA (Memory-Based Gravitational Search Algorithm) was proposed to overcome the ELD problem. In one of the research investigations [32], a novel Multiobjective Virus Colony Search (MOVCS) was proposed to elucidate the multiobjective dynamic economic emission dispatch (DEED) problem. This model aims to mitigate the emissions produced by fossil-fuel power generators and simultaneously reduce the cost incurred upon wind-thermal electrical energy costs. In [33], HOMER software is used to model and simulate a wind-solar hybrid system independent of the national grid in the northwest of Iran. A multiobjective particle swarm optimization technique is proposed to optimize the sizing of a green energy system connected to a randomly disrupted grid [34]. The energy cost for evaluating hybrid system economies, the loss of probability of power supply (LPSP) for reliability assessments, and the System Surplus Energy Rates (SSER) were considered objective functions for evaluating hybrid system compatibility and efficiency. A sustainable energy distribution configuration for microgrids integrated into the national grid using back-to-back converters in a renewable power system was examined in [35]. Different scenarios of several sustainability schemes of power management in microgrids were analyzed.

SGEO (Social Group Entropy Optimization) technique was proposed in [36] to resolve Fuel Constrained Dynamic Economic Dispatch (FCDED) with Demand Side Management (DSM). The technique combined the pumped hydrostorage plant with renewable energy sources. This research used a stochastic fractal search algorithm to overcome the biobjective combined heat and power economic dispatch (CHPED) problem [37].

In [38], mayfly algorithm (MA) was proposed by Dr. Konstantinos Zervoudakis in 2021. In this paper, an improved mayfly optimization algorithm was investigated with the help of a microgrid model under varying scenarios. The results were contrasted against recent state-of-the-art algorithms that also employed the same microgrid model.

The contributions of the current research paper are summarized herewith.

(i) An improved version of the mayfly optimization algorithm incorporating Levy flight is proposed to elucidate the microgrid test system's CEED problem.

(ii) An improved mayfly optimization algorithm is proposed in addition to Levy flight to overcome the CEED problem encountered in microgrid test system.

(iii) Levy flight has been leveraged in this study since it possesses huge advantages in not engaging local optimal. An optimal trade-off is provided by the proposed algorithm between exploration and exploitation phases.

(iv) The authors validated the supremacy of the proposed algorithm in terms of resolving the CEED problem under two different objective functions that involve advanced energy sources.

(v) Compared with existing population-based optimization tools such as PSO and GA, only a few control parameters exist in IMA. This feature helps in making it the best optimization procedure. IMA-based CEED was authenticated as a unique and robust technique since it incurred less total generation cost than the solution even after conducting multiple random trials.

2. Mathematical Formulation of CEED for Microgrid

2.1. Combined Economic Emission Dispatch (CEED). The simultaneous mitigation of economic and environmental dispatch objective functions remains the primary objective for the CEED problem in the microgrid. In other terms, the total fuel cost must be reduced, while at the same time the emission levels should also be mitigated without compromising the constraints. So, it is suggested to formulate CEED as a single optimization problem as given herewith [9].

$$\min\mathrm{TC} = \sum_{i=1}^{\mathrm{NG}} \{F_i(P_{\mathrm{Gi}}), E_i(P_{\mathrm{Gi}})\}. \tag{1}$$

Here, TC denotes the total operating cost that should be minimized. The fuel cost of the i^{th} generator is represented as $F_i(P_{\mathrm{Gi}})$, the emission level of the i^{th} generator is represented as $E_i(P_{\mathrm{Gi}})$, P_{Gi} denotes the i^{th} generating unit's output power, and finally, NG corresponds to the whole generating unit count.

2.1.1. Minimization of Fuel Cost. In general, the fuel cost function is denoted through the quadratic equation given below [9]:

$$F_t = \sum_{i=1}^{\mathrm{NG}} F_i\left(P_{\mathrm{Gi}}\right) = \sum_{i=1}^{\mathrm{NG}}\left(x_i + y_i P_{\mathrm{Gi}} + z_i P_{\mathrm{Gi}}^2\right). \tag{2}$$

Here, F_t corresponds to overall fuel cost incurred in terms of \$ and $x_i\,(\$/h)$, $y_i\,(\$/\mathrm{MWh})$, and $z_i\,(\$/\mathrm{MW}^2\mathrm{h})$ correspond to the cost coefficients of i^{th} generating unit.

2.1.2. Minimization of Emission. When fossil fuels are used, the generators emit different sorts of pollutants. So, pollution mitigation forms the primary goal of most power system operations. Equation (3) denotes the expression for total emission from [5, 9]:

$$E_t = \sum_{i=1}^{\mathrm{NG}} E_i\left(P_{\mathrm{Gi}}\right) = \sum_{i=1}^{\mathrm{NG}}\left(\alpha_i + \beta_i P_{\mathrm{Gi}} + \gamma_i P_{\mathrm{Gi}}^2\right). \tag{3}$$

Here, E_t denotes the overall emission and α_i (kg/h), $\beta_i\,(\mathrm{kg}/(\mathrm{MWh}))$, $\gamma_i\,(\mathrm{kg}/(\mathrm{MW}^2\mathrm{h}))$ correspond to the i^{th} generating unit's emission coefficients.

2.1.3. Total Generation Cost of CEED Problem. It is possible to convert the dual-objective optimization problem, focusing emission, and fuel cost, into a single-objective optimization problem with the induction of PPF (price penalty factor) as given earlier [9]:

$$\min\mathrm{TC} = F_t + \Lambda \times E_t. \tag{4}$$

Here, Λ denotes the price penalty factor (PPF), which is calculated as a ratio between fuel cost and the emission of the corresponding generating unit (\$/kg). PPFs are of different types, while in the current study, the authors use min-max types sourced from [5, 9] for comparison. Following is the equation for min-max type [9]:

$$\Lambda_i = \frac{F_t\left(P_{\mathrm{Gi}}^{\min}\right)}{E_t\left(P_{\mathrm{Gi}}^{\max}\right)}, \quad i = 1, 2, \ldots, \mathrm{NG}, \tag{5}$$

where $P_{\mathrm{Gi}}^{\max}$ and $P_{\mathrm{Gi}}^{\min}$ Here, Λ denotes the ratio between maximum fuel cost and maximum emission of the corresponding generator in \$/kg [9]. The maximum and minimum output power of the generator combines emission with fuel cost. Afterwards, TC corresponds to the total operating cost of \$.

The following is the list of steps to be followed to determine the price penalty factor for a specific load demand [39].

(i) The ratio between minimum fuel cost and the maximum emission of every generating unit should be determined.

(ii) Price Penalty Factor values are sorted out in ascending order.

(iii) The maximum capacity of every unit ($P_{\mathrm{Gi}}^{\max}$) one is added at a time that starts from the lowest Λ_i, until $\sum P_{\mathrm{Gi}}^{\max} \geq P_D$.

(iv) Then, Λ_i, which has an association with the lowest unit in this process, remains the tentative PPF value (Λ) for the load under consideration.

So, a modified PPF (Λ) is utilized to arrive at the exact value for specific load demand based on the interpolation of Λ values corresponding to their load demand values.

2.2. Cost Functions of Renewable Energy Sources. Across the globe, renewable energy sources are the foremost choice of transmission when energy is produced, compared to traditional generators. In such a scenario, solar and wind power can be denoted as negative loads and can be used to mitigate the total load demand in the system [9]. However, the economic dispatch solution is considered a base to distribute the rest of the load demands on traditional generators. In the current study, the CEED solution for the microgrid takes cost functions of wind and solar-powered generating units into account. From [9], the input data for cost and emission coefficients are considered.

2.2.1. Cost Function of Wind Power Generating Unit. Current economic analysis is conducted for wind-based power generation and the specific cost can be determined with inputs, operation, maintenance, and equipment costs. This cost function is expressed as per [5, 9]:

$$C_w\left(P_w\right) = \left(\frac{r}{\left[1-(1+r)^{-N}\right]} I^p + O^E\right) P_w. \tag{6}$$

Here, P_w denotes the wind power produced in terms of kW, r corresponds to the interest rate, a denotes the Annuitization coefficient, N corresponds to lifetime investment in terms of years, and I^p and O^E correspond to the costs incurred upon investment per unit installed power (\$/kW) and operating and maintenance costs per unit installed power $(\$/kW)$, respectively.

The 24-hour data for the wind power generating unit is considered from [5]. The parameters required for wind power cost function are chosen from [5, 9].

2.2.2. Cost Function of Solar Power Generating Unit. Similar to wind power, solar-powered generating unit's cost function is expressed as in [5, 9]:

$$C_s\left(P_s\right) = \left(\frac{r}{\left[1-(1+r)^{-N}\right]} I^p + O^E\right) P_s, \tag{7}$$

where P_s is the output power from the solar-powered generating unit, r corresponds to interest rate, N denotes lifetime investment in terms of years, and I^p and O^E correspond to investment costs made upon per unit installed power (\$/kW) and operating and maintenance costs per unit installed power (\$/kW), respectively.

The 24-hour data of the solar power generating unit is considered [5]. The parameters required for solar power cost function are chosen from [5, 9].

2.3. Total Cost of CEED for Microgrid. The equation for the total CEED cost for the microgrid is shown below. This value gets minimized based on the cost functions regarding wind and solar power [9].

$$C_T = F_t + \Lambda \times E_t + \left(\frac{r}{\left[1-(1+r)^{-N}\right]} I^P + O^E \right) P_w + \left(\frac{r}{\left[1-(1+r)^{-N}\right]} I^P + O^E \right) P_s. \tag{8}$$

2.4. Constraints. The researcher considered both generator capacity and power balance constraints to contrast with existing optimization algorithms.

2.4.1. Islanded Mode of Microgrid. The current study considered islanded mode microgrid to compare with the optimization results achieved in [5, 9]. There is no trade-off for the power between the main grid and the microgrid in this mode. So, the microgrid needs to fulfil the local or confined community load demands.

2.4.2. Power Balance Constraint. The load demand must be equal to that of the total power generation [9].

$$\sum_{i=1}^{\mathrm{NG}} P_{\mathrm{Gi}} + P_w + P_s = P_D. \tag{9}$$

Here, P_D corresponds to the total load demand.

2.4.3. Generation Capacity Constraints. Every generating unit's output power gets flanked by both lower and upper bounds [9].

$$P_{\mathrm{Gi}}^{\min} \le P_{\mathrm{Gi}} \le P_{\mathrm{Gi}}^{\max}. \tag{10}$$

Here, $P_{\mathrm{Gi}}^{\min}$ and $P_{\mathrm{Gi}}^{\max}$ correspond to the minimum and maximum output powers of the i^{th} generating unit correspondingly.

3. Improved Mayfly Algorithm

Mayfly algorithm takes its inspiration from the social behaviour of mayflies, especially how they mate with each other [38]. It is assumed that mayflies are instantly considered adults as soon as the eggs are hatched. Leaving beside the period of their life, only the fittest mayflies tend to survive. Each mayfly has a position in search space that corresponds to a solution that overcomes the problem. RAND functions are utilized in conventional mayfly algorithm to produce novel variables that lead to local optimal. To increase MA's searching ability and create an optimal solution, the researchers integrated MA with Levy flight. If a Levy flight-based approach is utilized for system identification, it achieves rapid convergence and does not entail derivative information [40], attributed to stochastic random search, in line with the Levy flight concept [41]. Levy flight contributes heavily to increasing the optimal solution's local search avoidance and local trapping [42]. The flowchart of the proposed IMA algorithm is shown in Figure 1.

The steps required for the proposed mayfly optimization algorithm works are described as follows:

Step 1. Two mayfly sets, each representing a male and a female population, should be generated randomly. Then, every mayfly is arbitrarily placed in problem space as a candidate solution which is denoted by a d-dimensional vector $P_{\mathrm{Gi}} = (P_{\mathrm{G1}}, \ldots, P_{\mathrm{Gd}})$. Then the performance is assessed based on the predefined objective function $f(C_T(P_{\mathrm{Gi}}))$.

Step 2. A mayfly's velocity $\mathbf{v} = (v_1, \ldots, v_d)$ is initialized through its positional change. Its direction is decided as a hybrid interaction between individuals and the social flying experiences. To be specific, every mayfly tends to alter its trajectory in alignment with its personal best position (pbest) so far. It also alters based on the best position achieved by any other mayfly present in the swarm so far (gbest).

Step 3. The population of the male mayflies is initialized as $P_{\mathrm{Gmi}}\,(i = 1, 2, \ldots, \mathrm{NG})$ with velocities v_{mi}. The male mayflies, gathered in swarms, denote that the position of every mayfly gets altered in alignment with its individual's experience and that of the neighbor's. P_{Gi}^t is assumed to be the current position of mayfly i in search space at time step t and the position gets altered with the addition of velocity v_i^{t+1}, to the current position. This notation is formulated as given herewith.

$$P_{\mathrm{Gmi}}^{t+1} = P_{\mathrm{Gmi}}^{t} + v_i^{t+1}. \tag{11}$$

Male mayflies are considered as present a few meters above the water, with $P_{\mathrm{Gim}}^0\, U(P_{\mathrm{Gmmin}}, P_{\mathrm{Gmmax}})$, performing nuptial dance. It can be assumed that these mayflies lack great speeds due to constant movements. This results in the calculation of a male mayfly's velocity i as follows [38]:

$$\begin{aligned} v_{\mathrm{ij}}^{t+1} &= g * v_{\mathrm{ij}}^{t} + a_1 e^{-\beta r_p^2}\left(\mathrm{pbest}_{\mathrm{ij}} - P_{\mathrm{Gmij}}^t\right) \\ &\quad + a_2 e^{-\beta r_g^2}\left(\mathrm{gbest}_j - P_{\mathrm{Gmij}}^t\right). \end{aligned} \tag{12}$$

Here, v_{ij}^t corresponds to mayfly i's velocity in dimension $j = 1, \ldots, n$ at time step t, P_{Gmij}^t denotes the mayfly's i^{th} position in dimension j at time step t, a_1 and a_1 correspond to positive attraction constants utilized in scaling up the contribution of cognitive and social components,

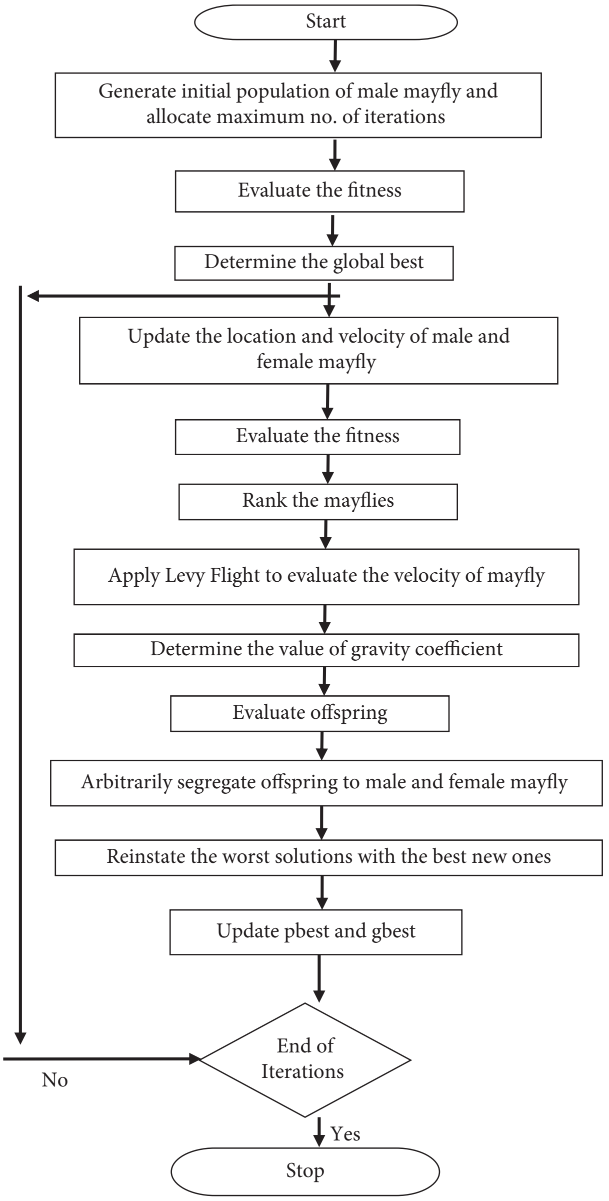

Start
Generate initial population of male mayfly and allocate maximum no. of iterations
Evaluate the fitness
Determine the global best
Update the location and velocity of male and female mayfly
Evaluate the fitness
Rank the mayflies
Apply Levy Flight to evaluate the velocity of mayfly
Determine the value of gravity coefficient
Evaluate offspring
Arbitrarily segregate offspring to male and female mayfly
Reinstate the worst solutions with the best new ones
Update pbest and gbest
End of Iterations
No
Yes
Stop

Figure 1: Flowchart of the proposed improved mayfly optimization algorithm.

respectively. Furthermore, pbest_i denotes the mayfly i^{th} best position which it had ever visited. Based on the minimization problems under consideration, the personal best position pbest_{ij} at the next time step $t + 1$ is determined as given herewith.

$$\text{pbest}_i = \begin{cases} P_{\text{Gmi}}^{t+1}, \text{iff}\left(P_{\text{Gmi}}^{t+1}\right) < f(\text{pbest}_i) \\ \text{is kept the same, otherwise} \end{cases}. \tag{13}$$

Following is the equation for the global best position gbest at time step t.

$$\begin{aligned} &\text{gbest} \in \{\text{pbest}_1, \text{pbest}_2, \ldots, \text{pbest}_N, | f(\text{cbest})\}, \\ &= \min\{f(\text{pbest}_1), f(\text{pbest}_2), \ldots, f(\text{pbest}_{\text{NG}})\}. \end{aligned} \tag{14}$$

Here β represents the fixed visibility coefficient used in (7). It is utilized to confine the visibility of the mayfly to others. Further, r_p denotes the Cartesian distance between P_{Gi} and pbest_i and r_g corresponds to the Cartesian distance between P_{Gi} and gbest. Following is the equation used to determine these distances.

$$\left\| P_{\text{Gmi}} - X_i \right\| = \sqrt{\sum_{j=1}^{n} \left(P_{\text{Gmij}} - X_{\text{ij}} \right)^2}, \tag{15}$$

where P_{Gmij} corresponds to the j^{th} element of mayfly i and X_i denotes the pbest_i or gbest. If the algorithm needs to function appropriately, then the best mayflies present in the swarm must continuously perform the up-and-down nuptial dance. So, the velocity of these best mayflies must be kept on changing which is calculated as follows [38]:

$$v_{\text{ij}}^{t+1} = v_{\text{ij}}^{t} + d \times r. \tag{16}$$

Here, d denotes the coefficient of nuptial dance whereas the random value in the range of $[-1, 1]$ is denoted by r.

Step 4. In this step, the female mayfly population is initialized $P_{\text{Gfi}} (i = 1, 2, \ldots, \text{NG})$ with velocities v_{fi}. Female mayflies tend not to gather as a swarm alike males. Instead, it tends to fly towards its male counterparts for mating. P_{Gfi}^{t} is assumed as the current position of female mayfly i in search space at time step t, while its position gets altered with the addition of velocity v_i^{t+1} to the current position, i.e.,

$$P_{\text{Gfi}}^{t+1} = P_{\text{Gfi}}^{t} + v_i^{t+1}. \tag{17}$$

Here, due to $P_{\text{Gfi}}^{0} U(P_{\text{Gfmin}}, P_{\text{Gfmax}})$ one cannot randomize the attraction process. So, the model is decided to be a deterministic process. As a result, their velocities are determined as given herewith in the presence of minimization problems [38].

$$v_{\text{ij}}^{t+1} = \begin{cases} g * v_{\text{ij}}^{t} + a_2 e^{-\beta \text{r}_{\text{mf}}^2} \left(P_{\text{Gmij}}^{t} - P_{\text{Gfij}}^{t} \right), \text{ if } f(P_{\text{Gfi}}) > f(P_{\text{Gmi}}), \\ g * v_{\text{ij}}^{t} + fl \times r, \text{ if } f(P_{\text{Gfi}}) \leq f(P_{\text{Gmi}}). \end{cases} \tag{18}$$

Here, v_{ij}^{t} corresponds to the female mayfly's velocity i in dimension $j = 1, \ldots, n$ at time step t, P_{Gfij}^{t} denotes the female mayfly i's position in dimension j at time step t, and a_2 denotes the positive attraction constant whereas it remains a fixed visibility coefficient. Further, the gravity coefficient is denoted by g, and r_{mf} corresponds to the Cartesian distance between male and female mayflies. Here fl corresponds to a random walk coefficient and r denotes the random value in the range of $[-1, 1]$. This value is determined based on (15).

Step 5. In this step, the Levy flight approach is involved in calculating the velocity of a mayfly candidate solution. Equation (19) is used to determine the velocity of the mayfly candidate solution [38].

$$v_{\mathrm{ij}}^{t+1} = \begin{cases} V_{\max}, & \text{if } v_{\mathrm{ij}}^{t+1} > V_{\max}, \\ -V_{\max}, & \text{if } v_{\mathrm{ij}}^{t+1} < -V_{\max}. \end{cases} \tag{19}$$

This stage uses the Levy flight approach to alter the position of the global finest component. Though the Levy flight method has been used for exploration purposes so far, it is associated with a specific search.

Here, $V_{\max}$ is calculated as follows:

$$V_{\max} = \mathrm{Levy}(\lambda) * \left(P_{\mathrm{Gmmax}} - P_{\mathrm{Gmmin}}\right). \tag{20}$$

Here, δ corresponds to a scale factor designed in alignment with the search space element. The author fixed δ as 1.

$$\mathrm{Levy}(\lambda) = 0.01\,\frac{r_5\sigma}{\left|r_6\right|^{1/\beta}}. \tag{21}$$

Further, σ is calculated as follows [42]:

$$\sigma = \left[\frac{\Gamma(1+\lambda)\sin(\pi(\lambda/2))}{(\Gamma((1+\lambda)/2)\lambda[2^{(\lambda-1)/2}])}\right]^{1/\lambda}. \tag{22}$$

Here, $\Gamma(\mathrm{x}) = (x-1)!$, r_5 corresponds to r_6 indiscriminate numbers that lie in the range of [0, 1], and $1 < \beta \leq 2$, where there is a constant value, i.e., 1.5 incorporated for β in the current study [40–42].

$\mathrm{Levy}(\lambda)$ denotes the step length, incorporated by Levy distribution with infinite variance and mean values with $1 < \lambda < 3$. λ corresponds to the distribution factor, whereas the gamma distribution function is denoted by $\Gamma(.)$.

Step 6. Gravity coefficient value calculation [38]:

Gravity coefficient g value can be considered a fixed number that lies in (0, 1].

$$g = g_{\max} - \frac{g_{\max} - g_{\min}}{\mathrm{iter}_{\max}} \times \mathrm{iter}, \tag{23}$$

where $g_{\max}$, $g_{\min}$ correspond to maximum and minimum values which can be taken for the gravity coefficient, and *iter* denotes the algorithm's current iteration, whereas the maximum count of iterations is denoted by $\mathrm{iter}_{\max}$.

Step 7. Mayflies are mated and the offspring are evaluated.

The mating process between the mayflies is discussed by the crossover operator as given herewith. From the male and female population, each one parent is selected through the same selection process, i.e., the attraction of females towards the males. Specifically, fitness function-based or random selection of the parents can be made. In terms of the fitness function, the best female mates with the best male, the second-best female with the second-best male, etc. This crossover results in two offsprings for which the formulation is given herewith [38]:

$$\begin{aligned} \mathrm{offspring1} &= L \times \mathrm{male} + (1-L) \times \mathrm{female}, \\ \mathrm{offspring2} &= L \times \mathrm{female} + (1-L) \times \mathrm{male}. \end{aligned} \tag{24}$$

Here, *male* denotes the male parent, and female corresponds to the female parent, while L is a random value within a specific range. The initial velocity of the offspring is fixed as zero.

3.1. The Pseudocode of the Improved Mayfly Algorithm. The pseudocode of the improved mayfly algorithm is devised as follows:

```
Formulate the objective function $f(C_T(P_{\mathrm{Gi}}))$, $P_{\mathrm{Gi}} = (P_{\mathrm{Gi1}}, \ldots, P_{\mathrm{Gi}d})^T$
Set the male mayfly population $P_{\mathrm{Gmi}}(i = 1, 2, \ldots, \mathrm{NG})$ and velocities $v_{mi}$
Set the female mayfly population $P_{\mathrm{Gfi}}(i = 1, 2, \ldots, \mathrm{NG})$ and velocities $v_{fi}$
Evaluate solutions
Determine global best gbest
Do While stopping criteria are not meet
    Update velocities and solutions of males and females
    Evaluate solutions
    Rank the mayflies
    Apply Levy flight approach to evaluate the velocity of a mayfly candidate solution
    Determine the value of the gravity coefficient
    Mate the mayflies
    Evaluate offspring
    Separate offspring to male and female randomly
    Reinstate worst solutions with the best new ones
    Update pbest and gbest
End While
Post-process results and visualization
```

4. Results and Discussion

The current study used a microgrid model with three conventional generators, solar, and wind units. One of the generators is a combined heat and power generator, whereas the other two conventional generators are synchronous. As per [9], three conventional generators and their daily load profile details were used in the current study. During the improved mayfly optimization algorithm implementation, various parameters were chosen for the optimal search process.

In order to assess the proposed IMA, various scenarios were considered as given herewith.

(i) All sources included

(ii) Thermal power generating units without renewable sources

(iii) Thermal power generating units with wind source only

(iv) Thermal power generating units with solar source only

Table 1: Optimal generation schedule of microgrid for case 1.

Time (h)	RGM cost ($/h) [5]	ACO cost ($/h) [5]	CSA cost ($/h) [5]	ISA cost ($/h) [5]	IHS cost ($/h) [9]	IAHS cost ($/h) [9]	MHS cost ($/h) [9]	MA ($/h)	IMA ($/h)
1	8529	7250	7153	7153	7090.5	7058.0	6942.8	6857.4	6824.6
2	8648	7511	7203	7203	7151.1	7130.2	7010.3	6904.7	6891.4
3	8675	7704	7278	7278	7170.8	7151.5	7100.7	7015.6	6994.8
4	8795	7742	7280	7285	7159.6	7130.8	7049.6	7020.3	7004.8
5	8758	8211	7545	7545	7528.2	7450.1	7377.2	7334.1	7309.7
6	8848	8459	7723	7679	7600.1	7572.2	7553.3	7544.2	7537.6
7	8964	8406	7457	7457	7444.2	7423.8	7294.1	7207.3	7189.5
8	9308	7923	7138	7138	7051.0	7050.3	6935.6	6879.7	6851.3
9	9609	9040	7731	7731	7660.3	7640.9	7576.4	7528.1	7505.8
10	10049	9599	7920	7937	7851.5	7845.4	7770.8	7752.6	7731.2
11	11520	11184	9231	9231	9152.0	9150.0	9073.4	9025.3	9006.8
12	12098	11616	9470	9470	9394.3	9381.3	9314.3	9271.4	9253.7
13	10676	10320	8482	8482	8400.3	8374.4	8326.2	8297.6	8273.9
14	9982	9707	8186	8186	8135.4	8119.9	8025.4	7758.9	7729.4
15	9569	9351	8154	8159	8100.6	8090.5	7984.4	7903.2	7892.7
16	9030	8469	7622	7626	7550.5	7539.6	7457.9	7419.7	7401.4
17	8872	8189	7526	7525	7470.6	7440.2	7362.6	7305.8	7291.3
18	9273	9061	8132	8131	8050.8	8040.4	7956.6	7904.3	7889.5
19	9990	9852	8652	8636	8549.6	8511.0	8462.3	8445.7	8436.1
20	12646	11897	9846	9811	9760.6	9710.0	9690.9	9681.4	9675.8
21	11496	11101	9383	9383	9249.9	9219.7	9221.6	9217.8	9216.1
22	9534	9488	8371	8370	8300.8	8281.4	8194.5	8146.7	8122.3
23	8667	8077	7572	7572	7463.7	7440.1	7403.1	7389.6	7378.4
24	8517	7498	7254	7262	7225.8	7195.7	7070.8	6991.3	6973.5
Total	232053	217655	192309	192250	190512.3	189947.5	188154.4	186802.7	186381.6

4.1. Case 1: All Sources Included. In the first case, the proposed IMO was utilized in the elucidation of the CEED problem in the microgrid and the case considered both wind and solar energy-powered generators. The optimization results were obtained using the proposed IMA and contrasted with the results obtained from other optimization algorithms. The generation cost calculated by the proposed IMA and other published methods is shown in Table 1 for comparative purposes. Figure 2 shows the convergence characteristics required to mitigate the total generation cost incurred from MA and IMA algorithms. As shown in Figure 2, when cross-verifying the proposed algorithm's cost convergence characteristic, a quicker and more smooth transition was obtained than other optimization techniques considered. Further, Figure 3 shows the comparison results of total generation cost saving from CEED problem when using the mayfly algorithm and other such optimization algorithms. The results achieved from the simulation reveal that the proposed IMA is superior to MA and other optimization algorithms. Moreover, the total generation cost obtained using the proposed IMA algorithm was less than other optimization algorithms. In Figure 4, the total generation cost was obtained using IMA with MA and other published algorithms. It is observed from Figure 4 that improved mayfly optimization algorithm enhanced the total generation cost by 19.68%, 14.37%, 3.08%, 3.05%, 2.17%, 1.88%, 0.94%, and 0.*23*% over RGM, ACO, CSA, ISA, HIS, IAHS, MHS, and MA, respectively.

4.2. Case 2: Thermal Power Generating Units without Renewable Sources. In this case, the mayfly optimization algorithm was utilized to resolve the CEED problem in the microgrid. The case took a total of 3 fossil-fuel-powered thermal generation units under consideration. The optimization results achieved by IMA were contrasted with other such algorithm results. Table 2 shows the total generation cost achieved by the proposed IMA and other such optimization algorithms. The cost convergence profile results are shown in Figure 5 for the proposed IMA and other optimization algorithms. From the results, it can be understood that IMA has promptly converged to the optimal outcome. The comparison results of total generation cost savings are shown in Figure 6 for the CEED problem when using IMA. It shows that IMA achieved better results than MA and other optimization algorithms. The comparison of total generation cost savings obtained using MA and other published algorithms is presented in Figure 7. Figure 7 infers that the total generation cost improved when using the IMA algorithm by 18.52%, 14.67%, 2.93%, 3.3%, 3.27%, 1.32%, and 0.94% over RGM ACO, CSA, ISA, MHS, and MA, respectively. Furthermore, the total generation cost obtained

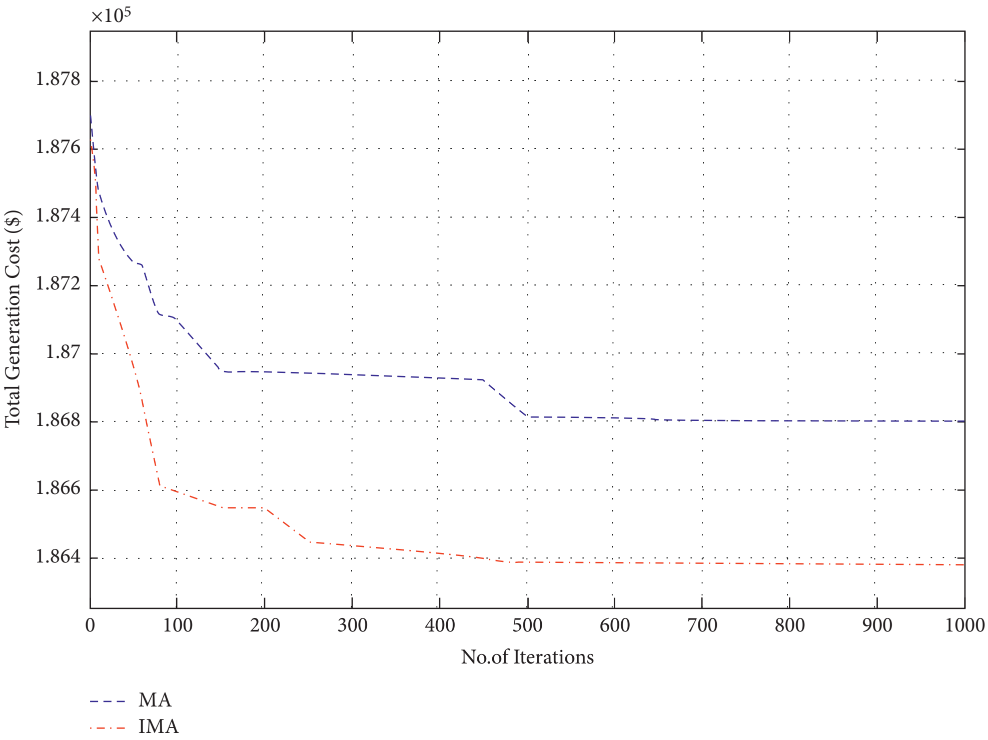


Figure 2: Convergence characteristics for total generation cost (case I).

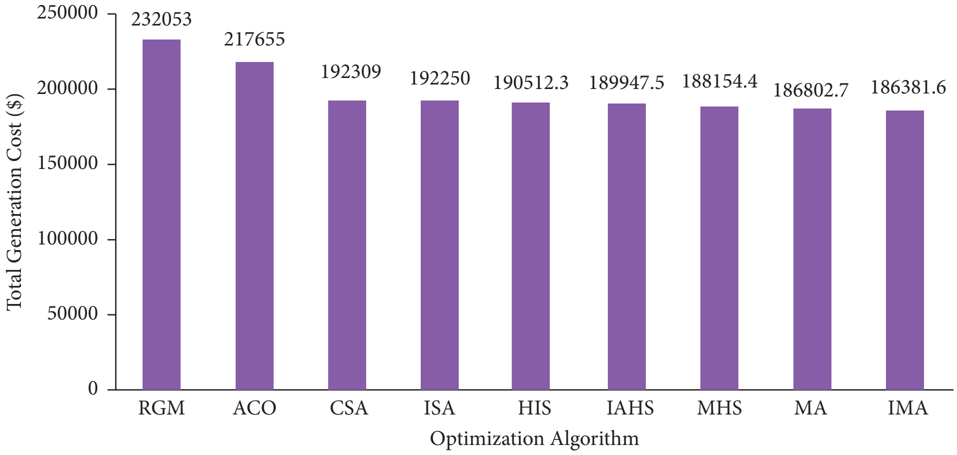


Figure 3: Comparison of total generation cost for case 1.

in case 1 can be less than in case 2 due to incorporating renewable energy sources in a microgrid.

4.3. Case 3: Thermal Power Generating Units with Wind Sources Only. Improved mayfly optimization algorithm was deployed in this case to resolve the CEED problem found in microgrids. This case considered fossil-fuel-powered thermal generators in addition to wind sources. IMA and other models (MA, IHS, CSA, and ISA algorithms) were simulated, and the results were compared. Table 3 shows the generation cost calculated for IMA and other such optimization algorithms. In Figure 8, the author shows the cost convergence characteristic for the optimization algorithms under comparison and the proposed IMA.

Further, Figure 8 also provides an inference; i.e., the convergence characteristic of the proposed LISA strategy II was smooth and quick compared to other strategies. In Figure 9, the researcher compared the total cost saving of IMA and other optimization algorithms from the CEED problem. It is observed from the application results that IMA yielded less total generation cost compared to MA, IHS, CSA, and ISA algorithms. Figure 10 shows the comparison results of operation cost savings obtained using IMA and

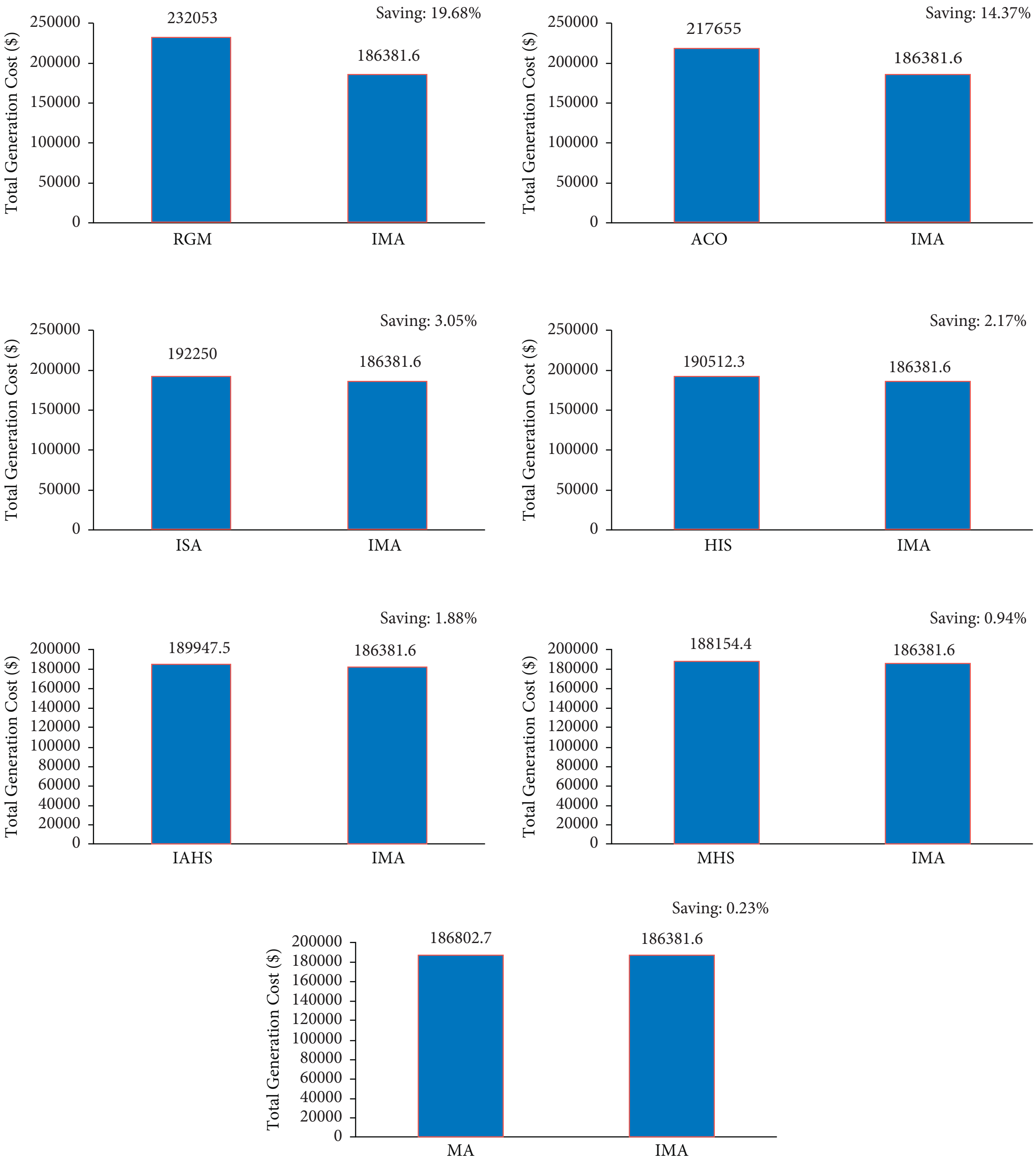


Figure 4: Total generation cost saving of CEED problem for case 1.

other published algorithms. It is observed from Figure 10 that IMA improved the operation cost by 3.66%, 3.64%, 1.6%, and 0.55% over CSA, ISA, MHS, and MA, respectively. Further, the total generation cost obtained in this case remains lower than in case 2 because of integrating a wind-powered energy source with the microgrid.

4.4. Case 4: Thermal Power Generating Units with Solar Source Only. The case scenario considered fossil-fuel-powered thermal generating units with solar sources. In this scenario, improved mayfly optimization algorithm was selected to resolve the CEED problem found in microgrids. Simulation results obtained using an improved mayfly optimization

Table 2: Optimal generation schedule of microgrid for case 2.

Time (h)	RGM cost ($/h) [5]	ACO cost ($/h) [5]	CSA cost ($/h) [5]	ISA cost ($/h) [5]	MHS ($/h) [9]	MA ($/h)	IMA ($/h)
1	8490	7317	7179	7179	6977.4	6849.7	6810.2
2	8528	7694	7365	7367	7194.4	7089.6	7061.3
3	8592	7922	7479	7499	7310.3	7223.8	7198.5
4	8675	8117	7598	7608	7429.5	7351.7	7319.3
5	8756	8318	7721	7722	7550.8	7448.2	7416.7
6	8878	8600	7849	7851	7675.4	7567.1	7534.3
7	9005	8768	7978	7978	7802.5	7731.8	7707.4
8	9167	8998	8110	8110	7933.7	7861.8	7829.1
9	10527	10406	8943	8943	8774.3	8697.4	8669.1
10	11867	11347	9540	9540	9380.9	9304.8	9275.4
11	12664	12032	9851	9850	9696.6	9612.5	9590.1
12	13511	12476	10170	10170	10020.0	9973.4	9942.8
13	12664	12032	9850	9746	9696.6	9668.2	9651.6
14	11160	10889	9238	9230	9074.4	8994.6	8971.3
15	10009	9936	8657	8675	8483.5	8401.7	8375.4
16	9167	8998	8110	8109	7933.7	7894.6	7869.2
17	8875	8599	7849	7849	7675.6	7632.8	7605.2
18	9347	9186	8244	8244	8067.5	7992.7	7969.1
19	10009	9936	8657	8657	8483.5	8401.8	8372.4
20	12664	12032	9851	9847	9696.6	9613.8	9589.5
21	11495	11197	9388	9388	9226.1	9148.7	9129.2
22	9540	9479	8377	8379	8203.0	8115.3	7991.5
23	8675	8117	7598	7598	7429.5	7349.3	7317.6
24	8515	7491	7265	7260	7082.8	7005.4	6974.1
Total	240780	229887	202867	202799	198798.4	196930.7	196170.3

algorithm are compared with the outcomes attained by MA and other such algorithms. The total generation cost of the improved mayfly optimization algorithm and other optimization algorithms is presented in Table 4. Figure 11 represents the convergence characteristics obtained to minimize total generation cost using MA and IMA. From Figure 11, it is concluded that the proposed IMA provides steady and quick convergence characteristics. Figure 12 show the comparison results of total generation cost saving achieved by IMA and other optimization algorithms for CEED problem found in microgrids. It is observed from the optimization results that an improved mayfly algorithm provides less total generation cost than other optimization techniques. Figure 13 shows the optimization results of total generation cost saving obtained using IMA and other published metaheuristic optimization algorithms. Figure 13 infers that the proposed IMA algorithm enhanced the total generation cost by 18.5%, 14.7%, 3.44%, 3.41%, 1.43%, and 0.41% over RGM, ACO, CSA, ISA, MHS, and MA, respectively. The authors also conclude that the total generation cost is less in this scenario than in case 2 because of the incorporation of solar-powered energy sources with the microgrid.

4.5. Comparison between the Cost Curves of All Scenarios. Figures 14 and 15 show the comparison results of total generation cost curves under all the scenarios compared to IMA and MA algorithms 24 hours a day. Furthermore, Figure 16 shows the quantitative comparative results of total cost under all the scenarios using IMA. One can notice from Figures 14–16 that case 1 provides a minimum generation cost compared to other scenarios. Also, it can be observed from case 2 that the highest generation cost is obtained in this case. This might be attributed to the reason that renewable energy sources function as negative loads, while the rest are provided by the fossil-fuel-powered thermal generating units only. It reduces the total generation cost. Furthermore, the total generation cost obtained was less in case 3 than in case 4. It could have occurred due to heavy investment costs incurred upon solar power compared to wind power.

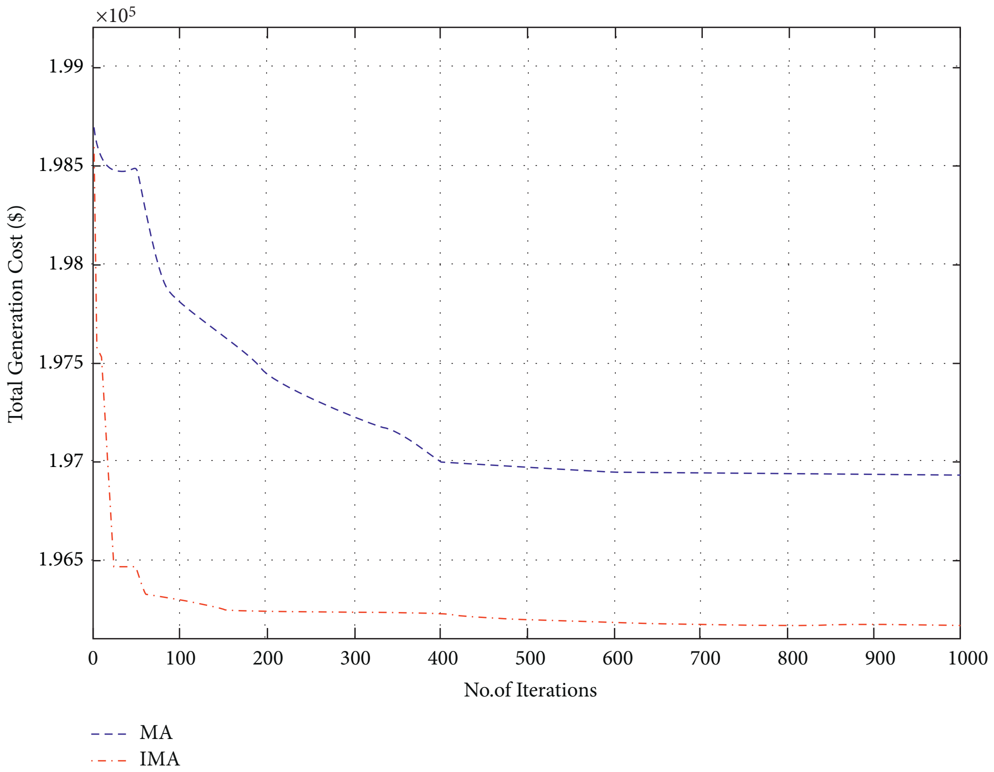


Figure 5: Convergence characteristics for total generation cost (case II).

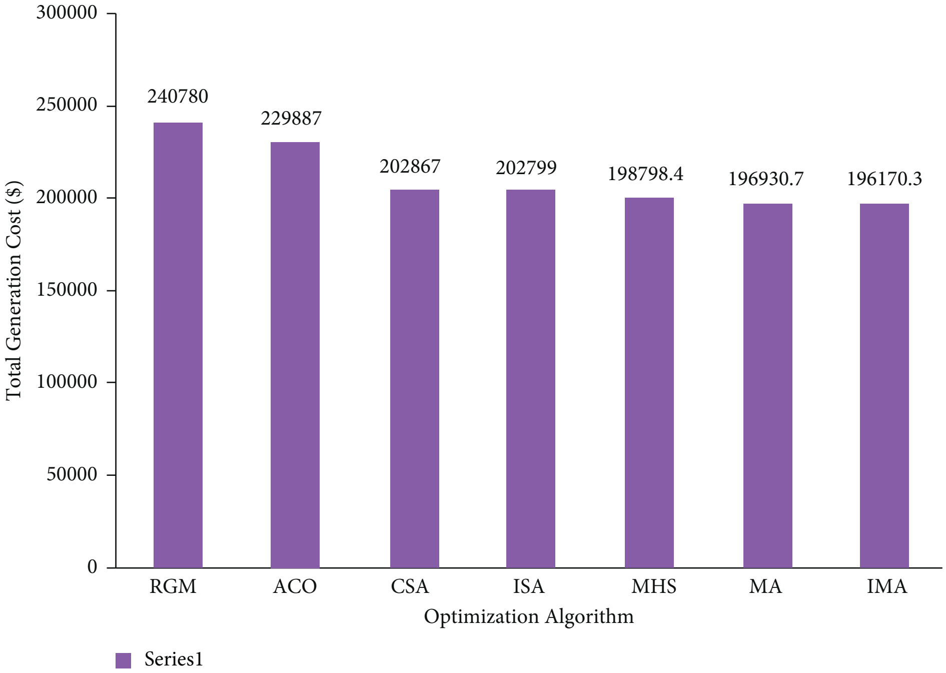


Figure 6: Comparison of total generation cost for case 2.

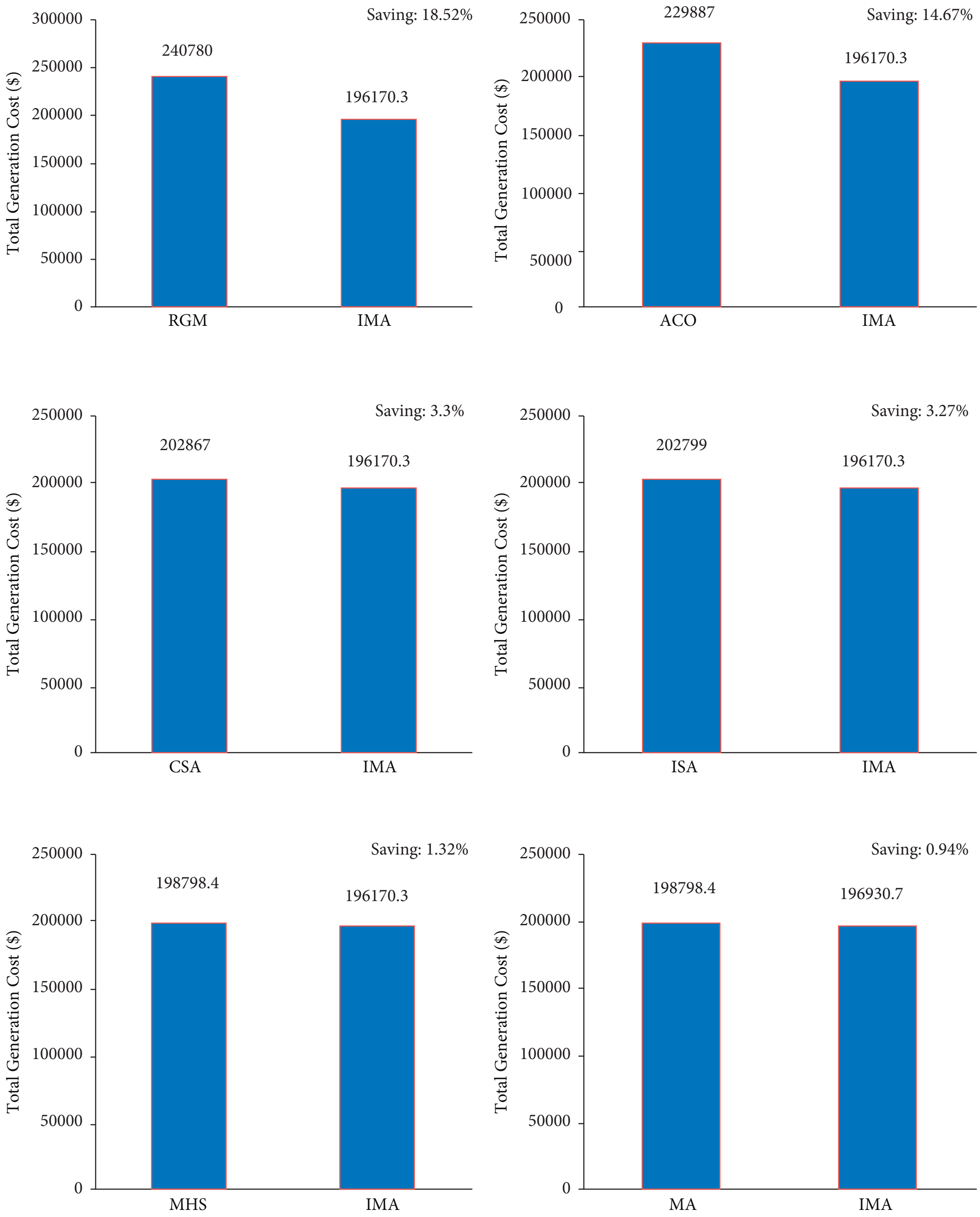


Figure 7: Total generation cost saving of CEED problem for case 2.

Table 3: Optimal generation schedule of microgrid for case 3.

Time (h)	CSA cost ($/h) [5]	ISA cost ($/h) [5]	MHS ($/h) [9]	MA ($/h)	IMA ($/h)
1	7153	7152	6943.4	6851.3	6819.6
2	7203	7199	7010.5	6917.4	6882.7
3	7279	7279	7099.6	7012.6	6990.8
4	7235	7235	7050.2	6972.8	6943.5
5	7544	7545	7377.2	7293.5	7269.3
6	7724	7724	7553.4	7476.8	7435.4
7	7606	7606	7439.4	7355.1	7321.9
8	7443	7443	7278.5	7194.3	7162.8
9	8364	8364	8190.1	8101.6	7784.2
10	9006	9006	8840.7	8753.4	8729.1
11	9454	9461	9295.8	9211.8	9178.6
12	9581	9581	9425.9	9346.9	9313.4
13	9408	9407	9248.7	9168.4	9132.8
14	8933	8933	8766.3	8679.5	8643.7
15	8427	8427	8252.3	8162.9	8123.1
16	7756	7758	7584.2	7495.4	7461.8
17	7761	7761	7590.1	7503.9	7481.6
18	8194	8193	8017.2	7927.1	7893.9
19	8636	8644	8461.9	8361.8	8329.2
20	9845	9842	9690.7	9613.9	9581.3
21	9383	9383	9221.8	9139.6	9107.9
22	8371	8325	8194.7	8127.4	8094.9
23	7572	7571	7403.7	7317.8	7293.5
24	7254	7254	7070.8	6976.7	6943.1
Total	197132	197093	193006.9	190962	189918

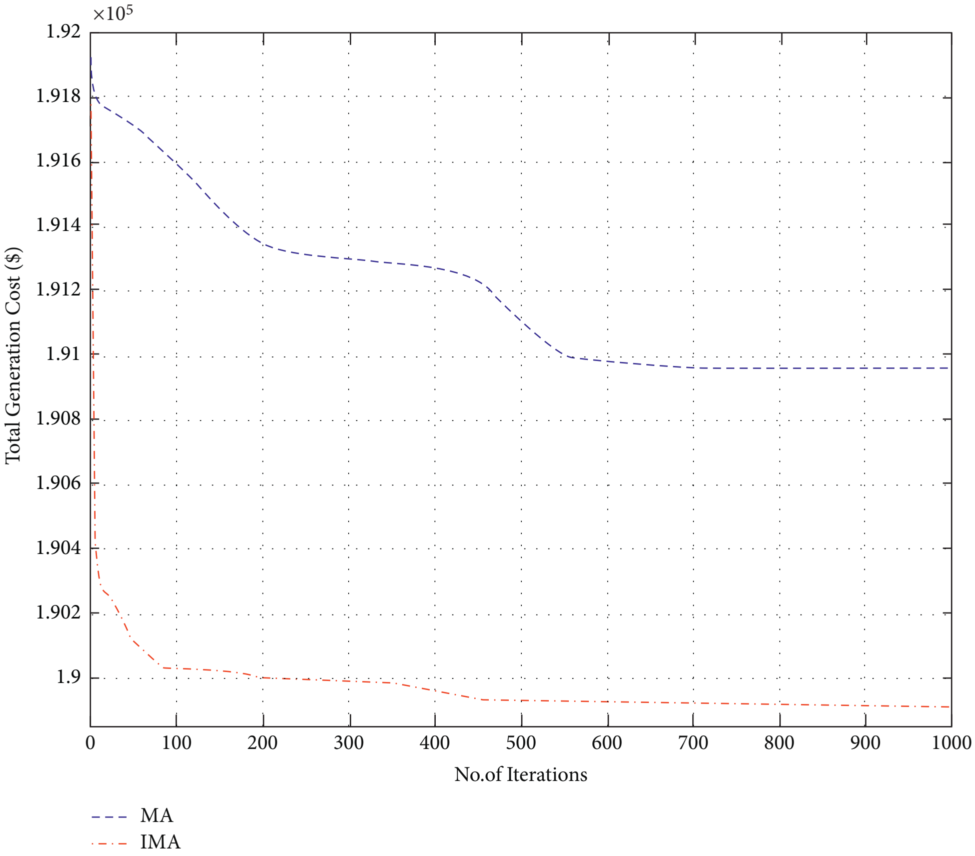


Figure 8: Convergence characteristics for total generation cost (case III).

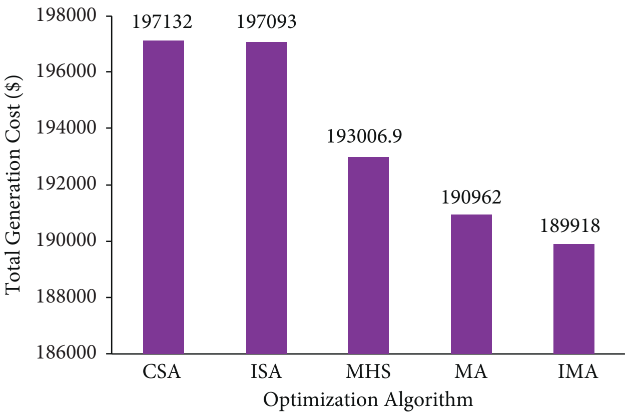


Figure 9: Comparison of total generation cost for case 3.

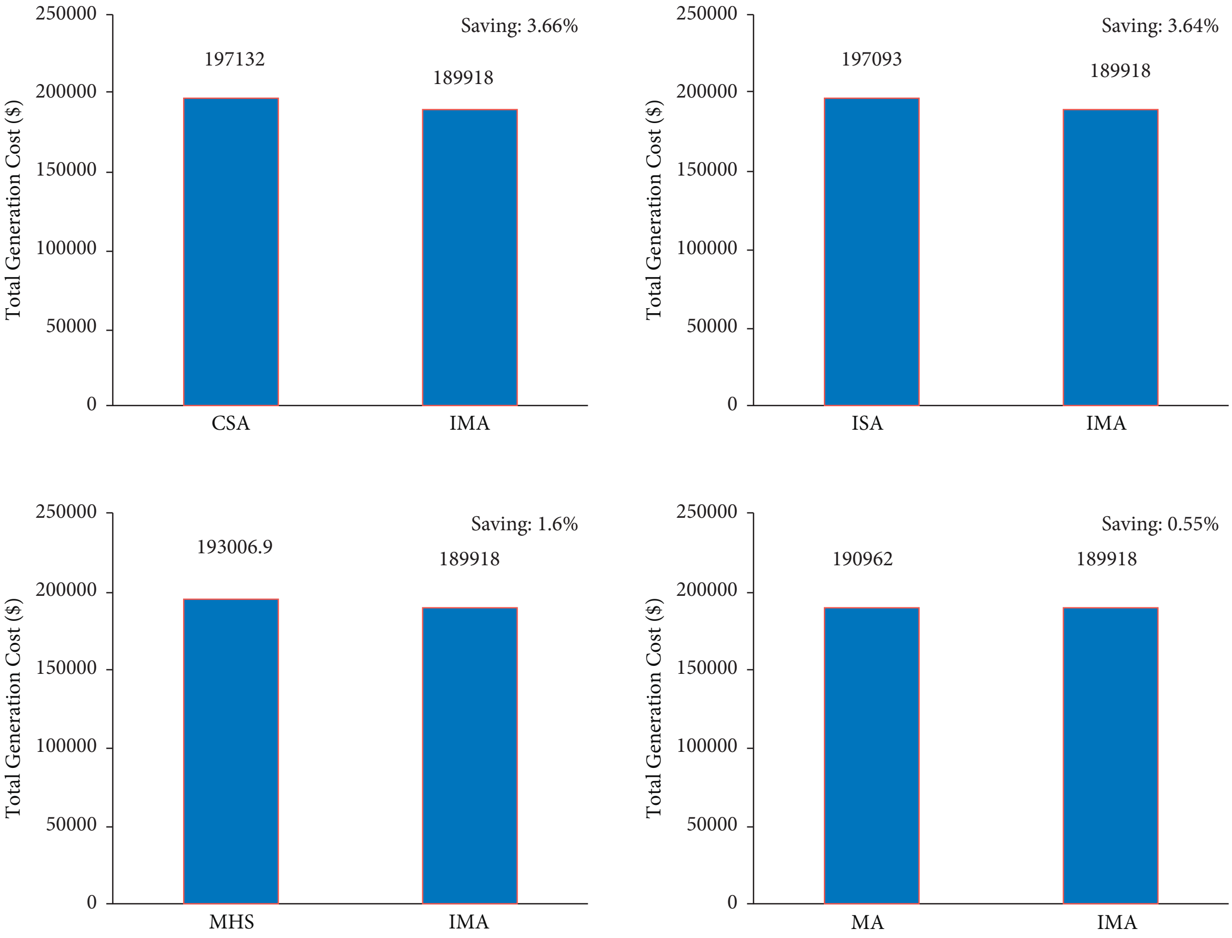


Figure 10: Total generation cost saving of CEED problem for case 3.

Table 4: Optimal generation schedule of microgrid for case 4.

Time (h)	RGM cost ($/h) [5]	ACO cost ($/h) [5]	CSA cost ($/h) [5]	ISA cost ($/h) [5]	MHS ($/h) [9]	MA ($/h)	IMA ($/h)
1	8490	7317	7179	7156	6977.4	6885.7	6827.4
2	8528	7694	7365	7364	7194.4	7106.1	7073.2
3	8592	7922	7479	7508	7310.3	7208.9	7174.3
4	8675	8117	7598	7599	7429.5	7335.2	7302.8
5	8756	8318	7721	7721	7550.8	7461.5	7423.9
6	8878	8600	7848	7841	7675.3	7589.8	7554.2
7	8849	8589	7816	7816	7647.2	7559.7	7523.4
8	8969	8559	7692	7692	7530.9	7447.5	7419.2
9	9788	9630	8269	8244	8105.9	8034.4	7994.2
10	10235	10139	8397	8337	8242.2	8187.3	8149.7
11	12153	11648	9620	9634	9465.9	9377.2	9341.8
12	13327	12336	10052	10053	9903.0	9847.2	9829.6
13	10957	10788	8887	8887	8734.9	8660.7	8639.1
14	10153	10012	8467	8467	8305.8	8227.1	8194.7
15	9707	9617	8377	8378	8206.8	8119.4	8094.9
16	9093	8829	7974	7970	7797.9	7707.6	7679.2
17	8810	8279	7608	7608	7444.4	7359.8	7336.2
18	9340	9137	8182	8182	8006.6	7912.4	7884.1
19	10009	9937	8657	8657	8483.5	8391.2	8357.4
20	12664	12032	9851	9849	9696.6	9617.2	9589.7
21	11495	11197	9388	9400	9226.1	9171.2	9143.8
22	9540	9479	8379	8379	8203.0	8112.3	8084.7
23	8675	8117	7598	7596	7429.5	7344.8	7312.2
24	8515	7491	7264	7263	7082.8	6990.4	6943.1
Total	234198	223784	197668	197601	193650.8	191654.6	190872.8

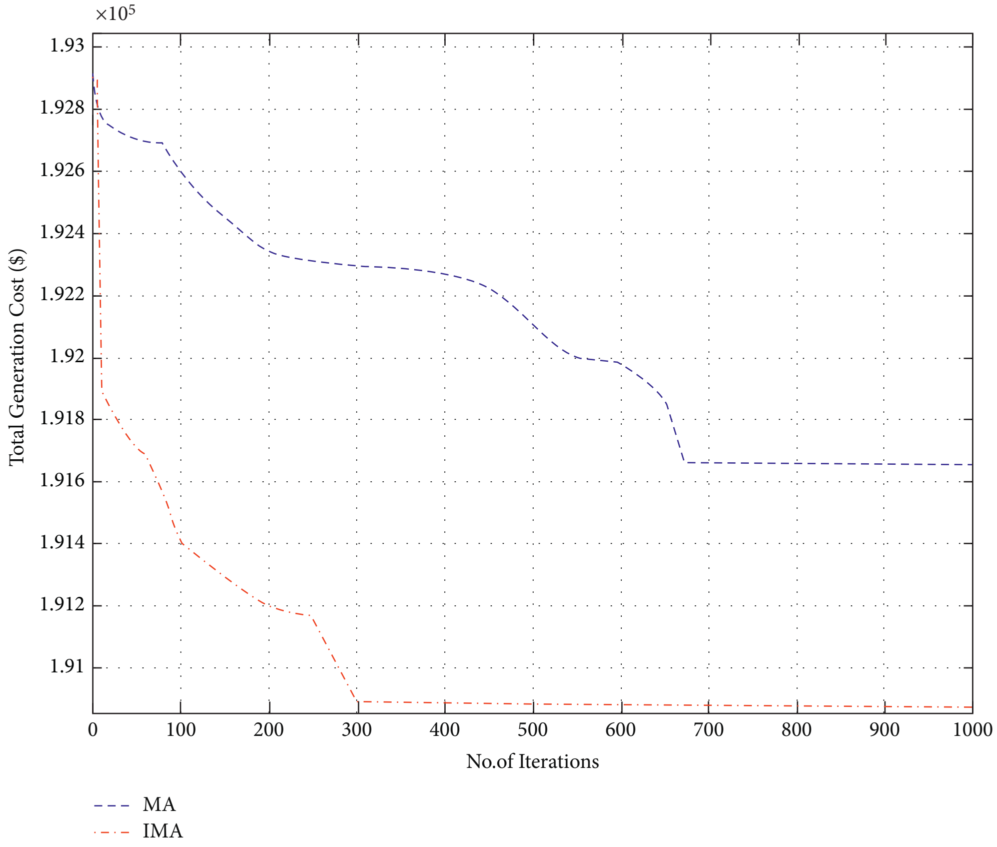


Figure 11: Convergence characteristics for total generation cost (case IV).

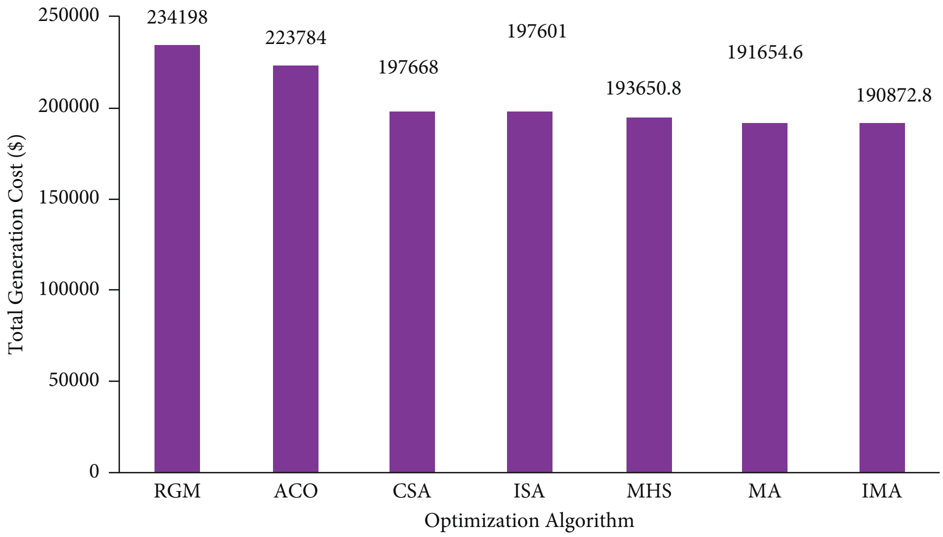


Figure 12: Comparison of total generation cost for case 4.

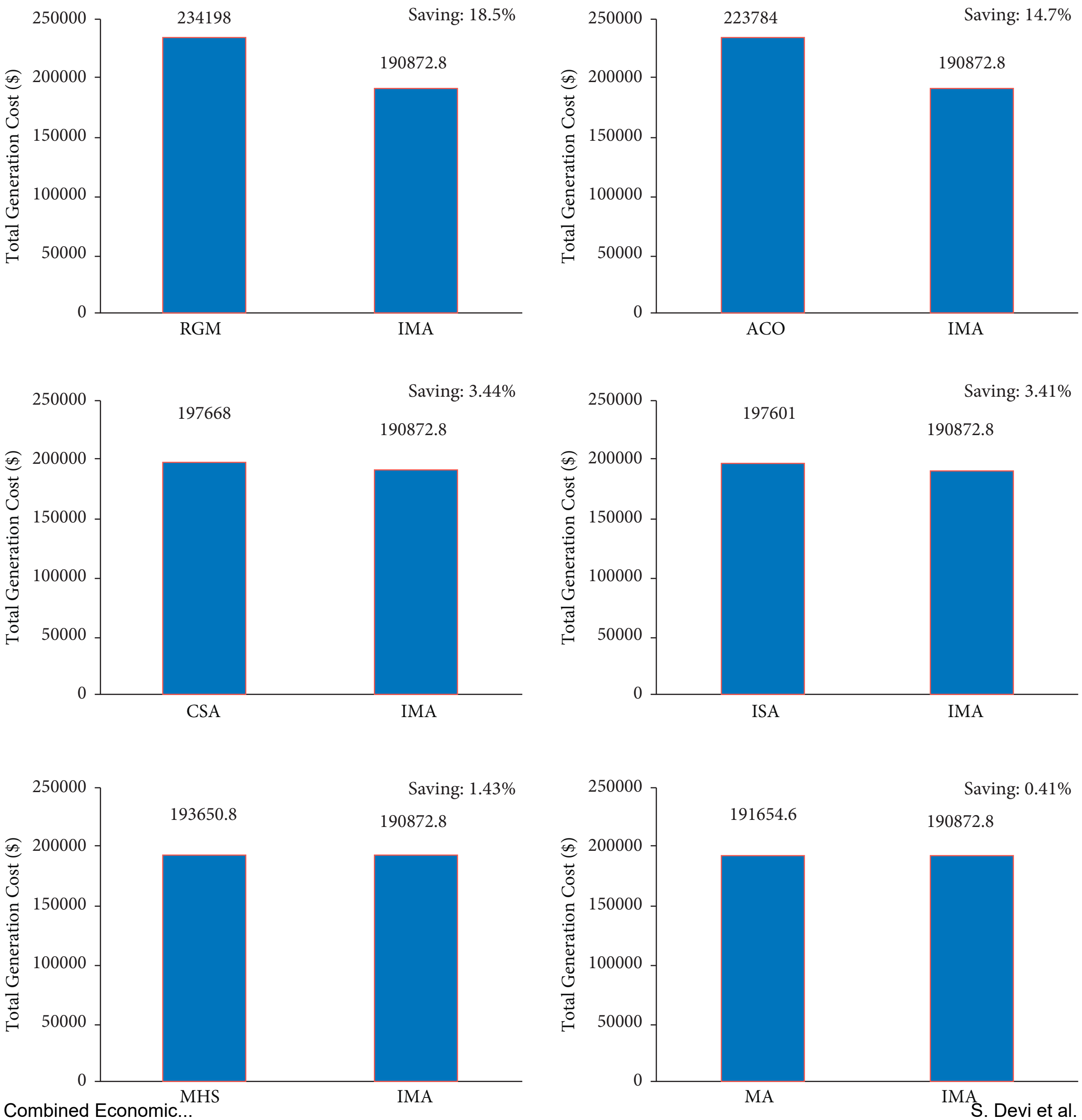


Figure 13: Total generation cost saving of CEED problem for case.

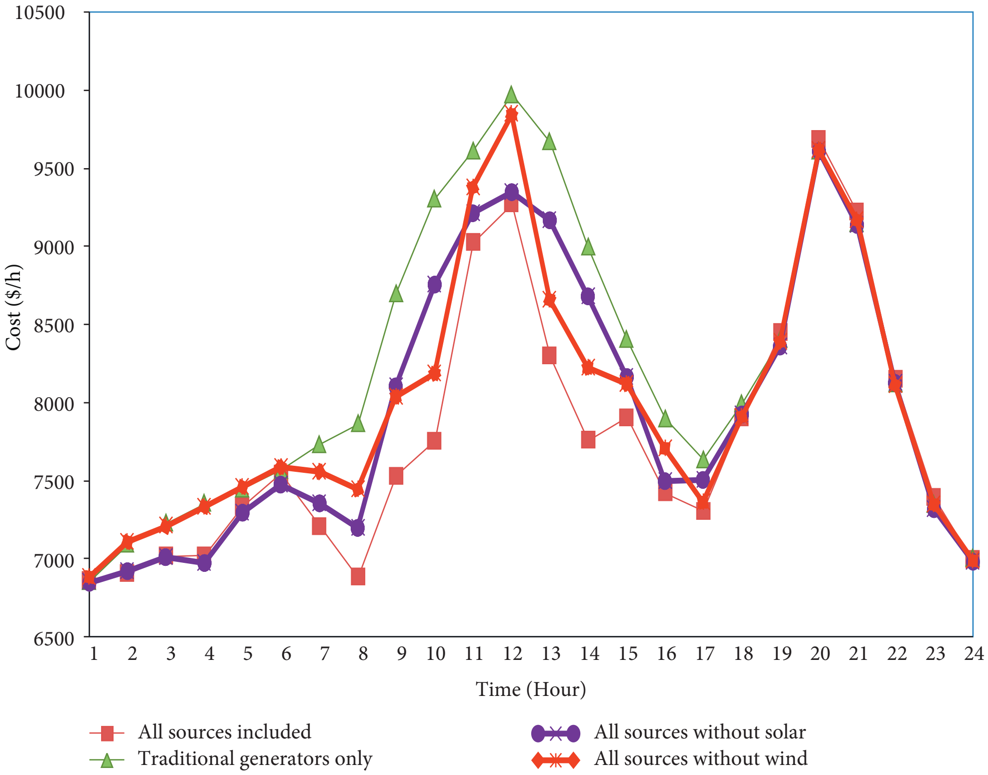


FIGURE 14: Comparison of the cost curve for all cases for 24 hours of a day using mayfly algorithm.

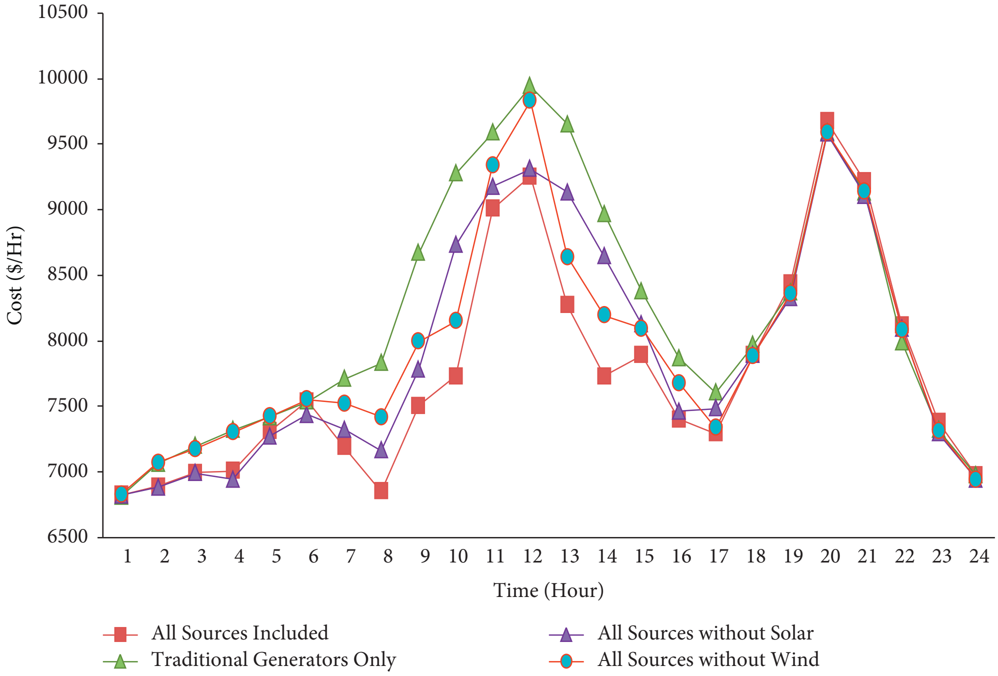


FIGURE 15: Comparison of the cost curve for all cases for 24 hours of a day using improved mayfly algorithm.

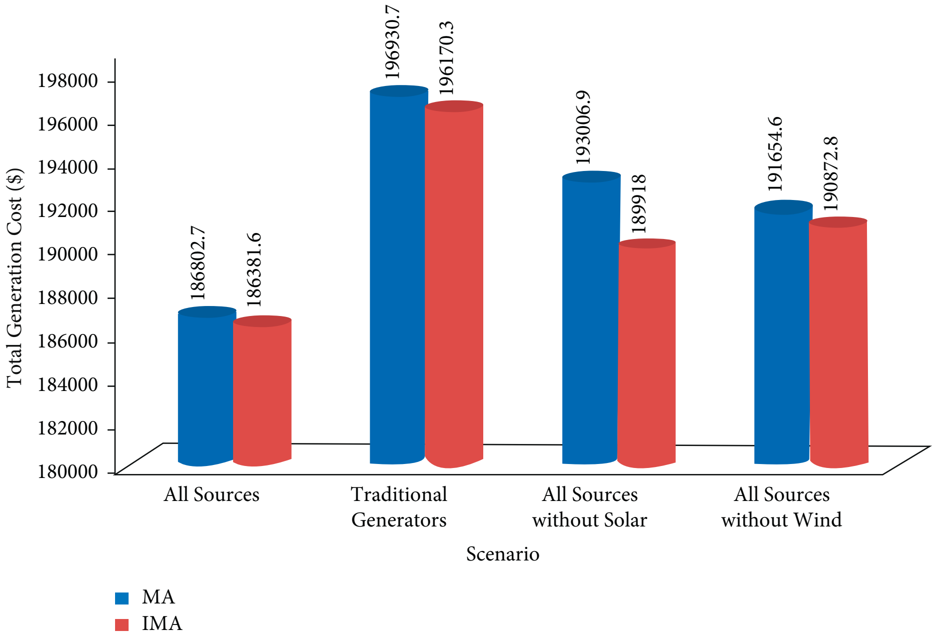


FIGURE 16: Comparison of total generation cost for all cases using MA and IMA algorithm.

5. Conclusion

In the current study, the improved mayfly optimization algorithm (IMA) has been implemented to resolve the combined economic emission dispatch (CEED) with renewable energy sources. The study incorporated the proposed IMA as a solution for the CEED problem encountered in the microgrid. Solar and wind power are considered as the cost functions in this study. The proposed IMA algorithm was validated for its supremacy and efficiency in a microgrid model under varying scenarios. The outcomes of IMA and other algorithms were compared and contrasted. The comparison results show that the proposed IMA algorithm is better in cost reduction under all the scenarios. This infers that the proposed IMA is superior, robust, and efficient over other metaheuristic optimization algorithms published earlier. In future, the improved mayfly optimization algorithm can be applied to tackle the CEED problem in grid-connected microgrids comprising battery storage and electric vehicles to accomplish single and multiobjective optimization.

Nomenclature

List of Symbols

a:	Annuitization coefficient
C_w:	Cost of wind power generating unit
C_s:	Cost of solar-powered generating unit
d:	Coefficient of nuptial dance
$E_i(P_{\mathrm{Gi}})$:	Emission level of the i^{th} generator
E_t:	Overall emission
fl:	Random walk coefficient
$F_i(P_{\mathrm{Gi}})$:	Fuel cost of the i^{th} generator
F_t:	Overall fuel cost incurred
g:	Gravity coefficient
gbest:	Global best
iter:	Current iteration
$\mathrm{iter}_{\mathrm{max}}$:	Maximum no. of iterations
L:	Random number
I^P:	Costs incurred upon investment per unit installed power
N:	Lifetime investment
NG:	No. of generating units
O^E:	Operating and maintenance costs per unit installed power
pbest:	Personal best
$P_{\mathrm{Gi}}^{\mathrm{max}}$:	Maximum output power of generator i
$P_{\mathrm{Gi}}^{\mathrm{min}}$:	Minimum output power of generator i
P_D:	Total load demand
P_{Gi}, P_{Gmi}, P_{Gfi}:	Power output of i^{th} generating unit
P_s:	Output power from solar-powered generating unit
P_w:	Output power from wind power generating unit
r:	Interest rate
r_p:	Cartesian distance between x_i and pbest_i
r_g:	Cartesian distance between x_i and gbest
r_{mf}:	Cartesian distance between male and female mayflies
TC:	Total operating cost
β:	Fixed visibility coefficient
Λ:	Price penalty factor
Λ_i:	Price penalty factor of i^{th} generator
Λ:	Distribution factor
δ:	Scaling factor
$\Gamma(.)$:	Gamma distribution function
σ:	Standard deviation.

Abbreviations

BSDE:	Bernstein-search differential evolution
CEED:	Combined economic emission dispatch
ChOA:	Chimp optimization algorithm
CHP:	Combined heat and power
CHPED:	Combined heat and power economic dispatch
CNO:	Collective neurodynamic optimization
DCHPED:	Dynamic combined heat and power economic emission dispatch
DDG:	Dispatchable distributed generator
DE:	Differential evolution
DEED:	Dynamic economic emission dispatch
DSM:	Demand side management
ECED:	Environment constrained economic dispatch
EED:	Economic environmental dispatch
ELD:	Economic load dispatch
EMA:	Exchange market algorithm
FC:	Fuel cell
FCDED:	Fuel constrained dynamic economic dispatch
FP:	Fractional programming
GA:	Genetic algorithm
GWO:	Grey wolf optimizer
HFA/HS:	Hybrid Firefly and Harmony Search
IGWO:	Improved grey wolf optimizer
IMA:	Improved mayfly algorithm
ISA:	Interior search algorithm
LPSP:	Loss of probability of power supply
MA:	Mayfly algorithm
MAACPSO:	Modified adaptive accelerated particle swarm optimization
MBGSA:	Memory-based gravitational search algorithm
MG:	Microgrid
MILP:	Mixed-integer linear programming
MOVCS:	Multiobjective virus colony search
MT:	Microturbine
PPF:	Price penalty factor
PSO:	Particle swarm optimization
PV:	Photovoltaic
QPSO:	Quantum particle swarm optimization
SFS:	Stochastic fractal search algorithm
SGEO:	Social group entropy optimization
SOS:	Sequential optimization strategy
SSER:	System surplus energy rates
WOA:	Whale optimization algorithm
WT:	Wind turbine.

References

[1] N. I. Nwulu and X. Xia, “Optimal dispatch for a microgrid incorporating renewables and demand response,” *Renewable Energy*, vol. 101, pp. 16–28, 2017.

[2] N. Karthik, A. K. Parvathy, R. Arul, and K. Padmanathan, “Economic load dispatch in a microgrid using interior search algorithm, international conference on power and advanced computing,” in *Proceedings of the 2019 Innovations in Power and Advanced Computing Technologies (i-PACT).*, Vellore, India, March 2019.

[3] N. Karthik, A. K. Parvathy, and R. Arul, “A review of optimal operation of microgrids,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, pp. 2842–2849, 2020.

[4] N. Karthik, A. K. Parvathy, R. Arul, and S. Baskar, “A review of optimization techniques applied to solve unit commitment problem in microgrid,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 3, pp. 1161–1169, 2019.

[5] I. N. Trivedi, P. Jangir, M. Bhoye, and N. Jangir, “An economic load dispatch and multiple environmental dispatch problem solution with microgrids using interior search algorithm,” *Neural Computing & Applications*, vol. 30, no. 7, pp. 2173–2189, 2018.

[6] N Karthik, A. K Parvathy, and R Arul, “Multi-objective economic emission dispatch using interior search algorithm,” *International Transactions on Electrical Energy Systems*, vol. 29, 2019.

[7] A. Rajagopalan, P. Kasinathan, K. Nagarajan, V. K. Ramachandaramurthy, V. Sengoden, and S. Alavandar, “Chaotic self-adaptive interior search algorithm to solve combined economic emission dispatch problems with security constraints,” *Int Trans Electr Energ Syst*, vol. 29, no. 2, Article ID e12026, 2019.

[8] N. Karthik, A. K. Parvathy, and R. Arul, “Non-convex economic load dispatch using cuckoo search algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, no. No. 1, pp. 48–57, 2017.

[9] E. Ehab, “Elattar, Modified harmony search algorithm for combined economic emission dispatch of microgrid incorporating renewable sources,” *Energy*, vol. 159, pp. 496–507, 2018.

[10] B. Dey, B. Bhattacharyya, and F. Pedro García Márquez, “A hybrid optimization-based approach to solve environment constrained economic dispatch problem on microgrid system,” *Journal of Cleaner Production*, vol. 307, Article ID 127196, 2021.

[11] M. Kharrich, O. H. Mohammed, S. Kamel, M. Aljohani, M. Akherraz, and M. I. Mosaad, “Optimal Design of Microgrid Using Chimp Optimization Algorithm,” in *Proceedings of the 2021 IEEE International Conference on*

Automation/XXIV Congress of the Chilean Association of Automatic Control (ICA-ACCA), pp. 1–5, Santiago, Chile.

[12] R. Alayi, M. H. Ahmadi, A. R. Visei, S. Sharma, and A. Najafi, "Technical and environmental analysis of photovoltaic and solar water heater cogeneration system: a case study of Saveh City," *International Journal of Low Carbon Technologies*, vol. 16, no. 2, pp. 447–453, 2021.

[13] M. I. Alomoush, "Microgrid dynamic combined power–heat economic-emission dispatch with deferrable loads and price-based energy storage elements and power exchange," *Sustainable Energy, Grids and Networks*, vol. 26, Article ID 100479, 2021.

[14] B. Dey, S. K. Roy, and B. Bhattacharyya, "Solving multi-objective economic emission dispatch of a renewable integrated microgrid using latest bio-inspired algorithms," *Engineering Science and Technology, an International Journal*, vol. 22, no. 1, pp. 55–66, 2019.

[15] M. I. Alomoush, "Microgrid combined power-heat economic-emission dispatch considering stochastic renewable energy resources, power purchase and emission tax," *Energy Conversion and Management*, vol. 200, Article ID 112090, 2019.

[16] T. Wang, X. He, T. Huang, C. Li, and W. Zhang, "Collective neurodynamic optimization for economic emission dispatch problem considering valve point effect in microgrid," *Neural Networks*, vol. 93, pp. 126–136, 2017.

[17] F. Moazeni and J. Khazaei, "Dynamic economic dispatch of islanded water-energy microgrids with smart building thermal energy management system," *Applied Energy*, vol. 276, Article ID 115422, 2020.

[18] M. Kharrich, S. Kamel, A. S. Alghamdi et al., "Optimal design of an isolated hybrid microgrid for enhanced deployment of renewable energy sources in Saudi arabia," *Sustainability*, vol. 13, no. 9, p. 4708, 2021.

[19] A. A. E. Tawfiq, M. O. A. El-Raouf, M. I. Mosaad, A. F. A. Gawad, and M. A. E. Farahat, "Optimal reliability study of grid-connected PV systems using evolutionary computing techniques," *IEEE Access*, vol. 9, no. 1, pp. 42125–42139, 2021.

[20] N. B. Roy and D. Das, "Optimal allocation of active and reactive power of dispatchable distributed generators in a droop controlled islanded microgrid considering renewable generation and load demand uncertainties," *Sustainable Energy, Grids and Networks*, vol. 27, Article ID 100482, 2021.

[21] X.-G. Zhao, Z.-Q. Zhang, Y.-M Xie, and M. Jin, "Economic-environmental dispatch of microgrid based on improved quantum particle swarm optimization," *Energy*, vol. 195, Article ID 117014, 2020.

[22] H. Hou, M. Xue, Y. Xu et al., "Multi-objective economic dispatch of a microgrid considering electric vehicle and transferable load," *Applied Energy*, vol. 262, Article ID 114489, 2020.

[23] W. Gil-González, D. Montoya, E. Holguín, A. Garces, and L. F. Grisales-Noreña, "Economic dispatch of energy storage systems in dc microgrids employing a semi-definite programming model," *Journal of Energy Storage*, vol. 21, pp. 1–8, 2019.

[24] M. Nemati, M. Braun, and S. Tenbohlen, "Optimization of unit commitment and economic dispatch in microgrids based on genetic algorithm and mixed integer linear programming," *Applied Energy*, vol. 210, pp. 944–963, 2018.

[25] P. Pourghasem, F. Sohrabi, M. Abapour, and B. Mohammadi-Ivatloo, "Stochastic multi-objective dynamic dispatch of renewable and CHP-based islanded microgrids," *Electric Power Systems Research*, vol. 173, pp. 193–201, 2019.

[26] X. Lu, K. Zhou, and S. Yang, "Multi-objective optimal dispatch of microgrid containing electric vehicles," *Journal of Cleaner Production*, vol. 165, pp. 1572–1581, 2017.

[27] R. Alayi, A. Kasaeian, and F. Atabi, "Thermal analysis of parabolic trough concentration photovoltaic/thermal system for using in buildings," *Environmental Progress & Sustainable Energy*, vol. 38, no. 6, Article ID 13220, 2019.

[28] R. Alayi, A. Kasaeian, and F. Atabi, "Optical modeling and optimization of parabolic trough concentration photovoltaic/thermal system," *Environmental Progress & Sustainable Energy*, vol. 39, no. 2, Article ID e13303, 2020.

[29] M. M. Samy, M. I. Mosaad, and S. Barakat, "Optimal economic study of hybrid PV-wind-fuel cell system integrated to unreliable electric utility using hybrid search optimization technique," *International Journal of Hydrogen Energy*, vol. 46, no. 20, pp. 11217–11231, 2021.

[30] F. Nazari-Heris, B. Mohammadi-ivatloo, and D. Nazarpour, "Network constrained economic dispatch of renewable energy and CHP based microgrids," *International Journal of Electrical Power & Energy Systems*, vol. 110, pp. 144–160, 2019.

[31] Z. Younes, I. Alhamrouni, S. Mekhilef, and M. Reyasudin, "A memory-based gravitational search algorithm for solving economic dispatch problem in micro-grid," *Ain Shams Engineering Journal*, vol. 12, no. Issue 2, pp. 1985–1994, 2021.

[32] Y. Zou, J. Zhao, D. Ding, F. Miao, and B. Sobhani, "Solving dynamic economic and emission dispatch in power system integrated electric vehicle and wind turbine using multi-objective virus colony search algorithm," *Sustainable Cities and Society*, vol. 67, Article ID 102722, 2021.

[33] R. Alayi, A. Kasaeian, and A. Njafi, "Optimization and evaluation of a wind, solar and fuel cell hybrid system in supplying electricity to a remote district in national grid," *International Journal of Energy Sector Management*, vol. 14, no. 2, pp. 408–418, 2020.

[34] M. M. Samy, M. I. Mosaad, M. F. El-Naggar, and S. Barakat, "Reliability support of undependable grid using green energy systems: economic study," *IEEE Access*, vol. 9, no. 1, pp. 14528–14539, 2021.

[35] R. Alayi, F. Zishan, M. Mohkam, S. Hoseinzadeh, S. Memon, and D. A. Garcia, "A sustainable energy distribution configuration for microgrids integrated to the national grid using back-to-back converters in a renewable power system," *Electronics*, vol. 10, no. 15, p. 1826, 2021.

[36] M. Basu, "Fuel constrained dynamic economic dispatch with demand side management," *Energy*, vol. 223, Article ID 120068, 2021.

[37] M. I. Alomoush, "Application of the stochastic fractal search algorithm and compromise programming to combined heat and power economic-emission dispatch," *Engineering Optimization*, vol. 52, no. 11, pp. 1992–2010, 2019.

[38] K. Zervoudakis and S. Tsafarakis, "A mayfly optimization algorithm," *Computers & Industrial Engineering*, vol. 145, Article ID 106559, 2020.

[39] A. Y. Abdelaziz, E. S. Ali, and S. M. Abd Elazim, "Implementation of flower pollination algorithm for solving economic load dispatch and combined economic emission

dispatch problems in power systems," *Energy*, vol. 101, pp. 506–518, 2016.

[40] X.-S. Yang and S. Deb, "Multiobjective cuckoo search for design optimization," *Computers & Operations Research*, vol. 40, no. 6, pp. 1616–1624, 2013.

[41] H. Hakli and H. Uguz, "A novel particle swarm optimization algorithm with Levy flight," *Applied Soft Computing*, vol. 23, no. 1, pp. 333–345, 2014.

[42] N. Karthik, A. K. Parvathy, R. Arul, and K. Padmanathan, "Multi-objective optimal power flow using a new heuristic optimization algorithm with the incorporation of renewable energy sources," *International Journal of Energy and Environmental Engineering*, vol. 12, no. 4, pp. 641–678, 2021.

Analysis of Artificial Neural Network: Architecture, Types, and Forecasting Applications

Achyutananda Panda, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, achutanada.panda23@gmail.com*

Sandip Kar Mazumdar, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, sk_mazumdar1@gmail.com*

Himanshu Sekhar Moharana, *Department of Mechanical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, hsmoharana26@gmail.com*

Achyutananda Panda, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, achutanada.panda23@gmail.com*

Abstract

The artificial neural network reduces humanity and society's burden to solve complex problems highly efficiently. Artificial neural networks resemble brain activities based on the acquired training samples used for various applications such as classification, regression, prediction, smart grid, natural language processing, image processing, medical diagnosis, and so on. This paper illustrates the different artificial neural network architectures, types, merits, demerits, and applications. Therefore, this paper provides valuable information to students and researchers to enrich their knowledge about an artificial neural network and research it. This paper also proposed a multilayer-perceptron-neural-network-based solar irradiance forecasting model, an improved backpropagation neural network-based rainfall forecasting model, and an Elman neural network-based temperature forecasting model. The performances of the proposed neural network-based forecasting models are analyzed with various hidden neurons and validated using the acquired real-time meteorological data. The proposed neural network forecasting models achieve rigorous results with reduced errors for the considered applications and aid sustainability.

1. Introduction

In the modern world, ANN is actively replacing the existing methods; this motivated us to address the issue of ANN and made interest in the quest of ANN and provided the complete guideline to the reader about the ANN types, architecture, and applications of ANN. The network that resembles or mimics the biological human brain functions to accomplish a given task is an artificial neural network. In a neural network, one neuron to the other neuron connection exists with some strength known as weight or synaptic weight. The on and off state of a neuron is decided by the threshold function. The perceptron concept was introduced in 1958 by Frank Rosenblatt [1], which is the ability to learn with the single-layer network. The limitation is if the data points are not linearly separable, it cannot solve the problem. Still, many research activities are required to address the perceptron network linear separability issue. Inputs usually are binary, bipolar, and the real-time value from the environment. Many forecasting models are attempted in the literature, but simple, feasible, easy to implement, and accurate forecasting is one of the thrust research fields. In neural network-based forecasting, an interpretable machine learning tool is important [2, 3], and feature selection/extraction is a preprocessing method used to extract important relevant input features [4], data-driven, hybrid, ensemble, and deep neural network aid effective solutions [5–10]. Recently, some researchers performed data-driven-based forecasting [11–13]. In forecasting applications, variability is caused because of the measurement shift and noise. The uncertainty about measurement shift and noise can be overcome by proper commissioning, data evaluation, quality check, sensor calibration, and usage of on-site measurement data [9].

1.1. Comparison between the Human Brain and AI (Artificial Intelligence)

(i) The machine system involves step-by-step procedures and instructions, but humans have fewer processing steps because of the massively parallel operation. In this aspect, humans are ahead of artificial intelligence.

(ii) Regarding the size and complexity aspect, the human biological brain has 10^{11} numbers of neurons approximately and 10^{15} numbers interconnecting with that brain size; this is many neurons. Interconnection is highly impossible based on AI; hence, complexity exists in the human brain both outside of dendrites and inside of cell body computation, but it delays the artificial intelligence process.

(iii) Regarding the strength (or) the synaptic weight of interconnection, information has been stored in the machine having replaceable storage, but the brain has an adaptable storage system.

(iv) The brain has a much high fault tolerance ability compared to a computer and artificial intelligence machine.

(v) Control mechanisms: the retrieval of corrupted information is complicated for the human brain than the machine, so the control mechanism is more difficult for the human brain than the machine.

1.1.1. Motivation. Practical examples: babies or kids can differentiate fruit; guava and green apple both are green in color and shape, but with respect to experience and unique features, the baby or kid determine green apple and guava if we include green mango, and green apple baby can identify similarly ANN-based acquired knowledge about the trained data sets. It has the generalization ability to identify the unknown data sets correctly.

The learning or training process is the stage; the network can acquire knowledge about the situation or environment. The acquired knowledge is stored in the synaptic weights. According to the structural interlinkages of neurons (computational elements), activation function, and weight computation process (learning algorithm), the artificial neural networks can be classified into various types, starting from Mcculloch and Pitts network to present hybrid neural networks.

In a single-layer neural network model, the hidden layer does not appear, so only single connection linkages exist from input to output. A multilayer neural network consists of more than one connection linkage from the input to the output, which can solve challenging and complex problems.

The network does not have feedback, and information can flow from the input layer to the output layer via one or more hidden layers known as a feedforward neural network. Examples of feedforward neural networks are backpropagation, multilayer neural network, radial basis neural network, etc.

The neural network consists of feedback, and information can flow from the input layer to the output layer via one or more hidden layers, and vice versa known as a feedback neural network. Examples of feedback neural networks are the Hopfield network, Elman network, and so on.

1.2. Contribution. This paper has the following contributions:

(i) Give a clear understanding of the comparison between the human brain and artificial intelligence.

(ii) Discuss the development background of an ANN, artificial neural network generalized procedural steps with diagrammatic explanation, typical structure, and applications of artificial neural networks.

(iii) Propose various types of artificial neural networks, like multilayer perceptron neural network, improved backpropagation neural network, and Elman neural network for solar irradiance forecasting, rainfall forecasting, and temperature forecasting, respectively, to aid sustainability.

1.3. Highlights of This Paper

(i) Acquire knowledge of the primary artificial neural network types and applications

(ii) Discuss various ANN architectures, types, and applications

(iii) Propose a solar irradiance forecasting model using MLPNN (Multilayer perceptron neural network)

(iv) Suggest a rainfall forecasting model using IBPNN (Improved Backpropagation Neural Network)

(v) Present a temperature forecasting model using ENN (Elman neural network)

(vi) Carry out various hidden neuron-based analyses

(vii) Propose proved validity of forecasting models in real-time data

(viii) Achieve rigorous forecasting outcomes with reduced errors regarding the proposed forecasting model

(ix) Analyze the different neural network models and familiarize the reader with the forecasting applications

1.4. Development Background of Artificial Neural Network. The motivation of Artificial Neural Network (ANN) is the biological system's parallel and distributed processing. In 1986, McClelland et al. [14] developed an intelligent machine with artificial intelligence, but searching for the solution is the problem with this model. Hence, many heuristic searches address the task accomplishment, and the rule-based approach addresses the representation problem. Table 1 represents the ANN development in the literature for more clarity.

1.4.1. Second Generation Neural Networks. The second generation neural networks are as follows:

(i) Perceptron: for specific learning rules, the network can learn with known target values (supervised learning rule)

Table 1: Development of ANN in the literature.

Year	Authors	Proposed model
1943	Mcculloch and Pitts [15]	Perceptron network with two artificial neurons.
1949	Hebb [16]	Hebbian learning rule.
1958	Rosenblatt [1, 17]	Perceptron network models.
1960	Widrow and Hoff [18]	Adaline neural network.
1962	Widrow [19]	Madaline neural network.
1964	Zadeh [20]	Fuzzy logic.
1982	Hopfield [21]	Hopfield network (recurrent).
1986	Rumelhart et al. [22]	Backpropagation neural network.
1988	Chang and Yang [23]	Cellular neural network (communication exists between only neighboring neurons).
1995	Cortes and Vapnik [24]	Support vector machine.
2002	Gerstner and Kistler [25]	Spiking neural network.
2012	Hinton [26]	Deep learning neural network.

(ii) Backpropagation: based on the various learning methods, the network can learn and adapt the learning data set to known target values (supervised learning rules)

(iii) Kohonen neural network or Self-Organizing Map (SOM): the network learns without knowing the target values (unsupervised learning rule)

(iv) Radial basis neural network (supervised learning rule)

(v) Adaptive resonance theory (unsupervised learning rule)

(vi) Elman neural network (supervised learning rule)

(vii) Hopfield neural network (unsupervised learning rule)

(viii) Special networks (unsupervised and supervised learning rules) like support vector machine

1.4.2. Third Generation Neural Network. Spiking neural network: in this neural network, the limitation of MLP, like cycle firing, has been avoided. This model acquires the firing of biological neurons based on spikes.

1.4.3. Fourth Generation Neural Network. The fourth-generation neural network can be classified into two types as follows:

(i) Deep learning neural network: the deep learning neural network can overcome the gradient vanishing problem over one number of hidden layers

(ii) Hybrid neural network: using both artificial neural network and optimization algorithms, a combination of physical and statistical methods, and so on, makes a hybrid neural network that overcomes the individual network's limitations

2. Generalized Algorithm for Artificial Neural Network

The algorithm in neural networks is nothing but a step-by-step procedure to accomplish a specific task. We showed the generalized algorithm of an artificial neural network in Figure 1, which comprises the following steps as follows:

Step 1. Start the neural network design phase, choose the appropriate neural network model of the artificial neural network (feedforward or feedback, or special neural network, or hybrid model).

Step 2. After selecting a neural network model design, the proposed model includes the number of input parameters, the number of hidden layers, the number of hidden neurons, and so on.

Step 3. After completion of a network design process, initialize the proposed model.

Step 4. After initialization, for the chosen application, collect the data set and perform the data normalization process to eliminate the data discrepancy and missing data and improve the output accuracy. Meanwhile, divide the collected data set into two portions: the training phase and the testing phase.

Step 5. After the data collection and normalization process, learn the proposed neural network with a training data set.

Step 6. Check the proposed neural network's output and whether there is a minimal lead error and verify if the performance is acceptable or not. If the proposed neural network model's performance during the training phase is satisfactory with minimal error, then go for the testing phase. Else, the performance was not up to the mark; then again go to Step 2. The process continues until a match with the set goal.

Step 7. After completing the training phase, perform the proposed neural network model's testing process on the testing data set (the testing data sets are unseen raw data specified during the training phase).

Step 8. After completion of the testing phase, check if the proposed neural network model can achieve generic performance and generalize it well or not. If it generalizes well, record the output of the proposed neural network model; else again perform the remodeling of the neural network to

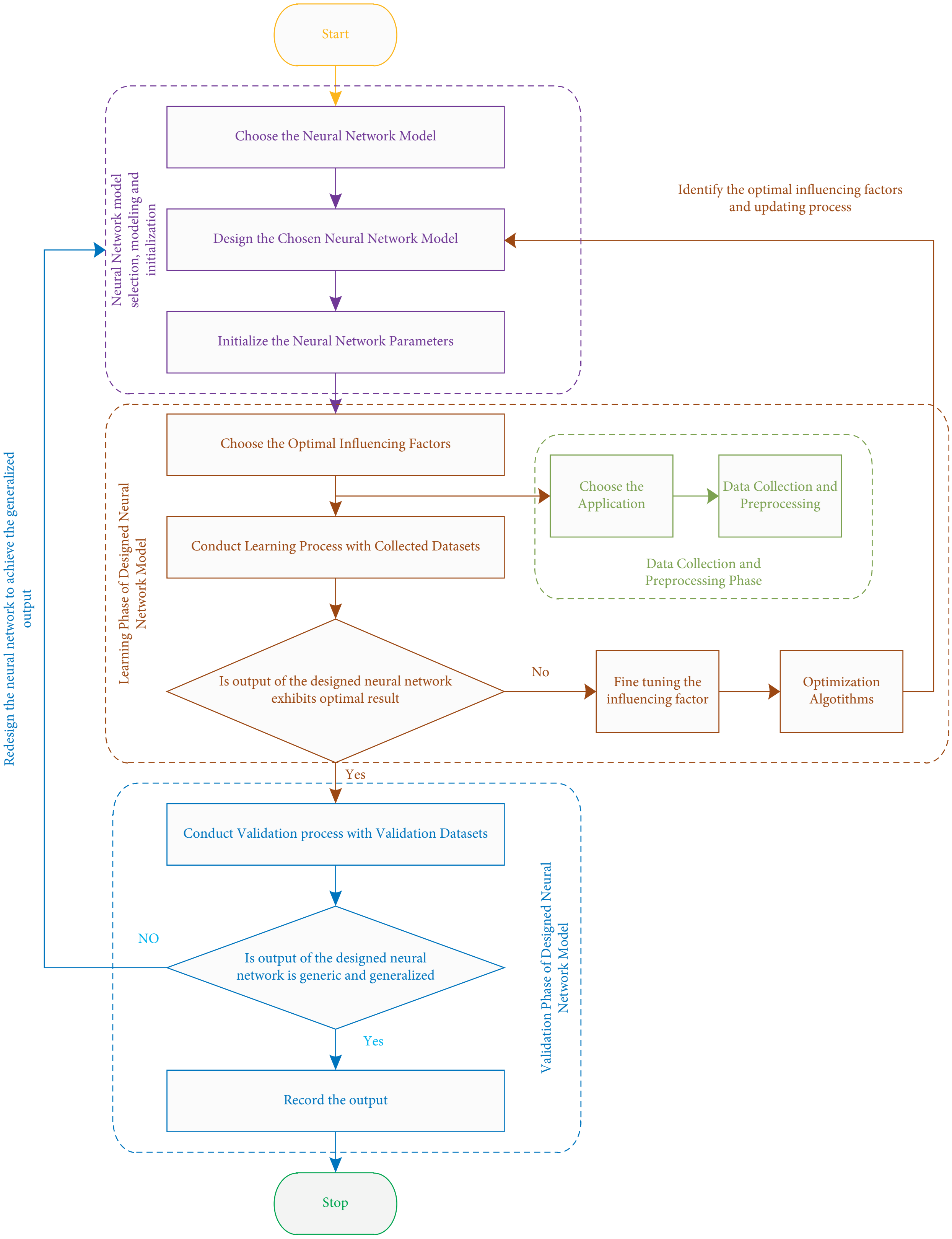

Start
Choose the Neural Network Model
Design the Chosen Neural Network Model
Initialize the Neural Network Parameters
Neural Network model selection, modeling and initialization
Identify the optimal influencing factors and updating process
Choose the Optimal Influencing Factors
Choose the Application
Data Collection and Preprocessing
Conduct Learning Process with Collected Datasets
Data Collection and Preprocessing Phase
Learning Phase of Designed Neural Network Model
Is output of the designed neural network exhibits optimal result
No
Fine tuning the influencing factor
Optimization Algotithms
Yes
Redesign the neural network to achieve the generalized output
Conduct Validation process with Validation Datasets
NO
Is output of the designed neural network is generic and generalized
Validation Phase of Designed Neural Network Model
Yes
Record the output
Stop

Figure 1: Generalized algorithm of artificial neural network.

achieve the generalized outcome (both training and testing phases lead to an optimal result with minimal error).

Step 9. Stop the process.

3. Applications of Artificial Neural Networks

Artificial neural network applications are not confined to the specific domain; it has a wide variety of applications. An artificial neural network has scope for various applications. Some applications are tabulated in Table 2. The artificial neural network has multiple applications, but not limited content, and it can cover all fields. Hence, it is an interdisciplinary field.

Virtual reality, decision support system (medical science and engineering fields), control engineering, data mining, computer vision, image, pattern recognition, human-machine interface, and few ANN applications. A system that can efficiently and intelligently solve the problem with computation ease is known as an expert system. It poses the ability to learn the environment, think, and apply the gained experience to complete the given task without assistance from the human being.

The system with many acquired knowledge based on many examples, wholly deriving the description of patterns and acquiring knowledge, is a complex problem in pattern recognition, natural language processing, speech, and computer vision applications. The human recalls the pattern, but the machine recall of the data pattern can be in the form of handwriting speech even though the sound of communication is different at various levels in the case of appropriate human identification . The pattern was not clearly defined but also was based on the knowledge humans identified; why not a machine can identify? This question arises from the leading research for the development of existing models in the field of ANN. The human identifies and recognizes the pattern or input based on various samples and examples' continuous learning ability.

In medical practice, ANN is used for medical disease diagnostics because identifying the disease is a challenging task for the medical practitioner and doctor because there will be an overlap of symptoms of various illnesses. Hence, there are no specific guidelines about disease identification based on the medical practice's experience and knowledge; doctors are suggesting the appropriate treatment for the patients. Sometimes, due to human error, patients were affected and they suffered unfair treatment because of the diseases' improper identification. To mitigate the above-said problems, nowadays, an artificial neural network occurs with a vital role in diagnosing patients' conditions.

A game-playing, self-regulated vehicle, self-control expert system, natural language processing, robotics, etc., are some thrust research fields of AI.

Forecasting, regression, classification, and diagnosis of diseases is based on the medical field's symptoms, computer vision, natural language processing (NLP), engineering, and science applications; the artificial neural network is widely used nowadays because of the promising solution to the challenging problem.

4. Proposed Various Artificial Neural Network for Various Forecasting Applications

Although artificial neural networks are suitable for various applications, this paper carries out modeling and analyzes artificial neural networks' effectiveness in forecasting applications like solar irradiance forecasting, rainfall forecasting, and temperature forecasting to aid sustainability. The roadmap of the proposed forecasting model is shown in Figure 2.

4.1. Proposed Multilayer Perceptron Neural Network for Solar Irradiance Forecasting. This paper proposed five meteorological input parameters based on a multilayer perceptron neural network [27]. The multilayer perceptron neural network (MLPNN) consists of one or over one hidden layer, which performs better in computational efficiency than a single-layer perceptron neural network. It belongs to the feedforward neural network and is associated with a supervised learning rule to explore synaptic weight values, and it has a complex problem-solving ability.

The proposed multilayer perceptron neural network is arranged into an input layer, one or more hidden layers, and an output layer. Regarding the hidden layers and nonlinear transfer function, the multilayer perceptron neural network can handle linear and nonlinear relationships between the input and output vectors. In the hidden layer, the hyperbolic tangent sigmoid activation function (nonlinear transfer function) is adopted, and the proposed neural network is trained by employing the backpropagation gradient descent learning rule. The proposed multilayer neural network induces sped-up convergence because it is a fully connected network.

The proposed multilayer perceptron neural network-based forecasting model for solar irradiance forecasting considers the solar irradiance impacting parameters as the inputs such as Solar Irradiance (SI), Temperature (TD), Wind Speed (WS), Dew Point (DP), and Cloud Cover (CC). These five meteorological input parameters are of much impact on solar irradiance. Hence, these parameters are accounted as the input parameters for the proposed neural network. The forecasted solar irradiance is viewed as the output neuron in the single output layer.

The proposed five input-based multilayer perceptron neural networks aim to achieve the best solar irradiance forecast to reduce the minimum error values. The proposed five input-based multilayer perceptron neural networks for solar irradiance structural design are depicted in Figure 3. Table 3 presents the proposed multilayer perceptron neural network design parameters. For the proposed neural network, model-independent computation is performed for each of the layers present in the neural network with respect to the received data. The computed outcomes are transferred to the next immediate layer as an input. Then, the neural network output is then obtained from the output layer; this process flow is clearly understood by Figure 3. During the training process, the involved computations are given as follows.

Table 2: Applications of ANN in various fields.

	Field	Application
ANN (Artificial Neural Network)	Computer science and engineering	NLP (natural language processing) Human machine interface Image processing Virtual reality Data mining Pattern recognition Image and pattern classification
	Electrical and electronics engineering	Fault identification Process and control system Systems integration Forecasting of energy and power
	Civil engineering	Design of optimal structure Material selection and decision making
	Mechanical engineering	Robotics Process optimization Material design
	Aeronautical engineering	Aircraft design Satellite and space application
	Medical science	Classification, identification, and diagnosis of diseases Decision support system
	Environmental engineering	Weather forecasting Rainfall forecasting
	Other fields	Economic forecasting Forex forecasting

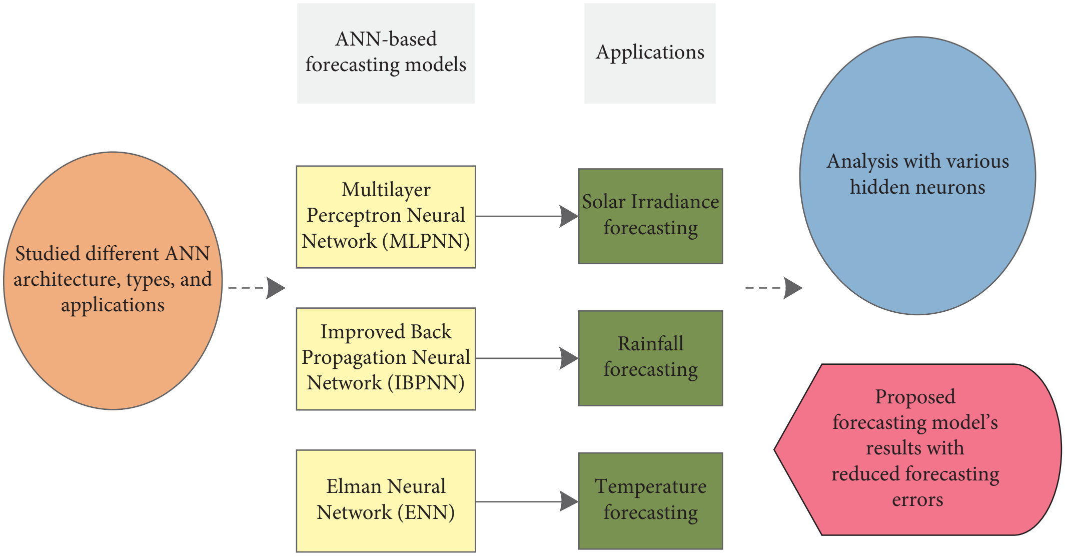

ANN-based forecasting models
Applications
Studied different ANN architecture, types, and applications
Multilayer Perceptron Neural Network (MLPNN)
Improved Back Propagation Neural Network (IBPNN)
Elman Neural Network (ENN)
Solar Irradiance forecasting
Rainfall forecasting
Temperature forecasting
Analysis with various hidden neurons
Proposed forecasting model's results with reduced forecasting errors

Figure 2: Proposed forecasting model road map.

The proposed MLPNN input vector:

$$K = [SI, TD, WS, DP, CC]. \tag{1}$$

The proposed MLPNN output vector:

$$J = [SI_f]. \tag{2}$$

The proposed MLPNN synaptic weight vectors between the input vector to the hidden vector:

$$SV = \begin{bmatrix} SV_{11}, SV_{12}, \ldots, SV_{1h}, SV_{21}, SV_{22}, \ldots, SV_{2h}, SV_{31}, SV_{32}, \ldots, SV_{3h}, \\ SV_{41}, SV_{42}, \ldots, SV_{4h}, SV_{51}, SV_{52}, \ldots, SV_{5h} \end{bmatrix}. \tag{3}$$

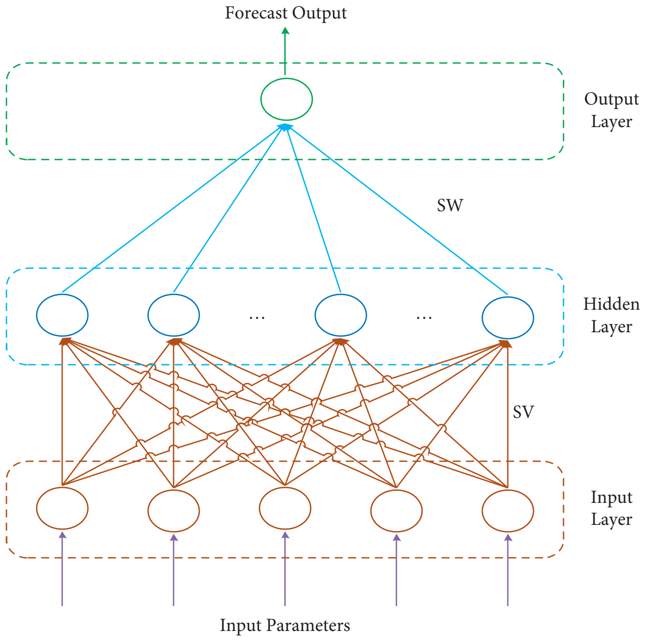


FIGURE 3: Structural design of the proposed MLPNN.

TABLE 3: Designed parameters of the proposed various artificial neural networks.

Improved backpropagation neural network	Multilayer perceptron neural network	Elman neural network
Input neuron = 5	Input neuron = 5	Input neuron = 5
Hidden layer = 1	Hidden layer = 1	Hidden layer = 1
Output neuron = 1 RF_f	Output neuron = 1 SI_f	Output neuron = 1 TD_f
Epochs = 2000	Epochs = 2000	Epochs = 2000
Learning rate = 0.01	Learning rate = 0.1	Learning rate = 0.1
Momentum factor = 0.9	Threshold = 1	Threshold = 1
Threshold = 1		

The proposed MLPNN obtains the net input of the hidden layer and subsequently the output of the hidden layer,

$$Y_q = f\left(\sum_{p=1}^{5}\sum_{q=1}^{h} K_p SV_{pq}\right), \tag{4}$$

where K is the input vector, SV is the synaptic weights between the input layer and hidden layer, and h is the number of hidden neurons.

The proposed MLPNN synaptic weight vectors between the hidden to output vector:

$$SW = [SW_1, SW_2, \ldots\ldots, SW_h]. \tag{5}$$

The proposed MLPNN obtains the net input of the output layer and subsequently its neural network final output:

$$Z = f\left(\sum_{q=1}^{h}\left(Y_q SW_q\right)\right), \tag{6}$$
$$q = 1, 2, \ldots.., h,$$

where SW is the synaptic weight between the hidden layer and output layer, and "f" is the activation function, namely, the tangent sigmoidal activation function.

4.2. Proposed Improved Backpropagation Neural Network for Rainfall Forecasting. The multilayer feedforward neural network with a momentum-based backpropagation learning algorithm is known as the proposed Improved Backpropagation Neural Network (IBPNN) [28]. IBPNN can balance between generalization and the network's memorization. The proposed improved backpropagation neural network is arranged into the input layer, hidden layer, and output layer. The proposed improved backpropagation neural network training stages are classified into three phases: feedforward stage, error computation stage, weight modification, and update stage. The processing elements present in the proposed feedforward neural networks perform an independent computation based on a considered set of input data and synaptic weights with a continuous differential activation function, and the obtained outcomes are passed to the successive layers, and then a final output is achieved from the output layer, which is compared with the target for error computation. Evaluated error is propagated backward via the output layer-hidden layer-input layer to achieve minimal error.

For the proposed improved backpropagation neural network based on the given set of training inputs and target pairs, the synaptic weights changed and updated, leading to accurate rainfall forecasting with minimal error. The structural design of the proposed five-input-based improved backpropagation neural network for rainfall forecasting is illustrated in Figure 4.

The rainfall is influenced by various variables such as temperature, precipitation of water content, wind speed, and relative humidity. Hence, the proposed neural networks consider these variables as the neural network's inputs to overcome the variance in the atmosphere. The proposed improved backpropagation neural network learning algorithm includes a momentum factor, making the neural network get faster convergence.

$(K_1, K_2, K_3, K_4, K_5\colon J) =$ (Rainfall, Precipitation of Water Content, Temperature, Relative Humidity, Wind Speed: Forecasted Rainfall).

$$\left(K_1, K_2, K_3, K_4, K_5\colon J\right) = \left(RF, PWC, TD, RH, WS\colon RF_f\right), \tag{7}$$

where RF_f is the forecasted rainfall.

The proposed IBPNN input vector:

$$K = [RF, PWC, T\ D, RH, WS]. \tag{8}$$

The proposed IBPNN output vector:

$$J = \left[RF_f\right]. \tag{9}$$

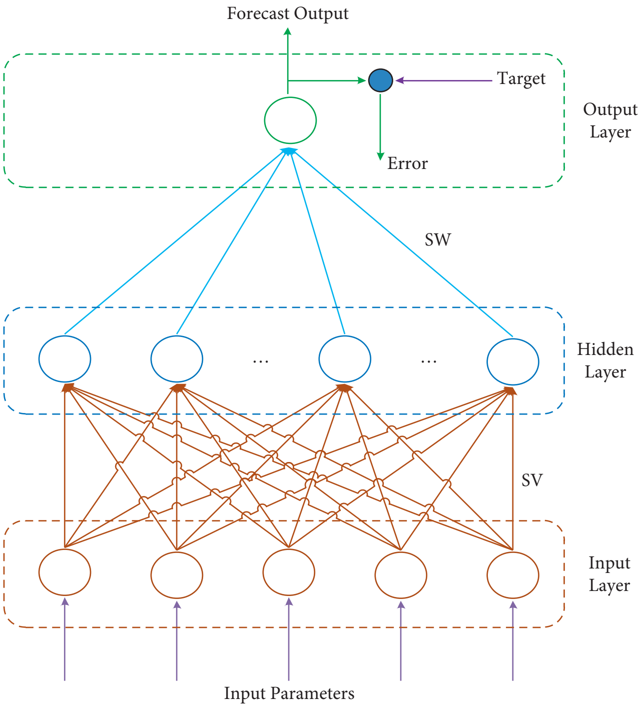


FIGURE 4: Structural design of IBPNN.

The proposed IBPNN synaptic weight vectors from the input layer to the hidden layer:

$$SV = \begin{bmatrix} SV_{11}, SV_{12}, \ldots, SV_{1h}, SV_{21}, SV_{22}, \ldots, SV_{2h}, SV_{31}, SV_{32}, \ldots, SV_{3h}, \\ SV_{41}, SV_{42}, \ldots, SV_{4h}, SV_{51}, SV_{52}, \ldots, SV_{5h} \end{bmatrix}. \tag{10}$$

The proposed IBPNN net input of the hidden layer:

$$Y_{inq} = \sum_{p=1}^{5} \sum_{q=1}^{h} K_p SV_{pq}. \tag{11}$$

The proposed IBPNN output of the hidden layer:

$$Y_q = f\left(\sum_{p=1}^{5} \sum_{q=1}^{h} K_p SV_{pq}\right). \tag{12}$$

where K is the input of IBPNN, SV is the synaptic weights between the input layer and hidden layer, and h is the number of hidden neurons.

The proposed IBPNN synaptic weight vectors from the hidden layer to the output layer:

$$SW = [SW_1, SW_2, \ldots\ldots, SW_h]. \tag{13}$$

The proposed IBPNN net input of the output layer:

$$Z_{in} = \sum_{q=1}^{h} \left(Y_q SW_q\right). \tag{14}$$

The proposed IBPNN output:

$$Z = f\left(\sum_{q=1}^{h} \left(Y_q SW_q\right)\right), \quad q = 1, 2, \ldots\ldots, h, \tag{15}$$

where SW is the synaptic weight between the hidden layer and the output layer and f is the activation function.

The proposed IBPNN computed error in the output layer:

$$E = (T_r - Z) f'(Z_{\text{in}}), \tag{16}$$

where $f'(Z_{\text{in}})$ is the derivative of the net input of the output layer.

The evaluated error (E) is propagated back to the hidden layer.

In the proposed IBPNN, each hidden neuron $(Y_q, q = 1, 2, \ldots, h)$ sums its delta inputs from the output layer neurons:

$$E_{\text{in}q} = \sum_{q=1}^{h} ESW_q. \quad (17)$$

The proposed IBPNN error in the hidden layer:

$$E_q = E_{\text{in}q} f'\left(Y_{\text{in}\,q}\right), \quad (18)$$

where $f'(Y_{\text{in}q})$ is the derivative of the net input of the hidden layer.

The computed error (E_q) is propagated back to the input layer to minimize the error during the backpropagation stage:

$$\text{The proposed IBPNN error in the output layer, } = [E]. \quad (19)$$

$$\text{For the proposed IBPNN error in the hidden layer, } = \left[E_j\right]. \quad (20)$$

The proposed improved backpropagation neural network mathematical equations for the synaptic weight updating process were as follows:

$$SW_q(n+1) = SW_q(n) + \alpha EY_q + \eta\left[SW_q(n) - SW_q(n-1)\right], \quad (21)$$

$$SV_{pq}(n+1) = SV_{pq}(n) + \alpha E_q k_p + \eta\left[SV_{pq}(n) - SV_{pq}(n-1)\right], \quad (22)$$

where α is the learning rate and η is the momentum factor.

The synaptic weights are updated and changed using the mathematical equations (21) and (22). The learning stages of the proposed neural network and weight updating process are continued in the proposed IBPNN until attaining the stopping condition (i.e., set value).

4.3. Proposed Elman Neural Network for Temperature Forecasting. The feedback neural network has advantages over the feedforward network; with feedback from the output, the neural network stability and performance can be improved. In a feedback neural network, Elman neural network (ENN) is a famous feedback neural network, which Elman suggested in 1990 [29–31]. Because of the superior performance, Elman network can be used for various applications such as forecasting, speech recognition, modeling, and control. Like the feedforward neural network arranged into the input layer, hidden layer, and output layer, one more layer is also added into the feedback neural network: the feedback layer or recurrent layer. The feedback storage and memory retaining have been done with the help of the recurrent layer. It also copies the one-step delay in the hidden layer. With the help of the internal connection proposed, Elman neural network dynamic characteristics are achieved. The proposed Elman neural network-based forecasting model hidden layer is associated with the hyperbolic tangent sigmoid activation function, and the output layer is associated with the purelin activation function.

The proposed temperature forecasting model is developed using an Elman neural network with five inputs such as Temperature $(T\ D)$, Dew Point (DP), Solar Irradiance (SI), Wind Speed (WS), and Relative Humidity (RH). The proposed Elman neural network complexity is reduced to a single hidden layer with various hidden neurons and one output layer with a single output neuron, i.e., forecast temperature. The proposed Elman neural network's objective is to achieve the accurate forecasting of temperature with reduced convergence time and minimal error. The structural design of the Elman neural network is shown in Figure 5.

During the training process, the involved computations are given as follows:

$(K_1, K_2, K_3, K_4, K_5\colon J)$ = (Temperature, Dew Point, Solar Irradiance, Wind Speed, Relative Humidity: Forecast Temperature).

$(K_1, K_2, K_3, K_4, K_5\colon J) = (T\ D, DP\ , SI, WS, RH\colon T\ D_f)$, where TD_f is the Forecast Temperature in Degrees.

Let SV_c be the synaptic weights between the context layer and the input layer.

Let SV be the synaptic weights between the input layer and the hidden layer.

Let SV_2 be the synaptic weights between the hidden layer and the recurrent link layer.

Let SW be the synaptic weights between the hidden layer and the output layer.

Let $h(\cdot)$ be the activation function, namely, the hyperbolic tangent sigmoid activation function adopted for the hidden layer.

Let $f(\cdot)$ be the activation function, namely, the purelin activation function which is adopted for the output layer.

Figure 5 infers that the proposed Elman neural network comprising each layer is performing independent computations on receiving data. The obtained output is passed to the next successive layer, and after that, finally, from the output layer, the neural network output is computed. The proposed Elman neural network has the ability that previous input influences the current input responses. The input $K(X-1)$ passes through the hidden layer that multiplies the synaptic weight (SV) with the association of the hyperbolic tangent sigmoid activation function. In addition to the previous state output $SV_c K_c(X)$, the current input $SVK(X-1)$ was also added, which aids the proposed feedback neural network to efficiently learn the function. The value $K(X)$ is passed through an output layer multiplied by the synaptic weight SW with the association of the purelin activation function.

The proposed ENN input vector:

$$K = [TD, DP, SI, WS, RH]. \quad (23)$$

The proposed ENN output vector:

$$J = \left[TD_f\right]. \quad (24)$$

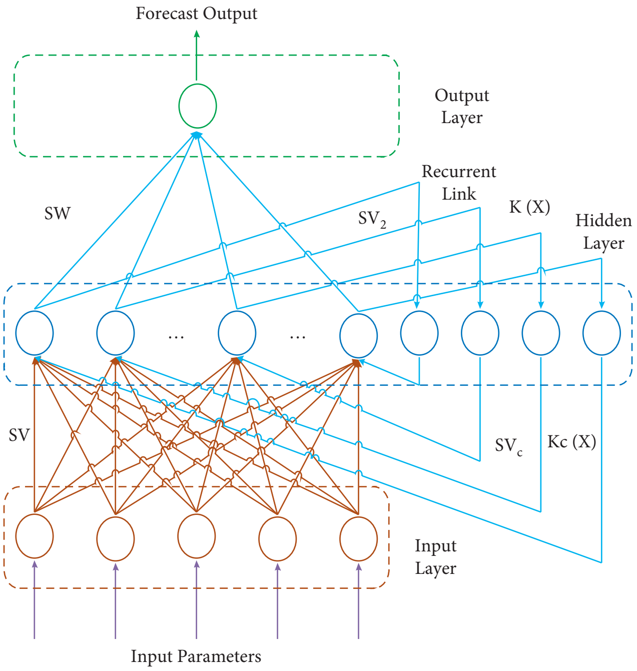


Figure 5: Structural design of ENN.

The proposed ENN synaptic weight vector between inputs to the hidden vector:

$$SV = \begin{bmatrix} SV_{11}, SV_{12}, \ldots, SV_{1h}, SV_{21}, SV_{22}, \ldots, SV_{2h}, SV_{31}, SV_{32}, \ldots, SV_{3h}, \\ SV_{41}, SV_{42}, \ldots, SV_{4h}, SV_{51}, SV_{52}. \ldots, SV_{5h} \end{bmatrix}. \tag{25}$$

The proposed ENN synaptic weight vector of the recurrent link layer vector:

$$SV_2 = [SV_{21}, SV_{22}, \ldots ., SV_{2h}]. \tag{26}$$

The proposed ENN synaptic weight vector of the output layer vector:

$$SW = [SW_{11}, SW_{21}, \ldots ., SW_{ho}]. \tag{27}$$

The proposed ENN Synaptic weight vector of the recurrent link layer with the input vector:

$$SV_c = \begin{bmatrix} SV_{c11}, SV_{c12}, \ldots, SV_{c1h}, SV_{c21}, SV_{c22}, \ldots, SV_{c2h}, SV_{c31}, SV_{c32}, \ldots, \\ SV_{c3h}, SV_{c41}, SV_{c42}, \ldots, SV_{c4h}, SV_{c51}, SV_{c52}, \ldots, SV_{c5h} \end{bmatrix}. \tag{28}$$

The proposed ENN input:

$$K(X) = h(SV_c K_c(X) + SVK(X-1)). \tag{29}$$

The proposed ENN output:

$$J(X) = f(SWK(X)). \tag{30}$$

The proposed ENN input of the recurrent link layer:

$$K_c(X) = K(X-1). \tag{31}$$

4.4. Proposed Artificial Neural Network Experimental Implementation. In the current year, forecasting is a vital tool because people can plan according to it. Solar irradiance forecasting is a crucial factor in the solar energy system. Based on the forecasted solar irradiance, solar energy can be estimated to operate better than the power system and the possibility of extracting maximum solar power. It carried out many research works in solar irradiance forecasting [32–34]. Although a more accurate solar irradiance forecasting model is still needed, it motivates the authors to develop multilayer perceptron neural network-based solar irradiance forecasting. The meteorological center tries to forecast the precise prediction of weather reports (Temperature) and rainfall. The predicted output does not match the current values because these values are highly influenced by the atmosphere variables like wind speed, cloud cover, precipitation of water content, etc. Some researchers endeavor to forecast the rainfall [35–37] and temperature [38–40] because of the volatile nature outperforming the generic forecasting model required. This fact motivates the development of the high accurate forecasting model for rainfall and temperature forecasting based on the proposed feedforward neural network (i.e., improved backpropagation neural network) and feedback neural network (i.e., Elman neural network), respectively.

The design and development stage, training stage, and testing stage are the three stages of the proposed artificial neural networks. The input-related data for the chosen applications are gathered, normalized into the range of zero to one through min-max normalization (variance eliminated), and then the training and testing process of the designed neural network is carried out with the acquired training and testing data sets, respectively. The validity of the proposed neural network-based forecasting model has been proven and analyzed based on the computed error validation process. It requires the proper design in neural network modeling because the improper selection of design parameters leads to poor performance.

4.41. Data Collection. For the considered solar irradiance forecasting applications, rainfall forecasting and temperature-related real-time input parameters are acquired from the National Oceanic and Atmospheric Administration, United States.

For the considered solar irradiance forecasting, the related inputs (Solar Irradiance (*SI*), Temperature (*TD*), Wind Speed (*WS*), Dew Point (*DP*), and Cloud Cover (*CC*)) data samples are acquired from the period of January 2014 to December 2019, which comprise 175200 total number of data samples of each input.

For the considered rainfall forecasting, the related inputs (Rainfall (*RF*), Precipitation of Water Content (*PCW*), Temperature (*T D*), Relative Humidity (*RH*), and Wind Speed (*WS*)) are acquired from the period of April 2014 to April 2019, which consist of 5256 0 data samples of each input.

For the considered temperature forecasting application, the related input parameters (Temperature (*TD*) , Dew Point (*DP*), Solar Irradiance (*SI*), Wind Speed (*WS*), and Relative Humidity (*RH*)) are acquired from the period of March 2012 to April 2019, which consists of 1051200 data samples of each considered input.

4 .4 .2.Data Normalization. The data normalization is required because the data collected from the resource center are real-time data that possess the variance with respect to various ranges and various units to remove the variance of the acquired real-time data. The data normalization process, collecting data irrespective of multiple ranges and different units, classifies the data in the range from 0 to 1. In data normalization, various methods are available for this proposed work; min-max normalization adopts the proposed artificial neural network-based forecasting model. The proposed artificial neural network's numerical computation and accuracy can be improved by employing data normalization. The following transformational mathematical equation is used for the normalization of the real-time collected data.

$$\text{Normalized input, } K'_p = \left(\frac{K_p - K_{\min}}{K_{\max} - K_{\min}}\right)\left(K'_{\max} - K'_{\min}\right) + K'_{\min}, \tag{32}$$

where K_p is the collected real-time input data, $K_{\min}$ is the minimal input data, $K_{\max}$ is the maximum input data, $K_{\min}{}'$ is the minimal target value, and $K_{\max}{}'$ is the maximum target value.

4.4.3. Proposed Artificial Neural Network Modeling. The designed parameters of the proposed artificial neural networks are presented in Table 3. The proposed various artificial neural networks modeling parameter dimensions such as the number of input neurons, hidden neurons, output neurons, the number of epochs, learning rate, momentum factor, and the threshold value are tabulated in Table 3.

The implemented multilayer-perceptron-neural-network- (MLPNN-) based solar irradiance forecasting model inputs are passed to the hidden layer that is multiplied by synaptic weight (SV) with hyperbolic tangent sigmoid activation function. The hidden layer's output is passed to the output layer that is multiplied by synaptic weight (SW) with the purelin activation function. We use the Levenberg-Marquardt training algorithm for the proposed MLPNN-based solar irradiance forecasting model training process.

The proposed improved-back-propagation-neural-network- (IBPNN-) based rainfall forecasting model inputs are transmitted to the hidden layer multiplied by the synaptic weight (SV) utilizing a hyperbolic tangent sigmoid activation function. The hidden layer's output is transmitted to the output layer multiplied by the synaptic weight (SW) with the tangential sigmoid activation function. The training algorithm used for IBPNN is Levenberg- Marquardt backpropagation training algorithm. The momentum factor is included in the learning algorithm, which leads to speed-up convergence.

The designed Elman-neural-network- (ENN-) based temperature forecasting model and the synaptic input weights (SV) are interconnected to the hidden layer using the hyperbolic tangent sigmoid activation function. The hidden layer's output is interconnected to the output layer with synaptic weight (SW) using the purelin activation function. As a result of training, the previous inputs get reflected in the Elman neural network. The training algorithm used for the proposed Elman neural network is gradient descent with momentum and an adaptive linear backpropagation training algorithm.

For all proposed artificial-neural-network-based forecasting models, training and testing are done through the normalized data set. The validation process is continued until the stopping condition is reached.

4.4.4. Selection of Number of Hidden Neurons in the Proposed Artificial Neural Networks. The most challenging process in the artificial neural network is selecting the required number of hidden neurons to place in the artificial neural network [41–45]. There is no generalized formulation and criterion available for selecting hidden neurons in an artificial neural network. The random selection and trial-and-error methods also take much more time. If the hidden neurons are too low and too high, both condition neural networks do not achieve optimal results. Hence, the proposed artificial neural network, namely, multilayer neural network, improved backpropagation neural network, and Elman neural network with a single hidden layer, is preferred because the neural network with a single hidden layer can solve the problem with less computational difficulty. In that single hidden layer, the hidden neurons are varied from one (1) to fifteen (15). The designed neural network is validated for each considered hidden neuron, and the obtained results are tabulated. According to Tables 4–6, appropriated numbers of hidden neurons are identified for the proposed artificial neural network based on the minimal error and minimal convergence time.

4.4.5. Training and Testing of the Proposed Artificial Neural Network Performance. The proposed solar irradiance forecasting model is built using the training data set. The proposed MLPNN-based solar forecasting model performance is verified based on the testing data set. The acquired data are classified into two sets, like training and testing. The training set comprises 70 percentages of the obtained data samples, and the testing set comprises the remaining unseen 30 percentages of acquired data samples. For the solar irradiance forecasting, the acquired data samples are 175200 real-time data samples, 70% data samples (122640) are used for the training stage, and the unseen data samples (52560) are used for the testing stage of the proposed neural network.

The proposed rainfall forecasting model is built on the training data set, and the proposed IBPNN based rainfall forecasting model performance is verified based on the testing data set. The acquired data are classified into two sets, training and testing, respectively. The training set comprises 70 percentages of the obtained data samples, and the testing set consists of the remaining unseen 30 percentages of obtained data samples. For rainfall forecasting, the acquired data samples are 52560 real-time data samples, 70% data samples (36792) are used for the training stage, and the unseen data samples (15768) are used for the proposed testing stage neural network.

The proposed temperature forecasting model is built on the training data set, and the proposed ENN-based temperature forecasting model performance is verified based on the testing data set. The acquired data are classified into two sets, like training and testing. The training set consists of 70 percentages of the obtained data samples, and the testing set consists of the remaining unseen 30 percentages of acquired data samples. For temperature forecasting, the acquired data samples are 1051200 real-time data samples, 70% of data samples (735840) are used for the training, and 30% of the unseen data samples (315360) are used for the testing stage of the proposed neural network.

The proposed artificial neural network performance is validated based on the training and testing set. The error qualifiers like R, MAPE, MSE, MAE, RMSE, MRE, and Time are used to verify the proposed artificial neural network performance. The number of hidden neurons that leads to reduced error is fixed as the optimal number of hidden neurons in the proposed artificial neural network.

4.4.6. Error Qualifier of the Proposed Artificial Neural Networks. The performance of the proposed various artificial neural networks such as multilayer perceptron neural network, improved backpropagation neural network, and Elman neural network is verified by the evaluated error qualifiers, namely, Correlation Coefficient (R), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Relative Error (MRE), and Time in Minutes. The proposed artificial neural network effectiveness is evaluated based on the error qualifiers equations (33)–(38). The following mathematical equations are used for the error computation:

$$MAPE = \frac{100}{N} \sum_{p=1}^{N} \left| \frac{\left(K_p' - K_p^f\right)}{\overline{K}_p} \right|, \tag{33}$$

$$MSE = \frac{1}{N} \sum_{p=1}^{N} \left(K_p' - K_p^f\right)^2, \tag{34}$$

$$MAE = \frac{1}{N} \sum_{p=1}^{N} \left(\left|K_p' - K_p^f\right|\right), \tag{35}$$

$$RMSE = \sqrt{\left(\frac{1}{N} \sum_{p=1}^{N} \left(K_p' - K_p^f\right)^2\right)}, \tag{36}$$

$$\mathrm{MRE} = \frac{1}{N} \sum_{p=1}^{N} \left| \frac{\left(K_p' - K_p^f\right)}{\overline{K}_p} \right|, \tag{37}$$

$$\mathrm{R} = 1 - \left(\frac{\sum_{p=1}^{N}\left(K_p' - K_p^f\right)}{\sum_{p=1}^{N} K_p^f}\right)^2. \tag{38}$$

where N is total number of data samples, K_p' is target output, $\overline{K}_p$is average target output, and $K_p^f$is forecasted output.

Table 4: The proposed five input-based multilayer perceptron neural network statistical performance analyses with various hidden neurons for solar irradiance forecasting.

Number of hidden neurons	Error qualifier						
	R	MAPE	MSE	MAE	RMSE	MRE	Time (min)
1	1	0.0029	7.1344*e*-05	0.0063	0.0084	2.9047*e*-05	3.17
2	1	0.0014	1.8264*e*-05	0.0031	0.0043	1.4451*e*-05	4.13
3	1	8.2990*e*-04	7.3104*e*-06	0.0018	0.0027	8.2990*e*-06	2.03
4	1	0.0020	4.3011*e*-05	0.0043	0.0066	1.9727*e*-05	2.32
5	1	7.2196*e*-04	5.7284*e*-06	0.0016	0.0024	7.2196*e*-06	3.32
6	1	0.0025	7.1011*e*-05	0.0055	0.0084	2.5175*e*-05	1.31
7	1	0.0026	7.4721*e*-05	0.0055	0.0086	2.5515*e*-05	2.03
8	1	8.2778*e*-04	7.9809*e*-06	0.0018	0.0028	8.2778*e*-06	1.48
9	1	7.3968*e*-04	6.2360*e*-06	0.0016	0.0025	7.3968*e*-06	3.32
10	1	7.5442*e*-04	6.5662*e*-06	0.0016	0.0026	7.5442*e*-06	3.50
11	1	0.0018	3.8319*e*-05	0.0040	0.0062	1.8279*e*-05	2.11
12	1	5.1926*e*-04	3.0656*e*-06	0.0011	0.0018	5.1926*e*-06	2.17
13	1	3.2149*e*-04	1.1717*e*-06	6.9914*e*-04	0.0011	3.2149*e*-06	5.04
14	1	**2.3034*e*-04**	**5.9791e-07**	**5.0093e-04**	**7.7324e-04**	**2.3034e-06**	**1.32**
15	1	0.0012	1.7649*e*-05	0.0027	0.0042	1.2348*e*-05	3.02

Bold implies the best results.

Table 5: The proposed five inputs based on improved backpropagation neural network statistical performance analysis with various hidden neurons for rainfall forecasting.

Number of hidden neurons	Error qualifier						
	R	MAPE	MSE	MAE	RMSE	MRE	Time (min)
1	0.98949	8.9725	657.7246	18.1105	25.6461	0.0897	1.21
2	0.1638	100	7.4967*e*+04	201.8431	273.8012	1.000	0.01
3	0.99991	0.7373	5.4600	1.4882	2.3367	0.0074	0.57
4	1	0.2696	0.6398	0.5441	0.7999	0.0027	1.59
5	1	0.0945	0.0864	0.1906	0.2939	9.4451*e*-04	4.15
6	1	0.0344	0.0143	0.0695	0.1194	3.4418*e*-04	1.54
7	**1**	**0.0158**	**0.0042**	**0.0320**	**0.0645**	**1.5841*e*-04**	**1.24**
8	1	0.0203	0.0077	0.0409	0.0880	2.0264*e*-04	1.36
9	1	0.0199	0.0079	0.0401	0.0869	1.9879*e*-04	1.49
10	1	0.0170	0.0066	0.0344	0.0815	1.7022*e*-04	1.52
11	1	0.0357	0.1512	0.0721	0.3889	3.5704*e*-04	1.13
12	1	0.0283	0.0842	0.0572	0.2902	2.8320*e*-04	3.47
13	1	0.0231	0.2452	0.0466	0.4956	2.3092*e*-04	3.10
14	1	0.0191	0.3369	0.0385	0.5804	1.9076*e*-04	5.40
15	1	0.0255	0.6127	0.0515	0.7828	2.5517*e*-04	8.31

Bold implies the best results.

4.5. Validation Experimental Results of the Proposed Artificial Neural Network for Various Forecasting Applications. The proposed artificial-neural-networks-based forecasting models are experimentally simulated in MATLAB platform version 2013 running on Acer computers with Pentium (*R*) Dual-Core processor running at 2.30 GHz with 2 GB of RAM.

4.5.1. Experimental Results of the Proposed Multilayer-Perceptron-Neural-Network-Based Solar Irradiance Forecasting. The designed multilayer-perceptron-neural-network-based solar irradiance forecasting model performance is validated experimentally, and the got results with various numbers of hidden neurons (1–15) are tabulated in Table 4 and analyzed. From the analysis of Table 4, it is observed that the proposed multilayer-perceptron-neural-network-based solar irradiance forecasting model performs better for various hidden neurons.

The proposed forecasting model output is changed drastically by varying the hidden neuron in the proposed multilayer perceptron neural network's hidden layer. The hidden neuron in the hidden layer increases further and makes the neural network unstable, while decreasing the hidden neuron further also leads the neural network to become unstable. Based on the achieved solar irradiance forecasting results among 1 to 15 hidden neurons, the proposed multilayer perceptron neural network contains a single hidden layer that possesses 14 hidden neurons that provide the best outputs with minimal error and reduced convergence time.

Therefore, the proposed multilayer perceptron neural network with five inputs, a single hidden layer, 14 hidden neurons in the hidden layer, and a single output neuron

Table 6: The proposed five input-based Elman neural network statistical performance analyses with various hidden neurons for temperature forecasting.

Number of hidden neurons	Error qualifier					
	MAPE	MSE	MAE	RMSE	MRE	Time (sec)
1	0.6612	0.1299	0.1488	0.3605	0.0066	22
2	1.2701	0.4218	0.2859	0.6494	0.0127	31
3	0.8163	0.1949	0.1837	0.4415	0.0082	29
4	0.6202	0.0684	0.1396	0.2615	0.0062	33
5	1.0752	0.3555	0.2420	0.5962	0.0108	34
6	0.4970	0.0821	0.1119	0.2866	0.0050	33
7	0.1069	0.0035	0.0241	0.0590	0.0011	44
8	0.1013	0.0015	0.0228	0.0384	0.0010	35
9	0.4501	0.0671	0.1013	0.2591	0.0045	42
10	0.9629	0.2676	0.2167	0.5173	0.0096	45
11	**0.1023**	**0.0011**	**0.0230**	**0.0332**	**0.0010**	**22**
12	0.3122	0.0319	0.0703	0.1787	0.0031	46
13	0.1715	0.0025	0.0386	0.0502	0.0017	39
14	0.4440	0.0473	0.0999	0.2175	0.0044	49
15	0.8641	0.2005	0.1945	0.4478	0.0086	51

Bold implies the best results.

structure is identified as the best framework. The obtained output plots based on this structural framework-associated forecasting model are shown in Figure 6. The number of data vs. solar irradiance is presented in Figure 7. The real-time target solar irradiance compared with forecasted solar irradiance is illustrated in Figure 8. Error vs. number of data is shown in Figure 9. Relationship between forecast solar irradiance and real-time target solar irradiance, respectively. Because of the space limitation, portions of the obtained results are shown in Figures 6–9. The 14 hidden neuron-based developed neural network (MLPNN) models forecasting solar irradiance are much matched with the real-time target solar irradiance. Hence, the error values are reduced to the minimal, clearly understood from Figures 7 and 8, respectively.

Figure 8 shows that the proposed neural network with 14 hidden neurons results in a minimal error on the considered data samples. The relationship between forecasted solar irradiance and real-time target solar irradiance is a linear relationship that illustrates that the forecasted solar irradiance accurately matches the real-time target. Hence, the proposed neural network (MLPNN) proved its validity, which is noticed in Figure 9.

4.5.2. Experimental Results of Proposed Improved Backpropagation Neural Network-Based Rainfall Forecasting. The proposed improved backpropagation neural network-based rainfall forecasting model performance is validated experimentally using the collected data, and the obtained results of the proposed neural network with various numbers of hidden neurons (1–15) are tabulated in Table 5. According to the obtained results in Table 5, it is noticed by careful analyses that the proposed rainfall forecasting model based on the improved backpropagation neural network performs very poorly for the hidden neurons 1 and 2 except that for the remaining hidden neurons results in the good output. The IBPNN-based forecasting model achieved outcome is changed drastically with the hidden neurons changes in the proposed improved backpropagation neural network's hidden layer.

In a neural network, the neural network stability and performance are highly affected by hidden neurons. Based on the analysis of the obtained results, it is observed that the proposed improved backpropagation neural network with a single hidden layer and seven hidden neurons in the hidden-layer-based design neural network (IBPNN) achieves superior performance in terms of reduced error and reduced convergence speed compared among other hidden neurons based on design neural networks.

Therefore, the proposed improved backpropagation neural network with five inputs, a single hidden layer, seven hidden neurons in the hidden layer, and a single output layer with one output neuron has been identified as the optimal structural framework of the proposed neural network. The obtained rainfall forecasting plots with seven hidden neurons-based design IBPNN are depicted in Figure 10. Rainfall in mm vs. data samples is shown in Figure 11. Original target rainfall compared with forecast rainfall is illustrated in Figure 12. Evaluation error metric vs. the number of data samples is presented in Figure 13, the relationship between forecast rainfall and original target rainfall, respectively. Because of the space limitation, portions of the obtained results are shown in Figures 10–13.

The considered data sample for the validation of the neural network is represented in Figure 10. In the seven-hidden-neurons-based developed backpropagation neural network model, where the forecast rainfall is a relative value compared with the original target rainfall, the forecasting accuracy is better which is understood from Figure 11. It is observed that the designed IBPNN achieves minimal errors for more clarity; the evaluation error with respect to the data samples is depicted in Figure 12. The relationship between forecast rainfall and original target rainfall is a linear relationship, which illustrates that the forecast rainfall is higher, accurately matched with the original target, which is clearly observed in Figure 13.

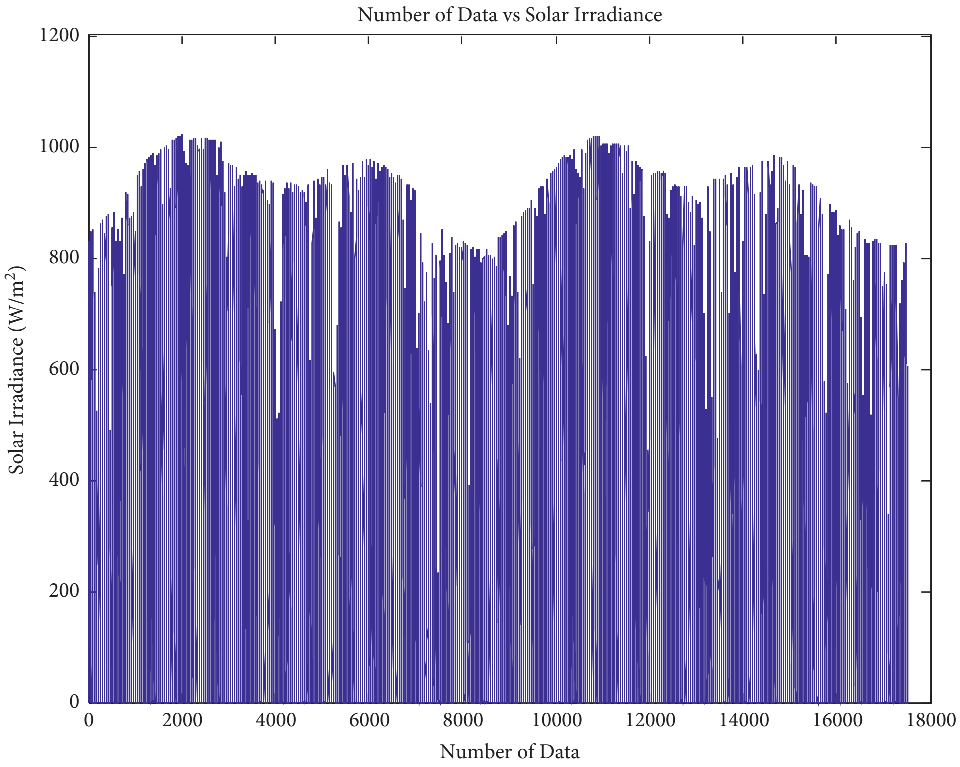


FIGURE 6: Number of data vs. solar irradiance.

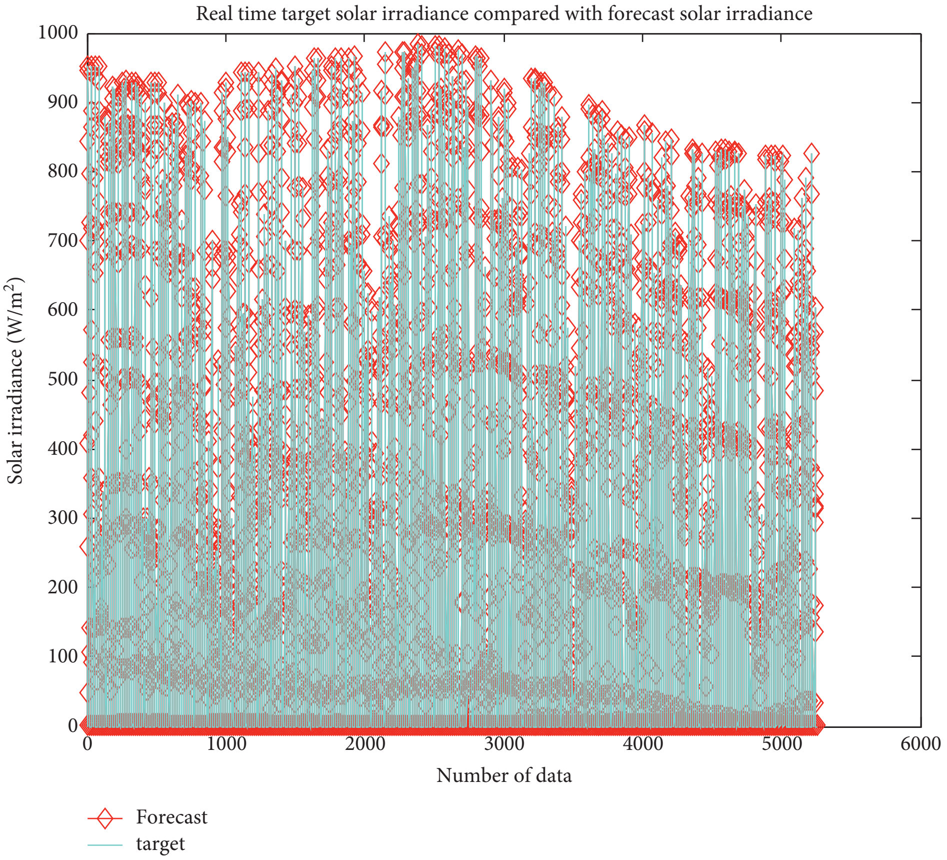


FIGURE 7: Real-time target solar irradiance compared with forecasted solar irradiance.

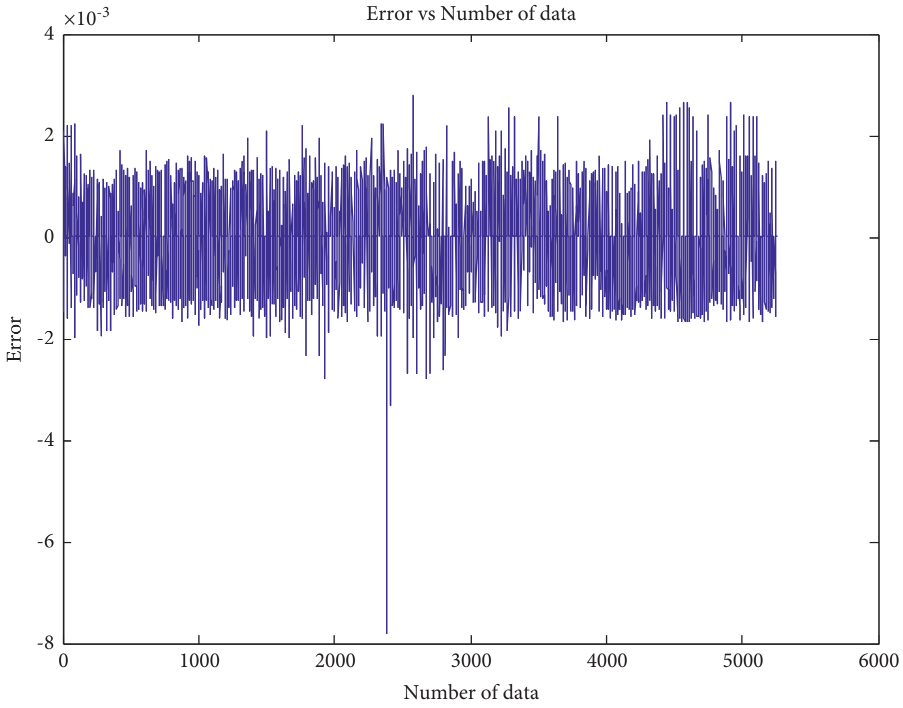


Figure 8: Error vs. the number of data.

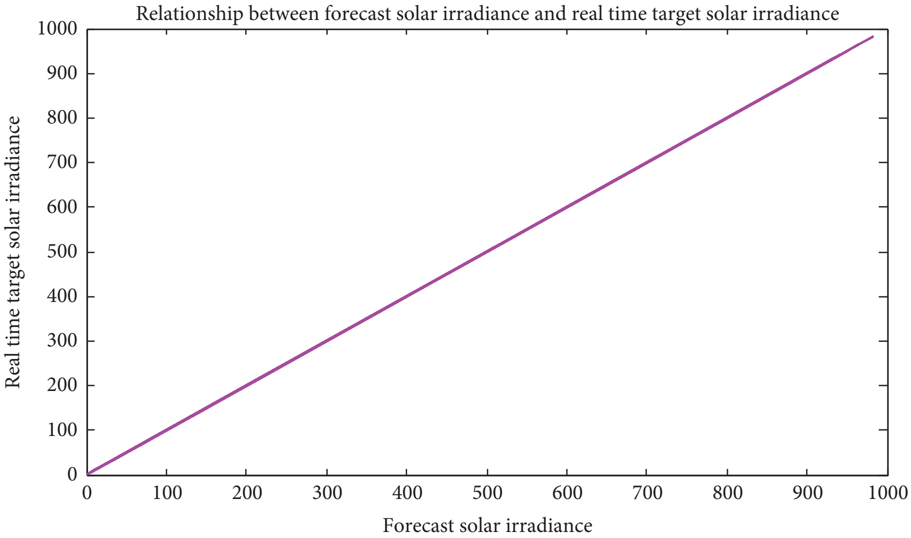


Figure 9: Relationship between forecast solar irradiance and real-time target solar irradiance.

4.5.3. Experimental Results of the Proposed Elman-Neural-Network-Based Temperature Forecasting. The proposed temperature forecasting models based on Elman neural network validated with the acquired data samples and achieved experimental outputs based on different numbers of hidden neurons from 1 to 15 are tabulated in Table 6. According to the achieved results in Table 6, it is observed that the proposed Elman-neural-network-based forecasting model performs well for all hidden neurons. The proposed Elman-neural-network- (ENN-) based temperature forecasting model output is changed drastically due to the number of hidden neurons varying in the proposed Elman neural network's hidden layer.

In a feedback neural network, hidden neurons profoundly influence the aspects of neural network stability and performance convergence. From the result analysis of the obtained outputs, it is noticed that the designed Elman neural network with a single hidden layer and 11 hidden neurons in the hidden layer achieves better performance in terms of reduced error and reduced convergence speed compared among other numbers of hidden neurons based on the designed neural networks. Therefore, the proposed

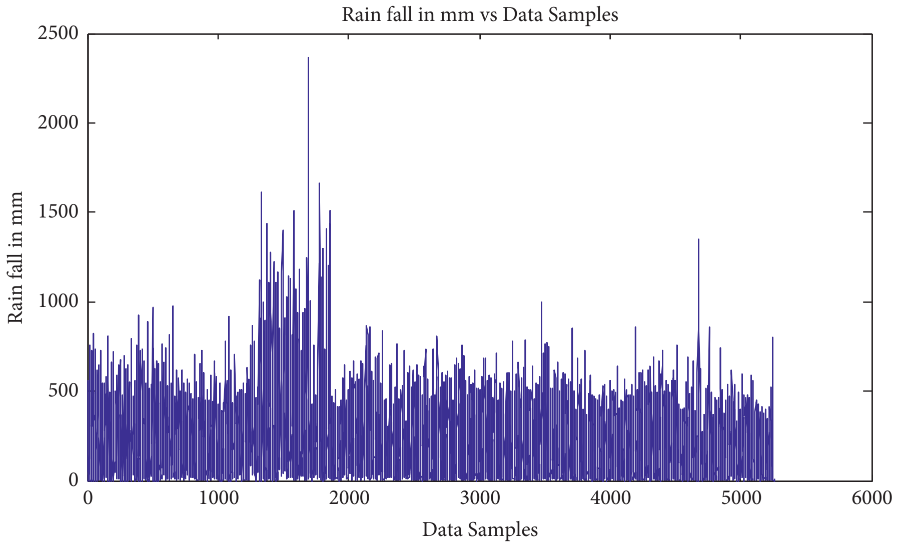


Figure 10: Rainfall in mm vs. data samples.

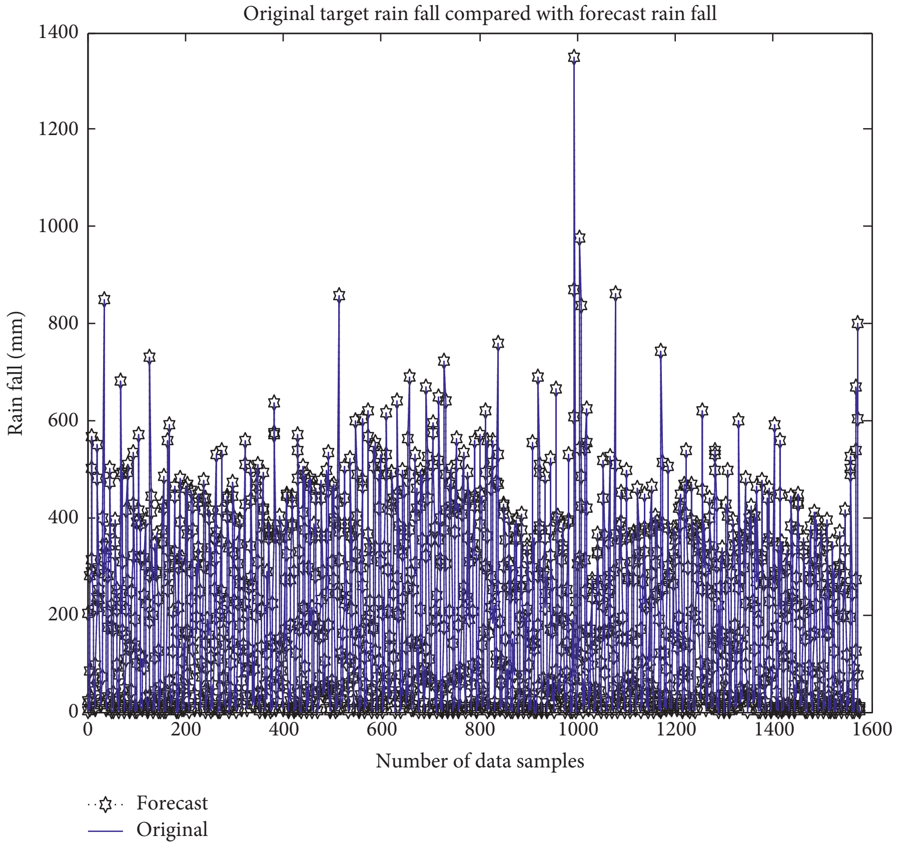


Figure 11: Original target rainfall compared with forecast rainfall.

temperature forecasting model based on Elman neural network with five inputs, a single hidden layer, 11 hidden neurons in the hidden layer, and a single output layer with one output neuron has been identified as the useful structural framework of the proposed neural network.

The obtained temperature forecasting plots with respect to the 11 hidden neuron-based design ENNs are shown in Figure 14. Temperature vs. data samples is presented in Figure 15. Comparison between target and forecasted temperature is presented in Figure 16. Error vs. number of data is shown in Figure 17, the relationship between target and forecast temperature, respectively. Due to the space limitation, portions of the obtained results are shown in Figures 14–17.

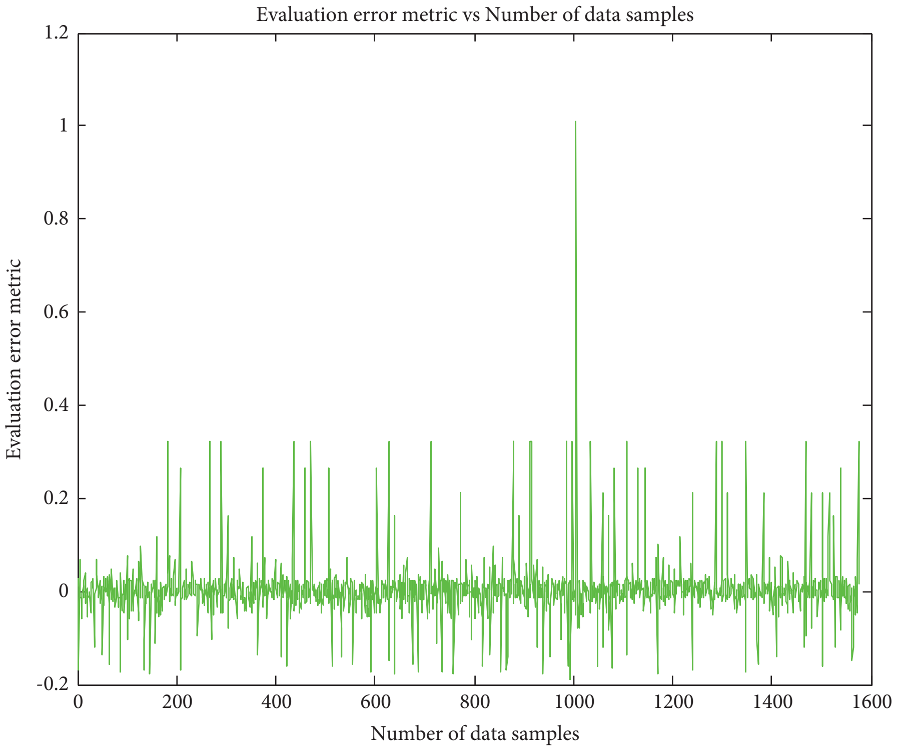


Figure 12: Evaluation error metric vs. the number of data samples.

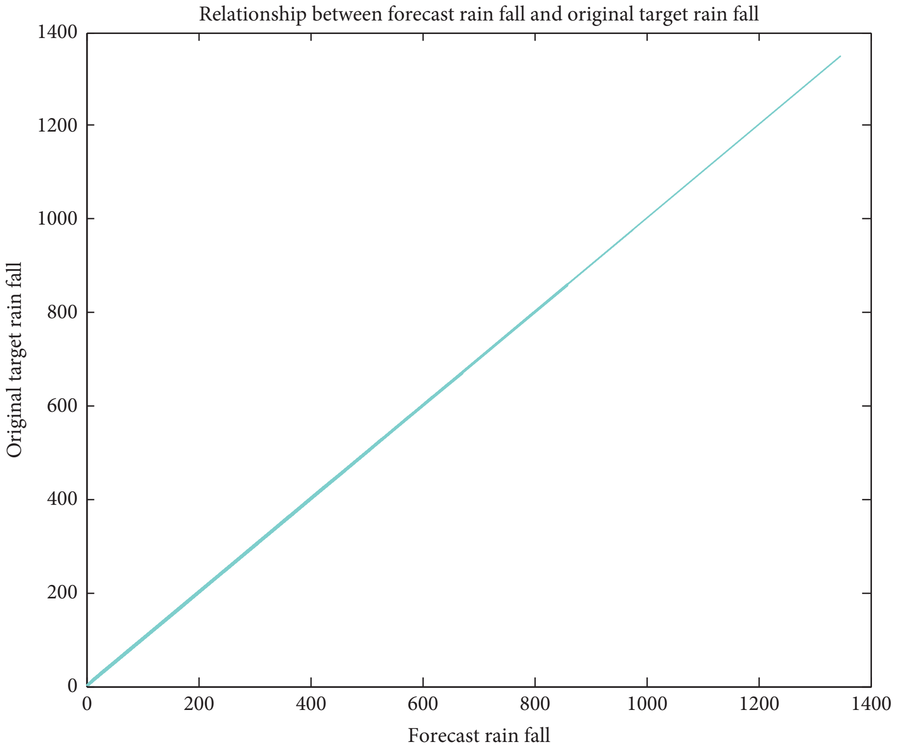


Figure 13: Relationship between forecast rainfall and original target rainfall.

The 11 hidden neurons associated with the single hidden layer Elman neural network forecast temperature match the target values. Hence, the error values are the least; it is clearly understood from Figures 15 and 16, respectively. The relationship between forecast temperature and the target temperature is linear; it is noticed from Figure 17.

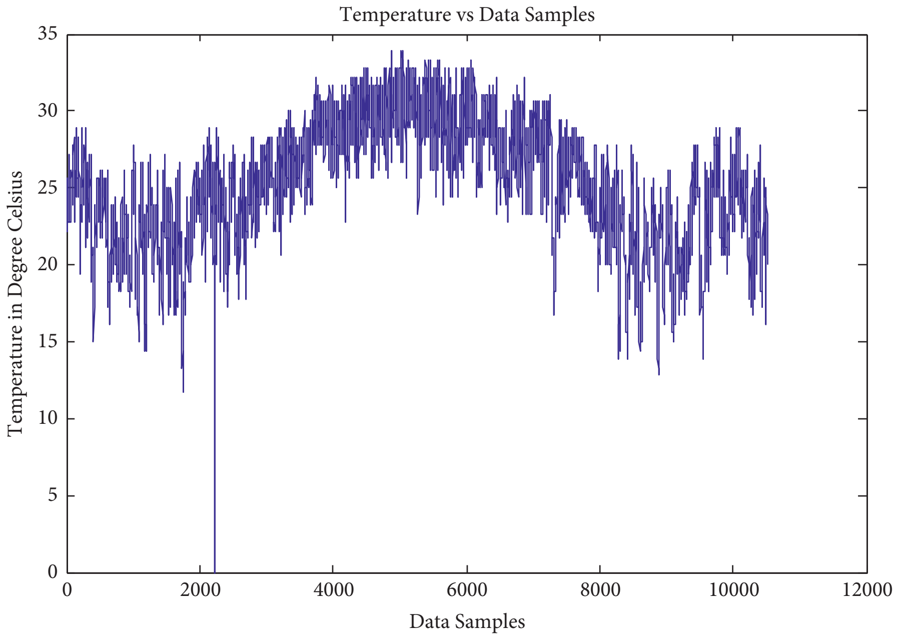


Figure 14: Temperature vs. data samples.

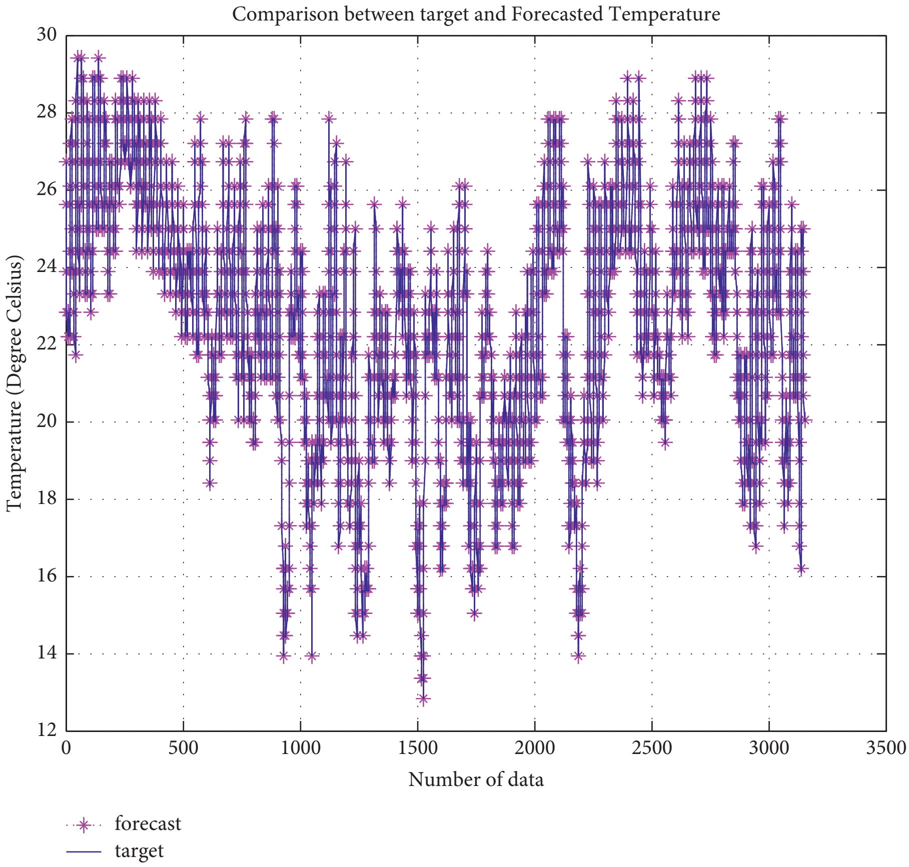


Figure 15: Comparison between target and forecasted temperature.

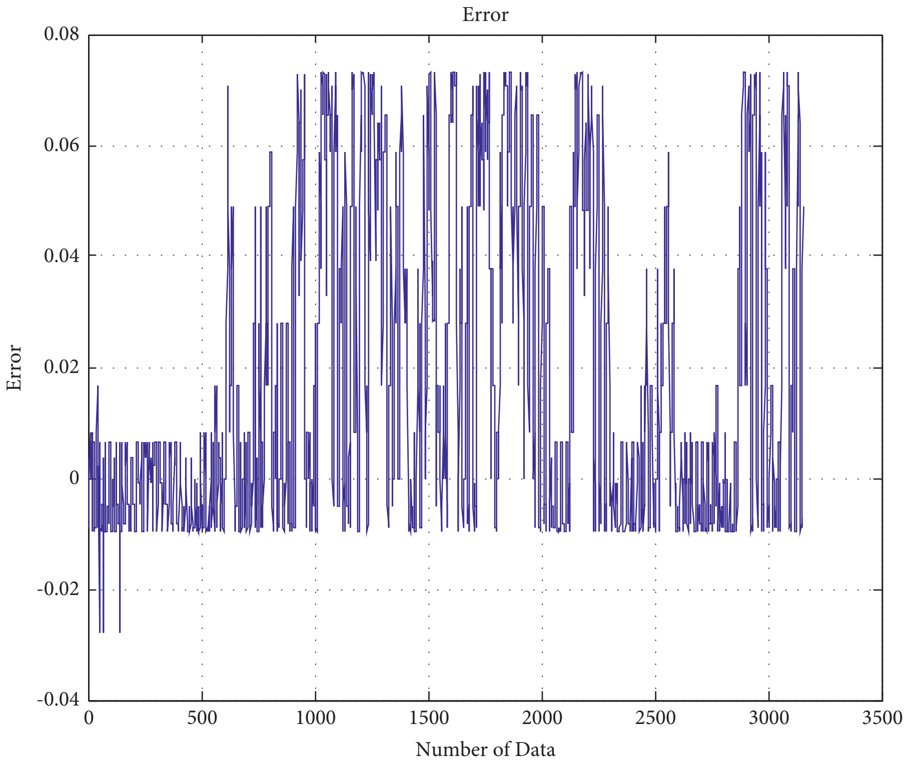


Figure 16: Error vs. number of data.

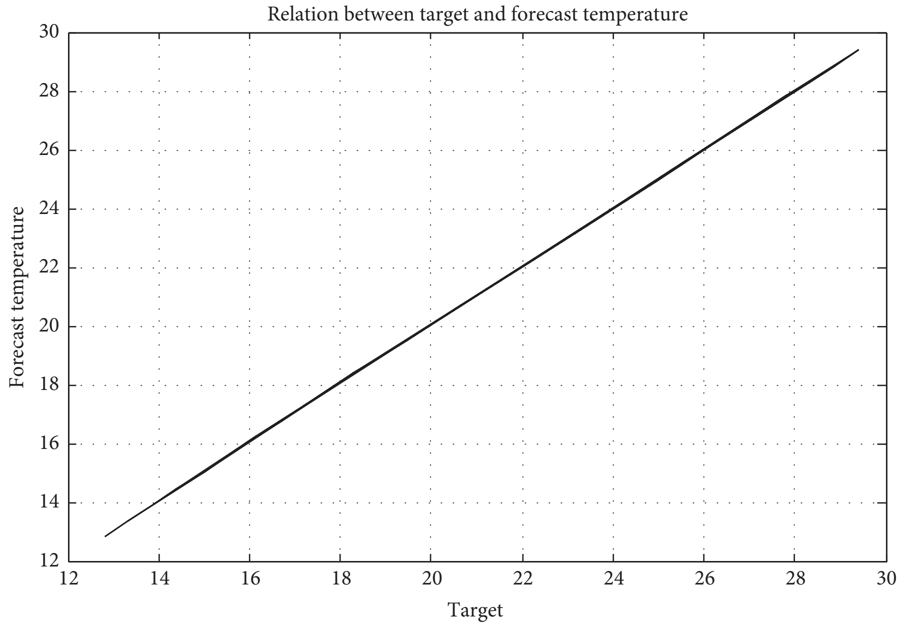


Figure 17: Relationship between target and forecast temperature.

5. Conclusion

Nowadays, human expectations and needs are increasing widely. All are interested in artificial intelligence to make their work easy and effective. This paper discusses the history of artificial neural networks, the generation of artificial neural networks, the generalized process involved in artificial neural networks, the various types, structural design, and artificial neural network applications that are elucidated in a detailed manner. The artificial neural network can address multiple applications, but this paper forecasting application is considered for performance analysis.

The highlights of the differences between the proposed models and the existing ones are as follows:

(i) The proposed models were developed with five years, five years, and seven years' data sets, respectively, for solar irradiance, rainfall, and temperature forecasting applications. Thus, we overcome the interannual variability-based uncertainty.

(ii) The proposed forecasting models possess minimal design complexity, are feasible to implement, and result in minor error qualifiers.

This paper proposed two feedforward neural networks, such as multilayer perceptron neural network (MLPNN) and improved backpropagation neural network (IBPNN), which could be used for forecasting applications like solar irradiance and rainfall forecasting, respectively. Moreover, the proposed one-feedback neural network, such as Elman neural network (ENN), can be used for the temperature forecasting application. The proposed artificial neural networks (MLPNN, IBPNN, and ENN) performances are statistically analyzed with various hidden neurons. The designed neural network–based forecasting model effectiveness is successfully validated using the acquired real-time training and test data set. Error qualifiers are used to analyze the performance of the proposed neural networks. According to the obtained results from the proposed artificial neural network–based forecasting model, it is observed that the proposed neural network–based forecasting model outperforms in all considered applications with much minimal error and reduced convergence time. The proposed multilayer perceptron neural network, which comprises 5 inputs, single hidden layer, and 14 hidden neurons achieves the minimal errors like $R = 1$, MAPE = 2.3034e-04, MSE = 5.9791e-07, MAE = 5.0093e-04, RMSE = 7.7324e-04, MRE = 2.3034e-06, and Time = 1.32 minutes for solar irradiance forecasting application. The suggested improved backpropagation neural network, which comprises five inputs, a single hidden layer, and seven hidden neurons, achieves minimal errors like $R = 1$, MAPE = 0.0158, MSE = 0.0042, MAE = 0.0320, RMSE = 0.0645, MRE = 1.5841e-04, and Time = 1.24 minutes for rainfall forecasting application. Similarly, the proposed Elman neural network, which comprises five inputs, a single hidden layer, and 11 hidden neurons, achieves minimal errors like MAPE = 0.1023, MSE = 0.0011, MAE = 0.0230, RMSE = 0.0332, MRE = 0.0010, and Time = 22 sec for temperature forecasting application. Hence, the proposed neural network–based forecasting models proved their validity, and they assist sustainability.

5.1. Proposed Forecasting Model Limitation and Future Work. To efficiently handle big data is one limitation of the proposed forecasting model. Thus, the authors can implement improved intelligent model-based forecasting in future work.

5.2. Recommendation of Future Direction and Research Scope. The building blocks of ANNs are neurons, linkages with weighted connection, activation function, and learning algorithms. Still, many research works focused on neural network performance improvement. The appropriated architecture selection was lacking in the field of artificial neural networks. There are no general guidelines available for the architecture framework. Artificial neural networks can effectively handle nonlinearity but obtain a feasible solution that is not generic. It can be overcome by the optimization algorithm associated with an artificial neural network (hybrid model). The readers can focus their research attention on the deep learning artificial neural network. However, ANN provided a promising result. Still, it has inefficiencies in some other applications like smart grid, natural language processing, speech recognition, computer vision, and so on, which lead to the quest to identify the optimal modeling of ANN.

The barrier to Growth in Artificial Neural Networks:

(i) Unique and privacy rights of the human being lost by an artificial neural network

(ii) Scarcity of job opportunities

(iii) Possibility to endanger humans and the environment

Artificial neural networks have many unique features and advantages; meanwhile, it has some barriers as well. Therefore, the advent of science and advancement should be healthy to benefit society, improve the economy and sustainability, and safeguard the environment and other living things.

References

[1] F. Rosenblatt, "The Perceptron — a perceiving and recognizing automaton," Report 85–460–1, Cornell Aeronautical Laboratory, New York, NY, USA, 1957.

[2] K. Liu, X. Hu, J. Meng, J. M. Guerrero, and R. Teodorescu, "RUBoost-based ensemble machine learning for electrode quality classification in Li-ion battery manufacturing," *IEEE*, 2021.

[3] K. Liu, Z. Wei, Z. Yang, and L. Kang, "Mass load prediction for lithium-ion battery electrode clean production: a machine learning approach," *Journal of Cleaner Production*, vol. 289, Article ID 125159, 2021.

[4] K. Liu, X. Hu, H. Zhou, L. Tong, D. Widanalage, and J. Marco, "Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification," *IEEE*, vol. 26, 2021.

[5] T. Hu, K. Li, H. Ma, H. Sun, and K. Liu, "Quantile forecast of renewable energy generation based on Indicator Gradient Descent and deep residual BiLSTM," *Control Engineering Practice*, vol. 114, Article ID 104863, 2021.

[6] X. Tang, K. Liu, K. Li, W. D. Widanage, and E. Kendrick, "Recovering large-scale battery aging data set with machine learning," *Patterns*, vol. 2, no. 8, Article ID 100302, 2021.

[7] K. Liu, Y. Shang, Q. Ouyang, and W. D. Widanage, "A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, pp. 3170–3180, 2020.

[8] M. Madhiarasan and S. N. Deepa, "Determination of adequate hidden neurons in combo neural network using new formulation and fine tuning with IMGWOA for enrich wind-speed forecasting," *International Journal of Applied Research on Information Technology and Computing*, vol. 9, no. 1, pp. 89–101, 2018.

[9] M. Madhiarasan, L. Mohamed, and P. P. Roy, "Novel co-operative multi-input multilayer perceptron neural network performance analysis with application of solar irradiance forecasting," *International Journal of Photoenergy*, vol. 2021, pp. 1–24, Article ID 7238293, 2021.

[10] M. Madhiarasan and S. N. Deepa, "Long-Term wind speed forecasting using spiking neural network optimized by improved modified grey wolf optimization algorithm," *International Journal of Advanced Research*, vol. 4, no. 7, pp. 356–368, 2016.

[11] C. Lyu, S. Basumallik, S. Eftekharnejad, and C. Xu, "A data-driven solar irradiance forecasting model with minimum data," in *Proceedings of the IEEE Texas Power and Energy Conference (TPEC)*, pp. 1–6, IEEE, College Station, TX, USA, February 2021.

[12] J. I. Athavale, M. Yoda, and Y. Joshi, "Comparison of data driven modeling approaches for temperature prediction in data centers," *International Journal of Heat and Mass Transfer*, vol. 135, pp. 1039–1052, 2019.

[13] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng, and S. Winkler, "A data-driven approach for accurate rainfall prediction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9323–9331, 2019.

[14] D. E. Rumelhart, G. E. Hinton, and J. L. McClelland, "A general framework for parallel distributed processing," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 45–76, MIT Press, Cambridge, MA, USA, 1986.

[15] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulltin of Mathematical Biophysics*, vol. 7, pp. 115–133, 1943.

[16] D. O. Hebb, *The Organization of Behavior*, Wiley & Sons, New York, NY, USA, 1949.

[17] F. Rosenblatt, *Principles of Neurodynamics*, Spartan Books, Washington, DC, USA, 1962.

[18] B. Widrow and M. E. Hoff, "Adaptive switching circuits," *1960 IRE WESCON Convention Record*, pp. 96–104, 1960.

[19] B. Widrow, "Generalization and information storage in networks of Adaline' neurons," in *Self-Organizing Systems 1962*, M. C. Yovitz, G. T. Jacobi, and G. Goldstein, Eds., Article ID 435461, Spartan Books, Washington, DC, USA, 1962.

[20] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[21] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the USA*, vol. 79, no. 8, pp. 2554–2558, 1982.

[22] D. E. Rumelhart, G. R. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition (V1 and V2)*, D. E. Rumulhart and J. L. McClelland, Eds., vol. 1, MIT Press, Cambridge, MA, USA.

[23] L. O. Chua and L. Yang, "Cellular neural networks: theory," *IEEE Transactions on Circuits and Systems*, vol. 35, pp. 1257–1272, 1988.

[24] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[25] W. Gerstner and W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity"*, Cambridge University Press, Cambridge, MA, 2002.

[26] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*, G. Montavon, G. B. Orr, and K. R. Müller, Eds., vol. 7700, Springer, Berlin, Germany, 2012.

[27] M. Madhiarasan and S. N. Deepa, "Comparative analysis on hidden neurons estimation in multi layer perceptron neural networks for wind speed forecasting," *Artificial Intelligence Review*, vol. 48, no. 4, pp. 449–471, 2017.

[28] M. Madhiarasan and S. N. Deepa, "A novel criterion to select hidden neuron numbers in improved backpropagation networks for wind speed forecasting," *Applied Intelligence*, vol. 44, no. 4, pp. 878–893, 2016.

[29] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[30] M. Madhiarasan and S. N. Deepa, "ELMAN neural network with modified grey wolf optimizer for enhanced wind speed forecasting," *Circuits and Systems*, vol. 7, no. 10, pp. 2975–2995, 2016.

[31] M. Madhiarasan and S. N. Deepa, "A novel method to select hidden neurons in ELMAN neural network for wind speed prediction application," *WSEAS Transactions on Power Systems*, vol. 13, pp. 13–30, 2018.

[32] M. Madhiarasan and S. N. Deepa, "Deep neural network using new training strategy based forecasting method for wind speed and solar irradiance forecast," *Middle-East Journal of Scientific Research*, vol. 24, no. 12, pp. 3730–3747, 2016.

[33] M. Madhiarasan and S. N. Deepa, "A new hybridized optimization algorithm to optimize echo state network for application in solar irradiance and wind speed forecasting," *World Applied Sciences Journal*, vol. 35, no. 4, pp. 596–614, 2017.

[34] M. Madhiarasan and S. N. Deepa, "Precisious estimation of solar irradiance by innovative neural network and identify exact hidden layer nodes through novel deciding standard," *Asian Journal of Research in Social Sciences and Humanities*, vol. 6, no. 12, pp. 951–974, 2016.

[35] J. Abbot and J. Marohasy, "Application of artificial neural networks to rainfall forecasting in queensland, Australia," *Advances in Atmospheric Sciences*, vol. 29, no. 4, pp. 717–730, 2012.

[36] G. Geetha and R. S. Selvaraj, "Prediction of monthly rainfall in Chennai using backpropagation neural network model," *International Journal of Engineering, Science and Technology*, vol. 3, no. 1, pp. 211–213, 2011.

[37] P. Zhang, Y. Jia, J. Gao, W. Song, and H. K. Leung, "Short-term rainfall forecasting using multilayer perceptron," *IEEE Transactions on Big Data*, vol. 6, no. 1, pp. 93–106, 2018.

[38] T. Cowan, M. C. Wheeler, O. Alves et al., "Forecasting the extreme rainfall, low temperatures, and strong winds associated with the northern Queensland floods of February 2019," *Weather and Climate*, vol. 26, Article ID 100232, 2019.

[39] A. Krzemień, "Fire risk prevention in underground coal gasification (UCG) within active mines: temperature forecast by means of mars models," *Energy*, vol. 170, pp. 777–790, 2019.

[40] B. Spencer, O. Alfandi, and F. Al-Obeidat, "Forecasting temperature in a smart home with segmented linear regression," *Procedia Computer Science*, vol. 155, pp. 511–518, 2019.

[41] M. Madhiarasan and S. N. Deepa, "Performance investigation of six artificial neural networks for different time scale wind speed forecasting in three wind farms of coimbatore region," *International Journal of Innovation and Scientific Research*, vol. 23, no. 2, pp. 380–411, 2016.

[42] M. Madhiarasan, "Accurate prediction of different forecast horizons wind speed using a recursive radial basis function neural network," *Protection and Control of Modern Power Systems*, vol. 5, no. 22, pp. 1–9, 2020.

[43] M. Madhiarasan, "Long-Term wind speed prediction using artificial neural network-based approaches," *AIMS Geosciences*, vol. 7, no. 4, pp. 542–552, 2021.

[44] M. Madhiarasan, M. Tipaldi, and P. Siano, "Analysis of artificial neural network performance based on influencing factors for temperature forecasting applications," *Journal of High Speed Networks*, vol. 26, no. 3, pp. 209–223, 2020.

[45] M. Madhiarasan, "Certain algebraic criteria for design of hybrid neural network models with applications in renewable energy forecasting," Ph. D. Thesis, Anna University, Chennai, India, 2018.

Intelligent Power Grid Video Surveillance Technology Based on Efficient Compression Algorithm Using Robust Particle Swarm Optimization

Sunil Kumar Tripathy, *Department of Mechanical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sktripathy1@gmail.com*

Soumya Datta Mohanty, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, soumya_datta.mohanty@gmail.com*

Chinmaya Ranjan Pradhan, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, cr_pradhan@outlook.com*

Satyajit Nayak, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, satyajit_nayak@gmail.com*

Abstract

Companies that produce energy transmit it to any or all households via a power grid, which is a regulated power transmission hub that acts as a middleman. When a power grid fails, the whole area it serves is blacked out. To ensure smooth and effective functioning, a power grid monitoring system is required. Computer vision is among the most commonly utilized and active research applications in the world of video surveillance. Though a lot has been accomplished in the field of power grid surveillance, a more effective compression method is still required for large quantities of grid surveillance video data to be archived compactly and sent efficiently. Video compression has become increasingly essential with the advent of contemporary video processing algorithms. An algorithm's efficacy in a power grid monitoring system depends on the rate at which video data is sent. A novel compression technique for video inputs from power grid monitoring equipment is described in this study. Due to a lack of redundancy in visual input, traditional techniques are unable to fulfill the current demand standards for modern technology. As a result, the volume of data that needs to be saved and handled in live time grows. Encoding frames and decreasing duplication in surveillance video using texture information similarity, the proposed technique overcomes the aforementioned problems by Robust Particle Swarm Optimization (RPSO) based run-length coding approach. Our solution surpasses other current and relevant existing algorithms based on experimental findings and assessments of different surveillance video sequences utilizing varied parameters. A massive collection of surveillance films was compressed at a 50% higher rate using the suggested approach than with existing methods.

1. Introduction

As discussed by Memos et al. [1], the number of Switch-Mode Power Supply is increasing, as are incentive-based switching activities at the end-user level. A high-resolution time-resolution monitoring system will be required for future smart grids' operational stability to properly examine the state of the electricity grid. When it comes to power grid measurement applications, kilohertz frequencies are used, but the degree of aggregation and the reporting rate is not the same. Instead of using a second rate for instantaneous data like the smart meters, they utilize a day or more rate for the cumulative consumption data. As a result of the consolidation, communication lines and storage space needs have been significantly reduced. Assessing power quality (PQ) as well as disaggregating loads necessitates more data, as Gao et al. have shown [2]. Several features can be added on top of Harmonics; however, they can only provide partial information. Changes in grid operating approaches, demand-side control, and the rise of decentralized generation have led to an unknown number of combinations of interruptions. Features-based approaches may be unreliable due to the fact that data gets destroyed, particularly when exciting short-lived occurrences. Some thresholds may be adjusted by the user in commercially available equipment for PQ measurement at sample rates ranging from 10 kHz to 100 megahertz; raw data is captured when an event happens. A future smart grid, on the other hand, will have hard-to-predict threshold values. There may be further insights to be gained by examining raw data from synchronized measurements at different locations—even if not all scattered sensors were able to classify events simultaneously and hence did not capture them at a high resolution was depicted by Tsakanikas et al. [3]. Deploying a continuous storage system for raw data will assist data-driven research that attempts to improve event classification and smart grid analysis algorithms; for example, when using lossless data compression, compressing and transmitting large volumes of data is considerably easier. Uncommonly, the raw data stream of a recording device has three voltage readings (from the 3 phases) and 4 current measurements (3 for the phase currents and 1 for the neutral conductor). Nominally sinusoidal voltage curves that are 120° out of phase make up a three-phase power system. A strong relationship has been established between the various modes of communication. The same applies to current lines, and leveraging this interaction allows for a particular minimization in data volume. As a result of their unique distortion, the waveforms are less connected. Conceived as a way to decrease correlation in current channels owing to phase-load distortions. It is important to note that waveforms only change at the equipment's contact state or load alterations in terms of distortions. In general, these operations are slow compared to the length of time. With only one load connection, waveform changes are rare. Because load currents are increased whenever a massive number of demands get linked to the sensing field of the grid, variations exist rapidly. Waveform compression is therefore conceivable. It is known that lossless compression methods exist for certain applications, such as music and video. However, no method has been identified that is specifically designed to take advantage of the periodicity and multichannel nature of electrical signals encountered in a stream compression methodology. An overview of lossy and nonlossy techniques is included in the book, as are the CR values from trials. Applications that focus on PQ-event compression are listed; these implementations were developed by Shidik and his colleagues [4] among themselves. There is no statistical analysis of lengthy original data. These models focused on extremely precise incident data to validate the applicability of algorithmic changes in their own unique contexts. In the majority of cases, data sources are not referenced or provided at all. We find the researchers do not have a benchmark against which to assess the feasibility of compression algorithms for grid wave information, regardless of whether they are using known techniques or new ones that have yet to be found. We have chosen to focus on the development of compression with no degradation techniques and performances for grid data at a lot of sampling to address these issues in the present contribution. When utilizing input data with a variety of ideas, we are considering new growth ideas made of natural time-series analysis. New lossless compression algorithms could be developed by using testing data and comparison parameters for the first thorough accessible standard. They can be used as a decision assistance tool by researchers dealing with data-intensive smart grid measures. The preprocessing phase entails changing the color space, after which the features may be retrieved using pseudo-component analysis. Then, utilizing Robust Particle Swarm Optimization, the encoding and decoding process may be completed. The main contribution of the research work is as follows:

(i) To design and develop a compression-based video surveillance technology based on the optimization approach

(ii) For the purpose of the authentication, run-length encoding and decoding were performed

The following is how the rest of the article is organized. In Section 2, a literature survey is being reported on strategies to reduce loss during video compression. The issue of lossless video compression mechanisms was then addressed in Section 3. Section 4 then poses the proposed mechanism over lossless video compression. The results of the suggested method and the conclusions were examined in Sections 5 and Section 6.

2. Related Works

In [1], the article looks into wireless sensor networks (WSNs) alongside the most recent research on social confidentiality and protection in WSNs. While adopting High-Efficiency Video Coding (HEVC) as a new media compression standard, a novel EAMSuS in the IoT organization is presented (HEVC). In [5], complete situational awareness is provided via real-time video analysis and active cameras. In [6], a new section of MPEG standards called Video Compression Modulation (VCM) has been suggested by the author. Video Coding for Machine Vision seeks to bridge the gap between machine vision feature coding and human vision video coding to accomplish collaborative compression and intelligent analytics. VCM's definition, formulation, and paradigm are provided first, corresponding with Digital Retina's rising compress instance. This is why they analyze video compression and features from the unique perspective of MPEG standards, which offers both academics and industry proof to accomplish the collaborative compression of the video shortly. In [7], using MapReduce, the author has developed UTOPIA Smart Video Surveillance for smart cities. From their end, we were able to incorporate smart video surveillance into our middleware platform. With the help of this article, we show that the system is scalable, efficient, dependable, and flexible. In [8], here when it comes to edge computing capabilities, the cloud object tracking and behavior identification system (CORBIS) was demonstrated. To increase distributed video surveillance systems' resiliency and intelligence, network bandwidth and reaction time between wireless cameras and cloud servers are being reduced in the Internet-of-things (IoT). In [9], an effective cryptosystem is used to create a safe IoT-based surveillance system. There are three parts to it. An automated summary technique based on histogram clustering is used to extract keyframes from the surveillance footage in the first stage. To compress the data, a discrete cosine transform is applied to it (DCT). Not to mention, a discrete fractional random transform is used to develop an efficient picture encryption approach in the suggested framework (DFRT). In [10], the author proposes a novel approach for compressing video inputs from surveillance systems. There is no way to reduce visual input redundancy using outdated methods that do not meet the demands of modern technologies. Video input storage needs to increase as a result, as does the time required to process the video input in real-time. To compress video inputs from surveillance systems, a unique technique is presented in this research paper. Visual input redundancy cannot be reduced using obsolete approaches that do not match the expectations of contemporary technology. This raises the storage requirements for video input and the processing time as a result. In [11], by using compressed sensing (CS), the author suggests creating security keys from the measurement matrix elements to secure your identity. Assailants cannot reconstruct the video using these. They are designed to prevent this. A WMSN testbed is used to analyze the effectiveness of the proposed security architecture in terms of memory footprint, security processing overhead, communication overhead, energy consumption, and packet loss, for example. In [12], a new binary exponential backoff (NBEB) technique was suggested by the author to "compress" unsent data that can preserve important information but recover the electronic trend as much as feasible. Data coming in may be temporally chosen and dumped into a buffer, while fresh data can be added to the buffer as it is received. As a result of the algorithm, the incoming traffic rate can be reduced in an exponential relationship with the transmitting failure times. In [13], the author suggested the lossless compression technique to handle the problem of managing huge raw data amounts with their quasiperiodic nature. The best compression method for this sort of data is determined by comparing the many freely accessible algorithms and implementations in terms of compression ratio, calculation time, and operating principles as well as algorithms for audio archiving; there are other algorithms for general data compression (Lempel–Ziv–Markov chain algorithm (LZMA), Deflate, Prediction by partial matching (PPMd), Burrows–Wheeler algorithm (Bzip2), and GNU zip (Gzip)) that are put to the test against one other. Deal with the challenge of managing enormous raw data quantities with their quasiperiodic nature by using lossless compression. Compression ratio, computation time, and operating principles are all taken into account when comparing publicly available algorithms and implementations to decide which is the most efficient. Additionally, generic data compression techniques such as LZMA, Deflate, PPMd, Bzip2, and Gzip are also put to the test. In [14], an efficient embedded image coder based on a reversibly discrete cosine transform is proposed for lossless. ROI coding with a high compression ratio (RDCT) was suggested. To further compress the background, a hierarchical (SPIHT) partitioning technique is used to combine the proposed rearranged structure with a lost zero tree wavelet coding. Results of the coding process indicate that the new encoder outperforms many state-of-the-art methods for still photo compression. In [15–17], the focus was based on the loss of video compression. Even at lower bit rates, the novel loss-compression method improves contourlet compression performance. Along with SVD, compression efficiency is improved by standardization and prediction of broken subband coefficients (BSCs) [18]. We measure the computational complexity of our solution with a better video quality. HCD uses DWT, DCT, and genetic optimization to improve the performance of transformed coefficients, among other techniques. This method works well with MVC to get the best possible rate distortion. The simulation results are produced using MATLAB Simulink R2015 to examine PSNR, bit rate, and calculation time for various video sequences using various wavelet functions, and the performance results are evaluated [19]. To solve the optimization issue of trajectory combination while producing video synopses, a new approach has been devised. When dealing with the optimization issue of motion trajectory combination, the technique makes use of the genetic algorithm's temporal combination methods (GA) [20]. The evolutionary algorithm is utilized as an activation function within the hidden layer of the neural network to construct an optimum codebook for adaptive vector quantization, which is proposed as a modified video compression model. The context-based initial codebook is generated using a background removal technique that extracts motion items from frames. Furthermore, lossless compression of important wavelet coefficients is achieved using Differential Pulse Code Modulation (DPCM), whereas lossy compression of low energy coefficients is achieved using Learning Vector Quantization (LVQ) neural networks [21]. This paper presents a rapid text encryption method based on a genetic algorithm. It is possible to use genetic operators Crossover and Mutation to encrypt data. By splitting up the plain text characters into pairs and using a crossover operation to obtain the encrypted text from the plain text, this encryption approach uses mutations to get its encrypted message.

From the literature survey, reviewed images and videos are compressed using transform-based and fractal approaches, along with other lossless encoding algorithms, which are now the most frequently used methods for still and video compression. Each technique has its own set of pros and downsides like breaking of the wavelet signal and low compression ratio; hence, it is important to choose the right one. It is most common for video-based images to be compressed using transform-based compression (TBC). In order to achieve compression, the signal or values are altered. Using various transformations, they convert a spatial domain representation into a picture. Brushlet is an example of an adaptive transformation (Verdoja and Grangetto 2017); bandelet (Raja 2018) (Erwan et al. 2005) and directionlet (Jing, et al. 2021) give information about the picture in advance. After applying these modifications to a picture, its essential function is altered. Hence, we are motivated to develop a methodology that overcomes all the existing video compression issues.

3. Problem Statement

Rapid advances are being made in compressing technology. As a challenging and essential topic, real-time video compression has sparked a lot of studies. This corpus of information has been included in the motion video standards to a large extent. Unanswered are several significant questions. According to the point of view of a compression algorithm, eliminating various redundancies from certain types of video data is a compression challenge. Thorough knowledge of the problem is needed, as well as a novel approach to solve all of the existing research gaps with irreversible video compression. Progress in other fields, such as artificial intelligence, has contributed to the breakthroughs in compression. A compression algorithm's success depends on the acceptance of a new generation of algorithms in addition to its technological excellence.

4. Proposed Work

As a result of the smart grid's usage of ICTs, the generation, distribution, and consumption of electricity are all more efficient (Information and Communication Technologies). For example, the transmission system and the medium-voltage level distribution system are monitored by Supervisory Control and Data Acquisition (SCADA) and wide-area monitoring systems (WAMS). It is important to remember that the primary objective of compression is to minimize the amount of data. That is if the compressed data retains most of its original content. Various scholars are currently involved in proposing effective techniques of data compression. Listed below are some of the most prevalent data compression techniques. With this analysis, we are focusing on compressing the PQ-event data in a video context in each successive frame to save space. To accomplish this, we must first identify the video frame object. Robust Particle Swarm Optimization is used to create a lossless video compression method. This is a diagram of the recommended technique shown in Figure 1.

4.1. Dataset. They used the UK Domestic Appliance-Level Electricity (UK-DALE) Dataset to conduct the experiments. A smart distribution system collects data on three-phase voltage, current, active and reactive power, and power factor from transformers at 54 substations as well as estimations of current and voltage at the inlets of three homes. The data is then analyzed and compared with the raw data from three homes. A 16 kHz sampling rate and a 24-bit vertical resolution were employed in the acquisition. There was a random selection of six FLAC-compressed recordings from 2014-8-08 to 2014-05-15, each having an hour of recordings. In a proprietary format, these data are recorded as four-byte floating-point numbers with timestamps at a sampling rate of 15 kHz. Voltage and current values are included in phase 2 of house 5. Every one of the four files contains 266 s. Large-scale databases hold all data transferred via a network. Raw data for three-phase voltages need 8.4 GB per day, whereas three-phase currents (including neutral) require 19.35 GB per day. To transmit the data, you need 0.8 Mbit/s and 1.8 Mbit/s in turn. This dataset was compiled in the following locations: as our institution's main power supply in Karlsruhe, Germany, we also have power outlets in our practical room and a substation transformer there. A total of seven channels consisting of four currents and three voltages are sampled at 12.8 and 25 kS/sec, respectively. There are seven channels, with four currents and three voltages sampled at 12.8 and 25 kS/sec, for a total of seven channels. Single-channel testing and dual-channel testing include measuring the current and/or voltage of a single phase in both situations, depending on which method is used. To save the data, raw 16-bit integers are stored in blocks of 60 s.

Electricity generation, transmission, and distribution in smart power systems are all affected by the analysis of the data. As a result, data exchange and memory requirements are expected to grow considerably, and data storage and bandwidth requirements for communication links in smart grids are also expected to increase. It is necessary to raise the sampling frequency to receive reliable and real-time information from the intelligent grid. There will be a greater emphasis on smart grid data compression in the future. Figure 1 illustrates the proposed compression approach. In areas of the grid with significant data volume, this approach can be used successfully.

4.1.1. Preprocessing. There are several steps to video compression, the first being preprocessing. Preprocessing is essential for a database's longevity and usefulness. For this reason, each stage in the video data processing workflow appears to be crucial. The procedure involves preprocessing, such as error detection or any other conversions that are not essential. Power grid video can cause picture frames to be split. The Bayesian motion subsampling approach may be used to create the video frame. This is a common method for removing frames from a movie. As the name implies, it is a computerized method used to enhance the frame creation process. For the most common sensitivities, the picture frame intensity range has been expanded, which results in a better image frame sensitivity value.

Let p denote the subsampled of each possible frame illustrated as

$$p^{y} = \frac{(\text{Number of the pixel frames with the } y \text{ intensity})}{\text{Total number of the pixel frames}}. \tag{1}$$

Here, $y = 0, 1, \ldots, y - 1$.

The separated pixel frames can be defined as depicted in [22]:

$$H_{i,j} = \text{base}\left((Y-1) \sum_{Y=0}^{b_{i,j}} p^{Y} \right), \tag{2}$$

where base represents the nearest integer. This is equivalent to transforming the pixel intensity [23]:

$$\frac{\partial N}{\partial x}\left(\int_{0}^{N} pN(x)\mathrm{d}z \right) = \left(\partial N(N)\left(x^{-1}\right)(N) \right) \frac{\mathrm{d}}{\mathrm{d}N}. \tag{3}$$

Here, finally, the probability distributed uniformity function can be represented as $\partial N / \partial x$.

When it comes to histograms, the equalization procedure can soften and enhance them. However, even though the histogram produced by the equalization is perfectly flat, it will be softened. After reducing the pictures' superfluous

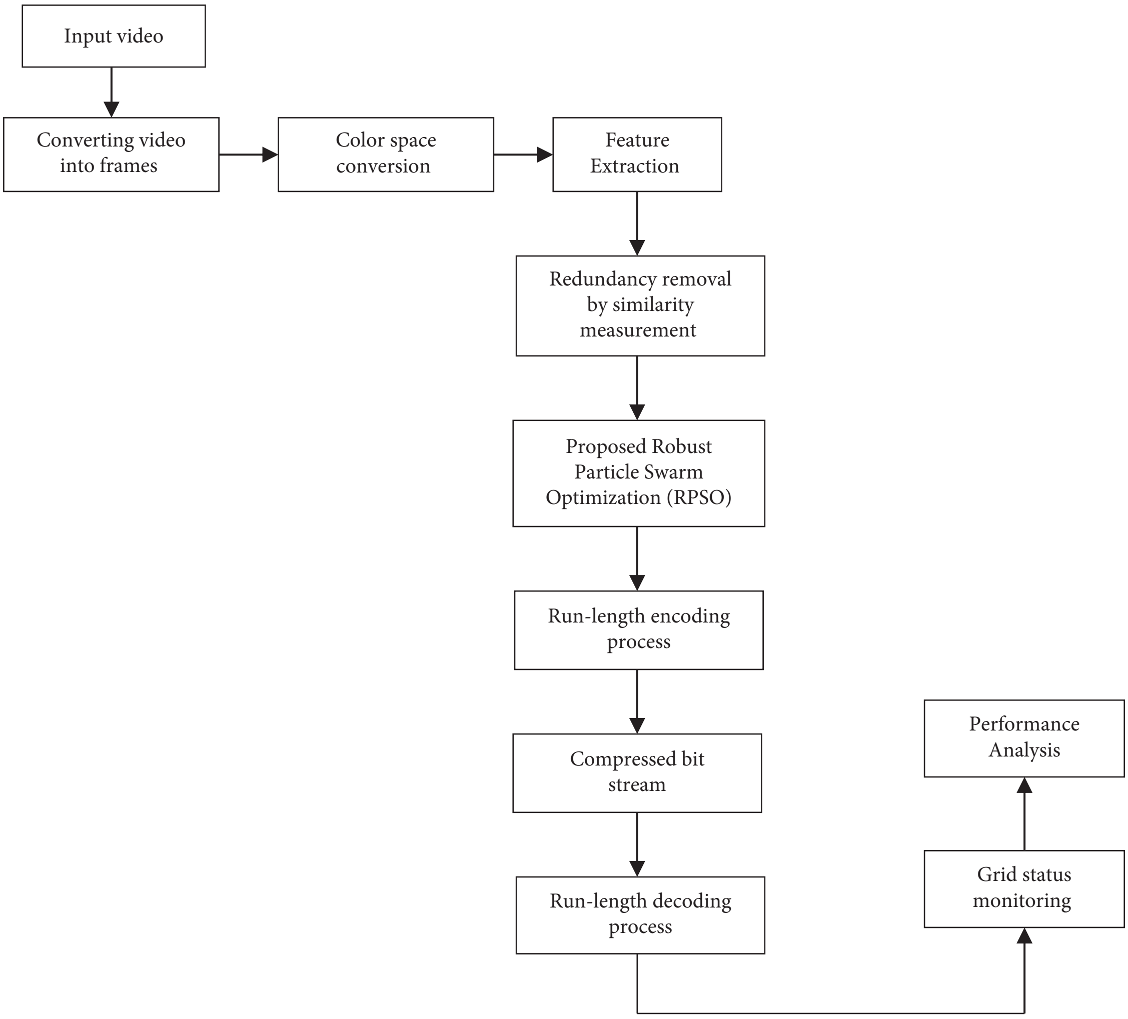

Input video
Converting video into frames
Color space conversion
Feature Extraction
Redundancy removal by similarity measurement
Proposed Robust Particle Swarm Optimization (RPSO)
Run-length encoding process
Compressed bit stream
Run-length decoding process
Grid status monitoring
Performance Analysis

Figure 1: Schematic representation of the suggested methodology.

noise, we apply a threshold technique to improve the refined frame acquired from the context. Thereafter, binary images are created, which streamlines the image processing process. As a result of the color space conversion, we see a shading effect in the majority of pictures. The picture contains three channels in most cases (red, green, blue). In the blue channel, there is no more information, but there is a great deal of contrast. Preprocessed green channel was deleted next. For example, here is how to extract the green channel [24]:

$$\begin{aligned} I_{\text{org}} &= f(\sigma, \mu, \beta), \\ I_{\text{red}} &= f(1, \mu, \beta), \\ I_{\text{Green}} &= f(\sigma, 2, \beta), \end{aligned} \tag{4}$$

where σ denotes the Red channel, μ denotes the Green channel, and β denotes the Blue channel.

Translation of color representation from one basis to another is called color space conversion (CSC). In most cases, this occurs while converting a picture from one color space to another. The use of a single threshold value for converting the color space is thus not recommended.

$$\theta \propto \text{Threshold}(E) \approx j * \left(\frac{u}{\left|v^{1}/3\right|}\right)\left(j_{\text{best}} - j_i\right), \tag{5}$$

where E represents converted the color space.

The color space is transformed to grayscale by keeping the brightness information. A grayscale picture frame can be represented as a collection of grayscale images by D_2.

$$ED_2 = \text{GS}\left(D_1\right) = \left\{d_2(1), d_2(2), \ldots, d_2(i), \ldots, d_2(|D|)\right\}. \tag{6}$$

After the frame gets preprocessed, the data can undergo the step of feature extraction.

4.1.2. Feature Extraction. We implemented the pseudo-component analysis in the feature extraction module to improve the compression performance and concentrate the image's information. The method for decreasing the size and complexity of data sets involves converting huge numbers of variables into smaller ones that retain the majority of the information contained in the large set. Naturally, limiting the number of parameters in sets of data lowers the information's accuracy, but the trick is to give up just a little precision for convenience. It is simpler to examine and interpret smaller data sets. Machine learning algorithms can also examine data more easily and quickly without dealing with extraneous issues. Each pseudo-redundancy component must be selected as a first stage in the process of feature extraction. In this module, the main goal is to extract the highlighted characteristics. Below are the configurations of this mechanism.

$$\begin{aligned} y_{\text{input}} &= \left[V_c^T y_c + B_c, V_s^T y_s + B_s\right], \\ \beta &= f_2\left(V_{\text{int}}^T f_1\left(y_{\text{input}}\right) + B_{\text{int}}\right). \end{aligned} \tag{7}$$

Here, $[V_c^T y_c + B_c, V_s^T y_s + B_s]$ Error! Bookmark not defined denotes the overall feature level; $V_c^T \in \mathbb{Z}^{C_c \times C_{\text{int}}}, V_s^T \in \mathbb{Z}^{C_s \times C_{\text{int}}}$, and $V_{\text{int}}^T \in \mathbb{Z}^{2C_{\text{int}} \times 1}$ represent feature weights; B_c, B_s, and B_{int} depict the associated features; C_c and C_s correspond to the sizes of the input medium of the categorization and feature sections, accordingly; and C_{int} denotes the internal input. Operations $f_1(y) = \max(y, 0)$ and $f_1(y) = 1/(1 + \exp(-y))$ associate to sigmoid activation operation, accordingly. The attention map is further standardized to [0, 1]. The outcome of the feature extraction is represented as depicted in [25]:

$$y_{\text{out}} = f_3([\beta \times y_c, y_s]). \tag{8}$$

Here, f_3 consists of a sequence of the feature components.

Pseudo- and nonpseudo-component characteristics can be selected using a property calculation technique. To determine pseudo-component characteristics, Hong correlations approaches, which employ averaging techniques, and Leibovici correlations, which use mixing principles, are used. In this approach, the phase fraction values are collected from a compositional system to minimize the difference between them. Pseudo- and nonpseudo-redundancy characteristics can be retrieved, as shown as follows [26]:

$$Lo_{\text{Pseudo,nonpseudo}} = 1 - \frac{2|A \cap B|}{|A| + |B|} = 1 - \frac{\sum_j^N p_j g_j + sm}{\sum_j^N p_j + \sum_j^N g_j + sm}, \tag{9}$$

where p_j represents the pseudo features, g_j represents the nonpseudo features, and sm represents the empirical constant.

4.1.3. Optimized Compression Process. Video data may be compressed without losing any information using this method. Concerning Robust Particle Swarm Optimization (RPSO) and run-length coding (RLC), the common RLC can be optimized by using the optimization algorithm and has been employed for the compression stage. We analyze the properties of compressed data using this technique. To maximize compression-related parameters, it is advised to use this method as a population-based approach. RPSO is initialized with the sample particles and modified with the optimal answer in each cycle. The resulting answer is called fitness and is referred to as the $_{\text{best}}$. The best solution obtained by a particle in a population is considered the world's top value monitored by the particle swarm optimizer (g_{best}). By the two p_{best} and g_{best} solutions, the positions of each particle change to global optima. The individual speeds and location functions of each particle are as follows. In a dimension search space D, there is a swarm composed of particles where each particle is represented by 'i' in a vector of $X_i = x_{i1}, x_{i2}, \ldots\ldots, x_{\text{id}}$ and the particle bet solution p_{best} is denoted as $p_i = p_{i1}, p_{i2}, \ldots\ldots, p_{\text{id}}$. Then, the best solution of the subset swarm is given by $g_{\text{best}} p_g = \left\{p_{g1}, p_{g2}, \ldots\ldots, p_{g\,d}\right\}$. The i^{th} particle velocity is represented as $V_i = V_{i1}, V_{i2}, \ldots\ldots, V_{\text{id}}$. The particle velocity and location are updated based on equations (10) and (11).

The weights updates are given by [27]

$$\begin{aligned} v_{\text{id}}(n+1) = W(it) * v_{\text{id}}^n + C_1 * \text{rand} * \left(p_{\text{id}} - x_{\text{id}}^n\right) \\ + C_2 * \text{rand} * \left(p_{g\,d} - x_{\text{id}}^n\right), \end{aligned} \tag{10}$$

where W represents the weighed features, C represents the cross features, n represents the constant, rand represents the random number, x_{id} represents the particle velocity, and p_{id} represents the particle motion. Here, depending on the feature extracted, the details can be updated depending upon the weightage, where $V_{\text{id}} = V_{\max}$, and it reflects the number of iterations between 1 and 10, where 10 is the maximum number of iterations. The random value of 0 to 1 is represented by the rand. C_1 and C_2 normally signify a nonnegative amount of an acceleration constant; here, C_1 and $C_2 = 1.05$. The particle orientation is also modified with [28]

$$x_{\text{id}}^{n+1} = x_{\text{id}}^n + v_{\text{id}}^{n+1}. \tag{11}$$

Any swarm obtains the health or objective f and each iteration provides the best solution; then if $f(x_i) < f(p_{\text{best}})$ and $f(g_{\text{best}})$, then p_{best} and $g_{\text{best}} = x_i$. The optimal measurement is obtained to maximize the curve transformation coefficients.

Video reconstruction can be done via run-length encoding once the optimized values are acquired. Using sequential data, this is a fairly simple operation to perform on your computer. For redundant data, it is a great tool. Running symbols are replaced by shorter ones in this technique. There are two ways to express the run-length code in grayscale images: V and R, where V represents the character count and R represents the run length. For optimized run-length optimal run-length encoding (ORLE), the following steps are required:

Step 1: Coefficient optimization

Step 2: Enter the string

Step 3: Give a unique value from the very first symbol or letter

Step 4: Otherwise, leave if the character or symbol is the final one in the string

Step 4: Additional symbols can be read and counted

Step 5: Until the preceding symbol subband has a nonmatching value, move on to step 3

Step 6: This will give you a count of the number of times a certain symbol appears in a given sentence

The suggested methodology uses a vector that contains a variety of scales to convert subbands that are optimized minimum and maximum to achieve the best result.

$$\text{Compressed}_{\text{Fitness value}} = \frac{-40 * q\left(-3 * \sqrt{\sum S_v}\right)}{2 - \exp\left(\sum \cos\left(3\pi * S_v\right)/d_b\right)} + 10 \text{ exp } , \tag{12}$$

where q denotes the compressed reconstructed value and S_v is the compressed score value that is obtained. Finally, the best rate of compression can get obtained. The RPSO reconstructs the data by using run-length decoding after refining the transforming Algorithm 1 curvelet parameters are as follows.

Finally, after compression, the status of the grid can get sorted out and it can get monitored and the irregular grid distribution can get identified.

5. Performance Analysis

Increasingly, data is being exchanged across smart grid sectors. Many types of data are created every day. For example, meteorological data such as the amount of sun or wind, humidity, or temperature are essential for optimal performance in many industries. Two phases in the data interchange procedure are encoding and decoding (or decryption). Numerous operations take place during the encoding phase to prepare data for transmission; when the data is encoded and decoded, it will be returned to its original form. In this section, you will learn about the complete process of performing experiments for performance evaluation. It is written in MATLAB, which is a programming language. Measurement data was collected over 24 hours in 1-minute, 5-second, 10-second, and 20-second intervals to assess the proposed compression methods. Readings from multiple meters were collected for each period in a data matrix.

Table 1 illustrates the effect of truncating small singular values on the compression ratio (CR) and percentage residual root difference. It can be seen that the minimum root mean square distance is obtained when eight singular values are considered. This leads to a reduction in the signal length. Compared to other sets of data, the calculated CR values for the 5-second time interval data are closer to the Total compression ratios (TCR) values in Table 1. Data obtained at 1-minute, 10-second, and 20-second intervals have generated CRs that deviate somewhat from the TCRs. Figure 2 illustrates the relationship between the number of significant singular values and TCR. According to the plotted data, the size of the data matrix has an impact on the ratio of compression (r), the number of singularly significant values. There are two different sizes of a data matrix: 5 seconds and 1 minute. A greater number of significant single values were required to match the TCR in 5-, 10-, and 20-second datasets than in the 1-minute dataset, as can be observed in Figure 2. As an alternative, selecting a shorter time interval, such as five seconds, will offer a better approximation on the number of significant singular values, resulting in the computed CRs being closer to the TCRs.

The mean error is a colloquial phrase that refers to the average of all mistakes in a collection. In this context, an "error" refers to a measurement uncertainty or the difference between the measured value and the correct/true value. Measuring error, often known as observational error, is the more formal word for error. According to Figure 3, there is a relationship between the related mean error for different time interval data and TCRs. As shown in Figure 3, the data consisting of measurements per 1-minute interval has the lowest mean error. The MAE found for greater matrix sizes is larger when the TCRs are higher.

According to Figure 4, there is a correlation between the number of significant singular values and the rate of mistake. For the first 100 single values, the 5-second dataset has the greatest MAE, followed by 10-seconds, 20-seconds, and 1-minute time interval dataset that has the lowest MAE. There is practically no inaccuracy in any dataset after the first 100 single values. A dataset's size has a substantial impact on singular values and the correctness of reconstructed data.

A smart distribution system's data is compressed in this part to see how well the approach works. To sum up, more singular values are required to fulfill TCR as a data set grows in size, as shown by the experimental findings. Nevertheless, increasing the number of singular values will reduce the amount of data that has to be compressed. As a result, there are fewer errors when the data is rebuilt after it has been compressed. As a result, a greater amount of data must be transferred through a wider range of communication channels. By compressing information with a high number of singular values to fulfill the TCR, you will have to send more data. The TCR must be matched to the quantity of data to be compressed to maximize the connection bandwidth when transferring the compressed data. The data reconstruction error can be calculated between the reconstructed data $g(i, j, s)$ and the original data $F(i, j, s)$ using

$$P(s) = \frac{1}{3MN} \sum_{i=0}^{a-1} \sum_{j=0}^{b-1} \sum_{s=0}^{2} \left\| g(i, j, s) - F(i, j, k) \right\|. \tag{13}$$

In addition, the Mean Average Error (MSE) (calculated by averaging squared error) is another way to assess reconstruction accuracy.

The MAE is defined as [29]

$$\text{MAE} = \sum_{i=0}^{a-1} \sum_{j=0}^{b-1} \sum_{s=0}^{2} \left\| g(i, j, s) - f(i, j, s) \right\|^2. \tag{14}$$

A measure of the quality of compression and reconstruction is the signal-to-noise ratio (SNR). There are two ways to define the peak SNR [30]:

$$\text{PSNR}(\text{dB}) = 10 \log_{10} \frac{(\text{Max}_i)}{\sqrt{\text{MAE}}}, \tag{15}$$

where Max_i is the maximum possible pixel value.

MD quantifies the greatest difference between original and reconstructed values. The average difference between original and reconstructed values is denoted as SSIM. For each of the formulas [31],

```
Input: Extracted features
Output: Compressed data $C_d$
To compute compressed value,
For $i = 1$: size$(D_{n\_\text{parameters}}, 1)$
   For $j = 1$: size$(D_{n\_\text{features}}, 1)$
   Weighed updates
$v_{\text{id}}(n+1) = W(it) * v_{\text{id}}^{n} + C_1 * \text{rand} * (p_{\text{id}} - x_{\text{id}}^{n})$
$+C_2 * \text{rand} * (p_{g\,d} - x_{\text{id}}^{n})$,
End
End
Data compressed features $d_{n\_\text{fea}} = [d_{n\_\text{fea}}\text{rand}]$
To compute, Run length encoding
Class label = unique(target)
$K$ = length(class label)
For $d = 1$: $k$
   Temp = total class mean($I$, :)
Run length decoding
Data grouping
Compressed$_{\text{Fitness value}}$ =
$(-40 * q(-3 * \sqrt{\sum S_v})/2 - \exp(\sum \cos(3\pi * S_v)/d_b))$
$+10$ exp,
End
```

Algorithm 1: (RPSO)

Table 1: Computed CR, and number of singular values, r of compression.

TCR	1 minute		20 seconds		10 seconds		5 seconds	
	r	CR	r	CR	r	CR	r	CR
100	6	96.68	8	103.62	9	101.89	10	96.84
80	7	82.87	10	82.90	11	83.36	12	80.70
60	10	58.01	14	59.21	15	61.13	16	60.53
50	12	48.34	17	48.76	18	50.94	19	50.97
30	19	30.53	28	29.60	31	29.58	32	30.26
10	58	10.00	83	9.99	92	9.97	97	9.98
5	116	5.00	166	4.99	183	5.01	194	4.99
4	145	4.00	207	4.00	229	4.00	242	4.00

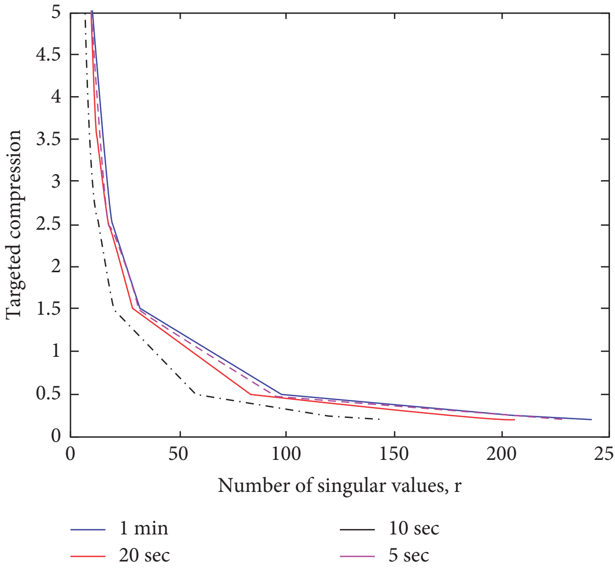


Figure 2: Plot of the number of singular values versus TCR for the different datasets.

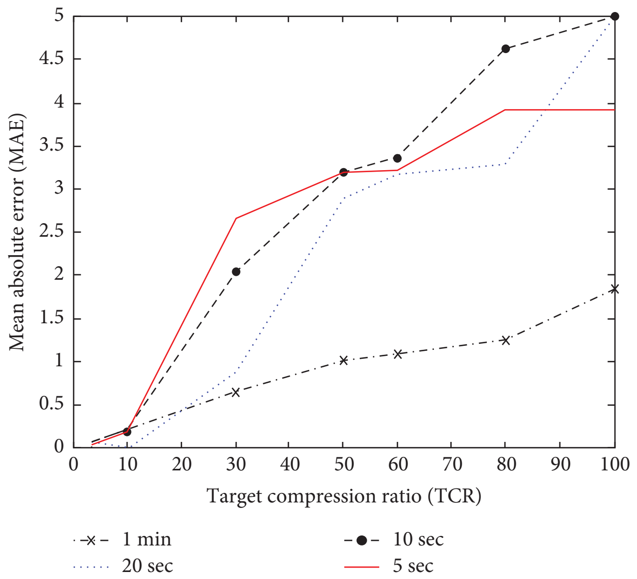


Figure 3: Plot of MAE versus TCR for different sampling rates.

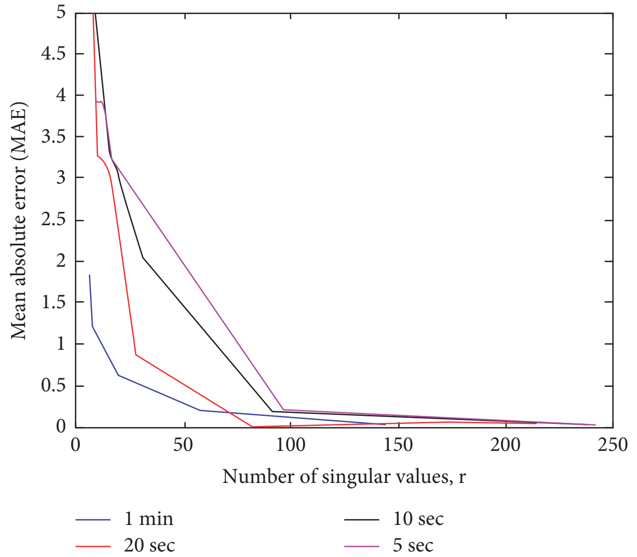


Figure 4: Plot of MAE versus the number of singular values, r for the dataset.

$$\mathrm{SSIM} = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \sum_{s=0}^{s-1} \| I(i, j, s) \|, \tag{16}$$
$$\mathrm{MD} = \max_{0 \le i \le M} \| I(i, j, s) \|.$$

The video reconstruction error (MSE), signal-to-noise ratio (PSNR), matching distance (MD), and percent compression ratio (PCR) values are obtained as depicted in Figure 5. The satisfying results are obtained over the compression as depicted in Table 2. From Table 2 and Figure 5, the suggested methodology shows the highest performance over PSNR, MSE, and MD. As illustrated in the PSNR contours for the testing set in Figure 5, the PSNR improves as the compressed image bit rate increases. The results demonstrate a rising pattern in PSNR values, whereas MSE drops progressively as the compressed image bit rate improves. As a result, a higher compressed image bit rate means higher resolution images and fewer mistakes.

The advantages of the existing mechanism in which the high compression ratio was obtained but it takes more time for compression. Hence, it can be overcome by the proposed mechanism.

5.1. Complexity Analysis. In general, the total number of states is approximately equal to 2^N for computing nth RLE number ($F(^N)$). Notice that each state denotes a function call to 'RPSO with RLE()$'$ which does nothing but makes another

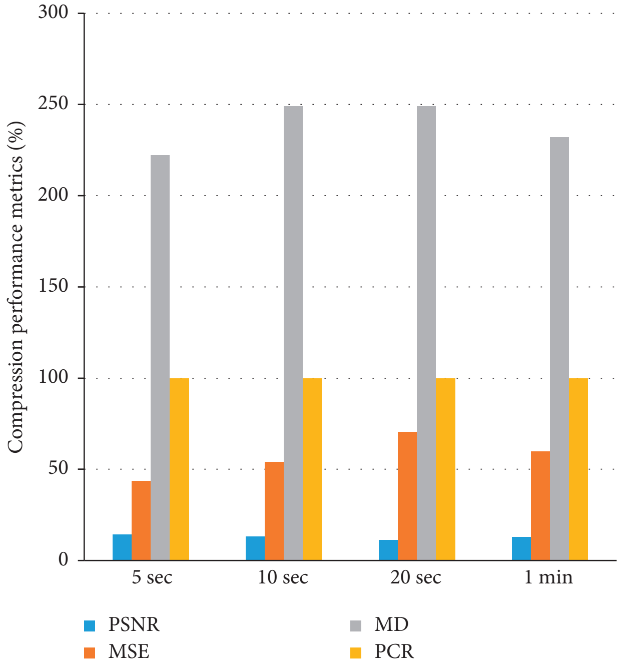


Figure 5: Image Quality metrics.

Table 2: Average Data Quality metrics.

Parameters	5 sec	10 sec	20 sec	1 min
PSNR	14.1913	13.2848	11.168	12.778
MAE	43.5359	54.1624	70.49	59.7
MD	222.1475	249.0821	249.05	232.1
PCR	100	100	100	100

Table 3: Average compression ratio.

Compression level	Number of images	Compression ratio
Discrete spatial multilayer perceptron (proposed)	100	10 : 1
Haar [17]	100	10 : 17
Cosine [17]	100	10 : 18

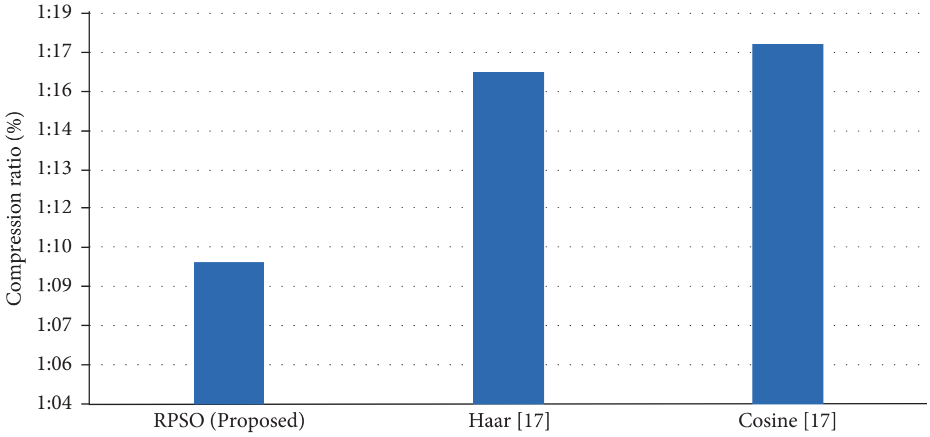


Figure 6: Compression ratio.

recursive call. Therefore, the total time taken to compute the nth number of the sequence is $O(2^N)$.

In digital file compression, duplication is the most important issue. If N_1 and N_2 signify the amounts of data holding units in the raw and encoded images, correspondingly, the compression ratio, CR, could be specified as CR $= N_1/\mathrm{N}_2$ as well as the data duplication of the original image as RD $= 1 - (1/\mathrm{CR})$. From Table 3 and Figure 6, the proposed methodology can acquire the exact ratio of compression (10 : 1) when compared to Haar [17] (10 : 16.5) and Cosine [17] (10 : 17.2) techniques.

6. Conclusion

Data compression techniques such as RPSO compression were examined and evaluated in this article. Data from a smart distribution system was used to evaluate the algorithm with 1-minute, 10-second, 20-second, and 5-second interval datasets. The results obtained demonstrate that the amount of the data has a considerable influence on the proposed approach. Larger datasets require more significant single values to achieve low error rates. When used to the smart grid, RPSO may be used as a simple and uncomplicated compression method. The significant singular values will provide a decent approximation when the compressed data has to be rebuilt using the recommended approach. Depending on the number of singular values used, RPSO compression can lower the volume of data. However, if you have a lot of data, you should consider using the proposed compression technique, which has a faster execution time and low error rates. Also, a lot of the pointed advantages exist. There will be some disadvantages also; in the proposed work, the order of bytes is independent. Compilation needs to be done again for compression. Errors may occur while transmitting data. We have to decompress the previous data. The disadvantages can be overcome in future work.

References

[1] V. A. Memos, K. E. Psannis, Y. Ishibashi, B.-G. Kim, and B. B. Gupta, "An efficient algorithm for media-based surveillance system (EAMSuS) in IoT smart city framework," *Future Generation Computer Systems*, vol. 83, pp. 619–628, 2018.

[2] Z. J. Gao and J. S. Wang, "Application of smart grid technology in the coalmine power system," *Applied Mechanics and Materials*, vol. 441, pp. 236–239, 2014.

[3] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736–753, 2018.

[4] G. F. Shidik, E. Noersasongko, A. Nugraha, P. N. Andono, J. Jumanto, and E. J. Kusuma, "A systematic review of intelligence video surveillance: trends, techniques, frameworks, and datasets," *IEEE Access*, vol. 7, pp. 170457–170473, 2019.

[5] A. Hampapur, L. Brown, J. Connell et al., "Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 38–51, 2005.

[6] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: a paradigm of collaborative compression and intelligent analytics," *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.

[7] C.-S. Yoon, H.-S. Jung, J.-W. Park, H.-G. Lee, C.-H. Yun, and Y. W. Lee, "A cloud-based UTOPIA smart video surveillance system for smart cities," *Applied Sciences*, vol. 10, no. 18, p. 6572, 2020.

[8] R. Rajavel, S. K. Ravichandran, K. Harimoorthy, P. Nagappan, and K. R. Gobichettipalayam, "IoT-based smart healthcare video surveillance system using edge computing," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2021.

[9] R. Hamza, A. Hassan, T. Huang, L. Ke, and H. Yan, "An efficient cryptosystem for video surveillance in the internet of things environment," *Complexity*, vol. 2019, Article ID 1625678, 11 pages, 2019.

[10] V. R. Prakash, "An enhanced coding algorithm for efficient video coding," *Journal of the Institute of Electronics and Computer*, vol. 1, pp. 28–38, 2019.

[11] S. A. Nandhini and S. Radha, "Efficient compressed sensing-based security approach for video surveillance application in wireless multimedia sensor networks," *Computers & Electrical Engineering*, vol. 60, pp. 175–192, 2017.

[12] T. Jiang, H. Wang, M. Daneshmand, and D. Wu, "Cognitive radio-based smart grid traffic scheduling with binary exponential backoff," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2038–2046, 2017.

[13] R. Jumar, H. Maaß, and V. Hagenmeyer, "Comparison of lossless compression schemes for high rate electrical grid time series for smart grid monitoring and analysis," *Computers & Electrical Engineering*, vol. 71, pp. 465–476, 2018.

[14] S. A. Elhannachi, N. Benamrane, and T.-A. Abdelmalik, "Adaptive medical image compression based on lossy and lossless embedded zero tree methods," *Journal of Information Processing Systems*, vol. 13, pp. 40–56, 2017.

[15] P. E. Sophia and J. Anitha, "Enhanced method of using contour let transform for medical image compression," *International Journal of Advanced Intelligence Paradigms*, vol. 14, no. 1/2, pp. 107–121, 2019.

[16] T. Kalidoss, L. Rajasekaran, K. Kanagasabai, G. Sannasi, and A. Kannan, "QoS aware trust based routing algorithm for wireless sensor networks," *Wireless Personal Communications*, vol. 110, no. 4, pp. 1637–1658, 2020.

[17] I. Yamnenko and V. Levchenko, "Video-data compression using wavelet analysis," in *Proceedings of the 2019 IEEE 20th International Conference on Computational Problems of Electrical Engineering (CPEE)*, pp. 1–4, Lviv-Slavske, Ukraine, September 2019.

[18] S. Rahimunnisha and G. Sudhavani, "Novel complexity reduction technique for multi-view video compression using HCD based genetic algorithm," *Design Engineering*, vol. 2021, no. 6, pp. 3219–3228, 2021.

[19] L. Xu, H. Liu, X. Yan, S. Liao, and X. Zhang, "Optimization method for trajectory combination in surveillance video synopsis based on genetic algorithm," *Journal of Ambient*

Intelligence and Humanized Computing, vol. 6, no. 5, pp. 623–633, 2015.

[20] S. M. Darwish and A. A. J. Almajtomi, "Metaheuristic-based vector quantization approach: a new paradigm for neural network-based video compression," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7367–7396, 2021.

[21] R. B. Abduljabbar, O. K. Hamid, and N. J. Alhyani, "Features of genetic algorithm for plain text encryption," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, p. 434, 2021.

[22] B. Azam, S. Ur Rahman, M. Irfan et al., "A reliable auto-robust analysis of blood smear images for classification of microcytic hypochromic anemia using gray level matrices and gabor feature bank," *Entropy*, vol. 22, no. 9, p. 1040, 2020.

[23] D. Xiang, T. Tang, L. Zhao, and Y. Su, "Superpixel generating algorithm based on pixel intensity and location similarity for SAR image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 6, pp. 1414–1418, 2013.

[24] S. Kwon, H. Kim, and K. S. Park, "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone," in *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2174–2177, San Diego, CA, USA, August 2012.

[25] T. Tuncer, S. Dogan, F. Ertam, and A. Subasi, "A novel ensemble local graph structure based feature extraction network for EEG signal analysis," *Biomedical Signal Processing and Control*, vol. 61, Article ID 102006, 2020.

[26] H. Xie, Y. Ren, W. Long, X. Yang, and X. Tang, "Principal component analysis in projection and image domains—another form of spectral imaging in photon-counting CT," *Institute of Electrical and Electronics Engineers Transactions on Biomedical Engineering*, vol. 68, pp. 1074–1083, 2020.

[27] W. Liu, Z. Wang, N. Zeng, Y. Yuan, F. E. Alsaadi, and X. Liu, "A novel randomised particle swarm optimizer," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 2, pp. 529–540, 2021.

[28] Z. Yong, Y. Li-Juan, Z. Qian, and S. Xiao-Yan, "Multi-objective optimization of building energy performance using a particle swarm optimizer with less control parameters," *Journal of Building Engineering*, vol. 32, Article ID 101505, 2020.

[29] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, pp. 79–82, 2005.

[30] Q. Huynh-Thu and M. Ghanbari, "The accuracy of PSNR in predicting video quality for different video scenes and frame rates," *Telecommunication Systems*, vol. 49, no. 1, pp. 35–48, 2012.

[31] A. K. Moorthy and A. C. Bovik, "Efficient motion weighted spatio-temporal video SSIM index," *Human Vision and Electronic Imaging*, vol. XV, Article ID 75271I, 2010.

Absorption Performance of Doped TiO_2-Based Perovskite Solar Cell using FDTD Simulation

Balagoni Sampath Kumar, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, bs.kumar456@gmail.com*

Sanjay Kumar Nayak, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, sk.nayak24@gmail.com*

Sunita Baral, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, sunita.baral95@gmail.com*

Aravinda Mahapatra, *Department of Mechanical Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, aravindamahapatra92@outlook.com*

Abstract

In the third generation of the solar cell era, significant trends in the development of perovskite solar cells (PSC) were observed. Exploring suitable materials for its wafer structure, such as perovskite and electron transport layers (ETL), were a major emphasis of high-performance PSC development. Because of its matching band structure to $MaPbI_3$, TiO_2 is the most often utilized material for ETL. However, in the application of TiO_2 to PSC, electron trapping and a wide energy gap become a drawback. The goal of this research is to improve the absorption performance of PSC employing ETL with Fe and Ta-doped TiO_2 as well as the thickness of the material. The interaction between the electromagnetic waves of light and the solar cell structure was calculated using Finite-Difference Time-Domain (FDTD) simulations, which resulted in the absorption spectra.

In comparison to pure TiO_2, which absorbs only 79.5% of the incident light, Fe-TiO_2 and Ta-TiO_2 as ETL in solar cells have increased absorption spectra to 81.7% and 81.2%, respectively. Finally, we may conclude that the optimum ETL layer parameters are 0.32% Fe doping and a thickness of 100 nm.

1. Introduction

Concerns about the environment, particularly the repercussions of the greenhouse effect for future generations, have prompted the development of alternative energy sources such as solar, wind, and biomass in recent decades [1]. Solar energy is regarded as the most abundant, sustainable, and ecologically beneficial renewable energy source among these [2, 3]. Because of their high efficiency of over 25%, perovskite solar cells (PSC) have attracted a lot of interest [4–8]. Perovskite is a hybrid organic-inorganic methylammonium lead halide $MaPbX_3$ ($MA = CH_3NH_3$; X = I, Br, Cr) material with excellent optoelectronic properties such as bandgap tunability, high light absorption coefficient, and long diffusion length [9, 10].

The development of PSC depends on several factors, including perovskite layer, the electron transport layer (ETL), and hole transport layer (HTL), as previously studied by Azri et al. and Hasanah et al. [11, 12]. The ETL was functioned to extract and transport the photogenerated electron and suppress the charge recombination to blocking hole as one of the significant activities in solar cell devices [13]. Therefore, the ETL is required to have high charge mobility, sufficient energy level alignment, and related morphology and interfacial properties [14]. TiO_2 is a well-known material that has widely used in PSC due to its good alignment of conduction band to the lowest unoccupied molecular orbitals (LUMO) of active layers of perovskite [11]. However, TiO_2havea low-coordinated Ti cation at the surface, which provided electron traps due to low-lying Ti 3d and this condition reduced the charge mobility [15]. The wide energy gap owned by TiO_2limits the absorption only at the UV-light region, which can affect the PSC absorption performance [16].

The enhancement of TiO_2 properties is commonly obtained by doping with noble metal ions, transition metal, and rare metal [17]. The dopants are function to reduce the number of electron trap sites because they can replace the low-coordinated Ti cations. The charge mobility can be improved by certain amounts of dopants that trigger exciton separation[17]. TiO_2 is an n-type semiconductor, and dopants for TiO_2 are classified into two types. To begin, pentavalent cations (Nb, Ta, and so forth) increase electron conductivity in TiO_2 and thus the energy gap [18]. Second, divalent/trivalent cations (Cr, Fe, Ni, Co, and so forth) can be used to convert TiO_2 to a p-type and decrease the energy gap [19].

The absorption performance of PSC with $MaPbI_3$ was investigated in this study using different types of TiO_2 doping as ETL. Tantalum (Ta) at concentrations of 1.8% and 3.8% and iron (Fe) at concentrations of 0.11% and 0.32% were employed as dopant materials. The thickness of TiO_2-based ETL was also simulated to investigate the effect of thin and thick ETL. The simulation was obtained using Finite-Difference Time-Domain (FDTD) methods, which use Maxwell's equations to calculate. The absorption spectra are characterized at wavelengths ranging from 300 to 1500 nm, and the results are presented as absorption percentages and indexes in arbitrary units. Finally, this research should provide the best parameters for doped TiO_2-based ETL in PSC.

2. Methods

The optical model was utilized to examine the absorption performance of doped TiO_2-based PSC in this study. Lumerical Ltd.'s Finite-Difference Time-Domain (FDTD) simulations use Maxwell's equations to represent the interaction between electromagnetic light waves and the solar cell structure. FDTD simulations have been widely employed in various applications, including photovoltaic (PV) research [20]. The incident light bandwidth employed in this work was 300–1500 nm, based on an AM1.5 spectrum with a resolution of 12 nm; photon flux was high in this bandwidth but low elsewhere [21]. Light absorption is expressed as $\mathrm{Abs}(\lambda)$ for each wavelength and is formulated using

$$\mathrm{Abs}(\lambda) = \int P_{\mathrm{Abs}}(\lambda) dV. \quad (1)$$

$P_{\mathrm{Abs}}(\lambda)$ is the power absorbed per unit volume for a given wavelength, and dV is the absorber volume, which may be calculated using

$$P_{\mathrm{Abs}} = \frac{1}{2}\omega\varepsilon'' \left|\vec{E}\right|^2, \quad (2)$$

where ω is the angular frequency of the light which corresponds to the wavelength and ε'' is the imaginary part of the dielectric permittivity and $|\vec{E}|^2$ is the electric field strength.

The performance of solar cells was determined using the FDTD simulation with the refractive index as the fundamental material parameter. ITO, TiO_2, $MaPbI_3$, CuSCN, and Au were used as the front contact, electron transport layer (ETL), absorption/perovskite layer, hole transport layer (HTL), and back reflector, respectively, in the PSC structure shown in Figure 1. The earlier work [22–24] refers to these materials complex refractive index and thickness. Meanwhile, the ETL is optimized in this study by varying the doping materials and thickness. The dopants used are 1.8% and 3.8% tantalum (Ta) to produce Ta-TiO_2 and 0.11 and 0.32% iron (Fe) to produce Fe-TiO_2. The thickness was also varied to 20 nm, 50 nm, 100 nm, 200 nm, and 500 nm. The refractive index of doped TiO_2 was calculated using the refractive index-bandgap relation obtained from Reddy and Ahammed [25, 26].

The FDTD simulations are carried out in the PSC structure 2-dimensional mesh cells, with a mesh accuracy rate of 3. The light source was used in the y-axis plane-wave mode with amplitude and wavelength of 1 and 300–1500 nm, respectively. The y-axis was used with perfect matching layers (PML) and boundary conditions (BC) to maximize incident light trapping. Meanwhile, the periodic BC was applied in the x-axis, assuming the structure's infinite periodicity [27, 28]. A 2D y-normal frequency-domain field and power monitor were clamped to the layer being studied to record the absorption. The general investigation of the structure was obtained by placed the monitor between ITO-TiO_2 and CuSCN-Au. The 3D frequency-domain field and power monitor were also used to record the electric field $|\vec{E}|^2$ profile to analyze the field distribution in the PSC structure. Finally, the absorption curve and E-field profile for the desired structure and layers can be generated by this simulation.

3. Results and Discussion

3.1. Contribution of PSC Layers to Absorption Performance. As descibed in the methods,the PSC structure in this study, consists of five layers with varying materials and functions. According to Shockley-Quisser theory, the direct bandgap of $MaPbI_3$ is 1.5–1.6 eV [10]. The absorption coefficient of $MaPbI_3$ with a thickness of 280 nm was 80% of incident light in the sun-ray range [29]. As a result, these perovskite materials are an excellent candidate for absorption layer to produce high PSC performance. CuSCN with a bandgap of 3.4 eV was used in the HTL because it had the highest efficiency compared to other HTL materials such as P3HT, Spiro-OMeTAD, CuI, and NiO. The efficiency of CuSCN in HTL is supported by the excellentalignment of its highest occupied molecular orbital (HOMO) level of the CuSCN with the valence band of perovskite layer [11]. The TiO_2 with bandgap 3.2 eV used in the ETL is also used to enhance the PSC performance. Azri et al. reported that TiO_2had

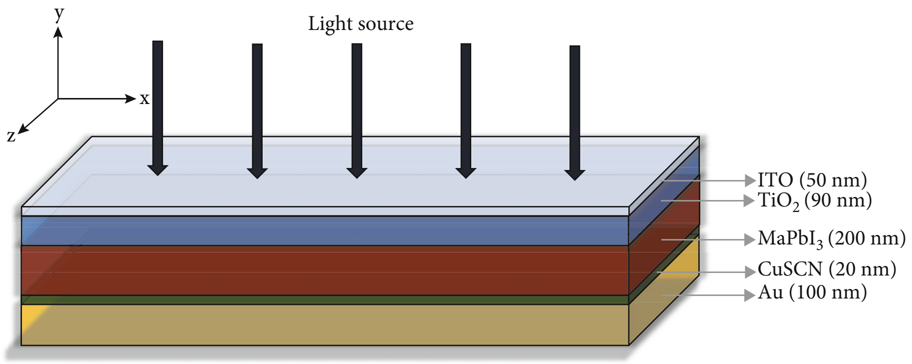


Figure 1: Sketch of PSC structure was simulated in this study; the TiO_2 layer as ETL was varied its doping and thickness.

obtained the best efficiency of 20.26% compared to ITO, PCBM, IGZO, and SnO_2 [11]. Similar to the HTL, the good ETL should have a sufficient conduction band alignment to the lowest unoccupied molecular orbitals (LUMO) of the perovskite active layers[11].

The absorption spectra for each material are simulated separately and displayed in Figure 2(a) to understand the contribution of each layer to the absorption performance of the proposed PSC structure. The gray shaded curve plane represents the absorption of the entire PSC structure, while the green, purple, red, and blue curve planes represent the absorption of each material layer, including ITO, TiO_2, $MaPbI_3$, and CuSCN. It can be seen that the $MaPbI_3$ layer contributes the most to absorption activity, which corresponds to its function as an active or absorption layer. The highest absorption reaches 0.92–0.95 at the 400–750 nm wavelength and then drops significantly to 0.1 at 800–1400 nm. On the other hand, ITO as a front contact has the lowest absorption of 0.04–0.20, indicating a suitable functionality because incident light is not trapped in the ITO and is transmitted to the active layer. TiO_2 and CuSCN have low absorption but produce a significant peak at specific wavelengths, with TiO_2 producing a peak at 312 nm and CuSCN producing a peak at 911 nm.

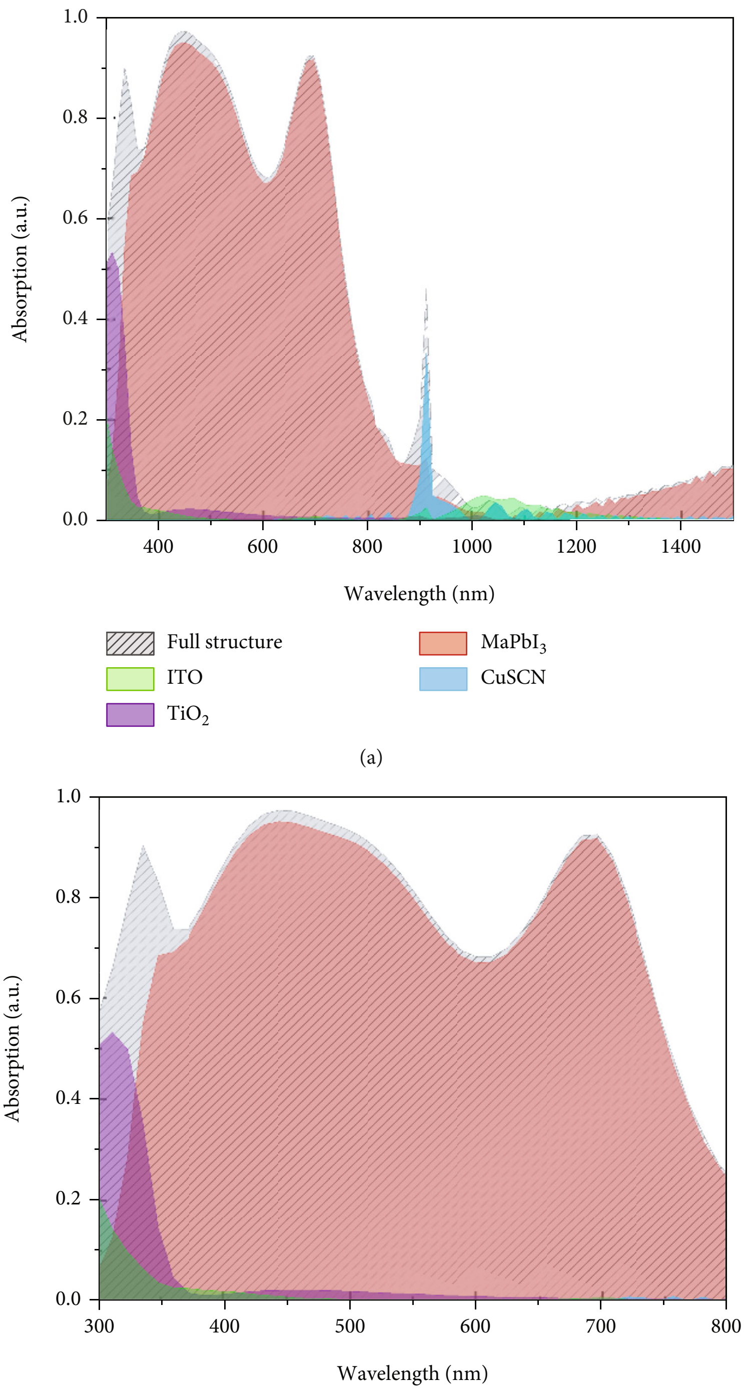
As can be seen, the absorption spectra collected in each layer support each other to construct the complete structural spectra. The total light absorbed by the structure reached 79.5% with an average absorption of 0.795 by examining the shorter wavelength range of 300–800 nm as shown in Figure 2(b) and assuming the white space as residual incident light from the reflectance loss. This reflectance loss is incident light that is not absorbed by the structure. We have shown in this section that TiO_2 contribution as an ETL layer is still minimal. The current study will optimize through doping and thickness modifications, and the enhancement effect will be studied by comparison in Figure 2.

3.2. Absorption Optimization using Doped TiO_2-based ETL.

In this study, two contrasting characteristics of materials, namely, Ta and Fe, were used for doping variation. Ta is considered a pentavalent cation that can increase the electron conductivity in TiO_2 and robust n-type characteristics [18]. Therefore, Ta doping can be expanded the energy gap of TiO_2, namely 3.5 eV and 3.7 eV, for 1.8% Ta-TiO_2 and 3.8% Ta-TiO_2, respectively [30]. Meanwhile, Fe doping which includes divalent/trivalent cations functions to modify TiO_2 into a p-type. Fe-doped TiO_2 can decrease the bandgap from 3.2 eV to 2.76 eV and 2.68 eV for 0.11% Fe-TiO_2 and 0.32% Fe-TiO_2, respectively [19]. In this section, to discover the effect of doping in ETL, the absorption spectra were compared to pure TiO_2, as shown in Figure 3. ETL thickness was kept constant at 90 nm for each doping study.

According to Figure 3, absorption in the active layer was increased for all doping modifications. The most substantial improvement was recorded around 350 nm, 520 nm, and 911 nm. Around the wavelength of 520 nm, the doped TiO_2-based ETL achieved an absorption of 0.98. When compared to TiO_2-based ETL with 0.88 absorption, indicates 0.2 less to the maximum absorption of 1.0 and 0.1 enhancement. In terms of total light absorption, the doped ETL increased to 81.7% for Fe-TiO_2 and 81.2% for Ta-TiO_2, respectively. This finding also revealed that Fe doping is slightly greater than Ta doping, with an average absorption of 0.816 and 0.810, respectively, between 300 nm and 800 nm. Higher Fe doping performance corresponds to a reduced bandgap, and decreasing TiO_2 bandgap extends the light absorption to visible light. Lower bandgap brings the material closer to the optimal bandgap for solar cells of 1.4–2.4 eV [31]. The absorption spectra differed only slightly because of the low doping concentration in Fe-TiO_2. The results, however, showed that raising the doping concentration resulted in more significant gains.

The doping effect can also be observed from the E-field profile of PSC structure in the z-normal plane. In Figure 4, a comparison of the E-field profile for the PSC structure with ETL TiO_2 and 0.32% Fe-TiO_2 was shown. The profile in Figure 4 occurred at a certain wavelength optimum based on absorption spectra in Figure 3 including 370 nm, 516 nm, 768 nm, and 912 nm. Referring to Equation (2); the E-field $|\vec{E}|^2$ is proportional to the absorption power, P_{Abs}. The color legend shows that the color shift from deep blue to red indicates a larger E-field. For each wavelength, the E-field intensity resulting by 0.32% Fe-TiO_2-based PSC is higher than TiO_2-based PSC. With 0.32% Fe-TiO_2, the high E-field is more congregates on the front contact surface until it enters the active layer as the wavelength increases.



(a)

(b)

FIGURE 2: The spectral absorption of each layer in TiO_2-based PSC in the wavelength of (a) 300–1500 nm and (b) 300–800 nm.

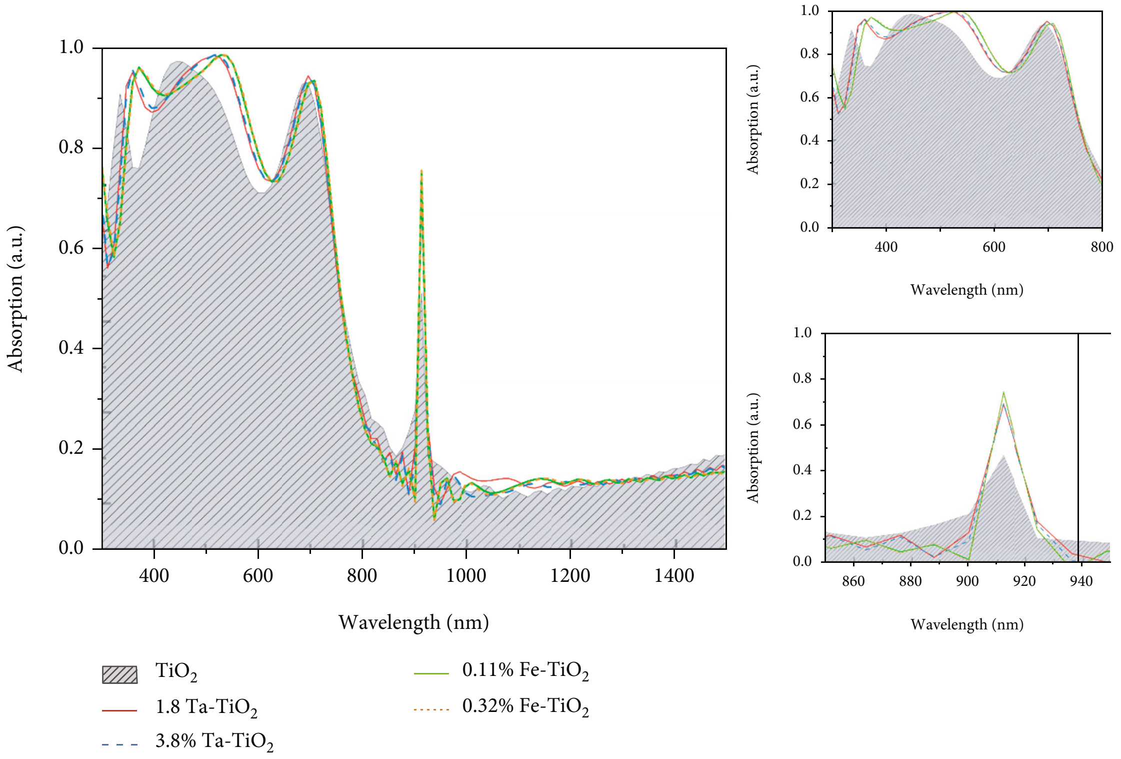


FIGURE 3: The spectral absorption of PSC with doping variations of TiO_2 as ETL layers.

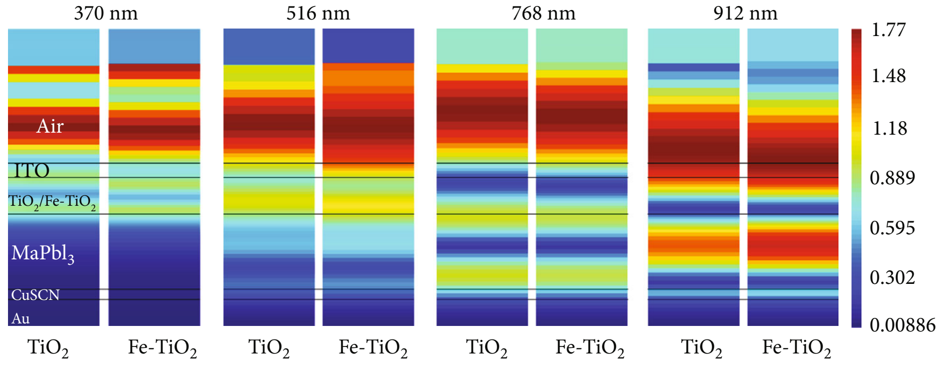


FIGURE 4: Comparison of E-field profile for PSC structure with ETL TiO_2 and 0.32% Fe-TiO_2 in wavelength of 370 nm, 516 nm, 768 nm, and 912 nm.

Generally, the light absorption in PSC can be enhanced by modifying the ETL to lead the resonance and optical enhancement. The nanostructure such as photonic crystal has been reported to result the high-performance PSC [1, 21]. Due to the strong absorption dependence on refractive index,thickness studies can also be a simple and efficient way to optimize the PSC. Tooghi et al. reported the influence of hole transport layer (HTL) thickness on the absorption, and it shows that the thinner the HTL, the better performance obtained [32]. Since the ETL was placed in front of the active layer, the thickness should not be too thick, which can block the incident light transmission nor should not be too thin to optimize the electron transport activity [32]. Several studies on PSC applied TiO_2-based ETL with thickness in the range 10–700 nm [33–37]. In this study, the thickness of 0.32% Fe-TiO_2-based ETL was varied by 20 nm, 50 nm, 100 nm, 200 nm, and 500 nm. The 500 nm ETL has a fluctuating absorption spectrum, as shown in Figure 5, indicating that incident light is reflected back into the air at certain wavelength. The thinner layer of ETL is resulting more stable absorption, but it has dropped at 650–700 nm of wavelength. The optimum absorption is provided by the 100 nm of thickness, the spectra were dropped around 750 nm, and the peak occurred around 911 nm. Finally, it can be concluded that the optimum thickness of Fe-TiO_2-based ETL is around 100 nm, with light absorption around 81%.

Finally, the proposed PSC has demonstrated excellent absorption performance. As shown in Table 1, Ta and Fe-

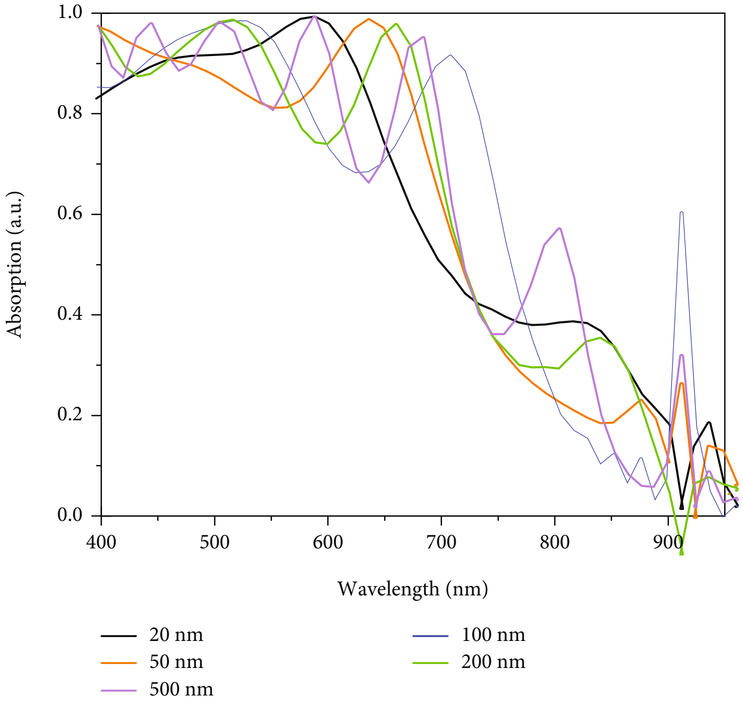


Figure 5: The spectral absorption of PSC with thickness variations of 0.32% Fe-TiO_2 as ETL layers.

Table 1: The comparison of absorption performance at 446 nm of wavelength for different TiO_2 dopant.

Structure	ETL thickness	Absorption	Reference
FTO/Er-TiO_2/$MaPbI_3$/Spiro-MeOTAD	1 μm	57%	[39]
FTO/W-TiO_2/$MaPbI_3$/Spiro-MeOTAD	5 μm	81%	[40]
FTO/Nb-TiO_2/MaPbI/Spiro-MeOTAD	40 nm	81%	[41]
FTO/Ru-TiO_2/m-TiO_2+$MaPbI_3$/Spiro-MeOTAD	50 nm	89%	[42]
ITO/Ta-TiO_2/$MaPbI_3$/CuSCN	20 nm	92%	This work
ITO/Fe-TiO_2/$MaPbI_3$/CuSCN	20 nm	92%	This work

doped TiO_2have a higher absorption percentage than other dopants reported in previous studies. Because of the higher intensity of the solar spectrum in AM1.5 [38], the comparison is specified in a wavelength of 446 nm.

4. Conclusions

In this study, the absorption enhancement of PSC-doped TiO_2 has been successfully achieved by optimizing the dopant, concentration, and thickness of ETL. The FDTD methods were used to simulate the PSC, which consist of five layers, including ITO (front contact), TiO_2 (ETL), MaPbI3 (active/absorption layer), CuSCN (HTL), and Au (back contact). The applied Fe and Ta doping into TiO_2 resulted in the higher incident light absorption of 81.7% and 81.2%, respectively, while the pure TiO_2 obtained lower absorption of 79.5%. However, the effect of doping concentrations cannot be further studied due to the slight performance difference. In general, 0.32% Fe-TiO_2 was the optimum ETL layer for the proposed PSC structure due to its ability to reduce the bandgap. The thickness optimization revealed that the thicker the ETL, the more fluctuated the spectra, and the thinner the ETL, the more the spectra dropped at short wavelength. Finally, Fe-TiO_2 with a thickness of 100 nm was concluded as the optimized ETL layers in the studied PSC. The proposed design is expected to deliver high solar cell performance in short circuit current, open-circuit voltage, fill factor, and power conversion efficiency.

References

[1] A. O. Salau, A. S. Olufemi, G. Oluleye, V. A. Owoeye, and I. Ismail, "Modeling and performance analysis of dye-sensitized solar cell based on ZnO compact layer and TiO_2 photoanode," *Materials Today: Proceedings*, vol. 51, pp. 502–507, 2022.

[2] M. A. Elrabiaey, M. Hussein, M. F. O. Hameed, and S. S. Obayya, "Light absorption enhancement in ultrathin film solar cell with embedded dielectric nanowires," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[3] L. Grinis, S. Kotlyar, S. Rühle, J. Grinblat, and A. Zaban, "Conformal nano-sized inorganic coatings on mesoporous TiO_2 films for low- temperature dye-sensitized solar cell fabrication," *Advanced Functional Materials*, vol. 20, no. 2, pp. 282–288, 2010.

[4] J. Jeon, T. Eom, E. Lee et al., "Polymorphic phase control mechanism of organic–inorganic hybrid perovskite engineered by dual-site alloying," *The Journal of Physical Chemistry C*, vol. 121, no. 17, pp. 9508–9515, 2017.

[5] K. C. Ko, S. T. Bromley, J. Y. Lee, and F. Illas, "Size-dependent level alignment between rutile and anatase TiO_2nanoparticles: implications for photocatalysis," *The Journal of Physical Chemistry Letters*, vol. 8, no. 22, pp. 5593–5598, 2017.

[6] F. De Angelis, D. Meggiolaro, E. Mosconi, A. Petrozza, M. K. Nazeeruddin, and H. J. Snaith, "Trends in perovskite solar cells and optoelectronics: status of research and applications from the PSCO conference," *ACS Energy Letters*, vol. 2, no. 4, pp. 857–861, 2017.

[7] C. W. Myung, S. Javaid, K. S. Kim, and G. Lee, "Rashba–Dresselhaus effect in inorganic/organic lead iodide perovskite interfaces," *ACS Energy Letters*, vol. 3, no. 6, pp. 1294–1300, 2018.

[8] S. Javaid, C. W. Myung, J. Yun, G. Lee, and K. S. Kim, "Organic cation steered interfacial electron transfer within organic–inorganic perovskite solar cells," *Journal of Materials Chemistry A*, vol. 6, no. 10, pp. 4305–4312, 2018.

[9] B. G. Krishna, D. S. Ghosh, and S. Tiwari, "Progress in ambient air-processed perovskite solar cells: insights into processing techniques and stability assessment," *Solar Energy*, vol. 224, pp. 1369–1395, 2021.

[10] A. M. Leguy, P. Azarhoosh, M. I. Alonso et al., "Experimental and theoretical optical properties of methylammonium lead halide perovskites," *Nanoscale*, vol. 8, no. 12, pp. 6317–6327, 2016.

[11] F. Azri, A. Meftah, N. Sengouga, and A. Meftah, "Electron and hole transport layers optimization by numerical simulation of a perovskite solar cell," *Solar Energy*, vol. 181, pp. 372–378, 2019.

[12] L. Hasanah, A. Ashidiq, R. E. Pawinanto et al., "Dimensional optimization of TiO_2 nanodisk photonic crystals on lead iodide ($MAPbI_3$) perovskite solar cells by using FDTD simulations," *Applied Sciences*, vol. 12, no. 1, p. 351, 2022.

[13] H. Zhou, Q. Chen, G. Li et al., "Interface engineering of highly efficient perovskite solar cells," *Science*, vol. 345, no. 6196, pp. 542–546, 2014.

[14] G. Yang, H. Tao, P. Qin, W. Ke, and G. Fang, "Recent progress in electron transport layers for efficient perovskite solar cells," *Journal of Materials Chemistry A*, vol. 4, no. 11, pp. 3970–3990, 2016.

[15] C. H. Hsu, K. T. Chen, L. Y. Lin et al., "Tantalum-doped TiO_2 prepared by atomic layer deposition and its application in perovskite solar cells," *Nanomaterials*, vol. 11, no. 6, p. 1504, 2021.

[16] C. Xu, D. Lin, J. N. Niu, Y. H. Qiang, D. W. Li, and C. X. Tao, "Preparation of Ta-doped TiO_2using Ta_2O_5as the doping source," *Chinese Physics Letters*, vol. 32, no. 8, article 088102, 2015.

[17] J. J. Carey and K. P. McKenna, "Screening doping strategies to mitigate electron trapping at anatase TiO_2surfaces," *The Journal of Physical Chemistry C*, vol. 123, no. 36, pp. 22358–22367, 2019.

[18] V. C. Anitha, A. N. Banerjee, and S. W. Joo, "Recent developments in TiO_2 as n- and p-type transparent semiconductors: synthesis, modification, properties, and energy-related applications," *Journal of Materials Science*, vol. 50, no. 23, pp. 7495–7536, 2015.

[19] S. Larumbe, M. Monge, and C. Gómez-Polo, "Comparative study of (N, Fe) doped TiO_2 photocatalysts," *Applied Surface Science*, vol. 327, pp. 490–497, 2015.

[20] M. G. Deceglie, V. E. Ferry, A. P. Alivisatos, and H. A. Atwater, "Design of nanostructured solar cells using coupled optical and electrical modeling," *Nano Letters*, vol. 12, no. 6, pp. 2894–2900, 2012.

[21] S. Haque, M. J. Mendes, O. Sanchez-Sobrado, H. Águas, E. Fortunato, and R. Martins, "Photonic-structured TiO_2 for high-efficiency, flexible and stable perovskite solar cells," *Nano Energy*, vol. 59, pp. 91–101, 2019.

[22] L. J. Phillips, A. M. Rashed, R. E. Treharne et al., "Dispersion relation data for methylammonium lead triiodide perovskite deposited on a (100) silicon wafer using a two-step vapour-phase reaction process," *Data in Brief*, vol. 5, pp. 926–928, 2015.

[23] J. R. DeVore, "Refractive indices of rutile and sphalerite," *JOSA*, vol. 41, no. 6, pp. 416–419, 1951.

[24] P. Pattanasattayavong, G. O. N. Ndjawa, K. Zhao et al., "Electric field-induced hole transport in copper(i) thiocyanate (CuSCN) thin-films processed from solution at room temperature," *Chemical Communications*, vol. 49, no. 39, pp. 4154–4156, 2013.

[25] N. M. Ravindra, P. Ganapathy, and J. Choi, "Energy gap-refractive index relations in semiconductors - an overview," *Infrared Physics & Technology*, vol. 50, no. 1, pp. 21–29, 2007.

[26] R. R. Reddy and Y. N. Ahammed, "A study on the Moss relation," *Infrared Physics & Technology*, vol. 36, no. 5, pp. 825–830, 1995.

[27] M. J. Mendes, S. Haque, O. Sanchez-Sobrado et al., "Optimal-enhanced solar cell ultra-thinning with broadband nanophotonic light capture," *IScience*, vol. 3, pp. 238–254, 2018.

[28] M. J. Mendes, A. Araújo, A. Vicente et al., "Design of optimized wave-optical spheroidal nanostructures for photonic-enhanced solar cells," *Nano Energy*, vol. 26, pp. 286–296, 2016.

[29] C. Roldan-Carmona, O. Malinkiewicz, R. Betancur et al., "High efficiency single-junction semitransparent perovskite solar cells," *Energy & Environmental Science*, vol. 7, no. 9, pp. 2968–2973, 2014.

[30] Z. Yong, P. E. Trevisanutto, L. Chiodo et al., "Emerging giant resonant exciton induced by Ta substitution in anataseTiO_2:

a tunable correlation effect," *Physical Review B*, vol. 93, no. 20, p. 205118, 2016.

[31] I. Ramiro and A. Martí, "Intermediate band solar cells: present and future," *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 7, pp. 705–713, 2021.

[32] A. Tooghi, D. Fathi, and M. Eskandari, "High-performance perovskite solar cell using photonic-plasmonic nanostructure," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[33] W. Ke, C. C. Stoumpos, J. L. Logsdon et al., "TiO_2–ZnS cascade electron transport layer for efficient formamidinium tin iodide perovskite solar cells," *Journal of the American Chemical Society*, vol. 138, no. 45, pp. 14998–15003, 2016.

[34] X. Sun, J. Xu, L. Xiao et al., "Influence of the porosity of the TiO_2 film on the performance of the perovskite solar cell," *International Journal of Photoenergy*, vol. 2017, Article ID 4935265, 10 pages, 2017.

[35] R. Teimouri, Z. Heydari, M. P. Ghaziani et al., "Synthesizing Li doped TiO_2 electron transport layers for highly efficient planar perovskite solar cell," *Superlattices and Microstructures*, vol. 145, p. 106627, 2020.

[36] Y. Yue, T. Umeyama, Y. Kohara et al., "Polymer-assisted construction of mesoporous TiO_2layers for improving perovskite solar cell performance," *The Journal of Physical Chemistry C*, vol. 119, no. 40, pp. 22847–22854, 2015.

[37] Y. You, W. Tian, L. Min, F. Cao, K. Deng, and L. Li, "TiO_2/WO_3 bilayer as electron transport layer for efficient planar perovskite solar cell with efficiency exceeding 20%," *Advanced Materials Interfaces*, vol. 7, no. 1, p. 1901406, 2020.

[38] K. Tanabe, "A review of ultrahigh efficiency III-V semiconductor compound solar cells: multijunction tandem, lower dimensional, photonic up/down conversion and plasmonic nanometallic structures," *Energies*, vol. 2, no. 3, pp. 504–530, 2009.

[39] H. Chen, W. Zhu, Z. Zhang, W. Cai, and X. Zhou, "Er and Mg co-doped TiO_2 nanorod arrays and improvement of photovoltaic property in perovskite solar cell," *Journal of Alloys and Compounds*, vol. 771, pp. 649–657, 2019.

[40] J. Liu, J. Zhang, G. Yue, X. Lu, Z. Hu, and Y. Zhu, "W-doped TiO_2 photoanode for high performance perovskite solar cell," *Electrochimica Acta*, vol. 195, pp. 143–149, 2016.

[41] G. Yin, J. Ma, H. Jiang et al., "Enhancing efficiency and stability of perovskite solar cells through Nb-doping of TiO2at low temperature," *ACS Applied Materials & Interfaces*, vol. 9, no. 12, pp. 10752–10758, 2017.

[42] S. Wang, B. Liu, Y. Zhu et al., "Enhanced performance of TiO_2-based perovskite solar cells with Ru-doped TiO_2 electron transport layer," *Solar Energy*, vol. 169, pp. 335–342, 2018.

Inhomogeneous Winding for Loosely Coupled Transformers to Reduce Magnetic Loss

Chinmaya Ranjan Pradhan, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, cr_pradhan@outlook.com*

Ajaya Kumar swain, *Department of Electrical and Electronics Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ajayaswain9@gmail.com*

Ipsita Samal, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, ipsitasamal55@gmail.com*

Subhrajit Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, subhrajitsahoo23@yahoo.co.in*

Abstract

Wireless power transfer has been proved promising in various applications. The homogeneous winding method in loosely coupled transformers incurs unnecessary intense magnetic field distribution in the center and causes extra magnetic loss. An inhomogeneous winding method is proposed in this paper, and a relatively homogeneous magnetic field distribution inside the core is achieved. This paper investigated the magnetic loss of homogeneous winding and inhomogeneous winding for wireless power transfer. A theoretical model was built to evaluate magnetic loss under inhomogeneous winding. The coupling coefficient and magnetic loss were investigated individually and comparisons were made between different width ratio combinations. Theoretical analysis was validated in experiments.

1. Introduction

Wireless power transfer eliminates the need for wires to connect the load from power source, and it has broad prospects in implantable medical devices, electric vehicles, etc. Several approaches to improve the energy efficiency of the wireless coupled coils have been developed [1–5].

Homogeneous winding, i.e., maintain the same distance between each turn, is widely applied in loosely coupled transformers [6–8]. And, this traditional winding method incurs the inhomogeneous internal magnetic field distribution; the magnetic induction intensity is concentrated in the central area, which results in greater loss in core center. An inhomogeneous winding method is proposed in this paper: coils were loosely winded in the center while tightly winded on two ends (Figure 1), so as to realize homogeneous magnetic field distribution inside the core (Figure 2) and reduce magnetic loss.

Steinmetz equation is the most used method to characterize core losses [9, 10]. However, for wireless power transfer system, the uneven flux density distribution in the core makes it difficult to employ Steinmetz equation directly. Moreover, if we divide the core into several sections, the core loss in each section is incurred not only by its own windings but also by its adjacent windings. Alternatively, FEM simulation software is widely applied to calculate core loss [11, 12]. However, this method is very time-consuming, especially for more accurate 3D models. In addition, the optimization of system parameters can only by realized by sweeping design parameters. The optimized point could be missed since it lacks an overall understanding of the whole optimization region. In the view of these problems, a magnetic circuit model [13] is proposed in this paper; it is valid for solenoid winding structure and convenient to obtain flux distribution.

It would be desirable to reduce the magnetic loss while maintaining tight and compact windings. In this paper, different winding parameters were investigated and compared, in terms of coupling coefficient and magnetic loss.

This paper is arranged as follows. A magnetic circuit model is proposed in Section 2 to calculate the magnetic loss under inhomogeneous winding. Section 3 investigates the influence of winding parameters. The experimental setup and results are discussed in Section 4. Conclusions are drawn in Section 5.

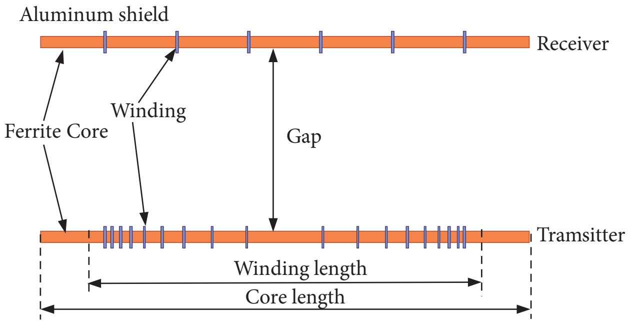


Figure 1: Structure of inhomogeneous winding for loosely coupled transformers.

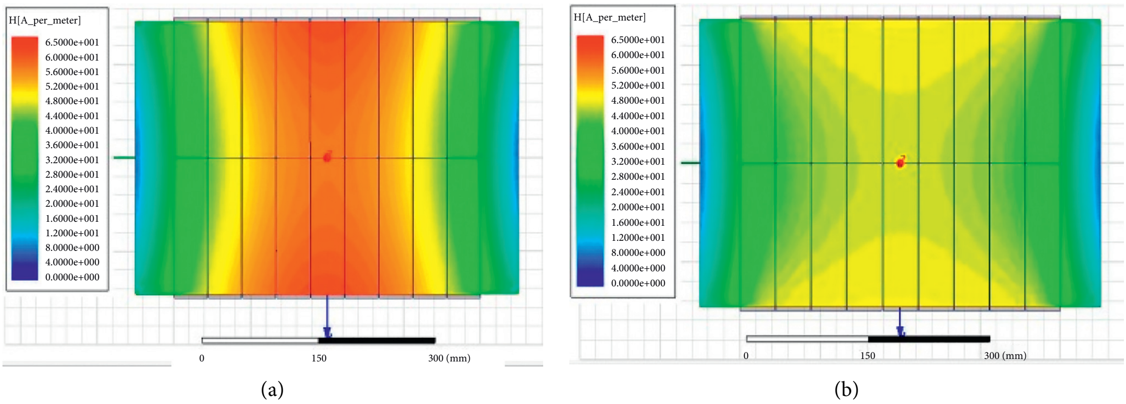


Figure 2: Internal magnetic field. (a) Homogeneous winding. (b) Inhomogeneous winding.

2. Magnetic Losses of Ferrites under Inhomogeneous Winding

A typical wireless power transfer system is illustrated in Figure 3. It contains a full-bridge inverter and rectifier and corresponding compensating topology. The loosely coupled transformer contains a transmitter coil, a receiver coil, and corresponding magnetic cores. In this paper, the transmitter and receiver coils are both solenoid winding. On the primary side, the resonant capacitor C_P connects in series with the transmitter coil, to form the resonant network. On the secondary side, the resonant capacitor C_S connects in parallel with the receiver coil.

A typical flux distribution of the solenoid structure is shown in Figure 4. The total flux concerns the internal leakage flux, external leakage flux, and mutual flux, among which the internal leakage flux comprises the majority of leakage flux since its path length is much shorter than others.

Corresponding equivalent magnetic reluctance network is analyzed in Figure 5. As seen in Figure 5, the core is divided into 7 parts in longitudinal direction. In order to clearly demonstrate the magnetic motive force and the magnetic reluctance in each flux path, at least 7 divisions have to be provided. With more divisions, theoretically, we can obtain a more accurate result, but the calculation complexity will increase dramatically. The discretization number is a trade-off between precision and complexity. Each core section is modelled as magnetic reluctance, while core section with excitation winding is modelled as a voltage source in series with magnetic reluctance. The voltage source corresponds to the number of turns and current excitation in the winding, represented by $\alpha \cdot i_p$ in Figure 5.

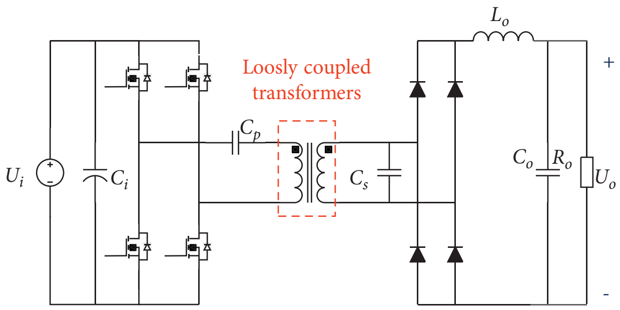


Figure 3: Typical wireless power transfer system.

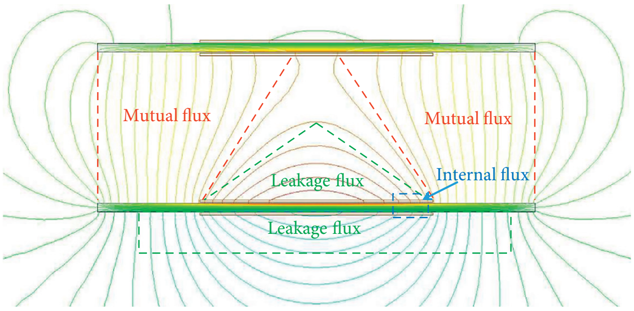


Figure 4: Flux distribution of the solenoid structure.

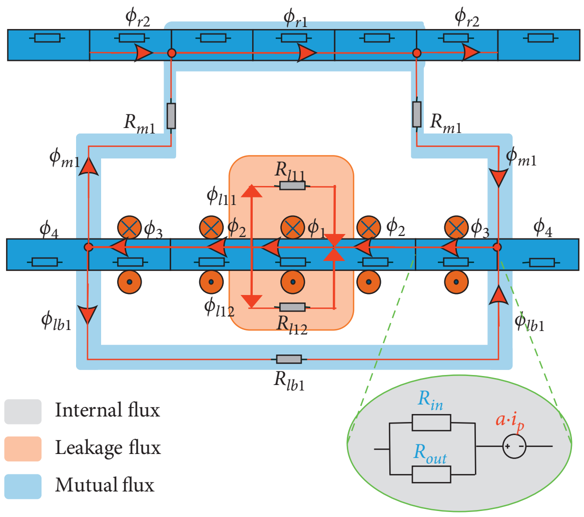


Figure 5: Equivalent magnetic reluctance network.

A lumped magnetic reluctance model was built in Figure 6, concerning voltage source and reluctance inside the core and in the air.

In the situation of uneven flux distribution, values of parameters $\{R_{l1}, R_{l2}, R_{l3}, R_{l4}\}$ and $\{R_{m1}, R_{m2}, R_{m3}, R_{m4}\}$ cannot be derived using the empirical equation. FEM simulation is applied once to obtain flux distribution $\{\phi_1, \phi_2, \phi_3, \phi_4\}$. Substitute the flux values into the model in Figure 6; the magnetic reluctance values can be derived according to Kirchhoff's voltage law.

In addition, the magnetic reluctance R_c can be calculated as

$$R_c = \frac{l_c}{\mu_0 \mu_r A_c}, \tag{1}$$

where l_c is the length of each individual core, μ_0 is the vacuum permeability, and μ_r is the relative permeability of cores.

The total flux density along y-axis is

$$B = \frac{\phi}{A_c}, \tag{2}$$

where A_c is the cross-sectional area of the core in the x-z plane.

After obtaining all of the parameters in the lumped magnetic reluctance model, the magnetic flux density under different working conditions can be acquired:

$$B = \frac{A^{-1} U}{A_c}, \tag{3}$$

where U is the magnetic motive force (MMF) matrix and A is the magnetic reluctance coefficient.

It is worth noting that the flux density in the primary core is excited not only by the primary winding but also by the secondary winding. By combining the results generated by primary and secondary excitations, the flux density distribution in the primary core can be calculated.

As a result, for cores under sinusoidal current excitation, the magnetic loss can be calculated using the Steinmetz equation:

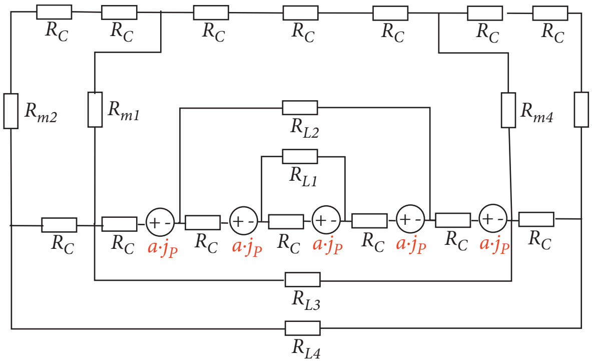


Figure 6: Lumped magnetic reluctance model.

$$P_V = k \cdot f^{\alpha} \cdot \widehat{B}^{\beta}, \tag{4}$$

where $\widehat{B}$ is the peak induction of a sinusoidal excitation with frequency f, P_V is the time-average power loss per unit volume, and k and α are material parameters which can be obtained from the material datasheet.

By substituting (3) into (4), the magnetic loss in each section can be calculated. Generally, the highest flux density is designed well under saturation; thus, the ferrite usually works in the linear region and the overall magnetic loss can be summed.

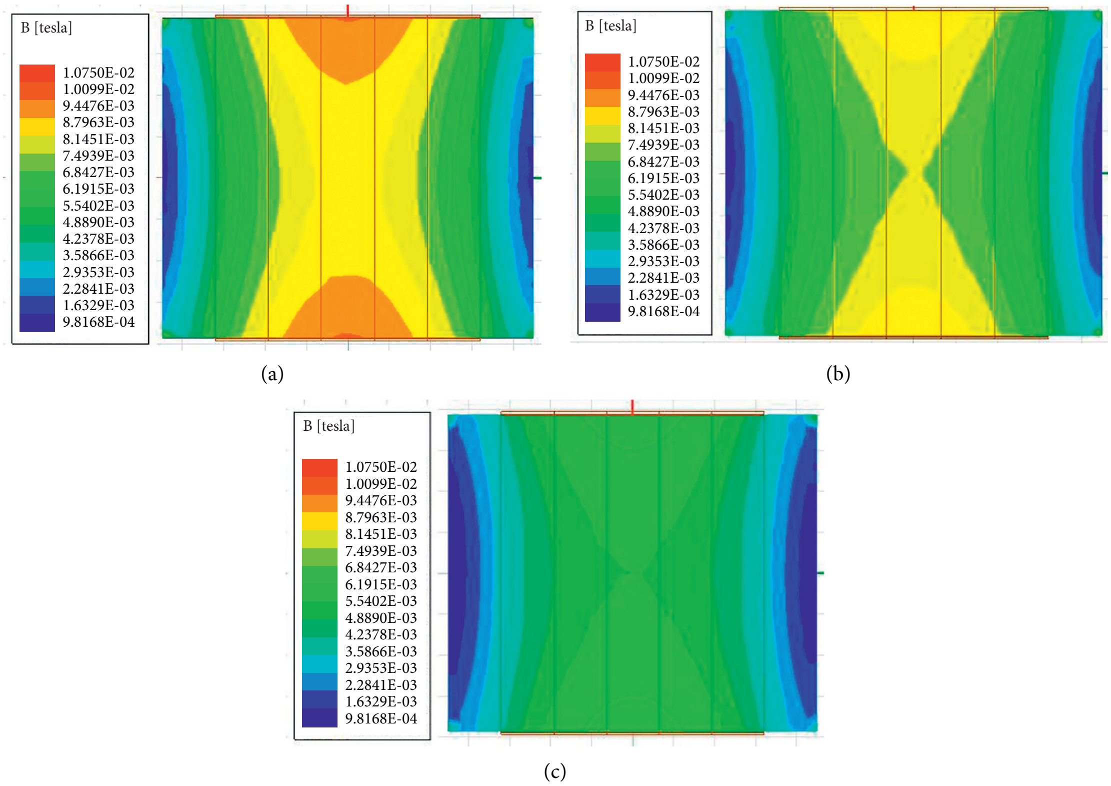
3. Influence of Winding Parameters

To investigate the influence of winding parameters, we constructed two coupled coils, each with the same number of turns in total and different winding spaces between each turn. The receiver coil is designed to be homogeneous winding, while the transmitter coil was constructed with different spaces between each turn.

It would be desirable to reduce the magnetic loss while maintaining tight and compact windings. In this paper, different coil winding width ratio combinations were investigated and compared, in terms of coupling coefficient and magnetic loss.

The system configurations are as follows. The magnetic core is made of ultra-low-loss soft magnetic material DMR47. The overall dimension of the core is $500 * 380 * 12$ mm, which is formed by small magnetic cubes ($50 * 38 * 6$ mm). The number of primary and secondary coil turns is both 45 turns. The air gap between primary and secondary coils is 200 mm. The system works at its resonant frequency 50 kHz. The input voltage is 160V and the load of the system is 50 Ω.

The coils were equally divided into 5 portions. The number of portions determines the number of combination possibilities of winding density. With more portions, we can obtain a much more accurate result, but the calculation complexity will increase dramatically. The discretization number is a trade-off between precision and complexity. The current in each section $\alpha \cdot i_p$ corresponds to the number of turns. Serially connect the 5 portions, and calculate the coupling coefficient of the receiver side with the transmitter side. Apply different current in each portion to realize the



(a)

(b)

(c)

Figure 7: Magnetic field distributions in core with different winding parameters. (a) Homogeneous winding (9 : 9 : 9 : 9 : 9). (b) Winding parameters (10 : 10 : 5 : 10 : 10). (c) Winding parameters (17 : 4 : 3 : 4 : 17).

effects of inhomogeneous winding. A one-row five-column array indicates the current value in 5 portions. As seen in Figure 7, (9 : 9 : 9 : 9 : 9) indicates 9 turns in each portion.

The receiver coil remains homogeneous winding, while different winding parameters of the transmitter coil were studied to seek for the optimal combination. Typical theoretical results were compared with simulations in ANSYS Maxwell; the effects of inhomogeneous winding were shown in Figures 7(a)–7(c) under the same magnetic induction intensity scale. The simulation results agree well with theoretical analysis, as shown in Table 1.

Table 1: Comparison of core loss in theoretical and simulation results.

Winding parameters	Theoretical loss (W)	Simulation loss (W)
(9 : 9 : 9 : 9 : 9)	16.05	16.08
(10 : 10 : 5 : 10 : 10)	15.55	15.58
(17 : 4 : 3 : 4 : 17)	15.10	15.12

In order to illustrate the effects of inhomogeneous winding on magnetic loss, the variations of coupling coefficient and core loss are depicted with different winding parameters, as shown in Figure 8.

The horizontal axis represents the coupling coefficient, while the vertical axis is 1/core loss. The desired winding parameters are with high coupling coefficients and low core losses, so points positioned in the top right region are preferred.

As seen from Figure 8, the range of coupling coefficient is limited between 0.118 and 0.122. Homogeneous winding (9, 9, 9, 9, 9) has the worst performance, with the lowest coupling coefficient and highest core loss, compared with other cases with coarse winding in center. Inhomogeneous winding effectively reduced the magnetic loss in the ferrite core. The optimal situation within consideration range, with winding parameter (17, 4, 3, 4, 17), reduced the core loss by 5.6% compared with the homogeneous case, while the coupling coefficient increased by 1.9%. When designing the winding parameters, relative low magnetic field density in the core center helps to reduce the overall core loss.

4. Experimental Result and Discussion

Measurements for wireless power transfer system under inhomogeneous winding were obtained to evaluate whether the power loss reduction found for inhomogeneous winding translated to improve power efficiency well.

Primary coil turns are winded around the core with different spaces between each turn. The total number of turns remains constant, and the working frequency is unchanged, so the copper loss is assumed to be the same, and so is the eddy-current loss. The circuit always works at the soft switching mode, so the switching loss remain unchanged. Therefore, with different winding parameters, the variation in core loss results in the change in system efficiency. The overview of the experiment system is shown in Figure 9.

The diagram of setups for measuring is shown in Figure 10. Test waveforms of the system are shown in Figure 11, including the primary current i_p, secondary current i_s, and

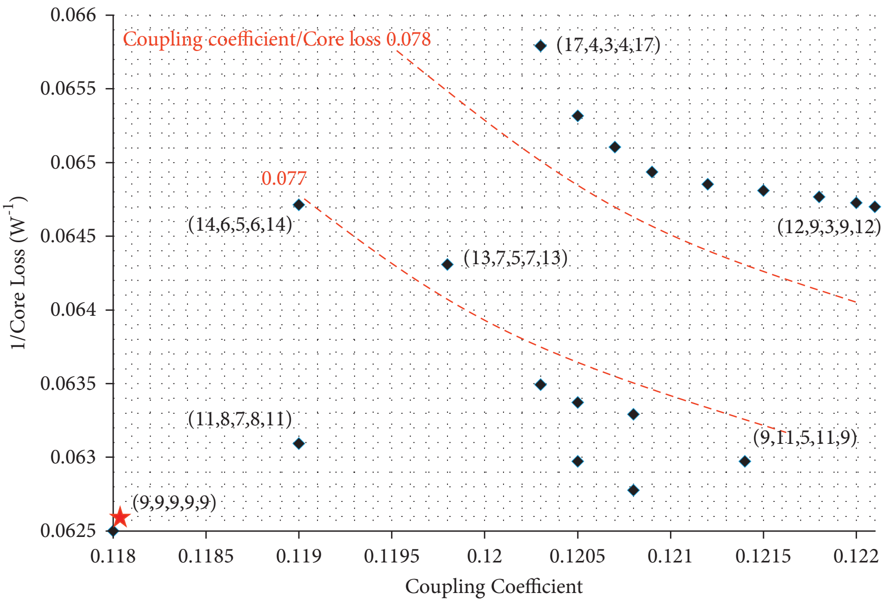


FIGURE 8: Variation of core loss and coupling coefficient with different winding parameters.

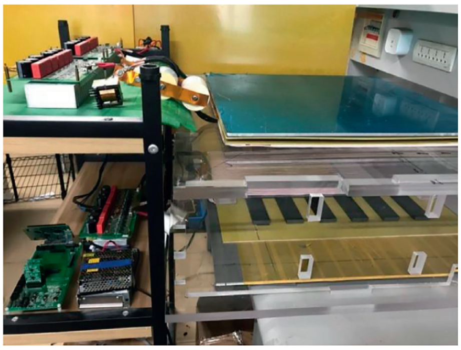


FIGURE 9: Experimental setup for wireless power transfer under inhomogeneous winding.

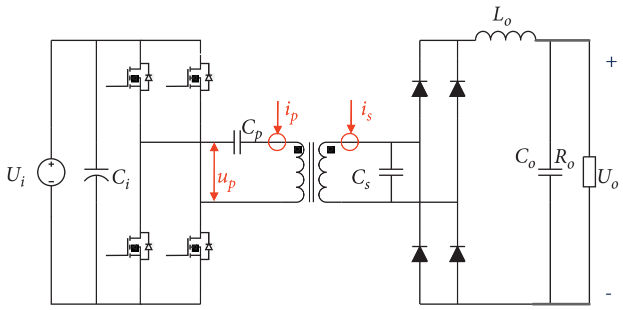


FIGURE 10: The diagram of the setups for measuring.

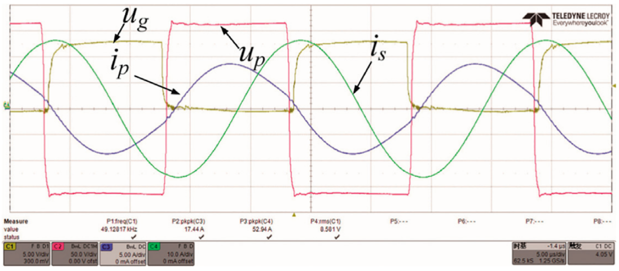


FIGURE 11: Test waveforms of the system.

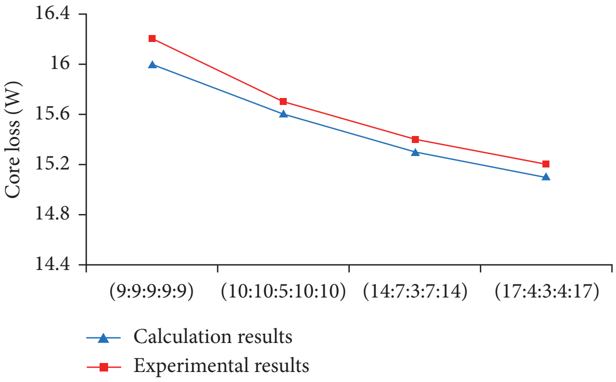


FIGURE 12: Variation of core loss with different winding parameters.

output voltage of the inverter u_p. u_g represent the gate driving signal for MOSFET in the inverter bridge. The system works under ZVS condition, as not much of a voltage spike is observed in the waveform.

The experimental and calculation results are compared in Figure 12. Since it is difficult to directly obtain the core loss, the experimental core losses were obtained by subtracting the measured total losses with the measured winding losses by using a LCR meter, the switching device losses, and the diode losses by integrating the voltage and current waveforms using a HDO4034 oscilloscope. The variation of measured core loss agrees well with the calculated results, assuming the winding loss and switching loss remain unchanged under the same frequency. The system power loss significantly reduces as coils are wound relatively coarsely in the center and tightly at two ends.

5. Conclusions

A novel inhomogeneous winding method for loosely coupled transformers is proposed in the paper. A magnetic reluctance model for solenoid structure is built to calculate the core loss under inhomogeneous winding. Sweeping maps with different primary winding parameters were provided to investigate the optimal combination. The obtained experimental results show great agreement with the presented optimization. Compared with traditional homogeneous winding, the new inhomogeneous winding method effectively reduces magnetic loss in the ferrite core, while maintaining tight coupling between primary and secondary coils.

References

[1] S. Lee, D. H. Kim, Y. Cho et al., "Low leakage electromagnetic field level and high efficiency using a novel hybrid loop-array design for wireless high power transfer system," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4356–4367, 2019.

[2] M. Y. Li, X. Y. Chen, H. Y. Gou et al., "Conceptual design and characteristic analysis of a sliding-type superconducting wireless power transfer system using ReBCO primary at 50Hz," *IEEE Transactions on Applied Superconductivity*, vol. 29, no. 2, Article ID 5501304, 2019.

[3] Z. H. Ye, Y. Sun, X. Dai, and C. Tang, "Energy efficiency analysis of U-coil wireless power transfer system," *IEEE Transactions on Power Electonics*, vol. 31, no. 7, pp. 4809–4817, 2019.

[4] Z. Yan, Y. Li, C. Zhang, and Q. Yang, "Influence factors analysis and improvement method on efficiency of wireless power transfer via coupled magnetic resonance," *IEEE Transactions on Magnetics*, vol. 50, no. 4, Article ID 4004204, 2014.

[5] J. Liu, Q. Deng, D. Czarkowski, M. K. Kazimierczuk, H. Zhou, and W. Hu, "Frequency optimization for inductive power transfer based on AC resistance evaluation in litz-wire coil," *IEEE Transactions on Power Electronics*, vol. 34, no. 3, pp. 2355–2363, 2019.

[6] Z. Cheng, Y. Lei, K. Song, and C. Zhu, "Design and loss analysis of loosely coupled transformer for an underwater high-power inductive power transfer system," *IEEE Transactions on Magnetics*, vol. 51, no. 7, Article ID 8401110, 2017.

[7] X. Liu, C. Liu, W. Han, and P. W. T. Pong, "Design and implementation of a multi-purpose TMR sensor matrix for wireless electric vehicle charging," *IEEE Sensors Journal*, vol. 19, no. 5, pp. 1683–1692, 2019.

[8] T. Gonda, S. Mototani, K. Doki, and A. Torii, "Effect of air space in waterproof sealed case containing transmitter and receiver of wireless power transfer in sea water," *Electrical Engineering in Japan*, vol. 206, pp. 24–31, 2019.

[9] J. Muhlethaler, J. Biela, J. W. Kolar, and A. Ecklebe, "Core losses under the DC bias condition basedon steinmetz parameters," *IEEE Transactions on Power Electronics*, vol. 27, no. 2, pp. 953–963, 2012.

[10] S. C. Tang and N. J. McDannold, "Power Loss analysis and comparison of segmented and unsegmented energy coupling coils for wireless energy transfer," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 3, no. 1, pp. 215–225, 2015.

[11] K. E. I. Elnail, X. L. Huang, X. Chen, L. L. Tan, and H. Z. Xu, "Core Structure and electromagnetic field evaluation in WPT systems for charging electric vehicles," *Energies*, vol. 11, no. 7, p. 1734, 2018.

[12] X. Zhang, S. L. Ho, and W. N. Fu, "Quantitative analysis of a wireless power transfer cell with planar spiral structures," *IEEE Transactions on Magnetics*, vol. 47, no. 10, pp. 3200–3203, 2011.

[13] Y. Tang, F. Zhu, and H. Ma, "Efficiency optimization with a novel magnetic-circuit model for inductive power transfer in EVs," *Journal of Power Electronics*, vol. 18, no. 1, pp. 309–322, 2018.

Energy Auditing in Three-Phase Brushless DC Motor Drive Output for Electrical Vehicle Communication Using Machine Learning Technique

Pranay Rout, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pranayrout93@gmail.com*

Madhulita Mohapatra, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar,madhulitamohapatra@gmail.com*

Alekha Sahoo, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, alekha.sahoo241@gmail.com*

Pratik Mohanty, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, pratikmohanty92@hotmail.com*

Abstract

Using predictive nonlinear optimal control, this model examines the output power of a three-phase brushless DC motor (BLDC) drive to ensure that it is stabilized (PNOC). A BLDC is a kind of electric motor that is used in a variety of applications and is one of the models of electric motors that are utilized in constant speed applications. In this motor, the movable component of the rotor created torque and the rotor rotated in a position of low reluctance; the location of the rotor is determined by the motor's maximum inductance value. The BLDC drive controls the motor via the converter circuit, and the converter circuit ensures that the motor receives the appropriate output power. The project manager should have a thorough discussion with the team about the demagnetization of the malfunctioning BLDC motor before beginning this job. It is possible to model a machine using many existing technologies, such as electrical equivalent circuit diagram (EEC), which are based on a number of assumptions that make the analysis process or the analysis approach simpler. Despite numerical methodologies, these approach scenarios give frequency domain loop (FDL) precision frequency domain, using a suitable weight strategy to deliver high power solution creation (NM). The purpose of this essay is to integrate these two technologies in order to make contributions via the development of a new hybrid EEC-FDL model closed-loop brushless DC motor. PNOC is a driving system that uses predictive nonlinear optimal control (PNOC). The generated model is subjected to simulations under both healthy and incorrect settings, respectively. MATLAB software is utilized to construct the simulation of the control circuit, and simulation outputs are validated by experimental findings. Predictive nonlinear optimal control (PNOC) is employed to eliminate torque ripple and improve system stability.

1. Introduction

Electrical motors are used in a variety of applications by all types of industrial, commercial, and other equipment. Research conducted by the Electric Power Research Institute (EPRI) found that motors are responsible for 51 percent of the world's total electricity generation [1]. Energy usage is measured in kilowatt hours. The amount of energy utilized by other industries is little in comparison. Lighting, for example, contributes for 19 percent of total expenditure, as does the heating and cooling system (16 percent), with information technology accounting for 14 percent. Unquestionably, there is a need for the use of an efficient and durable controller for motor control which will result in cost saving energy. Electric motors have an impact on practically every element of contemporary life, refrigerators, vacuum cleaners, elevators, air conditioners, washing machines, and other similar appliances. Electric motors are used in a variety of applications such as fans, computer hard disc drives, and industrial operations. In truth, motors utilize the majority of the energy, regardless of the situation at hand applicability in a residential, industrial, or commercial setting. Its energy efficiency is high [2]. The kind of motor determines the performance of the motor. Some are designed to be more energy efficient. Some are effective, while others are not. In addition, there has been a recent fast development of motor drives in the vehicle business, using modern hybrid technology; it has created a fantastic opportunity. Variable speed motor drives with excellent energy efficiency are in high demand. Precision and reliability are essential in many adjustable speed drive speed control that is continuous and stable throughout time, with excellent transient response and improved performance, as well as increased efficiency. Conventional direct current motors are quite efficient [3]. The linear torque-speed characteristic is efficient, easy to produce, and linear in torque. However, several of these criteria are not met by conventional direct current motors, mostly because of their size. DC motors necessitate periodic maintenance due to the presence of a commutator and brushes. Brush cleaning and replacement are important tasks. As a result, DC motors have a limited application, applications in the commercial sector. Furthermore, in the relationship between the supplied torque and the size, because of the motor's low performance, it can only be used in situations where size and weight are limited. Weight and size are critical considerations, particularly in electric cars and aeroplanes and applications. As a result, brushless direct current (BLDC) motors have emerged as a viable alternative, a more efficient alternative to traditional direct current motors [4].

Because of the depletion of fossil fuel and energy resources, as well as the increasing emphasis on society, the creation of the electric vehicle system is a promising and effective mode of personal and urban transportation. Because of the short battery life, electric motors for automotive applications are often designed to be more energy efficient while also having a smaller overall volume. Since the concept of progress and modern conversion technology, permanent magnet material has a certain Predictive nonlinear optimal control (PNOC) motor with inverter control which is widely used in automotive applications [5]. Permanent magnet material has a certain predictive nonlinear optimal control (PNOC) motor with inverter control.

In this particular instance, the BLDC motor type provides the optimum overall performance and efficiency of the motor system. Motors become popular much more quickly as a result of this than other types of motors [6]. These motors are well suited for applications requiring high safety and complexity, and they are often used in automotive and industrial applications. DC motors with styling angles that are comparable to brushless DC motors are available. When it comes to current and torque and voltage and speed, there is a nonlinear connection between them. BLDC motors can run at faster speeds than brushed motors because of an electronic control system that substitutes the typical mechanical transmission used in brushed motors [7]. There are several benefits to using a BLDC motor, including improved speed and torque characteristics, high dynamic responsiveness, great efficiency, noiseless operation, and increased speed. It is possible to get improved speed and torque characteristics while also decreasing electromagnetic interference using these (EMI). Control modelling, control scheme selection, simulation, and parameter modification of the PLDC motor drive system are all included in the practical application of the system [8]. PNOC (predictive nonlinear optimum control) of the drive system with optimal control of a complex process system must be implemented, and the system parameters must be adjusted as a result of this.

The speed control of the BLDC drive system was suggested based on the numerous controls that were tested. Most industrial process controllers, on the other hand, should linearly be, to varied degrees, a mathematical model of the system [9], the parameters of variability and uncertainty, and the parameters of uncertainty and variability. When it comes to controlling variables, selecting and adjusting them become more challenging and less reliable. The predictive nonlinear optimization control (PNOC) optimum control system is a simple system that relies on the prediction of the control device to achieve optimal control. Due to the ease with which it may be achieved and controlled, it is dependent on developing mathematical models. Additionally, the variable speed/torque application offers a straightforward, effective control-based system that is both sturdy and dynamically responsive [10].

2. Review of the Literature

Here, the properties of the BLDC motor and its prior control techniques are addressed for the goals of application and research. The characteristics of the BLDC motor and its previous control strategies are examined.

In the absence of a sensor for the brushless motor control system, the transmission signal serves as the primary indicator of the system's performance. We presented a reconstruction technique based on the terminal voltage-transfer compensation strategy [1] in order to increase the accuracy of the converter. The active control system takes the role of the conventional control system, which controlled the speed and current of the BLDC motor. The suggested Field Programmable Gate Arrays (FPGAs) are controlled directly on the inverter input power with the aid of the Dynamic Power Containment Technology (DPCT) control system [2], which is based on Dynamic Power Containment Technology (DPCT). When this occurs, the magnetic flux density, flux distribution, and torque of finite element models will only be utilized to evaluate motor performance and to analyse and approve the design of the system [3, 4].

The improved Tunicate Swarm optimization algorithm (ITSA) was employed in this work to optimise the design, which was carried out utilizing speed and torque controllers. As a result, the best gain parameters are calculated with the aid of the intended scope function, which is necessary to increase the controller function [4, 5].

These brushless DC motors have recently gained popularity among designers, who point to their benefits in terms of simplicity, high output (torque), long-term usage, and high-speed stability [6] as just a few of their many advantages. For the speed control of a BLDC motor, research has been conducted on the control rate, performance, and comparisons between a traditional Proportional-Integral (PI) controller and an artificial neural network (ANN). The design of a suitable PID controller for a BLDC motor speed controller is critical to the controller's proper functioning. Design and speed control automated tuning, as well as an enhanced PID algorithm, are all presented in this system, which was recently inspired by a speed control approach for a BLDC motor [11–13]. Research and development are carried out in this system to achieve speed control of the BLDC motor utilizing the PI controllers and the fuzzy controllers, respectively. In most industrial applications, the BLDC motor control speed is used in conjunction with a normal PI controller; however, this produces nonlinear circumstances with detrimental impacts that are present at various levels. As a result, the fuzzy logic control technique [14–18] is used to counteract this obnoxious behaviour by implementing the constant speed control method.

It is the primary focus of this article to explore the use of soft computing technology to manage the speed of BLDC motors. Instabilities in the control system include unwanted overshoot, a longer stabilisation period, and vibration as the system transitions from one state to the next [11]. When utilized in conjunction with a closed-loop controller setup, PID and self-adjusting fuzzy control technology may be employed to overcome maximal overshoot and a lengthy stabilisation time [19, 20]. In this work, the signal is communicated from the Android phone to the Arduino Uno through the Bluetooth module linked to the Arduino Uno, which is based on the wireless speed control technology and uses a DC motor that does not have a Bluetooth brush to convey the signal [12]. A pulse width modulation (PWM) approach is used to regulate the speed of a permanent magnet direct current (PLDC) motor [21–23].

A significant influence has been made on engineering applications by advances in artificial intelligence (AI) approaches such as neural networks, fuzzy logic, genetic algorithms, and particle swarm optimization techniques [13]. Recently, artificial intelligence approaches have had a significant influence on the field of electrical engineering, notably in the domain of power electronics application in motor drive systems. In its most basic form, artificial intelligence (AI) is a computer simulation of human thought, which is referred to as computational intelligence. They are used in a variety of fascinating applications in the fields of power electronics and motor drives.

Artificial intelligence approaches may be divided into four broad areas.

The expert system, fuzzy logic (FL) system, artificial neural network (ANN), and genetic algorithm are the four types of algorithms.

For expert systems, when it comes down to it, an expert system is just a computer software based on Boolean logic that is supposed to transmit the knowledge and competence of a human person across a variety of fields to solve complicated issues. Expert system computing is referred to as "hard" or exact computing, while FL, ANN, and GA computing are referred to as "soft" or approximation computing, respectively. The software for the knowledge base or the rule base is carefully structured in such a way that it is simple to learn, modify, and update its contents. Expert systems are used in a variety of applications, including controller parameter tweaking, problem diagnosis, and automated drive testing.

Fuzzy logic systems are a kind of logic system that is ambiguous.

Fuzzy logic control is another kind of artificial intelligence technology; however, it has a more recent history than evolutionary systems. A FLC is a heuristic method to nonlinear system design that incorporates knowledge and important features of human thinking into the process of creating nonlinear systems. In fuzzy set theory, a specific item has a degree of membership in a given set that may be anywhere between 0 and 1. The degree of membership in a given set can be anywhere in the range of 0 to 1.

Each fuzzy set is defined by a linguistic variable, which is in turn defined by a multivalued membership function, which is defined by a linguistic variable. The fuzzy set theory is used to design the FL controllers. As a result, the bounds of fuzzy sets are hazy and unclear, which makes them suitable for approximation models since they may be approximated.

Artificial neural networks (ANNs) are a kind of neural network that may be programmed to do certain tasks.

Artificial neural networks (ANNs) have been widely used in the field of function approximation. It is not necessary to use a mathematical model while using ANN approaches.

When correctly tuned, they provide better performance than traditional controllers while requiring less tuning work than conventional controllers. They have a straightforward design 6 approach that incorporates data from a real-world system, even in the lack of specialised knowledge or expertise. For applications in power electronics and motor drive, the ANN is used to estimate the rotor speed, flux, resistance, and position by measuring the position of the rotor.

For algorithm with genetic components, genetic algorithms are a subset of evolutionary computing algorithms that employ a probabilistic approach to solve optimization and search issues. Genetic algorithms are a subset of evolutionary computation algorithms. In GA-based solutions, an initial population is assumed, and then, the optimization is achieved after numerous generations of reproduction, crossover, and mutation operations are performed on the population. Numerous circumstances, such as solving nonlinear system transcendental equations and optimising FLC, call for the use of genetic algorithms. In addition, genetic algorithms are not influenced by the local minima since they do not use derivatives.

The FL control approach is used in the majority of consumer devices, including as washing machines, autofocus cameras, and air conditioners, among others. When compared to FLC and ANN, the use of the evolutionary system and the GA in the area of power electronics and motor drives is severely constrained [14].

3. Materials and Machines

This paper covers the design and analysis of a BLDC motor speed control system based on predictive nonlinear optimal control (PNOC) controllers. By regulating the magnetic flux and torque components of the BLDC motor drives, you may control the system's quick unstable reaction and hence control its rapid unstable response [15]. This system employs a predictive nonlinear optimal control (PNOC) controller in order to obtain the best possible dynamic performance for the driver system that has been suggested.

In the motor drive business, PI controllers are widely utilized because of their simplicity, robustness, clarity of operation, ease of control, and convenience of installation.

This controller performs well in a linear system as well as in systems with a predetermined set of known parameters or load conditions, as seen in the example below. When the parameters depart from their known values, the responsiveness of the system deteriorates, resulting in a loss of stability in the system [16]. When the transfer function of the motor drive system changes depending on the operating circumstances, the controller settings must be adjusted to account for this fluctuation. Although the settings of the PI controller are generally preset offline, the controller is unable to deal with the operating circumstances of the drive system when it is in operation [17].

The nonlinear VI characteristics of switches, nonlinear inductances, and electromagnetic couplings between components make BLDC motor drives very nonlinear owing to the fact that they are electronically commutated.

Nonlinearity is caused primarily by the omnipresent switching element, which renders all power electronic systems substantially nonlinear, even if all components are considered to be perfect in their performance.

Because of this, the controller's parameter must be modified in order to accommodate the operating state. Customizing settings manually using the trial and error approach or the Ziegler-Nichols tuning criterion raises the computational 84 loads on the system and causes it to operate at a slower rate during its whole operation. In this chapter, we will examine how to regulate the speed of the BLDC motor drive with the use of a PI controller, whose gains are gained by trial and error [18]. The next part shows how the GA was employed in the optimum tuning of the gains of the PI controller, as well as the results of the experiment. The performance of the drive with the manually tuned PI controller is compared to the performance of the drive with the genetic algorithm-based optimally tuned PI controller for a change in the set speed, load disturbances, and variation of inertia, among other things.

Figure 1 depicts the architectural framework for a sensorless BLDC motor that has been presented. It is vital to note that during BLDC motor operation, the VSI, or voltage source inverter, creates magnetic flux and is also set at the motor angle to make its functioning easier and to play an important role [19]. According to the BLOC motor theory, there is basically a supply line running between the two phases of the motor. One of the phases generates an electromagnetic field (EMF). Back EMF provides fundamental sensing capabilities, which aid in the detection of postzero crossings. It also creates pulses with location codes, which aid in the detection of postzero crossings [20]. As a result, the switch operation choice may be regulated based on the switching table information. Incorporating this loop with an advanced predictive nonlinear optimal control (PNOC) controller will significantly improve the entire function's stability.

3.1. Modelling of the BLDC Motor. The following four equations can describe the three-phase star connected to the BLDC motor. The three-phase BLDC motor equivalent circuit is shown in Figure 2.

The benefits of using KVL loop phase stator windings are given below.

$$V_a = R_a + L_a \frac{di_a}{dt} + M_{ab} \frac{di_b}{dt} + M_{ac} \frac{di_c}{dt} + e_a, \tag{1}$$

$$V_b = R_b + L_b \frac{di_b}{dt} + M_{bc} \frac{di_a}{dt} + M_{ba} \frac{di_c}{dt} + e_b, \tag{2}$$

$$V_c = R_c + L_c \frac{di_c}{dt} + M_{ca} \frac{di_a}{dt} + M_{cb} \frac{di_b}{dt} + e_c. \tag{3}$$

When the back EMF waveform e_a, e_b, and e_c and angular velocity of the rotor shaft function,

$$e = K_e \omega_m, \tag{4}$$

where K_e is the back EMF constant.

Thus, the computed model of the BLDC motor can be represented by the following formula in the frame matrix:

$$\begin{aligned} \begin{bmatrix} \mathrm{L_a} & \mathrm{M_{ab}} & \mathrm{M_{ac}} \\ \mathrm{M_{ba}} & \mathrm{L_b} & \mathrm{M_{bc}} \\ \mathrm{M_{ca}} & \mathrm{M_{cb}} & \mathrm{L_c} \end{bmatrix} \frac{d}{dt} \begin{bmatrix} \mathrm{i_a} \\ \mathrm{i_b} \\ \mathrm{i_c} \end{bmatrix} &= \begin{bmatrix} \mathrm{V_a} \\ \mathrm{V_b} \\ \mathrm{V_c} \end{bmatrix} \\ &- \begin{bmatrix} \mathrm{R} & 0 & 0 \\ 0 & \mathrm{R} & 0 \\ 0 & 0 & \mathrm{R} \end{bmatrix} \begin{bmatrix} \mathrm{i_a} \\ \mathrm{i_b} \\ \mathrm{i_c} \end{bmatrix} - \begin{bmatrix} \mathrm{e_a} \\ \mathrm{e_b} \\ \mathrm{e_c} \end{bmatrix}. \end{aligned} \tag{5}$$

If it is considered that the rotor has a mounting surface design, which is frequently the case in the operation of the BLDC motor, which is based on the fact that the stator inductance is independent of the rotor position, it follows that the rotor position is independent of the stator

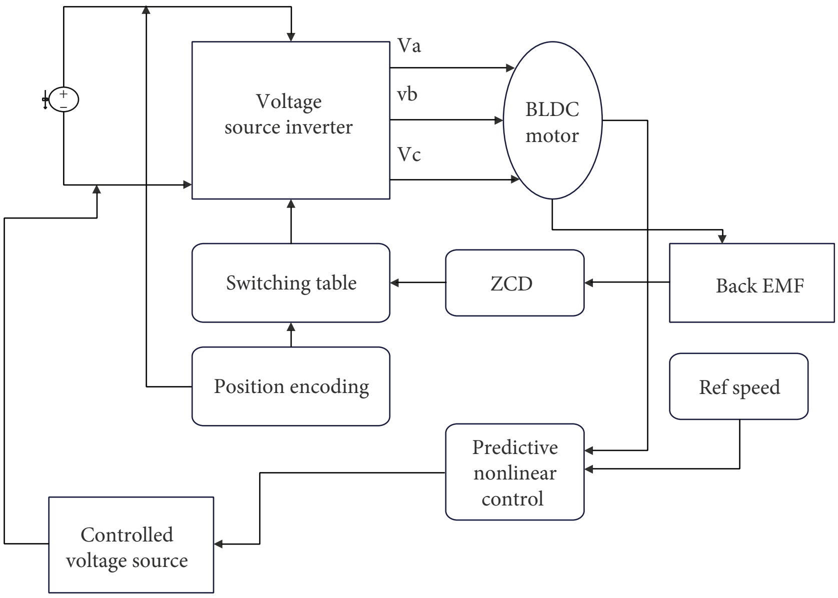


Figure 1: Proposed block diagram for the BLDC speed control.

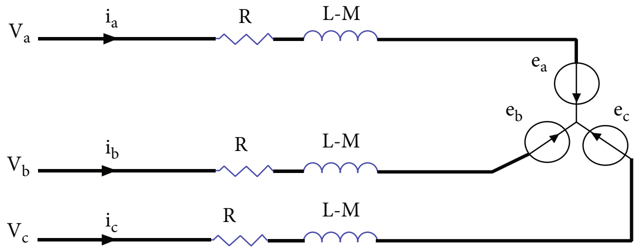


Figure 2: Three-phase BLDC motor equivalent circuit.

inductance:

$$L_a = L_b = L_c = L. \tag{6}$$

The mutual inductance formula is represented in the following form:

$$M_{ab} = M_{ac} = M_{ba} = M_{ca} = M_{cb} = M. \tag{7}$$

Suppose the three-phase system and the resistance of all phases are equal:

$$R_a = R_b = R_c = R. \tag{8}$$

Rearrange equation (5):

$$\begin{bmatrix} L & M & M \\ M & L & M \\ M & M & L \end{bmatrix} \frac{d}{dt} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} = \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} - \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} - \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix}. \tag{9}$$

The electromechanical torque is expressed as

$$T_{\text{em}} = J\frac{d_{\omega r}}{dt} + B_{\omega r} + T_L. \tag{10}$$

The electromagnetic torque of a three-phase BLDC motor, on the other hand, is dictated by the current speed and the waveform of the opposing electromagnetic force, respectively. In this way, they are instantly subjected to electromagnetic torque:

$$T_{\text{em}} = \frac{1}{\omega_m}\left(e_a i_a + e_b i_b + e_c i_c\right). \tag{11}$$

The phase voltages are represented by the symbols V, I, and e, while the phase current and back EMF power are represented by the symbols b and c, respectively. The resistance R and the inductance L represent the phases, and the values of the electrical torque and the load torque of T_e and T_L represent the corresponding values of the electrical torque and the load torque. J denotes the moment of inertia; k_f denotes continuous friction; and _*m* denotes the rotational speed. Back EMF and electrical torque are represented by the following equations:

$$e_a = \frac{K_e}{2}\omega_m F(\theta_e), \tag{12}$$

$$e_b = \frac{K_e}{2}\omega_m F(\theta_e - \frac{4\pi}{3}\Big), \tag{13}$$

$$T_e = \frac{K_t}{2}\left[F(\theta_e) + F\left[\theta_e - \frac{2\pi}{3}\right] + F\left[\theta_e - \frac{4\pi}{3}\right]\right]. \tag{14}$$

The counterelectromotive force constant and torque constant are denoted by the letters K_e and K_t, respectively.

It is common practice to change the machine model in order to simplify a rotating frame of reference and increase computing performance [22].

3.1.1. Brushless DC Motor Operation Method. Three-phase BLDC motors operate in a two-phase configuration. In two-phase energization, the third stage is closed in order to provide the greatest amount of torque. The rotor location has an impact on the energization of the first and second stages [23]. Each time 600 milliseconds passes, the data from the position sensor outputs a three-digit number (H1, H2, and H3), which changes (electric degrees).

The rotor is 1200 away from the stator field lines at the start of each interval and 600 away from the stator field lines at the conclusion of each interval. A simple representation of the magnetic field lines is that they are perpendicular to the highest torque. The current transfer is accomplished via a six-step inverter, as shown in Figure 3. The switching interval of the six-step inverter is shown in the following Table 1.

The switching sequence, the current direction, and the position sensor signal are all shown in the first row of Table 1. The proposed predictive nonlinear optimal control (PNOC) controller would adjust the pulse width modulation (PWM) based on the switching sequence, whose operation is explained below.

3.2. Proposed Predictive Nonlinear Optimal Control (PNOC) Algorithm-Based Motor Control. It is illustrated how a block diagram (PNOC) controller in combination with a speed controller may be used to demonstrate how the suggested predictive nonlinear optimum [24] control can be implemented. The speed of the BLDC motor in Plant *G* is regulated in the manner shown in Figure 4. There are two inputs in the plant that are used to take into account external influences. In addition to the two output terminals, which are designated by the letters *u* and *W*, this device has one for rated speed (*y*) and another for strong output (*Z*). It has three control signals: one control signal, one control signal, and one control signal, in addition to *W*.

In order to manage the speed of the motor as a result of the comparison between the anticipated and reference speeds, the error *e* output *u* is provided to the controller PNOC, which generates pulse width modulation (PWM) and controls the speed of the motor as a result of the comparison.

The goal is to find *K*, which satisfies the expression of the feedback system:

$$||T_{ZW}||\infty \triangleq \begin{bmatrix} W_1 S \\ W_1 R \\ W_1 T \end{bmatrix} \leq \gamma. \tag{15}$$

This type of mixture has the following properties: the sensitivity *S* is a sensitivity function in the transfer matrix from zero to one, *R*; *u* is a sensitivity function in the transfer matrix from one watt to the transfer matrix input [24] sensitivity function, and the complementary sensitivity *T* is a function, and the sensitivity *T* is a function. The sensitivity is necessary in order to get a low-frequency signal from the region of the outer immunity, and it is given by expression (16). The sensitivity must be kept as little as possible in order to obtain the signal, $\omega \longrightarrow 0$.

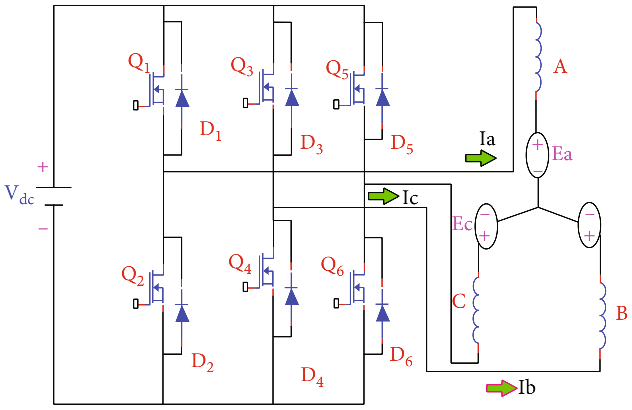


Figure 3: Simplified BLDC driver circuit.

Table 1: Switching sequence of three-phase inverter.

Switching	Input sequence	Sensor position	Switching level	Current level
0-60	0	1	W_1-W_6	Positive
60-120	1	1	W_1-W_2	Negative
120-180	2	0	W_3-W_2	Off
180-240	3	0	W_3-W_4	Positive
240-300	4	0	W_5-W_4	Negative
300-360	5	1	W_5-W_6	Off

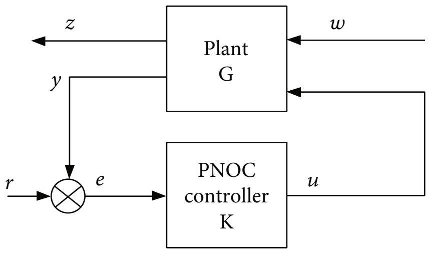


Figure 4: Block diagram for the predictive-nonlinear optimal control (PNOC) controller.

$$s = \frac{1}{(1 + Gk)}. \tag{16}$$

As a result of the existence of the high-frequency zone, the compensation of the sensitivity function $T = 1 - S$ should be limited to a bare minimum in order to minimise modelling error in the high-frequency zone.

After the PNOC ring has been closed, the instability pole of the chosen band is adjusted in accordance with the new value of the instability pole. The technique illustrates how to build a frequency domain design cycle with pinpoint accuracy by including the appropriate weight into the design cycle throughout the generation process. If the plant is frequency-dependent and has an extended weight, it is feasible to produce the signal with the assistance of the MATLAB script "Hinfsin" composite controller. Once this is accomplished, the signal may be fine-tuned to achieve the appropriate levels of performance and resilience. Following the selection of the requisite weights, the frequency domain loop design may be finished, as seen in the following example. Depending on the state location, the acceptable waiting function for PNOC issue and error signal E, control signal U, and output signal Y is raised by W_1 through W_3 in order to achieve an even better waiting function. Because W_1 through W_3 are raised by W_1 through W_3, the waiting function becomes much better. The functional block design for the predictive optimum control strategy is seen in Figure 5.

From Figure 5, the below equation is defined:

$$\begin{bmatrix} W_1 S \\ W_2 R \\ W_3 T \\ e \end{bmatrix} = \begin{bmatrix} W_1 & -W_1 G \\ 0 & W_2 \\ 0 & W_3 G \\ 1 & -G \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix}. \quad (17)$$

The computational model of improved plant P is attained as

$$P = \begin{bmatrix} W_1 & -W_1 G \\ 0 & W_2 \\ 0 & W_3 G \\ 1 & -G \end{bmatrix}. \quad (18)$$

The MATLAB function is used to generate advanced plant B as the augw equation.

$$P = \mathrm{augw}(G, W_1, W_2, W_3). \quad (19)$$

The MATLAB function "hinfsyn" PNOC controller K to equation (20) provided to P is an advanced plant which is used to integrate the team, and the related transfer function KT is obtained (21).

$$K = \mathrm{hinfsyn}(P), \quad (20)$$

$$\mathrm{KT} = \mathrm{tf}\,(K). \quad (21)$$

In order to develop a successful control design, it is critical to choose a suitable weight design. This is a vital consideration that should not be disregarded. The following criteria must be satisfied in order to choose components that are compatible with the standards of the different frequency ranges.

The main objective is to achieve a closed loop of tracking error with the least amount of interference in the low-frequency band, which is a difficult feat to do.

In order to operate safely in the middle frequency band, it is necessary to maintain a constant and acceptable margin of safety at all times.

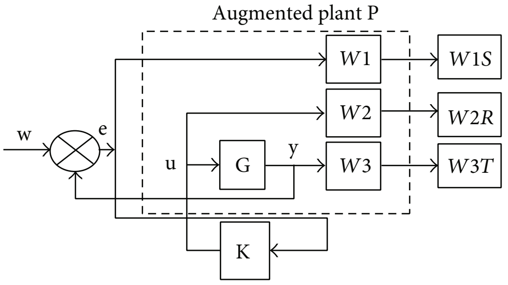


Figure 5: Functional block diagram for the predictive nonlinear optimal control (PNOC) controller.

The control signal must be kept within a certain range in order to function properly at higher frequency ranges.

In order to choose the most appropriate reaction time, it is first required to select the most appropriate sensitivity weighting function, which should be developed based on specific aspects of the desired response time. After that, the reaction time may be selected. Low-pass filters with high- and low-frequency gains are used to govern the exchange of overshoot by balancing the required fixed-level error as well as the high- and low-frequency gains.

In order to represent W_1 in equation (22) as a result of this decision, a simple low-pass filter is used to represent W_1 in the following equation:

$$W_1 = \frac{1}{M_s} \frac{\tau P^s + 1}{M_s \tau P^s + A/M_s}, \quad (22)$$

in such case, M_s.

Maximum sensitivity function A is used. The maximum permitted deviation as well as the standard state b system bandwidth is specified. In the selection of additional weights, the preservation [25] of the lower row controller is taken into consideration, since many weights may become typical selections. Furthermore, the constant is given to W_1 and W_2, both of which are constrained to keeping the control signal in their respective states.

The six coefficient parameters a, b, c, and d of W_1 and g and h of W_3 are carefully chosen in order to achieve the right weights for the corresponding coefficients. Some "thumb rules" for adjusting weight have been submitted for consideration. The use of these factor weights in this study results in significant control improvement for PNOC.

$$W_1 = \frac{a(s + b)}{\mathrm{CS} + d}, \quad (23)$$

$$W_2 = g, \quad (24)$$

$$W_3 = h. \quad (25)$$

3.2.1. PNOC Algorithm Steps. The rapid action of particle motion in PNOC allows for speedier fusion, which leads in the production of a higher-quality solution more quickly. A particle-initiated group is a term used to refer to a group of groups known as random persons in the PNOC setting.

As part of Step 1, the following parameters of the original problem are taken into consideration: (1)
In this section, the estimated number of particles (I a, I b, and I c) as well as the rotor speed (w r) is shown.
(ii) The total number of evaluations has been calculated.
(IV) Learning characteristics (C1, C2). (iv) Boundaries of the search space (iv).
Step 2: in the search space, start producing random particle levels to see where they go.
In Step 3, you will execute the predictive nonlinear optimal control (PNOC) controller assembly based on the weights determined from the particle levels you computed.
In Step 4, PNOC controller models are simulated and synthesised in order to estimate the fitness value of the controller.
Step 5: calculate the optimal global solution as well as the optimal local solution for each particle and for the whole system in order to maximise efficiency.
Step 6: error signals are framed by employing the variables I d and I q, as well as their associated reference esteems. When applying rotor charging flux, the I d reference is utilized to govern the amount of flux that is applied. The rotor output is regulated by the I q reference, which is used in conjunction with the rotor. The error signals indicate the contributions made by the controller to the computations made by the PNOC controller. In the output of the controller, there is a voltage vector that will be delivered to the motor via the controller, denoted by the letters v d and q.
When you get to Step 7, you should run the following procedures in the main loop.
Weight-to-weight ratio of each particle increases as a result of the first point.
Based on the particle velocity data acquired, modify the placement of the rotor as required.
Repetition of Steps 3 through 6 as many times as required until you have achieved the maximum number of iterations is finished in Step 8.
(9) The PNOC infinite controller has the optimal weight according to the best global solution to the entire issues1 (see Figure 1) (see Step 9).

Algorithm 1: Steps.

When it comes to optimization solutions, the search location is described as the collection of all viable optimization [26] solutions to the problem. Define the search area's perimeter. The maximum and minimum values of the W_1, W_2, and W_3 coefficients have been specified in advance of the experiment. Following the setup of the population members, the production of random locations and velocities for each particle is performed. Individuals' finest performance memories are used in determining the value of the target function. The best neighbours on a personal level are the most valuable when compared to the best neighbours on a global level, which is the highest value.

When particles move, their influence is felt more quickly in PNOC, allowing for faster integration and, ultimately, a more effective solution. A random human is referred to as a particle launched in the PNOC universe. In search engine optimization, search sites are defined as all of the potential solutions to optimization problems that have been discovered [27]. It is necessary to indicate the lowest and highest values of the W_1, W_2, and W_3 coefficients in the search area range in order to define the search area range. Once the members of the population have been identified, a random location and velocity for each particle are generated. The objective function value is determined using the individual's best performance memory as a starting point for the calculation. In comparison to one's own personal best results as well as the best value of one's neighbours, it is regarded as being the greatest wherever in the globe. Its personal best and the greatest particle motion in the world are on their way to achieving a new speed and current momentum and knowledge of the new situation, as well as his own personal best and the greatest particle motion in the world.

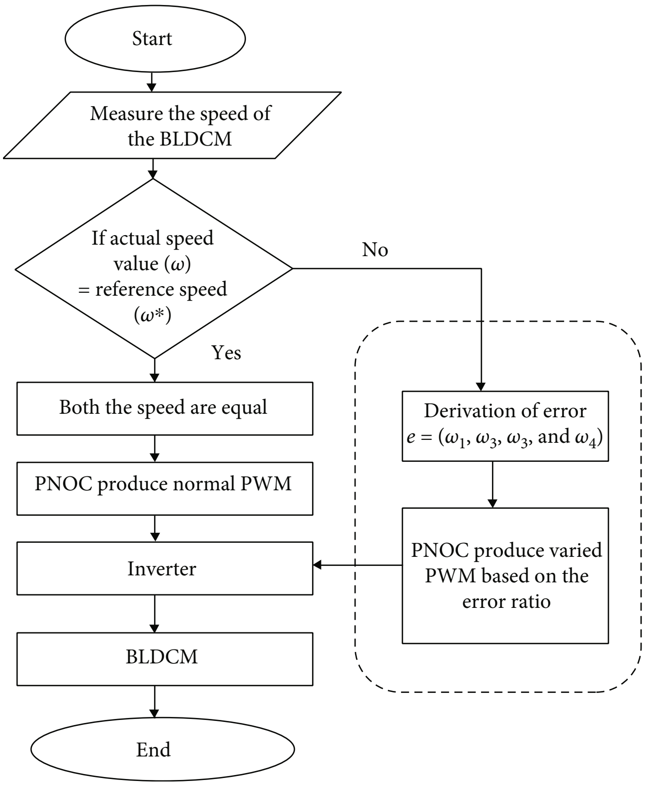

Start
Measure the speed of the BLDCM
If actual speed value (ω) = reference speed (ω*)
No
Yes
Both the speed are equal
PNOC produce normal PWM
Inverter
BLDCM
End
Derivation of error $e = (\omega_1, \omega_3, \omega_3, \text{and } \omega_4)$
PNOC produce varied PWM based on the error ratio

Figure 6: The flow chart for the PNOC-based BLDCM.

In order to get the best coefficient, the optimization aim is to identify the optimal global cost, which is equal to the absolute value for error reduction, by using the PNOC weighting function and obtaining the best coefficient. In order to get the controller's transfer function, it is

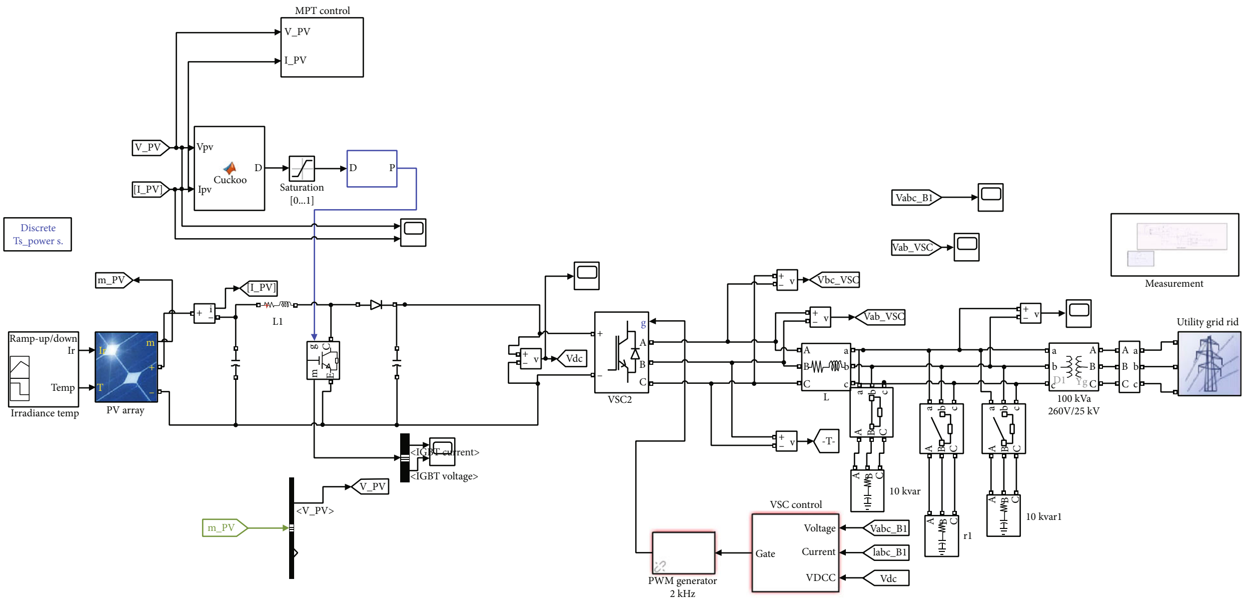


Figure 7: Final proposed system Simulink model.

important to generate the best prices in the world, namely, W_1, W_2, and W_3, using the best available data. In this application, the controller is used as a speed controller, with the output of the controller providing a reference value for torque.

Figure 6 represents the flowchart of the proposed work. The control operations of the PNOC controller for the BLDCM speed control operation are shown in Figure 6. The set speed and reference speed are determined by the PNOC controller in this control module, which is located on the main board. The PWM of the inverter will be modified in response to the value.

4. Result and Discussion

For the proposed BLDC driver system, which is being developed in MATLAB2017b software, the system function will be simulated and the system will be coordinated in order to test the operation of the proposed model's performance operation. Figure 7 depicts the suggested BLDCM model's implementation in MATLAB/Simulink, as well as the system's operation and the ability to show library SIM power supply system-wide layouts.

Figure 7 depicts the suggested predictive nonlinear optimal regulation (PNOC) strategy-based speed control of the BLDC motor, which is based on PNOC. The performance of the BLDC motor is evaluated using a variety of torque-speed combinations, the results of which are shown below.

On the basis of load torque variation, Figure 8 depicts the current waveform for a BLDC motor. As can be seen in this figure, the initial current variation is significant during time period (0-2), as can be seen in the previous figure.

The torque of a BLDC motor is seen in Figure 9. The load torque is applied to the suggested model for the differential speed analysis at a time length of 0.2 seconds in order to conduct the differential speed analysis. The speed and current of the stator are adjusted in response to the torque value.

The suggested predictive nonlinear optimal control (PNOC) controller would automatically modify the pulse width modulation (PWM) and maintain the target speed of the BLDC motor based on the changing of the load torque. Figure 10 shows how the load torque fluctuation is absorbed; the suggested PNOC would alter the PWM displayed.

Figure 11 shows the results of the BLDC motor speed stability study performed using the suggested motor. Using the predictive [28] nonlinear optimal control (PNOC) controller has been suggested. The driving cycle speed is set at 1500 rpm with a torque of 1 N-m at the moment (0-1), and the suggested PNOC controller will give suitable feedback to the inverter in order to maintain the desired speed, as seen in the preceding image [29].

Figure 11 depicts a comparative examination of many metrics, including steady-state error, recovery time, and peak overshoot time, in order to evaluate the performance of the BLDCM algorithm [30]. The results of the simulation are summarised in Table 2, and the findings reveal that the PNOC-based controller outperforms the current Proportional Integral Derivative (PID) controller when compared to the existing PID controller.

In Table 2 and in Figure 12, the control system parameters for varying various load torque situations are shown for each condition. The system will be evaluated in comparison to the proposed PNOC controller.

Following thorough examination of both practical and empirical guidelines, the following test was selected: The battery had been completely charged before to the commencement of the experiment and had not been recharged in

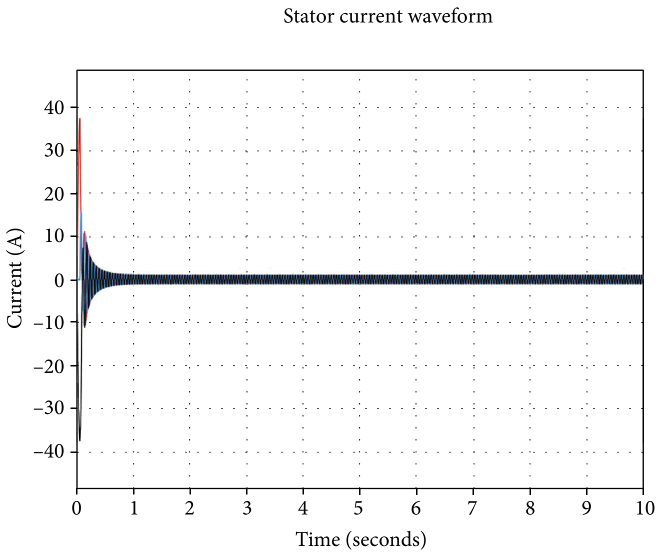


Figure 8: Output current of the proposed system.

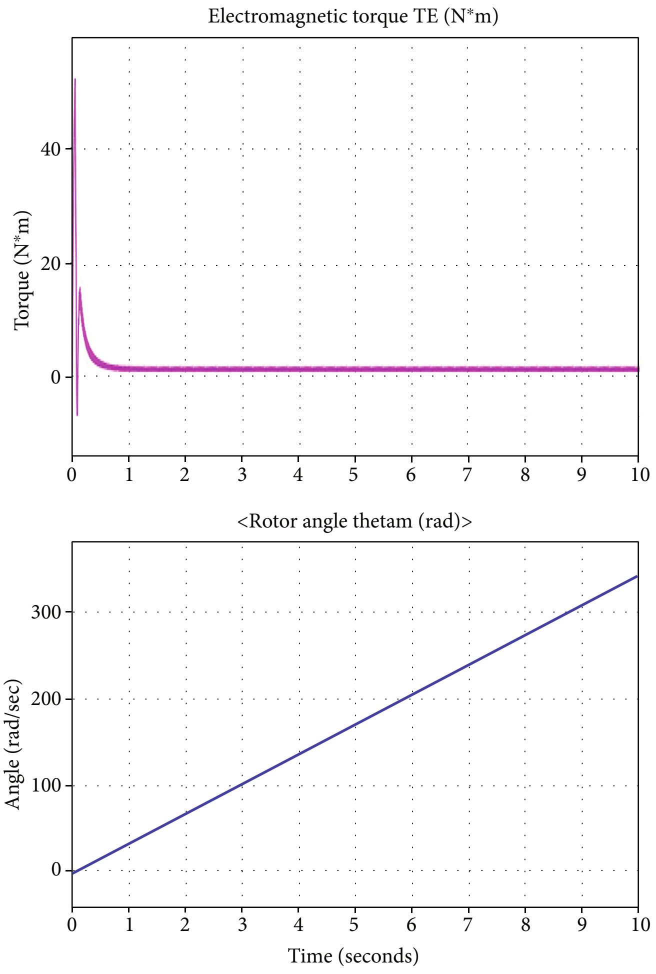


Figure 9: BLDC motor torque.

between the speed variations [30]. It was possible to measure the speed of the propeller using a laser speed calculator, which ranged from a minimum of 1400 rpm to a maximum of 10000 rpm. The servo tester was impossible [31] to be set on fixed speed values between 1400 and 3000 rpm since its discretion could not be set manually at low speed values

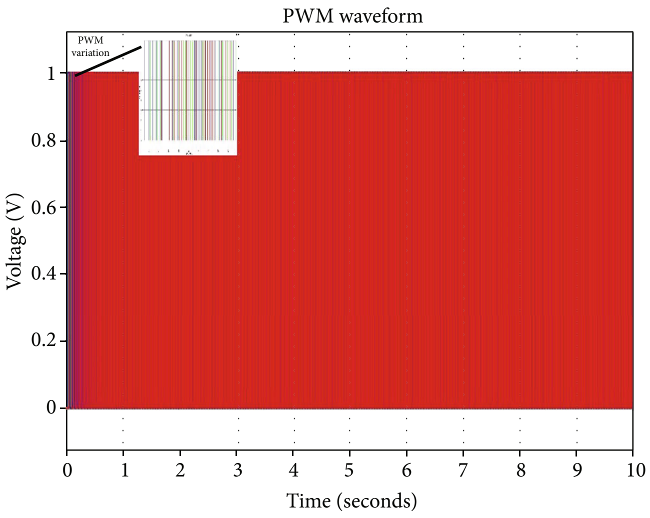


Figure 10: PWM waveform of the proposed controller.

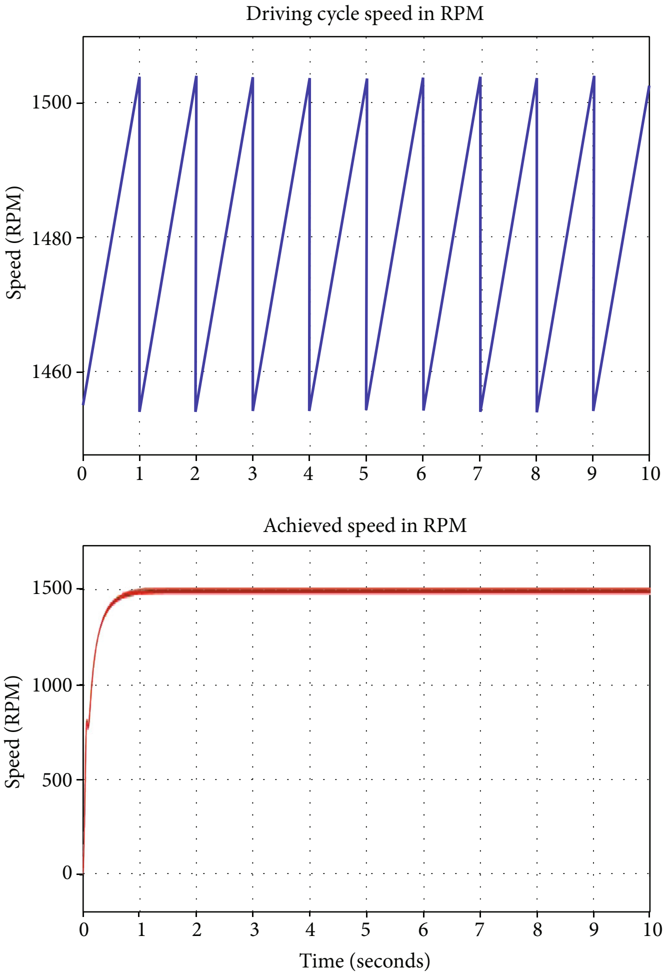


Figure 11: BLDC motor speed.

between these numbers. After 3000 rpm, a step of 1000 rpm every moment was applied to achieve the desired speed. The power consumption was recorded in amperes and grammes for each speed measurement, as well as the thrust in grammes for each speed moment. At the end, the cost of speed/watt per total € spent was computed in order to

Table 2: The performance analysis of the BLDC motor.

Parameters	PID	PNOC
Reference value (rpm)	1500	1500
Peak overshoot time (sec)	0.9564	0.5663
Recovery overshoot (%)	0.845	0.423
Recovery time (sec)	0.88	0.41
Steady state error value (rpm)	6.6	4.2
Steady-state error (%)	0.66	0.36

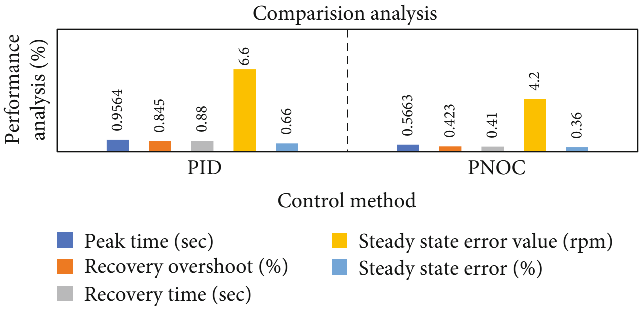


Figure 12: Comparison analysis.

provide an exact result of a cost and consumption efficient arrangement.

5. Conclusion

Using a predictive nonlinear optimal control (PNOC) controller, this paper presents the modelling and simulation of the speed control of BLDCM. Using MATLAB, we have validated the performance [32] of all of the controller settings, and we have produced a thorough analysis of the results. BLDC speed control working conditions are modified with the controller, as shown by the results obtained with this specific model from the simulation. The resulting steady-state value is 5.8 percent, which is much better than the traditional controllers. Based on the summary of data, it can be concluded that the proposed PNOC controllers have the distinct characteristics of error elimination and total reduction of steady-state error. The constant speed operation assures that the system will continue to operate indefinitely [33]. Achieving the highest steady-state value (rpm) of 4.2 by using sophisticated predictive nonlinear optimal control (PNOC) yields successful outcomes when used in the automobile industry.

Because of the benefits of sensorless control for BLDC motor drives, which have gained popularity in recent years, a sensorless controlled BLDC motor drive was selected for this research to make use of those advantages. According to the findings of the literature review, the back EMF zero crossing detection approach is excellent and is extensively utilized in the motor drive sector. Due to the fact that the work is limited to the investigation of artificial intelligence controller strategies for BLDC motors, the sensorless control strategy based on the reverse EMF zero crossing was chosen for the investigation.

We have improved upon the simple simulation model that we built for the BLDC motor drive to achieve sensorless speed control by using zero crossing of the back EMF. The BLDC motor drive is comprised of a PI speed controller and a current controller for controlling the motor speed. The sensorless controlled BLDC motor was subjected to a thorough investigation under a variety of operating situations, including beginning, load perturbations, and variations in the specified speed. In order to verify the reduced simulation model, a hardware setup was used. For the purpose of demonstrating the efficiency of the simplified modelling, the responses were compared to waveforms published in the literature as well as waveforms produced utilizing hall sensors. The findings of the simulation and the hardware were almost identical. As a result, the sensorless model was employed in the subsequent investigation of the suggested artificial intelligence controller to increase the dynamic performance of the BLDC motor drive system.

Algorithm 1 represents the flow of optimal function of the proposed algorithm.

Appendix

```
function [x_pop, fx_val]=PI_objfun_E(x_pop, options)
global sys_controlled
global time
global sysrl
%Splitting the chromosomes into two separate strings
Kp=x_pop(1);
Ki=x_pop(2);
%Creating the PI controller from the present values
pi_den=[1 0];
pi_num=[ Kp Ki];
pi_sys=tf(pi_num, pi_den);
```

References

[1] G. Li, T. Zhang, B. Li, T. Fu, and P. Duan, “Commutation compensation strategy for brushless DC motor based on terminal voltage reconstruction,” *Journal of Electrical Engineering and Technology*, vol. 16, no. 4, pp. 2031–2043, 2021.

[2] S. Subramanian, R. M. Mohan, S. K. Shanmugam, N. Bačanin, M. Zivkovic, and I. Strumberger, “Speed control and quantum vibration reduction of brushless DC motor using FPGA based

dynamic power containment technique," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2021.

[3] P. K. Shukla, M. Zakariah, W. A. Hatamleh, H. Tarazi, and B. Tiwari, "AI-driven novel approach for liver cancer screening and prediction using cascaded fully convolutional neural network," *Journal of Healthcare Engineering*, vol. 2022, Article ID 4277436, 14 pages, 2022.

[4] A. Kerem, "Design, implementation and speed estimation of three-phase 2 kW out-runner permanent magnet BLDC motor for ultralight electric vehicles," *Electrical Engineering*, vol. 103, no. 5, pp. 2547–2559, 2021.

[5] B. N. Kommula and V. R. Kota, "An integrated converter topology for torque ripple minimization in BLDC motor using an ITSA technique," *Journal of Ambient Intelligence and Humanized Computing*, vol. 49, pp. 1–20, 2021.

[6] A. Sampathkumar, M. Tesfayohani, S. K. Shandilya et al., ""Internet of Medical Things (IoMT) and Reflective Belief Design-Based Big Data Analytics with Convolution Neural Network-Metaheuristic Optimization Procedure (CNN-MOP)," *Computational Intelligence and Neuroscience*, vol. 2022, no. Article ID 2898061, p. 14, 2022.

[7] H. S. Sridhar, P. Hemanth, H. V. Soumya, and B. G. Joshi, "Speed control of BLDC motor using soft computing technique," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1162–1168, Trichy, India, Sept.2020.

[8] A. T. Hafez, A. A. Sarhan, and S. Givigi, "Brushless DC motor speed control based on advanced sliding mode control (SMC) techniques," *IEEE International Systems Conference (SysCon)*, 2019, pp. 1–6, Orlando, FL, USA, April 2019.

[9] V. K. Trivedi, P. Kumar Shukla, and A. Pandey, "Hue based plant leaves disease detection and classification using machine learning approach," in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 549–554, Bhopal, India, June 2021.

[10] S. Joshi, S. Stalin, P. K. Shukla, R. Bhatt, R. S. Bhadoria, and B. Tiwari, "Unified Authentication and Access Control for Future Mobile Communication-Based Lightweight IoT Systems Using Blockchain," *Wireless Communications and Mobile Computing*, , no. Article ID 8621230, p. 12.

[11] A. Varshney, D. Gupta, and B. Dwivedi, "Speed response of brushless DC motor using fuzzy PID controller under varying load condition," *Journal of Electrical Systems and Information Technology*, vol. 4, no. 2, pp. 310–321, 2017.

[12] A. Khare, R. Gupta, and P. K. Shukla, "Improving the protection of wireless sensor network using a black hole optimization algorithm (BHOA) on best feasible node capture attack," in *IoT and Analytics for Sensor Networks. Lecture Notes in Networks and Systems*, vol. 244, Springer, Singapore, 2022.

[13] V. Verma, N. S. Pal, and B. Kumar, "Speed control of the sensorless BLDC motor drive through different controllers," in *Harmony Search and Nature Inspired Optimization Algorithms. Advances in Intelligent Systems and Computing*, N. Yadav, A. Yadav, J. Bansal, K. Deep, and J. Kim, Eds., vol. 741, Springer, Singapore, 2019.

[14] W. Huazhang, "Design and implementation of brushless DC motor drive and control system," *Procedia Engineering*, vol. 29, pp. 2219–2224, 2012.

[15] R. Bhatt, P. Maheshwary, and P. Shukla, "Application of fruit fly optimization algorithm for single-path routing in wireless sensor network for node capture attack," in *Computing and Network Sustainability. Lecture Notes in Networks and Systems*, vol. 75, Springer, Singapore, 2019.

[16] D. Potnuru, K. Alice Mary, and C. Sai Babu, "Experimental implementation of flower pollination algorithm for speed controller of a BLDC motor," *Ain Shams Engineering Journal.*, vol. 10, no. 2, pp. 287–295, 2019.

[17] J. A. Prakosa, D. V. Samokhvalov, G. R. V. Ponce, and F. S. Al-Mahturi, "Speed control of brushless DC motor for QuadCopter drone ground test," in *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, vol. 2019, pp. 644–648, Saint Petersburg and Moscow, Russia, Jan 2019.

[18] N. Hussain, P. Maheshwary, P. K. Shukla, and A. Singh, "Detection of black hole attack in GPCR VANET on road network," in *International Conference on Advanced Computing Networking and Informatics. Advances in Intelligent Systems and Computing*, vol. 870, Springer, Singapore, 2019.

[19] A. Mamadapur and G. Unde Mahadev, "Speed control of BLDC motor using neural network controller and PID controller," in *2019 2nd International Conference on Power and Embedded Drive Control (ICPEDC)*, pp. 146–151, Chennai, India, Aug 2019.

[20] V. R. Walekar and S. V. Murkute, ""Speed control of BLDC motor using PI & fuzzy approach: a comparative study," 2018 International Conference on Information," in *2018 International Conference on Information, Communication, Engineering and Technology (ICICET)*, pp. 1–4, Pune, India, Aug 2018.

[21] P. K. Shukla, P. K. Shukla, M. Bhatele et al., "A novel machine learning model to predict the staying time of international migrants," *International Journal on Artificial Intelligence Tools*, vol. 30, no. 2, article 2150002, 2021.

[22] H. Hu, T. Wang, S. Zhao, and C. Wang, "Speed control of brushless direct current motor using a genetic algorithm–optimized fuzzy proportional integral differential controller," *Advances in Mechanical Engineering*, vol. 11, no. 11, Article ID 1687814019890199, 2019.

[23] D. Potnuru and S. Ch, "Design and implementation methodology for rapid control prototyping of closed loop speed control for BLDC motor," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 1, pp. 99–111, 2018.

[24] S. Stalin, P. Maheshwary, and P. Kumar Shukla, "Payback of image encryption techniques: a quantitative investigation," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1370–1380, Madurai, India, May 2019.

[25] S. Gobinath and M. Madheswaran, "Deep perceptron neural network with fuzzy PID controller for speed control and stability analysis of BLDC motor," *Soft Computing*, vol. 24, no. 13, pp. 10161–10180, 2020.

[26] P. Suganthi, S. Nagapavithra, and S. Umamaheswari, "Modeling and simulation of closed-loop speed control for BLDC motor," in *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pp. 229–233, Mallasamudram, India, March 2017.

[27] M. K. Ahirwar, P. K. Shukla, and R. Singhai, "CBO-IE: a data mining approach for healthcare IoT dataset using chaotic biogeography-based optimization and information entropy," *Scientific Programming*, vol. 2021, Article ID 8715668, 14 pages, 2021.

[28] K. S. Devi, R. Dhanasekaran, and S. Muthulakshmi, "Improvement of speed control performance in BLDC motor using

fuzzy PID controller," in *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pp. 380–384, Ramanathapuram, India, May 2016.

[29] E. Gowthaman, V. Vinodhini, M. Y. Hussain, S. K. Dhinakaran, and T. Sabarinathan, "Speed control of permanent magnet brushless DC motor using hybrid fuzzy proportional plus integral plus derivative controller," *Energy Procedia*, vol. 117, pp. 1101–1108, 2017.

[30] A. Bhattacharjee, G. Ghosh, V. Kumar Tayal, and P. Choudekar, "Speed control of BLDC motor through mobile application via secured Bluetooth," in *2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE)*, pp. 203–206, Noida, India, Oct 2017.

[31] R. Janarthanan, R. U. Maheshwari, P. K. Shukla, P. K. Shukla, S. Mirjalili, and M. Kumar, "Intelligent detection of the PV faults based on artificial neural network and type 2 fuzzy systems," *Energies*, vol. 14, no. 20, article 6584, 2021.

[32] T. Ebin Joseph, M. V. Sreethumol, and A. Dinesh Pai, "Speed control of BLDC motor drive under DTC scheme using OC with modified integrator," in *International Conference on Technological Advancements in Power and Energy (TAP Energy)*, pp. 79–84, Kollam, India, June 2015.

[33] D. Kumpanya and S. Tunyasrirut, "DSP-based speed control of brushless DC motor," in *In Asian Simulation Conference*, vol. 474, Springer, Berlin, Heidelberg, 2014.

Fault Diagnosis Method of Distribution Equipment Based on Hybrid Model of Robot and Deep Learning

Nabnit Panigrahi, *Department of Electrical and Electronics Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, nabnitpanigrahi@gmail.com*

Anil Sahoo, *Department of Electrical Engineering , Capital Engineering College, Bhubaneswar, anil_sahoo342@gmail.com*

Shashi Bhusan Mohanty, *Department of Electrical Engineering , Raajdhani Engineering College, Bhubaneswar, s.b.mohanty@gmail.com*

Smruti Ranjan Nayak, *Department of Electrical Engineering , NM Institute of Engineering & Technology, Bhubaneswar, smrutinayak@live.com*

Abstract

In view of the poor effect of most fault diagnosis methods on the intelligent recognition of equipment images, a fault diagnosis method of distribution equipment based on the hybrid model of robot and deep learning is proposed to reduce the dependence on manpower and realize efficient intelligent diagnosis. Firstly, the robot is used to collect the on-site state images of distribution equipment to build the image information database of distribution equipment. At the same time, the robot background is used as the comprehensive database data analysis platform to optimize the sample quality of the database. Then, the massive infrared images are segmented based on chroma saturation brightness space to distinguish the defective equipment images, and the defective equipment areas are extracted from the images by OTSU method. Finally, the residual network is used to improve the region-based fully convolutional networks (R-FCN) algorithm, and the improved R-FCN algorithm trained by the online hard example mining method is used for fault feature learning. The fault type, grade, and location of distribution equipment are obtained through fault criterion analysis. The experimental analysis of the proposed method based on PyTorch platform shows that the fault diagnosis time and accuracy are about 5.5 s and 92.06%, respectively, which are better than other comparison methods and provide a certain theoretical basis for the automatic diagnosis of power grid equipment.

1. Introduction

With the continuous improvement of social economy and people's living standards, the power demand is increasing day by day, and the scale of power system is growing day by day. It includes transmission and transformation networks of various voltage levels. Therefore, ensuring the safe and stable operation of complex power grid is an inevitable requirement to ensure the economic and social development. At the same time, the economic and social development has a great impact on the security and economy, and higher reliability is required [1, 2]. As the terminal of the whole power grid operation, distribution network is the part with the widest coverage and the largest scale in China's power system, and it is the key link to ensure that power can be "allocated and used" [3]. The distribution equipment will have some faults under long-term operation, resulting in abnormal temperature. Therefore, by detecting the temperature of the distribution equipment, the thermal fault diagnosis of the distribution equipment can be carried out quickly, which plays a great role in the safe operation of the power grid.

The infrared image of equipment is used for fault diagnosis with high efficiency, accurate judgment, safety, and reliability. At the same time, it is free from electromagnetic interference, fast detection speed, and no power failure of live equipment. Therefore, infrared diagnosis is widely used in the field of equipment fault monitoring and diagnosis technology [4]. However, due to the characteristics of large quantity and complex types of distribution equipment, if we only rely on manual work in the process of data acquisition, analysis, and processing, the workload is relatively large, the efficiency is low, and the accuracy is relatively low due to the high dependence on manual experience. Therefore, automatic image acquisition and analysis of distribution equipment are of great significance to ensure the safety and stability of distribution network [5].

In recent years, the automatic inspection technology of power distribution room has been popularized, and various automatic robots and UAVs have made great progress in the original data acquisition stage. However, the accurate and efficient processing of collected image data is still in its infancy. How to extract the features of interest from infrared images for power distribution equipment recognition is a problem to be solved [6]. Among them, the deep learning algorithm has made great achievements in image processing, speech recognition, and text analysis. By establishing a deep-seated neural network, high-level features are extracted from low-level features layer by layer, so as to achieve the effect of target classification and recognition [7]. Compared with the manually designed feature extraction method, the distributed features obtained by deep learning network model can better express the essence of data [8, 9]. Therefore, in order to improve the efficiency of thermal fault detection of distribution equipment, improve the intelligence of power grid, reduce the labor cost of detection, and reduce the false detection rate, a fault diagnosis method of distribution equipment based on the hybrid model of robot and deep learning is proposed, which effectively ensures the safe and reliable operation of distribution equipment.

2. Related Research

At present, there are many researches on fault diagnosis of distribution equipment at home and abroad, which can be divided into traditional fault identification and classification methods and machine learning based identification and classification methods [10]. Among them, the traditional fault identification and classification methods mainly include fuzzy clustering, discrete wavelet transform, and chaotic algorithm. For example, [11] proposed an infrared image segmentation algorithm based on intuitionistic fuzzy clustering algorithm based on spatial distribution information, which is suitable for power equipment. It can well suppress the strong interference of nontarget objects in infrared image to image segmentation, but the method is more traditional and has poor segmentation effect for complex intelligent power grid equipment. Reference [12] proposed an anomaly detection method based on spatial clustering applied by auxiliary feature vector and density noise. The auxiliary feature vector of each conditional variable is constructed for clustering to identify normal data patterns and different types of anomalies. Reference [13] proposed a data mining driving scheme based on discrete wavelet transform to realize high impedance fault detection in active distribution network, but the universality of the method is not high. Reference [14] proposed a method to obtain the vibration characteristics of circuit breaker based on time-frequency and chaos analysis to realize circuit breaker fault identification. The image analysis process is complex, resulting in the reduction of fault identification efficiency. Reference [15] proposes a method based on feature model for single-phase grounding fault in active distribution network system, which transforms the solution of nonlinear feature model into single-objective optimization of feature entropy, which can well identify single-phase fault, but the identification effect of equipment with feature type is not ideal.

With the continuous development of computer technology and the rapid development of 5G communication technology, machine learning algorithm has been widely used this year, especially the deep learning algorithm has certain advantages in the field of fault identification and classification. Reference [16] proposes artificial neural network algorithm to identify the insulator state and uses single-layer and multilayer perceptron artificial intelligence algorithm to classify the conditions of distribution insulators. This technology can make the automatic inspection of electrical system more accurate and efficient, but it lacks high reliable database for support. Reference [17] proposed a Mask *R* convolution neural network method and used transfer learning and dynamic learning rate algorithm to realize efficient recognition of annotated image data sets, but it relied too much on graphics annotation and lacked practical application value. In [18], appropriate traveling wave time-frequency characteristic parameters of fault current are selected as the input of adaptive depth belief network model to obtain the fault type, but only considering the fault current characteristics as the basis, the reliability needs to be further improved.

Based on the above analysis, aiming at the problems such as the complexity and diversity of smart grid distribution equipment and the unsatisfactory effect of most existing image recognition methods, a distribution equipment fault diagnosis method based on robot and deep learning hybrid model is proposed. Its innovations are summarized as follows:

(1) In order to obtain the image information of distribution equipment more comprehensively, the proposed method introduces the robot to construct the corresponding image knowledge database, which provides the basis for fault classification and fault location.

(2) In order to locate the equipment defect area in the infrared image of distribution equipment, the proposed method performs threshold segmentation on the infrared image in hue saturation value (HSV) space and uses OTSU method to extract the equipment defect area, so as to improve the accuracy of subsequent fault diagnosis.

(3) Aiming at the problem that the deep learning algorithm is prone to gradient disappearance and gradient explosion, the proposed method uses the residual network to improve the region-based fully convolutional networks (R-FCN) algorithm and applies it to the learning of faulty equipment, so as to

obtain the fault type and location with high accuracy and further improve the safety of equipment.

3. Proposed Method

3.1. Construction of Image Information Database of Distribution Equipment Based on Robot Inspection. The traditional equipment status is usually determined by manual analysis. The workload is huge and error-prone, which affects the judgment of system status, resulting in potential safety hazards. Therefore, the robot is used for patrol inspection to obtain the status image of distribution equipment and build the corresponding information base for the analysis of equipment status, so as to find the faulty equipment in time and ensure the reliable operation of power grid [19]. The construction process of distribution equipment image information base based on robot inspection is shown in Figure 1.

The basic data sources of the database mainly include production system, online monitoring system, and robot background inspection system. The relevant data of the state quantity of power equipment mainly comes from the power production management system (PMS), which can provide the real-time operation condition, historical operation state, historical maintenance record, historical test data, equipment account, equipment parameters, and other information of the equipment. The online monitoring system mainly relies on various sensors on each power equipment for real-time monitoring. The robot background inspection system can not only provide the observation of some state quantities, but also carry out corresponding state evaluation and analysis for different equipment states according to the automatic state evaluation system. In addition, the data composition of the system includes infrared temperature measurement, visible light reading, and telemetry reading.

The robot inspection cycle generally refers to the inspection plan formulated by the distribution network operation inspection center, and two inspection robots complete the tasks of infrared temperature measurement and data transcription of equipment in the area [20]. At the same time, the robot background uses the threshold out of limit judgment method to automatically evaluate the equipment status. In order to ensure sufficient charging time of the robot and avoid the daily patrol and infrared temperature measurement period, the special patrol at night is set in the nonbusy working period of the robot every day, with the upper limit of one time. The data reports collected by the special patrol at night and infrared temperature measurement are included in the database for screening and preprocessing.

In addition, the background of the inspection robot is equipped with a system server, which includes data analysis software terminal, data exchange server, data storage server, data operation server, and other modules. The data exchange server is responsible for collecting and classifying the production system, online monitoring system, and robot patrol data into the storage server. There are three-party databases, fault information base and knowledge information base in the data storage server.

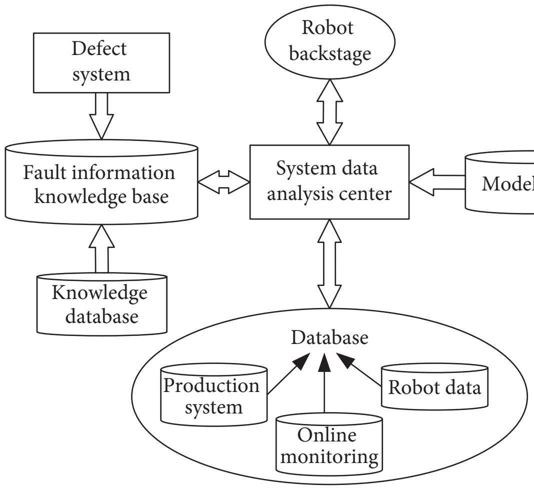


Figure 1: Construction process of image information database for distribution equipment.

(1) The data of the three-party platform includes the information required in the database structure table. After eliminating the redundant information, the integrated data in a unified format can be obtained, and the defect alarm data can be located and retrieved quickly.

(2) The fault information base is mainly taken from the defect system records and contains a large number of relevant equipment fault cases, including fault characteristics, solutions, expert opinions, and manufacturer records. At the same time, the maintenance record database and equipment account database are used to build a comprehensive database of fault information, so as to screen the fault inspection points.

(3) The knowledge information base is the engine for the system to evaluate the equipment status and judge the fault. The internal rules at all levels provide the logical basis for the system to judge the fault. The key is knowledge acquisition, that is, collecting and mining the knowledge at all levels to enrich the knowledge base.

3.2. Defect Feature Extraction of Distribution Equipment. When extracting the defect features of distribution equipment, it is necessary to perform threshold segmentation on the infrared image in HSV space, separate the infrared image background from irrelevant equipment and defective equipment, and then extract the equipment defect area [21].

3.2.1. OTSU Threshold Segmentation. OTSU is considered to be one of the best algorithms in image threshold segmentation. The threshold segmentation process of OTSU algorithm is as follows: firstly, the image is processed in gray level, the number of pixels in the whole image is counted, and the probability distribution of each pixel in the whole image is calculated; then, the gray level is traversed and searched in the whole image, and the interclass probability of the image foreground and background at the current gray level is calculated; finally, the threshold corresponding to the variance between classes and within classes is calculated by the given objective function.

Suppose there are D gray levels in the image, in which the number of pixels with gray value of i is N_i and the total number of pixels in the image is N. Then, the average gray value of the whole image is

$$\mu_{\sum} = \sum_{i=0}^{D-1} i \frac{N_i}{N}. \tag{1}$$

According to the gray characteristics of the image, the image is divided into foreground B_0 and background B_1. Then, $p_0(T)$ and $p_1(T)$ represent the probability of occurrence of foreground B_0 and background B_1 when the threshold is T, respectively. The calculation is as follows:

$$\begin{aligned} p_0(T) &= \sum_{i=0}^{T} \left(\frac{N_i}{N}\right), \\ p_1(T) &= 1 - p_0(T). \end{aligned} \tag{2}$$

Then, the mean values of foreground B_0 and background B_1 are

$$\begin{cases} \mu_0(T) = \dfrac{\sum_{i=0}^{T} i(N_i/N)}{p_0(T)}, \\ \\ \mu_1(T) = \dfrac{\mu_{\sum} - \sum_{i=0}^{T} i(N_i/N)}{p_1(T)}. \end{cases} \tag{3}$$

The interclass variance with threshold T in the gray histogram is calculated as follows:

$$\sigma_B^2(T) = p_0(T)\left[\mu_0(T) - \mu_{\sum}\right]^2 + p_1(T)\left[\mu_1(T) - \mu_{\sum}\right]^2. \tag{4}$$

The optimal threshold is defined as the T value corresponding to the maximum variance between classes, which is calculated as

$$\sigma_B^2(\widehat{T}) = \max_{0\le T\le D-1} \left\{\sigma_B^2(T)\right\}. \tag{5}$$

3.2.2. Defect Region Extraction Based on HSV. In order to improve the accuracy of equipment fault image classification, the defect region and background in the infrared image of fault power equipment are separated by using the defect region segmentation algorithm based on HSV. Since it is impossible to determine the defect type only by analyzing the fault area, it is necessary to segment the defect area based on mathematical morphology according to the location of the defect area. Through this method, the defective power equipment and the background area in the infrared image are separated, so as to reduce the interference of the background area in the infrared image on the defect type classification [22]. The flow of HSV based defect region extraction algorithm is shown in Figure 2.

When processing the infrared image of defective equipment, first merge the similar pixels corresponding to the area with the same temperature, and segment the image according to the threshold of the three components of the defective area in the HSV color space to extract the defective area. Then, the discrete defect regions are connected through the closed operation in mathematical morphology, and the threshold segmentation of the original image is carried out by OTSU method to separate the power equipment and the background region. Finally, the defect area is found in the binary image separated by OTSU method; that is, the defective power equipment is separated from other areas, so as to achieve the purpose of extracting defective power equipment and facilitate the identification and diagnosis of power equipment types and fault types.

3.3. Fault Diagnosis of Distribution Equipment Based on Deep Learning Hybrid Model

3.3.1. Defect Training Based on Deep Learning Hybrid Model. R-FCN algorithm architecture mainly includes backbone network, region proposal network (RPN), and region of interest (ROI) subnet [23]. When fault diagnosis of power distribution equipment is carried out, first input the collected infrared image of power equipment into convolution neural network and extract the convolution feature map of infrared image. In this process, deeper and more abstract image features can be extracted by using a larger backbone network (ResNet 101) to improve the recognition accuracy [24]. Then, the feature map is sent to the RPN network to generate anchors, which are marked with foreground and background, and the foreground area with high score is selected as the recommended area ROIs. These ROIs are sent to the ROI subnet for further training, and 300 recommended windows are generated for each infrared image of power equipment. At the same time, the characteristic map of the full convolution layer is calculated with the multilayer convolution kernel to generate a position sensitive score map. The ROI and Score Maps are input into the later Softmax layer for vote. Through the Softmax layer for classification, the ROI with the highest score is finally obtained, that is, the location and type of the object located and recognized. The architecture of R-FCN algorithm is shown in Figure 3.

(1) Residual Network. When the depth of the deep learning network reaches a certain degree, the problems of gradient disappearance and gradient explosion often appear during training. In order to solve this problem, the residual network (ResNet) is used to improve the R-FCN algorithm; that is, the residual network is selected as the backbone network. The residual element is essentially the mapping residual required for fitting through these stacked layers. Suppose that the network mapping is $H(x)$ and the residual mapping function of the network is $F(x)$, $F(x) = H(x) - x$. The so-called residual is the difference between the observed value

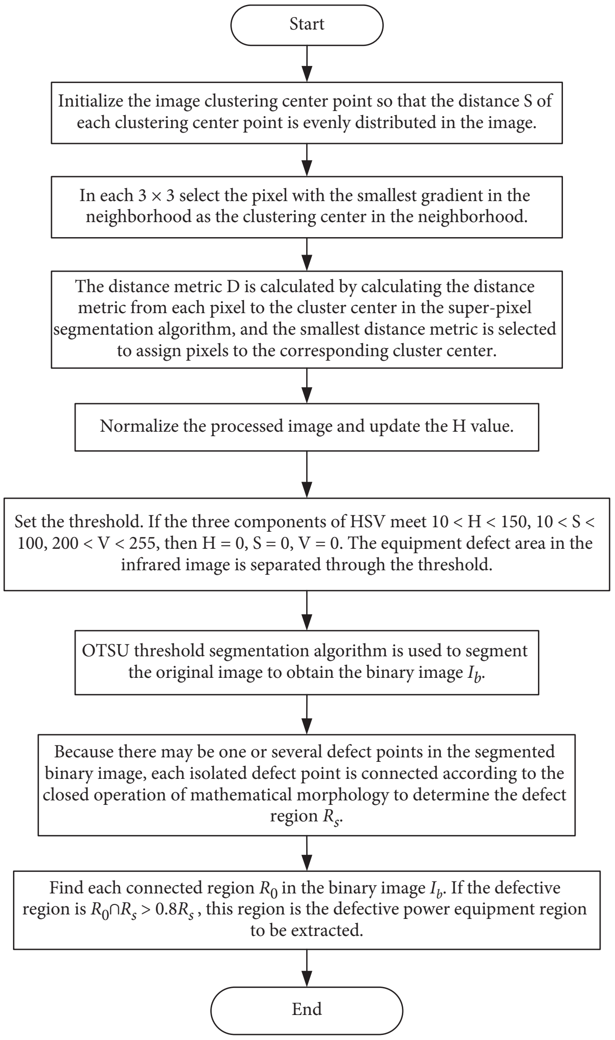

Start
Initialize the image clustering center point so that the distance S of each clustering center point is evenly distributed in the image.
In each 3 × 3 select the pixel with the smallest gradient in the neighborhood as the clustering center in the neighborhood.
The distance metric D is calculated by calculating the distance metric from each pixel to the cluster center in the super-pixel segmentation algorithm, and the smallest distance metric is selected to assign pixels to the corresponding cluster center.
Normalize the processed image and update the H value.
Set the threshold. If the three components of HSV meet 10 < H < 150, 10 < S < 100, 200 < V < 255, then H = 0, S = 0, V = 0. The equipment defect area in the infrared image is separated through the threshold.
OTSU threshold segmentation algorithm is used to segment the original image to obtain the binary image I_b.
Because there may be one or several defect points in the segmented binary image, each isolated defect point is connected according to the closed operation of mathematical morphology to determine the defect region R_s.
Find each connected region R_0 in the binary image I_b. If the defective region is $R_0 \cap R_s > 0.8R_s$, this region is the defective power equipment region to be extracted.
End

Figure 2: Defect region extraction process based on HSV.

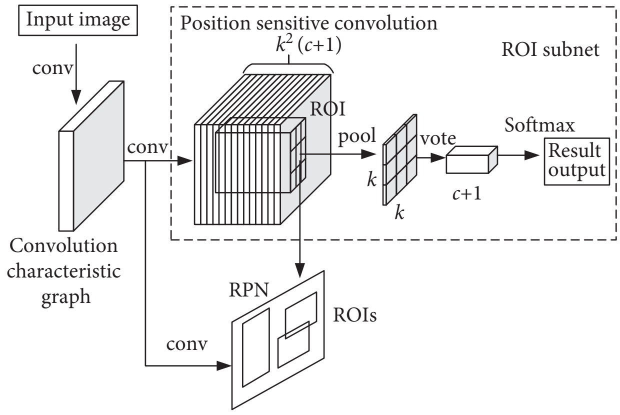


Figure 3: Architecture of R-FCN algorithm.

$H(x)$ and the estimated value x. The advantage of ResNet network is that it uses the stacking layer to fit $H(x)$ to get the mapping $H(x) = F(x) + x$. The advantage of this representation is that if the model has been fitted to the best state, it only needs to make $F(x) = 0$ to get $H(x) = x$, so as to avoid the disappearance and explosion of gradient.

The input x_m and output x_{m+1} of the m-th residual unit are expressed as follows:

$$x_{m+1} = f\left(h\left(x_m\right) + \delta\left(x_m, \omega_m\right)\right),$$
$$x_M = x_m + \sum_{i=1}^{M-1} \delta\left(x_i, \omega_i\right), \tag{6}$$

where δ is the ReLu activation function, ω is the weight between each unit, m, M respectively represent the shallow residual unit and deep residual unit, $h(\cdot)$ represents the identity mapping, and x_M is the final output of the residual unit.

In order to learn more and more abstract image features, the proposed method selects ResNet 101 network, and its configuration is shown in Table 1.

(2) RPN Network. The input of RPN network is image feature graph. According to anchor mechanism, 9 rectangular boxes with different sizes are generated for each point. When training the RPN network, compare the anchor with the manually calibrated true value area in the data set, mark the anchor frame with the largest overlap ratio as the foreground, and mark the anchor frame with an overlap ratio greater than 0.7 as the foreground sample. Mark the anchor box whose overlap ratio is less than 0.3 as the background sample. Select the positive and negative samples of anchor in proportion, use the maximum suppression method (NMS) and other methods to screen the top 250 ROIs with the highest score, and send these preliminarily screened preselected frames to the ROI subnet.

In addition, RPN network adopts anchor mechanism, which not only solves the problem of translation invariance, but also enables R-FCN algorithm to identify and locate targets with different overall dimensions. In the actual process of infrared image recognition of power distribution equipment, due to different equipment with different shape and structure, different sizes and variable aspect ratio, in order to ensure that there are targets in the receptive field corresponding to each sliding window on the feature map, multiscale anchor is required to ensure that the candidate frame is as complete as possible to select the target [25]. In the implementation of RPN network anchor, multiscale anchor can be obtained by setting the area of reference window (base_size), different area multiples and anchor aspect ratio, so that RPN can give more accurate foreground recommendation area.

(3) ROI Subnet. The role of ROI subnet is to correct the ROIs location obtained from RPN network, so as to obtain a more accurate target location of power equipment, so as to identify ROI. A position sensitive convolution layer is added after the last layer of the full convolution network, which can realize the translation variability of the algorithm and output the $k^2(c+1)$-dimensional position sensitive fractional graph. The RPN network filters out the characteristic map of ROI with size $a \times b$, which is divided into $k \times k$ parts (bin) and convoluted with $k^2(c+1)$ convolution cores; that is, each

Table 1: Structure table of ResNet 101.

Layer name	101-Layer
conv1	7*7, stride 2
conv2_x	3*3max pool, stride 2 $\begin{bmatrix} 1\times 1 & 64 \\ 3\times 3 & 64 \\ 1\times 1 & 256 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 1\times 1 & 128 \\ 3\times 3 & 128 \\ 1\times 1 & 512 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 1\times 1 & 256 \\ 3\times 3 & 256 \\ 1\times 1 & 1024 \end{bmatrix} \times 23$
conv5_x	$\begin{bmatrix} 1\times 1 & 512 \\ 3\times 3 & 512 \\ 1\times 1 & 2048 \end{bmatrix} \times 3$
Average pool, 1000-D FC, softmax	

part is mapped to a score map, in which $(c+1)$ is the number of categories plus the background to obtain the $k^2(c+1)$-layer position sensitive score map. The number of channels of bin in different positions is $a \times b \times (c+1)$, and the score of class $(c+1)$ in this position is stored. After the pooling process is completed, vote on the ROI. Sum the $k \times k$ parts bin to get the output of $(c+1)$ dimension, that is, the probability of category (score), and classify it through the softmax layer. At this time, the output result is the positioning coordinates and type of the object.

(4) OHEM. In RPN network, a large number of rectangular boxes are generated, and hundreds or thousands of regions will participate in the training of predicting target categories and locations. The proportion of power equipment is small, so the ratio of equipment background area to target area is too large, resulting in sample imbalance, which makes it difficult to identify distribution equipment. Therefore, online hard example mining (OHEM) method is used to train the network model [26]. During training, when there is a preselected area with large loss, the hard example can be trained and classified again, which can solve the problem of imbalance between positive and negative sample categories and improve the accuracy of infrared image recognition model of power equipment. The training framework of OHEM method is shown in Figure 4.

When OHEM carries out specific training, firstly, ResNet 101 is used to extract the image features of training samples, train the image classification and positioning branches, and calculate the classification loss L_{cls} and regression loss L_{reg} of each target. The loss function is calculated as follows:

$$L\left(s, d_{x,y,a,b}\right) = L_{cls}\left(s_{c^*}\right) + \zeta\left(c^* > 0\right) L_{\text{reg}}\left(d, d^*\right), \tag{7}$$

where $d_{x,y}$ is the upper left coordinate of the target area, d_a is the width of the target area, d_b is the height of the target area, ζ is the balance coefficient of classification loss and regression loss, and s is the image type.

Then, sort according to the loss value from high to low, and use the NMS method to select the first Ω samples with the largest loss value to screen out the difficult cases of this round of samples. Finally, these difficult samples are backpropagated to the network and trained again to update the weight of the whole network, and the penalty for low loss samples is ignored. Among them, the weight update is realized based on the random gradient descent method.

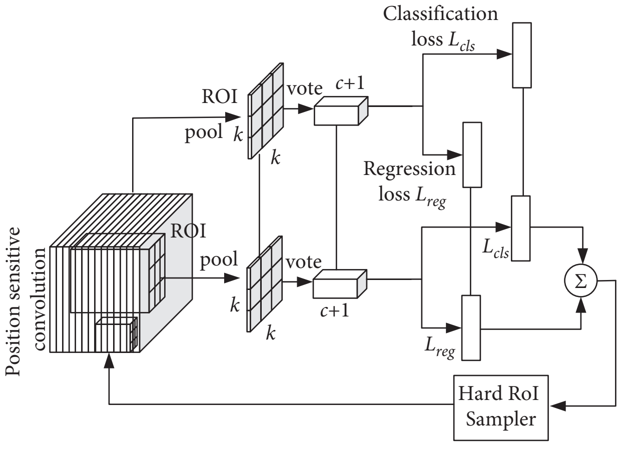

Classification loss L_{cls}
ROI
pool
k
k
vote
$c+1$
Regression loss L_{reg}
ROI
pool
k
k
vote
$c+1$
L_{cls}
Σ
L_{reg}
Hard RoI Sampler
Position sensitive convolution

Figure 4: Training framework of OHEM method.

3.3.2. Fault Diagnosis of Distribution Equipment. There are many kinds of equipment in the distribution network, such as circuit breaker, potential transformer (PT), current transformer (CT), lightning arrester, and transformer. These equipment can be classified into current heating type, voltage heating type, and comprehensive heating type according to the heating factors. Different diagnostic methods are selected for different types of equipment, and different diagnostic methods have different diagnostic criteria. The defects of the equipment, such as general defects, serious defects, and critical defects, are determined through various diagnostic methods.

(1) The heating of current heating equipment is mainly due to the current thermal effect. Generally, the surface temperature judgment method and relative temperature difference judgment method are used for fault diagnosis. The relative temperature

difference ∇_T and the maximum surface temperature $T_{\max}$ of different defect degrees of each distribution equipment are shown in Table 2.

(2) Voltage heating equipment is mainly due to voltage effect. The main equipment categories include zinc oxide arrester, high-voltage bushing, and coupling capacitor. This kind of equipment generally adopts image feature judgment method, similar comparison judgment method, and comprehensive analysis judgment method. Using the thermal image characteristics of the equipment and the image feature judgment method, the fault can be found quickly. If the similar comparison discrimination method and comprehensive analysis judgment method are used, the temperature difference ΔT shall be taken as the fault diagnosis index, in which the ΔT of zinc oxide arrester is 0.5~1 K, and the ΔT of high-voltage bushing, coupling capacitor, oil immersed PT and CT are 2~3 K.

(3) Comprehensive heating equipment needs to be diagnosed in combination with the diagnosis methods of voltage heating equipment and current heating equipment, mainly including insulators, generators, and transformers. In the actual thermal fault diagnosis, the fault diagnosis indexes of various methods should be combined to improve the efficiency and accuracy of fault diagnosis [27].

According to different types of distribution equipment, the corresponding thermal fault diagnosis and judgment methods can be selected, the corresponding parameters can be calculated, and the thermal fault diagnosis of distribution equipment can be carried out according to the diagnosis criteria. The fault diagnosis process of power distribution equipment is shown in Figure 5.

Firstly, the infrared image of distribution equipment is input into the detection model for image preprocessing, and the defect area is extracted based on HSV to divide the structure of distribution equipment. Then, the trained depth learning hybrid model is used to obtain the type and location of the target equipment, and the temperature information of each structural area on the infrared image of the distribution equipment is read at the same time. Finally, according to the selected fault diagnosis method, the thermal fault state, thermal fault level, and thermal fault location of distribution equipment are determined by using the diagnosis criterion. Different from the existing fault diagnosis methods of artificial equipment, the proposed method can use the deep learning hybrid model to realize the intelligent classification and fault type diagnosis of infrared images of distribution equipment, greatly reduce the workload of inspectors, and improve the automation level of fault diagnosis of distribution equipment.

4. Experiment and Analysis

The experiment is carried out in the 64 bit operating system environment of Ubuntu 6.04.4 LTS, in which the deep learning hybrid model uses the deep learning framework PyTorch v1.3 version, the programming language is Python 3.6.0, and third-party dependent libraries such as Open CV 4.0 and NumPy 1.3 are used for batch processing of data. At the same time, the network model is trained on the GPU of dual card Tesla P100, with a total video memory of 8 GB. In this way, large batch data can be set during training to improve the convergence speed of the model. Other hardware environments are as follows: 512 GB of memory resources and 1 TB of hard disk. When processing data and I/O operations on a large scale, this can realize parallel processing and high speed and ensure the training requirements of network model.

4.1. Experimental Data Set. The research scenario is the power equipment in the distribution network, so it is necessary to collect the picture materials of equipment faults in the distribution system and then construct a fixed format data set for test training. Through safety training and professional leadership, use mobile phones, cameras, and other equipment to shoot at the site of power distribution equipment. The scene of abnormal power grid equipment is selected, and a large number of positive samples are collected in multiple directions according to the shooting angle of video monitoring. In order to ensure the rationalization of data distribution, all types and forms of equipment anomaly types are covered in the acquisition process. After that, the abnormal categories and areas of power grid equipment are marked through the open-source and free wizard marking software.

The data set contains 2580 on-site abnormal images of RGB power grid equipment during the day, 6300 mark boxes in total, and 2769 on-site abnormal images of infrared power grid equipment at night, 7000 mark boxes in total, all from the real power distribution room, power equipment plant, etc. After the data annotation is completed, the annotation file in XML format is generated, which corresponds to the real image.

4.2. Comparative Analysis of Training Speed. The proposed method combines the hybrid model of robot and deep learning for equipment fault diagnosis. ResNet network is used to optimize the R-FCN algorithm, and the defect area is extracted based on HSV to improve the diagnosis effect. In order to demonstrate the improvement effect of the proposed method, it is compared with the diagnosis methods of ResNet network, Otsu threshold segmentation, and OHEM training. The diagnosis accuracy and training time are shown in Table 3.

It can be seen from Table 3 that extracting deeper equipment fault features using ResNet network can greatly improve the diagnosis accuracy, which is 9.88% higher than that of the model. However, due to the deepening of network layers, the training time is also increased, more than 10 s. At the same time, by integrating OHEM training depth learning model, the diagnostic accuracy continued to improve by 4.41%. Due to the simple and easy implementation of the training process, the training time is only increased by 0.7 s. It can be seen that the diagnostic accuracy of the proposed

Table 2: Fault diagnosis criterion of current heating equipment.

	General defect	Serious defect	Emergency defect
Circuit breaker	$35\% \leq \nabla_T < 80\%$	$80\% \leq \nabla_T < 95\%$	$\nabla_T \geq 95\%$
	—	$55\% \leq T_{\max} \leq 80\%$	$T_{\max} > 80\%$
Disconnecting switch	$35\% \leq \nabla_T < 80\%$	$80\% \leq \nabla_T < 95\%$	$\nabla_T \geq 95\%$
	—	$90\% \leq T_{\max} \leq 130\%$	$T_{\max} > 130\%$
CT	$35\% \leq \nabla_T < 80\%$	$80\% \leq \nabla_T < 95\%$	$\nabla_T \geq 95\%$
	—	$55\% \leq T_{\max} \leq 80\%$	$T_{\max} > 80\%$
Capacitor	$35\% \leq \nabla_T < 80\%$	$80\% \leq \nabla_T < 95\%$	$\nabla_T \geq 95\%$
	—	$55\% \leq T_{\max} \leq 80\%$	$T_{\max} > 80\%$
High voltage bushing	$35\% \leq \nabla_T < 80\%$	$80\% \leq \nabla_T < 95\%$	$\nabla_T \geq 95\%$
	—	$55\% \leq T_{\max} \leq 80\%$	$T_{\max} > 80\%$

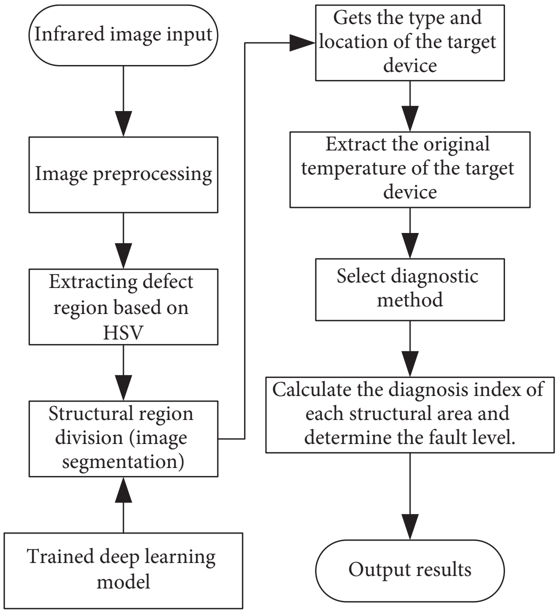

Infrared image input
Image preprocessing
Extracting defect region based on HSV
Structural region division (image segmentation)
Trained deep learning model
Gets the type and location of the target device
Extract the original temperature of the target device
Select diagnostic method
Calculate the diagnosis index of each structural area and determine the fault level.
Output results

Figure 5: Fault diagnosis process of distribution equipment.

Table 3: Fault diagnosis criterion of current heating equipment.

Method	Accuracy/%	Training time/s
Original model	82.49	4.6
Original model + OTSU threshold segmentation	85.82	5.9
Original model + OTSU threshold segmentation + ResNet network	92.37	10.5
Original model + OTSU threshold segmentation + ResNet network + OHEM	96.78	11.2

method is higher than that of the basic method, but the training time increases, and the training speed decreases.

In order to demonstrate the performance of the proposed method in training speed, it is compared with reference [11], reference [13], and reference [16]. The results are shown in Figure 6.

As can be seen from Figure 6, the training time of reference [13] is the shortest, only about 5S. Because its intuitionistic fuzzy clustering algorithm based on spatial distribution information for image recognition is simple and easy to implement, the training speed is fast. Reference [13] combines discrete wavelet transform and support vector machine algorithm to complete fault diagnosis, and [16] uses artificial neural network algorithm to classify faults. Both methods are complex and take a long time to calculate, so the training time is about 10 s. By using the improved R-FCN algorithm for fault diagnosis, the proposed method uses OHEM method to train it, which can simplify the data processing process, the training speed is fast, and the training time is about 5.5 s. At the same time, the robot background is used for data analysis, which can reduce the transmission time of image information.

4.3. Comparative Analysis of Fault Diagnosis Accuracy. The accuracy of fault diagnosis is a key judgment index. The accuracy of the proposed method and the methods in reference [11], reference [13], and reference [16] for

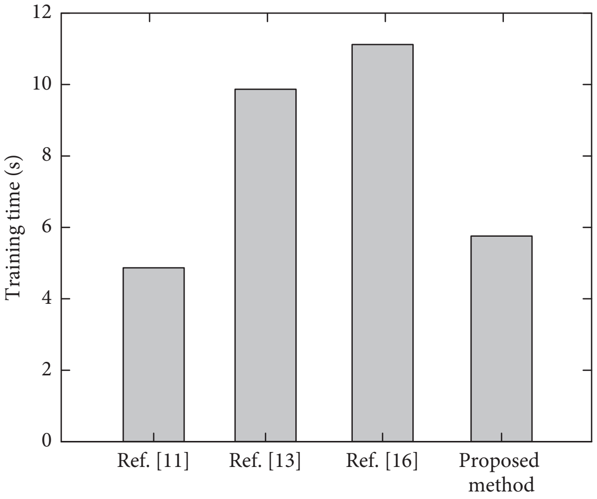


Figure 6: Training time of different methods.

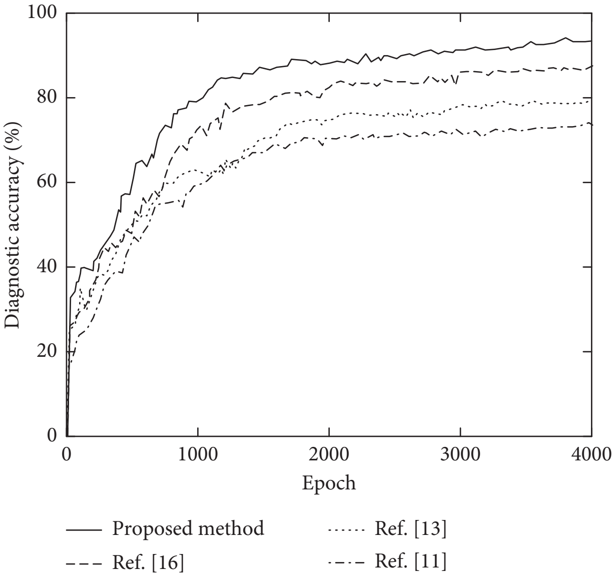


Figure 7: Diagnostic accuracy of different methods.

fault diagnosis of distribution equipment is shown in Figure 7.

As can be seen from Figure 7, compared with other methods, with the iteration of epoch, the fault diagnosis accuracy of the proposed method tends to be stable, about 92.06%. Due to its combination of robot and deep learning hybrid model, it deeply extracts the characteristics of various types of fault equipment for diagnosis, which further ensures the reliability of diagnosis results. Similarly, [16] uses artificial neural network algorithm for state recognition, but there is no efficient way to obtain the equipment state, and there is no complete database to support it. Therefore, the diagnosis accuracy is reduced by about 6% compared with the proposed method. Reference [13] adopts the improved support vector machine algorithm of genetic algorithm for fault detection, which has a good effect on high impedance fault diagnosis, but its universality is not high, so the diagnosis accuracy is about 80%. Reference [11] uses the traditional intuitionistic fuzzy clustering algorithm for graphic classification. The traditional method is difficult to apply to a large number of distribution equipment, so the diagnosis accuracy is low.

For the three fault types, the diagnostic accuracy of different methods is shown in Table 4.

It can be seen from Table 4 that the diagnostic accuracy of comprehensive heating equipment is generally lower than that of current heating equipment and voltage heating equipment. Taking the proposed method as an example, the diagnostic accuracy of comprehensive heating equipment is 89.31%, and the other two types are higher than 90%. Because the diagnostic criteria of comprehensive heating equipment are complex and easy to be confused, they affect the fault diagnosis. The recognition accuracy of current heating type defects is slightly higher, which may be due to

Table 4: Comparison results of diagnostic accuracy of each fault type.

Method	Current heating type (%)	Voltage heating type (%)	Comprehensive heating type (%)
Reference [11]	73.23	69.18	65.75
Reference [13]	82.84	80.36	79.04
Reference [16]	86.69	87.05	84.27
Proposed method	93.52	91.88	89.31

the obvious characteristics and large amount of data of infrared images of current heating type defects in the data set. However, the diagnosis accuracy of the proposed method is higher than that of other comparison methods. Taking the current heating equipment as an example, its diagnosis accuracy is as high as 93.52%, because it can well distinguish all kinds of faulty equipment by using the improved R-FCN algorithm to learn the equipment image features and evaluate the fault level according to the fault judgment. Other comparison methods only diagnose whether the equipment is faulty or not, but the diagnosis effect is poor for various specific fault types.

5. Conclusion

Nowadays, the construction of smart grid in China has entered a new stage of comprehensive and rapid development. The traditional manual detection methods have been difficult to deal with a large amount of infrared image data of power equipment. Therefore, a fault diagnosis method of distribution equipment based on the hybrid model of robot and deep learning is proposed. The image information database of distribution equipment based on robot inspection is constructed, and the OTSU method is used to extract the defect features of distribution equipment from the binary image of equipment based on HSV space. Then, the equipment defect characteristics are sent to the improved R-FCN algorithm for learning and analysis to obtain the fault type and location, and the fault level is obtained through the calculation of fault criterion. The experimental results based on PyTorch platform show that:

(1) Using the robot platform to build the image information database of distribution equipment can improve the accuracy of fault diagnosis of various equipment. The diagnostic accuracy of the proposed method for current heating equipment, voltage heating equipment, and comprehensive heating equipment is 93.52%, 91.88%, and 89.31%, respectively.

(2) Using OHEM method to train the improved R-FCN algorithm can shorten the model training time, improve the fault diagnosis efficiency, and further improve the diagnosis effect. The fault diagnosis time and accuracy are 5.5 s and 92.06%, respectively.

Since the extraction degree of the defective region will affect the recognition accuracy of the subsequent model, considering the subsequent semantic segmentation of the overall infrared image by using the convolution neural network optimized by conditional random field, the defective power equipment can be segmented in a more complex background to further improve the recognition accuracy of the subsequent model.

References

[1] S. Gangolu, P. Raja, M. P. Selvan, and V. K. Murali, “Effective algorithm for fault discrimination and estimation of fault location in transmission lines,” *IET Generation, Transmission & Distribution*, vol. 13, no. 13, pp. 2789–2798, 2019.

[2] N. Hu, H. Du, S. Liu, and Q. Lin, “Power equipment status information parallel fault diagnosis of based on MapReduce,” *Journal of Computational Methods in Science and Engineering*, vol. 19, no. 88, pp. 1–6, 2019.

[3] E. Bashar, Q. Han, R. Wu, L. Ran, O. Alatise, and S. Jupe, “Analysis of DC offset in fault current caused by machines in a medium voltage distribution network,” *Journal of Engineering*, vol. 2019, no. 17, pp. 3494–3499, 2019.

[4] H. Tian, P. Liu, S. Zhou et al., “Research on the deterioration process of electrical contact structure inside the ± 500 kV converter transformer RIP bushings and its prediction strategy,” *IET Generation, Transmission & Distribution*, vol. 13, no. 12, pp. 2391–2400, 2019.

[5] W. Luo, H. Wang, L. Wang, Z. Zhu, and H. Gao, “Faulted line location method for distribution systems based on the equipment’s information exchange,” *Dianli Xitong Baohu yu Kongzhi/Power System Protection and Control*, vol. 47, no. 4, pp. 73–82, 2019.

[6] N. Narasimhulu, D. Kumar, and M. V. Kumar, “Detection and classification of high impedance fault in power distribution system using hybrid technique,” *Journal of Circuits, Systems, and Computers*, vol. 29, no. 08, pp. 67–976, 2020.

[7] K. Chen, J. Hu, Y. Zhang, Z. Yu, and J. He, “Fault location in power distribution systems via deep graph convolutional networks,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 1, pp. 119–131, 2020.

[8] F. Wan, P. Madhika, J. Chwa, M. Mozumdar, and A. Ameri, “Automatic optimal synthesis of aircraft electric power distribution system,” *International Journal of Computing and Digital Systems*, vol. 9, no. 3, pp. 363–375, 2020.

[9] K. Jia, Q. Zhao, T. Feng, and T. Bi, “Distance protection scheme for DC distribution systems based on the high-frequency characteristics of faults,” *IEEE Transactions on Power Delivery*, vol. 35, no. 1, pp. 234–243, 2020.

[10] A. S. Alayande, I. K. Okakwu, O. E. Olabode, and O. K. Nwankwoh, “Analysis of unsymmetrical faults based on

artificial neural network using 11 kV distribution network of University of lagos as case study," *Journal of Advances in Science and Engineering*, vol. 4, no. 1, pp. 53–64, 2021.

[11] F. Hu, H. Chen, and X. Wang, "An intuitionistic kernel-based fuzzy C-means clustering algorithm with local information for power equipment image segmentation," *IEEE Access*, vol. 8, no. 6, pp. 4500–4514, 2020.

[12] H. Liu, Y. Wang, and W. Chen, "Anomaly detection for condition monitoring data using auxiliary feature vector and density-based clustering," *IET Generation, Transmission & Distribution*, vol. 14, no. 1, pp. 108–118, 2020.

[13] Youness, Mohammadnian, A. Turaj, and A. Soroudi, "Fault detection in distribution networks in presence of distributed generations using a data mining–driven wavelet transform," *IET Smart Grid*, vol. 2, no. 2, pp. 163–171, 2019.

[14] Q. Yang, J. Ruan, and Z. Zhuang, "Fault diagnosis of circuit breakers based on time-frequency and chaotic vibration analysis," *IET Generation, Transmission & Distribution*, vol. 14, no. 7, pp. 1214–1221, 2020.

[15] T. Zhang, H. Yu, P. Zeng, L. Sun, C. Song, and J. Liu, "Single phase fault diagnosis and location in active distribution network using synchronized voltage measurement," *International Journal of Electrical Power and Energy Systems*, vol. 117, no. 5, 2020.

[16] S. Frizzo-Stefenon, M. C. Silva, D. W. Bertol, L. H. Meyer, and A. Nied, "Fault diagnosis of insulators from ultrasound detection using neural networks," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 5, pp. 6655–6664, 2019.

[17] B. Wang, M. Dong, M. Ren et al., "Automatic fault diagnosis of infrared insulator images based on image instance segmentation and temperature analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 8, pp. 5345–5355, 2020.

[18] H. Liang, Y. Liu, G. Sheng, and X. Jiang, "Fault-cause identification method based on adaptive deep belief network and time-frequency characteristics of travelling wave," *IET Generation, Transmission & Distribution*, vol. 13, no. 5, pp. 724–732, 2019.

[19] M. Gholami, A. Abbaspour, M. Moeini-Aghtaie, M. Fotuhi-Firuzabad, and M. Lehtonen, "Detecting the location of short-circuit faults in active distribution network using PMU-based state estimation," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1396–1406, 2020.

[20] W. Hu, C. Ruan, H. Nian, and D. Sun, "Simplified modulation scheme for open-end winding PMSM system with common DC bus under open-phase fault based on circulating current suppression," *IEEE Transactions on Power Electronics*, vol. 35, no. 1, pp. 10–14, 2020.

[21] K. Zhu and P. W. T. Pong, "Fault classification of power distribution cables by detecting decaying DC components with magnetic sensing," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2016–2027, 2020.

[22] L. Romero, J. Blesa, V. Puig, G. Cembrano, and C. Trapiello, "First results in leak localization in water distribution networks using graph-based clustering and deep learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 16691–16696, 2020.

[23] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for volt-VAR control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008–3018, 2020.

[24] D. A. León-Vargas, V. A. Bucheli-Guerrero, and H. A. Ordoez, "Solar radiation prediction on photovoltaic systems using machine learning techniques," *Revista Facultad de Ingeniería*, vol. 29, no. 10, pp. 1–20, 2020.

[25] S. Yamane and K. Matsuo, "Adaptive control by convolutional neural network in plasma arc welding system," *ISIJ International*, vol. 60, no. 5, pp. 998–1005, 2020.

[26] M. Nabati, H. Navidan, R. Shahbazian, S. A. Ghorashi, and D. Windridge, "Using synthetic data to enhance the accuracy of fingerprint-based localization: a deep learning approach," *IEEE Sensors Letters*, vol. 4, no. 4, pp. 1–4, 2020.

[27] M. S. Erbnescu, N. C. Manea, L. Streba et al., "Automated gleason grading of prostate cancer using transfer learning from general-purpose deep-learning networks," *Romanian Journal of Morphology and Embryology*, vol. 61, no. 1, pp. 149–155, 2020.

Construction of a Big Data Platform for Economic Management Using Artificial Intelligence Algorithms Monitoring and Early Warning

Santosh Kumar Sharma, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, sk.sharma258@yahoo.co.in*

Purnya Prava Nayak, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, purnyaprava.nayak26@gmail.com*

Manisha Pradhan, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, manishapradhan456@gmail.com*

Jui Pattanaik, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, juipattanayak@gmail.com*

Abstract

The development of artificial intelligence and the emergence of big data have brought convenience to the development of various fields and also brought great influence to the economic field. There are many data sources for economic management, and the scale is huge, so how to manage these large-scale economic data has become an urgent problem to be solved. In addition, the issue of how to conduct security management on these large-scale data, protect the security of users' accounts and property, and effectively monitor and prewarn economic market risks is also an urgent issue. This article aims to build an economic market risk monitoring and early warning platform through advanced science and technology such as artificial intelligence and big data to realize an intelligent risk control platform in the economic and financial fields, as well as a data-driven risk management model to create intelligent risk early warning and prevention and the response system to enhance the intelligent level of risk assessment, early warning, prevention, and disposal. Experiments show that the artificial intelligence algorithm monitoring and early warning economic management big data platform constructed in this article shows that its accuracy of economic risk prediction can reach more than 90%.

1. Introduction

In recent years, the rapid development of information technology and communication networks and the vigorous development of the network economy have also brought greater difficulties to economic management. The network economy is that the laws of economic operation and some basic laws have changed; for example, online economic operation needs to be carried out according to the prescribed ways online and conform to the rules of online operation. But, because of the huge scale of economic data, the economic field cannot adjust its management system and management mode in time, making the economy face great management risks. In the process of transmission and communication on the Internet, the risk of inaccurate information being stolen or contaminated is also difficult to avoid [1]. In addition, the development of the network economy is also promoting the process of financial globalization and integration to the financial system. Therefore, with the development of the network economy, the risks faced by the economic field have also increased.

The artificial intelligence algorithm economic monitoring and early warning system constructed in this article monitors the development of economic information data in the big data platform and at the same time predicts future economic development trends based on the development and changes of these economic data and predicts the development of financial markets. It is an important part of the economic field. Decision-makers provide a reference data to make correct decisions; use big data, cloud computing, artificial intelligence, blockchain, and other mobile Internet technologies to provide the driving force for the integration of technology and economy to promote the innovation and transformation of economic management models [2]; build a three-dimensional, socialized, and information-based monitoring and early warning system, which can detect the trend of economic crimes in time, curb the trend of high incidence of illegal fund-raising, and avoid the losses caused by illegal means to the economic field; and improve the risk assessment of the economic field. The accuracy of this can make risk assessment intelligent and automated and promote the smooth development and operation of the entire economic field.

In order to grasp the development trend of the economic field, promote the stable development of the economy, and avoid possible risks in the economic field, many scholars have conducted research in this field. Among them, Pisareva, in view of the deployment of temporary processes in various socioeconomic systems and aggravating the crisis, proposed a formal method of setting to further improve the corresponding application tasks at all levels and in the field of socioeconomic development planning [3]. Lehrstuhl described the paradigm shift experienced in the field of macroeconomic research by adopting an evolutionary approach. The application of big data to economic problems can lead to new ways of thinking and research methods [4]. Lei et al. designed a new type of computer architecture, an accelerator based on an optical network chip (ONoC) to further accelerate the matching of citizens' supply and demand in the sharing economy [5]. Holm and Ploug believed that the governance model including the possibility of making metachoices can achieve the best balance between personal interests and public interests [6]. Li et al. discussed the key gaps and opportunities in the economic field, referred to the existing literature on decision-making, scenario analysis, and scientific philosophy under uncertainty, and developed the basic concepts guiding the future application of big data to energy economic modeling [7]. However, although these studies have a good reference value for economic development, they need to manually find out the risk control in the economic field. There is no complete warning system, and there are few studies on economic risk prediction.

This article has the following innovations: (1) Use big data technology to realize the integration of huge economic data information in the economic field, build a high-performance big data collection, storage, and analysis platform, and efficiently integrate various economic information data. Apply to risk control, reduce the overall risk of the online economy, and improve the ability to control economic risks. (2) Combine artificial intelligence, blockchain, and other mobile Internet technologies to design an economic management big data platform to realize the intelligent prediction of economic risks and the ability to control risks at the same time, improve the automation level of risk prediction and issue economic risk early warnings, and build artificial intelligence predictive warning system. (3) The newly constructed economic management big data platform based on artificial intelligence algorithm monitoring and warning has automatic learning function and automatic alarm device which is not available in the pain platform.

2. Construction Method of Economic Management Big Data Platform

2.1. Mobile Internet Technology

2.1.1. Artificial Intelligence Algorithm Monitoring and Early Warning. A scientific technology used by artificial intelligence to expand, simulate, and extend human intelligence is to make computers more likely to approach human behavior and IQ [8]. The application of artificial intelligence technology is shown in Figure 1.

The fields of artificial intelligence applications are far more than those listed in Figure 1. In the future, artificial intelligence technology will become more and more advanced, and the fields in which it can be used will become more and more extensive [9–11]. At present, the application of artificial intelligence in the economic field is becoming more and more extensive, such as the digital economy. Due to the emergence of the digital economy, it has brought us changes in the infrastructure in our lives and the development of emerging industries [12]. Of course, the development of the smart economy requires a complete set of smart forecasting systems to prevent economic risks and predict the development of future economic trends. Because the development of intelligent economy is affected by many uncertain factors, it is necessary for intelligent system to predict the development of intelligent economy in order to ensure the smooth operation of intelligent economy. For this reason, it is necessary to combine the data in the big data platform with corresponding algorithms, so that artificial intelligence can combine the corresponding algorithms to make predictions based on economic data. There are many algorithms used in artificial intelligence. In order to accurately predict various trends in economic management, this article uses a particle swarm algorithm that simulates the lost behavior of a flock of birds. This algorithm is used to reduce the error of predicting trends to ensure that the system is economical [13].

What the artificial intelligence particle swarm optimization algorithm needs is to use the economic data in big data to calculate and then reduce the error step by step. We regard economic data as a particle, and the amount of economic data is constantly changing and fluctuating. The records of the stored data are different. Every time the economic data will be presented in a different location, we remember this location as

$$C_i(t_i) = [C_1(t_1), C_2(t_2), \ldots, C_N(t_i)]^i, \quad i = 1, 2, 3 \ldots N. \tag{1}$$

In this formula, t is the time when the data is recorded, i represents the data label when the data is recorded, so that we can quickly find the data, and C represents the data in the server location.

In the process of data update and iteration, the economic data of the optimal prediction of the data at the t-th time can be obtained, which we call the individual extreme value, which is recorded as

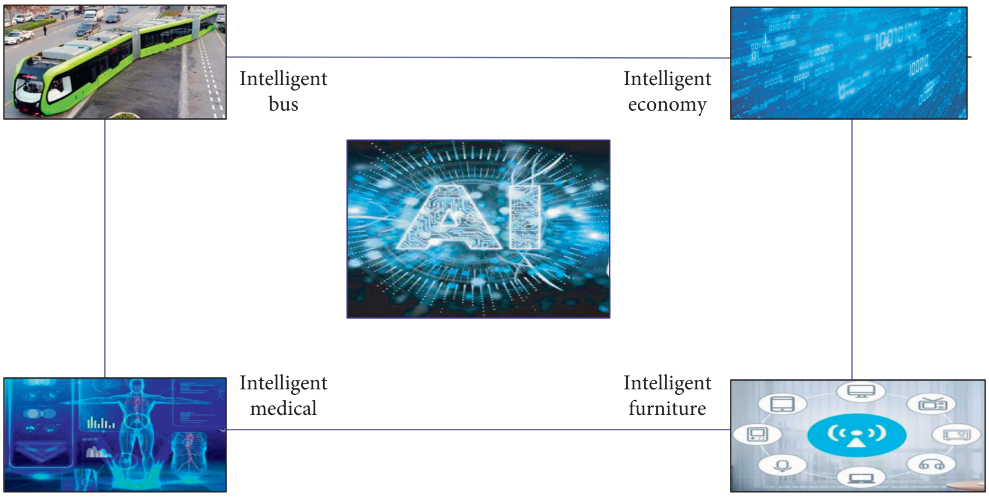


Figure 1: Application of artificial intelligence.

$$C_i^N(t_i) = \left[C_i^1(t_1), C_i^2(t_2), \ldots, C_i^N(t_i)\right]^i. \tag{2}$$

When we analyze and integrate economic data, we need to analyze all economic data to be able to correctly grasp the risks that may exist in the process of economic development and other aspects of economic development. Therefore, there will be an overall extreme value. We denote this extreme value as

$$C_i^Z(t_i) = \left[C_i^{(n+1)}\left(t_{(i+1)}\right), C_i^{(n+2)}\left(t_{(i+2)}\right), \ldots, C_i^Z\left(t_{(i+N)}\right)\right] \cdot S^t. \tag{3}$$

In the above equation, S is the weight that will be generated by the calculation of economic data in the whole. The speed of the i-th data in the t-th iteration update is

$$Q_i(t_N) = \left[Q_1(t_1), Q_2(t_2), \ldots, Q_i(t_N)\right]^i, \quad i = 1, 2, 3, \ldots, N. \tag{4}$$

Then the principle of the iterative update of the individual extreme value of the i-th data at the t-th time is as follows:

$$\begin{aligned} C_i^N(t_i + 1) &= C_i(t_1 + 1), Y\left(C_i(t_2 + 1)\right) \geq Y\left(C_i^N(t)\right), \\ C_i^N(t_i) &= C_i(t_1), Y\left(C_i(t)\right) \leq Y\left(C_i^N(t)\right). \end{aligned} \tag{5}$$

In the above equation, $Y(x)$ is the degree of adaptation function. All data is updated iteratively at time t, and all extreme values will also change according to the data update. The iterative update formula is as follows:

$$\begin{aligned} C_i^Z(t_i) &= G\left\{\min Y\left(C_i^Z(t_i)\right)\right\}, \\ x &= \left(C_i^Z(t_i)\right), \end{aligned} \tag{6}$$

where G represents the value of the independent variable x in the function $Y(x)$ and $C_i^Z(t_i)$ is the closest accurate actual economic data in the t-th iteration update, which we call the global optimal predicted economic data value.

In artificial intelligence, we use particle swarm optimization algorithm to find the trend of economic development. At the same time, we can also use particle swarm optimization algorithm to predict problems and risks in economic data, reducing our search for local data in solving huge economic data. It can predict the possible risks of problems and guide and prevent them in time.

(1). Blockchain technology. Before the large-scale economic data, the data information in our economic development needs to be absolutely confidential, so we need to choose an absolutely secure storage system to store this large-scale economic data information, and then add it to this storage server artificial intelligence password recognition algorithm [14]. The block storage mode in blockchain technology can ensure the security of these economic data. Block storage and central storage are shown in Figure 2.

As shown in Figure 2, the block storage is compared with the central storage. It can be found that it is divided into block storage, and the information and data are not aggregated into a total storage server, so, in this way, it can be guaranteed that the data in a single storage server does not have the risk of data leakage from other storage servers. The economic data information in other storage servers will not be exposed to the risk. This will bring the storage security of economic data to a higher level. If it is economic data stored in a central storage server, it is possible that all economic information leakage will cause huge economic losses [15].

Then, in the blockchain storage technology, we will use artificial intelligence technology to encrypt each blockchain. Through artificial intelligence technology, once information leaks, it will intelligently sound an alarm and immediately identify which part of the economic data leakage occurs to prevent greater economic losses [16]. Then the principle of adding a password to the blockchain is as follows:

When we set the password, it takes time t and the number of passwords set is d. When we enter the password, the password is passed into the system. The artificial intelligence technology will intelligently identify the entered password. If the password is entered incorrectly, it will not

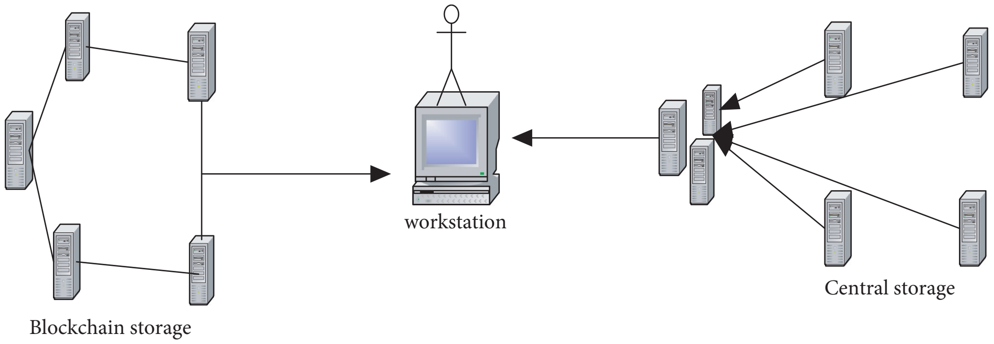


Figure 2: Block storage and central storage.

be possible to view economic data information; the principle is as follows:

$$\text{Code} = f(d)^{*} r_{s}. \tag{7}$$

In the above equation, r is the ability of artificial intelligence to recognize whether it is a password input. Once it is found that this is not a password but other means, it will immediately give an intelligent alarm and prevent illegal intrusion from the outside world. The recognition principle is as follows.

We assume that the number of passwords identified in artificial intelligence recognition is 4, and the password of one of the blockchains is set to 8888; then the first step of artificial intelligence recognition is to identify whether the number of passwords corresponds:

$$\text{Code} = f(4)^{*} r_{s} = \text{In}(4). \tag{8}$$

If the number of passwords corresponds, the artificial intelligence will monitor whether the password is correct, and if it does not correspond, it will perform artificial intelligence to close the password input program so that the password cannot be entered. After entering the correct number of passwords, artificial intelligence will recognize whether the password is correct, if the following forms are recognized:

$$\text{Code} \longrightarrow (8888) = \text{In}(8888) \xrightarrow{t} * v \text{ true}. \tag{9}$$

In this way, economic information data stored in one of our blockchains will be open for us to analyze and use. In order to ensure the security of the data in the data store, a password is set for each storage area [17, 18]. In this way, the risk of economic leakage can be better prevented.

2.2. The Monitoring and Early Warning of the Economy by Artificial Intelligence. The arrival of the information age not only brought convenience to economic development but also brought equal risks to the economic field [19]. For example, the financial risks brought by the development of our network economy will bring huge losses to the economic market, but the huge data information in the financial field cannot be detected manually, so it will greatly increase the occurrence of financial risks [20, 21]. Artificial intelligence is a computer technology that imitates the human brain. Therefore, with the help of artificial intelligence, we can quickly identify abnormal fluctuations in financial data and issue alarms in a timely manner, which also provides us with time to respond to financial risks [22]. The process of intelligently identifying economic risks is shown in Figure 3.

As shown in Figure 3, once the AI recognition system detects abnormal fluctuations in data or abnormal data, it will immediately issue an alarm. Before entering the recognition technology of artificial intelligence, economic data will undergo diversion analysis. The purpose of information diversion is to allow artificial intelligence to accurately identify economic data. The distribution of economic data is shown in Figure 4.

The data stored in the blockchain will be divided and summarized according to the number of block memories, and the data information stored in each block is different. Therefore, when the area is summarized, weight i will be generated. When there are W data streams, they will be summarized to the output layer M. The calculation formula is as follows:

$$\begin{aligned} M_{1} &= \sum_{i}^{x} \frac{W_{1}}{W_{2}} * \varpi, \\ M_{n} &= \sum_{i}^{x} \frac{W_{(n-1)}}{W_{n}} * \varpi. \end{aligned} \tag{10}$$

In the above equation, x is the internal threshold of the system, and ϖ is the correlation coefficient matrix generated when processing the data stream, and its form is as follows:

$$\varpi = \begin{bmatrix} x & n & i \end{bmatrix}. \tag{11}$$

Then the total amount of data recognized by intelligent artificial technology is calculated as follows:

$$H = \left(M_{1} + \cdots M_{n}\right)^{*} \varpi. \tag{12}$$

In this way, entering the artificial intelligence risk identification system can be identified at a faster speed, and, in this identification process, the artificial intelligence algorithm monitoring and early warning system will predict the future economic development trend based on the indicators of economic development surprise. The economic early warning system was created following a crisis in the capitalist market, in order to avoid the negative

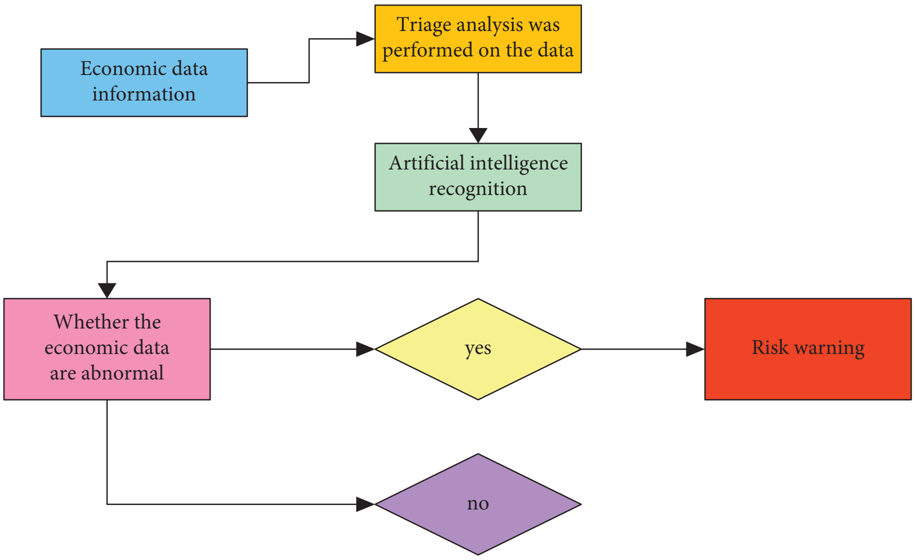

Economic data information
Triage analysis was performed on the data
Artificial intelligence recognition
Whether the economic data are abnormal
yes
Risk warning
no

Figure 3: Flow chart of intelligently identifying economic risks.

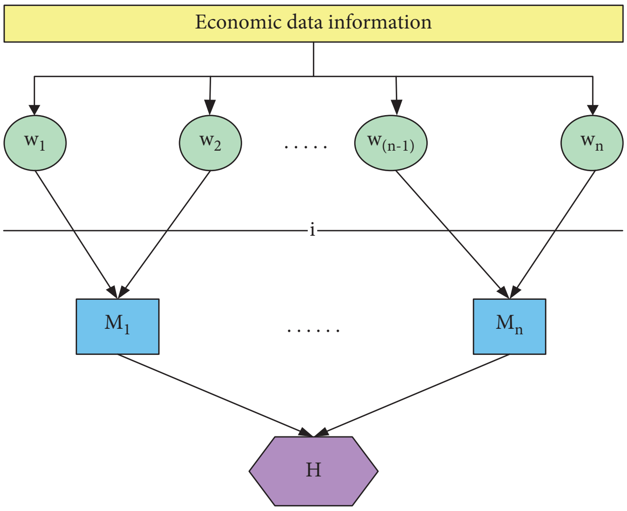


Figure 4: Economic data distribution diagram.

consequences of economic shrinkage and to avoid greater losses due to failure to respond to the corresponding countermeasures when the economy is bad. Nowadays, with the development of technology and economy, the huge amount of economic data and information makes it difficult for humans to analyze, and the development of the digital economy also brings greater risks of business cycle fluctuations, so the use of artificial intelligence algorithms to monitor and early warning can promptly discover potential risks and then make corresponding countermeasures [23].

In the artificial intelligence economic early warning system, in addition to predicting possible economic risks, we also need to provide early warning of the existing economic development boom, so that this big economic data platform can not only use artificial intelligence algorithm monitoring and early warning for risk prediction but also calculate the prosperity index of economic development, making it a way to detect economic risks and calculate the prosperity of economic development, so that we can better make corresponding countermeasures. Therefore, the current economic monitoring and early warning need to not only consider the economic development prosperity index but also monitor the abnormal fluctuations of economic data and potential risks in the economic field, so as to escort the economic development of our information age.

2.3. Construction of Big Data Monitoring and Early Warning Platform for Smart Economy. In order to allow the long-term and stable development of the economic market, we have rebuilt the economic monitoring and early warning system. This system combines artificial intelligence recognition technology and artificial intelligence algorithm monitoring and early warning. In combination with the big data platform to accommodate the large-scale economic information data, the newly built monitoring and early warning platform is shown in Figure 5.

As we can see in Figure 5, the newly constructed big data platform for economic management uses blockchain storage technology in the storage server for storing data, which can prevent the risk of economic information leakage in the first step of economic information storage. Of course, just in case, we still apply artificial intelligence monitoring and early warning to the blockchain storage server in the big data platform to ensure the security of economic data information [24]. In the artificial intelligence algorithm monitoring and early warning system, artificial intelligence algorithm technology is used to calculate the prosperity index of economic development, the prosperity index is an annual ranking based on factors such as wealth, economic growth, personal well-being, and quality of life, and artificial intelligence recognition technology is used to identify potential risks in the process of economic development, such as financial risks, management risks, fiscal risks, and industrial risks; artificial intelligence early warning technology can intelligently prevent and control some risks and send signals

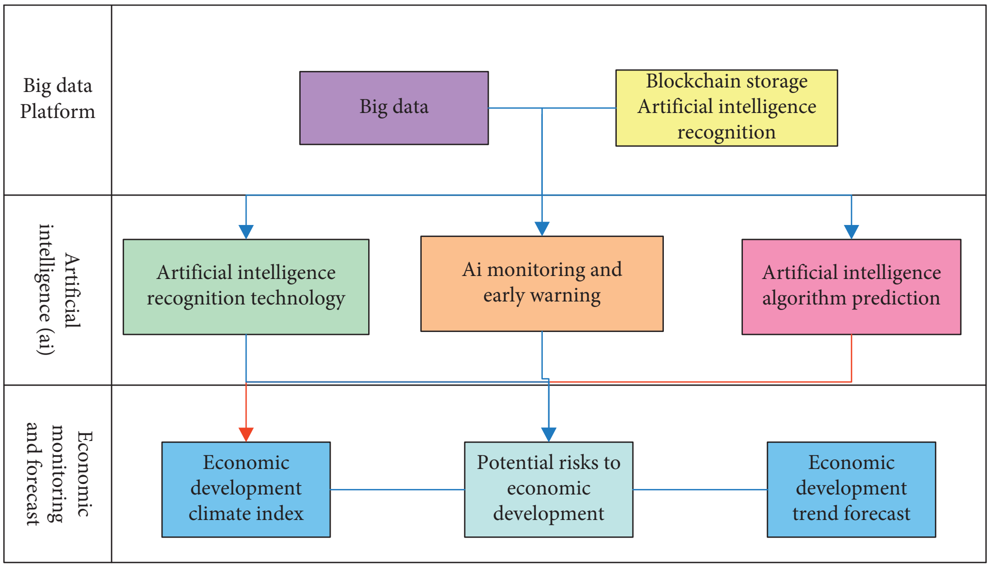

Big data Platform
Big data
Blockchain storage Artificial intelligence recognition
Artificial intelligence (ai)
Artificial intelligence recognition technology
Ai monitoring and early warning
Artificial intelligence algorithm prediction
Economic monitoring and forecast
Economic development climate index
Potential risks to economic development
Economic development trend forecast

Figure 5: Construction of monitoring and early warning platform.

through certain alarm systems to let us know potential risks and make corresponding countermeasures.

In order for artificial intelligence to more accurately predict the trend of economic development, we have built a new big data platform, the structure of which is shown in Figure 6.

This big data platform contains all the economic data, which is complicated. Therefore, in the big data, it is necessary to have an internal data distribution system to divert and manage the economic big data. The economic data will not be unorganized, so data is split on this big data platform, and a series of identifications are performed on the artificial intelligence platform to promote the speed and accuracy of economic data analysis by the intelligent platform. Although artificial intelligence technology has brought profound changes to the production and life of human society, it will affect not only the industrial structure but also the consumption structure, and it will also have a positive effect on the steady development of the intelligent economy. However, the collection, storage, and use of data may cause information leakage, which may cause serious privacy issues. Therefore, we need to use artificial intelligence to predict the risks in the smart economy to reduce the economic information leakage in the information age [25]. We have integrated artificial intelligence, blockchain technology, and big data into the economic monitoring and early warning platform. In order to verify the feasibility of this platform in the economic field, we have carried out a series of experiment analysis.

3. Experiments and Analysis of Smart Economy Big Data Monitoring and Early Warning Platform

3.1. Forecast of Potential Economic Risks. In this experiment, we will use a newly constructed platform to repredict the economic risk events of a listed company in the past 25 years. We first counted the various types of economic risks faced by the company's economic field (including entities and networks) in the past 25 years, including the risks that the company avoided and undetected risks. The risk events are shown in Table 1.

The above is the number of economic risk events that the company has experienced in 25 years. It is understood that, because of the huge scale of economic data, in the previous risk prediction system, only part of the split time was avoided, so some of the risks were recorded after the occurrence for future development of the company. However, due to the rapid development of the information technology era, the expansion of the company's business scale, and the dramatic increase in the scale of economic data, the risk prediction system cannot take into account all economic data, so only part of the risk is avoided. So this time we use the newly constructed system to bring the economic data of the past 25 years into the newly constructed platform to monitor how many economic risk events can be detected by this platform and compare them with the risk events discovered by the past forecasting system. The result is shown in Figure 7.

We can see in Figure 7 that the company's original early warning system has many deficiencies in the early warning of economic risks, and its accuracy of risk prediction is not high, so the company can avoid very few risks, which has caused a great deal to the company. The new risk early warning system studied in this article analyzes and recognizes the economic information data of the original 25 years. The detected risk events are close to the total number of risk events counted in the past 25 years. Its risk prediction accuracy rate can reach 98.21%.

At the same time, we also separately counted the number of events predicted by the new monitoring and early warning platform for real economic risks and network economic risks. The statistical structure is shown in Table 2.

Judging from the data in the table, it is more sensitive to the risk prediction of the network economy. The basic network economy risks have been predicted. Therefore, the newly constructed economic early warning system still has a

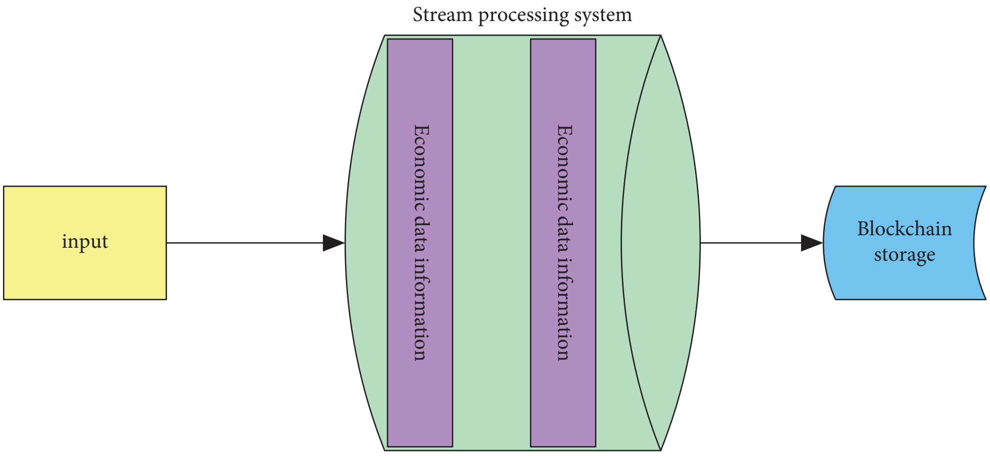


Figure 6: The newly constructed big data platform.

Table 1: The company's economic risk events.

	Tangible economy	Cybereconomy	Total
Corporate financial risk	10	16	26
Corporate management risk	4	24	28
Corporate industry risk	6	5	11
Economic information leakage risk	16	40	56

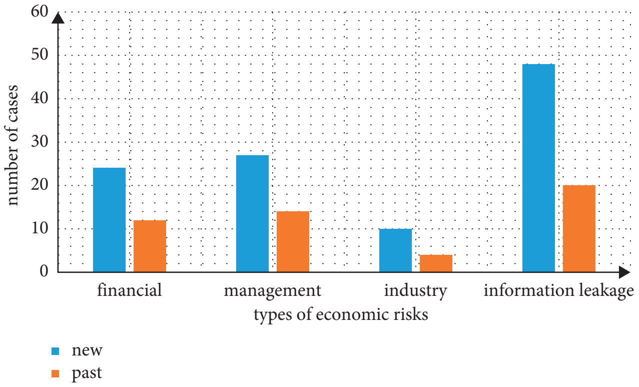


Figure 7: Comparison chart of economic risk prediction.

Table 2: The number of forecasted events for real economic risks and network economic risks.

	Tangible economy	Cybereconomy	Total
Corporate financial risk	9	15	24
Corporate management risk	3	24	27
Corporate industry risk	5	5	10
Economic information leakage risk	8	40	48
Total	25	84	109

high risk prediction accuracy, and, for every risk in the forecast, an alert will appear to remind us to take preventive measures to reduce the company's economic losses.

3.2. Early Warning Experiment and Analysis of Economic Development Prosperity Index. The economic development prosperity index is a barometer of economic development. With 100 as the critical value, the value is between 0 and 200. The confidence index is higher than 100, indicating that the economy is in a prosperous state and the economy is developing in a good direction. The confidence index is lower than 100, indicating that it is in a downturn and the economic operation is developing in an unfavorable direction. This experiment needs to use the artificial intelligence particle swarm algorithm in the newly constructed platform to predict the company's economic development boom early warning index for the next five years based on the company's past economic development. Table 3 shows the prosperous index of the company's economic development in the past 20 years.

In order to verify the economic development prosperity index of the system in this paper, we use the prosperity index from 2001 to 2014 as the analysis sample in the big data platform. The company's early warning prosperity index is compared with the data in Table 3. The calculated early warning prosperity index is shown in Figure 8.

In Figure 8, there is an optimization process of particle swarm algorithm, which is to reduce the calculation error of the early warning index, so that it can improve the accuracy of the early warning index [26]. We can see the comparison of the data in Figure 8, and we can find that basically the displayed prosperity index is greater than 100, which is similar to the development early warning index recorded in Table 3, so the company's economy is developing steadily. However, in 2019, there is a big error with the data in Table 3. It is because 2019 is affected by the company's internal factors. This is a factor that the platform cannot take into account, which leads to a big error.

3.3. Experiment Summary. Through experiment 1, we can see that the new intelligent platform is very sensitive to economic risk monitoring. Comparing its predicted data with actual data, basically all risk events can be accurately predicted, especially in the risk forecast of the network economy which can basically reach 99.2% of the forecast, while the accuracy of the real economy's forecast is not very high. It may be because the information and data of the real economy need to be manually counted in the data platform. The data records made here may not be very accurate, so the sensitivity of this platform to the risk prediction of the real economy is not high; the second experiment is to provide early warning and prediction of the economic development prosperity index of this platform, and it can be found that this platform is very effective. The accuracy of the early warning of the economic development prosperity index is also very high. Through the verification of Experiment 1 and Experiment 2, the economic management big data platform built in this article for monitoring and early warning of artificial intelligence algorithms has high accuracy in predicting economic risks, and the prediction and calculation of the prosperity index of economic development are also very high, but, in the forecasting of the early warning prosperity index, the instantaneous emergent factors will also be inadequately considered, causing great errors.

4. Discussion

This article first explains the artificial intelligence algorithm and establishes a theoretical basis for the later platform construction. Different types of risks in the economic field have a huge impact on economic development. Especially in the era of advanced information technology and rapid development of the network economy, the process of global economic integration is accelerating. Once economic risks appear, they will have a series of economic losses; that is, the world economy nowadays affects the whole body. Therefore, a more intelligent monitoring and early warning system is needed to monitor economic development. In addition, the development of the information age has led to a dramatic increase in the scale of economic data, so it is necessary to analyze and protect the security of these data with the help of advanced science and technology such as big data and blockchain.

Table 3: The company's economic development index in the past 20 years.

Year	Index	Year	Index
2001	120	2011	122
2002	110	2012	102
2003	102	2013	156
2004	87	2014	167
2005	92	2015	146
2006	99	2016	134
2007	100	2017	124
2008	105	2018	120
2009	107	2019	89
2010	118	2020	86

This article discusses the construction of an economic management big data platform with artificial intelligence algorithm monitoring and early warning. The previous economic early warning system has been improved to make it more suitable for the era of the prevailing network economy. In addition, the economic risk early warning system and the forecasting functions of economic development trends have been integrated on this platform, hoping to simplify the economic field's economic data. The repeatability of the analysis simplifies the work of economic data analysis. Improving the work efficiency in the economic field is to share the big data platform for risk prediction and economic development trends in the economic field to realize the corresponding function of the function. The newly constructed economic management big data platform relies on the blockchain storage technology, which can largely ensure the security of economic data and escort the development of economy.

This article verifies through experiments that the artificial intelligence algorithm monitoring and early warning economic management big data platform constructed in this article has high accuracy for economic risk prediction in the economic field and is especially suitable for the current era of network economy and the risks to the network economy. Forecasting is more sensitive than the risk prediction of the real economy, so I think this economic management big data platform needs to be improved. The risk prediction of the real economy can reach the same accuracy as the accuracy of the risk prediction of the network economy, accurately ensuring the smooth operation of the real economy. This article also has high accuracy for the economic development prospects' prosperity index. Therefore, although the economic management big data platform constructed in this article is lacking in the prediction of the real economy, it has high accuracy in other aspects. At the same time, the

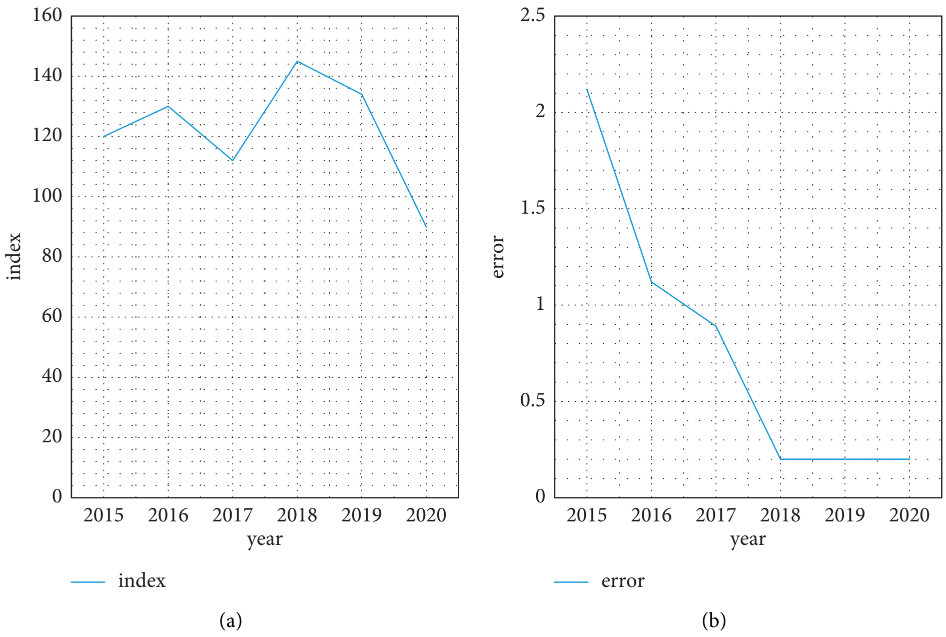


FIGURE 8: Early warning prosperity index. (a) Prosperity index. (b) Particle swarm optimization process.

blockchain technology in the platform can ensure the security of economic data information and improve the security protection index of economic information, which is very suitable for the current era of network economy.

5. Conclusions

This article describes the mobile Internet technology and explains how to ensure the security of economic data and information with artificial intelligence algorithms and blockchain technology. There are many factors that affect economic development, so the artificial intelligence algorithm monitoring and early warning economic management big data platform that this article studies combines big data and blockchain technology to strengthen the security of economic data storage to reduce the risk of information leakage. Then artificial intelligence early warning technology and recognition technology are used to identify the abnormal fluctuations of economic data in the cycle, and algorithm technology is used to calculate the economic development prosperity index, hoping to predict the economic development trend. This paper conducts experiments on this platform and it is found that the research in this paper is successful. The use of artificial intelligence to monitor and early-warn the various influencing factors in the economic field can find the potential factors affecting economic development with the greatest probability and make early response measures. However, the economic management big data platform constructed in this article is still slightly insufficient in the forecasting function of the real economy and needs to be improved. In this paper, although artificial intelligence algorithm warning technology is used in the construction of the big data platform for economic management, uncertain factors are still not taken into account. It is hoped that this aspect can be overcome in future research. Is the economic management of the big data platform more advanced?.

References

[1] E. Mas, D. Felsenstein, L. Moya, A. Y. Grinberger, R. Das, and S. Koshimura, "Dynamic integrated model for disaster management and socioeconomic analysis (DIM2SEA)," *Journal of Disaster Research*, vol. 13, no. 7, pp. 1257–1271, 2018.

[2] Y.-l. Qi, "Construction economic cost management under the market economy," *Procedia Economics and Finance*, vol. 2019, no. 2, pp. 156–160, 2021.

[3] O. M. Pisareva, "Goal-setting model in multilevel state strategic management of socio-economic development," *Economics of Contemporary Russia*, vol. 23, no. 1, pp. 52–76, 2021.

[4] F. Lehrstuhl, "Big data and complexity: is macroeconomics heading toward a new paradigm?" *Journal of Economic Methodology*, vol. 92, no. 3, pp. 1–20, 2017.

[5] G. Lei, Z. Ning, W. Hou, B. Hu, and P. Guo, "Quick answer for big data in sharing economy: innovative computer architecture design facilitating optimal service-demand matching," *Automation Science and Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 1494–1506, 2018.

[6] S. Holm and T. Ploug, "Big data and health research—the governance challenges in a mixed data economy," *Journal of bioethical inquiry*, vol. 14, no. 11, pp. 1–11, 2017.

[7] F. G. N. Li, C. Bataille, S. Pye, and A. O'Sullivan, "Prospects for energy economy modelling with big data: hype, eliminating blind spots, or revolutionising the state of the art?" *Applied Energy*, vol. 239, pp. 991–1002, 2019.

[8] V. Kisimov, D. Kabakchieva, A. Naydenov, and K. Stefanova, "Agile elastic desktop corporate architecture for big data," *Cybernetics and Information Technologies*, vol. 20, no. 3, pp. 15–31, 2020.

[9] X. Xiang, Q. Li, S. Khan, and O. I. Khalaf, "Urban water resource management for sustainable environment planning using artificial intelligence techniques," *Environmental Impact Assessment Review*, vol. 86, Article ID 106515, 2021.

[10] N. Man, K. Wang, and L. Liu, "Using computer cognitive atlas to improve students' divergent thinking ability," *Journal of Organizational and End User Computing*, vol. 33, no. 6, pp. 1–16, 2021.

[11] X. Yang, H. Li, L. Ni, and T. Li, "Application of artificial intelligence in precision marketing," *Journal of Organizational and End User Computing*, vol. 33, no. 4, pp. 209–219, 2021.

[12] M. Safa and L. Hill, "Necessity of big data analysis in construction management," *Strategic Direction*, vol. 35, no. 1, pp. 3–5, 2019.

[13] Y. Wu, S. Zheng, and J. Luo, "A Case Study of Chongqing," in *Proceedings of the 20th International Symposium on Advancement of Construction Management and Real Estate*, pp. 51–59, Singapore, 2017.

[14] J. Luo, "Research on the construction of tutorial system in rural teacher support service system: based on the practice of bijie and anshun in Guizhou Province," *DISP*, vol. 24, no. 95, pp. 52–55, 2018.

[15] Y. Wu, S. Zheng, and J. Luo, "Research on the application of BIM technology in tunnel project construction," in *Proceedings of the 20th International Symposium on Advancement of Construction Management and Real Estate*, pp. 391–404, Singapore, 2017.

[16] S. Mizuno, Y. Fujisawa, and N. Yamaki, "Construction of a comprehensive analysis platform for typology and its application," *Journal of Japan Industrial Management Association*, vol. 68, no. 2, pp. 99–108, 2017.

[17] K. Lengieza, "What we mean when we say platform inside the technology that's shaping the future of construction management," *Engineering news-record*, vol. 282, no. 6, p. 54, 2019.

[18] H. Lu and X. Xu, "Artificial intelligence and robotics," *Studies in Computational Intelligence*, vol. 752, pp. 267–275, 2018.

[19] Z. Yang and J. Wang, "A hybrid forecasting approach applied in wind speed forecasting based on a data processing strategy and an optimized artificial intelligence algorithm," *Energy*, vol. 160, pp. 87–100, 2018.

[20] E. Sharma, R. C. Deo, R. Prasad, and A. V. Parisi, "A hybrid air quality early-warning framework: an hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms," *The Science of the Total Environment*, vol. 709, pp. 135934.1–135934.23, 2020.

[21] Z. Dong, J. Wei, X. Chen, and P. Zheng, "Face detection in security monitoring based on artificial intelligence video retrieval technology," *IEEE Access*, vol. 8, no. 99, pp. 63421–63433, 2020.

[22] A. Zl, A. Dc, A. Rl, and B. Aa, "Artificial intelligence for securing industrial-based cyber–physical systems," *Future Generation Computer Systems*, vol. 117, pp. 291–298, 2021.

[23] A. H. Zaji, H. Bonakdari, and B. Gharabaghi, "Applying upstream satellite signals and a 2-D error minimization algorithm to advance early warning and management of flood water levels and river discharge," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 902–910, 2019.

[24] Z. Ye, J. Yang, N. Zhong, X. Tu, J. Jia, and J. Wang, "Tackling environmental challenges in pollution controls using artificial intelligence: a review," *The Science of the Total Environment*, vol. 699, pp. 134279.1–134279.28, 2020.

[25] X. Yang, P. Sui, X. Zhang et al., "Environmental and economic consequences analysis of cropping systems from fragmented to concentrated farmland in the North China Plain based on a joint use of life cycle assessment, emergy and economic analysis," *Journal of Environmental Management*, vol. 251, pp. 109588–109588.12, 2019.

[26] N. Badasyan, "Project feasibility analysis economic model for private investments in the renewable energy sector," *Built Environment Project and Asset Management*, vol. 8, no. 2, pp. 215–230, 2018.

For Consortium Blockchain, a Novel Semifragile Consensus Algorithm Based on Credit Space

Laxmi, *Department of Computer Science Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, laxmi3349@gmail.com*

Biraja Nayak, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, birajanayak21@gmail.com*

Namrata Khamari, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, namrayakhamari@outlook.com*

Susmita Mohapatra, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, susmitamohapatra963@gmail.com*

Abstract

Nowadays, blockchain is known as a new generation of secure information technologies for realizing business and industrial sustainability, and consensus algorithm is the key technology of blockchain. In order to solve the problem of "oligarchy" nodes and excessive punishment for nodes in existing credit consensus algorithms, a novel semifragile consensus algorithm based on the credit space for consortium blockchain is proposed in this paper. Firstly, the accounting node selection mechanism based on credit space is proposed. The credit value of the node is calculated according to a novel credit evaluation model, and then the credit space of the node is allocated according to the size of the credit value. Afterward, a random algorithm is used to select the accounting node in the credit space. This mechanism effectively inhibits the generation of "oligarchy" nodes and maintains the enthusiasm of nodes. Secondly, this paper proposes a semifragile hierarchical punishment mechanism, which punishes the malicious nodes with severe measures and gives the nonmalicious nodes the opportunity to continue participating in the consensus. So, this semifragile punishment mechanism solves the problem of excessive punishment of nodes. Experimental simulation results demonstrate that the proposed consensus algorithm has randomness while maintaining the credit incentive among nodes. In addition, the node's punishment mechanism is more reasonable. This algorithm has better security and can be well applied to consortium blockchain scenarios.

1. Introduction

In 2008, Satoshi Nakamoto publicly published Bitcoin [1]. Afterward, with the crazy of Bitcoin, blockchain as a core technology of Bitcoin has received extensive research attention [2]. Blockchain has the characteristics of decentralization, hard tamperability, traceability, and transparency, which solves the data monopoly and security problems current in the existing centralized platform [3]. At the same time, many studies have found that blockchain has many innovative applications in the field of IoT and sensor networks. For instance, Satapathy et al. [4] proposed a secure architecture based on open blockchain, which can solve some of the challenges in IoT applications, like issues with confidentiality and privacy of data; Mrinal et al. [5] proposed a blockchain-based wireless sensor network for secure vehicle tracking, reducing the need for an Internet connection and eliminating the use of continuous GPS tracking; that is, it can effectively protect the privacy of commuters and the security of collected data. Therefore, blockchain is known as a new generation of secure information technology. As the core part of the blockchain system, the consensus algorithm is the mechanism for each node of the blockchain to reach consensus on the block information of the whole network [6]. More precisely, it can ensure whether the latest block is correctly added to the blockchain. It is worth mentioning that the performance efficiency and security of the entire blockchain system will be affected by the merits of the consensus algorithm [7]. Similarly, consensus algorithm has always been the key technology of decentralized system, which is widely used in resource-constrained edge computing fields. For instance, Zeng et al. [8] proposed a scheme by utilizing the idle resources in volunteer vehicles to handle the overloaded issues in VEC servers; the scheme can reduce the offloading cost of vehicles and improve the utility of VEC servers; Zeng et al. [9] proposed a new vehicle edge computing framework based on software-defined networks, which introduces the reputation to measure the contribution of each vehicle. The proposed scheme not only brings more benefits to the edge server side but also reduces the average delay a lot.

Generally, different blockchain frameworks use different consensus algorithms. In summary, a common classification divides blockchain into three categories, including public blockchain, consortium blockchain, and private blockchain [10]. The number of nodes in the public blockchain is large, so the transaction speed will be slower. On the contrary, there are fewer nodes in private blockchain and consortium blockchain than in public blockchain, and the transaction speed will be faster [11]. However, the permissions in the private blockchain are controlled by a few nodes, which deviates from the original intention of decentralization [12]. Compared with the private blockchain, the permission design requirements in the consortium blockchain are more complicated and more credible. Now, relevant researches show that the consortium blockchain has more practical value in the fields of IoT applications and medical scenarios. For example, Thomas et al. [13] proposed an anonymous identity and access control system based on consortium blockchain, which improves the security of cross-domain identity authentication in the Internet of Things; Huang et al. [14] proposed a medical data privacy protection and safe sharing scheme based on consortium blockchain, which can effectively ensure the safety of patients' medical information and can safely share information.

At present, the consensus algorithm of consortium blockchain is mainly represented by the Practical Byzantine Fault Tolerance (PBFT) protocol [15]. PBFT has a high transaction speed; however, with the number of nodes increasing, the network overhead of PBFT will increase rapidly, and the consumption of computing power will be high [16]. Moreover, PBFT selects the leader node according to the continuous switching of view number, which may select malicious nodes as the leader node, resulting in poor system security [17]. As such, in order to solve the problem of malicious nodes becoming accounting nodes, researchers have proposed a credit mechanism to generate accounting nodes. The credit value was calculated on the basis of the node's performance in the system, and the node with a higher credit value preferentially became the accounting node [18]. For instance, Li et al. [19] proposed a consortium consensus algorithm based on credit (CCAC), which calculated the credit value of nodes by the contribution of node participation consensus, and selected a node to become an accounting node in turn according to the size of node credit. Notably, a consensus mechanism based on credit reduced the consumption of algorithm computing power and improved the efficiency of consensus. However, this is not effective for the node with a small credit value and easily leads to low enthusiasm of nodes. Wang et al. [20] proposed a proof of work algorithm based on credit model (CPoW) and designed a node credit model based on BP neural network, which effectively reduced the huge resource consumption of repeated calculation in the production process of new blocks. Unfortunately, generating new blocks according to the order of credit value was easy to produce "oligarchy" nodes. Li et al. [21] proposed a dynamic hierarchical Byzantine fault-tolerant consensus mechanism based on credit (DHBFT). The presented reward and punishment plan could effectively reduce the possibility of malicious nodes becoming the leader node, but it could easily cause node with high credit values to be selected as the master node, which lacks fairness and easily causes other nodes to be less motivated. Liu et al. [22] proposed a master-slave multichain blockchain consensus mechanism based on reputation, which introduced credit value evaluation into the consensus mechanism based on proof of stake. In addition, it designed a joint consensus mechanism that integrates multiple consensus mechanisms, which improved the throughput of the transaction and ensured the consistency and nontamperability of the data. However, the punishment for all malicious nodes was too heavy, resulting in nodes being unable to normally participate in the consensus for a long time. Bugday et al. [23] proposed a reputation-based consensus group learning model to calculate the credit value based on the weight value of all nodes in the trust committee, which could effectively avoid malicious nodes, but the weight value of malicious nodes is large. Once a node had malicious acts, the credit value of this node would fall to a very low level, and it was difficult to continue to join the consensus. Huang et al. [24] proposed a credit-based proof of work mechanism for IoT devices, which improved security and enhanced transaction efficiency. Similarly, the punishment for malicious nodes was to reduce the credit value directly to a negative value, which made it difficult for nodes to participate in normal consensus.

To sum up, although the existing consensus algorithm based on credit has improved the efficiency and security of consensus, there are still problems that it is easy to generate "oligarchy" nodes and the punishment for nonmalicious nodes is too large. In order to solve the above problems, this paper proposes a semifragile consortium blockchain consensus algorithm based on credit space. The main contributions of this paper are as follows:

(1) An accounting node selection mechanism based on credit space is proposed. A credit evaluation model is formulated to calculate the credit value of the node, and the credit space of the node is allocated based on the credit value. Based on the credit space, an algorithm for randomly selecting accounting nodes is designed. The nodes with large credit space have a high probability of becoming accounting nodes. At the same time, a threshold equation for the number of accounting nodes is set for the problem of "oligarchy" nodes so that the number of times of becoming accounting nodes is limited.

(2) A semifragile hierarchical punishment mechanism is designed. Nodes with good working conditions are

in the *normal* layer, and the nodes with malicious behaviours are placed in the *prison* layer for "custody." Furthermore, we judge whether the node is malicious or nonmalicious; for malicious nodes, the "custody" time will be longer, and for nonmalicious nodes, they can be returned to the *normal* layer beyond the "custody" time. Therefore, the nonmalicious nodes have the opportunity to participate in the following consensus, and this mechanism can reduce the existence rate of malicious nodes.

2. Problem Statement

2.1. Problem of "Oligarchy" Nodes. Among the existing consensus algorithms based on credit, most of the accounting nodes are selected according to the size of the credit value, which is easy to produce "oligarchy" nodes, and the incentive degree for nodes with small credit value is not enough, such as CCAC algorithm [19]. The credit value of each node is calculated after the credit evaluation of the node, then the credit value is sorted from largest to smallest, and an accounting node is selected in this order, which can easily lead to the production of "oligarchy" nodes and cause other nodes to be less motivated. In this paper, we test the proportion of "oligarchy" nodes as accounting nodes in the total consensus times for CCAC to verify the adverse effects of "oligarchy" nodes on the network, and the results are shown in Table 1.

It can be seen from Table 1 that, with the number of consensuses increasing, the number of "oligarchy" nodes becoming accounting nodes also accounts for an increasing proportion, which can easily cause other nodes to be less motivated to work. Therefore, this paper proposes a mechanism for selecting accounting nodes based on credit space, which can effectively inhibit the generation of "oligarchy" nodes and increase the enthusiasm of nodes.

2.2. Problem of Node's Excessive Punishment. In view of the existing consortium blockchain consensus algorithm based on credit, the punishment for malicious nodes is too severe. More precisely, they do not judge whether the malicious behaviour of a node is deliberate or not, and the credit value of the nodes is always severely reduced so that these nodes cannot continue to participate in the consensus, typically such as the consensus algorithm in [22]. A PoS consensus mechanism based on credit value is proposed, and a credit value evaluation method is designed. The punishment equation for the credit value of malicious nodes is as follows:

$$trust_h^i = -trust_{h-1}^i, \tag{1}$$

where $trust_h^i$ represents the credit value of node i at the end of the hth cycle and $trust_{h-1}^i$ represents the credit value of node i at the end of the h-1th cycle. It can be seen from equation (1) that the credit value of the malicious node will be directly reduced to a negative value, making it difficult for the node to continue to participate in the following consensus. Besides, references [23, 24] mentioned in the Introduction also have the same punishment for malicious nodes. Both have too harsh punishments for malicious nodes, and normal consensus cannot be carried out for a long time. In this paper, we compare these algorithms to test the change in credit value of nodes with malicious behaviours, and the experimental results are shown in Figure 1.

It can be seen from Figure 1 that the credit value of malicious nodes in [22] will rapidly decrease from positive value to negative value, which is difficult to continue to participate in consensus for a long time. Although the algorithm in [23] did not reduce to a negative value, the credit value is very close to 0 and cannot compete with the credit value of normal nodes. In [24], the credit value of the malicious node is always below 0, and it is difficult to continue the normal consensus. By comparing the changes in the credit value of nodes with malicious behaviour in these three algorithms, it can be seen that they cannot participate in normal consensus for a long time for nodes with malicious behaviour. Therefore, this paper proposes a semifragile hierarchical punishment mechanism. This mechanism can make it difficult for malicious nodes to participate in consensus again, but nonmalicious nodes can continue to participate in consensus within a short amount of time.

3. Proposed Algorithm

3.1. Credit Evaluation Model. The credit of nodes represents the working performance of nodes in the process of participating in consensus [18]. The credit evaluation model proposed by the CCAC only considers the number of valid and invalid blocks generated by the accounting node and the time required to add on the chain [19]. However, it does not consider the time when the node is passive and offline from the block. In this paper, the credit evaluation of nodes will be carried out according to the four indicators of the number of transactions in the valid block, the time of the chain, the off-chain time, and the generation of invalid blocks. The credit evaluation indicators are given in Table 2.

Combined with these credit indicators, the data will be standardized so that the data can be calculated uniformly. In this paper, the minimum-maximum planning method is used to standardize the data. This method is the linear transformation of the original data, and the maximum max and minimum min will be set. After calculation by the standardized equation, the data range will fall between [0, 1], and then the following credit value is calculated. The computation equation is as follows:

$$i' = \frac{i - \min}{\max - \min}, \tag{2}$$

where max and min are obtained by preprocessing. In particular, the result of preprocessing is based on 100 consensus experiments in this paper. In the process of consensus experiments, the data of these indicators will be obtained, and max and min are the maximum and minimum values of data in each indicator. When an indicator in a node needs to be measured, it is only necessary to put the data into

Table 1: Percentage statistics of "oligarchy" nodes become accounting nodes.

Total consensus number	The proportion of times that nodes become accounting nodes (%)
400	52.4
600	58.3
800	62.8
1000	78.1

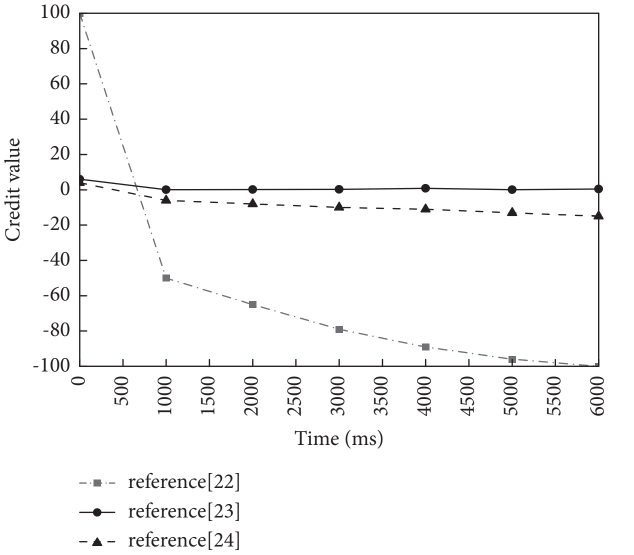


Figure 1: The credit value change diagram of the malicious nodes of each algorithm, recorded in a 6000 ms period. These dots represent the credit value of the malicious node taken every 500 ms.

the (2) to obtain the standardized value: i'_{num}, i'_{time}, $i'_{off-time}$, $i'_{invalid}$.

After getting the standardized data, some data may be positive or negative. Then, these data are added together, and finally, a value x that reflects the quality of the credit value is obtained, as shown in (3). Subsequently, the credit value $C(x_i)$ is to be accumulated or deleted by node i by (4).

$$x_i = i'_{num} + i'_{time} + i'_{off-time} + i'_{invalid}, \tag{3}$$

$$C(x_i) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx, -\infty < x < \infty, \tag{4}$$

where x_i represents the number after processing of the standardized data mentioned above. $C(x_i)$ represents the credit value of node i. Besides, μ is the mean value calculated from the data obtained in the preprocessing, and σ is the variance calculated after preprocessing.

Similarly, the credit value of the accounting node that works hard will be accumulated, as shown in (5). The initial credit value of each node is 1. When node i becomes an accounting node for the first time, its credit value is equal to the initial credit value plus the $C(x_i)$ calculated by (4). Moreover, when node i is selected as the accounting node again, its credit value is the sum of the newly calculated credit value and the previously obtained. The equation for calculating the credit value of node i as an accounting node for the nth time is as follows:

$$Credit_i^n = \begin{cases} 1 + C(x_i)\, n = 1, \\ Credit_i^{n-1} + C(x_i)\, n \neq 1. \end{cases} \tag{5}$$

In contrast, for nodes with malicious behaviour, it will be subtracted after calculating the corresponding credit value, as shown in (6). For the initial malicious nodes, its credit value is the initial credit value minus $C(x_i)$ calculated by (4), that is, the credit value obtained after work. But for the malicious nodes in the consensus process, the credit value subtracts the new credit value from the previous credit value. The calculation equation of the mth malicious credit value of node i is as follows:

$$Credit_i^m = \begin{cases} 1 - C(x_i)\, m = 1, \\ Credit_i^{n-1} - C(x_i)\, m \neq 1. \end{cases} \tag{6}$$

Through the credit evaluation model for node credit evaluation, the nodes in working well condition can get a larger credit value. That is, their opportunity to become an accounting node will be greater, which creates a benign network environment for nodes actively participating in consensus.

3.2. Credit Space. In this paper, the credit space is used as the basis for selecting an accounting node. After a node obtains its credit value, its corresponding credit space is allocated according to the proportion of the credit value in the entire space. That is, the greater the credit value of the node, the greater the allocated space, and the greater the probability of becoming accounting node. For this reason, this method can better motivate nodes to work. Furthermore, the random algorithm also ensures the randomness of the algorithm, and it does not mean that nodes with larger credit space will certainly become accounting nodes. The credit space of node i can be computed by

$$C\_Space_i = \frac{Credit_i}{\sum_{i=1}^{n} Credit_i} \times L, \tag{7}$$

where $Credit_i$ represents the credit value of node i and $\sum_{i=1}^{n} Credit_i$ is the sum of the credit values of each node in this round. L is the total length of the credit space. Figure 2 is a graph of the change of credit space when a node is selected as an accounting node. Figure 2(a) shows the distribution of credit space of a certain round of nodes. Assuming that the pointer is randomly selected to node 3, the credit value of node 3 becomes larger after being selected as an accounting node and packaged block successfully. Obviously, according to the calculation equation of credit space, its credit space will also become larger. So node 3 has a greater probability in the next selection of accounting node, which is like playing a

Table 2: Credit evaluation indicators.

Indicator name	Explanation
i_{time}	The time when the node generates a block
$i_{off-time}$	The time the node leaves the blockchain
i_{num}	The number of transactions in a valid block
$i_{invalid}$	The number of invalid blocks generated by nodes in consensus

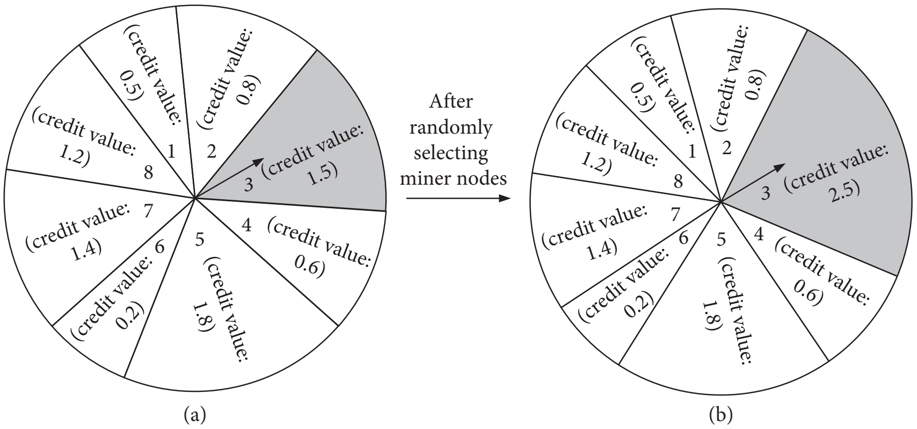


Figure 2: Credit space change diagram. Each sector in the figure represents the credit value of each node, and the pointer is like a turntable to represent random selection. (a) Distribution of credit space of each node in a certain round. (b) Distribution of credit space after selecting miner nodes.

roulette game. The greater the credit space of node 3, the greater the probability of the pointer pointing to node 3. Since the size of the whole credit space is fixed, each node's credit space is calculated according to the proportion of credit value, so the credit space of other nodes will be proportionally reduced.

What is more, after the node is selected as an accounting node, the corresponding packaging work must be completed. The packaging work involves the block structure, which is used to store and verify the credit value. The block structure includes the following:

(1) Blockhead: block version number, hash value of the previous block, timestamp, and random number.

(2) Blockchain time: record the time accounting nodes successfully package block into chain, which helps to verify whether the credit value is accumulated correctly.

(3) The hash value of the block's transaction data: record the transaction data generated by the accounting node of the block.

(4) Credit array in block: record the credit value obtained by nodes on blocks.

(5) Counting array in block: record the number of times a node becomes an accounting node.

Once completed the packaging work, the credit value will be calculated and accumulated in the original credit value. When the next accounting node selection begins, the credit space will be allocated according to the size of the credit value. However, there is a problem at present. Nodes with larger credit values may always be selected as accounting nodes, which leads to the generation of "oligarchy" nodes. Therefore, a threshold for becoming an accounting node is set. When it exceeds the current number threshold, it cannot continue to be selected as an accounting node. The threshold equation is as follows:

$$\tau = \frac{\sum_{i=1}^{Num} num_i}{Num} + t, \tag{8}$$

where τ is a threshold, Num represents the number of nodes, num_i denotes the number of node i becoming an accounting node, and $\sum_{i=1}^{Num} num_i$ represents the total number of times that all nodes become accounting nodes. The threshold calculated by the equation will change with the number of nodes becoming accounting nodes in the whole consensus network. When the threshold of this round increases, nodes may still be selected as accounting nodes. The constant t in the equation will be obtained through experiments, and the specific value is explained in the subsequent experimental part.

3.3. Semifragile Hierarchical Punishment Mechanism.

Generally, the punishment of malicious nodes in the existing credit mechanism is too severe, which directly reduces the credit value of malicious nodes and makes it too difficult to continue to participate in consensus. Consequently, this paper proposes a semifragile hierarchical punishment mechanism. Semifragile refers to the ability to distinguish whether a node with malicious acts is deliberate or nondeliberate. In our algorithm, the nodes judged as nondeliberate are given the opportunity to reparticipate in consensus.

In order to determine whether the malicious node is deliberate or nondeliberate, this paper judges the node by the number of malicious acts. When the number of malicious acts of a node is less than *m*, it is judged as a nondeliberate node, and vice versa. With respect to the value of *m*, this paper counts the number of malicious nodes through many experiments, and the results are shown in Table 3.

It can be seen from Table 3 that the nodes with the number of malicious acts less than or equal to 2 account for 98.53%, basically covering most of the nodes. Therefore, this paper selects 2 as the value of *m*, which is determined as the critical value of nonmalicious nodes.

The specific process of the semifragile hierarchical punishment mechanism is as follows; the general process is shown in Figure 3. First, all nodes will be placed in the *normal* layer, and then the credit space of nodes is calculated to select the accounting node. Moreover, the credit evaluation of the accounting node will be carried out. If the node has malicious behaviour, the node will be placed in the *prison* layer after calculating the credit value. It is worth mentioning that the nodes in the *prison* layer have no chance to be selected as accounting nodes and only the nodes in the *normal* layer have the chance to allocate the credit space to be selected as accounting nodes. The nodes in the *prison* layer will allocate the "custody" time according to the number of malicious acts. During this period, the nodes still need to participate in the data synchronization of the cluster. In particular, if the node is found to have malicious behaviour such as not performing block data synchronization or not working, it will continue to increase the "custody" time. After the time has passed, it is determined whether the node is deliberate or nondeliberate. If the node is a nondeliberate node, it will return to the *normal* layer and give it the opportunity to be selected as an accounting node again. Otherwise, the malicious node will continue to be punished.

About the node's "custody" time, when the number of nodes performing malicious acts increases, the time will increase obviously with the number of times. According to this characteristic, this paper uses the following function:

$$T = e^{x}. \qquad (9)$$

The function is monotonically increasing, where T is the "custody" time and x is the number of malicious acts. It can be seen from (9) that T is monotonically increasing, that is, when x increases, that is, when the number of malicious acts increases, the time increases exponentially. In contrast, for nonmalicious nodes, only one or two malicious activities are performed, and the "custody" time is relatively appropriate. It conforms to the principle of the proposed punishment mechanism, gives a good buffer to the nodes that do not deliberately perform malicious acts, and then gives the opportunity to participate in the consensus.

4. Algorithm Design

Firstly, the credit value of all nodes is initialized to 1. In the beginning, each node is placed in the *normal* layer, and each participating node is numbered. Moreover, the corresponding credit space is allocated according to the credit value of each node. Obviously, the size of the space allocated by each node is the same, and the total space is unchanged. Then the credit array *Cn* and count array *Cc* are constructed. *Cn* is used to store the credit value of the node, and *Cc* is the number of times the storage node has become an accounting node. Thereafter, begin the cycle of selecting the accounting nodes. The process of credit consensus is shown in Figure 4. As shown in Figure 4, the whole process can be divided into four steps: initialization stage, cyclic selection of accounting node stage, constructing block stage, and checking the new block stage.

Table 3: Statistics of the number of malicious acts. The proportion indicates the proportion of nodes with different times of malicious acts in the total nodes.

Number of malicious acts	Proportion of the number of nodes (%)
≤0	96.25
≤1	97.84
≤2	98.53
>2	1.47

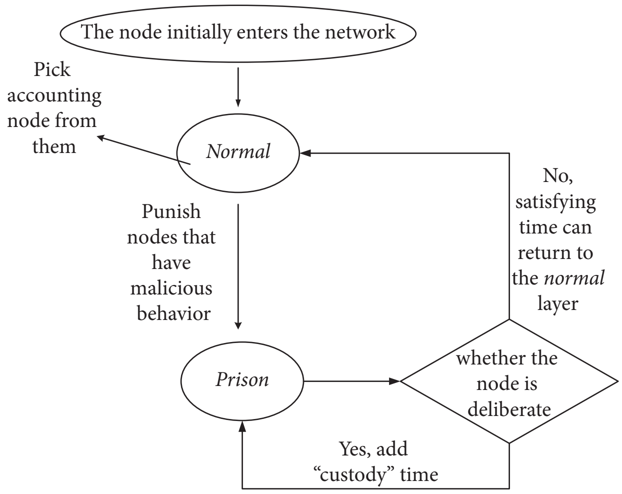

The node initially enters the network
Pick accounting node from them
Normal
Punish nodes that have malicious behavior
No, satisfying time can return to the *normal* layer
Prison
whether the node is deliberate
Yes, add "custody" time

Figure 3: Semifragile hierarchical punishment mechanism. Malicious nodes will judge whether the node is deliberate and take corresponding measures.

4.1. Initialization Stage. In the initial stage, the initial credit value of each participating node in the *normal* layer is set to 1, and the total credit space length is set to 100. The credit space of each node is calculated by equation (7), and the number of participating nodes is assigned. Thereafter, the credit array *Cn* and the count array *Cc* are constructed to store the credit value of the node and the number of nodes becoming accounting nodes, respectively. Algorithm 1 shows how to allocate the node's credit space.

4.2. Cyclic Selection of Accounting Node Stage. Through Algorithm 1, we have obtained the credit space of each node. Then, the algorithm randomly selects the accounting node.

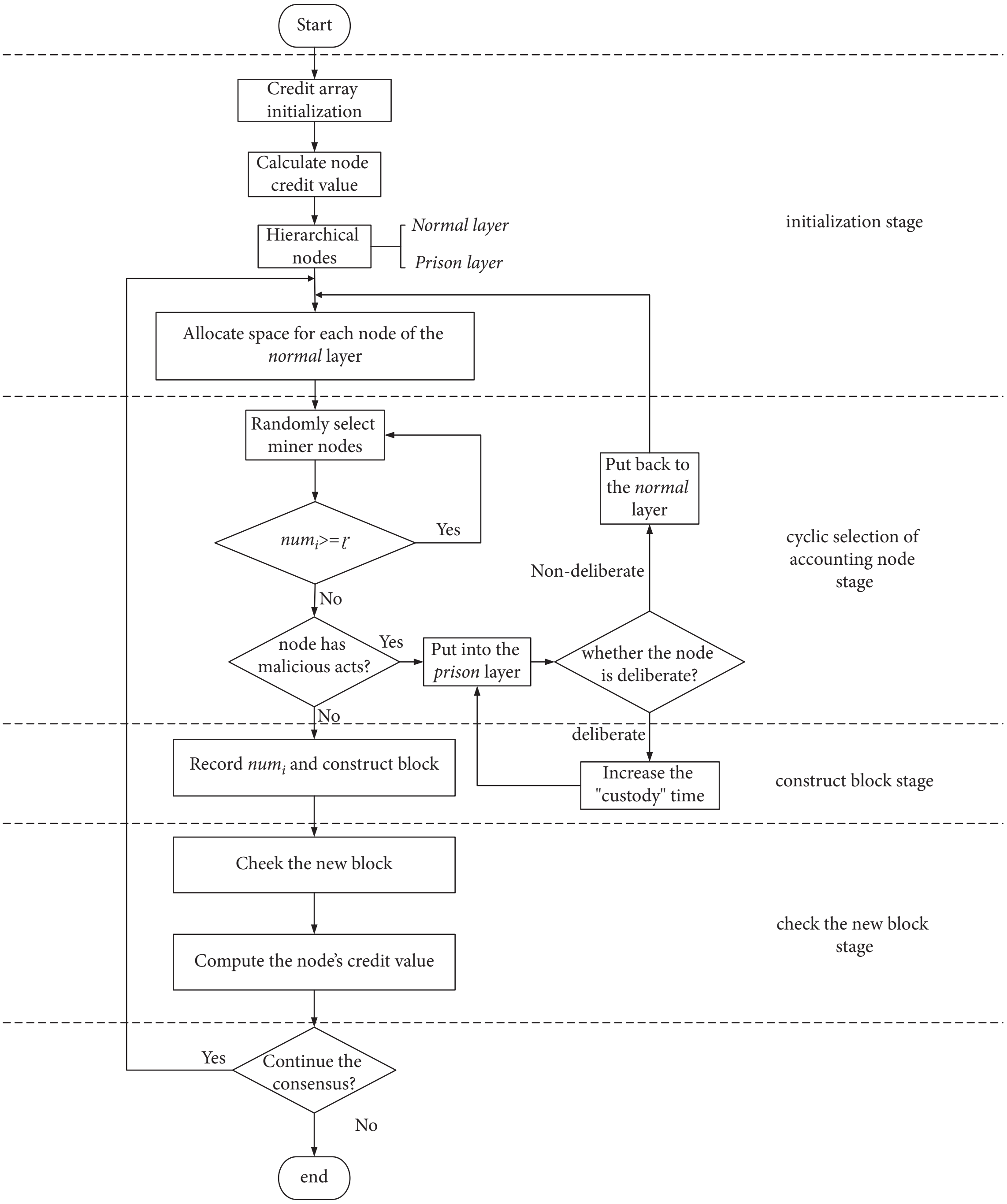

Start
Credit array initialization
Calculate node credit value
Hierarchical nodes
Normal layer
Prison layer
initialization stage
Allocate space for each node of the normal layer
Randomly select miner nodes
$num_i >= \tau$
Yes
No
cyclic selection of accounting node stage
Put back to the normal layer
Non-deliberate
node has malicious acts?
Yes
Put into the prison layer
whether the node is deliberate?
No
deliberate
Record num_i and construct block
Increase the "custody" time
construct block stage
Cheek the new block
Compute the node's credit value
check the new block stage
Yes
Continue the consensus?
No
end

Figure 4: Credit consensus process. This flowchart describes how nodes select accounting nodes and how to punish malicious nodes.

More precisely, each interval represents each node's credit space, and the algorithm selects accounting node by setting a random number and judging which interval the random number falls into. As a result, the node represented by this interval is selected as accounting node. Subsequently, the accounting node will complete the corresponding work and obtain the corresponding credit value. Generally, if an accounting node does not work, the node will be punished beyond the given time and enter the next accounting node's selection. When the next accounting node selection is conducted, the corresponding space will be allocated according to the credit value. If the credit value is larger, it is easier to obtain the packaging right. Since it is randomly selected, there will be nodes with low credit values that get the right to package. In order to avoid the generation of "oligarchy" node, τ is set. If num_i exceeds τ, node i cannot be selected as an accounting node. But it does not mean that i cannot be selected as an accounting node anymore because τ will change with $\sum_{i=1}^{Num} num_i$ in the whole consensus network. When τ becomes larger, node i may still be selected as the accounting node. Algorithm 2 gives the process of randomly selecting an accounting node.

If the node has malicious acts, the node will be placed in the *prison* layer for "custody." "Custody" time will be

```
Input: Cn (node's credit value array)
Output: spaceArray (node's credit space array)
spaceArray[] = {0}; //Initialization of Credit Space Array
  for $i = 1$ to $n$ do //Traversing all nodes
    if JudgePrsion (Cn[i]) //Judge if the node is in the prison layer, if it is, do not allocate space
      continue;
    end if
    if $i = = 0$ //When i is the first node in space
      spaceArray[i]=(Cn[i]/countSum (Cn)) * spaceLength; //Calculating the length of credit space
      continue;
    end if
    spaceArray[i] = (Cn[i]/countSum (Cn)) * spaceLength + spaceArray[i − 1]; //The length of credit space after becoming an accounting node, Cn is an array of normal layers
end for
```

Algorithm 1: Node layering and credit space allocation algorithm.

calculated according to (9). Then, it will determine whether the time has expired. If it has expired, determine whether the node is a malicious node. If it is not a malicious node, it will be released back to the *normal* layer. If it is a malicious node, continue to stay in the *prison* layer for "custody." If it has not expired, it will continue to stay in the *prison* layer. Algorithm 3 gives the penalty mechanism of malicious nodes.

4.3. Construct Block Stage. Once the accounting node is selected, the accounting node will broadcast the constructed block to all adjacent nodes. Afterward, adjacent nodes will receive the new block and broadcast it to the whole network after successful verification. When the block is verified, the block will be added to each node's blockchain copy. After all nodes have received and verified the block, the work of the next block construction will proceed.

4.4. Check the New Block Stage. After selecting the accounting node and completing the related transactions on the block, the node will broadcast the generated block to the whole network and then verify the credit value. Once the verification is correct, the node will obtain the corresponding credit reward. On the contrary, if the node has malicious behaviour, the behaviour will be recorded in the block, and the corresponding credit punishment will be carried out.

5. Experimental Results and Analysis

In our experiments, we use Golang programming language and JetBrainsGoLand 2020.3.4 for the simulation test. First, we use the Go language to write a single-machine multinode platform to simulate the consensus process. Then, we compare the performance of the consensus algorithm proposed in this paper with CCAC algorithm [19], CPoW algorithm [20], and master-slave multichain algorithm [22], and test the number of malicious nodes, punishment mechanism, and the consensus delay of nodes. Finally, the images are drawn according to the experimental data for comparative analysis.

5.1. Threshold Equation Constant Experiment. This experiment is to analyze the value of threshold equation constant. We selected 40 nodes for 600 consensuses and tested the average time consumption to select accounting nodes under different threshold equation constants.

It can be seen from Figure 5 that the average time consumption of selecting accounting nodes with a constant 3 is the least, while the average time consumption of other nodes is relatively high. Therefore, we choose constant 3 as the value of t in the threshold equation.

5.2. Statistics of Accounting Node Number. In this experiment, we test the number of times nodes become accounting nodes to verify the credit evaluation model and the mechanism of selecting accounting nodes. First, set 20 nodes, conduct consensus on them 600 times, and select accounting nodes. The experimental results are shown in Figure 6. As can be seen from the data in Figure 6, each node can become an accounting node in the consensus process. Some nodes have become accounting nodes only 5 or 6 times, and some nodes have become accounting nodes 18 or 20 times. This shows that the proposed algorithm can reflect the role of credit value and ensure the randomness of selecting accounting nodes through credit space.

In order to test whether the threshold τ can better limit the "oligarchy" node, this paper tests 20 nodes, carries out 1500 consensuses on them, selects the accounting node, and then records the number of rounds of threshold change and the highest number of accounting node in this round. The experimental results are shown in Figure 7. As shown in Figure 7, the threshold τ will change with the number of nodes becoming accounting nodes in the whole network, and the number of accounting nodes is also limited to the threshold τ. The number of accounting nodes increases more and more slowly and requires a longer consensus time. This

```
Input: spaceArray (Node's Credit Space Array)
Output: i (accounting node's serial number)
    while (nodeSelect) //nodeSelect is whether to select the miner to complete the identifier,
the initial value is true
        rand.Seed (time.Now().Unix()); //Set random number time seed
        randomSize = randomFloat (0, spaceLeangth); //Random number selected in space
        node = judgeSelect (spaceArray, randomSize); //Determine which node is selected
        if CoutArray [node] ≤ Exceeded //The requirement cannot exceed the threshold
            nodeSelect = false; //The selection is complete, jump out of the loop, otherwise
continue to choose
            Cc[i]++; //Count value plus 1
            return i;
    end if
    end while
```

Algorithm 2: Random selection accounting node algorithm.

```
Input: U (the set of malicious nodes)
Output: prisonArray (prison layer array)
    prisonArray[] = {}; //Initialize the prsion layer
    while (node in U)
    if JudgeMalicious (node) //Determine whether the node is malicious
        time = pow (e, x); //Calculate penalty time
        insert (node, prisonArray, time); //Put the node in jail and record the punishment time
        node++; //Pointer moved to the next malicious node
        continue;
    end if
    if JudgeTimeOut (node) //Determines whether the node penalty time expires
        if (maliciousCount ≤ 2) //Determines whether the node is a malicious node
                remove (node, prisonArray); //Remove the node
                node++; //Pointer moved to the next malicious node
                continue;
            else
                stayPrison (node); //Leave the node in the prison layer
                node++; //Pointer moved to the next malicious node
                continue;
            end if
    end if
    end while
```

Algorithm 3: Punishment algorithm for malicious nodes.

shows that the proposed threshold mechanism can effectively restrain the emergence of "oligarchy" nodes.

5.3. Semifragile Hierarchical Punishment Mechanism Experiment. In order to prove that the punishment mechanism proposed in this paper can effectively avoid malicious nodes destroying the consensus process, we do an experiment to test the number of malicious nodes in different algorithm. The experimental results are given in Figure 8. At first, 1000 nodes are set in the system, and 273 malicious nodes are set and labelled artificially in these nodes. With the increase of consensus times, it can be found that the number of labelled malicious nodes in each algorithm is gradually decreasing, but it should be noted that the number of malicious nodes in the proposed algorithm in this paper has a more obvious decline. From Figure 8, it can be seen that when the 70th consensus is carried out, the number of labelled malicious nodes in the algorithm proposed in this paper is reduced to 32, and the number of malicious nodes in other algorithms is more than that of this algorithm. This shows that the credit evaluation model proposed in this paper will gradually reduce the credit value of the malicious nodes. At the same time, the hierarchical punishment mechanism will also punish malicious nodes, which further restrains malicious nodes from doing evil. With the increase in the number of consensuses, the probability of selecting the malicious nodes as accounting nodes will be greatly reduced; as such, it will make the blockchain system more safe and reliable.

In order to further test the performance of the proposed semifragile hierarchical punishment mechanism, this paper does an experimental test to determine the malicious

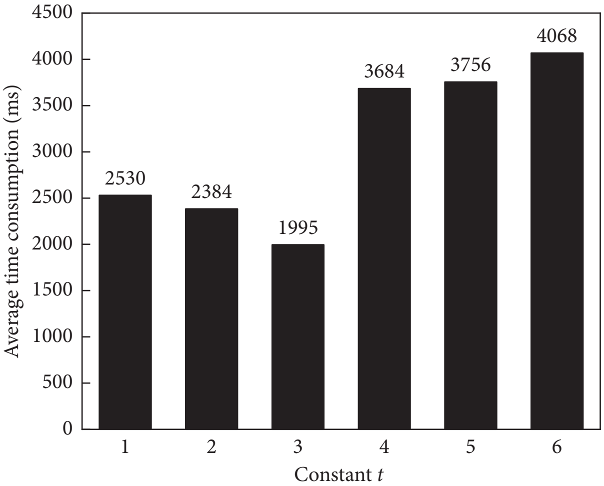


Figure 5: The experimental statistics of the constant *t*. The average time consumption of each constant in 600 consensuses is recorded.

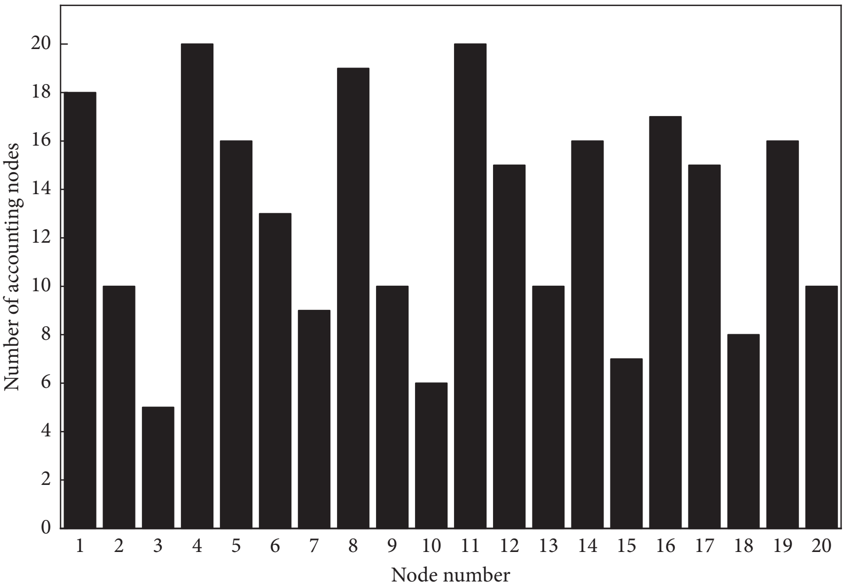


Figure 6: Statistics of the number of accounting nodes. The number of times each node becomes an accounting node in 600 consensuses is recorded.

behaviours of nodes. Firstly, a deliberate node and a nondeliberate node are marked, respectively, and they are placed in the *prison* layer. According to the proposed mechanism, nondeliberate nodes will be put back to the *normal* layer over time to continue to join the consensus, we record their credit values to observe the work of the node, and the results are shown in Figure 9.

It can be seen from Figure 9 that the credit value of a nondeliberate node decreases after malicious acts. After putting it into the *prison* layer, the credit value remains unchanged. If it is put back to the *normal* layer after exceeding the "custody" time, it can normally participate in the consensus. However, the credit value of the deliberate node declines after committing malicious acts. It is worth mentioning that if the deliberate node continues to commit malicious acts in the *prison* layer, the "custody" time will be double. It shows that the mechanism gives an opportunity to nondeliberate nodes and does not reduce its credit value to the point of being unable to participate in the consensus. That is, it makes nondeliberate nodes become normal nodes, while deliberate nodes are punished accordingly.

5.4. Consensus Delay. The consensus delay comparison results are shown in Figure 10. As can be seen from Figure 10, with the increase of consensus times, consensus delay increases gradually. The consensus delay of the CCAC algorithm is the lowest, the consensus delay of the proposed algorithm is only higher than that of CCAC, and the consensus delay of the CPoW algorithm is the highest. This is because the proposed algorithm in this paper selects accounting nodes based on the credit space and introduces the hierarchical punishment mechanism, which results in higher delay than CCAC. However, the consensus delay of the algorithm is within an acceptable range, and it does not

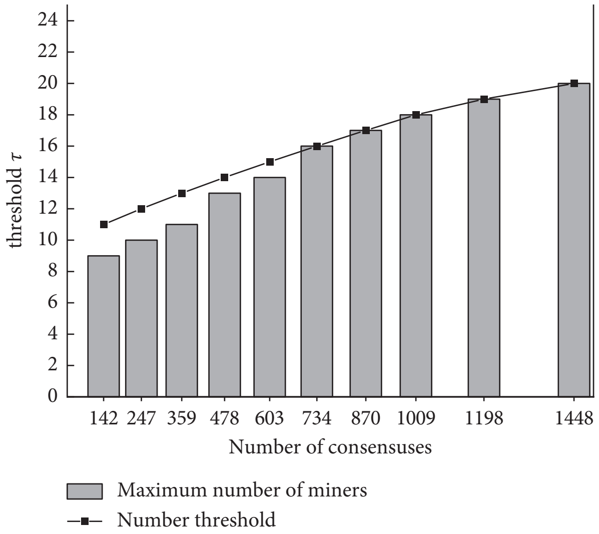


FIGURE 7: The change of the threshold τ. Each interval of the abscissa represents the current number of rounds when the number of thresholds has changed.

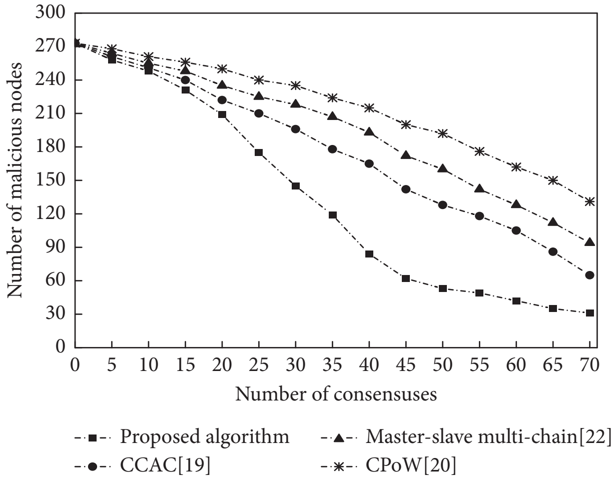


FIGURE 8: The number change of malicious nodes, recorded in 70 consensuses.

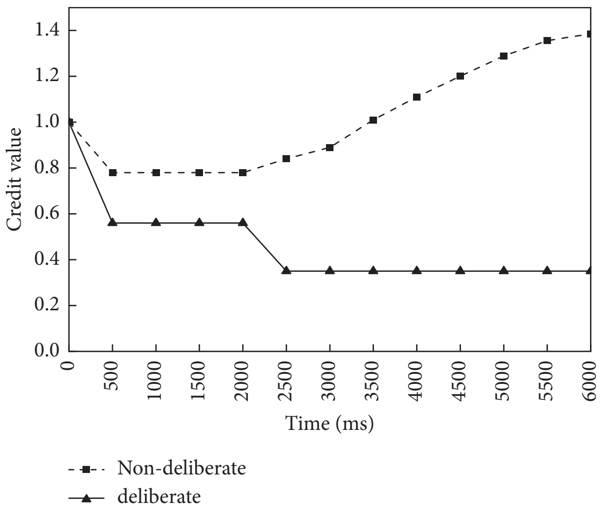


FIGURE 9: The credit value change of malicious nodes, recorded in a 6000 ms period.

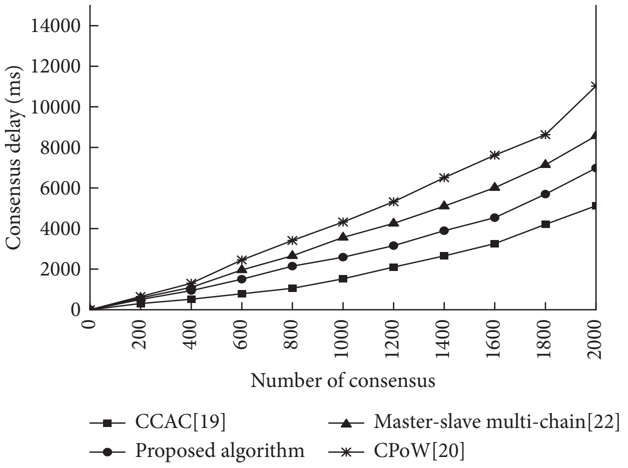


Figure 10: Comparison of consensus delay, recorded in 2000 consensuses.

affect the normal operation of the entire blockchain system. Compared with the proposed algorithm in this paper, CPoW is more difficult to solve the hash problem with the increase of blockchain length. Therefore, CPoW will consume a lot of computing power and have a high consensus delay. Because the master-slave multichain algorithm is based on the PoS algorithm, compared with CPoW, it saves a lot of energy consumption without mining. However, compared with the algorithm proposed in this paper, its consensus process is more complex and prone to bifurcation. Thus, the consensus delay is higher than that of the proposed algorithm in this paper.

5.5. Limitation. It can be seen from the results of the above experiments that the algorithm proposed in this paper can suppress the "oligarchy" nodes and deal with the deliberate nodes very well, but there are still some limitations. In this part of the credit evaluation model, the evaluation indicators set are not complete enough, so the evaluation of the nodes may not be comprehensive enough. This is a relatively limited point, and there is room for improvement in the future.

6. Conclusion

This paper proposed a semifragile consortium blockchain consensus algorithm based on credit space. According to the working situation of the node, we designed a credit evaluation model to calculate the credit value of the node and allocated the credit space. Besides, we proposed a randomly select mechanism for the accounting node based on the credit space, which solved the problem of insufficient incentive in the consensus algorithm and ensured the randomness of the node to become an accounting node. The experimental results show that the consensus mechanism in this paper has randomness while ensuring credit incentive; it enhances the security of the algorithm. In addition, it is more reasonable for the node penalty mechanism and has better performance in consensus efficiency, which is suitable for consensus in the consortium blockchain. Nonetheless, the algorithm still has shortcomings in the determination of malicious nodes and the design of the "custody" time equation of the semifragile hierarchical punishment mechanism. The next step will continue to conduct in-depth research on these two aspects.

References

[1] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," 2008, https://bitcoin.org/bitcoin.pdf.

[2] S. Zhang and .-H. Lee, "Analysis of the main consensus protocols of blockchain," *ICT express*, vol. 6, no. 2, pp. 93–97, 2020.

[3] M. X. Du, X. F. Ma, Z. Zhang, W. Xiangwei, and C. Qijun, "A Review on Consensus Algorithm of blockchain," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2567–2572, Banff, AB, Canada, October 2017.

[4] U. Satapathy, B. K. Mohanta, S. S. Panda, S. Sobhanayak, and D. Jena, "A secure framework for communication in internet of things application using hyperledger based blockchain," in *Proceedings of the 2019 10th international conference on computing, communication and networking technologies (ICCCNT)*, pp. 1–7, Kanpur, India, July 2019.

[5] M. Mrinal, A. Garg, V. D. Maikandavel, and A. Panja, "Blockchain secured vehicle tracking using wireless sensor network," *Ilköğretim Online*, vol. 20, no. 1, pp. 2472–2480, 2021.

[6] G. T. Nguyen and K. Kyungbaek, "A survey about consensus algorithms used in blockchain," *Journal of Information processing systems*, vol. 14, no. 1, pp. 101–128, 2018.

[7] Y. A. Min, "A study on performance evaluation factors of permissioned blockchain consensus algorithm," *Jouranl of Information and Security*, vol. 20, no. 1, pp. 3–8, 2020.

[8] F. Zeng, Q. Chen, L. Meng, and J. Wu, "Volunteer assisted collaborative offloading and resource allocation in vehicular edge computing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3247–3257, 2021.

[9] F. Zeng, Y. Chen, L. Yao, and J. Wu, "A novel reputation incentive mechanism and game theory analysis for service caching in software-defined vehicle edge computing," *Peer-to-Peer Networking and Applications*, vol. 14, no. 2, pp. 467–481, 2021.

[10] S. M. H. Bamakan, A. Motavali, and A. Babaei Bondarti, "A survey of blockchain consensus algorithms performance evaluation criteria," *Expert Systems with Applications*, vol. 154, no. 9, Article ID 113385, 2020.

[11] S. Pahlajani, A. Kshirsagar, and V. Pachghare, "Survey on Private Blockchain Consensus Algorithms," in *Proceedings of the International Conference on Innovations in Information and Communication Technology (ICIICT)*, pp. 1–6, Chennai, India, April 2019.

[12] A. Zhang and X. Lin, "Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain," *Journal of Medical Systems*, vol. 42, no. 8, pp. 1–18, 2018.

[13] H. Thomas and P. Alex, "Verifiable Anonymous Identities and Access Control in Permissioned Blockchains," Massachusetts Institute of Technology," 2016, http://arxiv.org/abs/1903.04584.

[14] H. Huang, P. Zhu, F. Xiao, X. Sun, and Q. Huang, "A blockchain-based scheme for privacy-preserving and secure sharing of medical data," *Computers & Security*, vol. 99, no. 12, pp. 102010–102023, 2020.

[15] S. J. Alsunaidi and F. A. Alhaidari, "A Survey of Consensus Algorithms for Blockchain Technology," in *Proceedings of the 2019 International Conference On Computer And Information Sciences (ICCIS)*, pp. 1–6, Sakaka, Saudi Arabia, April 2019.

[16] Y. Wu, P. Song, and F. Wang, "Hybrid consensus algorithm optimization: a mathematical method based on POS and PBFT and its application in blockchain," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[17] X. Zheng, W. Feng, M. Huang, and S. Feng, "Optimization of PBFT algorithm based on improved C4. 5," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[18] D. Wang, C. Jin, H. Li, and M. Perkowski, "Proof of activity consensus algorithm based on credit reward mechanism," in *Proceedings of the International Conference on Web Information Systems and Applications*, pp. 618–628, Guangzhou, China, September 2020.

[19] S. Z. Li, L. Huang, X. H. Deng, Z. Q. Wang, and H. W. Liu, "Consortium chain consensus algorithm based on credit," *Application Research of Computers*, vol. 38, no. 8, pp. 2284–2287, 2021.

[20] Z. Wang, Y. L. Tian, Q. X. Li, and X. YANG, "Proof of work algorithm based on credit model," *Journal on Communications*, vol. 39, no. 8, pp. 185–198, 2018.

[21] F. Li, K. Liu, J. Liu, Y. Fan, and S. Wang, "DHBFT: dynamic hierarchical Byzantine fault-tolerant consensus mechanism based on credit," in *Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference On Web And Big Data*, pp. 3–17, Tianjin, China, August 2020.

[22] H. Z. Liu, S. S. Li, W. L. Lv, and S. J. Wei, "Master-slave multiple-blockchain consensus based on credibility," *Journal of Nanjing University of Science and Technology*, vol. 44, no. 3, pp. 325–331, 2020.

[23] A. Bugday, A. Ozsoy, S. M. Öztaner, and H. Sever, "Creating consensus group using online learning based reputation in blockchain networks," *Pervasive and Mobile Computing*, vol. 59, no. 10, pp. 101056–101070, 2019.

[24] J. Huang, L. Kong, G. Chen, L. Cheng, K. Wu, and X. Liu, "B-IoT: Blockchain Driven Internet of Things with Credit-Based Consensus Mechanism"" in *Proceedings of the 2019 IEEE 39th International Conference On Distributed Computing Systems (ICDCS)*, pp. 1348–1357, Dallas, TX, USA, October 2019.

Topology-Aware Strategy for MPI-IO Operations in Clusters

Sanjay Kumar Padhi, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sk.padhi2@gmail.com*

Prasanna Kumar Chhotaray, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pkchhotaray85@gmail.com*

Ashis Singh, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, asish.singh2156@gmail.com*

Premananda Sahu, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, premanandasahu8@live.com*

Abstract

This paper presents the topology-aware two-phase I/O (TATP), which optimizes the most popular collective MPI-IO implementa-tion of ROMIO. In order to improve the hop-bytes metric during the file access, topology-aware two-phase I/O employs the Linear Assignment Problem (LAP) for finding an optimal assignment of file domain to aggregators, an aspect which is not considered in most two-phase I/O implementations. The distribution is based on the local data stored by each process, and its main purpose is to reduce the total hop-bytes of the I/O collective operation. Therefore, the global execution time can be improved. In most of the considered scenarios, topology-aware two-phase I/O obtains important improvements when compared with the original two-phase I/O implementations.

1. Introduction

A large class of scientific applications access a high volume of data frequently during their execution. Scalable solutions for efficient and concurrent access to storage are offered by parallel file systems such as Lustre, PVFS, and GPFS. The scientific applications access these parallel file systems through interfaces such as POSIX and MPI-IO [1] or high-level libraries which are based on MPI-IO. In this paper we target optimizing the implementation of MPI-IO interface inside ROMIO, which is the most popular MPI-IO distribution.

Most parallel applications do the computation and I/O alternatively. During the I/O phase, each process often issues a large amount of small noncontiguous I/O requests to access a common data set. These requests usually cause severe overall I/O performance degradation. In order to optimize the performance of the I/O system, the two-phase I/O algorithm is used to merge small individual requests into larger continuous requests. In this work we focus on improving two-phase I/O technique. We have designed and evaluated the topology-aware two-phase I/O technique in which file data access is not only dependent on the data distribution of each process but also dependent on the mapping of processes to computing resources. The comparison with other version of two-phase I/O shows that an important reduction of the run time can be obtained through our technique.

Cluster systems now are moving towards exascale with the high performance interconnection network and many-core architectures. Such systems are getting more and more hierarchical in their interconnection network and node architecture. Processes have different performance levels when communicating at various hierarchies. It is therefore critical for the MPI-IO libraries to reasonably handle the communication demands during the I/O procedure of high performance computing (HPC) applications on such hierarchical systems. MPI-IO is the predominant I/O standard for HPC applications in clusters. During the collective I/O procedure defined in MPI-IO, multiple aggregators exchange data with specific processes. However current MPI-IO optimization strategies do not take the communication pattern and network topology into consideration. In this work, we have designed the topology-aware two-phase I/O, which can improve the shuffle phase of collective I/O operations by carefully placing the aggregators on proper nodes. We have integrated the node physical architecture with network topology and used graph theory inside MPI-IO library to override the current trivial implementation.

On massively parallel clusters, parallel jobs typically acquire a fraction of the available nodes, which are discontinuous and do not correspond to any regular topology, even when the cluster does. On the other hand for modern machines, contention on specific links limits the communication performance. By suitably assigning processes on proper nodes of clusters, substantial communication and performance improvements on large parallel machines can be achieved. Recently the hop-bytes metric [2, 3], defined as the sum over all the messages of the product of number of hops the message has to traverse and the message size, has attracted much attention. For cluster, this equals the total communication volume. The reason of using the hop-bytes metric is that if the total communication volume is high, then the contention for specific links is also much more likely to increase, and the links would then become communication bottlenecks. During MPI-IO procedure, selection of aggregators and assignment of file domains that taking hop-bytes into account can significantly reduce communication overhead. Although the communication bottleneck caused by link contention is not directly measured by this metric, low values of this metric mean smaller communication overheads. When using this metric, we only have to measure the machine topology; the routing information is not necessary.

This paper is structured as follows. Section 2 introduces the related work. The implementation detail of two-phase I/O is described in Section 3. Section 4 gives the description of the topology-aware two-phase I/O. Section 5 overviews the evaluated application, in addition to the evaluation results that compare the topology-aware two-phase I/O with the original version of two-phase I/O.

2. Related Work

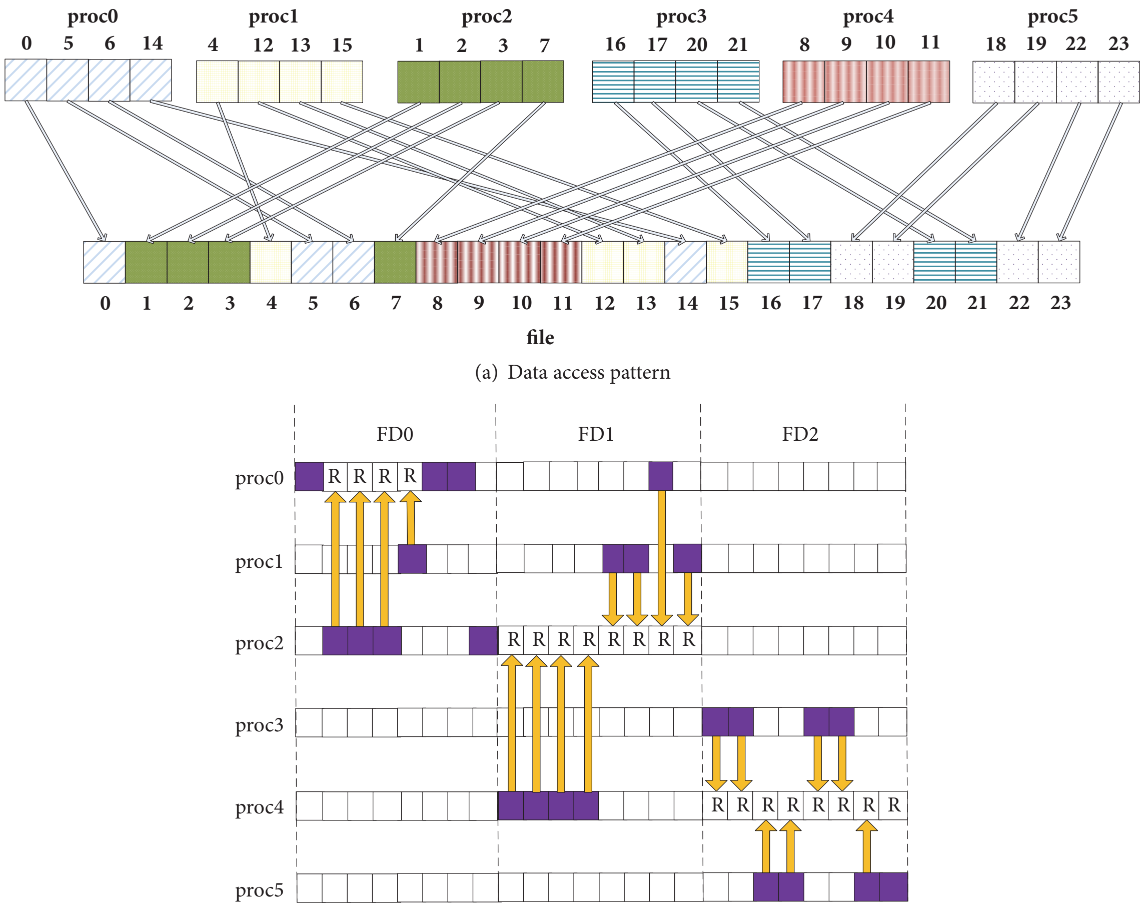
Due to the increasing requirements of applications for data movement to memory or storage, parallel I/O is an active research topic now. From the perspective of file system, highly scalable parallel file systems such as GPFS [4] or Lustre [5] are widely used. At the application level, parallel I/O libraries MPI-IO, which is part of the MPI-2 standard, is commonly deployed. With MPI-IO, collective I/O allows achieving improved performance. Various collective I/O write algorithms are evaluated by Chaarawi et al. [6]. Some researches try to optimize collective I/O with techniques such as automatic collective I/O tuning with machine learning [7] and process placement based on the I/O pattern [8]. Two-phase I/O is the de facto collective I/O algorithm [9]. It adds a shuffle phase in collective I/O phases by aggregating data on a subset of processes (aggregators) before writing it onto the parallel file system. ROMIO [10] is a popular implementation of MPI I/O using two-phase I/O and it has been included in MPICH, Open MPI, IBM MPI, NEC MPI, SGI MPI, and HP MPI. Some researches try to improve the two-phase I/O algorithm [11]. Approaches based on double buffers using multithreading to overlap shuffle phase and I/O phase have been studied in [12, 13]. Properly setting the buffer size and the aggregator number is still an important topic [14]. Finally, placing the aggregators on proper cores is a well-known problem. Certain approaches focus on discovering data locality and using a polynomial-time file domain to aggregator assignment algorithm to minimize communication between computing processes and aggregators [15]. Other researchers try to take the routing mechanism into consideration when issuing sparse data access on BG/Q [16]. Previous IBM supercomputers BG/P adopt a general method designed to increase the I/O bandwidth of collective I/O [17]. Tessier et al. [18] use a different approach which combines an optimized buffering system and a topology-aware aggregators mapping strategy targeting any kind of architecture and being extensible to address new tiers of storage. However their approach does not take link contention into consideration and also does not try to minimize the execution time of all aggregators.

3. Internal Structure of Two-Phase I/O

After reviewing the related work on MPI-IO optimization, we introduce two-phase I/O in detail. As the name indicates, two-phase collective I/O consists of an I/O phase and a shuffle phase. In the I/O phase, contiguous data block transfers are performed from or to the parallel file system. In the shuffle phase, by interprocess communication, small file requests of different processes are grouped in larger ones. Before the two phases, the file region which is contained between the minimum and maximum offsets of all file requests is divided into a configurable number of file domains (FD), and each FD is assigned to a chosen process which is called aggregator. All the data that locates inside an FD is aggregated by the related aggregator which is responsible for transferring the FD from or to the parallel file system.

In summary, the procedure of two-phase I/O can be divided into the following stages. Offsets and lengths calculation stage (st1): In this stage each process calculates the offsets and lengths of its access requests and communicates its start and end offsets to other processes. In the end of this stage, all processes have global file access information and the involved file interval can be calculated. File domain assignment stage (st2): In this stage the involved file interval is divided into file domains (FDs) among aggregators. In this way, each aggregator only accesses the data associated with its FD in the following stages. Access request calculation (st3): The portions of the access request of each process are analyzed and which file domains they locate can be calculated. Other processes' requests which lie in the file region of each aggregator are also calculated. Buffer writing (st4): Processes send their data to appropriate aggregators, and the data are stored in the buffer of each aggregator. Disk accessing (st5): For collective write, aggregators collect data from other processes and write them to the file domain; for collective read, aggregators read data from its file domain and send them to other processes. This stage is made as many times as the following calculus indicates: the file domain size of each aggregator divided by size of the collective buffer size.

In the previous implementation of two-phase I-O, the assignment of FD to each aggregator in st2 does not take the network topology and the distribution of data into consideration. Our technique tries to modify the FD assignment



(a) Data access pattern

(b) Two-phase I/O

FIGURE 1: Two-phase I/O. Optimizing collective MPI-IO writes.

strategy, so that the initial distribution of data in the cluster is considered. By means of this strategy it is possible to improve the communication overhead and, therefore, reduce the overall I/O time. Figure 1 details the two-phase I/O technique by an example of 6 processes that write a vector of 24 elements to a file in parallel. In this example, each process writes 4 noncontiguous data blocks. P0, P2, and P4 are chosen as aggregators, and the accessed file region is divided into 3 FDs, which are assigned to them. For this example, in the shuffle phase each aggregator receives 8 data blocks from the target processes and writes them to the file in the I/O phase.

Parallel file systems usually partition a large file into blocks which are striped across many I/O servers. A client should exclusively access a file region to guarantee the cache coherence. A locking mechanism is used by Lustre and GPFS to implement the exclusive access. The lock granularity of these parallel file system is set to their block size. If the involved file region is divided into equal size file domains, a block may expand over two file domains. So when two aggregators simultaneously modify that block, the access requests must be serially served. Liao et al. [19] improved ROMIO and proposed some methods which align each file domain to a lock boundary to prevent lock contention. Some strategies are also designed to avoid lock conflicts in ROMIO's Lustre-specific code. As Figure 2 shows, the minimum and maximum offsets of the accessed file region are aligned to the lock boundaries. The optimization strategy requires that the aggregator number should be a value that can divide the OST number exactly. In Figure 2, the stripes to be read are allocated round robin among aggregators P0 and P1, and a file domain consists of the assigned stripes of an aggregator. The collective buffer size is set to the stripe size. In the I/O phase each aggregator reads a stripe into its collective buffer, and they distribute the cached data to other processes in the shuffle phase.

4. Topology-Aware Strategy for MPI-IO Implementation

In previous implementations, the assignment of the FDs to each process does not take the I/O pattern and network topology into consideration. With the topology-aware two-phase I/O, the assignment of the FDs is dependent on the initial data distribution and the locations of the processes in the cluster. The new strategy tries to minimize the communication workload or latency during the shuffle phase, and it solves the problem based on the Linear Assignment Problem. In this section, we present details about the topology-aware

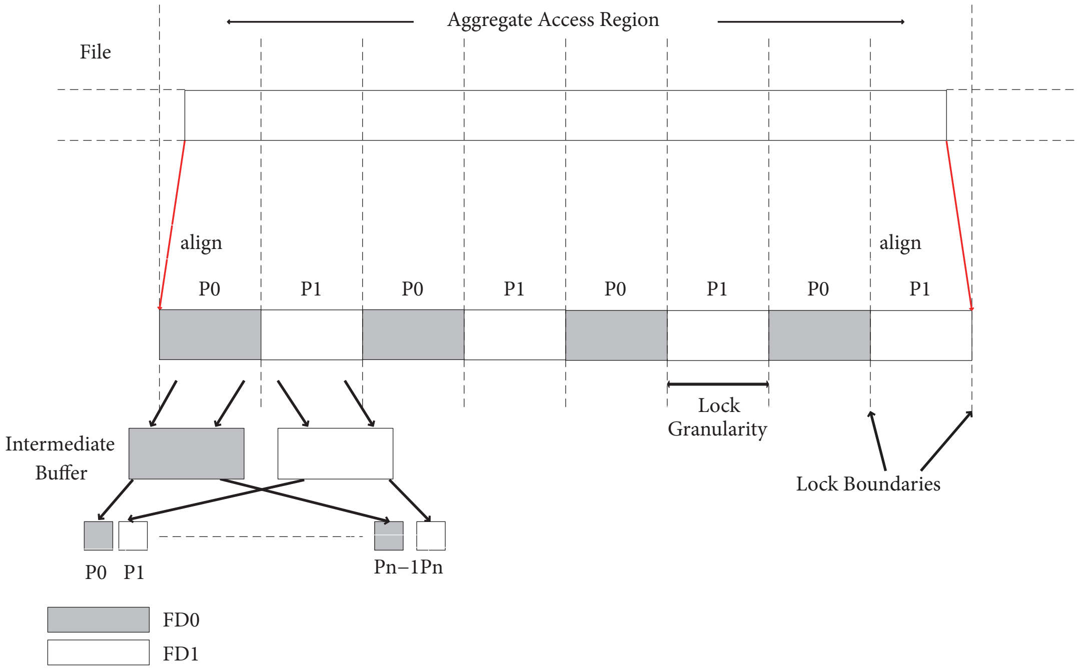


FIGURE 2: The implementation of two-phase I/O on supercomputers with Lustre installed.

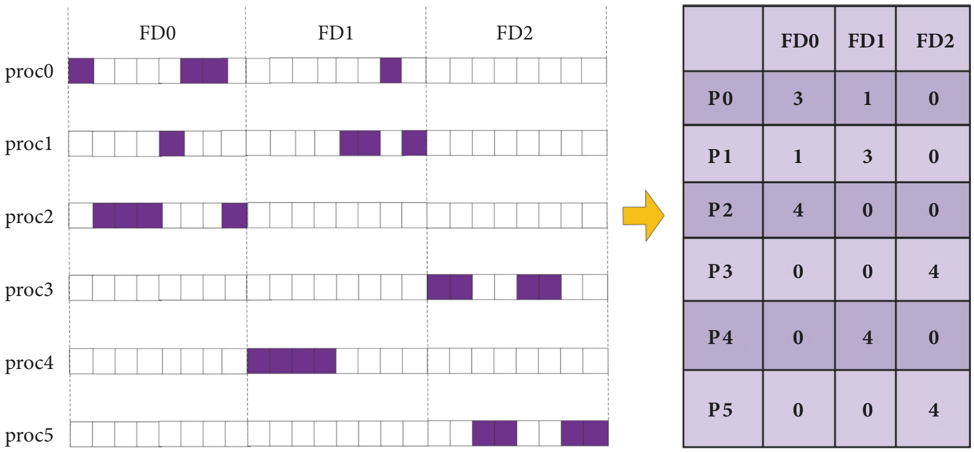
two-phase I/O. At first we show how we get the virtual and the physical topologies. These topologies are constructed as matrices whose elements represent the interprocess communication volume and the network distances between any two cores, respectively.

4.1. Collecting the Applications Communication Pattern Data. To optimize the communication performance, the vital piece of information that we need to get is the target application's communication pattern in the shuffle phase. In this case, the communication pattern is stored in a $p \times a$ matrix (p is the process number and a is the aggregator number) which consists of the volume of data exchanged between each aggregator and their target processes. Fortunately, for two-phase I/O, we can get its communication pattern during the shuffle phase in st3 and we do not need to preliminary run the application.

$$C = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 4 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \qquad (1)$$

In order to explain this, we use the example with the same data size and access pattern as shown in Figure 1. The number of intervals in which the file can be divided is set to be equal to the aggregator number, and the FDs can be calculated. The next step is constructing the communication matrix C, which is with so many rows as processes, and so many columns as FDs. Each matrix entry indicates the number of elements of a FD that will be accessed by a process. Matrix C shows the result for our example. Each single FD is assigned to one aggregator, and all processes communicate with the aggregators during the shuffle phase; thus the matrix gives the communication pattern. Figure 3 shows how to calculate the communication matrix. In matrix C, the rows indicate different processes and the columns indicate different aggregators.

4.2. Gathering the Computing Resources Topology. The node architecture and network architecture constitute the computing resource's physical topology. Higher communication performance can be expected between the cores with shorter network distance. The hardware locality (hwloc) library [20] can provide the underlying machine architecture abstraction. It detects the nodes' architectural components such as caches, cores, processor sockets, memory, NUMA, and SMT architecture and represents them as a tree with cores at the leaves and nodes at the top. We need a network discovery module to get a view of the computing resources' physical topology. InfiniBand subnet manager tools such as ibtracert [21] can help us discover the network distance between computing nodes interconnected by InfiniBand network. The module should get the distance information with the tools and merge the result with the nodes' architectural information got by hwloc. The result distance matrix will be used to make FDs assignment. For MPI applications running in a homogeneous cluster, the process with the minimum rank in the communicator will extract the distance between any two nodes using ibtracert. The process then scatters the distance



	FD0	FD1	FD2
P0	3	1	0
P1	1	3	0
P2	4	0	0
P3	0	0	4
P4	0	4	0
P5	0	0	4

FIGURE 3: The communication between aggregators and their target processes.

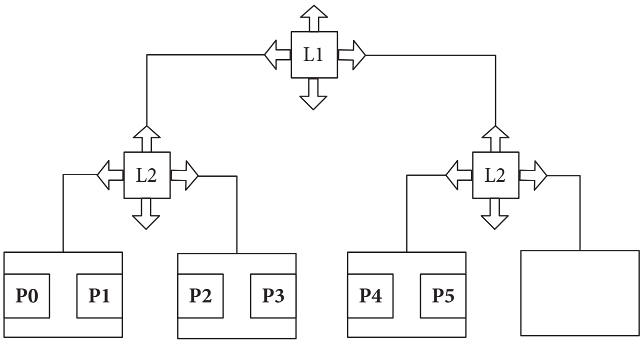


FIGURE 4: 4 nodes (here, 2-way 1-core) are connected through switches in a network with tree switches.

information to other processes which will integrate it into the node architecture to get the computing resources' full architecture.

$$D = \begin{pmatrix} 0 & 1 & 2 & 2 & 4 & 4 \\ 1 & 0 & 2 & 2 & 4 & 4 \\ 2 & 2 & 0 & 1 & 4 & 4 \\ 2 & 2 & 1 & 0 & 4 & 4 \\ 4 & 4 & 4 & 4 & 0 & 1 \\ 4 & 4 & 4 & 4 & 1 & 0 \end{pmatrix} \quad (2)$$

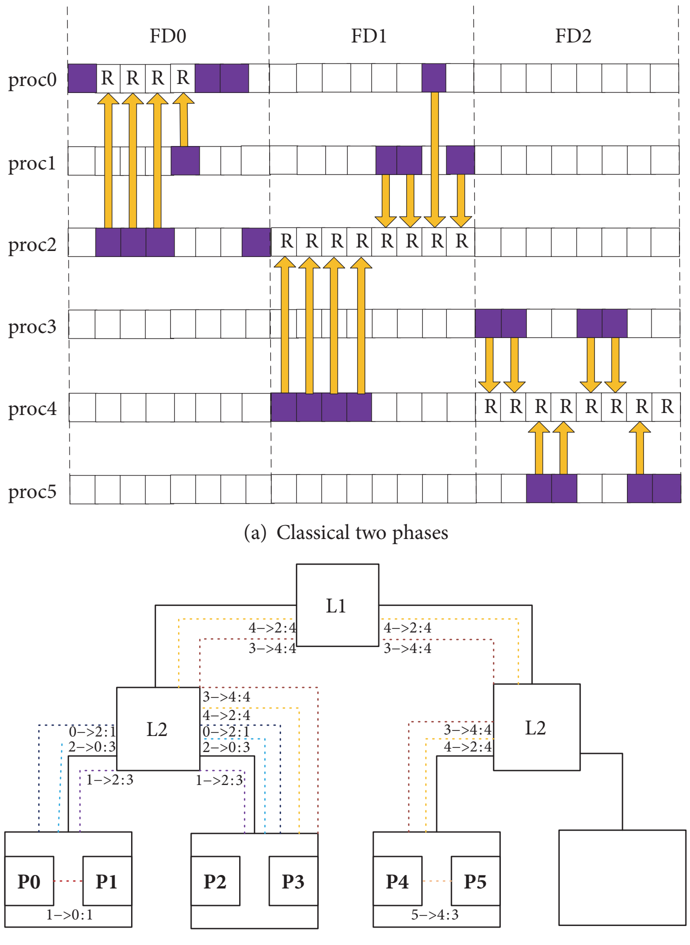
Topology matrix element represents the number of hops (length of the communication path) between two nodes. The farthest nodes get a value of the maximum network hop count for their topology matrix element. The intranode matrix elements are assigned a value that is always smaller than or equal to 1, indicating the fact that intranode communication is faster than internode communication which requires more than 1 network hop. Matrix D shows how the various distance values of Figure 4 are assigned based on the network topology and system architecture. 4 nodes (here, 2-way 1-core) are connected through 3 switches in a network. We assign the minimum distance value for cores on the same node. The next minimum distance value is 2: for instance P0 is able to communicate with P2 with 2 hops. We can assign distance values for other node pairs in the same way.

For classical two-phase I/O, the 3 FDs will be assigned to 3 aggregators running on 3 nodes; in this example they are P0, P2, and P4. The resulting network communication is shown in Figure 5. In this case the maximum link congestion is 15, and the overall I/O performance is affected by the over congested link.

4.3. Topology-Aware Two-Phase I/O. As shown in Figure 5, if the FDs are not properly assigned to the aggregators, the resulting network communication will cause high link congestion; some new mechanisms are needed to reduce the link congestion during the shuffle phase. Our technique is based on minimizing the communication workload and latency. The assignment of the FD to each aggregator in the proposed technique is different from the original version. Now each FD is assigned based on the total hop-bytes of the interprocess communication during the shuffle phase. The number of FDs into which the file can be divided is set by the hints, and the topology matrix D and the communication matrix C have already been calculated out. The next step consists in assigning each interval to each process efficiently.

$$W = C * D = \begin{pmatrix} 9 & 16 & 24 \\ 11 & 17 & 24 \\ 8 & 24 & 20 \\ 12 & 24 & 16 \\ 32 & 16 & 20 \\ 32 & 20 & 16 \end{pmatrix} \quad (3)$$

We use the hop-bytes metric to estimate the communication workload. To minimize the workload, we first calculate the work load matrix W with so many rows as processes and so many columns as FDs. When the jth FD is assigned to process i, entry e_{ij} of the work load matrix indicates the total hop-bytes of the related interprocess communications aiming to collect or scatter the data of the jth FD to or from



(a) Classical two phases

(b) Classical two-phase flow

FIGURE 5: The classical two-phase I/O and the resulting network communication.

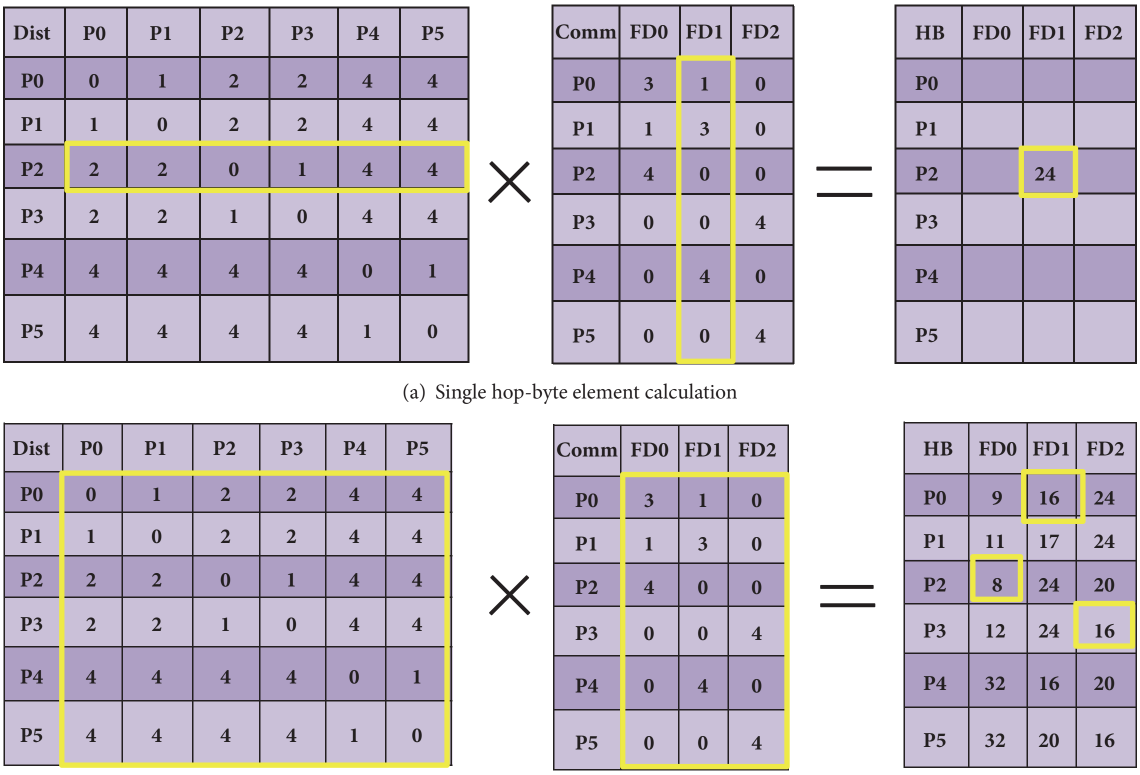
process i; thus e_{ij} is the inner product of the ith row of the topology matrix D and the jth column of the communication matrix C, respectively, so $W = C * D$. Figure 6(a) shows the result for our example. If the second FD is assigned to P2, all processes will communicate with P2 during the shuffle phase to send the data of the second FD stored by them. The message sizes are indicated by the second column of matrix C and the third row of matrix D indicates the distance between the processes. In this example, e_{21} equals 24. We can calculate all the other elements of matrix W in the same way as shown in Figure 6(b).

After getting the workload matrix W, we need to assign all the FDs to proper aggregators. The FD assignment strategy in our work is a Linear Assignment Problem (LAP) which has been well studied in combinatorial optimization and linear programming. LAP is about how to assign n items to n elements given a cost matrix in the best way. In our research, the cost matrix has been got through multiplying two matrices that record the interprocess communication volume and the interprocess distance, respectively. We have to reduce the contention on specific links, and the hop-bytes metric should be minimized. Each of the file domains can be assigned to only one aggregator, and one aggregator can access only one file domain during the I/O phase. In other words, we need to select n elements from the cost matrix, so that in each row and each column there is exactly one selected element, and the sum of them is minimum.

A large number of algorithms have been developed for LAP. The problem of finding the best FDs assignment to some particular processes can be based on the existing solutions of LAP. We have selected for our work the following algorithms, considered to be the most representative ones:

(i) Hungarian algorithm [22]: This is the first polynomial-time primal-dual algorithm solving the assignment problem. It was invented and published by Harold Kuhn in 1955 and has a $O(n^4)$ complexity.

(ii) Jonker and Volgenant algorithm [23]: A shortest augmenting path algorithm was developed to solve the Linear Assignment Problem. This algorithm contains new initialization routines and an implementation of Dijkstra's shortest path method. This algorithm is shown to be uniformly faster than the best algorithms from the literature for both dense and sparse problems computational experiments. It has a $O(n^3)$ complexity.



Dist	P0	P1	P2	P3	P4	P5
P0	0	1	2	2	4	4
P1	1	0	2	2	4	4
P2	2	2	0	1	4	4
P3	2	2	1	0	4	4
P4	4	4	4	4	0	1
P5	4	4	4	4	1	0

×

Comm	FD0	FD1	FD2
P0	3	1	0
P1	1	3	0
P2	4	0	0
P3	0	0	4
P4	0	4	0
P5	0	0	4

=

HB	FD0	FD1	FD2
P0			
P1			
P2		24	
P3			
P4			
P5			

(a) Single hop-byte element calculation

Dist	P0	P1	P2	P3	P4	P5
P0	0	1	2	2	4	4
P1	1	0	2	2	4	4
P2	2	2	0	1	4	4
P3	2	2	1	0	4	4
P4	4	4	4	4	0	1
P5	4	4	4	4	1	0

×

Comm	FD0	FD1	FD2
P0	3	1	0
P1	1	3	0
P2	4	0	0
P3	0	0	4
P4	0	4	0
P5	0	0	4

=

HB	FD0	FD1	FD2
P0	9	16	24
P1	11	17	24
P2	8	24	20
P3	12	24	16
P4	32	16	20
P5	32	20	16

(b) All hop-byte elements calculation

FIGURE 6: Calculate the hop-bytes when different FDs are assigned to different processes.

(iii) APC and APS algorithms [24]: These algorithms implement the Lawler O(n^3) version of the Hungarian algorithm by Carpenato, Martello, and Toth. APC works on a complete cost matrix, while APS works on a sparse one.

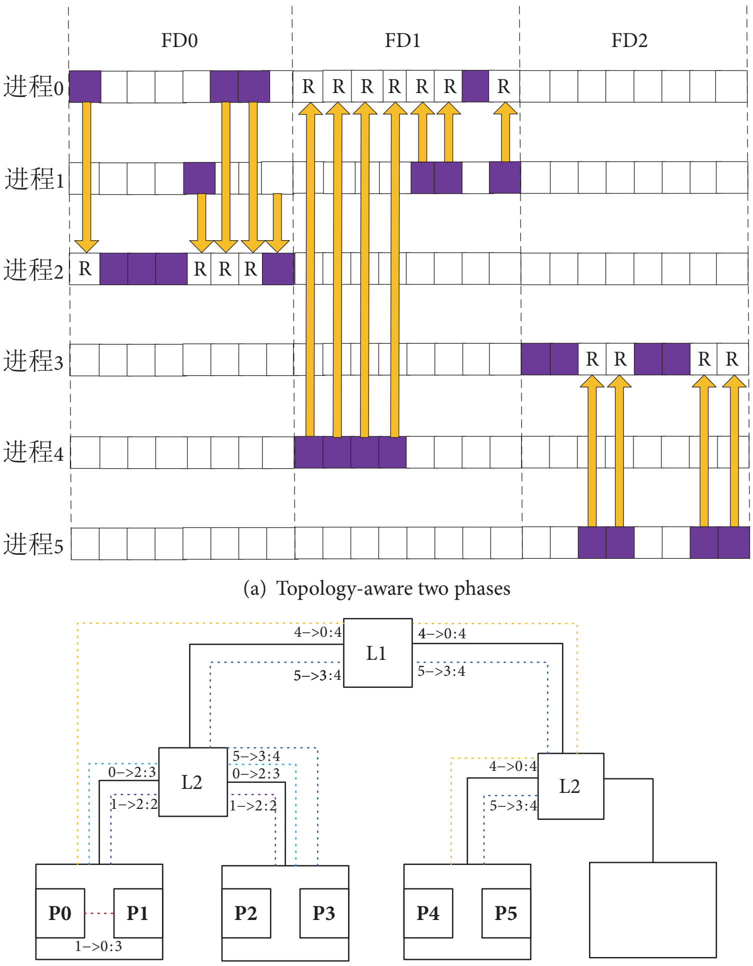
Previous evaluation [15] shows that the fastest algorithm is the Jonker and Volgenant one, and for this reason we have chosen to apply it in our topology-aware two-phase I/O. As shown in Figure 6(b), for our topology-aware two-phase I/O, the three file domains are assigned to P2, P0, and P3, respectively. Each process selected as aggregator writes to file a consecutive data set. As shown in Figure 7, in this case the maximum link congestion decreases to 9, and the overall I/O performance is significantly improved.

5. Performance Evaluation

The evaluations in this paper were performed by using two I/O benchmarks. We have compared topology-aware two-phase (TATP) I/O with the original version of two-phase (OTP) I/O implemented in MPICH and locality-aware two-phase (LATP) I/O implemented by Filgueira [15].

5.1. The Experimental Platforms. The tests have been made in our Inspur cluster and Sunway BlueLight running in National Supercomputer Center in Jinan. The Inspur cluster is organized with tree topology; it consists of 4 racks, each of which is composed of 20 Inspur computing nodes interconnected by a 40 Gb/sec InfiniBand switch. All the 4 racks are connected together by a 20 Gb/sec InfiniBand switch. Each computing node runs Red Hat 5.0 with a kernel of 2.6.18 and has two six-core processors and 8 GB memory. The parallel file system installed on the Inspur cluster is Lustre. Sunway BlueLight is organized with fat-tree topology, the water-cooled 9-rack system has 8704 ShenWei SW1600 (16 cores, 140GFLOPS) processors organized as 34 super nodes (each consisting of 256 compute nodes), 150 TB main memory, and 2 PB external storage. The system runs on its own operating system, Sunway RaiseOS, which is based on Linux. The parallel file system installed on Sunway BlueLight is also Lustre. All the computing nodes have the same distance to the I/O server of Lustre file system, so they have almost the same I/O performance in the I/O phase and improving the shuffle phase of two-phase I/O can significantly improve the overall I/O performance.

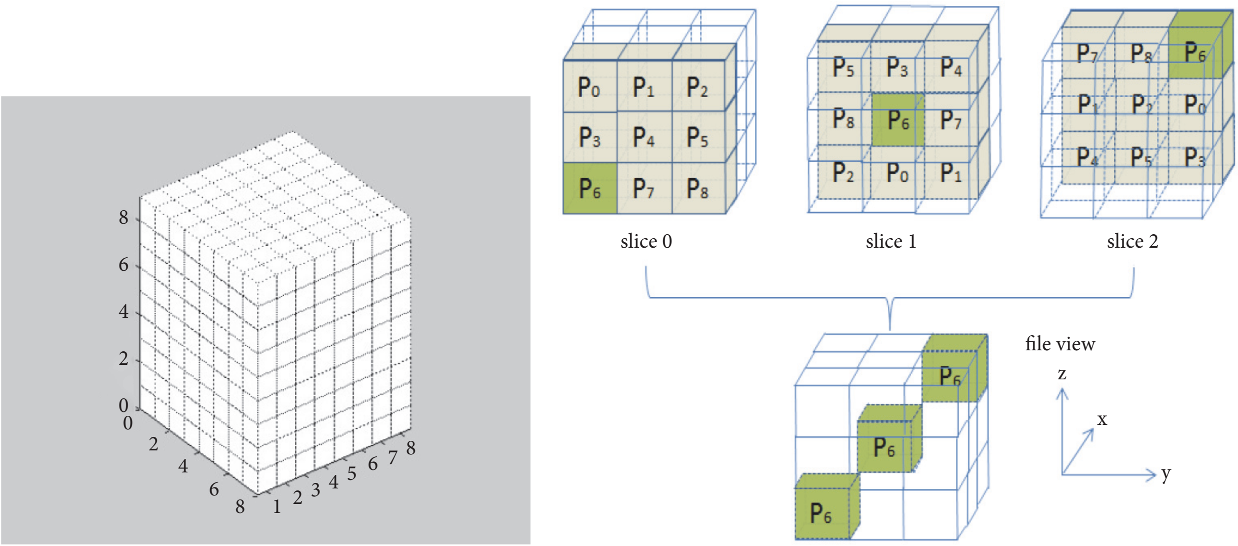
5.2. The I/O Benchmarks. We have run some benchmarks in our previous work [14], and we test those benchmarks with our new topology-aware two-phase I/O. The first benchmark has collective write/read operations. It reads/writes a three-dimensional double array from/to a file which stores the global array in row-major order. Its access pattern is shown in Figure 8(a). The 2000×2000×2000 double array is chosen and the total file size is about 60GB. Different number of processes are started to run the first parallel I/O benchmark, respectively. The data set is divided into cubes, each of which



(a) Topology-aware two phases

(b) Topology-aware two-phase flow

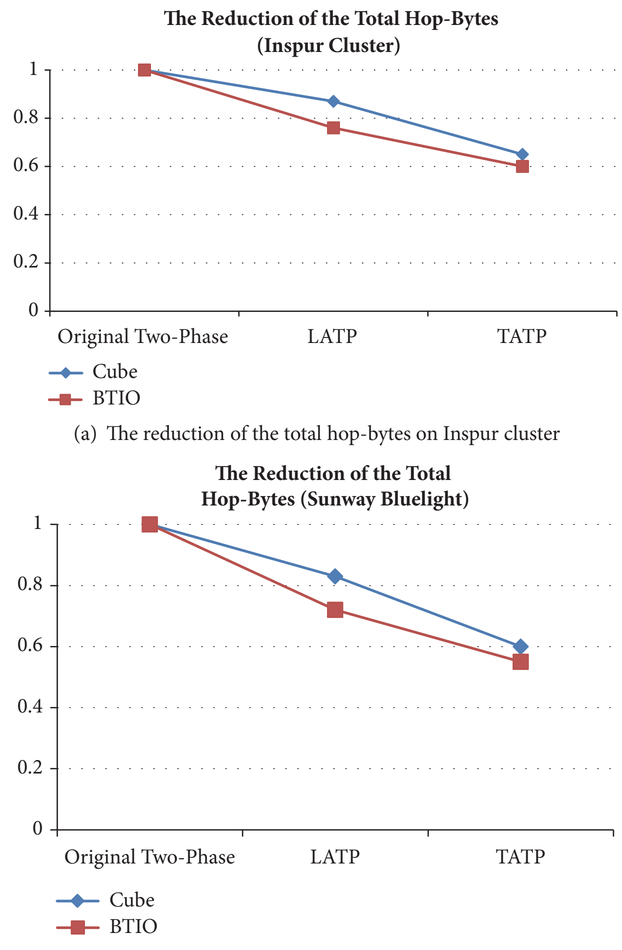
FIGURE 7: The topology-aware two-phase I/O and the resulting network communication.



(a) Cube I/O

(b) BTIO

FIGURE 8: The tested benchmarks' access pattern.



(a) The reduction of the total hop-bytes on Inspur cluster

(b) The reduction of the total hop-bytes on Sunway BlueLight

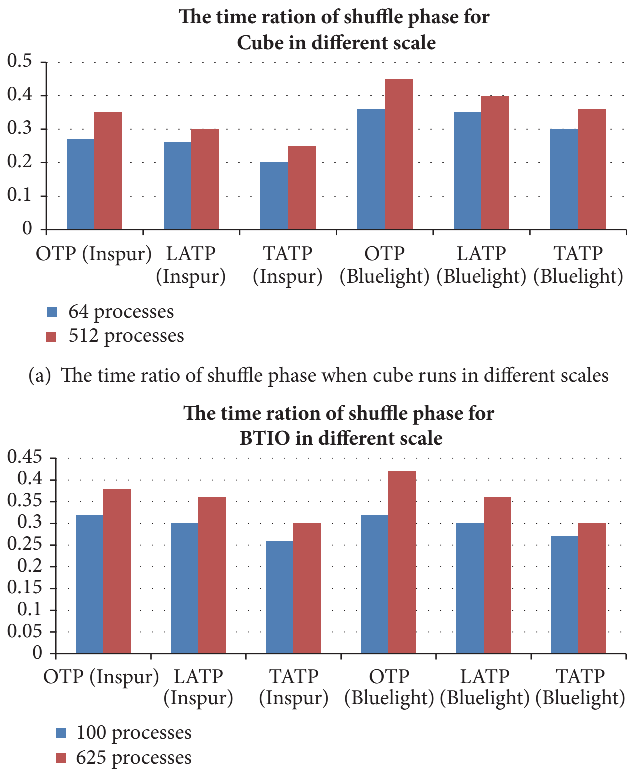
FIGURE 9: The reduction of the total hop-bytes.

is assigned to one process, and a process needs to access plenty of discontinuous data pieces in the file. File view is created to describe the access pattern and application accesses the file by collective MPI-IO operations. This benchmark represents the I/O patterns of many applications such as volume visualization which displays a 2D projection of a 3D discretely sampled data set.

In the second benchmark we implement the file access of BTIO [25] with MPI-IO functions. In this benchmark, a three-dimensional array is partitioned in a block-tridiagonal pattern and assigned across a square number of processes. Each process is responsible for many (the square root of the number of participating processors) subsets of the entire data set. When the process number is nine, Figure 8(b) illustrates how the data set is partitioned. Take process 6 as example, the cube in row 2 and column 0 of slice 0 is assigned to it, in the next slice the cube in row 1 ((2 − 1) mod 3) and column 1 ((0 + 1) mod 3) is assigned to it, and so on. Every process sets the file view and writes or reads all data subsets with one collective MPI-IO operation. We started different number of processes to run the test, respectively, and set the size of the global double array to 2000×2000×2000.

5.3. Performance Evaluation of Topology-Aware Two-Phase I/O. During the test, the aggregator number and collective buffer size were set to the default value, so on each node the process with the minimum process id will be chosen as aggregator. Firstly we started 512 and 625 processes to run cube and BTIO, respectively. Figure 9 shows the reduction of the total hop-bytes for topology-aware two-phase (TATP) I/O over original two-phase (OTP) I/O and locality-aware two-phase (LATP) I/O for the two benchmarks. We can see that when topology-aware two-phase I/O is applied, the volume of the total hop-bytes is considerably reduced. As we can see in Figures 9(a) and 9(b), for the two benchmarks, the topology-aware two-phase I/O and LATP I/O reduce more hop-bytes on Bluelight than on Inspur cluster. This condition can be explained. The Bluelight has larger scale and runs more jobs than the Inspur cluster. So its jobs usually run on very discrete nodes. With this case, the topology-aware two-phase I/O can obviously reduce the total hop-bytes and improve the collective communication during the I/O procedure.

We have represented the different phases of two-phase I/O as metadata calculation, metadata transformation, data shuffle, and data I/O. We tested the time ratio of data shuffle



(a) The time ratio of shuffle phase when cube runs in different scales

(b) The time ratio of shuffle phase when BTIO runs in different scales

Figure 10: The time ratio of shuffle phase when the two benchmarks run in different scales.

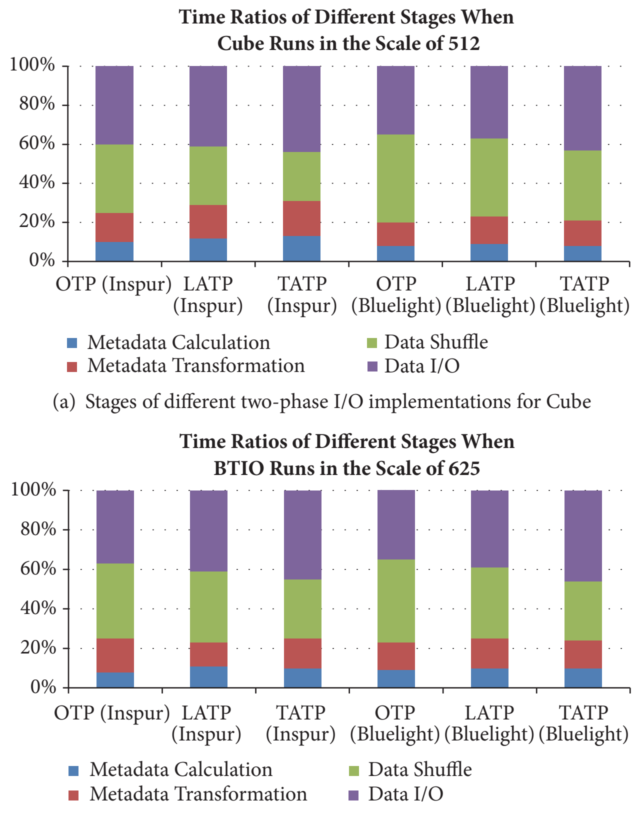
phase when the two benchmarks run in different scale. As we can see in Figures 10(a) and 10(b), the cost of the shuffle stage increases with the process number. Figure 11 represents the time ratios of different two-phase I/O stages when the two benchmarks run in the scale of 512 and 625 processes, respectively. As we can see in Figures 11(a) and 11(b), the slowest stages are data shuffle and data I/O. Based on this, we conclude that the shuffle stage is a bottleneck in two-phase I/O. For this reason, the TATP I/O technique with the aim of reducing the communication cost is necessary. This technique reduces the global hop-bytes of two-phase I/O, so the congestion is reduced and the I/O performance is improved.

We also tested how the process number affects the execution of the benchmarks. 64, 125, or 512 processes were, respectively, started to run the cube, and 100, 625, or 900 processes were, respectively, started to execute BTIO. Figure 12 shows the aggregated I/O speed when different number of processes did the collective I/O operations with different two-phase I/O implementations. We can see that the I/O performance is significantly affected by the process number. As we can see in Figure 12, topology-aware two-phase I/O has the best I/O performance. Note that when using topology-aware two-phase I/O, the total hop-bytes is also the minimum. The I/O speed increases with the decreasing of hop-bytes. These figures show the relevance of the I/O performance and the total hop-bytes. With this technique we can significantly increase the global I/O speed.

We also tested how the aggregator number affects the benchmarks' execution. With the default configuration, on each node, there is only one aggregator. To do the test, 512 processes were started to run cube and 625 processes were started to run BTIO. During the test, the number of aggregators on each node is set to 1, 2, and 4, respectively, and TATP I/O is used. The aggregated I/O speed is shown in Figure 13(a) and the comparison of the resulting hop-bytes is shown in Figure 13(b). Note that the I/O speed increases with the decreasing of hop-bytes. These figures show the relevance of the aggregator number and the total hop-bytes. Based on the design principle of two-phase I/O, once the aggregator number is set, no matter how we set the collective buffer size, the total hop-bytes during the shuffle phase remain unchanged. So properly setting the aggregator number is important for reducing the total hop-bytes. In our previous work we study how to automatically set the aggregator number and collective buffer size [14]. With this technique we can significantly increase the global I/O speed.

6. Conclusion

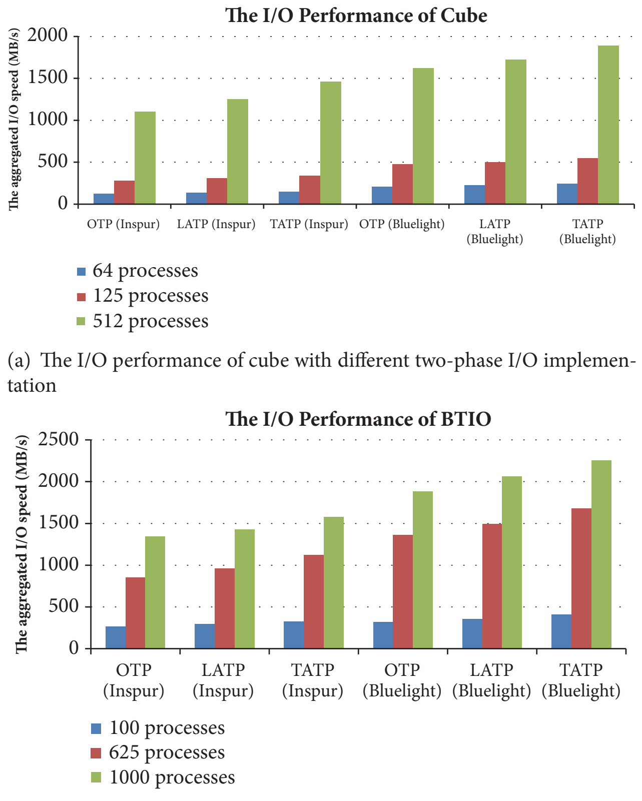
In this paper we have presented the topology-aware two-phase I/O (TATP), which optimizes the most popular collective MPI-IO implementation of ROMIO. With the topology-aware two-phase I/O (TATP), the assignment of the FDs depends on the initial data distribution and the locations



(a) Stages of different two-phase I/O implementations for Cube

(b) Stages of different two-phase I/O implementations for BTIO

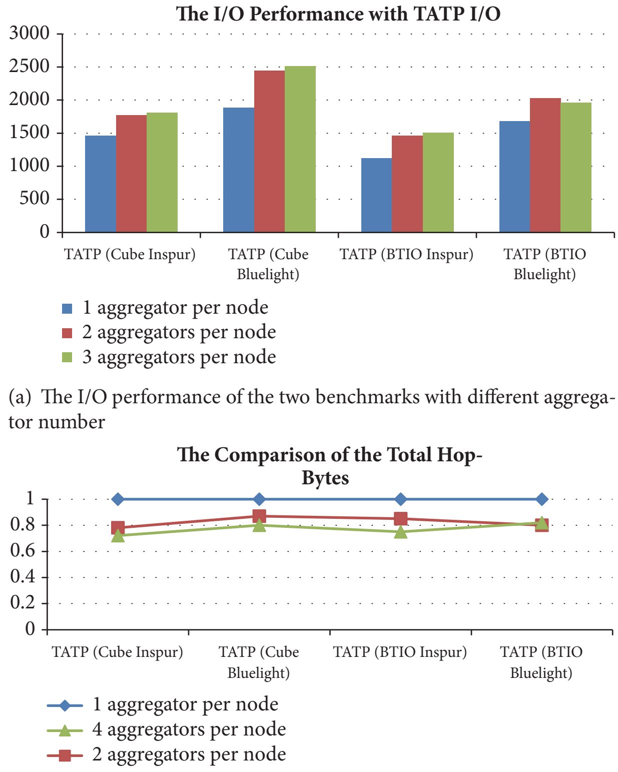
FIGURE 11: Stages of different two-phase I/O implementations for the two benchmarks.



(a) The I/O performance of cube with different two-phase I/O implementation

(b) The I/O performance of BTIO with different two-phase I/O implementation

FIGURE 12: The I/O performance of the two benchmarks with different two-phase I/O implementation.



(a) The I/O performance of the two benchmarks with different aggregator number

(b) The comparison of the total hop-bytes of the two benchmarks with different aggregator number

FIGURE 13: The relevance of the aggregator number and the total hop-bytes.

of the processes in the cluster. We use the hop-bytes metric to estimate the communication workload. If the total communication workload is low, then the contention for specific links is also much more likely to decrease. The Linear Assignment Problem (LAP) is employed to find an optimal assignment which can minimize the communication workload. As far as we know, this is the first work trying to improve the performance of two-phase I/O through reducing the total hop-bytes. Experiment results show that topology-aware two-phase I/O obtains important improvements when compared with the original two-phase I/O implementations.

Properly setting the aggregator number can significantly reduce the total hop-bytes. In our previous work an auto-tuning framework is used to automatically evaluate different aggregator numbers, respectively, but this approach takes long time to find the best configuration. In the future we will design an algorithm to compute the best configuration directly.

References

[1] R. Buyya, T. Cortes, and H. Jin, "Overview of the mpiio parallel i/o interface," in *Proceedings of the IPPS '95 Workshop on Input/Output in Parallel and Distributed Systems*, pp. 476–487, IEEE, 1995.

[2] C. D. Sudheer and A. Srinivasan, "Optimization of the hop-byte metric for effective topology aware mapping," in *International Conference on High Performance Computing*, IEEE, 2012.

[3] A. Bhatele, I. Chung, and L. V. Kale, "Automated mapping of structured communication graphs onto mesh interconnects," 2010.

[4] F. B. Schmuck and R. L. Haskin, "Gpfs: A shared-disk file system for large computing clusters," in *FAST '02 Proceedings of the 1st USENIX Conference on File and Storage Technologies*, vol. 2, pp. 19–32, 2002.

[5] P. Schwan, "Lustre: Building a file system for 1000-node clusters," in *Proceedings of the 2003 Linux Symposium*, vol. 2003, 2003.

[6] M. Chaarawi, S. Chandok, and E. Gabriel, "Performance evaluation of collective write algorithms in mpi i/o," in *Proceedings of the International Conference on Computational Science*, pp. 185–194, 2009.

[7] F. Isaila, P. Balaprakash, S. M. Wild et al., "Collective I/O Tuning Using Analytical and Machine Learning Models," in *Proceedings of the Ieee/rsj International Conference on Intelligent Robots and Systems*, vol. 3, pp. 2392–2397, 2015.

[8] V. Venkatesan, R. Anand, J. Subhlok, and E. Gabriel, "Optimized process placement for collective I/O operations," in *Proceedings of the European Mpi Users' Group Meeting*, pp. 31–36, 2013.

[9] J. M. D. Rosario, R. Bordawekar, and A. Choudhary, "Improved parallel I/O via a two-phase run-time access strategy," *ACM*, 1993.

[10] R. Thakur, W. Gropp, and E. Lusk, "A Case for Using MPI's Derived Datatypes to Improve I/O Performance," in *Proceedings of the 1998 ACM/IEEE Conference on Supercomputing, SC '98*, pp. 1–10, 1998.

[11] R. Thakur, W. Gropp, and E. Lusk, "Optimizing noncontiguous accesses in MPI-IO," *Parallel Computing*, vol. 28, no. 1, pp. 83–105, 2002.

[12] Y. Tsujita, H. Muguruma, K. Yoshinaga, A. Hori, M. Namiki, and Y. Ishikawa, "Improving collective i/o performance using pipelined two-phase i/o," in *Proceedings of the 2012 Symposium on High Performance Computing*, 2012.

[13] Y. Tsujita, K. Yoshinaga, A. Hori, M. Sato, M. Namiki, and Y. Ishikawa, "Multithreaded two-phase I/O: Improving collective MPI-IO performance on a lustre file system," in *Proceedings of the Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pp. 232–235, 2014.

[14] W. Liu, M. Gerndt, and B. Gong, *Model-Based MPI-IO Tuning with Periscope Tuning Framework*, John Wiley and Sons Ltd, 2016.

[15] R. Filgueira, D. E. Singh, J. C. Pichel, F. Isaila, and J. Carretero, "Data locality aware strategy for two-phase collective," in *Proceedings of the High Performance Computing for Computational Science-VECPAR 2008*, pp. 137–149, 2008.

[16] H. Bui, J. Leigh, E. Jungy, V. Vishwanathy, and M. E. Papka, "Improving Data Movement Performance for Sparse Data Patterns on the Blue Gene/Q Supercomputer," in *Proceedings of the International Conference on Parallel Processing Workshops*, pp. 302–311, 2015.

[17] V. Vishwanath, M. Hereld, V. Morozov, and M. E. Papka, "Topology-aware data movement and staging for I/O acceleration on Blue Gene/P supercomputing systems," in *Proceedings of the High PERFORMANCE Computing, Networking, Storage and Analysis*, 2011.

[18] F. Tessier, V. Vishwanath, and E. Jeannot, "TAPIOCA: An I/O Library for Optimized Topology-Aware Data Aggregation on Large-Scale Supercomputers," in *Proceedings of the IEEE International Conference on CLUSTER Computing*, pp. 70–80, 2017.

[19] W. keng Liao and A. Choudhary, "Dynamically Adapting File Domain Partitioning Methods for Collective I/O Based on Underlying Parallel File System Locking Protocols," in *Proceedings of the ACM/IEEE Conference on Supercomputing*, pp. 313–344, 2008.

[20] F. Broquedis, J. Clet-Ortega, S. Moreaud et al., "Hwloc: a generic framework for managing hardware affinities in HPC applications," in *Proceedings of the 18th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP '10)*, pp. 180–186, 2010.

[21] A. Bermúdez, R. Casado, F. J. Quiles, T. M. Pinkston, and J. Duato, "Evaluation of a subnet management mechanism for InfiniBand networks," in *Proceedings of the 2003 International Conference on Parallel Processing, ICPP 2003*, pp. 117–124, 2003.

[22] S. S. Blackman, "Multiple target tracking with radar applications," *Dedham Ma Artech House Inc P*, vol. 1, pp. 204-205, 1986.

[23] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, 1987.

[24] G. Carpaneto, S. Martello, and P. Toth, "Algorithms and codes for the assignment problem," *Annals of Operations Research*, vol. 13, no. 1-4, pp. 193–223, 1988.

[25] P. Wong and R. F. V. der Wijngaart, "NAS Parallel Benchmarks I/O Version 2.4," 2003.

IoT Location Privacy Protection Addresses in Wireless Sensor Networks Anonymity

Amiya Kumar sahoo, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, aksahoo336@gmail.com*

Shriram Agarwal, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, shriram_agarwal@gmail.com*

Ipsit Joshi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, ipsit.joshi29@gmail.*com

Srimanta Mohapatra, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srimantamohaparta66@gmail.com*

Abstract

Location privacy is very important for event-triggered type of Wireless Sensor Networks (WSNs) applications such as tracking and monitoring of wild animals. Most of the security schemes for WSNs are designed to provide protection for content privacy. Contextual privacy such as node identity anonymity has received much less attention. The adversary can fully explore such contextual information to disclose the location of critical components such as source nodes or base station. Most existing schemes provide location privacy at network layer. As no measures are taken to provide node identity anonymity at data link layer, the adversary can launch traffic analysis attacks to jeopardize location privacy. In this paper, a scheme named HASHA is proposed to defend against traffic analysis attacks through hashed one-time addresses. Hashed results of payload are used to create dynamic one-time MAC addresses between the communication pairs. Because of inevitable wireless frame errors, it is impossible for adversaries to track dynamic addresses. Therefore, HASHA can provide strong node identity anonymity, which makes traffic analysis attacks much more difficult and provides better location privacy. Simulations and analysis results show that HASHA can provide better location privacy with limited communication overheads, which is particularly suitable for resource-limited WSNs.

1. Introduction

A typical Wireless Sensor Network (WSN) is composed of dozens to thousands of tiny, low-cost, and resource-constrained sensor nodes that are self-organized as an ad hoc network to monitor the physical world. One type of applications of WSNs is wildlife habitat monitoring, in which all sensor nodes are deployed randomly to monitor the target of interests [1]. Detection events are reported from the source node to the base station in a multihop fashion. Unattended operation and open wireless communication channel make WSNs vulnerable to attacks. However, as sensor node has limited memory, energy, and communication resources, traditional security techniques cannot be used in WSNs. Light-weighted schemes are required to achieve secure communication for WSNs [2].

Security for WSNs has focused on security services that provide authentication, confidentiality, integrity, and availability [3, 4]. Such techniques belong to content privacy. Now, however, there is a growing interest in contextual privacy, which focuses on hiding the contextual information of WSNs. Location information of key components is one of the most important contextual privacy parts that should be protected.

In the wildlife monitoring application, all sensor nodes detect occurrence of the target animal to the base station. In the case that one sensor node (source node) detects target, a packet is generated and sent to the base station hop by hop to report occurrence of the target. In such applications, geographic locations of the source node and base station are sensitive information that should be protected [5]. The base station is the only gateway to outside networks, and the source node reveals physical location of wildlife. If the location of base station is disclosed by the adversary, the capture of the base station can make the entire network nonfunctional. And if the location of source node is disclosed, the adversary can find the animal easily because the geographic location of source node and the target must be very close. Therefore, providing location privacy of source node and base station is of great importance in such applications.

Existing techniques provide location privacy at network layer. Random Walk has packets that follow random route while forwarding the packets from the source node to base station [6, 7]. As it is difficult to the adversary to backtrack to the source node while random route is used, location privacy of source node is achieved. Dummy Data Source scheme invites some fake source nodes into the WSNs to confuse the adversary and provide location privacy [8, 9]. However, both schemes introduce additional communication overheads, which consume much more energy. For example, if the average hops count from the source node to base station is twice than that of shortest path, the energy consumption is twice too. For the same reason, if one more fake source node is added to the network, the power consumption doubled.

The adversary may launch traffic analysis attacks to find the geographic locations of the source node. As the content information is protected by the encryption techniques, the adversary cannot decrypt its contents without keys. However, as the contextual information is not well protected, it can be used to launch successful traffic analysis attacks.

The adversary first captures frames around the base station. The structure of frame at the data link layer data is <DA||SA||payload||FCS>. Payload is content from upper layer, FCS is the frame checksum, DA is receiver address, and SA is transmitter address. Supposing that the captured frame is <BS||B||payload||FCS>, the adversary cannot get any information from the payload because it is encrypted. However, two addresses indicate that the frame is from node B to the base station. To find the geographic location of node B, the adversary captures a series of data packets from node B at different locations and moves towards locations where stronger Received Signal Strength (RSS) presents. After finding the geographic location of node B, the adversary can find the next node by the same way. To find the source node, the adversary continues such process until no more next nodes are detected. Source location privacy was compromised.

Two steps are used repeatedly by the adversary to locate the source node. The first is forwarding relationship analysis. The adversary knows the address of base station by traffic analysis. Then, it knows the forwarding node closer to the source node by analyzing frames to the base station. The second step is to move closer to the forwarding node by analyzing RSS. Apparently, the addresses in data link layer frames are vital for successful traffic analysis attacks. It is much more difficult or impossible for the adversary to launch traffic analysis attacks if the addresses in the frame are well protected.

One way to hide node address is to break the relevance between physical node and the address of the node. For example, if node X and node X′ in WSN have the same address IDx, as two nodes have the same address but are deployed at different geographic location, the adversary cannot locate the node(s) by analyzing the RSS [9]. Thus, traffic analysis attacks can be eliminated. Of course, above simple scheme introduces great trouble to normal operation of networks. But breaking the relevance between physical node and the address is an effective way to defend against traffic analysis attack and provide location privacy [10–13].

Another way to break the relevance between physical node and the address is introducing more addresses to node that cannot be distinguished by the adversary. If node X communicates with base station using a serial of identities <X1, X2, X3, . . ., Xn>, and only node X and the base stations know that the addresses belongs to node X, the adversary cannot learn the communication relationship to track the source node by traffic analysis attacks [14].

Based on such observations, this paper proposes a novel scheme to provide location privacy at data link layer. The contributions of this paper are threefold: First, the proposed scheme protects location privacy at data link layer, which is more effective to defend against traffic analysis attacks. As compared to schemes at network layer, the proposed scheme introduces negligible communication overheads. Second, in tracking and monitoring applications, location privacy of the source node and that of base station are both very important. Exposure of base station will endanger the whole WSN, while source node location discloses the position of the target. Source location privacy and base station privacy are both provided in the proposed scheme. Existing schemes emphasize on either the source location privacy or base station location privacy. Third, the proposed scheme defends traffic analysis attacks through address anonymity, which can provide location privacy against both inner attackers and outside attackers. Protection against inner attackers is particularly important, because node compromise is fairly easy for unattended WSNs and the compromised node can be an inner attacker with some software modifications. Existing address anonymity schemes can only defend against outside attackers.

2. Related Works

Phantom Routing belongs to Random Walks type schemes that provide location privacy for WSNs [13–15]. To prevent being located by step-by-step tracing, the source node sends each packet to a randomly selected forward node. This forward node is called a Phantom node. On receiving the packet to be forwarded, the Phantom node routes the packet to the base station using broadcasting. Suppose that an adversary launches traffic analysis attacks to find the geographic location of the source node. As the Phantom node sends packets to base station via broadcasting instead of unicasting, it is fairly difficult for the adversary to trace to the Phantom node using traffic analysis. However, energy consumption in Phantom Routing is much greater than unicasting type of schemes, because broadcasting is used to forward packets from Phantom node to base station. To reduce power consumption of broadcasting, another scheme named Phantom Single-path Routing scheme (PSRS) is proposed [16]. Different from original Phantom Routing, the Phantom node in PSRS routes packets to base station via unicast. As the source node selects different Phantom node for each packet, different paths are used for different packets. Therefore, it is still very difficult to locate source node via traffic analysis attacks. The PSRS can reduce power consumption because broadcasts are eliminated. But the randomly selected paths are much more power-consuming than the shortest ones.

Another type of schemes that provides location privacy is dummy source node [17, 18]. Fake source nodes are introduced to obfuscate real source node. The basic idea of such schemes is quite simple. There are many source nodes in the WSN, only one node is real source node, and other nodes are fake source nodes. The adversary can no longer see which one is real source node, even if success traffic analysis attacks are launched. Obviously, one additional fake source node introduces additional network traffic, which corresponds to additional energy consumption. The more fake source nodes introduced, the more power consumption.

Simple Anonymity Scheme (SAS) is the first scheme proposed to provide location privacy at data link layer by hiding the address. Each node communicates with neighbor using a pseudonym [19]. A large range of pseudonyms are used, and each node is assigned with a subspace of the pseudonym space. Both nodes of the communication pair at data link layer know each other's pseudonym spaces. Both nodes use different pseudonym within its pseudonym spaces. Therefore, the adversary cannot identify the physical node if the pseudonym space is unknown to it. The main drawback of SAS is that it cannot protect address anonymity if there is an internal attacker. For example, if the adversary has the full pseudonym space and the subspace allocation for each node, it can capture frame and compare each address with the pseudonym space and finally find out the physical node for each address. Another drawback of SAS is that each node must store pseudonym space for each neighbor, which introduces great storage overheads if many neighbors exist.

Cryptographic Anonymity Scheme (CAS) uses a keyed hash function to generate the pseudonym used for communication between the communication pairs at data link layer [20, 21]. Before deployment, the communication pairs are assigned a key k for pseudonyms generation. After deployment, the communication pairs create pseudonyms with a random number r and a sequence number *seq*. The ith pseudonym can be expressed as $\mathrm{ID}i = \mathrm{H}_k(r \oplus seq)$. Before frame transmission, a different sequence number *seq* is used, so each frame has different pseudonym. CAS reduces storage overhead at the expense of additional computation overheads. Apparently, CAS cannot prevent internal attackers from finding out that some pseudonyms belong to a physical node if the key k is stolen by the adversary via compromising.

The schemes mentioned above either cannot protect location privacy in the presence of inner attackers or consume too much energy resource because of communication overheads. And most of the schemes proposed focused on protecting location privacy of source node. In this paper, the proposed scheme protects location privacy at data link layer by address anonymity. The address anonymity can resist traffic analysis attackers launched by both outside attackers and inner attackers. With a modification to the network layer, the scheme can provide location privacy with much less energy consumption. Location privacy of both base station and source node can be protected with the proposed scheme.

3. Network and Adversary Models

3.1. Network Model. In this paper, it is assumed that many nodes are randomly deployed to monitor the geographic location of the target. Each node is capable of communication, computation, and sensing. All nodes in the network are powered by batteries and work in an unattended manner [21]. Therefore, power efficiency is the most important design consideration for both software and hardware. There is only one base station in the WSN, which is the gateway to outside networks.

All nodes in the WSN are working coordinately to detect the presence of a target. Any tracking approaches can be used to detect the target, provided that they are power efficient. The node that detects the target is called the source node. On detecting the target, the source node sends packets to the base station to report the information of the target. The source node reports to the base station for fixed time interval until the target moves outside the detection radius. Other nodes in the WSN sleep unless they are requested to forward the packets from the source node to the base station.

3.2. Adversary Model. Location privacy of the source node and that of base station are both important [21–23]. We consider two types of adversary. The first type of adversary is interested in catching the animals that are monitored by the WSN. Because the network traffic from the source node is an excellent guide to find the animals, the adversary attempts to find the node closer to the source node (and the animal) through traffic analysis attacks. The second type of adversary attempts to find the base station and damage it, which will make the entire WSN useless. With the same approach, the adversary can find the base station.

Only local adversary is considered in this paper. The reason is that global adversary requires much more expensive devices than the local adversary. Some researches suppose that the adversary is equipped with wireless devices that can cover the whole WSN. Such devices should be very expensive. Many nodes that are geographically separated in a WSN may transmit simultaneously without collision at the respective receivers. But as the wireless device of adversary can hear many simultaneous transmissions, collision may occur at the adversary. Expensive wireless device may not necessarily lead to better attack results. Therefore, we only consider local adversary.

Only passive adversary is considered in this paper. That means the adversary never transmits to avoid being detected by WSNs. To locate the source node and base station, the adversary captures and analyzes frames to get the communication relationship among nodes. To move closer to the source node or base station, the adversary may move closer to a node by comparing RSS from different locations.

The adversary may launch another traffic analysis attack named time correlation [24–27]. After detecting the target, the source node sends a packet to the next node closer to the base station to notify the event. The next node also relays the packet to a node closer to the base station. The adversary can observe the correlation in transmitting time between one node and the next node to find the route to the source node or base station. For a simple example, if the adversary notices that after node A transmits a packet, node B transmits a packet with the same size, it can learn that node A is closer to the source node, and node B is closer to the base station. The reason is that, in a typical tracking and monitoring application, only the source node generates packets and the base station is the only destination.

As nodes in a WSN are frequently deployed in unattended environment, node may be captured by the adversary. The adversary can analyze the software and hardware of the node. It is possible for the adversary to get the pairwise shared keys or other sensitive information [28, 29]. Even more, modification to the software is also possible if the adversary has enough skills [30–34]. The captured node then becomes an internal attacker. Protecting attacks launched by an internal attacker is much more difficult than that launched by outside attackers [35–37].

4. Proposed Location Privacy Scheme

4.1. Address Anonymity Scheme. A node may have different identity at different layers of the network protocol stack. Identity at network layer and upper layers can be protected by cryptographic system. However, identity at data link layer has not been well protected in popular wireless standards such as 802.15.4 and Lora [31, 32]. Without introducing confusion, identity and Media Access Control (MAC) address are used interchangeably in this paper. The frame structure at data link layer can be illustrated in Figure 1.

DA	SA	Payload	FCS

Figure 1: DATA frame at the data link layer is composed of destination address, source address, payload, and frame checksum.

DA	SA	FCS

Figure 2: ACK frame at the data link layer is used for reliable communication.

DA is destination address of the frame. SA is the address of the sender. Payload is data from upper layer. Upper layer of data link layer is network layer. Therefore, payload at data link layer is usually packet at network layer plus control information. FCS is frame checksum.

As wireless channel is error prone, Automatic Repeat request (ARQ) is used to provide reliable data transmission. On receiving DATA frame, the receiver responses an ACK frame to inform the sender that it has received the DATA frame successfully. Structure of ACK frame is illustrated in Figure 2.

As compared to DATA frame, the ACK frame is much shorter. But both destination address and source address are included in the ACK frame. Destination address is the address of DATA frame sender, and source address is the address of receiver.

As elaborated in the adversary model, SA and DA of each node are known to the adversary who captures frames through eavesdropping. By analyzing these addresses, the adversary knows how many nodes in the WSN and the MAC address of each node. Furthermore, based on such captured frames, the adversary can deduce the routing information of the network or even locate a certain node in the network. Therefore, unprotected addresses at data link layer are the root factor jeopardizing location privacy.

To protect the addresses in the DATA frame and ACK frame, a hash function Hash() and a keyed hash function HMAC() are used. For DATA frames from node a to node b, both nodes keep the following variables: Key[a- > b], IDS[a- > b], IDD[a- > b], where Key[a- > b] is a secret key to protect the addresses in MAC frames. IDS[a- > b] is the source address assigned to DATA frame and IDD[a- > b] is the destination address assigned to DATA frame.

Nodes in the WSN know each other by beacon broadcasting. For example, node a knows IDb after receiving beacons from node b. Node a and node b initialize these variables as follows:

(i) Key[a- > b] ←0xFFFFFFFF

(ii) IDS[a- > b] ←HMAC(Key[a- > b], IDa)

(iii) IDD[a- > b] ←HMAC(Key[a- > b], IDb)

For the first DATA frame from node a to node b, two hashed addresses IDS[a- > b] and IDD[a- > b] are used. After ACK frame from node b to node a, both nodes update key and addresses:

(i) Key[a- > b] ←Key[a- > b] ⊕ Hash(payload)

(ii) IDS[a- > b] ←HMAC(Key[a- > b], IDa)

(iii) IDD[a- > b] ←HMAC(Key[a- > b], IDb)

As payload of the first frame is received successfully by node b, it has the same Key[a- > b], IDS[a- > b], and IDD[a- > b] as node a. We call this a secret key update process.

As it is well known, wireless channel is error prone. Both DATA frame and ACK frame may be corrupted. On receiving corrupted DATA frame, node b will not acknowledge node a with ACK frame. Both nodes will not update the key. Node a retransmits the DATA frame using the old key as described above.

In another scenario, DATA frame is received correctly by node b, but the ACK frame to node a is corrupted or lost. Node a retransmits the DATA frame as it does not receive the ACK frame correctly. But node b has already updated the key. Key mismatch problem occurs.

To address key mismatch problem, two temporary addresses are used by node b to avoid key mismatching. Node b keeps a copy of old address on receiving DATA frame successfully. If the received address in next DATA frame does not match the **new** address, it will try to match the **OLD** temporary address. If the old one matches, that means this DATA frame is a retransmission. Just reply node a with the ACK frame that already transmitted.

Node *a* and node *b* repeat such process for all the frames from node *a* to node *b*. Such process creates one-time source address and destination address. We call it a dynamic address or hashed address (HASHA) (Algorithm 1).

HASHA updates key for the communication pairs after a successful data transmission. And the one-time secret is further used to update the addresses, which creates dynamic addresses. Such process can create great difficulty to the adversary.

Figure 3 illustrates a typical scenario that an adversary captures frames from node A to node B. Initially, as the adversary knows MAC addresses node A and node B through capturing beacons. The adversary knows the initial value of Key[*a*- > *b*]. Both node B and the adversary receive DATA1 and DATA2 successfully.

At time *t1*, ACK3 is corrupted and node B does not receive it correctly; node B receives the retransmission with backup key. This will not introduce trouble to the adversary.

At time *t3*, DATA4 is not received correctly by the adversary; as the adversary is passive attacker, it cannot ask node A for retransmission. Thereafter, the adversary cannot trace frame from node A to node B after time *t3*. The reason is that the addresses used by node A and node B are created by HMAC function with key5. Key5 is created by all previous payloads from node A to node B. The result is that the dynamic one-time addresses of the following frames from node A to node B are indistinguishable to the adversary. Address anonymity is achieved.

As wireless frames are error prone because of collision and interference, corrupted frames at the adversary will prevent it from identifying nodes in the WSN. Therefore, with HASHA, the eavesdropping adversary cannot identify number of nodes in the WSN. Therefore, it cannot retrieve the routing information. Without forward routing information, the adversary cannot trace the source node and base station.

4.2. Possible Attacks against HASHA and Countermeasures. Even though the addresses are hidden by address anonymity, the adversary can still launch two types of attacks to jeopardize the source node and base station location privacy. The first attack is time correlation attack. The adversary can deploy several attack nodes in the target WSN. These nodes are carefully deployed so as all communications in the WSN can be captured. The geographic coordinates of these nodes are recorded in a center control point. The attack nodes can communicate with each other to report captured frames to the center control point. Time synchronization algorithm can be used to distribute global time to these nodes. Therefore, the resulting attack network can be used to detect transmission all over the network.

In a typical event-triggered monitoring type of WSN, network traffic in the networks is triggered by event detected by the source node. The source node reports event to the base station with the help of the forwarding nodes. On forwarding the event to the base station, transmission time of the forwarding nodes may disclose the location of source node and the base station.

As illustrated in Figure 4, nodes a1, a2, and a3 are nodes of the attack network to monitor network traffic. Node S is source node and node D is base station. Node A and node B are relay nodes. To report event from node S to base station D, node S sends packet to node A, and node A sends packet to base station D with the help of node B. The transmission time is illustrated in Figure 5. By analyzing the transmissions time serial, the adversary can find that node S is the source node and node D is the base station, which are located near node a1 and node a3, respectively. The location privacy of source node and that of base station are jeopardized. Of course, with the help of address anonymity, the adversary cannot identify node S, node A, and node B. But it can still detect that the source node is close to a1 and the base station is close to node a3. As locations of node a1 and node a3 are known to the adversary, address anonymity cannot eliminate such time correlation attacks.

Time correlation attacks use the pattern of occurrence of transmissions along the forwarding path to find the source node and base station. For example, for each event, transmission of node S is always followed by transmission of node A, because node A is the next hop of the forwarding path. Transmission serial {S1, A1, B1}, {S2, A2, B2}, and {S3, A3, B3} disclose forwarding relationship among nodes, which can be used to jeopardize source node and base station location privacy. Breaking the transmission pattern is important to eliminate time correlation attacks. The solution is to introduce random delay while forwarding packet. As illustrated in Figure 6, node S and node A delay random time for packets. The resulting transmissions serial {S1, A1, S2, A2, S3, B1, B2, A3, B3} does not disclose any forwarding relationship anymore. With the help of address anonymity, it is more difficult for the adversary to locate the source node and base station via time correlation attacks.

The formal description of random delay can be expressed as follows.

Each node forwards packets with random delay, which is effective to prevent time correlation attacks. Of course, random delays may introduce delay to event reporting to base station. In some applications, timely delivery of important packet to base station is very important. To provide higher priority to such important data, a smaller random delay *rand* in Algorithm 2 can be selected.

Another traffic analysis attack is traffic outlining attack. As address anonymity and random delay are used to prevent traffic analysis attack and time correlation attack, respectively, it is much difficult for adversary to launch attacks based on node address and forwarding relationship. But the adversary can still attack the target network via traffic outlining attack. As mentioned above, the adversary can deploy many attack nodes in the network to launch a distributed attack. For example, in the network illustrated in Figure 7, source node S reports to base station B. The adversary can deploy many attack nodes to monitor network traffic. As network traffic of event-triggered WSN is characterized from source node to base station, it is impossible for the adversary to outline the traffic without the help of distributed attack nodes. All attack nodes report to the adversary only in the presence of traffic in a certain time period. As the geographic location of attack nodes is known to the adversary, the adversary knows geographic

```
if node is sender then
    Key[a- > b] $\leftarrow$ 0xFFFFFFFF;
    if node has frame to be transmitted then
        IDS[a- > b] $\leftarrow$ HMAC(Key[a- > b], IDa);
        IDD[a- > b] $\leftarrow$ HMAC(Key[a- > b], IDb);
        Send DATA frame { IDS[a- > b], IDD[a- > b],payload, FCS};
        Create timer Stimer with a certain timeout;
    end
    if an ACK frame is received then
        for each neighbors do
            if FCS checksum OK and IDS[a- > b] from ACK frame = = IDS[a- > b] then
                //generate new key
                Key[a- > b] $\leftarrow$Key[a- > b] $\oplus$ Hash(payload);
                //generate new addresses
                IDS[a- > b] $\leftarrow$HMAC(Key[a- > b], IDa);
                IDD[a- > b] $\leftarrow$HMAC(Key[a- > b], IDb);
                kill timer Stimer;
            end
        end
    end
    if Stimer timeout then
        Send DATA frame { IDS[a- > b], IDD[a- > b],payload, FCS};
    end
end
if node is receiver then
    Key[a- > b] $\leftarrow$ 0xFFFFFFFF;
    IDS[a- > b] $\leftarrow$ HMAC(Key[a- > b], IDa);
    IDD[a- > b] $\leftarrow$ HMAC(Key[a- > b], IDb);
    IDSo[a- > b] $\leftarrow$ HMAC(Key[a- > b], IDa);
    IDDo[a- > b] $\leftarrow$ HMAC(Key[a- > b], IDb);
    if a DATA frame is received then
        for each neighbor do
            if FCS checksum OK and IDD[a- > b] from frame = = IDD[a- > b] then
                //save old addresses
                IDSo[a- > b] $\leftarrow$ IDS [a- > b];
                IDDo[a- > b] $\leftarrow$ IDD [a- > b];
                //generate new key
                Key[a- > b] $\leftarrow$Key[a- > b] $\oplus$ Hash(payload);
                //generate new addresses
                IDS[a- > b]<-HMAC(Key[a- > b], IDa);
                IDD[a- > b]<-HMAC(Key[a- > b], IDb);
                deliver frame to upper layer;
                send ACK frame { IDDo[a- > b], IDSo[a- > b], FCS};
            else if FCS checksum OK and IDD[a- > b] from frame = = IDDo[a- > b] then
                //DATA frame already delivered
                send ACK frame { IDDo[a- > b], IDSo[a- > b], FCS};
            end
        end
    end
end
```

Algorithm 1: Address anonymity.

distribution of traffic in a certain time period. If the attack nodes are deployed dense enough, the network traffic outline can be drawn by the adversary. Figure 7 illustrates such attack. Obviously, traffic outlining attack cannot be eliminated by address anonymity and random delay.

The solution to traffic outlining attack is circular traffic, which is illustrated in Figure 8. Network traffic from the source node to the base station follows two semicircle paths. And the two semicircle paths form a circular path. Source node selects one of the two semicircles randomly to forward packet. As to the adversary, traffic outlining attack cannot find the source node and base station because traffic in the networks forms a circular path (Algorithm 3).

The proposed solution can eliminate traffic outlining attacks effectively, because the source node and base station are hidden in a circular traffic outline. Combined with

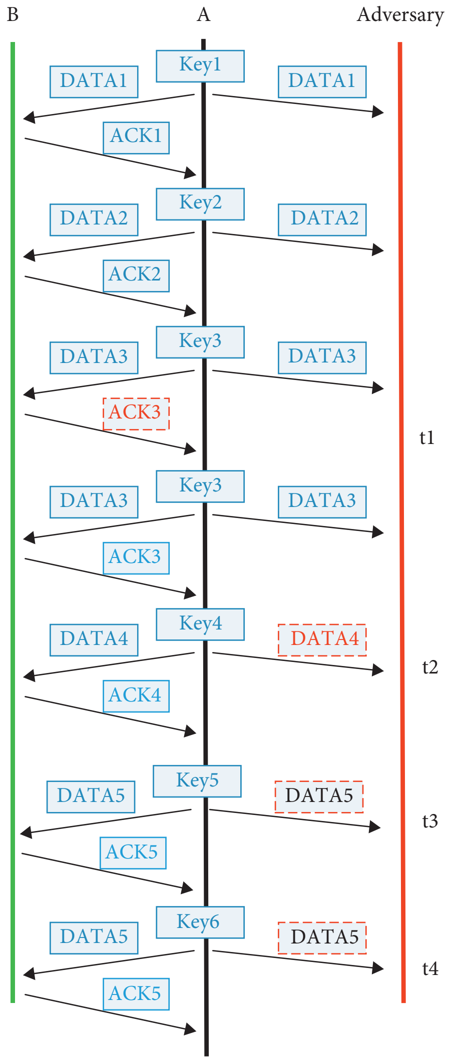


Figure 3: Frame error leads to address confusion.

address anonymity and random delays, traffic analysis attacks can be well addressed.

4.3. Performance Analysis. Efficiency is among the most important design considerations for data link layer schemes. Two different hash functions are used in HASHA, hash(), and HMAC(). hash() is used to generate hash value of payload, and the output is further used to create the key for HMAC(). According to the characteristics of HMAC() function, the computation complexity of brute-force attacks on hash key is 2^k, where k is the length of the key. Long key improves security strength. Therefore, the hashed results of hash() should be long enough to defend against brute-force attacks. Tiger/192 [38] is a good candidate for hash() because it is almost as fast as CRC32, but the width of the output is 192 bits. As HMAC() is used to create one-time address, performance is the top design consideration for HMAC() selection. Efficient MAC functions such as UMAC/32 [39] are a good candidate for HMAC().

Suppose that UMAC/32 is used for HMAC() and Tiger/192 is used for hash() in HASHA. According to the performance analysis of hash functions [38], the performance of UMAC/32 is 1 cycle per byte and Tiger/192 is 8.1 cycles per byte. From the illustrated HASHA process, hash() is called 1 time and HMAC() is called 2 times for both the sender and the receiver to transmit one frame. Supposing that the length of frame is *len* bytes and the MAC address is fixed to 6 byte, HASHA requires $len \ * 8.1 + 2 * 1$ cycles for one frame transmission and reception.

5. Performance Simulation

We use ns-2 to evaluate the energy consumption of HASHA. Several nodes are deployed over 200 m ∗ 200 m network field, and the base station is located at the center. The nodes' radio transmission radius is 50 m.

We deploy only one source node to report event to the base station. The total number of nodes in WSN changes from 50 to 400 in 50 steps. We record the average power

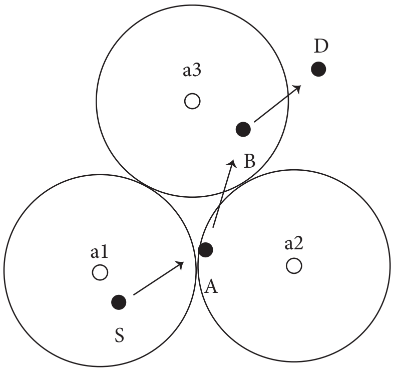


Figure 4: Example topology for time correlation attacks.

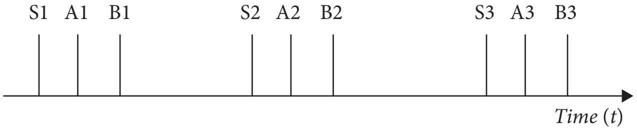


Figure 5: Time serial for reporting event from the source node to base station.

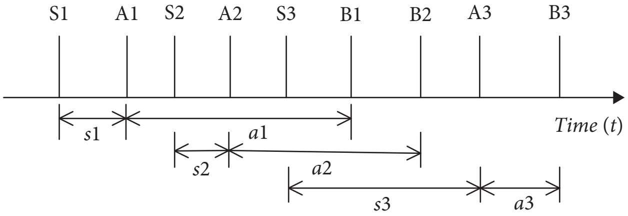


Figure 6: Nodes forwarding packets with random delays.

```
Node maintains a table with entry <data, time_to_transmit> to store data to be forwarded;
Node maintains clock timer, which is used for data transmission;
For data requested to be transmitted:
  Generate a random time rand;
  Insert into the table with entry <data, timer + rand>;
Node search the table to find data that could be transmitted:
  for each entry in the table do.
    if timer ≥ entry. time_to_transmit then
      Transmit the data;
    end
  end
```

Algorithm 2: Forwarding random delay.

consumption of HASHA and Phantom Routing [19], a well-known random location privacy preserve scheme. The power consumption of hash functions and wireless transmission and reception is listed in Table 1.

Figure 9 illustrates the overall power consumption of HASHA and Phantom Routing under different network size. While the size of the network is small (for example, 20 or 50 nodes), HASHA consumes more power than Phantom Routing. The reason is that hash operation is required for both transmitter and receiver, which introduce additional power consumption. Phantom Routing creates routing path longer than the shortest path. But as the network size is fairly small, the additional energy consumption for additional path is much less than hash operation. Therefore, the energy consumption of Phantom Routing is lower. As the size of network increased, the energy wasted on additional path increased dramatically. And that portion of energy cost is much greater than energy cost for hash operations.

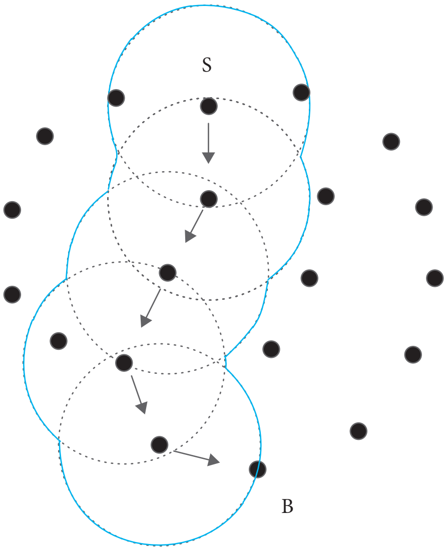


Figure 7: Traffic outlining attacks against event-triggered WSNs.

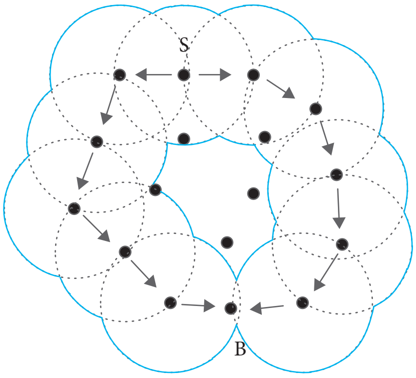


Figure 8: Circular traffic against traffic outlining attacks.

(1) Find the shortest path from the source node to base station according to routing protocol such as dijkstra.
(2) The base station calculates hops **n** from source node to base station and requests the node **n/2** hops away to initiate a circular forwarding path.
(3) The selected node broadcasts beacons which includes a counter with initial value **n/2**.
(4) All nodes that received the broadcasts decrease the value and forward it.
(5) All nodes that received the broadcasts with value 0 are candidates for circular forwarding.
(6) On having data to be sent, the source node selects one of the paths randomly to forward the data to the base station.

Algorithm 3: Circular forwarding against traffic outlining attacks.

Table 1: Power consumption of key operation.

Operation	Consumed energy
UMAC/32 hashing	0.143 uJ/byte
Tiger/192 hashing	0923 uJ/byte
Transmitting	5.623 uJ/byte
Receiving	6.39 uJ/byte

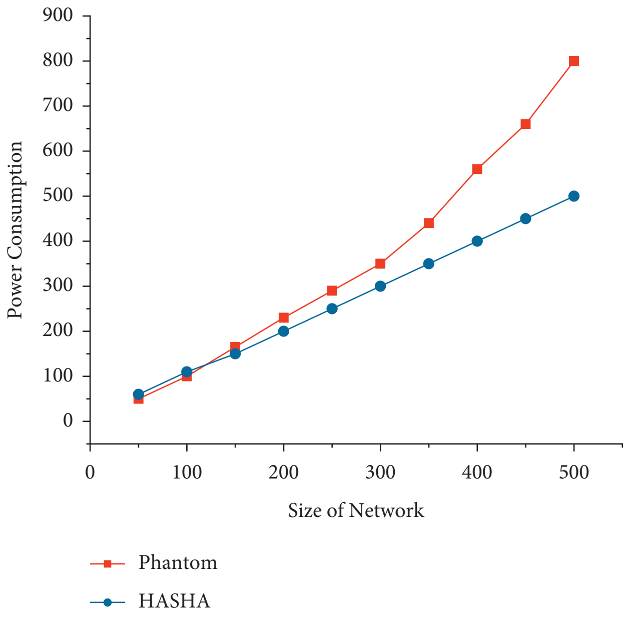


FIGURE 9: Power consumption of HASHA and Phantom Routing.

6. Conclusions

In this paper, we have identified that location privacy cannot be preserved efficiently at network layer, because address at data link layer is not protected well. The address at data link layer exposes node identity and packet routing information to the adversary. Traffic analysis attacks can be easily launched to jeopardize location privacy. HASHA scheme, which hides the addresses at data link layer, is proposed to protect location privacy. Analytical and simulation results show that HASHA is more energy efficient than traditional approaches [40, 41].

References

[1] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, “Applications of wireless sensor networks: an up-to-date survey,” *Applied System Innovation*, vol. 3, no. 1, p. 14, 2020.

[2] S. Ali, T. Al Balushi, Z. Nadir, and O. K. Hussain, *WSN Security Mechanisms for CPS”, Cyber Security for Cyber Physical Systems*, pp. 65–87, Springer, New York, NY, USA, 2018.

[3] W. Al Shehri, “A survey on security in wireless sensor networks,” *International journal of Network Security & Its Applications*, vol. 9, no. 1, pp. 25–32, 2017.

[4] Q. Shafi, “Cyber physical systems security: a brief survey,” in *Proceedings of the 12th IEEE International Conference on Computational Science and Its Applications*, pp. 146–150, Brazil, June 2012.

[5] Y. Li and J. Ren, “Preserving source-location privacy in wireless sensor networks,” in *Proceedings of the 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 1–9, IEEE, Italy, June 2009.

[6] L. Zhou and Y. Shan, “Multi-branch source location privacy protection scheme based on random walk in WSNs,” *IEEE*, in *Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis*, pp. 543–547, IEEE, China, April 2019.

[7] M. Conti, J. Willemsen, and B. Crispo, “Providing source location privacy in wireless sensor networks: a survey,” *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1238–1280, 2013.

[8] A. Tsitroulis, D. Lampoudis, and E. Tsekleves, “Exposing WPA2 security protocol vulnerabilities,” *International Journal of Information and Computer Security*, vol. 6, no. 1, pp. 93–107, 2014.

[9] C. Ozturk, Y. Zhang, and W. Trappe, “Source-location privacy in energy-constrained sensor network routing,” in *Proceedings of the 2nd ACM workshop on Security of Ad hoc and Sensor Networks, ser. SASN ’04, ACM*, Washington, DC, USA, October 2004.

[10] D. Braginsky and D. Estrin, "Rumor routing algorithm for sensor networks," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications, ser. WSNA '02, ACM*, vol. 9, pp. 22–31, ACM, New York, NY,USA, September 2002.

[11] E. Ekici, S. Vural, J. McNair, and D. Al-Abri, "Secure probabilistic location verification in randomly deployed wireless sensor networks," *Ad Hoc Networks*, vol. 6, no. 2, pp. 195–209, 2008.

[12] Y. Li, L. Lightfoot, and J. Ren, "Routing-based source-location privacy protection in wireless sensor networks," in *Proceedings of the IEEE International Conference on Electro/Information Technology*, vol. 6, pp. 29–34, IEEE, Canada, January 2009.

[13] I. Shaikh, H. Jameel, B. dAuriol, H. Lee, S. Lee, and Y.-J. Song, "Achieving network level privacy in wireless sensor networks," *Sensors*, vol. 10, no. 3, pp. 1447–1472, 2010.

[14] W. Yang and W. T. Zhu, "Protecting source location privacy in wireless sensor networks with data aggregation," *Ubiquitous Intelligence and Computing*, vol. 10, pp. 252–266, 2010.

[15] B. Alomair, A. Clark, J. Cuellar, and R. Poovendran, "Towards a statistical framework for source anonymity in sensor networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 12, pp. 1–12, 2011.

[16] M. Mahmoud and X. Shen, "A cloud-based scheme for protecting source-location privacy against hotspot-locating attack in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 10, pp. 1805–1818, 2012.

[17] P. Kamat, Y. Zhang, W. Trappe, and C. Ozturk, "Enhancing source-location privacy in sensor network routing," in *Proceedings of the 25th IEEE international conference on distributed computing systems*, pp. 599–608, IEEE, June 2005.

[18] M. Shao, Y. Yang, S. Zhu, and G. Cao, "Towards statistically strong source anonymity for sensor networks," in *Proceedings of the The 27th Conference on Computer Communications, ser. INFOCOM 2008*, vol. 4, pp. 51–55, IEEE, Columbus, OH, USA, March 2008.

[19] H. Chen and W. Lou, "From nowhere to somewhere: protecting end-to-end location privacy in wireless sensor networks," in *Proceedings of the Performance Computing and Communications Conference, IEEE 29th International*, vol. 12, pp. 1–8, IEEE, Piscataway, USA, December 2010.

[20] S. Tilak, N. B. Abu-Ghazaleh, and W. Heinzelman, "A taxonomy of wireless micro-sensor network models," *ACM SIGMOBILE - Mobile Computing and Communications Review*, vol. 6, no. 2, pp. 28–36, 2002.

[21] Y. Yang, M. Shao, S. Zhu, B. Urgaonkar, and G. Cao, "Towards event source unobservability with minimum network traffic in sensor networks," *ACM* in *Proceedings of the first ACM conference on Wireless network security, ser. WiSec '08*, vol. 4, pp. 77–88, ACM, New York, NY, USA, March 2008.

[22] P. Kamat, Y. Zhang, W. Trappe, and C. Ozturk, "Enhancing source-location privacy in sensor network routing," in *Proceedings of the 25th IEEE International Conference on Distributed Computing Systems*, pp. 599–608, Columbus, OH, USA, June 2005.

[23] X. Hong, P. Wang, J. Kong, Q. Zheng, and J. Liu, "Effective probabilistic approach protecting sensor traffic," in *Proceedings of the IEEE Military Communication Conference*, pp. 169–175, Atlantic City, NJ, USA, 2005.

[24] S. Jiang and N. H. Vaidya, W. Zhao, "Routing in packet radio networks to prevent traffic analysis," in *Proceedings of the IEEE Information Assurance and Security Workshop*, West Point, NY, USA, February 2000.

[25] J. Deng, R. Han, and S. Mishra, "Intrusion tolerance and anti-traffic analysis strategies for wireless sensor networks," in *Proceedings of the IEEE International Conference on Dependable Systems and Networks*, Florence, Italy, July 2004.

[26] J. Deng, R. Han, and S. Mishra, "INSENS: intrusion-tolerant routing for wireless sensor networks," *Computer Communications*, vol. 29, no. 2, pp. 216–230, 2006.

[27] J. Deng, R. Han, and S. Mishra, "Decorrelating Wireless sensor network traffic to inhibit traffic analysis attacks," *Pervasive and Mobile Computing*, vol. 2, no. 2, pp. 159–186, 2006.

[28] H. Gao, C. Liu, Y. Yin, Y. Xu, and Y. Li, "A hybrid approach to trust node assessment and management for VANETs cooperative data communication: historical interaction perspective," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021.

[29] H. Gao, L. Zhou, J. Y. Kim, Y. Li, and W. Huang, "The behavior guidance and abnormality detection for A-MCI patients under wireless sensor network," *ACM Transactions on Sensor Networks*, 2021.

[30] S. Olariu, M. Eltoweissy, and M. Younis, "ANSWER: autonomous wireless sensor network," in *Proceedings of the 1st ACM International Workshop on Quality of Service & Security in Wireless and Mobile Networks (Q2SWinet'05)*, pp. 88–95, Montreal, Quebec, Canada, October 2005.

[31] J. Kong, X. Hong, and M. Gerla, "An identity-free and on-demand routing scheme against anonymity threats in mobile ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 8, pp. 888–902, 2007.

[32] Y. Yanchao Zhang, W. Wei Liu, W. Wenjing Lou, and Y. Fang, "MASK: anonymous on-demand routing in mobile ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 9, pp. 2376–2385, 2006.

[33] H. Gao, Y. Zhang, H. Miao, R. J. D. Barroso, and X. Yang, "SDTIOA: modeling the timed privacy requirements of IoT service composition: a user interaction perspective for automatic transformation from bpel to timed automata," *Mobile Networks and Applications*, vol. 26, no. 6, pp. 2272–2297, 2021.

[34] H. Gao, X. Qin, R. J. D. Barroso et al., "Collaborative learning-based industrial IoT API recommendation for software-defined devices: the implicit knowledge discovery perspective," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 1, pp. 66–76, 2022.

[35] J. Al-Muhtadi, R. Campbell, A. Kapadia, and M. Dennis, "Routing through the mist: privacy preserving communication in ubiquitous computing environments," in *Proceedings of the 22nd International Conference on Distributed Computing Systems*, p. 74, Vienna, Austria, July 2002.

[36] Y. Huang, H. Xu, H. Gao, X. Ma, and W. Hussain, "SSUR: an approach to optimizing virtual machine allocation strategy based on user requirements for cloud data center," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, pp. 670–681, 2021.

[37] X. Ma, H. Xu, H. Gao, and M. Bian, "Real-time multiple-workflow scheduling in cloud environments," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4002–4018, 2021.

[38] J.-H. Park, Y. Jung, H. Ko, J. Kim, and M. Jun, "A privacy technique for providing anonymity to sensor nodes in a sensor network," in *Proceedings of the International*

Conference on Ubiquitous Computing and Multimedia Applications, Springer, Berlin, Heidelberg, 2011.

[39] J. Black, S. Halevi, H. Krawczyk, T. Krovetz, and P. Rogaway, "UMAC: fast and secure message authentication," in *Advances in Cryptology - CRYPTO' 99*, M. Wiener, Ed., vol. 1666, pp. 216–233, Springer, Berlin, Germany, 1999.

[40] S. Misra and G. Xue, "Efficient anonymity schemes for clustered wireless sensor networks," *International Journal of Sensor Networks*, vol. 1, no. 1/2, pp. 50–63, 2006.

[41] K. Zhang and Q. Zhang, "Preserve location privacy for cyber-physical systems with addresses hashing at data link layer," in *Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications*, pp. 1028–1032, Exeter, UK, June 2018, https://ieeexplore.ieee.org/document/8622908.

Real-Time Information Exchange Strategy for Large Data Volumes Based on IoT

Rasmi Sarangi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, rashmisarangi14@gmail.com*

Ashok Muduli, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, ashok.muduli29@yahoo.co.in*

Sushree Sangita Jena, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sushreesangita665.com*

Prakash Dehury, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, prakashdehury1@hotmail.com*

Abstract

In this paper, we study and analyse the real-time information exchange strategy of big data in the Internet of Things (IoT) and propose a primitive sensory data storage method (TSBPS) based on spatial-temporal chunking preprocessing, which substantially improves the speed of near real-time storage and writing of microsensory data through spatial-temporal prechunking, data compression, cache batch writing, and other techniques. The model is based on the idea of partitioning, which divides the storage and query of perceptual data into the microperceptual data layer and the perceptual data layer. The microaware data layer mainly studies the storage optimization and query optimization of raw sensory data and cleaned valid data; the aware data is the aggregation and statistics of microaware data, and the aware data layer mainly studies the storage optimization and query optimization of aware data. By arranging multiple wireless sensors at key monitoring points to collect corresponding data, building the core data service backend of the system, defining multifunctional servers, and constructing an optimal database model, we effectively solve the parameter collection and classification aggregation processing of different devices. To address the requirement of reliable and secure transmission in the process, we design a highly concurrent and high-performance TCP-based socket two-layer transmission framework and introduce the asymmetric encryption method (RSA) and data integrity verification method to design a transmission protocol that is both reliable and secure. The integration of big data and IoT is bound to bring the intelligence of human society to a new level with unlimited development prospects.

1. Introduction

As the number of IoT sensing devices proliferates, the scale and impact of IoT are expanding, making the market for IoT also growing [1]. The services of IoT applications are based on the collected data, so the core of IoT is data. With the increase of the number of IoT sensing devices, the sensing devices in various industries generate a huge amount of sensing data every day. The basis of IoT development is extended and expanded based on Internet, and its ultimate development goal is to achieve comprehensive sensing, reliable transmission, and intelligent processing [2]. The network architecture of IoT can be divided into three layers: the first is the perception layer, whose main role is to collect information, information processing, and other operations (through radio frequency identification devices, infrared sensors, card readers, etc.); the second is the network layer, whose main role is to transmit information (through mobile networks, the Internet, broadband networks, wireless networks, etc.); and the third is the application layer, whose main role is to complete the analysis and processing of information and control and decision-making. Among them, the network layer is the link between the sensing layer and the application layer for information exchange. Through processing and sharing of sensing information, the application layer provides powerful resources to support the processing of various businesses, thus truly realizing the intelligence and informatization of various industries. So far, the development of IoT technology has been very extensive; for example, smart city, intelligent medical care, intelligent transportation, intelligent home, intelligent agriculture, and many other fields are used in IoT [3].

The Internet of Things (IoT) continues to evolve as increased people use more easily connected devices, modified to the current time. The experiment is to get the maximum writing speed and the maximum average speed of the two writing methods, where the maximum writing speed refers to the maximum instantaneous writing speed that can be achieved during the writing process, and the average maximum writing speed refers to maintaining stability. Under the premise of not exceeding the Redis cache threshold, the maximum write speed can be maintained all the time. The result is an exponential distribution of data availability. With this vast amount of information, how can we find truly valuable data? The development of IoT systems has led to the rapid development of deep learning, and the successful application of vision-based target tracking in the fields of autonomous driving, behavioural analysis, intelligent surveillance, and virtual reality has gradually made it the focus of research in the field of deep learning and IoT technologies, and these applications require the processing of large amounts of spatial-temporal data [4]. For example, in the field of autonomous driving, the target tracking algorithm should be able to detect passers-by walking on the road and follow moving cars in real-time and successfully predict and judge their subsequent speed, trajectory, and other spatial-temporal data information; in the field of virtual reality, real-time human-machine interaction should be completed based on the motion trajectory captured by the camera [5]. However, practical applications of the system will often suffer from system lag, untimely feedback, and abnormalities in the collected spatial-temporal data, while the IoT system needs to be able to quickly provide feedback and processing of these collected data [6, 7].

Therefore, to achieve target tracking and an IoT search system that can provide fast and correct feedback, it is especially important to design an efficient IoT data processing method. Based on the diversity of spatial-temporal data of IoT system, a large amount of data, real-time, sensor node instability, and other characteristics, this paper proposes a target tracking-oriented IoT data processing technology, which can use a deep learning model to quickly classify the spatial-temporal data collected by IoT, clean the abnormal spatial-temporal data, and finally design and implement an efficient spatial-temporal data processing-based target tracking IoT search system.

1.1. Current Status of Research. Chin et al. proposed an Ethernet-based hybrid simulation technology solution for the Industrial Internet of Things, which uses PLCs to connect devices via Ethernet and uses a virtual environment running in parallel with the plant floor equipment as a reference to analyse performance, evaluate manufacturing system performance in real-time, and transfer data and coordinate actions [8]; Sankar et al. investigated some selected test methods for real-time Ethernet technology closely related to CNC system performance and also gave test scenarios and Ether CAT case studies to illustrate the feasibility of the designed test system with methods that can simply evaluate real-time Ethernet used in CNC systems to ensure the performance of CNC systems [9]; Li et al. studied the performance of Ethernet networking approach for Linux NC open-source CNC systems, which solves the real-time communication problem between system components and provides a new approach to integrate real-time Ethernet into Linux NC, realizing a CNC system that is completely based on open-source software and has performance that can compete with proprietary embedded CNCs [10]. With the continuous development of big data storage and query technology, IoT-aware big data technology is also moving forward, but there is still a lack of system solutions for efficient storage and fast query of perceptual big data [11]. In this paper, we study the high throughput writing technology and fast query technology for IoT-aware big data and based on the idea of data hierarchy and according to existing big data technology and theory, this paper implements a hierarchical storage and query system model (IoT-HSQM) for IoT-aware big data, which provides a solution for near real-time storage and fast-statistical analysis of IoT-aware big data [12].

Researchers at Virginia State University designed the Snuggle system, in which entities are described using a set of keywords (text messages) stored in each sensor node, using keywords to interrelate with physical entity sensors, and using keyword information to represent IoT entities, so that users can directly use keywords to search for IoT hardware nodes that match the query target hardware nodes; the system can then return the query to the most relevant and matching spatial-temporal data information collected by the K IoT hardware nodes [13–15].

With the increasing number of IoT hardware nodes, the spatial-temporal data collected by the hardware nodes have the characteristics of high dimensionality, complexity, and real-time at the same time, which leads to the increasing amount of data transmitted by the IoT system, the increasing difficulty of data processing, and the increasingly complex network structure between IoT nodes. And due to the instability of network transmission and the characteristics of hardware nodes and unreasonable storage mechanisms, it is easy to cause a variety of problems such as abnormal data saved to the background and slow system search efficiency. Besides, when target tracking is performed, the target tracking fails due to light, occlusion, and oversized network model. The data is stored into different data blocks according to the different characteristics of the spatial-temporal data collected by the sensors. When the data collected by sensors need to be transmitted to the background in real-time, information entropy is used to classify and store the data quickly, and a method is proposed for processing the transient abnormal data collected by IoT nodes, using the Influx DB time-series database with timestamps for partial IoT data storage to facilitate the construction of IoT systems and improve the search efficiency.

2. Analysis of Real-Time Exchange Strategy for IoT Large Data Volume

2.1. Analysis of IoT Data Processing Methods. IoT spatial-temporal data refers to data with spatial and temporal dimensions, and the data includes thematic attributes, time, and space (geographic location information) and has characteristics such as multisource sensing, large data volume, and high real-time. However, processing this spatial-temporal data is complex, and in contrast to static data where the image of the same car does not vary much between two adjacent frames, spatial-temporal data is completely different, with more image variation [16–18].

With the rapid rise of the Internet of Things (IoT), its technology and infrastructure are gradually improving, thus using the characteristics of IoT can bring us a different life: the current mutual communication of information in only 2 dimensions (e.g., any time, any place) can be extended to another dimension through IoT, i.e., the 3rd dimension: communication between any objects, as shown in Figure 1.

The main part of the IoT reference model can be divided into 4 layers: the application layer, the business/application support layer, the network layer, and the device layer. Also, cross-layer management capabilities and security capabilities are included. Among them, the application layer refers to the wide variety of IoT applications that users can eventually see. The business/application support layerincludes two kinds of capabilities: general-purpose support capability and dedicated support capability . At the same time, the abnormal IoT data collected in real-time is cleaned, and then different types of IoT data are stored in different Influx DB data blocks. The system adopts a distributed storage architecture to speed up the indexing rate. The management capability of IoT contains fault management, configuration management, settlement, performance, security, etc. The management capabilities of IoT can also be divided into generic management capabilities and dedicated management capabilities due to the difference in the needs of vertical and public industries. Security capabilities exist in the application layer, network layer, business/application support layer, and device layer. Security capabilities can also be divided into generic security capabilities and dedicated security capabilities due to the difference in demand between vertical industries and public industries.

Based on the existing network infrastructure, the business capability layer (the middlebox) provides new IoT capabilities for traditional industry devices and customers' existing IT systems to meet the needs of IoT services. Both the network domain and the terminal domain contain the application service layer, the business capability layer, and the network connectivity layer that connects the two. The management support system exists in the application service layer, business capability layer, and network connectivity layer of both the network domain and terminal domain.

Due to the IoT real-time search system, the collected data is continuous and usually, not much variation is found between adjacent numbers.

$$X_{tn} = \frac{\alpha x_{t-2} - \beta x_{t-1} - \gamma x_{t+2} + x_t}{5},$$
$$f_t = \left\{ \delta\left(W^{(f)} x_{t+1} + U^f h_{t+1}, \delta\left(W^{(f)} x_{t+1} - U^f h_{t+1}, C_t = \left(1 + f_t\right) \cdot C_t' - f_t\right) * C_t'. \right.\right. \tag{1}$$

A single update gate is generated by combining the input gate with the forget gate.

$$\begin{aligned} y &= x_t \sin\left(h_t - 1\right), \\ y &= x_t \sin\left(h_t - 1\right). \end{aligned} \tag{2}$$

When analysing the operation of industrial equipment, it is often impossible to know in advance the variation of an indicator, i.e., the distribution that the indicator as a whole follows, to calculate the mean, variance, extreme deviation, etc., of the data collected, and obtaining this distribution through hypothesis testing. Besides, it is possible to design a safety range check of the indicator with the help of the mean and variance.

$$\begin{aligned} x &= \frac{ii}{n} \sum_{i=1}^{n} x_i^2, \\ S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 + x^2, \\ S &= \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} x_i^2 + x^2}, \\ r &= x_n^* + x_1^*. \end{aligned} \tag{3}$$

By establishing a mapping function between the controllable and target variables, the attribute characteristics of the data are depicted and presented based on the time stamp. The main technical means is to establish a regression model of the data to obtain the corresponding regression function so that, given an input value to obtain the corresponding target value, the target value is compared with the measured value to obtain the foreseen result. It is usually used for studies of trends in indicators, forecasting of target values, and correlations between variables.

2.2. Analysis of Real-Time Information Exchange Strategies for Large Data Volumes. For this reason, we build the logical structure of Influx DB-based IoT spatial-temporal data as shown in Figure 2, where the real-time IoT data is first stored to the edge nodes, and the EPLSN algorithm calculates and classifies the real-time collected data on the edge nodes and cleans up the abnormal IoT data collected in real-time. The system adopts a distributed storage architecture. The system adopts distributed storage architecture to accelerate the indexing rate [19].

The data from the IoT nodes are monitored and tracked, such as hydrogen sulphide gas concentration, dairy farm temperature and humidity, and real-time target tracking [20].

For the data writing throughput of the microaware data layer, different amounts of data from the Beijing cab sensor

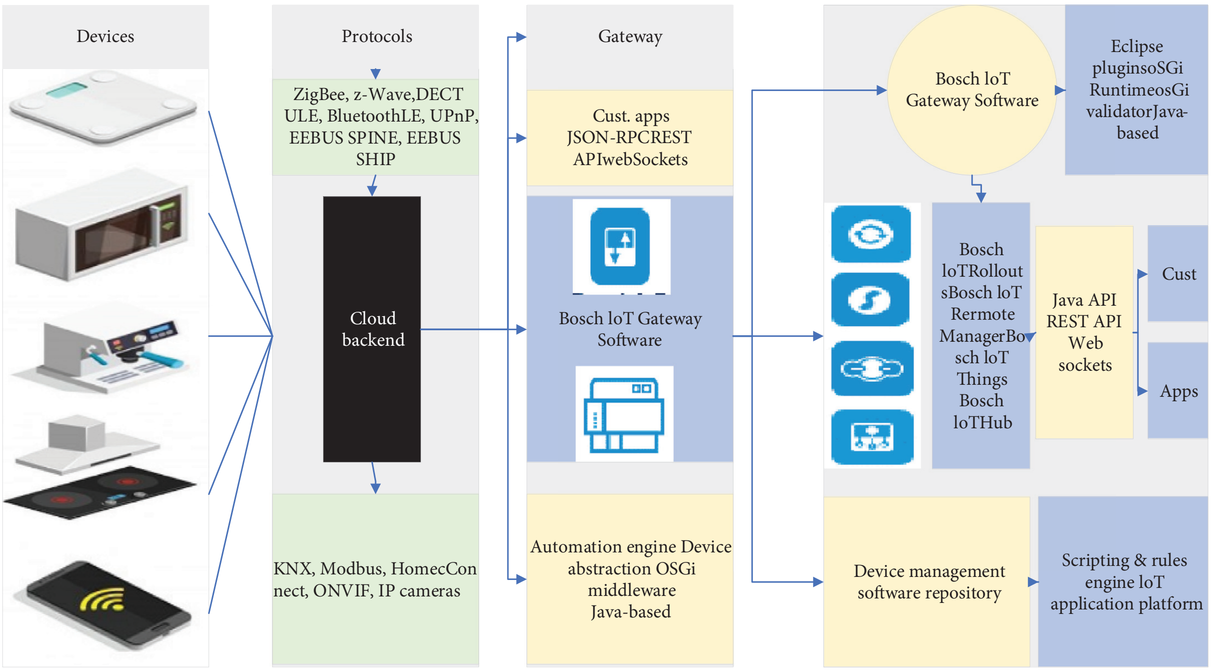

Devices
Protocols
Gateway
ZigBee, z-Wave,DECT ULE, BluetoothLE, UPnP, EEBUS SPINE, EEBUS SHIP
Cloud backend
KNX, Modbus, HomeCon nect, ONVIF, IP cameras
Cust. apps JSON-RPCREST APIwebSockets
Bosch loT Gateway Software
Automation engine Device abstraction OSGi middleware Java-based
Bosch loT Gateway Software
Eclipse pluginsoSGi RuntimeosGi validatorJava-based
Bosch loTRollout sBosch loT Rermote ManagerBo sch loT Things Bosch loTHub
Java API REST API Web sockets
Cust
Apps
Device management software repository
Scripting & rules engine loT application platform

Figure 1: IoT model architecture.

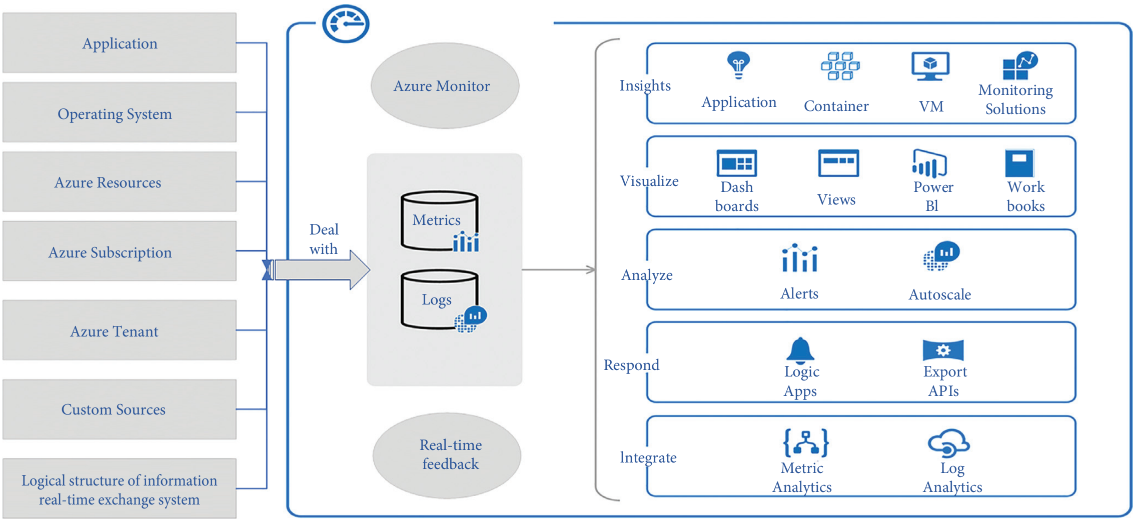


Figure 2: Logical structure of information real-time exchange system.

history are written to HDFS using the directory structure and file naming of the IoT-HSQM model. The average and maximum data write speeds are compared between the direct data storage in HDFS and the time-space block-based caching and clustering approach. To ensure the accuracy of the test results, most of the raw perceptual data are cached in memory in advance, and the occurrence time of the data is modified to the current time in advance when writing out.

The main goal of real-time extraction is to guarantee the real-time nature of data extraction and aggregation of aware data through big data processing techniques. Big data processing techniques are divided into two categories: batch processing techniques and streaming computing were big data batch processing techniques, suitable for offline processing of historical data. Streaming computing is a microbatch data processing method that slices data according to time intervals and processes them with multiple small batch tasks to achieve low latency and near real-time by rapidly executing multiple small tasks. To solve this problem, the accuracy of real-time classification of the Internet of Things is enhanced. The microaware data is streamed into the aware data extraction model, in which the real-time microaware data is appended to a multidimensional data table, and the aggregated computation module obtains the data directly from the multidimensional data table for computation to obtain the aware data.

To meet the different query statistics requirements of the upper layer applications, different aware data need to be preaggregated, but some aggregation calculations have to be performed based on the existing aware data; i.e., new aware data are obtained by reaggregating on top of the aware data as needed. For the convenience of distinction, the data aggregated from the microaware data layer is called fine-grained aware data, and the data reaggregated from the fine-grained aware data is called coarse-grained aware data, e.g., to get the results of a frequent query from the upper layer application that requires a time-consuming statistical analysis of one or more tables of fine-grained data [21]. Support capabilities refer to the basic service capabilities used by various IoT applications, such as basic data processing capabilities and basic data storage capabilities. The coarse-grained aware data is obtained by reaggregating the basic aware data and saving the new data after aggregation.

When extracting from fine-grained aware data, the data source of the data extraction model is changed to the corresponding aware data, and the extraction process is the same as the extraction from the microaware data layer. If the extracted aware data is stored in Druid, you can use Druid's preaggregation function in addition to the real-time extraction model. To use Druid's preaggregation feature, you need to predefine the aggregation granularity size in Druid. When the data is ingested, Druid aggregates the data according to the predefined granularity size; i.e., the data is grouped by timestamp column, dimension column, aggregation column, and aggregation granularity.

The sampled data from sensor sensing devices are transmitted to the data centre through the network, and there will be inconsistencies between the data reaching order and the data generating order. Similarly, the valid data after real-time cleaning, in addition to being stored in the microaware data layer, will be used as the data source for real-time extraction of aware data, so the data timing problem should also be handled. This will help traffic authorities to develop effective policies to reduce traffic congestion. In addition, the data can be used to study the behaviour of taxi drivers, and effective systemscan be designed to detect abnormal behaviour, increase the likelihood of finding new passengers, and take the best route to their destination. In the real-time extraction model, the data sliding time window is designed to store the data received in the past period in the data time window, and the extraction model simply moves the time window forward when the aware data extraction is performed. The data stored in the aware data layer is aggregated aware data with structured or semistructured characteristics, so it is not necessary to perform the complex transformation of aware data for statistical analysis based on aware data.

When caching larger data, the cache tends to be filled up quickly, which will lead to frequent cache replacement actions in the process of continuing the query. Therefore, in equation (6), the larger the cached data block the lower the cache weight, and large blocks of data exceeding the data block size threshold are not cached. If there are queries with long computation time and large and frequent result sets, they should be aggregated and stored as aware data. HRPB method records and saves the query history in the form of logs set periodic timing tasks to analyse the query logs offline and identifies and saves the regular query patterns among them. If the queries that occur frequently together are identified, the method is used in the cache model to use the cache more effectively. If they are not accessed again for some time, they are identified as a phase silent pattern. For example, each time a query is made for the most recent week's statistical results, the data for the past few weeks of the month is queried. After having performed a week of statistics results, none of the weekly statistics will be performed again for a few days. Such as in social media, healthcare, agriculture, transportation, and climate science, a large amount of spatiotemporal data is collected for spatiotemporal data processing and information search. This regular pattern was found to help further improve the cache management in the HRPB method and increase the query speed.

3. Analysis of Results

3.1. Analysis of IoT System Test Results. The data transfer system is tested on the assumption that Kafka and Zookeeper are up and running, the network connection is normal, and the data collection system is working properly. The test mainly examines the throughput and latency performance of the data transfer system in actual use. After starting the system, we monitored and recorded the data of the online system and plotted the throughput and latency curves as shown in Figure 3. The low throughput and high latency at the beginning were due to the initialization and caching work after the system was cold started.

Figure 3 shows that the data transmission rate of the transmission system can reach about 75,000 items/second under stable working conditions, and the delay is also maintained within 40 ms, which can fully meet the application scenario of the manufacturing workshop. The visualization system is tested on the premise that the manufacturing big data processing system works normally. The system efficiency of real-time search of the Internet of Things is accelerated, and an Internet of Things search system is finally realized. The visualization system adopts B/S architecture, and the load is mainly on the server-side, so we focus on the performance of the server-side. Starting from zero loads of the system, we gradually added connected clients and plotted CPU load and memory occupancy as shown in Figure 4. The resource occupation at zero loads in Figure 4 represents the system idle occupation rate, which is occupied by other programs of the server; the resource occupation at a single client connection reflects the resources consumed for system initialization.

The visualization system has low initialization resource occupancy, and the incremental resource occupancy is small and smooth when the number of clients increases. The deep learning model can be used to quickly classify the spatiotemporal data collected by the Internet of Things, clean the abnormal spatiotemporal data, and finally design and implement an efficient target tracking Internet of Things search system based on spatiotemporal data

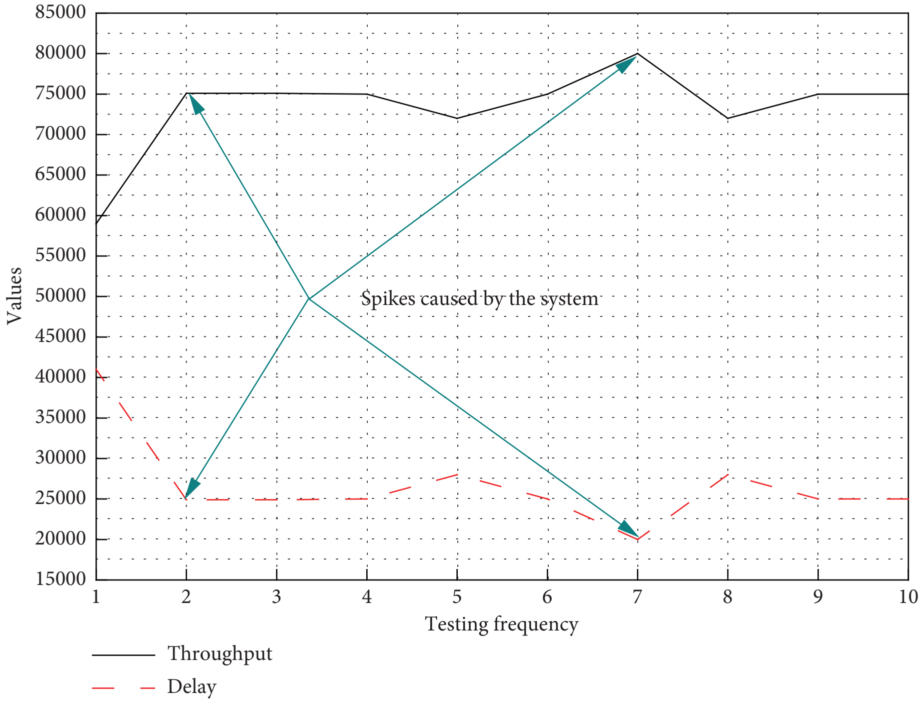


FIGURE 3: Data transmission system throughput vs. latency.

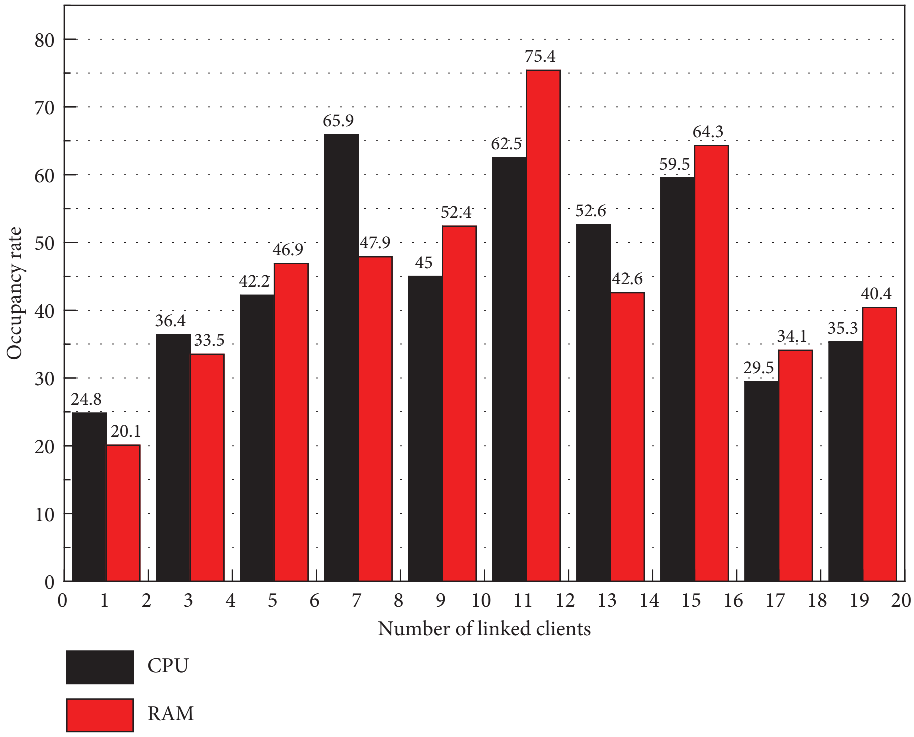


FIGURE 4: CPU and memory usage of the data visualization system.

processing. To evaluate the real-time effect of streaming computing, the visualization system refresh frequency is plotted against the streaming computing frequency as shown in Figure 5, and the minimum computing frequency is greater than the maximum refresh rate to meet the real-time requirement.

To validate the whole system, a hardware experimental platform is built and the steps of deploying and running each component of the system are described in detail. Subsequently, this chapter also demonstrates the operation effect of each module of the system to further verify the system effectiveness. Finally, to guarantee the correctness and

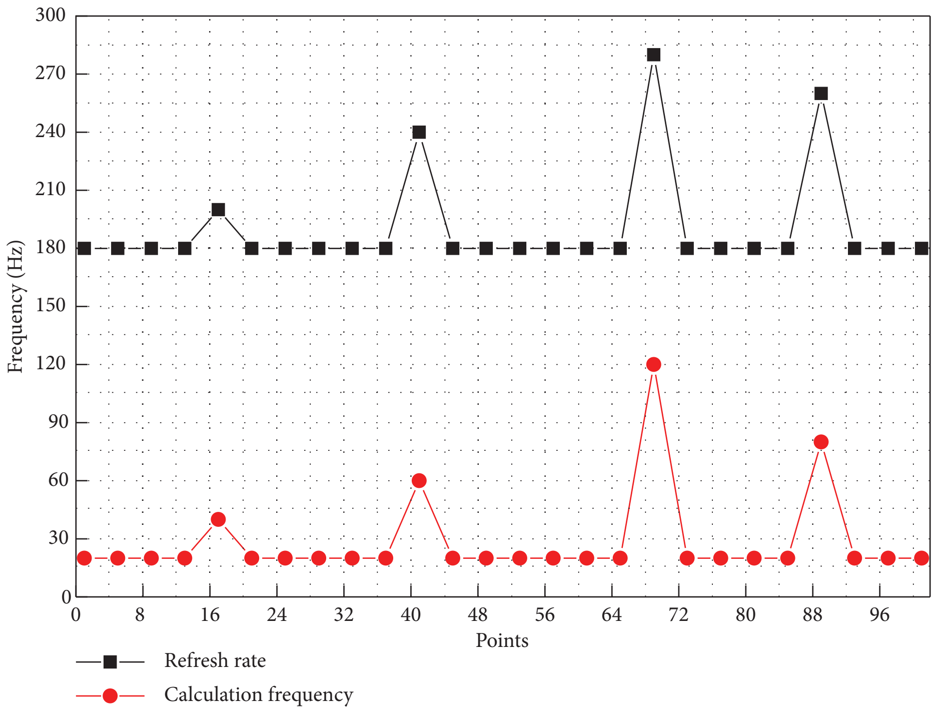


Figure 5: Comparison of streaming frequency and refresh frequency.

stability of the system, several software tests are designed and implemented, including functional tests for verification purposes and performance tests for stress testing purposes. The tests show that the digital production workshop big data processing platform developed in this paper can operate normally and achieve the expected results. A series of tests and practices prove that the digital manufacturing workshop big data processing platform developed in this paper can operate correctly and meet the expected design objectives. However, in practical systems, there are often problems such as system freezes, untimely feedback, and abnormal temporal and spatial data collection. However, IoT systems need to be able to quickly send feedback and process these collected data. A heterogeneous equipment network structure is designed and a unified equipment data collection system is developed. Based on the in-depth study of the characteristics of manufacturing big data, a four-layer network topology is designed, and the role of each layer and the development and implementation methods are elaborated for each network. And through data collection and comparative analysis of field devices, the correctness of the collection system is proved. Finally, various tests were conducted on the experimental platform, and the designed network topology and the developed data acquisition system were able to accomplish the expected objectives.

3.2. Analysis of Experimental Results. There is a limit on the number of connections that can be written at the same time on HDFS, so only some of the file writing connections are kept open, and when the data is continuously written, it looks for open file connections based on the data, and if not, it closes a recently unused connection and looks for the corresponding file on HDFS, establishes a connection, and then writes it. Therefore, a lot of time is spent on establishing and closing connections, as shown in Figure 6.

Figure 7 gives the relationship between the size of the assisted cache space and the terminal cost function, and it can be seen that the cost function of the terminal increases and then decreases, in line with the analysis of the terminal cost function in the previous section, where there exists a minimal value. From the figure, it can be seen that the optimal cache space is different for each terminal since different terminals have different response times, but it can be seen that the shorter the response time, the more the space available to participate in collaborative caching and thus the lower the cost function of the terminal. Terminal 11, when not participating in the assistance cache, has a cost of 1.93, while when participating in the assistance cache space greater than 0.7, its cost increases instead, so its optimal collaboration cache space is 0.7.

A large amount of collected data necessitates a correspondingly powerful underlying framework to support not only the storage and querying but also the valid data extracted from it. To meet the requirements of performing a variety of unique needs, data with various categories of learning capabilities is required, to truly realize the intelligence and informatization of various industries. So far, the development of Internet of Things technology has been very extensive, such as smart cities, smart medical care, smart transportation, smart homes, smart agriculture, and many other fields. For example, the ability to build aggregate models by analysing historical data is one of the most common requirements. However, this can be a very complex process considering that the data may be stored in different networks, let alone in different machines.

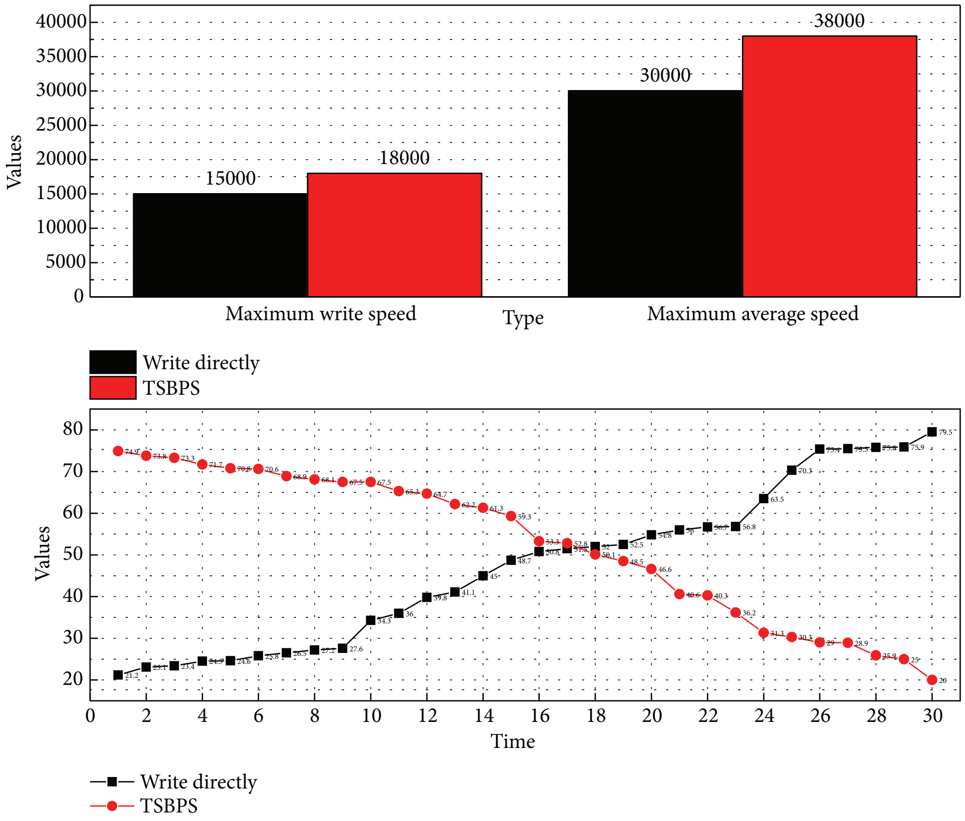


Figure 6: Write speed comparison chart.

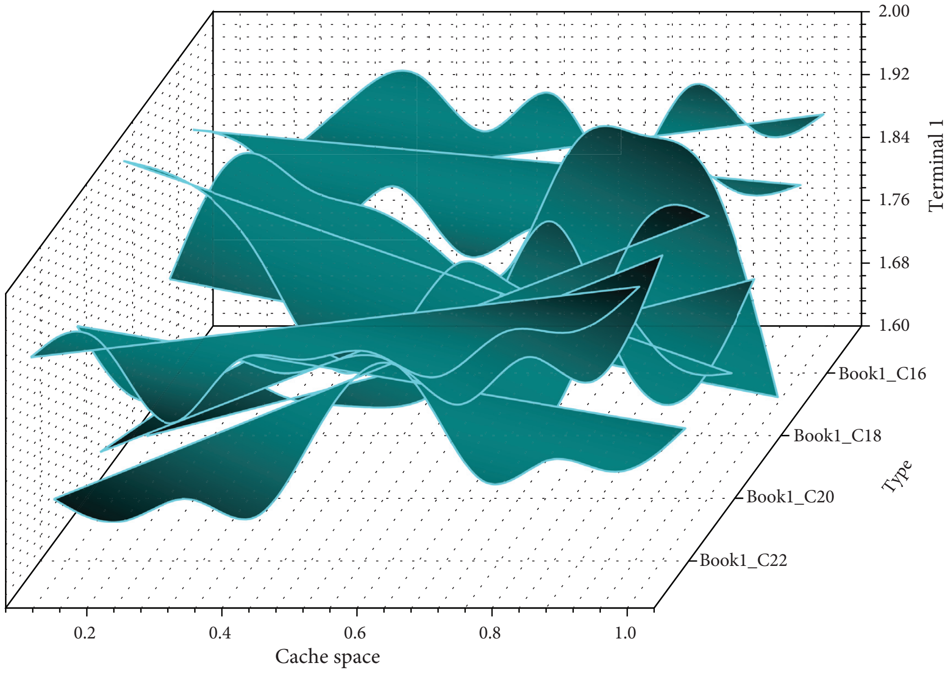
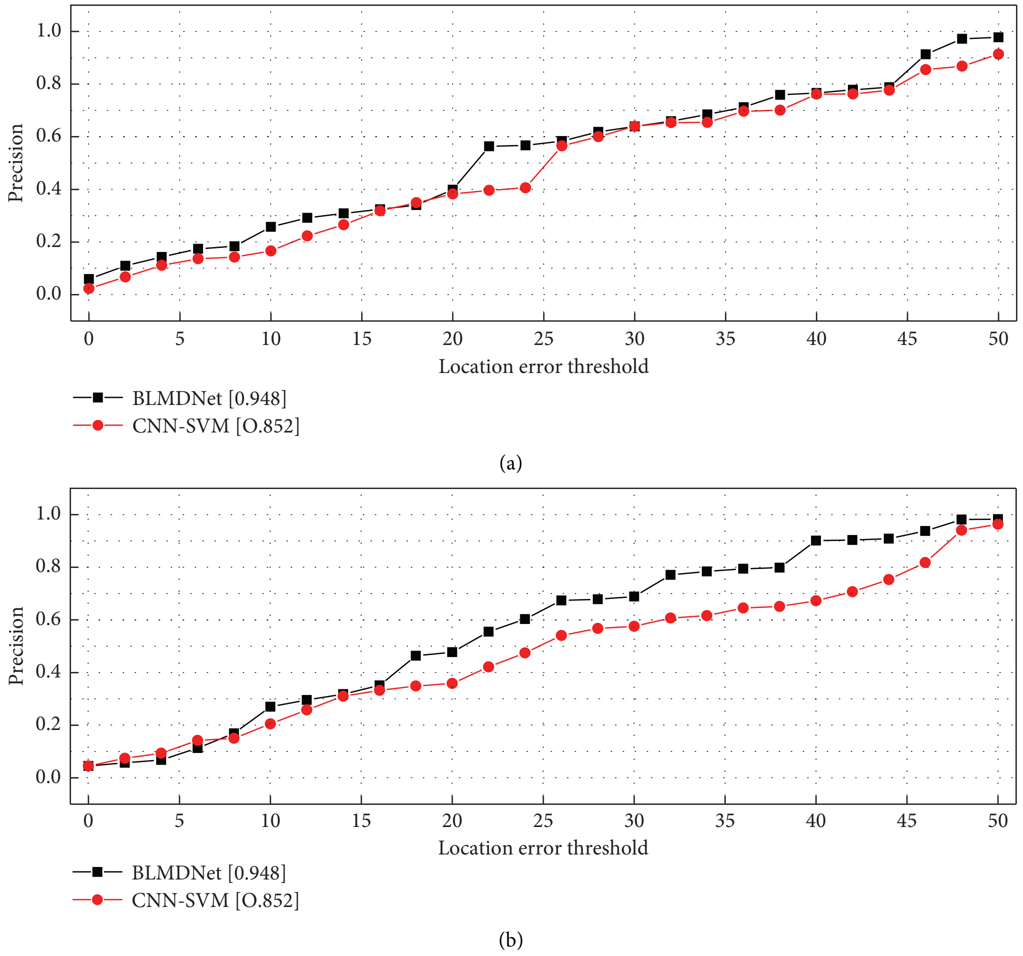


Figure 7: Trend graph of assisted cache space and end cost function.

The BLMDNet model Figure 8 with increased smoothness is obtained by using multiple identical small convolutional layers instead of one larger convolutional layer, which does not introduce more computational effort during the computation. At the same time, because of the feature of smoothness of target motion, the difference between two



(a)

(b)

Figure 8: Graph of predicted rate results.

adjacent frames is small, and the training set of images is expanded, which is higher than the target tracking accuracy of CNN-SVM prediction by about 0.2 for dynamic moving objects. And as the number of training sets of the model increases, the target tracking accuracy also increases. Data is obtained from various sources, which is one of the most significant features of IoT. The services of IoT applications are based on the collected data. Therefore, the core of IoT is data. With the increase in the number of IoT sensing devices, sensing devices in various industries generate massive amounts of sensing data every day. Combining and analysing heterogeneous data is a major challenge. Efforts to standardize data have enabled communication protocols to be developed to enable data exchange.

4. Conclusion

When big data is integrated into IoT, it will inevitably improve the intelligence of human production and life, and its application can be involved in almost all aspects. The microaware data layer proposes a TSBPS method for storing raw perceptual data based on spatial-temporal chunking preprocessing, which significantly improves the speed of storing and writing microaware data in near real-time through spatial-temporal prechunking, data compression, cache batch writing, and other techniques. The real-time aware data extraction model aggregates the aware data in real-time and stores them in the Druid and HBase storage systems in the aware data layer, and the fast-statistical analysis model based on the aware data provides efficient query and statistical analysis of the aware data. Mesosensing data is the aggregation and statistics of microsensing data, and the mesosensing data layer mainly studies the storage optimization and query optimization of mesosensing data. The result data of query and statistical analysis are cached in the aware data layer using Redis, and an HRPB caching method based on historical weights is proposed, which can effectively identify phase hot data to improve the cache hit rate. There are many types of IoT sensing data, and the data with moving sensors and the sensor data with fixed locations are distinguished by whether the sensor location is moving or not. The present model applies to both kinds of data, but it can still be optimized and customized differently according to the type of data stored, and how to make targeted optimization to further improve the performance of the system is one of the directions that can be studied next.

References

[1] E. Keane, K. Zvarikova, and Z. Rowland, "Cognitive automation, big data-driven manufacturing, and sustainable industrial value creation in internet of things-based real-time production logistics," *Economics, Management, and Financial Markets*, vol. 15, no. 4, pp. 39–48, 2020.

[2] S. AlZu'bi, B. Hawashin, M. Mujahed, Y. Jararweh, and B. B. Gupta, "An efficient employment of internet of multimedia things in smart and future agriculture," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 29581–29605, 2019.

[3] B. Ma, Z. Wu, S. Li et al., "Development of a support vector machine learning and smart phone Internet of Things-based architecture for real-time sleep apnea diagnosis," *BMC Medical Informatics and Decision Making*, vol. 20, no. Suppl 14, Article ID 298, 2020.

[4] E. Nica, K. Janoškova, and M. Kovacova, "Smart connected sensors, industrial big data, and real-time process monitoring in cyber-physical system-based manufacturing," *Journal of Self-Governance and Management Economics*, vol. 8, no. 4, pp. 29–38, 2020.

[5] G. Ghosh, M. Kavita, M. Sood, and S. Verma, "Internet of things based video surveillance systems for security applications," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 6, pp. 2582–2588, 2020.

[6] S. Y. Moon, J. H. Park, and J. H. Park, "Authentications for internet of things security: threats, challenges and studies," *Journal of Internet Technology*, vol. 19, no. 2, pp. 349–358, 2018.

[7] D. N. Jha, K. Alwasel, A. Alshoshan et al., "IoTSim-Edge: a simulation framework for modeling the behavior of Internet of Things and edge computing environments," *Software: Practice and Experience*, vol. 50, no. 6, pp. 844–867, 2020.

[8] W.-L. Chin, W. Li, and H.-H. Chen, "Energy big data security threats in IoT-based smart grid communications," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 70–75, 2017.

[9] S. P. Sankar, T. D. Subash, N. Vishwanath, and D. E. Geroge, "Security improvement in block chain technique enabled peer to peer network for beyond 5G and internet of things," *Peer-to-Peer Networking and Applications*, vol. 14, no. 1, pp. 392–402, 2021.

[10] J. Li, A. Maiti, M. Springer, and T. Gray, "Blockchain for supply chain quality management: challenges and opportunities in context of open manufacturing and industrial internet of things," *International Journal of Computer Integrated Manufacturing*, vol. 33, no. 12, pp. 1321–1355, 2020.

[11] M. Habib, I. Aljarah, and H. Faris, "A modified multi-objective particle swarm optimizer-based lévy flight: an approach toward intrusion detection in internet of things," *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6081–6108, 2020.

[12] K. E. Jeon, J. She, P. Soonsawad, and P. C. Ng, "Ble beacons for internet of things applications: survey, challenges, and opportunities," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 811–828, 2018.

[13] S. N. Matheu, J. L. Hernández-Ramos, A. F. Skarmeta, and G. Baldini, "A survey of cybersecurity certification for the internet of things," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1–36, 2020.

[14] N. Miloslavskaya and A. Tolstoy, "IoTBlockSIEM for information security incident management in the internet of things ecosystem," *Cluster Computing*, vol. 23, no. 3, pp. 1911–1925, 2020.

[15] N. B. Gaikwad, H. Ugale, A. Keskar, and N. C. Shivaprakash, "The internet-of-battlefield-things (IoBT)-Based enemy localization using soldiers location and gunshot direction," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11725–11734, 2020.

[16] H. Habibzadeh, K. Dinesh, O. R. Shishvan, A. Boggio-Dandry, G. Sharma, and T. Soyata, "A survey of healthcare Internet of Things (HIoT): a clinical perspective," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 53–71, 2019.

[17] S. N. Mahapatra, B. K. Singh, and V. Kumar, "A survey on secure transmission in internet of things: taxonomy, recent techniques, research requirements, and challenges," *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6211–6240, 2020.

[18] V.-P. Hoang, M.-H. Nguyen, T. Q. Do, D.-N. Le, and D. D. Bui, "A long range, energy efficient Internet of Things based drought monitoring system," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 2, pp. 1278–1287, 2020.

[19] B. Ç. Uslu, E. Okay, and E. Dursun, "Analysis of factors affecting IoT-based smart hospital design," *Journal of Cloud Computing (Heidelberg, Germany)*, vol. 9, no. 1, pp. 67–23, 2020.

[20] P. R. Kumar, A. T. Wan, and W. S. H. Suhaili, "Exploring data security and privacy issues in internet of things based on five-layer architecture," *International Journal of Communication Networks and Information Security*, vol. 12, no. 1, pp. 108–121, 2020.

[21] T. R. Wanasinghe, R. G. Gosine, L. A. James, G. K. I. Mann, O. de Silva, and P. J. Warrian, "The internet of things in the oil and gas industry: a systematic review," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8654–8673, 2020.

Establishing Humanitarian Support Centers in Large Regions at Risk in Mexico Using an Extended GRASP-Capacitated K-Means Clustering Algorithm

Vidya Mohanty, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, vidyamohanty22@outlook.com*

Ashis Acharya, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, ashisacharya12@gmail.com*

Rakhi Jha, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, rakhijha91@yahoo.co.in*

Sidhanta Kumar Balabantaray, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, sk.balabantaray111@gmail.com*

Abstract

Mexico is located within the so-called Fire Belt which makes it susceptible to earthquakes. In fact, two-thirds of the Mexican territory have a significant seismic risk. On the other hand, the country's location in the tropical zone makes it susceptible to hurricanes which are generated in both the Pacific and Atlantic Oceans. Due to these situations, each year many communities are affected by diverse natural disasters in Mexico and efficient logistic systems are required to provide prompt support. This work is aimed at providing an efficient metaheuristic to determine the most appropriate location for support centers in the State of Veracruz, which is one of the most affected regions in Mexico. The metaheuristic is based on the K-Means Clustering (KMC) algorithm which is extended to integrate (a) the associated capacity restrictions of the support centers, (b) a micro Genetic Algorithm μGA to estimate a search interval for the most suitable number of support centers, (c) variable number of assigned elements to centers in order to add flexibility to the assignation task, and (d) random-based decision model to further improve the final assignments. These extensions on the KMC algorithm led to the GRASP-Capacitated K-Means Clustering (GRASP-CKMC) algorithm which was able to provide very suitable solutions for the establishment of 260 support centers for 3837 communities at risk in Veracruz, Mexico. Validation of the GRASP-CKMC algorithm was performed with well-known test instances and metaheuristics. The validation supported its suitability as alternative to standard metaheuristics such as Capacitated K-Means (CKM), Genetic Algorithms (GA), and Variable Neighborhood Search (VNS).

1. Introduction

A *phenomenon* or *disturbing agent* is defined as an aggressive and potentially harmful physical event, natural or derived from human activity, which can cause loss of life or injury, material damage, serious disruption of social and economic life, or environmental degradation. Thus, these agents can have the following origins [1, 2]:

(a) Natural: geological, hydrometeorological, and astronomical.

(b) Anthropogenic: chemical-technological, health-ecological, and social-organizational.

Mexico is located within the so-called *Fire Belt* of the Pacific and within the tropical zone. This makes the country susceptible to a great variety of disturbing agents of natural origin [3]:

(a) Two-thirds of the country have significant seismic risks.

(b) Coastal regions are frequently affected by hurricanes which are generated in the Pacific and Atlantic Oceans.

Due to its geographical location, geological characteristics, and the complex morphology of its territory, the State of

Table 1: Declarations of natural disasters: 2014-2017 [5].

Year	Description	FONDEN (Mexican Pesos)
2014	01. Severe rain and fluvial flood in October 13-16	193,636,015.00
2015	01. Severe rain in March 11-12	1,610,707,729.00
	02. Severe rain in March 21-23 and severe rain and fluvial flood in March 25-27	
	03. Severe rain and fluvial flood in June 11-14	
	04. Hillside movement in July 9-13	
	05. Hillside movement in September 16-18	
	06. Severe rain and fluvial and rain flood in October 16-21	
	07. Severe rain and fluvial and rain flood in October 18-24	
2016	01. Hillside movement in August 5-7	860,195,408.00
	02. Severe rain and fluvial flood in August 5-7	
	03. Severe rain in September 27-28	
2017	01. An earthquake with magnitude 8.2 on September 7	496,947,819.20
	02. Hurricane “Katia” - severe rain and fluvial flood in September 8-12	
	03. Hillside movement from September 27 to October 9	
	04. Severe rain and fluvial flood from September 27 to October 9	
	05. Fluvial flood in October 11-15	
	06. Severe rain in October 11	
	Total	3,161,486,971.20

Veracruz in Mexico is exposed to natural phenomena such as earthquakes, volcanic eruptions, floods, and landslides. The presence of hydrometeorological phenomena is very common in Veracruz, which leads to frequent affectations. In response to the presence of disturbing natural phenomena, in Mexico the Natural Disasters Fund (FONDEN) was created. This is a financial instrument whose purpose is to provide relief supplies and assistance in emergency and disaster situations. In Veracruz, the rules of the Fund for the Prevention of Natural Disasters (FOPREDEN) are an instrument that aims to revitalize initiatives aimed at preventing disasters and seeks to optimize the use of available financial resources and magnify the results linked mainly to the preservation of the life and physical integrity of people, as well as that of public services and infrastructure and the environment [4].

As of 2017 FONDEN has authorized resources for more than three billion of Mexican pesos to support the road, educational sectors, forestry, hydraulic, naval, housing, and urban infrastructure in Veracruz due to the significant occurrence of natural disasters within the period 2014-2017. Table 1 presents an overview of the historical phenomena within this period and the resources provided.

Standard protocols to be performed before, during, and after a disaster involve different logistic processes. These are performed in the different phases of a disaster [6, 7]:

(1) Interdisaster phase: processes are performed in which the elaboration of the map of risks for the community is highlighted. Also the Plans of Emergency, which consist of inventory and location planning of resources, are performed.

(2) Pre-impact phase: warning to the population based on prediction mechanisms and implementation of mitigating measures are performed.

(3) Disaster impacts the community.

(4) Emergency phase: isolation, rescue, and external assistance are performed. It is often the phase in which local resources are overwhelmed and external aid is required to reduce the number of fatalities.

(5) Reconstruction phase: activities focused on recovering the normal duties of the community are performed.

Before the disaster occurs, it is important to have facilities with an optimal inventory of products of first necessity to support the survival of the people who will be affected. Also, after the disaster occurs, it is important to have the infrastructure to resupply the facilities and transport affected people to other facilities as needed.

Hence, among the most critical decisions and resources to provide relief to the affected communities in Veracruz, prepositioning of warehouses must be performed. This allows the protection of supplies and the efficient and timely supply of products to cover the basic needs of the people affected by the disturbing phenomenon. Within the activities to be performed in these warehouses or support centers, the following can be mentioned:

(1) Identification, labeling, and location of the necessary supplies to attend the emergency.

(2) Consolidation of load and change of means of transport.

(3) Delivery scheduling for the supplies.

Likewise, the warehouse must have an information and inventory control system which must be updated, through the control of inventories. The activation of a prepositioned warehouse is the responsibility of State Civil Protection with selection criteria for its allocation such as (a) being located outside the risk area, (b) having a solid and roofed construction in compliance with safety parameters, (c) being accessible through favorable conditions for transport loading and unloading, (d) being ventilated, illuminated, and without water seepage risk, (e) being located far away from flood-prone areas, (f) being free of pollution or plague, and (g) having space to facilitate the mobility, cleaning, and classification of products [8, 9]. Minimization of distance between the affected regions and the prepositioned warehouses is an important aspect of humanitarian relief planning because communities must be able to reach these centers within short periods of time and distances due to the severity of the disasters.

In this regard, humanitarian logistics (HL) formally addresses the "process of planning, implementing, and controlling the efficient, cost-effective flow of and storage of goods and materials as well as related information from point of consumption for the purpose of meeting the end beneficiary's requirements" [10, 11]. The need of HL for strategic planning has been recognized by important organizations such as the U.S. Federal Emergency Management Agency (FEMA) and the United Nations (UN) [11, 12]. In contrast to commercial logistics (CL), the main focus of HL is to save lives and provide beneficiaries with aid instead of maximizing profits. However, due to this characteristic, HL has disadvantages when compared to CL as it faces lower technology, challenging inventory control, unstable demand patterns, zero lead time, and unpredictable supply resources [13, 14].

Hence, different strategies have been developed within the field of HL for the optimal operation of all the aspects of the supply chain (SC) for the delivery of goods to affected communities considering these disadvantages. In this context, humanitarian relief organizations (HRO) have been identified as the best suited organizations for preparedness and recovery when compared to commercial and military organizations [13]. An important aspect of preparedness is the prepositioning of inventories or warehouses for postdisaster relief. Among the most recent strategies, which are focused on transportation, planning, policies and procedures, and inventory/warehousing [15], the following can be mentioned:

(i) In [16], a stochastic model was developed to determine the location of Emergency Medical Service (EMS) systems. In order to solve this model, exact and approximate (metaheuristics) methods were proposed.

(ii) The facility location problem was also addressed by [17] that presented a multiobjective optimization model to solve a multidepot emergency facilities location-routing problem. Due to the inherent computational complexity of this model an approximate method based on the metaheuristic of Genetic Algorithms (GAs) was developed.

(iii) Prepositioning or relief assets was studied in [11] to optimize the transportation of affected people to relief centers. The proposed stochastic model considered optimization of resources such as personnel and vehicles to minimize casualties.

(iv) In [18] the aspect of considering containers as storage facilities was studied, and a mathematical model was proposed to determine the locations of supply points and the quantity of containers and relief supplies assigned to each supply point under the minimum distance criteria.

(v) A stochastic inventory control strategy was proposed in [19] for the uncertain requirements of goods for postdisaster conditions, in order to have the adequate stock to serve the vital needs of affected communities.

(vi) A conceptual model that integrated the aspect of agility in HL was presented in [20] to improve on the response of HL to disaster scenarios. While no mathematical model was presented or discussed, the roles of people, processes, and technology were identified as agility enablers for the success of general models within HL.

In general for the determination and location of facilities (i.e., support centers, warehouses, prepositioned inventory, etc.) the following mathematical models have been considered: the Capacitated p-Median Problem (CPMP) [21, 22] and the Capacitated Centered Clustering Problem (CCCP) [23, 24]. Both models are focused on determining the location of p facilities in order to minimize the total weighted distance from the facilities to all demand points (customers). A demand point cannot be assigned to more than one facility, and the points assigned to a facility cannot exceed its capacity. The main difference between both models is about the features of the locations of the p facilities. In the CPMP the location is determined at a median point while for the CCCP the location is determined at a centroid.

Both models are difficult to be solved to optimality due to their NP-hard computational complexity. Particularly for large problems, this has led to the development of metaheuristics to provide near-optimal solutions [25]. In the literature, metaheuristics based on Clustering Search (CS) have been reported as the most competitive methods for the CCCP [24]. However, in this work we focus on providing an alternative to standard methods which are commonly implemented for practical situations. In the case of humanitarian relief actions, fast implementation is required, and we are considering the situation of Veracruz in Mexico, where 3837 communities with 526,954 people are at risk.

Thus, in this work a metaheuristic based on the integration of the Greedy Randomized Adaptive Search Procedure

(GRASP) and the K-Means Clustering (KMC) algorithm is presented to provide a suitable location planning for the support centers (prepositioned warehouses) for these communities in Mexico. In order to provide more accurate solutions for large CCCP instances than those of standard methods, the proposed metaheuristic has the following features:

(a) capacity restrictions to the KMC algorithm for the assignation of communities to support centers (Capacitated K-Means Clustering, CKMC);

(b) micro Genetic Algorithm (μGA) that performs single executions of the CKMC to estimate a search interval for the most suitable number of support centers;

(c) variable number of assigned communities to centers in order to add flexibility to the assignation task through iterative executions of the CKMC;

(d) conditional decision process to perform insertion, deletion, and exchange of communities between centers for further improvement of the final assignments;

(e) Earth's arc length as distance metric to locate centers within measurable distance in kilometers.

The details of this metaheuristic, termed as GRASP-CKMC, are presented as follows: in Section 2 the technical details of the GRASP-CKMC and its validation are presented. Then, in Section 3 the results on the instance of Veracruz are presented and analyzed. Finally, our conclusions are presented in Section 4.

2. GRASP-CKMC

A Greedy Randomized Adaptive Search Procedure (GRASP) is a metaheuristic which consists of two main phases: (a) the Construction Phase which consists in providing a feasible solution by combining a greedy function with a method of random selection and (b) the Local Search Phase which consists in iteratively improving the feasible solution [26, 27].

As presented in Figure 1 the proposed GRASP manages three main algorithms for these phases:

(i) Constructive Phase: a μGA is performed to determine the lower and upper limits for the most suitable number of clusters. Random selection is performed for the creation of the initial population and the number of V nearest points to extend the KMC to comply with capacity restrictions (CKCM).

(ii) Local Search Phase: the CKMC is iteratively performed with uniform random variation in V and the number of clusters K restricted by the lower and upper limits identified in the previous phase.

(iii) Random decision process to exchange locations between capacity-complying assignments for further improvement of the final CCCP solution.

In the following sections the details of the main algorithms used for the phases of the GRASP are presented and discussed.

2.1. Capacitated KMC. K-Means is one of the basic unsupervised learning algorithms that solve the well-known clustering problem [28–30]. This model is similar to the also well-known K-Nearest Neighbor (KNN) search algorithm [31]. The KMC follows a simple procedure to classify a given data set through a certain number of clusters K [28, 32]. Within the context of the CCCP or CPMP the facility is located at the cluster's centroid or median point, respectively. For multiple facilities, the first problem to be solved is the consolidation of clusters (i.e., groups of points) and the second is the determination of the median point or centroid. Both problems can be addressed simultaneously by the K-Means Clustering (KMC) algorithm. Figure 2 presents the details of the standard KMC algorithm.

As presented in Figure 2 clustering involves the unique assignment of a point to the nearest cluster based on its center (defined as the median point or the centroid). The locations of the centers must be reestimated each time that new assignments are performed, and new assignments can be generated each time that the reestimation process is performed as they affect the closeness of the centers to the considered points.

For the purpose of determining the locations of the support centers and their assigned communities, the standard KMC algorithm must integrate capacity restrictions. However this adds complexity to the assignment task because not all nearest points to a certain center can comply with its capacity restriction (thus, not all nearest points can be assigned to this center).

Approaches have been proposed to address the capacitated task. In contrast to the circular regions shown in Figure 2, in [32] a rectangular region around the center was considered to determine the candidates for clustering. This reduces the number of points to be assigned to the cluster and thus reduces the likelihood of not complying with the capacity restriction. The points located outside the rectangular region are omitted by this initial assignment process. After this process is performed, a priority is assigned to the omitted points in order to be assigned to the clusters with available capacity in a final assignment process. Other approaches involve an average distance for the reassignment of points [33].

The assignment of close points and reassignment of omitted points are procedures which can be performed with some randomness to add flexibility to the local search process of KMC. Thus, the proposal to extend the KMC to perform the capacitated task consists in including a uniform random variable to control the ratio of acceptance for the KMC algorithm (and thus, of the V nearest locations). This proposal is similar to the Variable Neighborhood Search (VNS) principle [34].

Figure 3 presents the general structure of the proposed capacity-restricted KMC algorithm (CKMC). As presented in Figure 1, this CKMC algorithm is used in both phases of the GRASP-CKMC metaheuristic. This is the reason of the adjustments stated in Figure 3:

(i) In the Constructive Phase, the CKMC is executed only once for two random K values which will be used

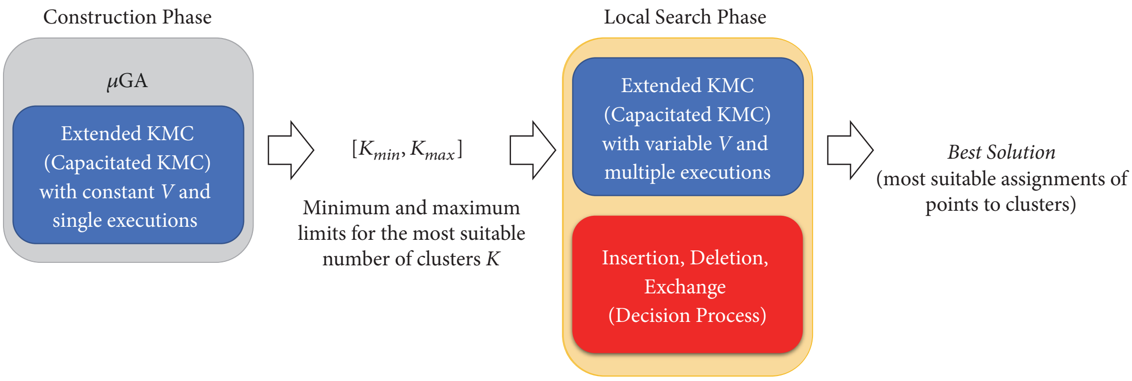

Construction Phase
μGA
Extended KMC (Capacitated KMC) with constant V and single executions
$[K_{min}, K_{max}]$
Minimum and maximum limits for the most suitable number of clusters K
Local Search Phase
Extended KMC (Capacitated KMC) with variable V and multiple executions
Insertion, Deletion, Exchange (Decision Process)
Best Solution
(most suitable assignments of points to clusters)

FIGURE 1: General structure of the GRASP-CKMC algorithm.

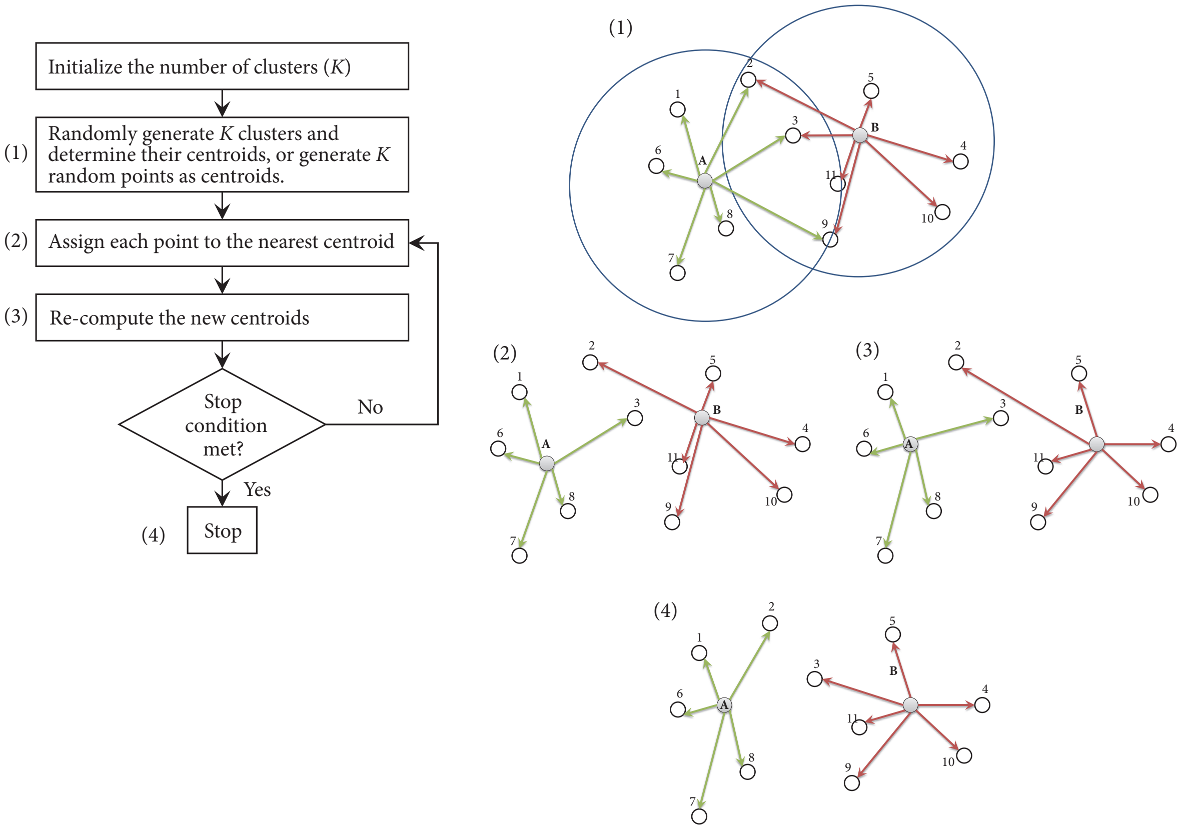

Initialize the number of clusters (K)
(1)
Randomly generate K clusters and determine their centroids, or generate K random points as centroids.
(2)
Assign each point to the nearest centroid
(3)
Re-compute the new centroids
Stop condition met?
No
Yes
(4)
Stop
(1)
(2)
(3)
(4)
A
B
1
2
3
4
5
6
7
8
9
10
11

FIGURE 2: Structure of the standard K-Means Clustering (KMC) algorithm.

by the μGA to determine the lower and upper limits (K_{min}, K_{max}) for K. Also, the number of nearest points to each center (V) is constant given by X.

(ii) In the Local Search Phase, the CKMC is iterated P times, and at each iteration, different values of K (within K_{min} and K_{max}) and the ratio of acceptance V (which has an upper limit given by X) are considered. At each iteration, the best assignment of points (locations) to clusters (centroids) (as measured by its objective function value G) is saved. After the P iterations of the CKMC algorithm are executed, the best found solution is improved by means of insertion, deletion, and exchange operations which are controlled by a decision process.

2.2. μGA. The standard KMC algorithm considers that the quantity of clusters is known *a priori* [30]. Within the context of the CCCP, the minimization of the objective function (total distance from each cluster to each assigned point) depends on finding the most suitable number of clusters. Hence, the proposed GRASP-CKMC includes an evolutionary mechanism to determine the suitable range of clusters which can minimize the total distance to the affected communities.

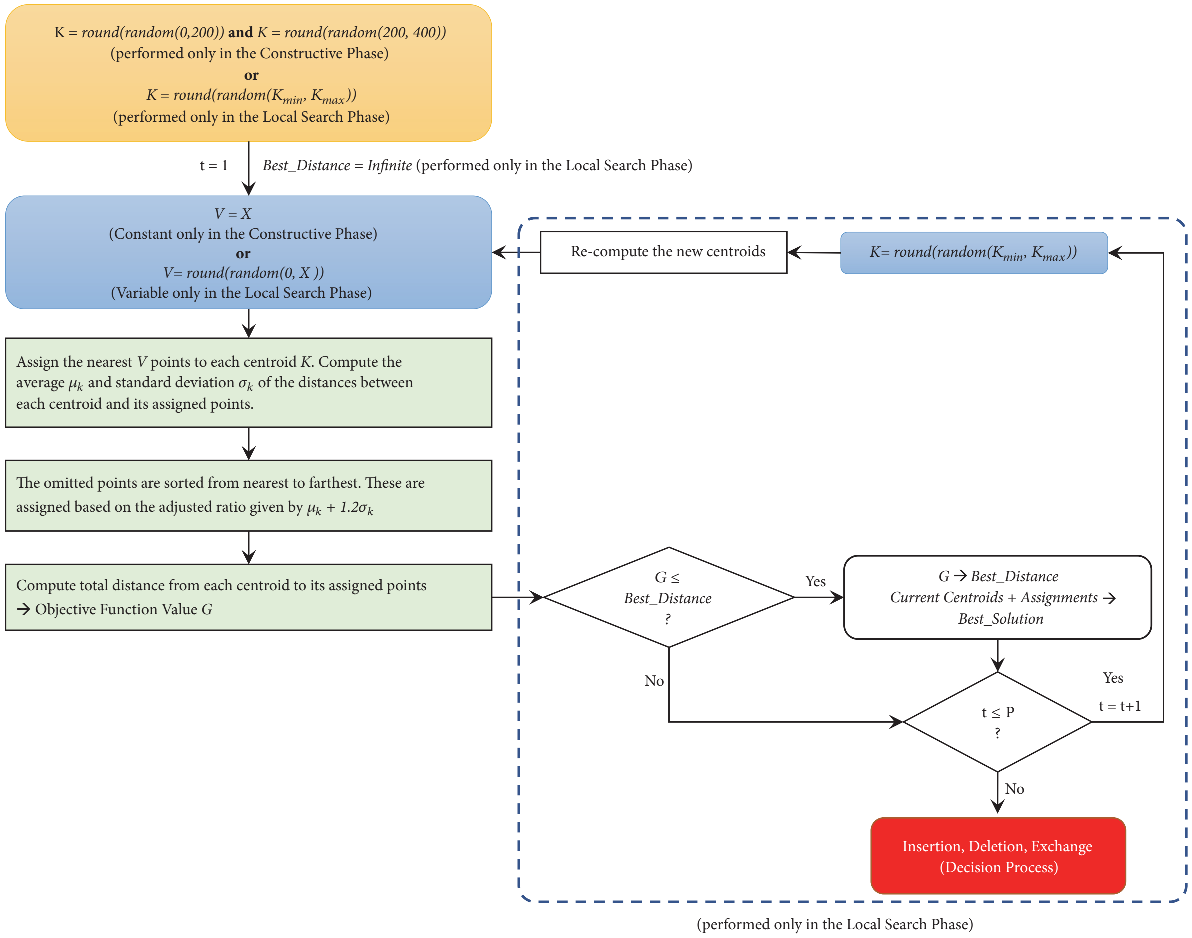

K = round(random(0,200)) and K = round(random(200, 400))
(performed only in the Constructive Phase)
or
K = round(random(K_{min}, K_{max}))
(performed only in the Local Search Phase)
t = 1
Best_Distance = Infinite (performed only in the Local Search Phase)
V = X
(Constant only in the Constructive Phase)
or
V= round(random(0, X))
(Variable only in the Local Search Phase)
Re-compute the new centroids
K= round(random(K_{min}, K_{max}))
Assign the nearest V points to each centroid K. Compute the average μ_k and standard deviation σ_k of the distances between each centroid and its assigned points.
The omitted points are sorted from nearest to farthest. These are assigned based on the adjusted ratio given by $\mu_k + 1.2\sigma_k$
Compute total distance from each centroid to its assigned points → Objective Function Value G
G ≤ Best_Distance ?
Yes
G → Best_Distance
Current Centroids + Assignments → Best_Solution
No
t ≤ P ?
Yes
t = t+1
No
Insertion, Deletion, Exchange
(Decision Process)
(performed only in the Local Search Phase)

FIGURE 3: Structure of the proposed Capacitated K-Means Clustering (CKMC) algorithm.

Figure 4 presents the general structure of the μGA which was developed to address this task. The μGA is characterized by small populations which can lead to achieving faster convergence with less storing memory [35, 36]. In this case, the individuals of the population of the μGA consist only of pairs of values (K_{min}, K_{max}) that can define the lower and upper limits of a range that may contain the K value that can lead to a total minimum distance on a single execution of the CKMC algorithm. By using the random mutation and the linear crossover operators a diversification on these bounds is obtained to estimate an interval for the local search of K within the main GRASP-CKMC algorithm. An estimate for (K_{min}, k_{max}) is obtained after m generations (in this case, m = 100) of the μGA are executed, and within this range, K is randomly selected.

2.3. Insertion, Deletion, and Exchange. As presented in Figures 1 and 3, the best solution found by the CKMC in the Local Search Phase is improved by a decision algorithm which performs insertion, deletion, and exchange of points between clusters. A conditional decision process was designed to avoid unnecessary tasks due to the random selection of points which can be inserted, deleted, or exchanged. The description of this improvement process is presented in Figure 5.

Finally, the implementation of the metaheuristic was performed with Octave and MATLAB in a HP Workstation with Intel Zeon CPU at 3.40 GHz with 8 GB RAM.

2.4. Assessment. Before proceeding to obtaining a solution for our instance, we assessed the performance of the GRASP-CKMC metaheuristic with a selection of CCCP instances. Due to the size of the instance (3837 communities), we considered the following SJC and DONI instances [37, 38]: SJC1 (100 points), SJC2 (200 points), SJC3a (300 points), SJC4a (402 points), DONI1 (1000 points), DONI2 (2000 points), DONI3 (3000 points), DONI4 (4000 points), and DONI5 (5000 points) [39]. For comparison purposes, the performance of the GRASP-CKMC metaheuristic was compared to standard and most recent methods, including the latest best known solutions as follows:

(i) Best known solutions as reported in [24].

(ii) Best results obtained by CKM and GA as reported in [38].

(iii) Best results obtained by VNS as reported in [23].

(iv) Best results obtained by TS (Tabu Search) and CS (Clustering Search) as reported in [24].

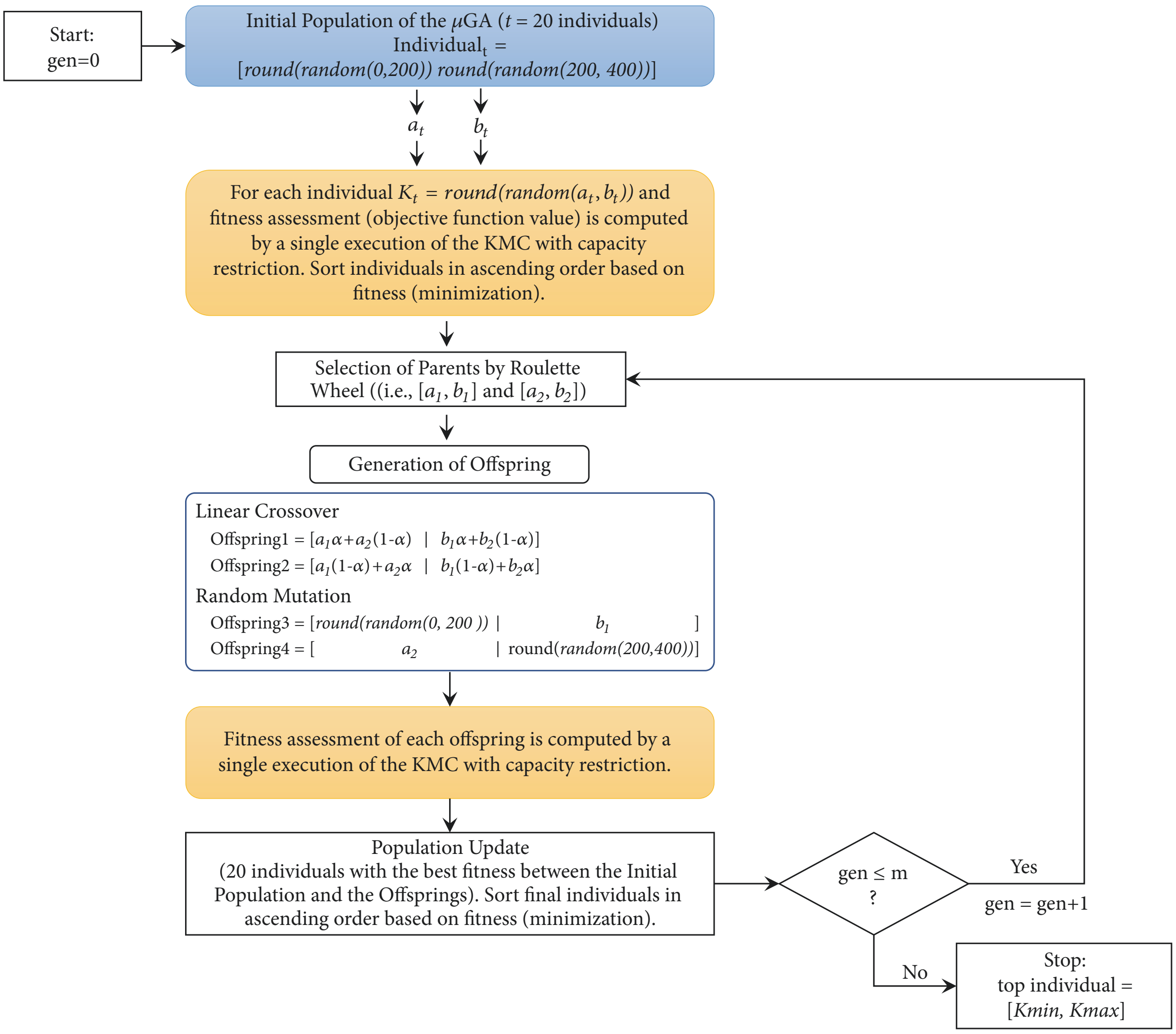

Start:
gen=0
Initial Population of the μGA (t = 20 individuals)
Individual_t =
[round(random(0,200)) round(random(200, 400))]
a_t
b_t
For each individual K_t = round(random(a_t, b_t)) and fitness assessment (objective function value) is computed by a single execution of the KMC with capacity restriction. Sort individuals in ascending order based on fitness (minimization).
Selection of Parents by Roulette Wheel ((i.e., [a_1, b_1] and [a_2, b_2])
Generation of Offspring
Linear Crossover
Offspring1 = [a_1α+a_2(1-α) | b_1α+b_2(1-α)]
Offspring2 = [a_1(1-α)+a_2α | b_1(1-α)+b_2α]
Random Mutation
Offspring3 = [round(random(0, 200)) | b_1]
Offspring4 = [a_2 | round(random(200,400))]
Fitness assessment of each offspring is computed by a single execution of the KMC with capacity restriction.
Population Update
(20 individuals with the best fitness between the Initial Population and the Offsprings). Sort final individuals in ascending order based on fitness (minimization).
gen ≤ m
?
Yes
gen = gen+1
No
Stop:
top individual =
[Kmin, Kmax]

Figure 4: Structure of the μGA for initialization of K.

(v) Best results obtained by the latest method known as Adaptive Biased Random-Key Genetic Algorithm (A-BRKGA) as reported in [24].

Table 2 presents the parameters for the GRASP-CKMC algorithm. As mentioned in [24], metaheuristics have no optimal values of parameters. Thus, recommended ranges are usually considered for these cases. For the GRASP-CKMC algorithm it was considered to have a lean execution due to the size of the instance and the diverse algorithms which were developed. Thus, small values were considered for X (the upper limit for the number of nearest points to each cluster), the executions of the CKMC in the Local Search Phase (P) and the number of pairs of points to be considered for exchange, deletion, or insertion (Y). Regarding the Mersenne Twister random number generator, it was considered as recommended by the MATLAB documentation.

Table 2: Parameters of the GRASP-CKMC.

Parameter	Value
X	10
P	50
Y	10000
Random Number Generator	Mersenne Twister

Table 3 presents the results obtained for 10 runs of the algorithm. It is observed that the average of the best results is 3.03% while the average of the worst results is 5.59%. Particularly for the instances DONI3 and DONI4, which have a similar number of points to the considered instance of 3837 communities, the GRASP-CKMC metaheuristic is able to obtain solutions with errors smaller than 5.0% (3.93% and 4.64%, respectively) within 10 runs.

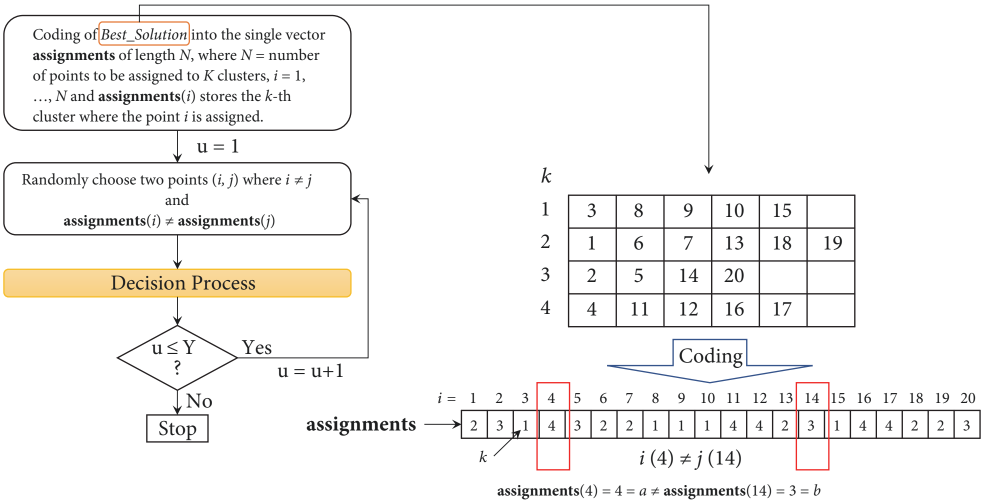
Table 4 presents the comparison of the best results obtained with the reviewed methods. When compared to CKM the proposed metaheuristic outperforms it in all instances. This is observed in the average error which is significantly higher for CKM in comparison to GRASP-CKMC (10.58% $>$ 3.03%). The average performance of the GA is similar to the performance of CKM (10.27% $\approx$ 10.00%).

TABLE 3: Results of 10 runs of the GRASP-CKMC metaheuristics on the SJC and DONI instances.

Instance	Best					Runs						Best	Worse	Average	Best	Worst	Average
	Known	1	2	3	4	5	6	7	8	9	10				Error(%)	Error(%)	Error(%)
SJC1	17359.75	17807.73	18230.49	17742.19	18088.71	17791.75	18054.23	18169.28	17948.46	18132.05	17789.13	17742.19	18230.49	17975.40	2.20%	5.02%	3.55%
SJC2	33181.65	33810.48	34077.98	33568.98	33529.65	33673.30	33342.50	33745.19	33330.23	34084.06	33731.71	33330.23	34084.06	33689.41	0.45%	2.72%	1.53%
SJC3a	45356.35	46993.46	46165.69	46789.82	46503.93	46559.25	46459.06	46944.45	46697.09	46250.59	46030.70	46030.70	46993.46	46539.40	1.49%	3.61%	2.61%
SJC4a	61931.60	62774.78	62710.35	63064.16	62962.47	62844.49	62899.00	63000.86	63357.25	62926.63	63369.95	62710.35	63369.95	62990.99	1.26%	2.32%	1.71%
DONI1	3021.41	3088.71	3165.04	3104.18	3144.54	3169.48	3141.08	3095.66	3109.24	3157.25	3158.03	3088.71	3169.48	3133.32	2.23%	4.90%	3.70%
DONI2	6080.70	6487.85	6499.84	6498.32	6490.31	6459.86	6498.46	6487.18	6484.72	6483.75	6420.06	6420.06	6499.84	6481.04	5.58%	6.89%	6.58%
DONI3	8343.49	8706.45	8971.18	8953.30	8964.51	8677.99	8691.23	8894.92	8782.56	9010.17	8671.70	8671.70	9010.17	8832.40	3.93%	7.99%	5.86%
DONI4	10777.64	11540.02	11508.46	11704.86	11568.63	11541.06	11534.28	11702.91	11277.44	11392.09	11685.20	11277.44	11704.86	11545.49	4.64%	8.60%	7.12%
DONI5	11114.67	12016.16	11991.16	12034.59	11846.86	12022.82	11724.42	11934.19	11937.12	11938.20	11893.37	11724.42	12034.59	11933.89	5.49%	8.28%	7.37%
														Average =	3.03%	5.59%	4.45%

Table 4: Performance of CKM, GA, VNS, GRASP-CKMC, A-BRKGA, TS, and CS on the SJC and DONI instances when compared to best-known solutions.

Instance	Best-Known	CKM	GA	VNS	GRASP-CKMC	A-BRKGA	TS	CS
SJC1	17359.75	17.18%	0.02%	1.94%	2.20%	0.00%	0.00%	0.00%
SJC2	33181.65	6.12%	0.83%	0.73%	0.45%	0.00%	0.00%	0.00%
SJC3a	45356.35	11.54%	3.29%	5.80%	1.49%	0.00%	0.00%	0.00%
SJC4a	61931.60	11.87%	4.92%	7.68%	1.26%	0.00%	0.10%	0.00%
DONI1	3021.41	7.06%	3.88%	0.00%	2.23%	-0.13%	0.12%	0.21%
DONI2	6080.70	10.06%	14.88%	0.00%	5.58%	4.78%	5.00%	4.81%
DONI3	8343.49	17.42%	15.70%	5.10%	3.93%	0.41%	0.00%	1.14%
DONI4	10777.64	7.58%	23.66%	6.85%	4.64%	0.12%	0.00%	1.62%
DONI5	11114.67	6.42%	25.24%	4.68%	5.49%	0.54%	0.00%	0.86%
	Average =	10.58%	10.27%	3.64%	3.03%	0.64%	0.58%	0.96%


Coding of Best_Solution into the single vector assignments of length N, where N = number of points to be assigned to K clusters, i = 1, …, N and assignments(i) stores the k-th cluster where the point i is assigned.
u = 1
Randomly choose two points (i, j) where i ≠ j and assignments(i) ≠ assignments(j)
Decision Process
u ≤ Y ?
Yes
u = u+1
No
Stop
Coding
assignments
k
i (4) ≠ j (14)
assignments(4) = 4 = a ≠ assignments(14) = 3 = b

k						
1	3	8	9	10	15	
2	1	6	7	13	18	19
3	2	5	14	20		
4	4	11	12	16	17	

i =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
assignments	2	3	1	4	3	2	2	1	1	1	4	4	2	3	1	4	4	2	2	3

- Compute the distances between the points (*i*, *j*) and their centroids (*a*, *b*): [d_{ia} d_{ib} d_{ja} d_{jb}]

Decision Process

- If $d_{ia} < d_{ib}$ and $d_{jb} < d_{ja}$ → The current assignments of *i*-*a* and *j*-*b* are suitable, no need to change.
- If $d_{ia} > d_{ib}$ and $d_{jb} < d_{ja}$ → The current assignment *j*-*b* and the new assignment *i*-*b* are more suitable to minimize distance (*insertion* of point *i* to cluster *b* = *deletion* of point *i* from cluster *a*)
- If $d_{ia} > d_{ib}$ and $d_{jb} > d_{ja}$ → The new assignments *j*-*a* and *i*-*b* are more suitable to minimize distance (*exchange* of point *j* to cluster *a*, and point *i* to cluster *b*)
- If $d_{ia} < d_{ib}$ and $d_{jb} > d_{ja}$ → The current assignment *i*-*a* and the new assignment *j*-*a* are more suitable to minimize distance (*insertion* of point *j* to cluster *a* = *deletion* of point *j* from cluster *b*)
- *Best_Solution* and *G* are updated if the changes in the assignments of (*i*, *j*) comply with the capacity restriction.

Figure 5: Structure of the decision process of the GRASP algorithm.

However this is observed because the GA outperforms the CKM method for medium instances (SJC1-DONI1) while the CKM significantly outperforms the GA for large instances (DONI2-DONI5). Better performance is observed with the VNS method with an average error of 3.64%. Also, in two instances the VNS method obtained the best known solutions (error = 0.0%). Even though the GRASP-CKMC metaheuristic is not able to obtain the best known solution, overall performance is better than VNS (3.03% < 3.64%). Particularly for instances SJC3a, SJC4a, DONI3, and DONI4, the GRASP-CKMC metaheuristic outperforms the VNS, GA, and CKM methods.

When comparing the performance of the GRASP-CKMC with more updated metaheuristics such as TS, CS, and A-BRKGA, these reported a better performance with average errors smaller than 1.0%. This is expected as the proposed metaheuristic is based on the GRASP and KMC principles and as such, it is proposed as an alternative to similar metaheuristics such as GA, KMC, and VNS. In general terms, the GRASP-CKMC performs in the middle between the

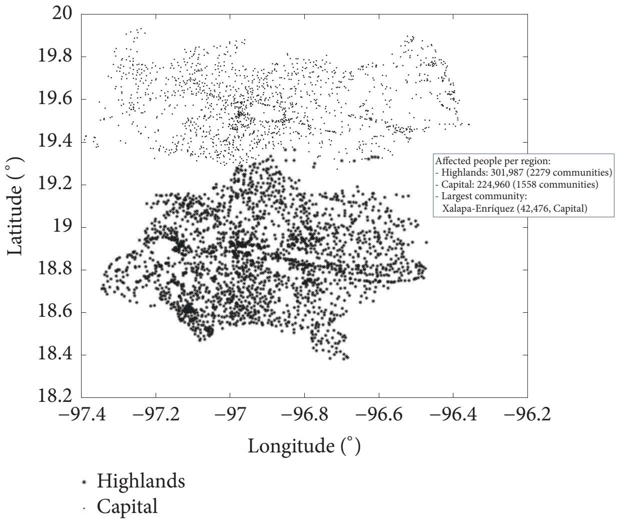

20
19.8
19.6
19.4
19.2
19
18.8
18.6
18.4
18.2
Latitude (°)
−97.4
−97.2
−97
−96.8
−96.6
−96.4
−96.2
Longitude (°)
Affected people per region:
- Highlands: 301,987 (2279 communities)
- Capital: 224,960 (1558 communities)
- Largest community:
Xalapa-Enríquez (42,476, Capital)
Highlands
Capital

FIGURE 6: Affected communities in Veracruz: capital and highlands regions.

standard and the most recent methods for the CCCP with an average best error of approximately 3.0%.

Due to these results, the proposed metaheuristic is considered suitable to address the location of the support centers or prepositioned warehouses for the communities of Veracruz.

3. Proposed Locations for Communities at Risk

Figure 6 presents general statistics regarding the people affected by disasters in the considered regions of Veracruz, Mexico. In total, in the capital and highlands regions, there are 526,947 people at risk throughout 3837 communities where the community of Xalapa-Enríquez has the largest amount with 42,476 people. Because support centers are considered to supply resources for a maximum of 10,000 people, larger communities (such as Xalapa-Enríquez) were segmented into equally-sized smaller communities. This led to a total of 3844 communities.

Due to the importance of minimizing the distance between the affected communities and the location of the centers, a reliable distance metric must be used. In this case, the geographic arc length metric is considered because it can provide accurate distances in kilometers based on the spherical model of Earth's surface which has a radius of R= 6,371 Km. With this metric, the arc length (distance) between two locations ($d_{i,j}$) with geographic coordinates (ϕ_i, θ_i) and (ϕ_j, θ_j), where ϕ is the latitude and θ is the longitude in radians, is estimated as follows [40]:

$$d_{i,j} = R \times \alpha_{i,j} = R \times \text{Arccos}\left[\cos\phi_i \cos\phi_j \cos\left(\theta_i - \theta_j\right) + \sin\phi_i \sin\phi_j\right]. \quad (1)$$

With this data, the GRASP-CKMC metaheuristic determined a set of 260 centers to provide support to the 3837 communities (or extended 3844 communities) with minimum average total distance. The general results are presented in Table 5 while Figure 7 presents the visualization of the assignments.

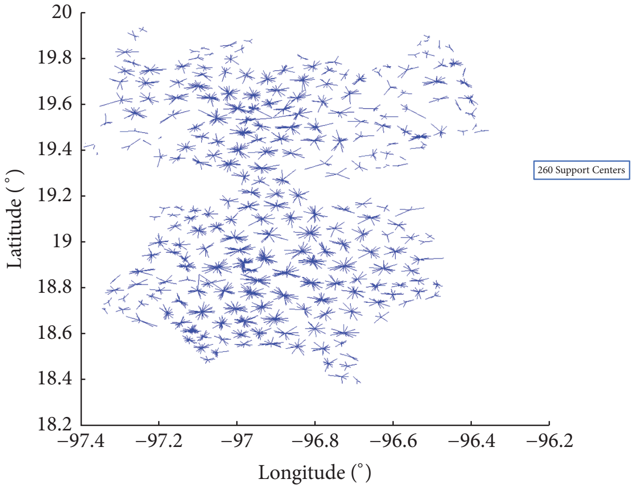


FIGURE 7: Affected communities in Veracruz: assignment of support centers.

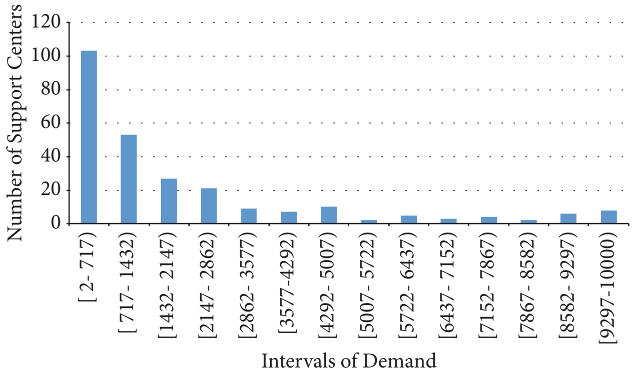


FIGURE 8: Affected communities in Veracruz: number of centers vs. intervals of demand.

Based on these results, it was determined that the mean distance that people at risk must travel from their community to its assigned center is approximately 2.08 Km with a standard deviation of 0.60 Km. In this case, humanitarian relief can be provided within a short period of time.

These results also provide information to determine the most suitable capacities for each center. Although the GRASP-CKMC imply the establishment of 260 centers to provide supplies to a maximum of 10,000 affected people, the people at risk within the communities assigned to each center can be considered to determine its most suitable capacity. Figure 8 presents a histogram that represents the number of centers assigned for each interval or range of people at risk. As observed, 103 centers serve communities with a minimum and maximum of 2 and 717 affected people, respectively. In contrast, just 8 centers serve communities with a minimum and maximum of 9297 and 10000 people, respectively. These results can be considered to make a better estimation of the capabilities of the prepositioned warehouses and, thus, of the necessary inventory.

TABLE 5: Affected communities in Veracruz: number of centers, number of locations (communities), and people at risk assigned to each center, total distance, and average distance from the locations to the assigned center.

Center	Locations	People at Risk	Total Distance	Average Distance	Center	Locations	People at Risk	Total Distance	Average Distance	Center	Locations	People at Risk	Total Distance	Average Distance	Center	Locations	People at Risk	Total Distance	Average Distance
1	12	2476	27.05	2.25	66	18	1927	44.33	2.46	131	9	1074	23.23	2.58	196	19	3227	53.64	2.82
2	5	68	8.38	1.68	67	12	656	29.48	2.46	132	11	926	20.38	1.85	197	19	320	46.62	2.45
3	10	4437	18.75	1.87	68	17	1034	39.65	2.33	133	15	713	37.88	2.53	198	12	741	28.42	2.37
4	15	602	44.27	2.95	69	13	1163	26.15	2.01	134	5	123	9.42	1.88	199	10	1924	25.27	2.53
5	3	167	5.16	1.72	70	25	9411	63.99	2.56	135	19	1995	45.68	2.40	200	21	2385	39.45	1.88
6	9	362	18.19	2.02	71	23	1269	54.62	2.37	136	17	312	50.70	2.98	201	5	460	9.81	1.96
7	23	3203	55.71	2.42	72	32	2259	78.58	2.46	137	8	175	18.37	2.30	202	4	81	4.22	1.05
8	14	213	25.19	1.80	73	41	8874	82.90	2.02	138	9	716	16.62	1.85	203	16	691	42.14	2.63
9	5	584	15.96	3.19	74	26	911	79.92	3.07	139	12	729	23.00	1.92	204	12	215	16.81	1.40
10	4	194	5.09	1.27	75	17	8374	30.90	1.82	140	13	2962	29.88	2.30	205	10	991	31.85	3.18
11	16	1410	34.09	2.13	76	33	4714	92.23	2.79	141	25	5207	55.78	2.23	206	3	198	4.01	1.34
12	2	137	1.19	0.59	77	14	2301	26.56	1.90	142	23	2576	39.97	1.74	207	3	62	5.03	1.68
13	2	88	3.03	1.52	78	7	639	13.16	1.88	143	6	213	9.57	1.59	208	11	911	20.30	1.85
14	19	4482	42.91	2.26	79	26	9941	50.56	1.94	144	4	544	5.23	1.31	209	23	3683	43.62	1.90
15	17	1561	41.25	2.43	80	3	14	3.58	1.19	145	7	185	6.71	0.96	210	15	920	32.07	2.14
16	10	267	23.04	2.30	81	5	70	8.30	1.66	146	18	2996	46.03	2.56	211	6	972	11.23	1.87
17	3	1233	2.46	0.82	82	19	1491	32.65	1.72	147	3	10	5.38	1.79	212	12	1711	21.40	1.78
18	7	584	24.53	3.50	83	16	828	38.64	2.42	148	41	5836	111.30	2.71	213	5	144	11.74	2.35
19	16	488	33.97	2.12	84	16	279	27.52	1.72	149	15	790	38.17	2.54	214	7	379	10.39	1.48
20	28	2371	71.08	2.54	85	6	263	6.15	1.02	150	4	76	3.58	0.89	215	28	150	50.60	1.81
21	6	921	7.91	1.32	86	17	434	36.45	2.14	151	3	157	1.93	0.64	216	14	585	27.98	2.00
22	4	242	5.73	1.43	87	14	1139	31.64	2.26	152	2	58	1.38	0.69	217	17	1270	31.63	1.86
23	15	1194	34.99	2.33	88	5	357	6.89	1.38	153	9	4179	19.71	2.19	218	17	1875	40.34	2.37
24	7	1105	11.96	1.71	89	23	1035	71.20	3.10	154	7	3404	19.11	2.73	219	3	585	4.06	1.35
25	17	321	29.46	1.73	90	24	1860	52.37	2.18	155	4	1235	14.53	3.63	220	12	627	26.54	2.21
26	5	326	8.13	1.63	91	9	258	15.56	1.73	156	10	756	20.69	2.07	221	6	174	11.57	1.93
27	6	779	14.21	2.37	92	21	4511	46.10	2.20	157	9	116	13.90	1.54	222	30	652	70.32	2.34
28	16	10000	52.02	3.25	93	3	804	9.60	3.20	158	8	1445	15.22	1.90	223	3	1508	1.91	0.64
29	24	2130	72.72	3.03	94	9	25	15.34	1.70	159	19	380	56.62	2.98	224	16	1250	36.62	2.29
30	29	1444	82.77	2.85	95	22	304	41.04	1.87	160	7	639	12.56	1.79	225	4	60	4.40	1.10
31	20	2256	36.76	1.84	96	16	1658	45.35	2.83	161	21	1564	41.40	1.97	226	23	9367	66.26	2.88
32	23	6639	53.35	2.32	97	4	385	7.06	1.77	162	27	9448	84.59	3.13	227	4	95	2.70	0.67
33	19	2312	42.10	2.22	98	23	9011	42.20	1.83	163	6	518	9.81	1.63	228	5	120	5.66	1.13

Table 5: Continued.

Center	Locations	People at Risk	Total Distance	Average Distance	Center	Locations	People at Risk	Total Distance	Average Distance	Center	Locations	People at Risk	Total Distance	Average Distance	Center	Locations	People at Risk	Total Distance	Average Distance
34	17	2277	25.31	1.49	99	16	2071	41.36	2.59	164	2	2	2.12	1.06	229	8	306	12.83	1.60
35	39	8353	109.04	2.80	100	5	538	6.88	1.38	165	2	33	0.23	0.12	230	6	137	10.67	1.78
36	16	9118	30.44	1.90	101	4	249	7.08	1.77	166	7	388	15.00	2.14	231	6	362	15.15	2.52
37	8	1337	21.39	2.67	102	38	3314	99.33	2.61	167	6	88	15.03	2.51	232	13	2334	26.04	2.00
38	19	945	31.18	1.64	103	25	1519	53.60	2.14	168	7	9191	11.35	1.62	233	13	1960	35.09	2.70
39	12	2176	28.03	2.34	104	7	176	21.04	3.01	169	29	1399	72.16	2.49	234	19	476	51.24	2.70
40	14	878	39.35	2.81	105	13	2766	21.44	1.65	170	9	240	18.24	2.03	235	26	3872	61.26	2.36
41	12	810	26.20	2.18	106	3	225	1.51	0.50	171	4	137	4.48	1.12	236	4	162	4.63	1.16
42	5	661	11.09	2.22	107	22	6305	51.83	2.36	172	12	561	25.68	2.14	237	19	4613	52.11	2.74
43	10	1215	19.68	1.97	108	5	39	8.05	1.61	173	22	8947	54.23	2.46	238	44	9940	63.23	1.44
44	16	322	41.62	2.60	109	11	1172	36.57	3.32	174	28	7414	52.33	1.87	239	12	69	27.39	2.28
45	8	941	13.11	1.64	110	16	1271	39.64	2.48	175	11	1894	16.74	1.52	240	19	2173	39.79	2.09
46	4	404	5.87	1.47	111	15	1041	33.29	2.22	176	12	1187	17.39	1.45	241	15	4723	44.96	3.00
47	20	2326	49.74	2.49	112	22	1034	64.23	2.92	177	28	3846	81.85	2.92	242	28	2218	76.62	2.74
48	14	588	26.33	1.88	113	16	2714	39.87	2.49	178	5	226	6.91	1.38	243	7	244	16.24	2.32
49	10	1138	21.05	2.10	114	6	2319	11.74	1.96	179	19	2849	43.91	2.31	244	43	5216	87.43	2.03
50	24	1335	54.73	2.28	115	13	1285	29.71	2.29	180	18	1298	42.73	2.37	245	12	1646	27.97	2.33
51	9	1479	15.07	1.67	116	33	4990	88.85	2.69	181	11	44	17.58	1.60	246	6	355	16.88	2.81
52	8	243	14.90	1.86	117	60	7460	144.50	2.41	182	5	904	13.07	2.61	247	40	6829	105.91	2.65
53	38	4231	94.32	2.48	118	29	1599	67.88	2.34	183	4	19	2.76	0.69	248	34	7534	86.96	2.56
54	22	2053	48.93	2.22	119	51	1966	78.85	1.55	184	25	4846	79.86	3.19	249	24	5724	54.87	2.29
55	16	925	41.17	2.57	120	14	696	35.84	2.56	185	34	7175	83.06	2.44	250	19	2167	29.69	1.56
56	5	383	4.55	0.91	121	19	2926	26.93	1.42	186	18	2028	44.00	2.44	251	23	3379	46.16	2.01
57	35	4994	71.98	2.06	122	11	597	26.20	2.38	187	12	787	19.73	1.64	252	3	579	5.51	1.84
58	14	6182	27.62	1.97	123	15	1831	39.75	2.65	188	9	448	13.68	1.52	253	6	1300	18.06	3.01
59	5	292	9.44	1.89	124	15	663	38.52	2.57	189	24	2859	72.49	3.02	254	13	884	27.78	2.14
60	22	10000	35.81	1.63	125	21	9613	48.10	2.29	190	24	6301	55.37	2.31	255	20	3998	50.98	2.55
61	9	8725	16.34	1.82	126	6	13	8.60	1.43	191	14	1603	30.47	2.18	256	17	1281	29.53	1.74
62	23	1191	57.54	2.50	127	5	47	8.31	1.66	192	43	4528	100.74	2.34	257	27	3751	63.21	2.34
63	24	7015	37.22	1.55	128	7	441	13.20	1.89	193	20	2260	47.66	2.38	258	9	68	12.89	1.43
64	8	154	25.07	3.13	129	7	389	13.08	1.87	194	11	105	28.17	2.56	259	13	1273	24.20	1.86
65	8	853	9.31	1.16	130	12	2991	29.61	2.47	195	15	780	30.59	2.04	260	9	2005	18.30	2.03

4. Conclusions and Future Work

The present work addressed the location planning for prepositioned warehouses or support centers for communities at risk in Veracruz, Mexico. This was addressed by means of the Capacitated Centered Clustering Problem (CCCP) [23] because minimization of distances between the affected regions and the prepositioned warehouses is an important aspect of humanitarian relief planning.

Due to the large set of communities (3837) and people at risk (526,947), a metaheuristic was developed to provide a suitable solution for this problem. This metaheuristic integrated the principles of GRASP, GA, and KMC to provide more suitable solutions than those obtained by similar local search metaheuristics. When tested with well-known large facility location instances, the metaheuristic termed as GRASP-CKMC was able to obtain a mean best error of 3.03%. Although more complex algorithms such as CS and A-BRKGA reported better results with errors smaller than 1.00%, the performance of the GRASP-CKCM metaheuristic was more competitive when compared to standard methods such as GA, KMC, and VNS. Thus, the GRASP-CKCM can be considered as a more suitable strategy when compared to these methods.

When the GRASP-CKMC was applied on the real instance with 3837 communities, the metaheuristic determined a set of 260 centers to provide full coverage to all communities. These results also provided insights regarding the utilization of these centers considering the actual communities assigned to them. Based on these insights, it was determined that the facility location task could also support the decisions regarding the characteristics of the support centers by obtaining the estimation of the communities assigned to each one of them. Thus, smaller centers or prepositioned warehouses can be considered for some regions. This can optimize the use of resources and improve relief efforts.

Optimization of the supply chain for humanitarian relief efforts is an extensive field which requires continuous advances in the logistics and production planning processes. Thus, as future work, the following aspects are considered:

(i) Extending the CCCP model to consider heterogeneous capacities for the centers.

(ii) Integrating route planning on the facility location problem to optimize the two-echelon supply chain.

(iii) Multicriteria optimization to extend on the facility location problem.

(iv) Integrating the use of Artificial Neural Networks (ANNs) to dynamically determine the number of clusters to improve speed and convergence of the CKMC algorithm.

(v) Integrating the principles of the CS method to enhance the performance of the GRASP metaheuristic.

References

[1] E. V. Arteaga-Vega and I. P. López-Delfín, *Gaceta Oficial: Ley Número 856 de Protección Civil y la Reducción del Riesgo de Desastres para el Estado de Veracruz*, Gobierno del Estado de Veracruz, 2013.

[2] V. E. R. Protección, *Instrumentos y Programas de Protección Civil*, Gobierno del Estado de Veracruz: Secretaría de Protección Civil, 2017.

[3] CENAPRED, *Diagnóstico de Peligros e Identificación de Riesgos de Desastres en México: Atlas Nacional de Riesgos de la República Mexicana*, Secretaría de Gobierno: Centro Nacional de Prevención de Desastres (CENAPRED), 2014.

[4] Instituto Nacional de Ecología y Cambio Climático, *Vulnerabilidad al cambio climático*, Secretaría de Gobierno: Instituto Nacional de Ecología y Cambio Climático, 2016, https://www.gob.mx/inecc/acciones-y-programas/vulnerabilidad-al-cambio-climatico-80125.

[5] Coordinación Nacional de Protección Civil, *Declaratorias de Desastre Natural: Fideicomiso No. 2003 FONDEN*, Coordinación Nacional de Protección Civil: Dirección General para la Gestión de Riesgos, 2018, http://www.proteccioncivil.gob.mx/work/models/ProteccionCivil/Resource/36/26/images/DGGR-RA2017-27MAR2018.pdf.

[6] P. I. Arcos González, R. Castro Delgado, and F. Del-Busto Prado, "Desastres y salud pública: Un abordaje desde el marco teórico de la epidemiología," *Revista Española de Salud Pública*, vol. 76, no. 2, pp. 121–132, 2002.

[7] A. Cozzolino, S. Rossi, and A. Conforti, "Agile and lean principles in the humanitarian supply chain: The case of the United Nations World Food Programme," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 2, no. 1, pp. 16–33, 2012.

[8] Sistema Nacional de Protección Civil, *Centro de Acopio*, Sistema Nacional de Protección Civil (SINAPROC), 2012, http://sismos.gob.mx/en/sismos/Centro_de_acopio.

[9] Sistema Nacional para el Desarrollo Integral de la Familia, *Manual Operativo, Atención a la Población en Riesgo o Condición de Emergencia APCE*, Sistema Nacional para el Desarrollo Integral de la Familia (SNDIF): Unidad de Atención a la Población Vulnerable, Dirección General de Alimentación y Desarrollo Comunitario, 2011, http://www.dif.gob.mx/dgadc/media/Manual%20de%20Operaci%C3%B3n%20APCE%202011.pdf.

[10] A. Thomas and M. Mizushima, "Logistics training: Necessity or luxury?" *Forced Migration Review*, vol. 22, pp. 60-61, 2005.

[11] J. Salmerón and A. Apte, "Stochastic optimization for natural disaster asset prepositioning," *Production Engineering Research and Development*, vol. 19, no. 5, pp. 561–574, 2010.

[12] A. Cozzolino, "Humanitarian logistics and supply chain management," in *Springer Briefs in Business: Humanitarian Logistics*, pp. 5–16, Springer, Berlin, Heidelberg, 2012.

[13] T. E. Fernandez and N. Suthikarnnarunai, "Control aspects in humanitarian logistics," *International Journal of Logistics Systems and Management*, vol. 28, no. 3, pp. 267–286, 2017.

[14] G. Kovács and K. Spens, "Identifying challenges in humanitarian logistics," *International Journal of Physical Distribution and Logistics Management*, vol. 39, no. 6, pp. 506–528, 2009.

[15] R. E. Overstreet, D. Hall, J. B. Hanna, and R. Kelly Rainer, "Research in humanitarian logistics," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 1, no. 2, pp. 114–131, 2011.

[16] P. Beraldi and M. E. Bruni, "A probabilistic model applied to emergency service vehicle location," *European Journal of Operational Research*, vol. 196, no. 1, pp. 323–331, 2009.

[17] B. Zhang, H. Li, S. Li, and J. Peng, "Sustainable multi-depot emergency facilities location-routing problem with uncertain information," *Applied Mathematics and Computation*, vol. 333, pp. 506–520, 2018.

[18] A. Şahin, M. Alp Ertem, and E. Emür, "Using containers as storage facilities in humanitarian logistics," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 4, no. 2, pp. 286–307, 2014.

[19] K. Ozbay and E. E. Ozguven, "Stochastic humanitarian inventory control model for disaster planning," *Transportation Research Record*, vol. 2022, no. 1, pp. 63–75, 2007.

[20] C. L'Hermitte, M. Bowles, P. Tatham, and B. Brooks, "An integrated approach to agility in humanitarian logistics," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 5, no. 2, pp. 209–233, 2015.

[21] B. Goldengorin, D. Krushinsky, and P. M. Pardalos, "The p-Median Problem," in *Cell Formation in Industrial Engineering: Theory, Algorithms and Experiments*, pp. 25–73, Springer, 2013.

[22] F. Stefanello, O. de-Araújo, and F. Müller, "Matheuristics for the capacitated p-median problem," *International Transactions in Operational Research*, 2014.

[23] M. Negreiros and A. Palhano, "The capacitated centred clustering problem," *Computers & Operations Research*, vol. 33, no. 6, pp. 1639–1663, 2006.

[24] A. A. Chaves and L. A. Nogueira Lorena, "Hybrid evolutionary algorithm for the Capacitated Centered Clustering Problem," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5013–5018, 2011.

[25] A. Martínez-Gavara, D. Landa-Silva, V. Campos, and R. Martí, "Randomized heuristics for the Capacitated Clustering Problem," *Information Sciences*, vol. 417, pp. 154–168, 2017.

[26] R. Interian and C. C. Ribeiro, "A GRASP heuristic using path-relinking and restarts for the Steiner traveling salesman problem," *International Transactions in Operational Research*, vol. 24, no. 6, pp. 1307–1323, 2017.

[27] R. Dias, B. Guazzelli, and P. Henrique, "A GRASP heuristic in the choice of clusterheads for wireless sensor networks provided as a service," in *Proceedings of the International Conference on High Performance Computing & Simulation*, pp. 160–167, IEEE, 2017.

[28] M. C. Hung, J. Wu, J. H. Chang, and D. L. Yang, "An efficient k-means clustering algorithm using simple partitioning," *Journal of Information Science and Engineering*, vol. 21, pp. 1157–1177, 2005.

[29] K. Liao and D. Guo, "A Clustering-based approach to the capacitated facility location problem," *Transactions in GIS*, vol. 12, no. 3, pp. 323–339, 2008.

[30] M. Javadi and J. Shahrabi, "New spatial clustering-based models for optimal urban facility location considering geographical obstacles," *Journal of Industrial Engineering International*, vol. 10, no. 54, pp. 10–12, 2014.

[31] C.-M. Pintea and P. C. Pop, "An improved hybrid algorithm for capacitated fixed-charge transportation problem," *Logic Journal of the IGPL. Interest Group in Pure and Applied Logics*, vol. 23, no. 3, pp. 369–378, 2015.

[32] R. Sahraeian and P. Kaveh, "Solving capacitated p-median problem by hybrid k-means clustering and fixed neighborhood search algorithm," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, pp. 1–6, 2010.

[33] G. R. Sutanto, S. Kim, D. Kim, and H. Sutanto, "A heuristic approach to handle capacitated facility location problem evaluated using clustering internal evaluation," *IOP Conference Series: Materials Science and Engineering*, vol. 332, no. 332, Article ID 012023, pp. 1–9, 2018.

[34] J. Brimberg, P. Hansen, N. Mladenoviċ, and E. Taillard, "Improvements and comparison of heuristics for solving the multisource weber problem," Technical Report IDSIA-33-97, 1997.

[35] R. Batres, "Generation of operating procedures for a mixing tank with a micro genetic algorithm," *Computers & Chemical Engineering*, vol. 57, pp. 112–121, 2013.

[36] P. C. Ribas, L. Yamamoto, H. L. Polli, L. Arruda, and F. Neves-Jr, "A micro-genetic algorithm for multi-objective scheduling of a real world pipeline network," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 1, pp. 302–313, 2013.

[37] A. Palhano, M. Negreiros, and G. Laporte, "A constrained k-median procedure for the capacitated centered clustering problem," in *Proceedings of the XIV Congreso Latino Ibero Americano de Investigacion de Operaciones (CLAIO '08)*, 2008.

[38] A. Chaves, J. Goncalves, and L. Nogueira, "Adaptive biased random-key genetic algorithm with local search for the capacitated centered clustering problem," *Computers & Industrial Engineering*, vol. 124, pp. 331–346, 2018.

[39] L. Nogueira, *Problem Instances*, INPE-Instituto Nacional de Pesquisas Espaciais, http://www.lac.inpe.br/lorena/instancias.html, last accessed on July 9th 2018.

[40] L. Cazabal-Valencia, S.-O. Caballero-Morales, and J.-L. Martínez-Flores, "Logistic model for the facility location problem on ellipsoids," *International Journal of Engineering Business Management*, vol. 8, pp. 1–9, 2016.

A *K*-Means and Ant Colony Optimization-Based Routing in Underwater Sensor Networks

Subrat Dash, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, subratdash43@gmail.com*

Srimanta Mohapatra, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, srimantamohaparta66@gmail.com*

Major Das, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, major.das556@gmail.com*

Pravat Kumar Rautray, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, pravatrautray216@gmail.com*

Abstract

Reliable data transfer seems a quite challenging task in underwater sensor networks (UWSNs) in comparison with terrestrial wireless sensor networks due to the peculiar attributes of UWSN communication. Therefore, *K*-means and ant colony optimization-based routing (KACO) is proposed in this paper. In KACO, network area under water is divided into layers with regard to the depth level. And nodes of each layer are divided into clusters by the optimized *K*-means algorithm. The *K*-means algorithm is used to cluster nodes. Considering the shortcoming of *K*-means clustering, an improved *K*-means clustering is used to select the initial cluster center. In the stage of selecting cluster heads, the remaining energy of nodes and the distance from the sink node are used to calculate the competing factors of nodes, and then, the cluster heads are selected according to the competing factors. In the intercluster routing, the ant colony optimization (ACO) was improved by introducing the Gini coefficient, and the intercluster routing based on improved ACO is proposed. The simulation results show that the proposed KACO routing can effectively reduce the energy consumption of nodes and improve the efficiency of packet transmission.

1. Introduction

With the development of modern science and technology, human beings are more fully aware of the use and development of the ocean, due to its rich resource reserves and research value, which has prompted us to continuously explore the underwater space [1]. A group of interconnected sensor nodes through acoustic channel form a underwater sensor networks (UWSNs) [2]. UWSNs have been capturing attention from the scientific and industrial communities [3]. The use of underwater sensor nodes, capable with wireless communication capabilities, has the power to detect real-time underwater monitoring.

In UWSNs, sensor nodes are deployed in rivers or seas to detect the characteristic present in the water environment, such as temperature, pressure, and water quality [4]. Then, they forward their data to surface sink [5]. Received data of all nodes, the sink node will preprocess these data and forward toward offshore data center.

Figure 1 shows the typical structure of UWSNs, which consists of underwater sensor nodes, sink nodes, and offshore data center. These could be a sink node, or there could be multiple sink nodes, depending on application.

Considering the characteristics of underwater communication environment [6], underwater sensor nodes are provided with acoustic modems to communicate with each other wirelessly and transmit the data toward sink node. While each sink node has both acoustic and radio modems. The acoustic modem is used to receive data from the underwater sensor nodes. And the radio modem is used to transmit data to the offshore data center [7].

Efficient data collection is tedious in UWSNs. However, routing protocols implemented in terrestrial wireless sensor network (WSN) cannot be directly executed in UWSNs since that transmission medium of underwater environment is different with that of terrestrial WSN. In addition, acoustic communication itself has significant limitations, for example, high propagation delay, slow data rates, and

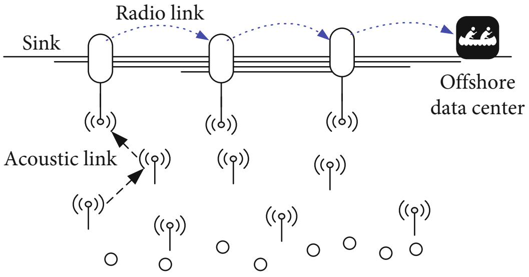


Figure 1: Typical structure of UWSNs.

absorption losses. Besides, sensor nodes are battery-powered. And it is hard to replace their batteries. Once energy of node is exhausted, it will immediately fail, which will cause network coverage vulnerability and greatly affect performance of network.

Therefore, these challenges can motivate the researches to design reliable and energy-efficient UWSN routing protocols. Clustering protocols have been widely used in terrestrial WSN. The basic idea of clustering protocol is to divide sensor nodes into some clusters. Each node in a cluster has chance to act as a cluster head. Cluster head (CH) takes charge of collecting data from other nodes in its cluster. The introduction of CHs can avoid long distance communication, and much energy is saved.

Therefore, *K*-means and ant colony optimization-based routing (KACO) is proposed in this paper. In KACO, the *K*-means algorithm is used to form clusters. Considering that the traditional *K*-means algorithm generates the initial cluster center in random way, node degree-based initial cluster center selection is introduced. In each cluster, the nodes ran for the cluster head in a distributed way based on residual energy of nodes and distance from sink node.

In addition, the improved ACO is used to construct the intercluster routing, and energy Gini coefficient is applied to construct the transition probability in order to balance energy consumption of clusters.

The rest of this paper is organized as follows: Section 2 presents related works. Section 3 describes the background. The proposed routing is described in Section 4. Section 5 deals with the simulation setup and discussion of the proposed routing. Finally, we conclude the paper in Section 6.

2. Related Works

Reference [8] has discussed the application of clustering routing in UWSN and has confirmed that clustering routing is also suitable for underwater environment. For example, reference [9] has proposed layers and unequal clusters-based energy efficient clustering routing (LUER). In LUER, routing decisions are made based on both link quality and residual energy of nodes. Unfortunately, location of nodes has not taken into account in routing process.

Except for clustering routing, researchers also have proposed other routing protocol. For example, reference [10] has proposed depth-based routing (DBR) protocol. In DBR, routing decisions are made based on depth of nodes. However, it has not taken full advantage of clustering technique, such as data transmission efficiency and reducing energy consumption of nodes.

Reference [11] proposed an energy-efficient multilevel adaptive clustering routing algorithm (ACUN). The algorithm adopts multilevel hierarchical network structure to determine the size of the competition radius. Node with larger residual energy is selected as CH. It can avoid early death of CHs. However, ACUN algorithm has not discussed intercluster routing. In fact, the energy consumed by CH in the routing phase is negligible. In addition, reference [12] proposed sparsity-aware and energy-efficient clustering algorithm. The power control mechanism is used to improve energy efficiency. The algorithm balances the energy consumed in view of power control mechanism rather than routing mechanism.

Reference [13] proposed an efficient metaheuristic-based clustering with routing protocol. The goal of the protocol is to elect an efficient set of CHs and route to destination. The protocol involves the designing of cultural emperor penguin optimize-based clustering techniques to form clusters. However, cultural emperor penguin optimize-based clustering techniques are too complex. In addition, the uniformity of cluster distribution is not guaranteed.

3. Background

3.1. Network Model. N underwater sensor nodes (nodes) are deployed randomly in-three dimensional network field $\ell \times \ell \times \ell$. These nodes form a set $\mathcal{S} = \{s_1, s_2, s_3, \cdots, s_N\}$. They are equipped with acoustic modems in order to communicate with other nodes in underwater environment.

Without loss of generality, we assumed that our network scenario was similar to the networks in [14–17]. The assumptions of the network can be described as follows: (1) each node knows its location upon first deployment [17]; (2) each node acquires its current level of depth with the help of pressure sensor; and (3) All nodes have the same initial limited energy, except the sink node, which has an unlimited power supply.

A sink node at the surface has acoustic and radio modems. The sink node firstly collects data from nodes in underwater environment through acoustic link, and transmits the collected data to offshore data center through radio link, as shown in Figure 2.

In order to improve the efficiency of transmitting data, the depth of deployment has been divided into layers. The number of layer is given by $M = \lceil \ell/(2r) \rceil$, where ℓ is the depth of network deployment and r is the transmission range of the node and $\lceil \cdot \rceil$ represents ceil function.

In initialization phase, M is loaded into hello message, and sink node broadcasts the hello message in the whole network. Once received hello message, the node extracts M from it. Therefore, each node can compute its layer

$$m_i = \left\lceil h_i \frac{M}{2r} \right\rceil, \qquad (1)$$

where h_i is the depth of node s_i and m_i represents the number of layer of node s_i.

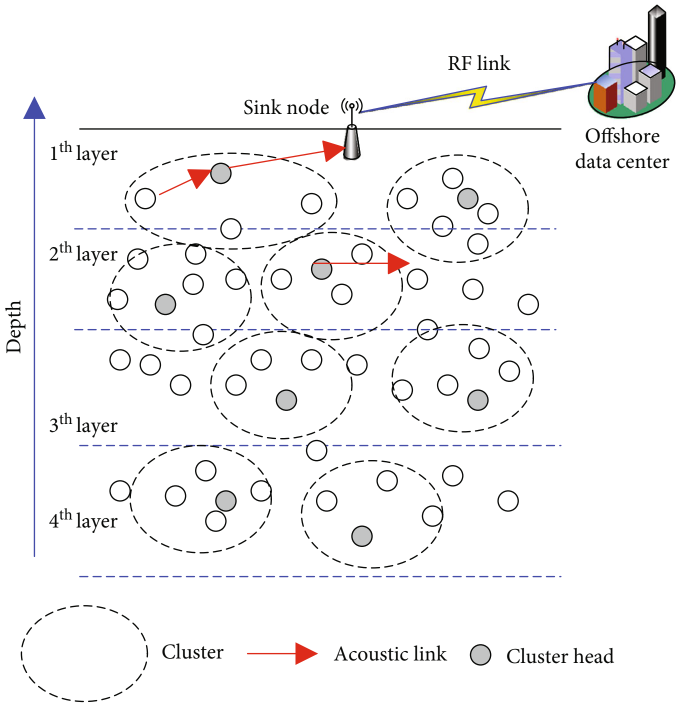


FIGURE 2: Network model.

Sink node locates in the first layer since that it is at the surface. Nodes in each layer are divided into clusters, and each cluster consists of a CH and cluster members, as shown in Figure 2.

3.2. Energy Consumption Model. Energy consumption model implemented in terrestrial WSN cannot be directly executed in UWSN since that characteristic of acoustic wave in underwater environment is different from that of radio wave. The energy consumption model was similar to the models in [18, 19].

Energy consumption of node is denoted as $E_{Tx}(k,d)$ when it transmits k bits of data over distance d.

$$E_{Tx}(k,d) = k \cdot E_{\text{elec}} + \frac{k}{R} \cdot P_{tx}, \tag{2}$$

where "$\cdot$"represents multiplication operation, E_{elec} is the energy consumption to transmit a bit data, and R is data rate. P_{tx} is the power of transmitting data, which is defined as

$$P_{tx} = P_0 \cdot d^2 \cdot 10^{(\alpha_{TL}(f)/10)}, \tag{3}$$

where P_0 is the minimum power required at the receiver and d is the distance between transmitter and receiver. $\alpha_{TL}(f)$ is the absorption coefficient, which is defined as

$$\alpha_{\text{TL}}(f) = 0.11 \frac{f^2}{1+f^2} + 44 \frac{f^2}{4100+f^2} + 2.75 \cdot 10^{-4} f^2 + 0.003, \tag{4}$$

where f is frequency.

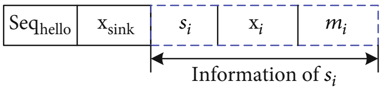


FIGURE 3: Structures of hello message.

Energy consumption of node is denoted as $E_{Rx}(k)$ when it has received k bits of data, which is defined as follows:

$$E_{Rx}(k) = k \cdot P_r, \tag{5}$$

where P_r is energy consumption parameter of received device, which is a constant dependent on the device.

Energy consumption of a CH or sink node is denoted as $E_{\text{DA}}(k)$ when it has fused k bits of data, which is defined as

$$E_{\text{DA}}(k) = k \cdot E_{\text{da}}, \tag{6}$$

where E_{da} is the energy consumed by fusing a single bit of data.

4. KACO Routing

KACO routing is mainly composed of four stages: initialization phase, cluster formation phase, CH selection phase, and intercluster routing establishment phase.

4.1. Initialization Phase. Initially, sink node broadcasts a hello message including serial number of message and location of sink node, as shown in Figure 3. $\mathbf{x}_{\text{sink}} = (x_{\text{sink}}, y_{\text{sink}}, z_{\text{sink}})$ is position vector of sink node.

Received hello message from the sink node, the sensor node s_i checks whether it is the first time to received it. If

```
Input: $N_k, k = 1, 2, \cdots, M$ ; $\mathscr{L}_k = \{s_1^k, s_2^k, \cdots, s_{|\mathscr{L}_k|}^k\}$, $C \longleftarrow \phi$
Output: Centroid position of all clusters: $C$
Step1: $N_k$ random centroids are selected from $\mathscr{L}_k$ and these centroids are located into initial centroid set $C$
Step2: centroids are located into initial centroid set
        Other sensor nodes compute the distance from these centroids, the centroid with minimum distance is selected to cluster
Step3: Centroid position of each cluster is calculated by Equation (8)
Step4: $E$ is calculated by Equation (10), and decide $E < \varepsilon$
        If not, go to Step 2.
        Otherwise, iteration is stopped, and output $C$
```

Algorithm 1

it is, the sensor node will extract the position vector of sink node from it and store the position vector. Subsequently, the position vector, ID, and layer level of the node are loaded into the hello message. Otherwise, the hello message is discarded. Figure 3 summarizes the details of the hello message structures used.

Afterward, node s_i transmits the hello message toward its neighbor nodes. Node s_i will discard the hello message if it is not received the message for the first time.

Received hello message forwarded by other sensor nodes, the sensor node s_i, extracts the relevant information of the sender from it ,and constructs its on-hop neighbor nodes set, including position vector, ID, and layer level of node. The same layer neighbor node set of node is denoted as $\mathscr{N}_i$. In other words, the layer of any node $s_j \in \mathscr{N}_i$ is the same to that of nodes$_i$.

4.2. Clustering Process. The clustering process is the method used to divide nodes in the same layer into groups such that energy consumption is more balanced.

Number of clusters is vital key in clustering process, which should be assigned prior to the clustering process. Therefore, the number of nodes in each layer is used to compute the accurate optimal number of clusters. Let $\mathscr{L}_k$ represents the node set in the k layer, where $k = 1, 2, \cdots, M$. The optimal number of clusters is given by N_k:

$$N_k = \sqrt{|\mathscr{L}_k|\ell/(2\pi r)}, \tag{7}$$

where $|\mathscr{L}_k|$ is the number of sensor nodes in $\mathscr{L}_k$.

4.2.1. Node Degree-Based K-Means Clustering Algorithm. The basic principle of K-means algorithm [20] is that the objects are divided into groups. It is implemented in iterative way, namely, iteration will stop until the square error criterion value is minimum or the maximum number of iterations is reached.

Therefore, the nodes in each layer are divided into clusters with the help of K-means algorithm as a result of its simplicity and efficiency. The specific implementation process is shown in Algorithm 1.

Input of Algorithm 1 is followed as (1) node set of k^{th} layer, denoted as $\mathscr{L}_k = \{s_1^k, s_2^k, \cdots, s_{|\mathscr{L}_k|}^k\}$, and (2) optimal number of clusters in k^{th} layer, denoted as N_k. Initially, N_k random centroids are selected from $\mathscr{L}_k$, and these centroids are located into initial centroid set, denoted $\mathscr{C} = \{C_1, C_2, \cdots, C_{N_k}\}$.

Other nodes compute the distance from these centroids, and the centroid with minimum distance is selected to be a cluster center and then form a cluster set $\mathscr{W} = \{\Omega_1, \Omega_2, \cdots, \Omega_{N_k}\}$.

Afterward, the centroid position of each cluster is calculated, which is defined as

$$C_i = \frac{1}{|\Omega_i|} \sum_{s_j \in \Omega_i} X_j^i, \tag{8}$$

where X_j^i is the position of node s_j in Ω_i and $|\Omega_i|$ is the number of nodes in Ω_i.

Square error criterion function is defined as

$$E = \sum_{i=1}^{N_k} \sum_{j=1}^{|\Omega_i|} \left| X_j^i - C_i \right|^2. \tag{9}$$

Iteration will stop when the square error criterion value is minimum or the maximum number of iterations is reached. If the above conditions are not met, Step 2 is done, as shown in Algorithm 1. ε is the threshold value.

Although K-means algorithm is simple and efficient, it still has a problem: initial centroid set is selected in a random way. These selected points may be isolated point or remote points, or distance among the selected points is too close, so that the distribution deviation of clusters is too large. The clustering result is wrong, and even the convergence time of the algorithm in the later period of operation is longer.

Therefore, Algorithm 1 is improved, namely, in Step 1 in Algorithm 1, the initial centroid set is selected according to node degree rather than random way. The specific process is as follows.

First, each node in k layer computes the number of neighbor nodes, which is called node degree. Then, the set $\mathscr{L}_k$ is sorted on node degree in descending order. Let $\tilde{\mathscr{L}}_k$ represents the sorted $\mathscr{L}_k$ on node degree in descending order. The node with the largest node degree was selected as the first point in initial centroid set. In other words, the first node in $\tilde{\mathscr{L}}_k$ is considered to be the first point in initial centroid set.

```
Input: $N_k, k = 1, 2, \cdots, M$ ; $\mathscr{L}_k = \{s_1^k, s_2^k, \cdots, s_{|\mathscr{L}_k|}^k\}, C \longleftarrow \phi$
Output: $C$
Step1: The set $\mathscr{L}_k$ is sorted on node degree order, denoted $\tilde{\mathscr{L}}_k$
Step2: The first node $\tilde{s}_i^k$ in $\tilde{\mathscr{L}}_k$ is added into $C$, namely, $C \longleftarrow C \cup \{\tilde{s}_i^k\}$
Step3: Update $\tilde{\mathscr{L}}_k$ set, namely, $\tilde{\mathscr{L}}_k \longleftarrow \tilde{\mathscr{L}}_k/\tilde{\mathscr{N}}_i^k$
Step4: Decide $|C| = N_k$? If Yes, interation is stopped, and output $C$. If not, go to Step 2.
```

Algorithm 2

For simplicity, let $\tilde{s}_i^k$ represents first node in $\tilde{\mathscr{L}}_k$. The one-hop and same layer neighbor node of node $\tilde{s}_i^k$ are removed from $\tilde{\mathscr{L}}_k$, namely, $\tilde{\mathscr{L}}_k \longleftarrow \tilde{\mathscr{L}}_k/\tilde{\mathscr{N}}_i^k$, where $\tilde{\mathscr{N}}_i^k$ is the one-hop and same layer neighbor node set of $\tilde{s}_i^k$.

Then, the first node in $\tilde{\mathscr{L}}_k$ is added into initial centroid set $\mathscr{C}$. Repeat the above process until there are N_k nodes in set $\mathscr{C}$. Algorithm 2 shows the process of constructing initial centroid set $\mathscr{C}$.

4.2.2. Cluster Head Selection Process. Next, each node in cluster calculates its competition factor, which is defined as

$$W_i = \omega_1 \frac{\ell - d_i}{\ell} + \omega_2 \frac{E_i^{\text{res}}}{E_{\text{init}}}, \tag{10}$$

where W_i is the competition factor, d_i is the distance between node s_i and sink node, E_i^{res} is residual energy of node s_i, E_{init} is the initial energy of each node in network, and ω_1 and ω_2 denote the coefficient such that $\omega_1 + \omega_2 = 1$.

From the definition, the farther the distance from sink node is, the greater $(\ell - d_i)/\ell$ of is. The more residual energy is, the greater $E_i^{\text{res}}/E_{\text{init}}$ of s_i is. Therefore, the nodes with large competition factor are preferentially selected as CH.

In KACO, the CHs are selected using a back-off timer [21]. The timer value is inversely proportional to competition factor of node. For instance, the back-off timer value will be low if the competition factor is high, and vice versa.

Therefore, the timer value of node s_i is set, which is defined as

$$\text{Time}(i) = \left\lfloor \frac{T}{W_i} \right\rfloor, \tag{11}$$

where T is the shortest wait time. $\text{Time}(i) > T$ due W_i is less than 1.

Obviously, back-off timer value will be expired soon for the nodes with competition factor. Once the back-off timer reaches zero, the node broadcast ADV_CH message within its cluster and declares itself as cluster head. When any of the nodes in cluster have received ADV_CH message before its timer expires, it gives up competing for cluster head in this round.

The entire process is depicted as shown in Figure 4. Firstly, the competition factor is calculated using Equation (10). Then, the back-off timer value is set using Equation (11). Each node listens to decide if the ADV_CH message is transmitted by a neighbor node. The node will give up competing for CH in this round if the ADV_CH message has been transmitted. Otherwise, the node waits and sends an ADV_CH message when it its timer expires.

Once forming cluster, the CH schedules the transmission of all cluster members as a time division multiple access (TDMA).

4.3. Multihop Routing. When a CH has collected data from all nodes in its cluster, it needs to transmit these data toward sink node. The CH directly transmits these data if the distance between the CH and sink node is less than r. Otherwise, the CH transmits these data in multihop routing. In this case, the cluster head needs to select a relay node (next-hop forwarding node).

In the intercluster routing, the ant colony optimization (ACO) was improved by introducing the Gini coefficient [22], and the intercluster routing based on improved ACO is proposed.

4.3.1. Gini Coefficient-Based State Transition Optimization. To balance energy consumption among CHs is the key to intercluster routing. Gini coefficient is a statistical index in economics to measure the balance degree of income distribution in a region. It can effectively and abstractly represent the difference of income distribution among individuals. Therefore, Gini coefficient is introduced into intercluster routing. And energy Gini coefficient is used to estimate energy equilibrium degree of clusters.

Firstly, energy Gini coefficient is defined as

$$G_{(\text{CH}_i,\text{CH}_j)} = \sum_{\text{CH}_j \in \mathscr{H}_i} \frac{\sum_{k=1}^{N(\text{CH}_i)} \sum_{h=1}^{N(\text{CH}_j)} \left| E_k^{\text{res}}(\text{CH}_i) - E_h^{\text{res}}(\text{CH}_j) \right|}{2\left[N(\text{CH}_j)\right]^2 E_{\text{CH}_j}^{\text{ave}}}, \tag{12}$$

where CH_i is the i^{th} CH. Assume that a relay node is selected by CH_i. The relay node is charge of forwarding data of CH_i toward sink node.

$\mathscr{H}_i$ is the neighbor CHs set, namely, relay nodes set. $N(\text{CH}_j)$ and $N(\text{CH}_i)$ are the number of nodes in CH_i and CH_j, respectively. $E_k^{\text{res}}(\text{CH}_i)$ is the residual energy of s_k in CH_i. $E_h^{\text{res}}(\text{CH}_j)$ is the residual energy of s_h in CH_j. $E_{\text{CH}_j}^{\text{ave}}$ is the average energy of node in CH_j.

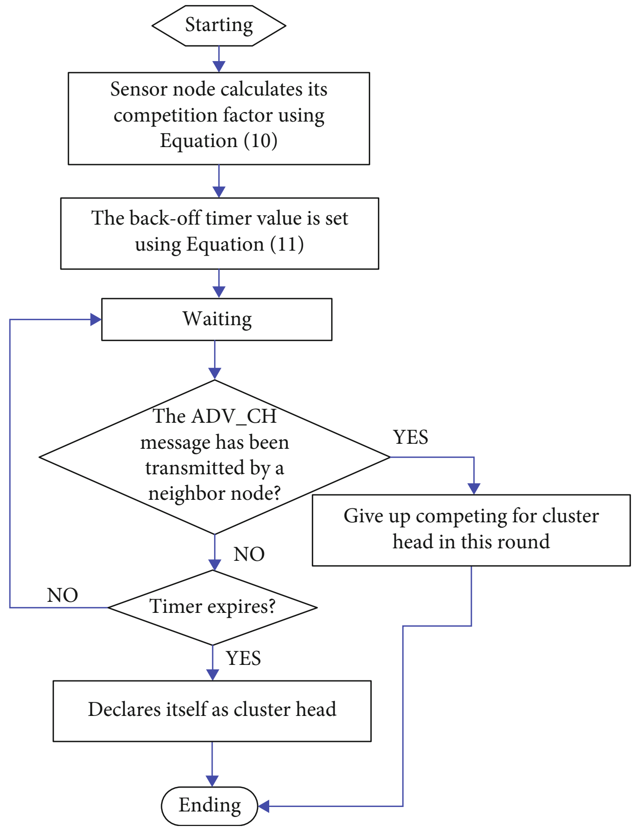

Starting
Sensor node calculates its competition factor using Equation (10)
The back-off timer value is set using Equation (11)
Waiting
The ADV_CH message has been transmitted by a neighbor node?
YES
Give up competing for cluster head in this round
NO
NO
Timer expires?
YES
Declares itself as cluster head
Ending

Figure 4: Flowchart of CH selection process.

At time t, forward ant $\mathscr{X}$ calculates the transition probability from CH_i to CH_j, which is defined as

$$P(\mathrm{CH}_i, \mathrm{CH}_j) = \frac{\left[\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t)\right]^{\alpha}\left[\eta_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t)\right]^{\beta}\left[G_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t)\right]^{\lambda}}{\sum_{\mathrm{CH}_k\in\mathscr{H}_i}\left\{\left[\tau_{(\mathrm{CH}_i,\mathrm{CH}_k)}(t)\right]^{\alpha}\left[\eta_{(\mathrm{CH}_i,\mathrm{CH}_k)}(t)\right]^{\beta}\left[G_{(\mathrm{CH}_i,\mathrm{CH}_k)}(t)\right]^{\lambda}\right\}}, \tag{13}$$

where $\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t)$ is the amount of pheromone trail on path from CH_i to CH_j at time t and α and β are parameters that determine the relative influence of the pheromone trails. $\eta_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t)$ represents visibility value, which expression is given by

$$\eta_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t) = \frac{e^{E^{\mathrm{res}}_{\mathrm{CH}j}}}{d_{\mathrm{CH}_i,\mathrm{CH}_j} + d_{\mathrm{CH}_j,\mathrm{sink}}}, \tag{14}$$

where $E^{\mathrm{res}}_{\mathrm{CH}_j}$ is the residual energy of CH_j, $d_{\mathrm{CH}_i,\mathrm{CH}_j}$ is the distance between CH_i and CH_j, and $d_{\mathrm{CH}_j,\mathrm{sink}}$ is the distance between CH_j and sink node. As known from Equation (14), both residual energy of candidate CHs and position of candidate CHs are used in the definition. Balance energy consumption of nodes in the selected CHs is ensured by introducing energy Gini coefficient in state transition function.

4.3.2. Updating Pheromone. In order to optimize the efficiency of ACO algorithm, the updating pheromone process is divided into local pheromone updating process and global pheromone udpating process [23].

The rule of updating pheromone is defined as follows:

$$\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t+1) = (1-\rho)\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t) + \rho\Delta\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t), \tag{15}$$

where ρ is the evaporation rate of pheromone and $\Delta\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t)$ is the quantity of pheromone laid on path from CH_i toCH_j.

When the forward ant $\mathscr{X}$ has moved from CH_i to CH_j, the local pheromone updating rule is applied:

$$\Delta\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t) = \frac{Q_L E^{\mathrm{res}}_{\mathrm{CH}_j}}{N(\mathrm{CH}_i) + d_{\mathrm{CH}_i,\mathrm{CH}_j}}, \tag{16}$$

where Q_L is the local pheromone concentration.

It will automatically disappear, and the corresponding backward ant $\widehat{\mathscr{X}}$ is generated when the forward ant $\mathscr{X}$ has reached the sink node. The backward ant $\widehat{\mathscr{X}}$ returns to source along reverse path, and global pheromone updating rule is applied:

$$\Delta\tau_{(\mathrm{CH}_i,\mathrm{CH}_j)}(t) = \frac{Q_G E^{\min}_{\mathrm{CH}}}{L_{\widehat{\mathscr{X}}}}, \tag{17}$$

where Q_G the global pheromone concentration, $L_{\widehat{\mathscr{X}}}$ is the length of the ant's path, and $E^{\min}_{\mathrm{CH}}$ is the minimum energy of all cluster heads in the path.

The flowchart of discovering intercluster routing is shown in Figure 5. Ant on the source selects next-hop CH according to transition probability shown in Equation (13) and updates local pheromone. When the ant has reached the sink node, $E^{\min}_{\mathrm{CH}}$ is computed. Then, the global pheromone has updated. Accordingly, to discover intercluster routing is completed, a new round of discovering intercluster routing will be started. Repeat the process until the maximum number of iterations is reached. In Figure 5, κ is the iteration number, and $N_{\max}$ is the maximum number of iterations.

5. Performance Evaluation

5.1. Simulation Environment. Simulations are conducted to evaluate the performance of KACO routing using MATLAB 2016a on local PC with an Intel i7 8th generation 3.20GHz process, 16G of RAM, and the Windows 10 Platform.

N underwater sensor nodes are distributed randomly in 500 m × 500 m × 500 m area. A sink node is located at the center of the surface, as shown in Figure 2. Table 1

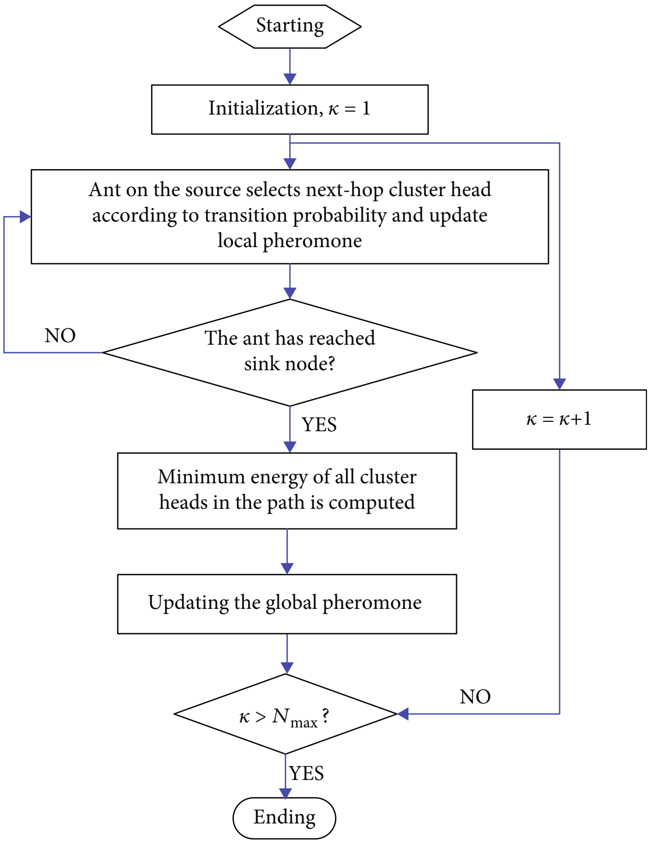

Starting
Initialization, $\kappa = 1$
Ant on the source selects next-hop cluster head according to transition probability and update local pheromone
NO
The ant has reached sink node?
YES
$\kappa = \kappa+1$
Minimum energy of all cluster heads in the path is computed
Updating the global pheromone
NO
$\kappa > N_{\max}$?
YES
Ending

Figure 5: Flowchart of discovering intercluster routing.

Table 1: Simulation parameter.

Parameter	Value
Network size	500 m × 500 m × 500 m
Number of nodes	50 to 450
Physical modern	LinkQuest UWM1000
Transmission radius of sensor node	150 m, 200 m
Acoustic frequency	30 kHz
Transmit power	2 W
Receive power	0.1 W
Data packet size	200 bits
Initial node energy	5 J
ω_1, ω_2	0.5,0.5
α, β, λ	3,6,3
ρ	0.7
Maximum number of iterations	20

summarizes the simulation parameters [17, 24] used in simulation.

Simulation results of proposed work are presented against two existing state of the art schemes: DBR and LUER. The performance is evaluated based on number of dead nodes, residual energy, and received packets at the sink node (RPSN).

5.2. Simulation Results and Discussions

5.2.1. Number of Dead Nodes. Firstly, a number of dead nodes of KACO, DBR, and LUER are analyzed when the number of nodes varies from 50 to 450. Dead node is that have consumed 95% of their energy.

Figures 6 and 7 shows the influence of number of nodes on number of dead nodes, where transmission radius of node is 200 m in Figure 6 and transmission radius of node is 150 m in Figure 7.

As can be seen from Figures 6 and 7, the number of dead nodes is less than DBR routing and LUER routing. The reason is as follows: in KACO routing, the *K*-means algorithm is used to form clusters so that the distribution of cluster is more even.

In addition, the improved ant colony algorithm is applied in discovering intercluster routing process, and the energy Gini coefficient is used to make the energy consumption among clusters more balanced. These factors make energy consumption of nodes in the network more balanced so that the number of dead nodes is reduced.

In DBR routing, the cluster has not been considered. And the next-hop forwarding node is selected based on depth of nodes. So, the path of transmitting data is longer to increase energy consumption of nodes. In LUER, routing decisions are made based on both link quality and residual energy of nodes. Unfortunately, location of nodes has not taken into account in routing process.

In addition, by comparing Figure 6 with Figure 7, it is not difficult to see that an increase in transmission radius is conducive to reducing the number of dead nodes. The reason is as follows: The larger the transmission radius of a node is, the larger the transmission range of the node is, and the shorter the hop number of the path is, which reduces the energy consumption of transmitting data.

5.2.2. Residual Energy of Nodes. In this section, average residual energy of nodes is analyzed when the number of nodes varies from 50 to 450. Figures 8 and 9 show the influence of average residual energy of nodes on number of dead nodes, where transmission radius of node is 200 m in Figure 8 and transmission radius of node is 150 m in Figure 9.

Average residual energy of is reduced when the number of nodes are increased, as in Figures 8 and 9. The reason is that the more nodes there are, the more packets generated and the more packets the have to transmit, which increases the energy consumption of nodes. Compared with DBR and LUER routing, KACO routing can effectively reduce energy consumption of nodes.

In addition, by comparing Figure 8 with Figure 9, it is not difficult to see that an increase in transmission radius is conducive to reducing energy consumption of nodes. For example, average residual energy of nodes in KACO is increased to 3 J when the transmission radius of node is from 150 m to 200 m and number of nodes is 450.

5.2.3. Received Packets at the Sink Node (RPSN). Finally, the RPSN of KACO, DBR, and LUER routing is analyzed.

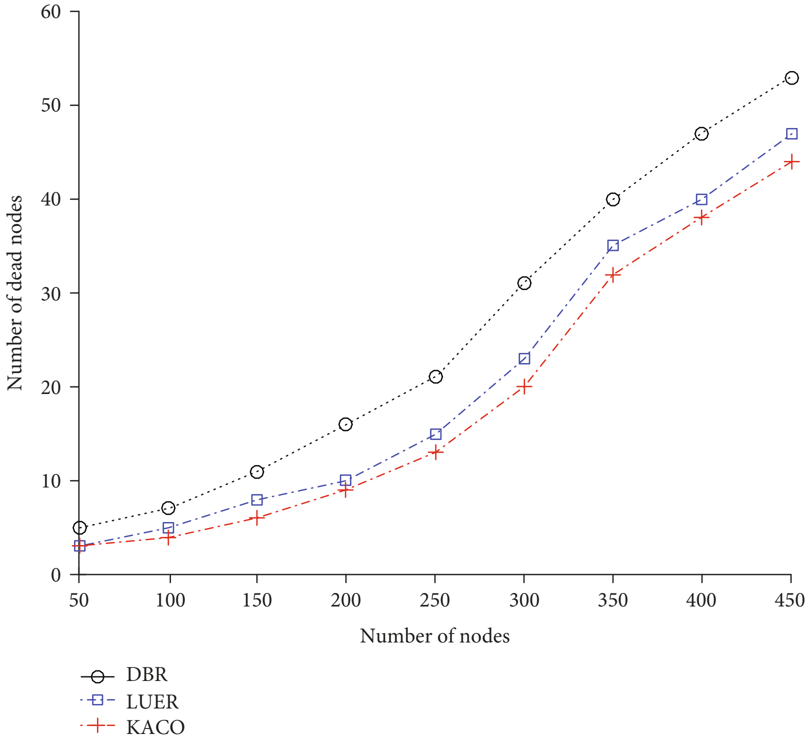


Figure 6: Number of dead nodes ($r = 200$ m).

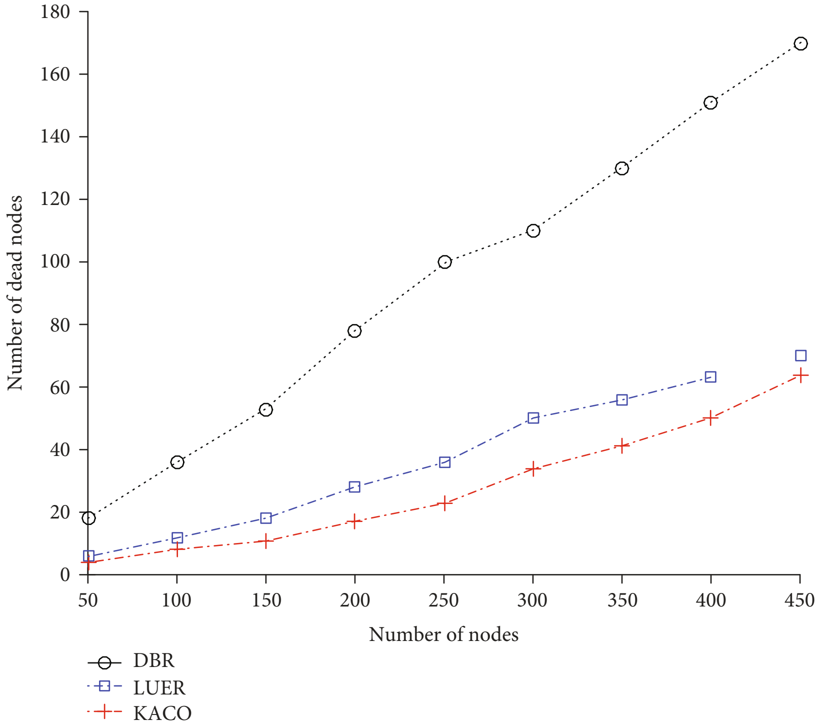


Figure 7: Number of dead nodes ($r = 150$ m).

The more packets the sink node receives, the better the routing performance is. In Figures 10 and 11, the total received packets at the sink node are evaluated in two cases. In Figure 10, transmission radius of sensor node is 200 m, and in Figure 11, transmission radius of sensor node is 150 m.

As known from Figure 10, the more nodes are, the more packets the sink node receives, which is as expected. The

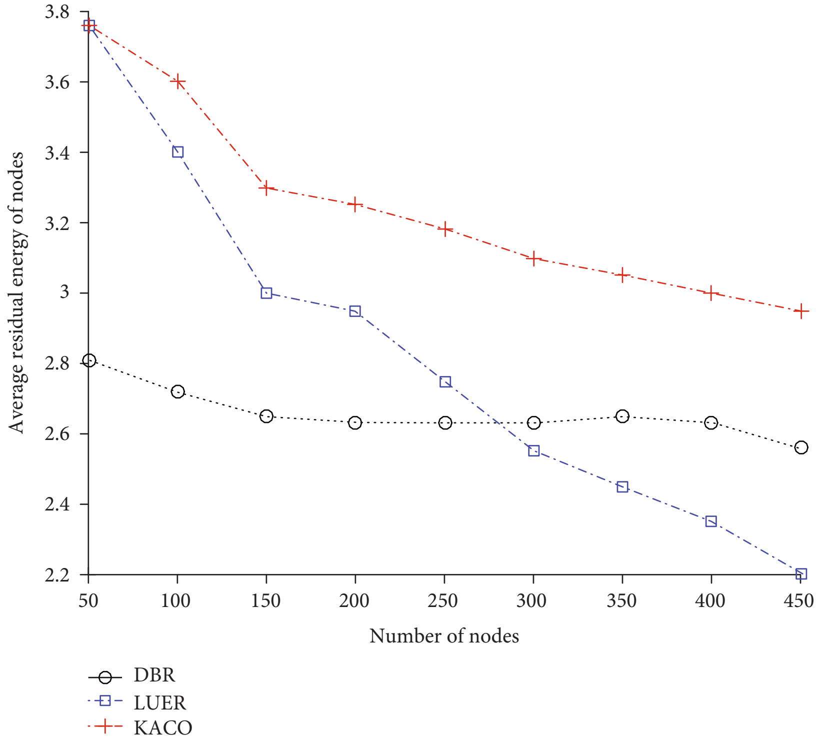


Figure 8: Average residual energy of nodes ($r = 200$ m).

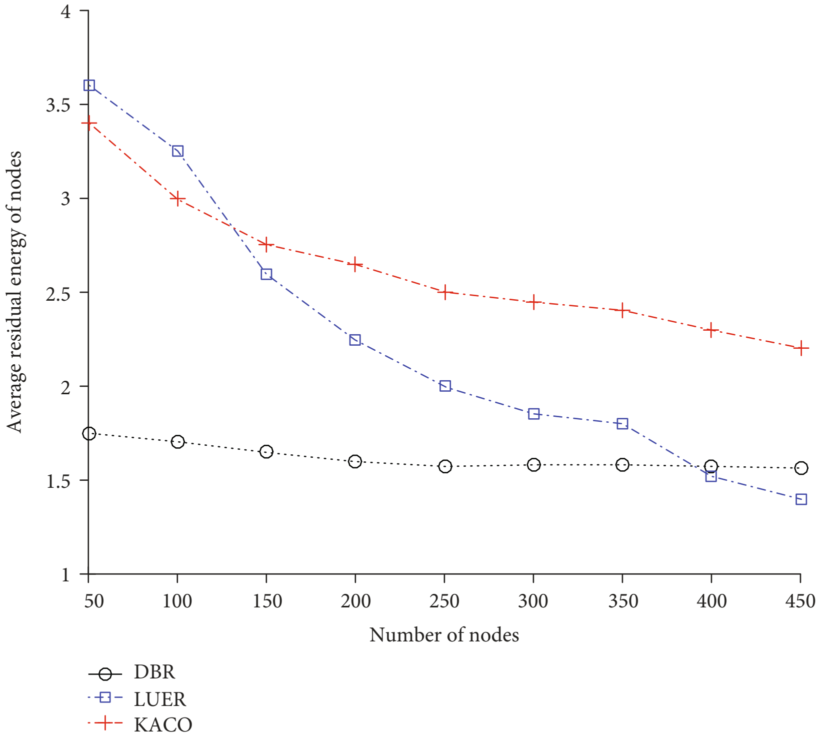


Figure 9: Average residual energy of nodes ($r = 150$ m).

more nodes are, the more packets will be generated, and the more nodes are, the better the connectivity of the network will be, which is beneficial for the sink node to receive packets. However, as the number of nodes increases, the RPSN rises slowly. The reason is that when the number of nodes increase to a certain number, the

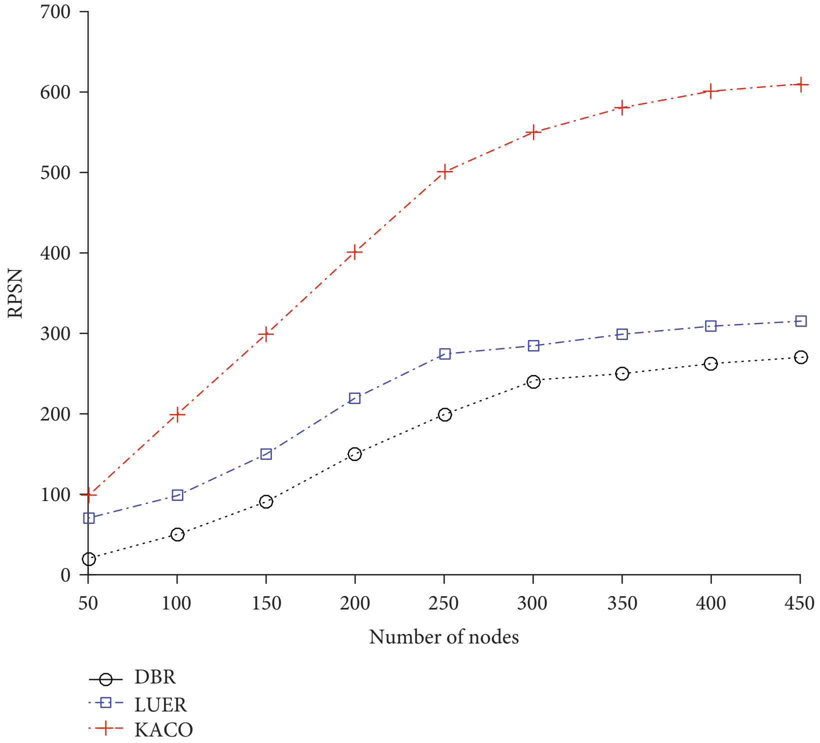


FIGURE 10: Received packets at the sink node ($r = 200$ m).

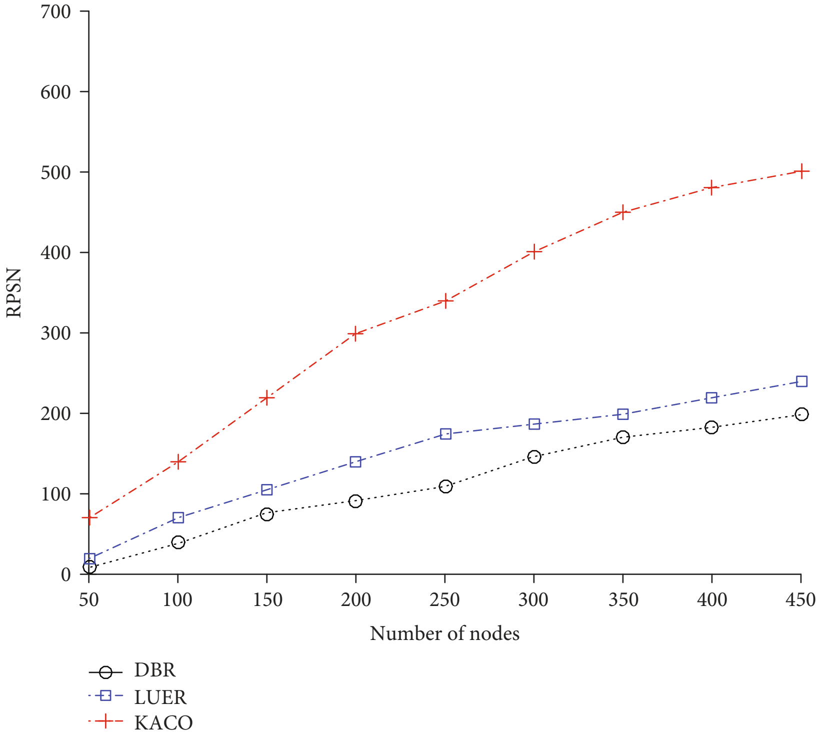


FIGURE 11: Received packets at the sink node ($r = 150$ m).

heavier the network burden is, the more energy the nodes consume, which result in the number of dead nodes (as shown in Figures 7 and 8).

As known from Figure 11, compared with DBR and LUER routing, the proposed KACO routing enables the sink node to receive more packets. This is attributed to the fact that the energy consumption of nodes is reduced, and the number of dead nodes is reduced by establishing intercluster routing based on improved ant colony optimization algorithm, so that the more packets are successfully transmitted to sink nodes. In addition, the communication range of nodes is extended, and the received packets at the sink increase when the transmission range increase.

6. Conclusions

Aiming at the routing problem in UWSNs, *K*-means and ant colony optimization-based Routing (KACO) has been proposed. In KACO, the clustering and depth of nodes are used to construct energy-efficient cluster routing. So, efficiency of transmitting data is improved to balance energy consumption among nodes.

The research work is limited to isomorphic networks (all nodes are same). In fact, the different types of nodes may have different data priorities. In the future, the issue will be addressed. In addition, data aggregation techniques can be designed for UWSN in the future.

References

[1] W. Zhang, J. Wang, G. Han, X. Zhang, and Y. Feng, "A cluster sleep-wake scheduling algorithm based on 3D topology control in underwater sensor networks," *Sensors*, vol. 19, no. 1, p. 156, 2019.

[2] A. Bagchi, "Hierarchical neighbor graphs: a topology control mechanism for data collection in heterogeneous wireless sensor networks," *Ad-hoc & Sensor Wireless Networks*, vol. 26, no. 4, pp. 171–191, 2015.

[3] Z. Rahman, F. Hashim, M. Rasid, M. Othman, and K. Ali Alezabi, "Normalized advancement based totally opportunistic routing algorithm with void detection and avoiding mechanism for underwater wireless sensor network," *Access*, vol. 8, pp. 67484–67500, 2020.

[4] N. T. Nguyen, T. T. Le, H. H. Nguyen, and M. Voznak, "Energy-efficient clustering multi-hop routing protocol in a UWSN," *Sensors*, vol. 21, no. 2, p. 627, 2021.

[5] H. Maqsood, N. Javaid, A. Yahya, B. Ali, Z. A. Khan, and U. Qasim, "MobiL-AUV: AUV-aided localization scheme for underwater wireless sensor networks [C]," in *International Conference on Innovative Mobile & Internet Services in Ubiquitous Computing*, pp. 170–175, IEEE, 2016.

[6] H. Jun and M. Zhengrong, "Efficient relay selection algorithm for cooperative routing in underwater sensor networks [J]," *Chinese Journal of sensor and actuators*, vol. 32, no. 3, pp. 458–462, 2019.

[7] M. Zhang, W. Cai, X. Zheng, and L. Zhou, "Energy aware and delay optimized pressure based routing protocol for underwater sensor networks [J]," *Chinese Journal of sensor and actuators*, vol. 32, no. 8, pp. 121–126, 2019.

[8] D. N. Sandeep and V. Kumar, "Review on clustering, coverage and connectivity in underwater wireless sensor networks: a communication techniques perspective," *IEEE Access*, vol. 5, pp. 11176–11199, 2017.

[9] A. Khan, I. Ali, A. Ghani et al., "Routing protocols for underwater wireless sensor networks: taxonomy, research challenges, routing strategies and future directions," *Sensors*, vol. 18, no. 5, pp. 1619–1628, 2018.

[10] Y. Hai, Z. J. Shi, and J. H. Cui, "DBR: depth-based routing for underwater sensor networks [C]," *International Ifip-tc6 Networking Conference on Adhoc & Sensor Networks*, pp. 34–42, 2008.

[11] W. Zhiping, L. Shaojiang, N. Weichuan et al., "An energy-efficient multi-level adaptive clustering routing algorithm for underwater wireless sensor networks [J]," *Cluster Computing*, vol. 22, no. 6, pp. 14651–14660, 2019.

[12] I. A. Hayder, S. N. Khan, F. Althobiani et al., "Towards controlled transmission: a novel power-based sparsity-aware and energy-efficient clustering for underwater sensor networks in marine transport safety," *Electronics*, vol. 10, no. 7, p. 854, 2021.

[13] N. Subramani, P. Mohan, Y. Alotaibi, S. Alghamdi, and O. I. Khalaf, "An efficient metaheuristic-based clustering with routing protocol for underwater wireless sensor networks," *Sensors*, vol. 22, no. 2, pp. 415-416, 2022.

[14] F. Ahmed, Z. Wadud, N. Javaid, N. Alrajeh, M. S. Alabed, and U. Qasim, "Mobile sinks assisted geographic and opportunistic routing based interference avoidance for underwater wireless sensor network," *Sensors*, vol. 18, no. 4, pp. 1062–1091, 2018.

[15] R. Coutinho and A. Boukerche, "OMUS: efficient opportunistic routing in multi-modal underwater sensor networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 9, pp. 5642–5655, 2021.

[16] W. A. N. G. Xin, X. U. Hai tao, J. Hua, and Q. Qin, "Multi-objective optimization opportunity routing for underwater sensor networks [J]," *Computer Engineering and Design*, vol. 41, no. 11, pp. 17–22, 2020.

[17] S. Karim, F. K. Shaikh, B. S. Chowdhry et al., "GCORP: geographic and cooperative opportunistic routing protocol for underwater sensor networks," *Access*, vol. 9, no. 4, pp. 27650–27667, 2021.

[18] M. Jouhari, K. Ibrahimi, H. Tembine, and J. Ben-Othman, "Underwater wireless sensor networks: a survey on enabling technologies, localization protocols, and internet of underwater things," *IEEE Access*, vol. 7, no. 8, pp. 96879–96899, 2019.

[19] M. Y. Durrani, R. Tariq, F. Aadil, M. Maqsood, Y. Nam, and K. Muhammad, "Adaptive node clustering technique for smart ocean under water sensor network (SOSNET)," *Sensors*, vol. 19, no. 5, pp. 1145–1156, 2019.

[20] M. Wazid and A. K. Das, "An efficient hybrid anomaly detection scheme using K-means clustering for wireless sensor networks," *Wireless Personal Communications*, vol. 90, no. 4, pp. 1971–2000, 2016.

[21] S. Arjunan, S. Pothula, and D. Ponnurangam, "F5N-based unequal clustering protocol (F5NUCP) for wireless sensor

networks," *International Journal of Communication Systems*, vol. 31, no. 17, pp. 1–14, 2018.

[22] W. Zongshan, Z. Yifan, L. Bo, Y. Jundong, and D. Hong wei, "Clustering routing algorithm for WSN based on energy balance and high efficiency [J]," *Computer Engineering and Design*, vol. 42, no. 10, pp. 2701–2710, 2021.

[23] L. Hong and L. Haowei, "Uneven clustering routing algorithm based on ant colony optimization [J]," *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, vol. 46, no. 8, pp. 50–55, 2018.

[24] S. Karim, F. K. Shaikh, K. Aurangzeb, B. S. Chowdhry, and M. Alhussein, "Anchor nodes assisted cluster-based routing protocol for reliable data transfer in underwater wireless sensor networks," *IEEE Access*, vol. 9, pp. 36730–36747, 2021.

Application of Optimized Convolution Neural Network Model in Mural Segmentation

Kisan Prayag Patro, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, kp_patro@gmail.com*

Arabinda Dash, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, arabindadash56@hotmail.com*

Prasanna Kumar Chhotaray, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, pkchhotaray85@gmail.com*

Sudesh Kumar Patel, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, sk.patel331@yahoo.co.in*

Abstract

To address the problems of blurred target boundaries and inefficient image segmentation in ancient mural image segmentation, a multi-classification image segmentation model MC-DM (Multi-class DeeplabV3+ MobileNetV2) that fuses lightweight convolutional neural networks is proposed. The model combines the Deeplabv3+ structure and MobileNetV2 network and adopts the unique spatial pyramid structure of DeeplabV3+ to process convolutional features for multi-scale fusion, which reduces the loss of detail in the mural segmentation images. Firstly, the features calculated at any resolution in MobileNetV2 network are extracted by hole convolution, the input step is expressed as the ratio between the input image resolution and the final resolution, and the density of encoder features is controlled according to the budget of computing resources. Then, the spatial pyramid pool structure is used to fuse the previously calculated features at multiple scales to enrich the semantic information of the feature image. Finally, the same convolution network is used to reduce the number of channels and filter the density feature map. The filtered features are fused with the features after multi-scale fusion again to obtain the final output. In total, 1000 scanned images of murals were adopted as datasets for testing under the JetBrains PyCharm Community Edition 2019 environment. The obtained experimental results indicate that MC-DM improves the training accuracy by 1 percentage point compared with the conventional SegNet-based image segmentation model, and by 2 percentage points compared with the PspNet network-based image segmentation model. The PSNR (peak signal-to-noise ratio) of the MC-DM model is improved by 3–8 dB on average compared with the experimental model. This confirms the effectiveness of the model in mural segmentation and provides a novel method for ancient mural image segmentation.

1. Introduction

Ancient murals are mediums of Chinese culture and have significant historical value; however, under natural and human impact, the ancient murals of the distant past have been exposed to various degrees of damage, while their content have been severely damaged. Hence, image restoration in murals has become one of the most difficult problems faced by cultural workers and historical researchers in the course of analyzing ancient murals. Mural segmentation is the first step of image analysis and plays a very important role in image engineering. Image feature extraction, target recognition and target detection all depend on the quality of image segmentation in the later image process stage. Similarly, as a key step of mural digital protection, mural segmentation is the basis of mural classification and restoration. The segmentation results directly affect the process of cultural relics protection. Therefore, the research on mural segmentation methods has attracted more and more attention.

Deep learning, a learning method based on artificial neural networks that imitates the human brain to process and interpret data, is a new field in machine learning research and is widely used in several fields such as image and sound processing. Deep learning can combine neural networks with probabilistic models to improve the inference ability of image models; hence, in the field of image segmentation, various image segmentation models based on deep learning have been proposed to effectively solve series of problems such as blurred image edge segmentation and missing information of segmented images in conventional segmentation methods. Based on the above, this paper proposed a new DeeplabV3+ model by improving the deep learning model and applied it to the segmentation of ancient murals.

2. Related Works

About the application of deep learning model in the field of image segmentation, researchers adopted fully convolutional networks (FCN) [1] or improved FCN networks for image segmentation initially, which featured a fully connected layer of convolutional neural networks (CNN) [2] replaced with a convolutional layer to adapt to arbitrary size inputs and output low-resolution segmented images. However, this method has significant limitations: the edge segmentation of FCN is poor, and the contour of the segmented image is blurred. Chen et al. [3] proposed the DeeplabV1 model in 2015 to address this problem, which utilizes a fully-connected conditional random field (CRF) to optimize boundary segmentation and effectively solve the edge contour segmentation problem inherent in FCN. The DeeplabV3+ [4] model is a modification of the previous generation DeeplabV3 model; it is a novel and improved scheme that can help researchers refine segmentation results and fares better in the delineation of object boundaries. In 2019, Ren et al. [5] combined the DeeplabV3+ model with the super pixel segmentation algorithm, simple linear iterative cluster (SLIC), and experimentally demonstrated that DeeplabV3+ has a better image detail restoration ability than FCN and SegNet segmentation [6] models. Image segmentation technology based on deep learning has been developing. Especially since the 2020 COVID-19, image segmentation technology has developed rapidly in medicine [7–10], which provides a new way of thinking for ancient murals.

In ancient mural image segmentation, conventional segmentation methods are mostly used, and these segmentation models are not universally applicable. Conventional mural segmentation methods apply various approaches, and one of them involves using the fuzzy C-mean (FCM) [11–13]. This objective-based fuzzy clustering algorithm is widely used, and its algorithm theory is mature; however, if this algorithm is employed in the mural segmentation field, it will be affected by sample imbalance. When the sample capacity of different classes is not consistent, it will cause a certain class of segmentation samples to encounter difficulties in approaching the target sample, which triggers poor segmentation results. The second approach is the mean drift algorithm-mean shift [14–17], which is essentially an estimation algorithm for kernel density; however, this algorithm runs slowly, and is only applicable to feature data point sets for which standard features have been established in mural segmentation; in addition, it is prone to the presence of images outside the target or missing parts of the target, and has a limited effect when performing batch segmentation. The third conventional mural segmentation algorithm, graph cuts [18–20], adopts a graph form to solve the energy function and assign corresponding weights to the edges of the graph and transform the energy function into an S/T graph for complete image segmentation. However, this method exhibits a poor segmentation effect when handling noise or occlusion, and it requires manual labeling of a number of front and back view pixel points, which has a series of problems such as manual intervention [21].

Based on the efficiency of deep learning neural networks, this study proposes a multi-class lightweight network segmentation model (Multi-class DeeplabV3+ MobileNetV2, MC-DM) that combines the lightweight convolutional neural network MobileNetV2 [22] with the DeeplabV3+ model. The model uses the DeeplabV3 + structure to collect multi-scale information of the image, effectively circumventing the missing semantic information of the image; in addition, it adopts the MobileNetV2 convolutional neural network to extract features, improve the efficiency of mural segmentation, and reduce the influence of hardware conditions on the segmentation effect. Experiments indicate that the method has different degrees of segmentation accuracy and efficiency in the mural image segmentation process and exhibits optimal robustness in terms of image segmentation edge continuity.

3. Materials and Methods

The improved MC-DM model involves lightweight neural network MobileNetV2, DeeplabV3+ model, ASPP structure and other related concepts. Therefore, we divide two parts to introduce: Relevant theories and Mural segmentation model MC-DM. Relevant theories focuses on the working principle of relevant network and model structure; Mural segmentation model MC-DM introduces the improvements and excellent characteristics of the proposed model.

3.1. Relevant Theories. In this part, we discuss the network structure, convolution mode of MobileNetV2 network and the working principle of DeeplabV3+ model respectively.

3.1.1. MobileNetV2. The MobileNetV2 convolutional neural network is proposed to solve the problems of large convolutional neural networks and insufficient hardware training that emerge during the training of image models. It is an important approach to addressing the hardware memory limitation of deep learning models deployed in mobile devices [23]. It is another important invention after SqueezeNet [24], ShuffleNet [25], Xception [26], and other lightweight neural convolutional networks. The core part of the network is depthwise separable convolution, and its operation comprises two parts: depthwise (DW) and pointwise (PW) convolution. With a 3×3 convolution kernel and a large number of channels, DW separable convolution can reduce the computational effort by approximately 9 times less than normal convolution.

Based on the first-generation lightweight network MobileNetV1, the MobileNetV2 network introduces the concepts of inverted residuals and linear bottlenecks; this limits the feature extraction in terms of the number of input channels because DW convolution does not change the number of channels. These two parts adopt the low dimensional compression as input, expand it to a high dimension, and then filter it using lightweight deep convolution; subsequently, the resulting features are represented by linear convolution projection into low dimension. The net structure of the MobileNetV2 is presented in Table 1.

In Table 1, *t*, *c*, *n*, and *s* denote the dilation factor, number of output channels, number of repetitions of the convolution layer, and step size, respectively. The first layer of each sequence has a step size, all other layers adopt a step size of 1, and all spatial convolutions employ a 3 × 3 convolution kernel. Each bottleneck contains three parts: dilation, convolution, and compression, with each row describing one or more sequences, repeated *n* times, and all layers in the same sequence having the same number of output channels. The MobileNetV2 network facilitates a significant reduction in the memory footprint problem required during inference by not fully specifying the intermediate tensor, and its application to mural segmentation can reduce the need for main memory accesses in most embedded hardware designs.

3.1.2. Conventional DeeplabV3+ Model. The DeeplabV3+ model is an improvement of the DeeplabV3 model with the residual neural network (ResNet) network as the underlying network, which also encodes an encoder-decoder structure to obtain clear object boundaries by recovering spatial information to optimize boundary segmentation. The ResNet network or Xception network is used for the feature extraction of the input image, after which, the image special is fused via atrous spatial pyramid pooling (ASPP) to prevent information loss. In the DeeplabV3+ model, the DeeplabV3 model is adopted as the encoder part with an external simple and effective decoder module to obtain clear results.

Null convolution with multiple null rates (rates) is employed in Deeplabv 3+ to efficiently extract contextual information in parallel, and this structure adopts the ASPP model to provide multi-scale information. The structure of this model is illustrated in Figure 1.

The ASPP module comprises a 1 × 1 convolutionand three 3 × 3 null convolutions with sampling rates of 6, 12, and 18, respectively. In the Deeplabv 3+ model, the input image is divided into two parts after passing through the backbone deep neural convolutional network, with one part going into the decoder and the other part going into the parallel null convolutional structure i.e., the ASPP model. Separate feature extraction is performed with different rates of void convolution, which is then merged, followed by 1 × 1 convolution, for which feature compression is performed. Then, the compressed feature map is upsampled four times via bilinear interpolation, to pass into the decoder.

3.2. Mural Segmentation Model MC-DM. In this section, we focus on the improvement of MC-DM model and the working principle of each part of the model.

Table 1: Structure of MobileNetV2 network.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	—	32	1	2
$112^2 \times 32$	Bottleneck	1	16	1	1
$112^2 \times 16$	Bottleneck	6	24	2	2
$56^2 \times 24$	Bottleneck	6	32	3	2
$28^2 \times 32$	Bottleneck	6	64	4	2
$14^2 \times 64$	Bottleneck	6	96	3	1
$14^2 \times 96$	Bottleneck	6	160	3	2
$7^2 \times 160$	Bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1 × 1	—	1280	1	1
$7^2 \times 1280$	Avgpool 7 × 7	—	—	1	—
$1 \times 1 \times 1280$	conv2d 1 × 1	—	k	—	

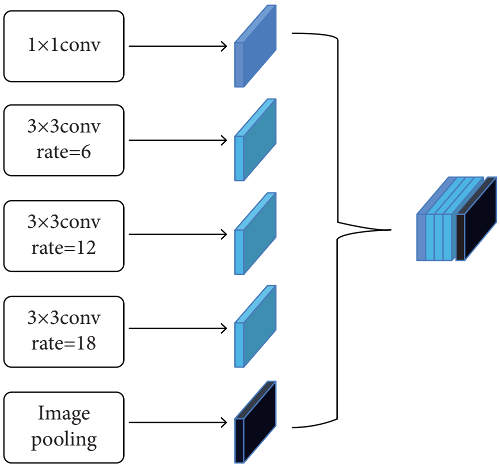


Figure 1: ASPP model structure diagram.

3.2.1. DeeeplabV3+ MC-DM Incorporating MobileNetV2. The DeeplabV3+ underlying network is highly adaptive. To achieve segmentation accuracy, researchers have incorporated ResNet. Although such models have high classification accuracy, their model depth keeps deepening, which triggers an increase in model complexity. Complex segmentation models are constrained by hardware memory and are demanding for mobile or embedded devices, which cannot satisfy the segmentation requirements of low latency and high response rate in specific scenarios. To address this problem, a segmentation model that combines the lightweight neural network MobileNetV2 with the segmentation model DeeplabV3+ is proposed. The encoder module in the model is employed to reduce feature loss and capture higher-level semantic information, while the decoder module is used to extract details and recover spatial information. The model decomposes the convolution into two independent layer factors to replace the full convolution operator, performs light filtering by applying a single convolution filter to each input channel, and later constructs new features via the linear combination of the input channels. The changes to the convolutional network optimize the performance of the DeeplabV3+ decoder module in recovering detailed object boundaries.

With the same dataset, the MC-DM model adopts a network that has a significant advantage in segmentation efficiency over convolutional networks such as ResNet and Xception. The most significant difference between this model and the conventional DeeplabV3+ is that instead of using standard convolution to extract features, the DW convolution that can perform feature extraction in high dimensions is adopted. The advantage of this method is that it makes the MC-DM model substantially less computationally intensive than the conventional DeeplabV3+ model, which can be applied in the mural segmentation field to satisfy the efficient requirements of mural segmentation while ensuring accuracy. The improved model is illustrated in Figure 2.

The first improvement of the model is the combination of hole convolution structure and deep separable network structure. In Figure 2, structure A represents the null convolution, which extracts features computed at arbitrary resolution from the MobileNetV2 network, expresses the input step size as the ratio obtained from the input image resolution to the final resolution, and controls the density of encoder features based on the budget of computational resources, to control the budget for encoder computational resources. For the semantic segmentation task, an output with a step size of 16 is used for more intensive feature extraction after discarding the span in the last one or two blocks. This approach is taken because when the decoder output step size is 8, the segmentation performance is improved relative to the output step size of 16, and although the performance is improved, it increases the computational complexity. Therefore, in the MC-DM, the output step size used for the encoder module is 16, which has the advantage of balancing segmentation accuracy and speed.

The second improvement of the model is combining the spatial pyramid pool with MobilNetV2, as shown in structure B. The structure uses hole convolution with different hole rates to fuse the features calculated by MobilNetV2 at multiple scales, which enriches semantic information and effectively balances accuracy and running time.

The third improvement of the model is using the same convolution network to reduce the number of channels and modify the output step of the model, as shown in structure *C*. The use of the same convolution network solves the problem of training difficulty caused by a large number of channels in low-level features. Secondly, we modify the output step setting and set its value to 4, which can make appropriate trade-offs for density feature mapping and simplify the Decoder module under the condition of limited GPU resources, so as to improve the image segmentation efficiency of the model.

3.2.2. Description of the Algorithm. The workflow of the MC-DM segmentation model is presented in the following steps.

Step 1: Input the mural image of the fixed size and resolution into the segmentation model.

Step 2: Perform the feature extraction of the image using improved depth separable network and retain the mural image detail information using null convolution.

Step 3: Shunt low-level features into the ASPP and Decoder structures, respectively, to retain image feature information to a great extent.

Step 4: Multi-scale-fuse the feature information passing through the ASPP structure via 1×1 convolution and feed the fusion result into the decoder structure; the low-level features that initially enter the decoder structure are refined by different convolution layers to refine the features.

Step 5: Upsample the encoder output feature map via bilinear interpolation and maintain the same size as that of the feature map after the feature refinement in decoder. The sampled results are fused with the refined results and features to obtain a more feature-rich mural image.

Step 6: Upsample the feature fusion image again to obtain a segmented image with the same parameters as the input image, and then complete the segmentation process.

4. Analysis of Experimental Results

4.1. Experimental Environment and Data Sources. The personal computer environment for the experiments is Windows 10 with Intel Core i7-9750H CPU, NVIDIA GeForce 1660Ti GPU, and 8G RAM. The TensorFlow deep learning framework was used to train and test the semantic segmentation model in the text.

The dataset of DeeplabV3+ employed a single-channel annotated map, and the experimental images were obtained from the scanned images of the album "The Complete Collection of Dunhuang Murals in China," while the image annotation of the scanned images was developed into a dataset by the graphic user interface annotation software, LabelMe. The sample dataset graph is presented in Figure 3.

Figure 3(a) represents the scanned image, based on which the edges of the scanned image are labeled point by point using floating points and the labeled points are connected to form the result presented in Figure 3(b). Subsequently, based on the original and annotated images, a single-channel grayscale image was trained and merged with the scanned image to form the dataset. The dataset contains 1000 images divided into five categories: animals, houses, people, auspicious clouds, and Buddha images, with 200 images in each category. The images were pre-processed using the letterbox function to prevent missing frames in the images during the training process. To reduce the occurrence of overfitting triggered by few images, experiments were performed to enhance the obtained images. The enhancement was carried out by changing the color of the image, increasing the noise, and changing the brightness. Figure 4 presents the image obtained from the data enhancement.

The original image is presented in column (a) of Figure 4, while the last four columns depict the enhanced images. The obtained results need to be tested several times owing to the stochastic nature of the enhancement with functions. In the experimental phase, 90% of the dataset was adopted for training and 10% for prediction. The experiments were

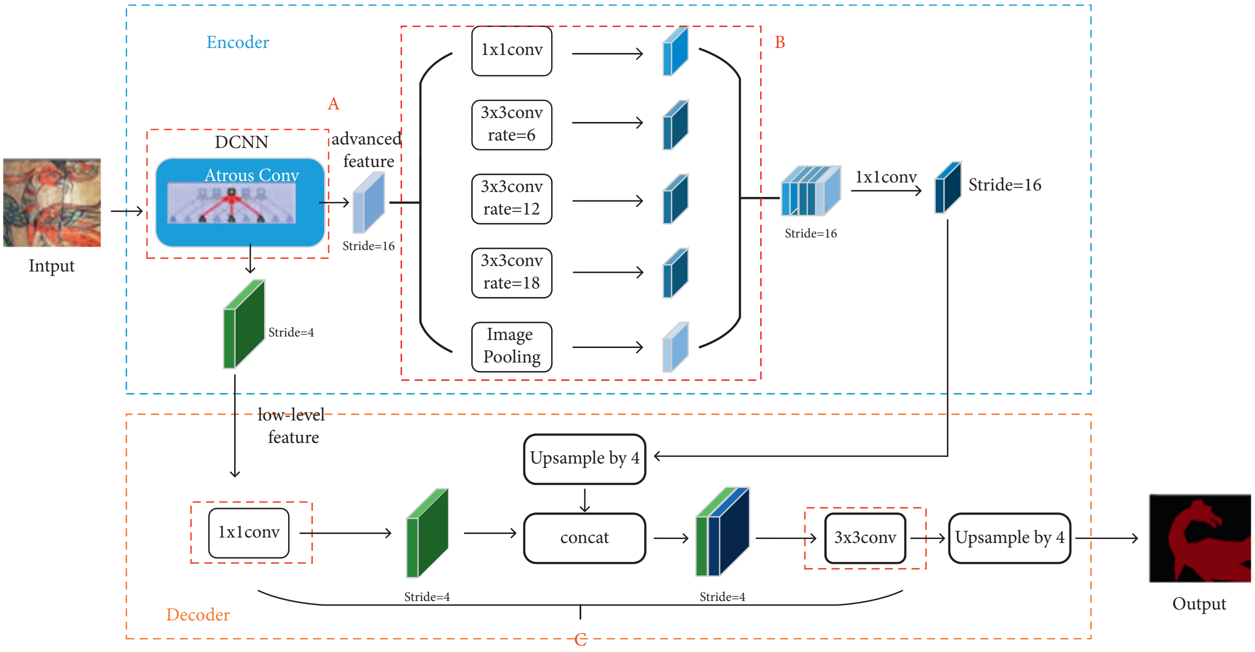


Figure 2: MC-DM diagram.

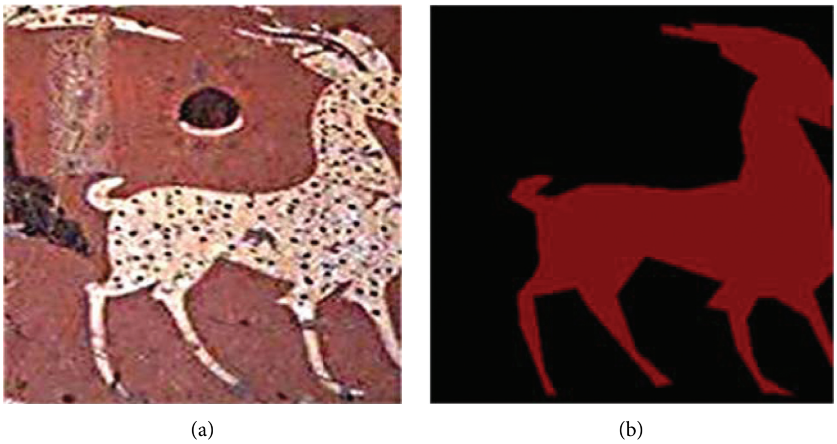


Figure 3: Sample plot of DeeplabV3+ dataset. (a) Scanned image of an ancient mural. (b) Label image of an ancient mural.

limited to the accuracy of the test set, and when the loss value val_loss of the test set did not decrease twice in a row, the learning rate was reduced to continue the training. Training was cut off when the loss value stabilized, and the obtained data were saved every 30 generations. The variation of the splitting accuracy is presented in Figure 5.

To improve the experimental training accuracy, the first 10 generations of test set loss values took a wide range, which triggered large fluctuations in the experimental test set training accuracy. After 10 generations, the overall accuracy of the experiments and the test set training accuracy gradually increased and stabilized at the 40th generation, while the learning rate reached the optimum.

4.2. Comparative Experiments. Three different image segmentation models were designed for comparisons with the models presented in this study, based on the homemade dataset. First, the MobileNetV2 network was combined with the models in [27, 28] as comparison Models 1 and 2, respectively. The model in [29] was adopted as comparison Model 3. All three models were altered to ensure that one part of the combined model remained unchanged and optimally comparable. Four images of different types were selected from the dataset for segmentation, and the segmentation results were labeled at pixel level to obtain the visual comparison effect, and the obtained results are presented in Figure 6.

In Figure 6, column (a) shows the image to be segmented, columns (b), (c), and (d) illustrates the segmented images of the original image under Models 1, 2, and 3, respectively, and column (e) presents the segmentation results of the MC-DM model. In Model 1, owing to the use of continuous downsampling resulting in a large amount of

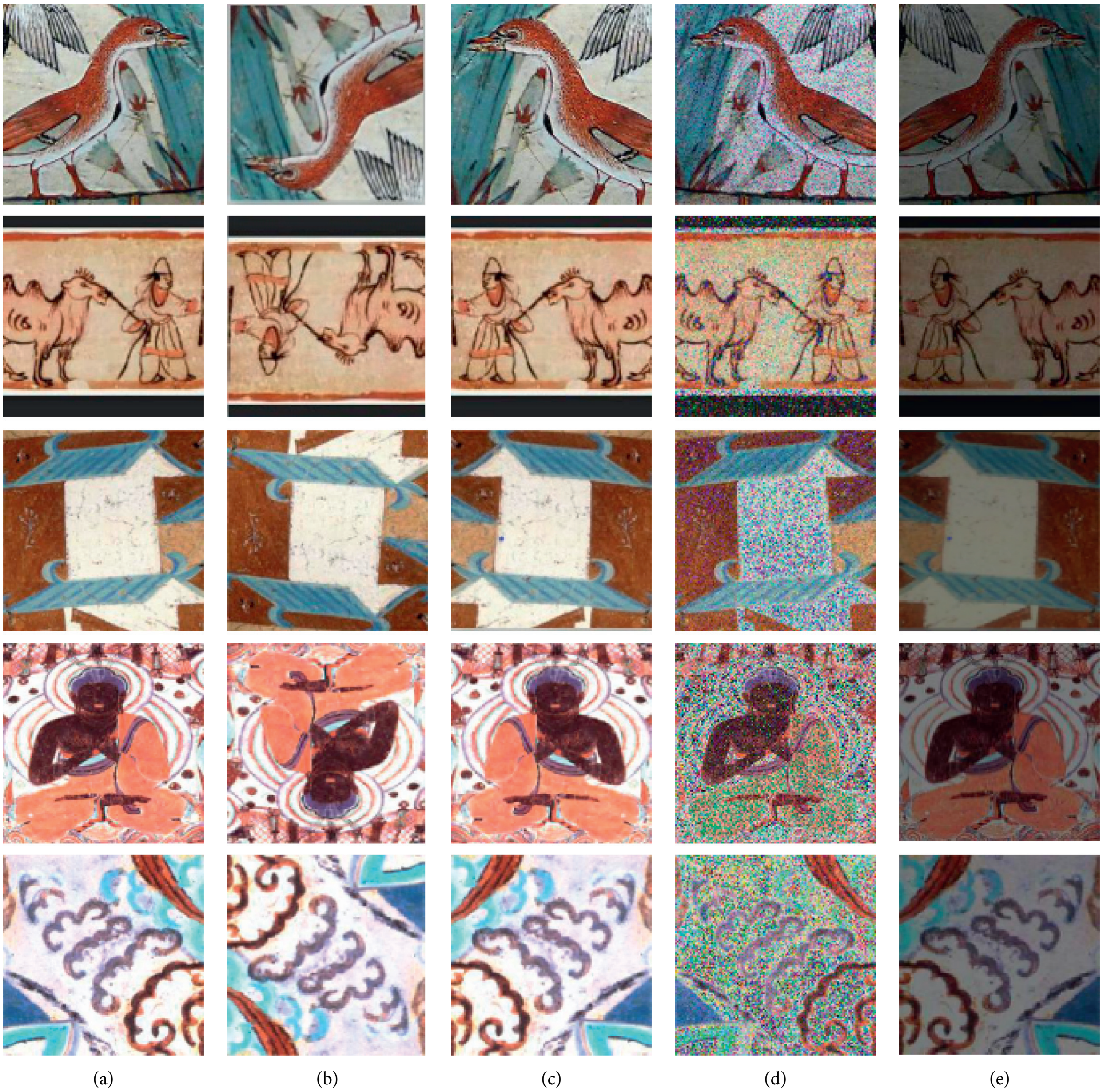


Figure 4: Data enhancement image. (a) Original image. (b) Upside down. (c) Flipped. (d) Noise(a). (e) Darken.

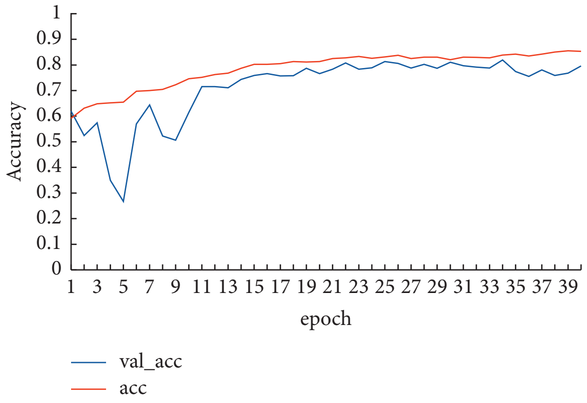


Figure 5: Accuracy variation graph.

spatial information in the input image overlapping each one-pixel on the output feature map, multiple image spatial information with lossy boundary information is not conducive for image segmentation. Model 2 first performs multi-scale pooling of the input feature information, after which the pooling results are upsampled, and then stitched. The advantage of this approach is that the information of different sensory fields can be utilized to enrich the image content; however, it easily triggers a situation where severe loss of single-category image information occurs, and the segmented edges do not match the real edges, as illustrated in Figure 6(c). Model 3 combines the DeeplabV3+ model and Xcepton network, which ignorantly increases the number of convolutional network parameters, thereby increasing the difficulty of image training; in addition, the

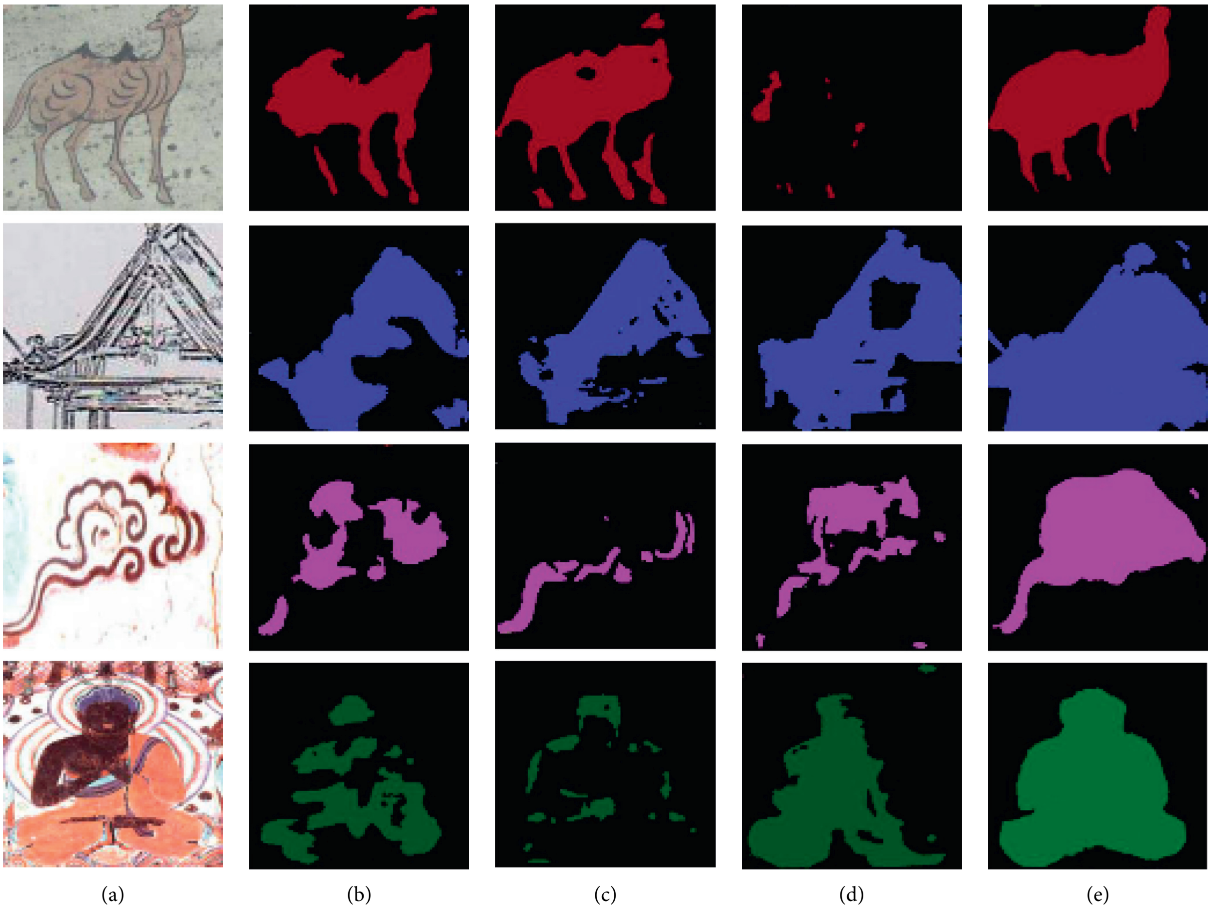


Figure 6: Comparison of segmentation effects. (a) Sample image. (b) Model 1. (c) Model 2. (d) Model 3. (e) MC-DM.

image segmentation results are influenced by the hardware equipment, and the loss of details in the center of the segmented image is severe. The MobileNetV2 network used by the MC-DM segmentation model reduces the number of networks. Furthermore, increasing the decoder structure extracts the image details, and the segmentation effect is the best among the four models.

The peak signal-to-noise ratio (PSNR) is adopted as an objective indicator, and the magnitude of the value represents the frame loss rate of the segmented image. Accordingly, a higher value represents a better image segmentation effect. The results of the PSNR values for four randomly selected samples are presented in Table 2.

Table 2: PSNR (dB) comparison.

Sample	Model 1	Model 2	Model 3	MC-DM
1	25.86	26.39	21.42	26.66
2	16.08	15.91	17.85	20.00
3	16.67	14.61	15.98	21.98
4	20.85	18.62	21.98	30.73

In Sample 1, the sample image lines are simple, the four segmentation models have similar effects, and the MC-DM segments the image with the highest PSNR value, which is 1 dB higher than the comparison model. Samples 2 and 3 have relatively complex image contours, and partial fusion exists between the target and background, while MC-DM segments the image with a significant increase in the PSNR value, which is 5 dB higher than the comparison model on average. Sample 4 has an image with a complex structure and more background information, which exerts a more significant impact on the segmentation results of the image. MC-DM performs well in the segmentation results of this sample, and the PSNR values are improved by 10 dB on average, compared with the comparison model; in addition, the experiment verifies the feasibility of this model in mural segmentation. The training accuracies of the four models are presented in Table 3.

Model 1 adopts deconvolution and up-pooling, which can only barely recognize the image shape, and its segmentation results are coarse. Model 2 has more missing details in the center of the image, although features of different sizes are obtained by multi-scale pooling. Model 3 improves the underlying network of the model; in addition, using depth-separable convolution, it optimizes the feature extraction method in the mural image segmentation process, but its segmentation results are poor for a single-category image. The improved model, MC-DM, is the most efficient in the mural segmentation process and it addresses the problem of missing details in Model 2. Compared with Model 1, MC-DM-segmented image edges are completely preserved, and the loss of image information is not

Table 3: Training accuracy table.

Segmented model	Accuracy/%
Model 1	84.07
Model 2	83.59
Model 3	84.88
MC-DM	85.30

significant. MC-DM model exhibits better applicability than Model 3 and does not exhibit the phenomenon of large differences in segmentation results due to different image types. Hence, it can be inferred from the two experimental parameters of PSNR and training accuracy that the segmentation effect of MC-DM is better than that of the other three models, and the model segmentation contour tends to the ideal contour without causing a large number of missing details.

5. Conclusions

Ancient Chinese murals are an important witness of Chinese civilization and an inseparable part of the development of the history of world civilization. Due to the long history, murals are negatively affected by many factors such as environment and man-made, and there are many problems such as image deformity, falling off and cracks. How to effectively preserve these precious cultural relics is the top priority at present. In this paper, the deep learning model is integrated into mural image segmentation, and the powerful learning ability of neural network is used to improve the problems of traditional segmentation methods, such as fuzzy image edge segmentation, which is a new exploration in ancient mural image processing. The main contributions of this paper are reflected in the following two aspects: (1) MC-DM model is proposed. The model uses hole convolution and lightweight neural network to extract mural image features, adjust the output step of Decoder structure, and balance the accuracy and speed of network segmentation. The same convolution network is used to reduce the number of channels. The density feature mapping is properly selected to reduce the difficulty of model training in the case of limited GPU resources. (2) The proposed MC-DM model is applied to the segmentation of ancient murals to solve the problems of unclear segmentation target boundary and low segmentation efficiency of traditional mural segmentation models. Based on the idea of deep learning, this paper carries out image segmentation of ancient murals, makes a systematic research on feature extraction and feature fusion on the basis of the original research, improves the ability of image feature restoration, effectively interprets the mural image meaning, and provides a new idea for the research of digital protection of ancient cultural relics.

However, there are still some problems in the experimental process, such as small data scale, lack of feature information, poor effect of multi sharp point image segmentation and so on. The model proposed in this paper still needs to be improved to meet the changing practical requirements. Therefore, the future work will be carried out from the following two aspects: (1) In the experimental stage, because DeeplabV3+, like other models in the Deeplab family, requires a specific dataset, the samples need to be manually labeled in the early stage, which is a huge workload. This problem can be solv ed by continuing to collect high-quality mural images with high quality and rich image information, and constantly expanding the number of images in the data set. The continuous improvement of the data set will make the training of the model more sufficient and better, which could av oid the problems of ov er fitting and under fitting caused by the lack of data. (2) The problem of blurred edges emerges in the multi-category image segmentation due to the geometric reduction of the experimentally encoded output feature maps relative to the input images. This is also a problem that needs to be further addressed in the future segmentation of ancient wall paintings.

References

[1] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, no. 6, pp. 119–133, 2019.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[3] L.-C. Chen, Y.-K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, Munich, Germany, September 2018.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[5] F.-L. Ren, X. He, Z.-H. Wei, L. You, and L. I. Mu-Yu, "Semantic segmentation based on DeepLabV3+ and superpixel optimization," *Optics and Precision Engineering*, vol. 27, no. 12, pp. 2722–2729, 2019.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[7] S. Dash, M. R. Senapati, P. K. Sahu, and P. S. R. Chowdary, "Illumination normalized based technique for retinal blood

vessel segmentation," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, 2020.

[8] A. Ranganath, M. R. Senapati, and P. K. Sahu, "Classification of textures using pixel range calculation method," in *Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, February 2021.

[9] A. Ranganath, P. K. Sahu, and M. R. Senapati, "Detection of COVID from chest X-ray images using pivot distribution count method," in *Proceedings of the 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 373–378, Noida, India, August 2021.

[10] A. Ranganath, P. K. Sahu, and M. R. Senapati, "A novel approach for detection of coronavirus disease from computed tomography scan images using the pivot distribution count method," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 10, no. 2, pp. 145–156, 2021.

[11] R. Lan and Y. Lin, "Suppressed nonlocal space Intuitionistic Fuzzy C-Means Image segmentation algorithm," *Journal of electronics and information*, vol. 41, no. 6, pp. 1472–1479, 2019.

[12] J.-J. Sun and Y. Xu, "Underdetermined mixed matrix estimation based on weighted improved fuzzy C-means clustering," *Computer Applications*, vol. 40, no. 6, pp. 1769–1773, 2020.

[13] S. Mishra, P. Sahu, and M. R. Senapati, "MASCA–PSO based LLRBFNN model and improved fast and robust FCM algorithm for detection and classification of brain tumor from MR image," *Evolutionary Intelligence*, vol. 12, no. 4, pp. 647–663, 2019.

[14] S.-N. Zhao and W.-J. Wang, "Image segmentation algorithm based on SVM and fast mean shift," *Journal of Small Microcomputer Systems*, vol. 38, no. 7, pp. 1614–1618, 2017.

[15] I. A. Iswanto, T. W. Choa, and B. Li, "Object tracking based on meanshift and particle-kalman filter algorithm with multi features," *Procedia Computer Science*, vol. 157, pp. 521–529, 2019.

[16] M. Lu and Y. Xu, "Overview of target tracking algorithms," *Acta Automatica Sinica*, vol. 45, no. 7, pp. 1244–1260, 2019.

[17] S. Das, A. Patra, S. Mishra, and M. R. Senapati, "A self-adaptive fuzzy-based optimised functional link artificial neural network model for financial time series prediction," *International Journal of Business Forecasting and Marketing Intelligence*, vol. 2, no. 1, p. 55, 2015.

[18] B. Martin, D. Paulusma, and E. J. van Leeuwen, "Disconnected cuts in claw-free graphs," *Journal of Computer and System Sciences*, vol. 113, pp. 60–75, 2020.

[19] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, and S. Pandiyan, "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes Classifier," *Measurement*, vol. 163, Article ID 107922, 2020.

[20] A. Ranganath, M. R. Senapati, and P. K. Sahu, "Estimating the fractal dimension of images using pixel range calculation technique," *The Visual Computer*, vol. 37, no. 4, pp. 635–650, 2021.

[21] M. Antonello, S. Chiesurin, and S. Ghidoni, "Enhancing semantic segmentation with detection priors and iterated graph cuts for robotics," *Engineering Applications of Artificial Intelligence*, vol. 90, Article ID 103467, 2020.

[22] M. Sandler, A. Howard, M.-L. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[23] Q.-H. Li, C.-P. Li, J. Zhang, H. Chen, and S.-Q. Wang, "Survey of compressed deep neural network," *Computer Science*, vol. 46, no. 9, pp. 1–14, 2019.

[24] K. Nan, S. Liu, J. Du, and H. Liu, "Deep model compression for mobile platforms: a survey," *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 677–693, 2019.

[25] F. Daghero, D.-J. Pagliari, and M. Poncino, "Energy-efficient deep learning inference on edge devices," *Advances in Computers*, vol. 122, pp. 247–301, 2020.

[26] J.-J. Jiang, Y.-F. Xiong, and X. Xia, "A manual inspection of Defects4J bugs and its implications for automatic program repair," *Science China (Information Sciences)*, vol. 62, no. 10, pp. 31–46, 2019.

[27] S.-Q. Luo, Z.-C. Zhang, and Q. Yue, "Image semantic segmentation based on improved SEGNET model," *Computer Engineering*, vol. 47, no. 4, pp. 256–261, 2021.

[28] J. Zhou, M. Hao, D. Zhang, P. Zou, and W. Zhang, "Fusion PSPnet image segmentation based method for multi-focus image fusion," *IEEE Photonics Journal*, vol. 11, no. 6, pp. 1–12, 2019.

[29] X. Tian, L. Wang, and Q. Ding, "Image semantic segmentation based on deep learning," *Journal of Software*, vol. 30, no. 2, pp. 40–46, 2019.

Speech as a Biomarker for COVID-19 Detection Using Machine Learning

Rasmi Sarangi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, rasmisarangi226@yahoo.co.in*

Rakhi Jha, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, rakhijha91@yahoo.co.in*

Umakanta Dash, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, umakanta.das221@gmail.com*

Tapas Ranjan Baitharu, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar,tkbaitharu11@gmail.com*

Abstract

The use of speech as a biomedical signal for diagnosing COVID-19 is investigated using statistical analysis of speech spectral features and classification algorithms based on machine learning. It is established that spectral features of speech, obtained by computing the short-time Fourier Transform (STFT), get altered in a statistical sense as a result of physiological changes. These spectral features are then used as input features to machine learning-based classification algorithms to classify them as coming from a COVID-19 positive individual or not. Speech samples from healthy as well as "asymptomatic" COVID-19 positive individuals have been used in this study. It is shown that the RMS error of statistical distribution fitting is higher in the case of speech samples of COVID-19 positive speech samples as compared to the speech samples of healthy individuals. Five state-of-the-art machine learning classification algorithms have also been analyzed, and the performance evaluation metrics of these algorithms are also presented. The tuning of machine learning model parameters is done so as to minimize the misclassification of COVID-19 positive individuals as being COVID-19 negative since the cost associated with this misclassification is higher than the opposite misclassification. The best performance in terms of the "recall" metric is observed for the Decision Forest algorithm which gives a recall value of 0.7892.

1. Introduction

The most basic functions of the human body are usually monitored by measuring the vital signs—temperature, heart (pulse) rate, respiratory (breathing rate), and blood pressure [1]. These are usually measured using medical devices, but nowadays, easy to use and low-cost devices and smart gadgets are available which allow measuring temperature, pulse rate, and blood pressure at home, even by nonmedical professionals. Various types of sensors, present in such devices, sense some signal generated by the human body, process the signal, and provide a reading of the vital sign in a simple and easy to interpret format. It is therefore necessary to connect the devices/sensors at appropriate points on the human body to obtain the desired measurements. The placement of such sensors on the body is invasive and intrusive and causes inconvenience to the patient/individual. This is particularly true when monitoring professional athletes and sports persons while they are performing intense exercise/training. Moreover, it does not allow measurement of body parameters without attaching the device/its sensors (probes) to the body or from a remote location; i.e., the patient has to be at the same location as the medical device. This article investigates the use of speech as a biomedical signal to detect COVID-19 based on a statistical analysis of speech and binary classification using machine learning. Speech characteristics of an individual get altered as a result of physiological and emotional changes [2–6]. Other factors that can cause physiological changes in the body are changes in health conditions, aging, stress, pollution exposure, and physical activity. There is significant evidence from the literature that clearly establishes a correlation between the characteristics of human speech and the physiological parameters of the speaker. A correlation between heart-related parameters such as heart rate, electrocardiogram (ECG) features, and the influence of heart function on speech characteristics has been shown in [7–13]. In [14], the variation of speech characteristics due to physical activity is demonstrated, and the effect of physical activity and fitness level on heart rate is shown in [15]. Physiological changes due to physical activity also depend on the regularity, duration, and intensity of the activity performed [16]. A correlation between speech and blood pressure is established in [17], and a method to detect emotions from speech is described in [15]. It is also established in the literature that tiredness can also affect an individual's speech and can cause speech to become slurred (dysarthria) [18]. Noncontact methods to measure physiological parameters based on image and video processing have also been investigated and reported in the literature [19,20]. Noncontact methods based on speech, images, and video can facilitate remote monitoring, telemedicine, and smart healthcare which are expected to play a major role in future healthcare infrastructures.

The COVID-19 pandemic era has necessitated and triggered an enormous amount of research into such noncontact-based diagnostic methods to detect COVID-19 using machine learning and deep learning [21–30]. A review of COVID-19 diagnostic methods along with prevention tools and policy decisions for COVID-19 management is provided in [31]. Artificial intelligence-based COVID-19 diagnosis tools are not without pitfalls. A critique of AI-based tools being given emergency authorization by regulatory bodies indicates that many such tools have been developed using small or low-quality datasets [32], concluding that AI could be useful in dealing with the COVID-19 pandemic but requires more detailed investigation and validation. Several research studies based on artificial intelligence are ongoing not only to detect COVID-19 but also to predict and understand the effects of the pandemic and be prepared for eventualities. A method to detect COVID-19 using machine learning on symptoms is proposed in [33]. COVID-19 detection based on the application of AI on X-ray and computerized tomography (CT) images is reported in [34–41]. In [42–44], detection of COVID-19 by applying machine learning to routine blood examination data has also been reported. AI-based systems to predict the deterioration of COVID-19 patients toward severe disease have been presented in [45–47], and prediction of mortality risk among COVID-19 infected individuals using AI is also available in the literature [48–50]. The COVID-19 pandemic has indeed necessitated and highlighted the need for interdisciplinary and transdisciplinary approaches to diagnose, treat, and manage not just COVID-19 patients but also to address medical problems in general [51]. The challenges involved in the use of AI for COVID-19 are elaborated in [52]. Several research groups [53–56] are actively investigating the use of speech sounds, cough sounds, and breathing/respiratory sounds to detect COVID-19 by analyzing these sounds using artificial intelligence algorithms.

2. Materials and Methods

2.1. Data Used in This Study. Speech recordings used in this study comprise two categories—speech from healthy individuals with no known preexisting medical conditions at resting heart rate and speech from asymptomatic COVID-19 positive individuals. Heart rate and blood oxygen saturation level (SpO_2) are measured simultaneously at the time of recording the speech using an off-the-shelf pulse oximeter. It should be noted that the pulse rate measured by the pulse oximeter is exactly equal to the heart rate [57]. The total number of speech recordings of healthy individuals at resting heart rate is 84. All the healthy volunteers are in the age group of 25–45 years. Speech samples along with heart rate and SpO_2 measurements were also obtained from 22 individuals who had tested COVID-19 positive following contact tracing, but with no conspicuous symptoms. The youngest in this category is 32 years and the oldest is 57 years. It should be noted that obtaining speech data of COVID-19 patients was challenging and hence the relatively small set of samples.

Speech recording was made using a Logitech headphone equipped with a noise cancellation microphone. While the samples of all the healthy individuals were recorded using the same microphone in the same environment, the speech samples of COVID-19 individuals were recorded with different microphones of the same make and model (Logitech H540) and under different ambient conditions for each. Hence, any variations in speech characteristics arising due to the difference in recording device and ambiance are not taken into consideration. It is reasonable to ignore these variations since the recordings were made in quiet rooms using microphones of the same make and model and therefore have the same technical specifications. Of course, the acoustic effects of the room and background noise, albeit small, are not taken into consideration as it was not possible to bring the COVID-19 patients to the laboratory settings where the recording of healthy individuals was made. Each individual was asked to read the sentence "A quick brown fox jumped over the lazy dogs" which was recorded by turning the microphone "ON": for 5 seconds. The recording was made in stereo format at a sampling rate of 16000 samples per second (sps) which is the standard sampling rate for wideband representation of speech [58]. The recording is quantized using 2^{16} quantization levels resulting in an audio bit rate of 256 kbps and stored on a computer in uncompressed .WAV format. Heart rate and SpO_2 level are also concurrently measured at the time of speech recording using a pulse oximeter. The attributes of the data are highlighted in Table 1.

2.2. Preprocessing of Speech Data. Unwanted components such as DC bias, which usually gets introduced by PC audio cards [59] and silence intervals due to pauses made by the

Table 1: Attributes of data used in this study.

	Age group (years)	No. of recordings	Sampling rate (sps)	Quantization depth (bits)	Audio bit rate (kbps)	Audio format	Other parameters measured
Healthy	25–45	84	16000	16	256	.wav	Heart rate, SpO_2
COVID+	32–57	22					

speaker, are removed by preprocessing each of the speech recordings. DC bias is removed using a 1^{st}-order infinite impulse response (IIR) filter, whereas silence intervals are removed by applying a voice activity detection (VAD) mechanism which extracts speech frames containing voice activity. VAD also mitigates noise effects by applying a posteriori signal to noise ratio (SNR) weighting to emphasize reliable segments of voice activity even at low SNR. DC bias removal and VAD are applied as per the implementation provided in [60]. A block diagram of preprocessing steps is shown in Figure 1.

The features of speech that have been used in this study are the short-term Fourier Transform (STFT) coefficients. Features are defined as characteristics of a signal that enables some algorithm to detect an inherent pattern associated with the signal [61]. The premise of detecting COVID-19 from speech features stems from the fact that speech is produced by moving air from the lungs through the vocal cavity. Since there is an interaction between the lungs and heart for the oxygenation of the blood, cardiovascular responses are influenced by activities such as reading and speaking [62]. It is shown in [63] that breathing pattern is affected by the process of speech production. Changes in breathing patterns, in turn, have an effect on the heart rate, and this effect is termed respiratory sinus arrhythmia (RSA) [64]. Several techniques for feature extraction have been proposed in the literature for various applications but predominantly for speaker/speech recognition and speech enhancement. Linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), perceptual linear prediction (PLP), Mel frequency cepstral coefficients (MFCC), Mel frequency discrete wavelet coefficients (MFDWC), feature extraction using principal component analysis (PCA), and wavelets based features are some of the common features that have been reported in the literature [65–70]. STFT represents the time-varying spectral properties of a signal, and for this study, STFT coefficients with a high spectral resolution are used in order to capture subtle differences between closely spaced frequency components. The high spectral resolution is achieved by computing the STFT of long segments of speech, i.e., segment length greater than 500 ms. The high spectral resolution is achieved at the expense of temporal resolution. Since a correlation between physiological parameters and spectral features of speech is evident from existing literature, STFT coefficients with high spectral resolution have been used in this study. The STFT coefficients are used as input features to machine learning algorithms to classify the speech signal as that of COVID-19 positive or not.

2.3. Statistical Modeling of Speech Features. The most common symptoms of COVID-19 are fever, tiredness, and dry cough, and these may not be conspicuous until about 14 days after getting infected with an average of 5-6 days for the symptoms to become conspicuous. In this article, it is shown using statistical modeling of speech features that it is possible to detect COVID-19 from an individual's speech much before the symptoms become conspicuous so that the person can be quarantined, tested, and provided with medical support at an early stage. At their onset, while symptoms may not be conspicuous to the affected individual or to observers, physiological changes occur in the individual that cause variations in speech characteristics which can be analyzed by artificial intelligence (AI) algorithms. Signal processing and AI can be applied to speech to detect physiological changes which have a direct or indirect relation to one or more of the COVID-19 symptoms. The existence of a correlation between speech characteristics and physiological, psychological, and emotional conditions is well established in the literature. It is therefore possible to detect COVID-19 infection from speech samples of individuals and this possibility is investigated in this article. The relationship between the most common symptoms of COVID-19 and affected physiological parameters is illustrated in Figure 2.

A statistical analysis of speech spectral features is performed by applying maximum likelihood estimation (MLE) to obtain the best statistical distribution along with the distribution parameters that best characterize speech STFT coefficients, statistically. It has been shown in [58] that for speech samples at resting heart rate, STFT coefficients having high spectral resolution are accurately modeled by a Laplacian distribution (LD) with the estimated LD parameters exhibiting small RMS error. The Laplacian distribution is defined as

$$\mathbf{p}(\mathbf{x}) = \frac{1}{2\mathbf{b}}\mathbf{exp}\left(\frac{-|\mathbf{x}-\boldsymbol{\mu}|}{\mathbf{b}}\right) \tag{1}$$

parameter and $b > 0$ is the scale parameter. The procedure to estimate μ and b, the RMS error associated with the estimation of b, and its lower bound defined as the Cramer-Rao bound (CRB) are provided in [58].

2.4. Binary Classification of Speech Samples Using Machine Learning. From the statistical analysis of speech STFT coefficients at the high spectral resolution, it is evident that the RMS error of fitted LD increases as a result of COVID-19 infection. Based on this finding, binary classification of speech signals as COVID-19 positive or COVID-19 negative is investigated by using STFT coefficients as input features to machine learning algorithms. In order to train and develop the AI models, speech samples of healthy as well as COVID-19 positive individuals are used. The trained AI model can

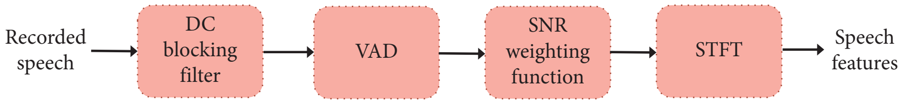


Figure 1: Speech preprocessing.

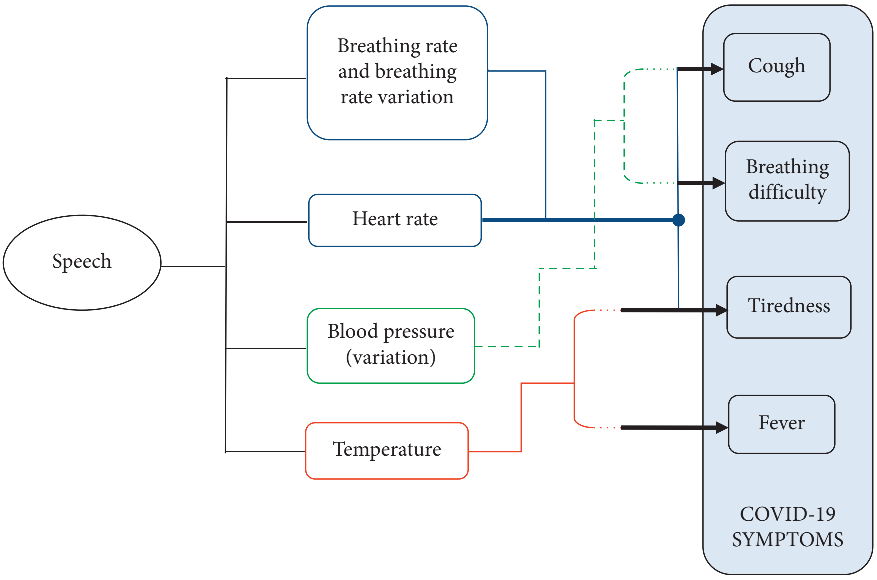


Figure 2: Biological parameters correlated to speech and COVID-19 symptoms.

then be incorporated into a mobile "app" for early detection of COVID-19, once the desired level of accuracy is achieved and regulatory approvals are obtained. If speech can be used to detect COVID-19, the functionality of the "smartphone" which already has wide proliferation and ubiquitous presence can be extended to alleviate the challenges posed by the pandemic. The results reported in the literature [71–78] are quite promising, providing exciting and interesting answers, giving confidence that research on this topic can lead to the development of mobile applications which can be used not only to detect COVID-19 from human sounds but also for other medical diagnostic/monitoring purposes. COVID-19 diagnosis using only cough recordings is presented in [71]. However, it uses biomarker information such as muscular degradation, vocal cords, sentiment, and lungs/respiratory tract function along with the cough recordings for diagnosis. The relation between COVID-19 symptoms and respiratory system function is highlighted in [72] along with a survey of AI-based COVID-19 diagnoses using human audio signals. Cough and respiratory sounds are used to classify COVID-19 and non-COVID-19 individuals in [73]. Furthermore, it is shown that cough from COVID-19 can be distinguished from healthy individuals' cough as well as cough of asthmatic patients. A project in progress [74] investigates the detection of COVID-19 from human audio sounds using AI. A news feature article in Nature [75] highlights research interest and progress among academic as well as commercial organizations to use the human voice for various diagnostic purposes including COVID-19. AI4COVID-19 is an app that runs an AI algorithm in the cloud to detect COVID-19 from cough sounds and reports promising results, encouraging further collection of labeled cough sounds [76]. An overview of the possibilities, challenges, and use cases of computer audition is presented in [77], which clearly highlights the potential of using sound analysis using AI for COVID-19 diagnosis. A support vector machine (SVM) based method to detect COVID-19 from speech signals is presented in [78] which combines voice signals and symptoms reported by the patient. In contrast to the research available in the literature, this article uses only speech signals without any side information such as symptoms or other biomarker information.

The STFT coefficients are labeled as coming from the speech of COVID-19 negative, i.e., healthy (Class 0) and COVID-19 positive (Class 1) individuals. Microsoft Azure Machine Learning Studio (MAMLS) cloud platform is used in this study to perform binary classification, and the performance of classification is analyzed and compared for five state-of-the-art classification algorithms available in MAMLS. Machine learning techniques produce a model for the data by learning the statistical relationship between input data (e.g., STFT coefficients extracted from speech signals) and output data (e.g., class label). The hyperparameters of the produced model are tuned optimally to minimize the classification error in an independent test dataset, resulting in a generalized model that can perform well on the test data set as well. The tuning is performed manually by adjusting the model hyperparameters until the highest value for the "recall" metric is achieved. Good performance on only the training dataset would result in an overfitting solution. A brief description of the five algorithms used for binary classification is provided here for completeness. The block diagram of the methodology used in the work is shown in Figure 3. An overview of the used ML algorithms follows.

Boosted Decision Tree (BDT) is an ensemble learning technique, wherein the succeeding tree corrects the errors of the previous tree to minimize classification error. The complete ensemble of trees is used for correctly predicting the binary class to which the input data belongs [79]. Another classification algorithm based on ensemble learning is the Decision Forest (DF) algorithm, wherein the most popular class is selected depending on the vote from each of the generated trees [80]. Neural Networks (NNs) are a network of interconnected layers of processing units called neurons. A typical NN consists of neurons aggregated into three layers. The first layer is formed by the input feature set which is linked to the output layer via an interconnection of several hidden layers in the middle. Each neuron processes its input variables and its output is passed to the neuron in the subsequent layer [81]. Logistic Regression (LoR) is a statistical technique for analyzing data when a dichotomous outcome is determined by one or more independent variables [82]. Support Vector Machines (SVMs) are based on the principle of recognizing patterns in a multidimensional hyperplane to estimate the maximum margin between samples of binary classes using a multidimensional input feature space [83].

These algorithms have relatively fast training and good performance and are robust to overfitting and have therefore been chosen in this study. The performance of classification models based on each of these algorithms is evaluated using the evaluation metrics listed in Table 2. These evaluation metrics are standard in machine learning literature [84].

The input features used for binary classification are the STFT coefficients of speech from each of the 84 healthy individuals and the 22 COVID-19 positive individuals. Each individual's speech sample comprises 8 segments and STFT coefficients obtained from each speech segment are used as input features for binary classification. As mentioned in Section 4, STFT coefficients are complex numbers; hence, each speech segment comprises "real" and "imaginary" parts of STFT coefficients. The number of frequency points used in the computation of STFT coefficients is 8192, which is obtained as the next power of 2 greater than the segment length. Thus, for each individual speech sample, a matrix of 8192 rows × 8 columns is generated. The real and imaginary parts of the complex STFT coefficients are separated resulting in two separate matrices having dimensions of 8192 rows × 8 columns each. Class label is assigned to each row of these matrices as "Class 0" for healthy individuals' speech samples and "Class 1" for COVID-19 positive individuals. Thus, there are $8192 \times 84 = 688{,}128$ rows of STFT coefficients (real part) labeled as Class 0 and $8192 \times 22 = 180{,}224$ rows of STFT coefficients (real part) labeled as Class 1. Correspondingly, an equal number of rows are available under each class label containing the "imaginary part" of STFT coefficients.

Each row of the real/imaginary part of the STFT coefficients matrix corresponds to a frequency point in the STFT computation, and each column represents a segment of speech. The rows are treated as examples and columns as features since each column of the STFT matrix represents the time-localized spectral features of the speech signals. Every real and imaginary part "x" of STFT coefficients is normalized to lie in the interval [0,1] using a MinMax normalizer as follows:

$$\text{Normalized value} = \frac{x - \min(x)}{[\max(x) - \min(x)]}. \tag{2}$$

Since the statistical distribution for both the real and imaginary parts of speech STFT coefficients is Laplacian, these are treated together without distinction in the context of this work. Thus, the entire dataset comprises 1,736,704 labeled rows, half of which comprise the real part of STFT coefficients and the other half comprise the imaginary part of STFT coefficients, which is saved in .csv format. For binary classification using machine learning, only the rows corresponding to the real part of STFT coefficients are utilized to reduce the time taken for training and cross-validation. The data is split in an 80 : 20 ratio; i.e., 80% of the rows are used for training and the remaining 20% are used for testing. Since the dataset used in this study is highly imbalanced—Class 0 constitutes nearly 80% of the dataset and Class 1 constitutes a little over 20% of the dataset—data splitting is performed with "stratification." Stratification ensures that each subset of split data has the same class distribution as the entire dataset. The ratio of healthy: COVID-19 + samples in terms of speech recordings is $84 : 22 = 3.8181 : 1$. In terms of the STFT coefficients also, this ratio remains the same. Since only the real part of STFT coefficients has been used for binary classification, the ratio healthy: COVID-19 + samples in terms of STFT coefficients is $688128{:}180224 = 3.8181 : 1$. Since stratification has been used, both training and testing data contain STFT coefficients of "healthy" and "COVID+" individuals in the same proportion as $84 : 22 = 3.8181 : 1$. Furthermore, the train-test split with stratification at the STFT level ensures permutation of the labeled STFT coefficients across all "individuals"—Class 0 as well as Class 1. Even though the number of speech samples used in this study is small, the number of frequency points (rows) of STFT is large due to the high spectral resolution adopted. The performance evaluation metrics are computed following a 10-fold cross-validation process. Furthermore, as in the data splitting process, stratification is used in the cross-validation process as well to ensure that the class distribution of the training data set is maintained in each fold of cross-validation.

3. Results

3.1. Classification of COVID-19 Samples Based on Statistical Distribution Fitting. LD fitting based on MLE is applied to all the speech samples used in this study—healthy individuals without COVID-19 as well as COVID-19 positive individuals. A comparison of statistical properties of speech STFT coefficients of the two categories of speech samples is performed. The statistical distribution of speech STFT coefficients of healthy individuals, i.e., not

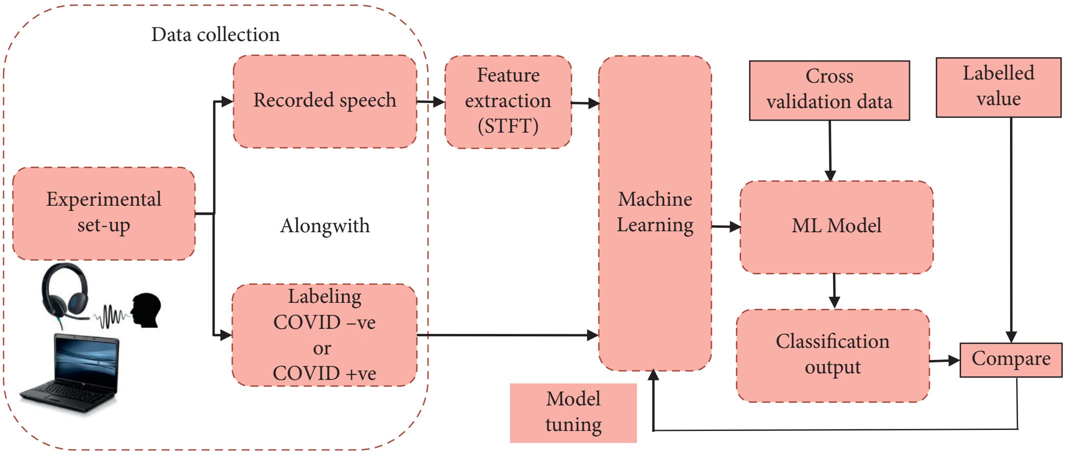


Figure 3: Block diagram of the methodology used.

Table 2: Evaluation metrics for binary classification.

Evaluation metric	Definition	Notations
	Binary classification	
Precision (PRE)	PRE = $t_p/t_p + f_p$	tp–Total no. of true positive samples
Recall (REC)	REC = $t_p/t_p + f_n$	tn–Total no. of true negative samples
Accuracy (ACC)	ACC = $t_p + t_n/t_p + t_n + f_p + f_n$	fp–Total no. of false positive samples
F1-score	$F1 = 2 * P_{RE} * R_{EC}/P_{RE} + R_{EC}$	fn – Total no. of false negative samples
Area under RoC curve	(AUC) AUC = $\int_0^1$ RoC	RoC - receiver operating characteristic curve

infected by COVID-19, is shown in Figure 4 and that of a COVID-19 positive individual is shown in Figure 5.

It is found from Figure 4 that spectral features of the speech of healthy individuals are accurately modeled by LD, with small RMS error as has been established in the literature [58]. In the case of speech samples of COVID-19 positive individuals without any conspicuous symptoms, while the statistical distribution of the STFT coefficients is still closely modeled by LD, the RMS error of the fitted distribution has a nearly 10-fold increase as compared to non-COVID-19 individuals. This increase in the RMS error of the fitted distribution indicates a variation of speech characteristics as a result of COVID-19 infection. The PDFs are obtained by plotting the envelope of histograms of STFT coefficients. The "estimated" PDF represents the fitted distribution based on estimated Laplacian distribution parameters "μ" and "b" and the "actual" PDF is the actual distribution of the STFT coefficients. The Laplacian distribution is therefore a suitable distribution for speech STFT coefficients as the RMS error between the actual and fitted distributions is small.

3.2. Performance Evaluation of Binary Classification. The model hyperparameters for each algorithm are tuned and optimized to achieve the best performance in binary classification in terms of the "Recall" metric. The performance evaluation metrics for the five binary classification algorithms used in this study, along with their optimal parameterization, are listed in Table 3. The best performance in terms of precision, recall, accuracy, and F1 score is achieved for the DF algorithm. For the classification application considered in this work, classifying speech spectral features as COVID-19 positive or not, the cost associated with misclassification is very high for "false negative" as compared to "false positive."

Since "recall" provides a measure of correctly predicted positives against the total number of positive

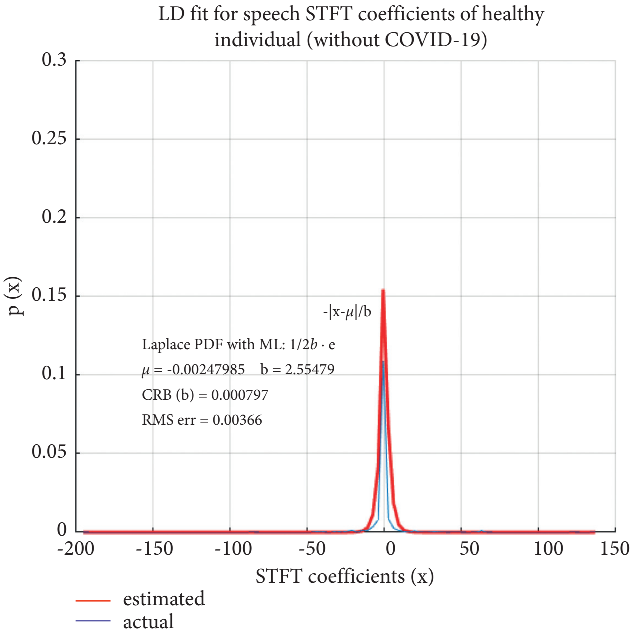


Figure 4: Statistical properties of speech STFT coefficients of a healthy person (without COVID-19).

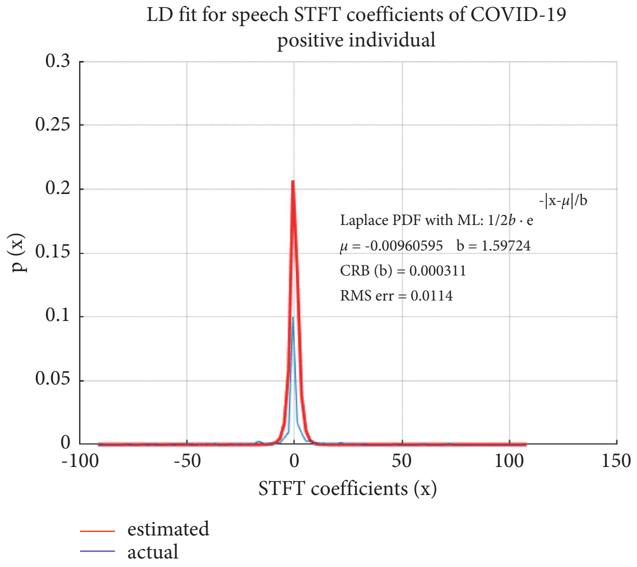


Figure 5: Statistical properties of speech STFT coefficients of an infected person (with COVID-19).

examples, it is important for our classification problem to have a high value for this metric. This will minimize misclassifying a COVID-19 positive example as not being COVID-19 positive. While other evaluation metrics have also been determined, "recall" is the more important metric in the context of this work. In Table 3, the values within brackets are the standard deviations of the metrics. Small values for standard deviation indicate that the models are verified with an unbiased dataset which has been achieved by using stratification in the train-test split as well as in cross-validation.

4. Discussion

4.1. Statistical Distribution of COVID-19 Positive and COVID-19 Negative Speech Samples. The average RMS error for the fitted LD averaged over all the speech samples belonging to each of the two categories is shown in Table 4. This increase in RMS error of the fitted LD in COVID-19 positive samples is attributed to the physiological changes associated with COVID-19 infection which affect the characteristics of speech.

Since the symptoms are not conspicuous among the samples used in this study, the distribution of STFT coefficients of speech is still Laplacian, albeit with a higher RMS error. STFT coefficients being complex numbers, the above findings are valid for both the "real" as well as "imaginary" parts of STFT coefficients and the same has also been shown in [58]. It remains to be seen if the distribution deviates significantly from being Laplacian or even ceases to be Laplacian when the symptoms become more pronounced and conspicuous. The increase in RMS error indicates such a trend. It should be noted that the data used in this study is unbalanced—the dataset from COVID-19 positive individuals is smaller than that of healthy individuals. The results discussed in this section clearly indicate that the statistical properties of speech spectral features are altered as a result of COVID-19 infection. However, as it was not possible to obtain samples of COVID-19 positive individuals whose symptoms are more pronounced and conspicuous, this shall be a subject of future investigation, once such samples are obtained. Due to the prevailing COVID-19 restrictions, access to such individuals has not been possible.

4.2. COVID-19 Detection on Test Data Using the Binary Classification Models. Finally, the optimally parameterized classification algorithms discussed in Section 2.4 have been tested on the test dataset. As discussed in Section 3.2, the algorithms have been parameterized to optimize the "recall" performance metric. The classification results of 20 samples (rows) from the test data are shown in Table 5 which contains the "actual" and "predicted" classes for the 20 test samples by each of the five classification algorithms. It can be observed from Table 5 that the misclassification of COVID-19 positive as "not positive", i.e., class 1 being misclassified as Class 0, is the lowest for the DF algorithm. The misclassified values are highlighted in bold and underlined.

For future investigation, concurrently at the time of recording speech samples of individuals, biomedical parameters such as heart rate (pulse oximeter), oxygen saturation (pulse oximeter), blood pressure (digital BP monitor), and temperature (infrared thermometer) have also been measured. These shall be used for future research to develop machine learning-based regression algorithms to predict these biomedical parameters from speech signals. The variations of these parameters among COVID-19 negative and COVID-19 positive individuals shall be analyzed to improve the accuracy of detecting COVID-19 from speech samples. The devices used to measure the biomedical parameters are shown in Figure 6. The e-health sensor platform shown in Figure 6 facilitates direct recording of the biomedical parameter to a PC, thus avoiding the manual entry of data.

A limitation of the work presented in this article is that it cannot distinguish between similar symptoms which may

Table 3: Performance metrics for binary classification algorithms.

Classification algorithms	Optimal Parameterization	Performance metrics Mean value (standard deviation)				
		A_{CC}	P_{RE}	R_{EC}	F1 score	AUC
BDT	No. of Leaves: 16 Learning rate: 0.05 No. of trees: 100	0.724 (0.048)	0.714 (0.037)	0.7037 (0.063)	0.7088 (0.052)	0.717 (0.053)
DF	Random split Count: 128 Maximum Depth: 32 No. of decision trees: 16	0.7317 (0.021)	0.7421 (0.017)	0.7892 (0.081)	0.7649 (0.025)	0.755 (0.017)
NN	Learning rate: 0.001 No. of hidden Nodes: 314	0.711 (0.031)	0.7271 (0.043)	0.7188 (0.018)	0.7229 (0.029)	0.7616 (0.095)
LoR	Optimization Tolerance: 1e-06 L1 regularization weight: 1 Memory size for L-BFGS: 18	0.6741 (0.019)	0.6805 (0.024)	0.6161 (0.027)	0.6467 (0.019)	0.6874 (0.065)
SVM	Lambda – 0.001	0.694 (0.017)	0.673 (0.074)	0.6027 (0.019)	0.6359 (0.011)	0.6619 (0.037)

Table 4: Average RMS error of the fitted LD distributions.

Category	Average RMS error of fitted LD
Without COVID-19	0.00354
With COVID-19	0.01271

Table 5: Test evaluation for binary classification

Test sample	Actual class	Predicted class				
		BDT	DF	NN	LoR	SVM
1	0	0	1	0	0	1
2	1	1	1	**0**	1	1
3	1	1	1	1	**0**	**0**
4	0	1	0	1	1	1
5	1	1	1	**0**	1	1
6	0	0	0	0	0	0
7	1	**0**	**0**	1	1	1
8	1	1	1	1	**0**	**0**
9	1	**0**	**0**	**0**	1	1
10	0	1	0	0	1	1
11	1	**0**	1	1	1	1
12	0	0	1	1	0	0
13	0	1	0	1	1	0
14	1	1	1	1	**0**	1
15	0	0	0	0	0	1
16	0	1	1	0	0	1
17	1	1	1	**0**	**0**	**0**
18	1	**0**	1	1	1	1
19	0	0	0	1	0	1
20	1	1	1	1	0	1

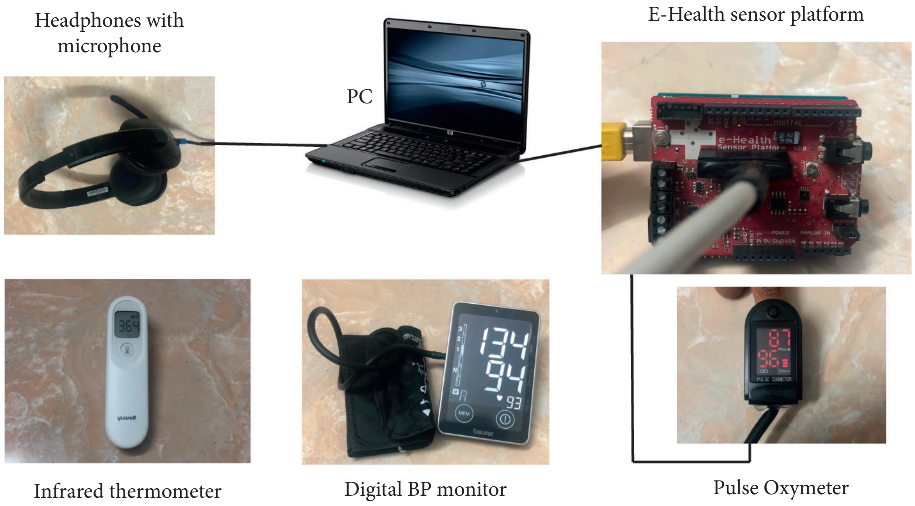


FIGURE 6: Devices used for measuring biomedical parameters along with speech for future analysis.

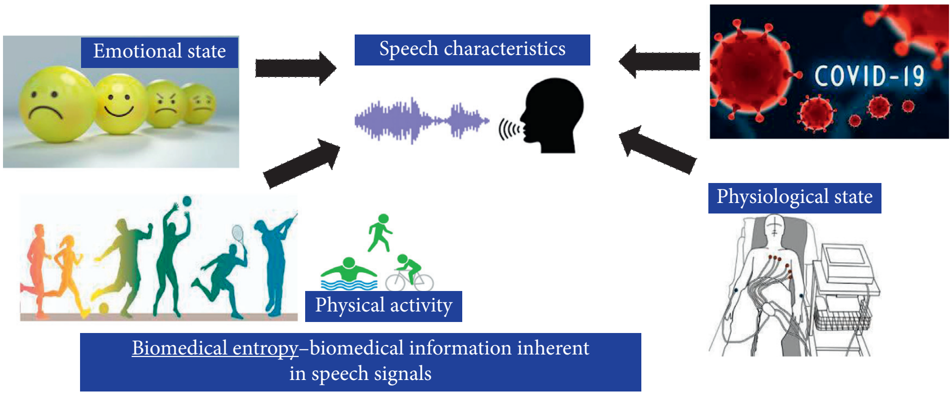


FIGURE 7: Possible applications of the proposed research.

appear due to multiple different causes such as influenza or myocarditis. It requires further research involving speech data from patients with various illnesses that have symptoms similar to COVID-19. This shall be a subject of future work.

5. Conclusions

This article investigates the statistical properties of speech spectral features for samples taken from healthy as well as asymptomatic COVID-19 positive individuals. While the statistical distribution for both is Laplacian, the RMS error of the fitted Laplace distribution is higher in the case of asymptomatic COVID-19 positive speech samples. This indicates that spectral properties of speech get altered as a result of physiological changes caused due to COVID-19 infection. It is therefore deduced that there is an associated entropy in speech which can be used to detect COVID-19. STFT coefficients of speech are then used as input features of machine learning-based classification algorithms and the classification performance of five state-of-the-art classification algorithms has been evaluated. All the five classification algorithms exhibit a moderate level of performance having their evaluation metrics values around 70% of their maximum values. The best performance is observed for the DF algorithm which has the highest value for the "recall" metric with a value of 0.7892. "Recall" is the metric used while training the model hyperparameters as a higher recall value means minimizing misclassification of the "false negative" category. The cost of misclassifying a COVID-19 positive sample as a COVID-19 negative is high and hence the choice of recall as the evaluation metric is to be maximized while tuning the model parameters. It is also noted from Table 5 that the misclassification of Class 1 (COVID-19 positive) as Class 0 (COVID-19 negative) is least for the DF algorithm when tested on previously unseen test data. The results obtained are promising and provide evidence that COVID-19 infection can be detected from speech signals of individuals.

Speech can be used as a biomedical signal to diagnose various physical and emotional disorders. It can be used to monitor the performance/health conditions of individuals while performing physical activity. The focus of this work, however, is to detect COVID-19 infection by analyzing a person's speech signal. This is possible because speech characteristics of individuals get altered by these conditions as depicted in Figure 7.

The results presented are concurring with similar approaches available in published literature. For example, 100% sensitivity is reported in [71] for asymptomatic cases, but it uses additional biomarker information along with cough sounds. A maximum "recall" value of 0.72 is reported in [73,78] while, in [76], the highest accuracy of 92.85% is reported for binary classification using deep transfer learning.

The results presented in this work can be improved by using a larger dataset comprising different classes of human vocal sounds which should also include samples of individuals of different languages, dialects, and other health conditions. It was intended to collect large datasets by encouraging community participation but that could not be achieved due to regulatory procedures and limitations of funding. Hence, the results presented in this work are based on a small dataset but the findings are encouraging. Future work shall consider using the magnitude of STFT coefficients rather than just the real/imaginary part and also consider the use of other types of audio signal features such as MFCC as input features for ML-based classification. The detection of COVID-19 using speech can facilitate real-time, remote monitoring of infected yet asymptomatic individuals. This will allow early detection of COVID-19 symptoms and help manage the ongoing COVID-19 situation better. It should be noted that the fundamental idea presented in this article is not limited to detecting COVID-19 symptoms only but has broader applications in medical diagnosis and patient monitoring/care. AI can detect changes in human vocal sounds not discernible to the human ear. Smartphone apps that use AI algorithms to analyze human vocal sounds for diagnosis, screening, and monitoring can be extremely useful and are expected to play a vital role in future healthcare technologies.

References

[1] Vital Signs (Body Temperature, Pulse Rate, Respiration Rate, Blood Pressure) [Internet], "Johns Hopkins Medicine," 2021, https://www.hopkinsmedicine.org/health/conditions-and-diseases/vital-signs-body-temperature-pulse-rate-respiration-rate-blood-pressure.

[2] A. Reynolds and A. Paivio, "Cognitive and emotional determinants of speech," *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 22, no. 3, pp. 164–175, 1968.

[3] L. A. Ramig, "Effects of physiological aging on vowel spectral noise," *Journal of Gerontology*, vol. 38, no. 2, pp. 223–225, 1983.

[4] J. Trouvain and K. P. Truong, *Prosodic Characteristics of Read Speech before and after Treadmill Running*, International Speech Communication Association (ISCA), Baixas, France, 2015.

[5] T. D. Borkovec, R. L. Wall, and N. M. Stone, "False Physiological Feedback and the maintenance of speech anxiety," *Journal of Abnormal Psychology*, vol. 83, no. 2, pp. 164–168, 1974.

[6] ScienceEncyclopedia, "Speech - the Physiology of Speech - Air, Vocal, Words, and Sound - JRank Articles [Internet]," 2019, https://science.jrank.org/pages/6371/Speech-physiology-speech.html.

[7] D. Skopin and S. Baglikov, "Heartbeat feature extraction from vowel speech signal using 2D spectrum representation," in *Proceedings of the 4th Inernational Conference Information Technology*, p. 6, Doha, Qatarp, June 2009.

[8] B. Schuller, F. Friedmann, and F. Eyben, "Automatic recognition of physiological parameters in the human voice: heart rate and skin conductance," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7219–7223, IEEE, Vancouver, Canada, Vancover 2013.

[9] A. Mesleh, D. Skopin, S. Baglikov, and A. Quteishat, "Heart rate extraction from vowel speech signals," *Journal of Computer Science and Technology*, vol. 27, no. 6, pp. 1243–1251, 2012.

[10] J. Kaur and R. Kaur, "Extraction of heart rate parameters using speech analysis," *International Journal of Science and Research*, vol. 3, no. 10, pp. 1374–1376, 2014.

[11] M. Sakai, "Modeling the relationship between heart rate and features of vocal frequency," *International Journal of Computer Application*, vol. 120, no. 6, pp. 32–37, 2015.

[12] R. F. Orlikoff and R. J. Baken, "The effect of the heartbeat on vocal fundamental frequency perturbation," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 3, pp. 576–582, 1989.

[13] B. Schuller, F. Friedmann, and F. Eyben, "The munich biovoice corpus: effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 1506–1510, Reykjavik: European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014.

[14] M. Usman, "On the performance degradation of speaker recognition system due to variation in speech characteristics caused by physiological changes," *International Journal of Computing and Digital Systemss*, vol. 6, no. 3, pp. 119–127, 2017.

[15] A. P. James, "Heart rate monitoring using human speech spectral features," *Human-centric Computing and Information Sciences*, vol. 5, no. 1, pp. 1–12, 2015.

[16] D. A. Burton, K. Stokes, and G. M. Hall, "Physiological effects of exercise," *Continuing Education in Anaesthesia, Critical Care & Pain*, vol. 4, no. 6, pp. 185–188, 2004.

[17] M. Sakai, "Feasibility study on blood pressure estimations from voice spectrum analysis," *International Journal of Computer Application*, vol. 109, no. 7, pp. 39–43, 2015.

[18] D. Griswold and M. Abraham, "Slurred Speech From Anxiety: Causes and Treatments [Internet]. CalmClinic," 2021, https://www.calmclinic.com/anxiety/symptoms/slurred-speech.

[19] C. G. Scully, J. Jinseok Lee, J. Meyer et al., "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 303–306, 2012.

[20] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[21] Y. Mardian, H. Kosasih, M. Karyana, A. Neal, and C.-Y. Lau, "Review of current COVID-19 diagnostics and opportunities for further development," *Frontiers of Medicine*, vol. 8, 2021.

[22] N. M.-H. Tayarani, "Applications of artificial intelligence in battling against covid-19: a literature review," *Chaos, Solitons & Fractals*, vol. 142, Article ID 110338, 2021.

[23] A. S. Albahri, R. A. Hamid, J. k. Alwan et al., "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review," *Journal of Medical Systems*, vol. 44, no. 7, p. 122, 2020.

[24] N. Alballa and I. Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review," *Informatics in Medicine Unlocked*, vol. 24, Article ID 100564, 2021.

[25] J. Bullock, A. Luccioni, K. H. Pham, C. Sin Nga Lam, and M. Luengo-Oroz, "Mapping the landscape of artificial intelligence applications against COVID-19," *Journal of Artificial Intelligence Research*, vol. 69, pp. 807–845, 2020.

[26] I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham, "Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5330, 2020.

[27] T. Aishwarya and V. Ravi Kumar, "Machine learning and deep learning approaches to analyze and detect COVID-19: a review," *SN Computer Science*, vol. 2, no. 3, p. 226, 2021.

[28] H. Swapnarekha, H. S. Behera, J. Nayak, and B. Naik, "Role of intelligent computing in COVID-19 prognosis: a state-of-the-art review," *Chaos, Solitons & Fractals*, vol. 138, Article ID 109947, 2020.

[29] L. Wynants, B. Van Calster, G. S. Collins et al., "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, Article ID m1328, 2020.

[30] W. T. Li, J. Ma, N. Shende et al., "Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 247, 2020.

[31] M. Allam, S. Cai, S. Ganesh et al., "COVID-19 diagnostics, tools, and prevention," *Diagnostics*, vol. 10, no. 6, p. 409, 2020.

[32] The Lancet Digital Health, "Artificial intelligence for COVID-19: saviour or saboteur?" *Lancet Digit Heal*, vol. 3, p. e1, 2021.

[33] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *Npj Digital Medicine*, vol. 4, no. 1, p. 3, 2021.

[34] M. Irfan, M. A. Iftikhar, S. Yasin et al., "Role of hybrid deep neural networks (HDNNs), computed tomography, and chest X-rays for the detection of COVID-19," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3056, 2021.

[35] S. Masoud Rezaeijo, M. Ghorvei, and M. Alaei, "A machine learning method based on lesion segmentation for quantitative analysis of CT radiomics to detect COVID-19," in *Proceedings of the 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 1–5, IEEE, Mashhad, Iran, December 2020.

[36] S. M. Rezaeijo, R. Abedi-Firouzjah, M. Ghorvei, and S. Sarnameh, "Screening of COVID-19 based on the extracted radiomics features from chest CT images," *Journal of X-Ray Science and Technology*, vol. 29, no. 2, pp. 229–243, 2021.

[37] A. Borkowski, "Using artificial intelligence for COVID-19 chest X-ray diagnosis," *Federal Practitioner*, vol. 37, no. 9, pp. 398–404, 2020.

[38] L. Brunese, F. Martinelli, F. Mercaldo, and A. Santone, "Machine learning for coronavirus covid-19 detection from chest x-rays," *Procedia Computer Science*, vol. 176, pp. 2212–2221, 2020.

[39] Y. Karbhari, A. Basu, Z. W. Geem, G.-T. Han, and R. Sarkar, "Generation of synthetic chest X-ray images and detection of COVID-19: a deep learning based approach," *Diagnostics*, vol. 11, no. 5, p. 895, 2021.

[40] Y. Qiblawey, A. Tahir, M. E. H. Chowdhury et al., "Detection and severity classification of COVID-19 in CT images using deep learning," *Diagnostics*, vol. 11, no. 5, p. 893, 2021.

[41] S. Chattopadhyay, A. Dey, P. K. Singh, Z. W. Geem, and R. Sarkar, "COVID-19 detection by optimizing deep residual features with improved clustering-based golden ratio optimizer," *Diagnostics*, vol. 11, no. 2, p. 315, 2021.

[42] M. Kukar, G. Gunčar, T. Vovko et al., "COVID-19 Diagnosis by Routine Blood Tests Using Machine Learning," *Scientific Reports*, vol. 11, Article ID 10738, 2020.

[43] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study," *Journal of Medical Systems*, vol. 44, no. 8, p. 135, 2020.

[44] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, "Ensemble learning model for diagnosing COVID-19 from routine blood tests," *Informatics in Medicine Unlocked*, vol. 21, Article ID 100449, 2020.

[45] F. E. Shamout, Y. Shen, N. Wu et al., "An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department," *Npj Digital Medicine*, vol. 4, no. 1, p. 80, 2021.

[46] Z. Dai, D. Zeng, D. Cui et al., "Prediction of COVID-19 patients at high risk of progression to severe disease," *Frontiers in Public Health*, vol. 8, Article ID 574915, 2020.

[47] M. Carlile, B. Hurt, A. Hsiao, M. Hogarth, C. A. Longhurst, and C. Dameff, "Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department," *Journal of the American College of Emergency Physicians Open*, vol. 1, no. 6, pp. 1459–1464, 2020.

[48] C. Hu, Z. Liu, Y. Jiang et al., "Early prediction of mortality risk among patients with severe COVID-19, using machine learning," *International Journal of Epidemiology*, vol. 49, no. 6, pp. 1918–1929, 2021.

[49] A. López-Escobar, R. Madurga, J. M. Castellano et al., "Risk score for predicting in-hospital mortality in COVID-19 (RIM score)," *Diagnostics*, vol. 11, no. 4, p. 596, 2021.

[50] Y. Gao, G.-Y. Cai, W. Fang et al., "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nature Communications*, vol. 11, no. 1, p. 5033, 2020.

[51] F. Tretter, O. Wolkenhauer, M. Meyer-Hermann et al., "The quest for system-theoretical medicine in the COVID-19 era," *Frontiers of Medicine*, vol. 8, 2021.

[52] W. Naudé, "Artificial intelligence vs COVID-19: limitations, constraints and pitfalls," *AI & Society*, vol. 35, no. 3, pp. 761–765, 2020.

[53] M. W. Tobias, "AI and Medical Diagnostics: Can A Smartphone App Detect Covid-19 from Speech or A Cough? Forbes," 2020.

[54] J. Chu, *Artificial Intelligence Model Detects Asymptomatic Covid-19 Infections through Cellphone-Recorded Coughs*, MIT News, Cambridge, MA, USA, 2020.

[55] M. Scudellari, *AI Recognizes COVID-19 in the Sound of a Cough*, IEEE Spectrum - The Institute, USA, 2020.

[56] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," *BMJ Innovations*, vol. 7, no. 2, pp. 356–362, 2021.

[57] M. MacGill, "Heart rate: What is a normal heart rate? [Internet]," 2017, https://www.medicalnewstoday.com/articles/235710.php.

[58] M. Usman, M. Zubair, M. Shiblee, P. Rodrigues, and S. Jaffar, "Probabilistic modeling of speech in spectral domain using maximum likelihood estimation," *Symmetry*, vol. 10, no. 12, p. 750, 2018.

[59] P. Partila, M. Voznak, M. Mikulec, and J. Zdralek, "Fundamental frequency extraction method using central clipping and its importance for the classification of emotional state," *Advances in Electrical and Electronic Engineering*, vol. 10, no. 4, pp. 270–275, 2012.

[60] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.

[61] J. J. Wolf, "Speech signal processing and feature extraction," in *Spoken Language Generation and Understanding*Dordrecht: Springer, Netherlands, Europe, 1980.

[62] K. J. Reilly and C. A. Moore, "Respiratory sinus arrhythmia during speech production," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 1, pp. 164–177, 2003.

[63] C. Von Euler, "Speech motor control," in *Proceedings of the International Symposium on Speech Motor Control, Wenner -Gren Center International Symposium Series*, pp. 95–103, Stockholm, Sweden, May 1981, https://www.sciencedirect.com/science/article/pii/B978008028892550013X.

[64] F. Yasuma and J.-i. Hayano, "Respiratory sinus arrhythmia," *Chest*, vol. 125, no. 2, pp. 683–690, 2004.

[65] S. B. Magre, R. R. Deshmukh, and P. P. Shrishrimal, "A comparative study on feature extraction techniques in speech recognition," in *Proceedings of the International Conference on Recent Advances in Statistics and Their Applications*, Aurangabad, Maharashtra, India, December 2013.

[66] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, Hoboken, NJ, USA, 2001.

[67] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[68] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[69] Z. Tufekci and J. N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in *Proceedings of the IEEE SoutheastCon 2000 "Preparing for the New Millennium" (Cat No00CH37105)*, pp. 116–123, IEEE, Nashville, TN, USA, April 2000.

[70] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[71] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[72] G. Deshpande and B. W. Schuller, "Audio, Speech, Language, & Signal Processing for COVID-19: A Comprehensive Overview," *Pattern Recognit*, vol. 122, Article ID 108289, 2020.

[73] C. Brown, J. Chauhan, A. Grammenos et al., "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3474–3484, ACM, New York, NY, USA, August 2020.

[74] I. Trancoso, "Project to Detect COVID-19 from Coughs and Speech," 2021, https://www.inesc-id.pt/project-to-detect-covid-19-from-coughs-and-speech/.

[75] E. Anthes, "Alexa, do I have COVID-19?" *Nature*, vol. 586, no. 7827, pp. 22–25, 2020.

[76] A. Imran, I. Posokhova, H. N. Qureshi et al., "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, Article ID 100378, 2020.

[77] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "COVID-19 and Computer Audition: An Overview on what Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis," *Digit. Health*, vol. 3, 2021.

[78] J. Han, C. Brown, J. Chauhan et al., "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," in *Proceedings of the IEEE ICASSP 2021*, Toronto, Canada, June 2021.

[79] P. Bühlmann and B. Yu, "Boosting with theL2Loss," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[80] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, 2011.

[81] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[82] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352–359, 2002.

[83] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of Electronic Imaging*, vol. 16, no. 4, Article ID 049901, 2007.

[84] S. Roychowdhury and M. Bihis, "AG-MIC: azure-based generalized flow for medical image classification," *IEEE Access*, vol. 4, pp. 5243–5257, 2016.

E-LPDAE: An Edge-Assisted Lightweight Power Data Aggregation and Encryption Scheme

Biraja Nayak, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, birajanayak21@gmail.com*

Rudra Prasad Nanda, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, rudraprasad858@gmail.com*

Sachikanta Pati, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, sachikantapati98@outlook.com*

Rajesh Tripathy, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, rajeshtripathy1@outlook.com*

Abstract

In smart grid systems, electric utilities require real-time access to customer electricity data; however, these data might reveal users' private information, presenting opportunities for edge computing to encrypt the information while also posing new challenges. In this paper, we propose an Edge-assisted Lightweight Power Data Aggregation Encryption (E-LPDAE) scheme for secure communication in a smart grid. First, in the edge privacy aggregation model, the data of smart meters are rationally divided and stored in a distributed manner using simulated annealing region division, and the edge servers of trusted organizations perform key one-time settings. The model encrypts the data using Paillier homomorphic encryption. It then runs a virtual name-based verification algorithm to achieve identity anonymization and verifiability of the encrypted data. The experimental results indicate that the E-LPDAE scheme reduces overall system power consumption and has significantly lower computation and communication overhead than existing aggregation schemes.

1. Introduction

In recent years, with the rapid development of modern science and technology and urbanization, the combination of power systems and information technology has produced a new concept—Smart Grid [1]. Smart Grid is the intelligence of the power grid. Building a smart grid can optimize resource allocation, reduce consumption, and increase efficiency. In smart grid applications, smart meters are deployed in all households in a residential area, each smart meter can collect the user's electricity consumption data and report it to the control center periodically (for example, every 15 minutes), and the control center can perform actions based on the reported data and real-time data analysis and take corresponding measures to ensure the health of the power system. Therefore, in the process of data transmission, a large number of real-time electricity consumption data of users is interacted with and calculated on the transmission line [2].

By using container technology, edge computing [3] is able to collect heterogeneous data in real time across a wide range of devices and can provide elastic computing resources for deep learning models. The resource configuration of edge computing can satisfy offline processing and analysis of small-area data, thereby ensuring the safe transmission and processing of various data. In addition, edge computing can reduce network latency and improve the utilization of network transmission bandwidth with the help of high-speed communication technology. In the implementation process of smart grid, the introduction of edge computing has a good development prospect, as shown in Figure 1.

Interaction and calculation of real-time electricity consumption provide a great convenience for power companies to fully grasp the electricity consumption of their customers but, at the same time, pose serious security and privacy risks. As pointed out by the National Institute of Standards and Technology (NIST) in the United States, there are more and richer data in smart grid systems. While bringing convenience to services, data leakage will also bring many security threats. Once the real-time electricity consumption information is stolen by the attacker, through the

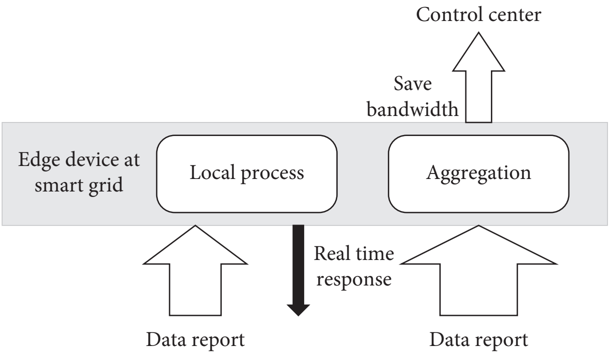


Figure 1: The edge computing paradigm extends cloud computing capabilities to the edge of the network to provide real-time response to local processes, as well as aggregated bandwidth savings.

analysis of the data, the user's detailed family life habits and other information can be obtained. Therefore, how to protect user privacy and data security in smart grids has become a research hotspot in recent years [4].

In order to overcome the above challenges, we propose an edge-assisted lightweight power data aggregation and encryption scheme. The main contributions of this paper are summarized as follows:

(i) An edge privacy aggregation model is proposed. The model uses simulated annealing (SA) to propose a segmentation algorithm for smart meters, Simulated Annealing Region Division (SARD). The algorithm can generate optimal area division according to the energy consumption of electricity meters, which is convenient for data collection and analysis of cluster electricity meters. The realization of distributed data storage is conducive to the privacy protection of smart meter data.

(ii) The Trusted Organization (TO) can set all keys in the system at one time, improve the efficiency of the smart grid, and reduce the power consumption of the system. Since a trusted organization stores a large amount of sensitive information such as keys, if it is stolen by an attacker, it will seriously threaten the data privacy and security of users. Such issues can be resolved by using edge servers, which are relatively trustworthy.

(iii) A virtual name-based authentication algorithm is proposed. The algorithm uses an encryption mechanism combining chameleon signature and Paillier cryptography to encrypt and verify the data to ensure the security of transmitted data while reducing the communication overhead; a selection strategy is developed using an attribute decision tree to improve the value of the data. Finally, the aggregated encrypted data is sent to the Cloud Power Distribution Center (CPDC). The CPDC decrypts the data in order to obtain the final result.

The rest of this paper is organized as follows. Section 2 summarizes the related work. In Section 3, we do some preparatory work. In Section 4, we describe the procedure and algorithm of the scheme. The results of the experimental analysis are reported in Section 5. Finally, the conclusions are discussed in Section 6.

2. Related Works

Although many secure communication schemes to protect the privacy of smart grid users have been introduced over the years, not many privacy-preserving aggregation schemes such as [5–8] have been proposed so far. Electricity consumption data collection is an important process in smart grid communication systems. However, a report from the Netherlands argues that frequent reading of smart meters is problematic from a legal point of view [9], violates the European Convention on Human Rights, and generates many load issues. Fortunately, integrating edge computing into smart grids and designing data aggregation schemes that protect privacy can avoid these problems. First, Pacific Northwest National Laboratory first proposed "edge computing" in an internal report in 2013. With the rapid growth of the Internet of Things, edge computing has received a lot of attention. Shi et al. summarize typical examples of the smart home and collaborative edge and present some of the challenges and opportunities in the area of edge computing [10]. It moves some of the workloads used in the cloud to the edge nodes. The security of sensitive data stored on cloud servers through edge nodes will be of great concern to users. Therefore, consideration should be given to the resource requirements of edge devices, as well as the privacy of smart grid users.

To address these issues, we use data aggregation technology to solve the transmission conflict problem of a large number of data packets for smart grids in edge computing. To improve the security of the data aggregation model, traditional secure data aggregation schemes use hop-by-hop aggregation encryption [11]. However, frequent encryption and decryption operations may affect the aggregation efficiency and increase the corresponding additional energy consumption and the delay of the data aggregation process. An efficient privacy-preserving aggregation scheme (EPPA) for smart grid communication [7] was proposed by Lu et al. They used a super-incremental sequence to construct multidimensional data and encrypted the data with Paillier homomorphic encryption [12]; however, the scheme has security flaws. Shi et al. used an untrusted aggregator to differentially aggregate multiple time slots, which is more costly based on computationally intensive systems [13]. Fan et al. [14] proposed a secure power usage data aggregation scheme for smart grids, but it critically requires a third-party trust mechanism for distribution, adding an additional burden. Li et al. proposed a distributed incremental data aggregation approach where they used homomorphic encryption to solve the repetitive regular data aggregation task [5]. Garcia and Jacobs used homomorphic encryption to ensure the privacy of users and gave a measurement method [6]. Lu et al. proposed a lightweight privacy-preserving data aggregation scheme called Lightweight Privacy-Preserving Data Aggregation (LPDA), but it cannot achieve identity anonymization [15]. Hua et al. proposed an effective smart grid aggregation scheme against malicious data mining attacks but increased the computational overhead and communication overhead [16].

3. Preparation

This section reviews the main basic concepts related to our work, including Paillier homomorphic encryption, simulated annealing region partition [17], chameleon hash function [18], and Attribute decision tree.

3.1. Paillier Homomorphic Encryption. Paillier cryptography is an additive homomorphic public key cryptography, which has been widely used in the field of encrypted signal processing or third-party data processing. Its homomorphic property is that the corresponding arithmetic operation can be performed on the ciphertext directly after encryption, and the result of the operation is the same as that of the corresponding operation in the plaintext domain. Its probabilistic property is that for the same plaintext, different ciphertexts can be obtained by different encryption processes, thus ensuring the semantic security of the ciphertext. The mechanisms used for encryption and decryption are as follows:

(1) Key generation: randomly select two large prime numbers p and q , calculate their product N and the least common multiple of $p-1$ and $q-1$, and then randomly select an integer that satisfies the following conditions:

$$\gcd\left(L\left(g^{\lambda} \bmod N^{2}\right), N\right) = 1. \tag{1}$$

Among them, function $L(u) = (u-1)/N$ and function $\gcd(.)$ are used to calculate the greatest common divisor of two numbers. Z_{N^2} is the set of integers less than $x \in Z_{p}^{*}$, while $Z_{N^2}^{*}$ is the set of integers coprime with N^2 in $Z_{N^3}^{*}$. (N, g) and λ are public key and private key, respectively.

(2) Encryption process: a random integer $r \in Z_i$ is selected. For any plaintext $m \in Z_w$, the corresponding ciphertext c is obtained by using public key (N, g) encryption:

$$\begin{aligned} c &= E[m, r] \\ &= g^{m} r^{N} \bmod N^{2}. \end{aligned} \tag{2}$$

According to the properties of the Paillier encryption system, when ciphertext $c \in Z_{N^2}^{*}$ is encrypted with the same public key, because the selection of ciphertext r is random, different ciphertext c can be obtained for the same plaintext m , but the same plaintext m can be restored after decryption, thus ensuring the semantic security of ciphertext.

(3) Decryption process: decrypt ciphertext c with private key n to get the corresponding plaintext m.

$$\begin{aligned} m &= D[c] \\ &= \frac{L\left(c^{\lambda} \bmod N^{2}\right)}{L\left(g^{\lambda} \bmod N^{2}\right)} \bmod N. \end{aligned} \tag{3}$$

3.2. Simulated Annealing Region Division

3.2.1. Regional Division. For a given smart meter, the division of area Q is expressed as follows:

$$Q \equiv \sum_{s=1}^{s_Q} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^{2} \right]. \tag{4}$$

s_Q is the number of regions, L is the number of links between smart meter nodes in the smart grid, l_s is the number of regions in region Q, L is the number of links between smart meter nodes in the smart grid, l_s is the number of links between smart meter nodes in region Q, and d_s is the sum of degrees of smart meter nodes in region Q. First, we use equation (4) to randomly place smart meters on the device layer into the area. Finally, we use a simulated annealing algorithm to find the optimal partition.

3.2.2. Simulated Annealing Algorithm. It is a general probabilistic algorithm that is used in our scheme to find the optimal solution to the zoning problem, where one can find low-cost smart meter regions, but not local minima for high-cost smart meter regions. We introduce the energy consumption T_e of smart meters to achieve this. Starting from high T_e, it gradually decreases and the system gradually approaches the minimum cost, avoiding the high-cost local minima.

The purpose of identifying modules is to maximize the use of modules, where costs $C = -Q$ and Q are the areas defined in equation (4). We update each energy consumption randomly, and the probability is expressed as

$$p = \begin{cases} 1 C(S') \leq C(T_e), \\ \exp\left(-\dfrac{C(S') - C(T_e)}{T} \right) C(S') > C(T_e), \end{cases} \tag{5}$$

where $C(S')$ is the cost after the update and $C(T_e)$ is the cost before the update, $\Delta C = C(S') - C(T_e)$.

3.3. Chameleon Hash Function. Traditional cryptographic hash functions are difficult to find collisions. But the chameleon hash function can artificially set up a “back door”: if you master it, you can easily find collisions. This breaks the collision resistance of the hash function, but for most people, these properties remain, and the hash is still secure. Accenture applied the characteristics of the chameleon hash function and applied for a patent on an editable blockchain.

Although the decentralization and irrevocability of the blockchain are damaged to a certain extent, on the other hand, it also expands the application scenario of the blockchain and meets part of the needs of the government's regulatory requirements [19].

Principle description: suppose there exist two prime numbers p, q , and $q = kp + 1$ is large enough. The private key of the chameleon hash function is $x \in Z_p^*$, Z_p^* is the group of order q, and g is its generating element. The public key is $h = g^x \bmod p$. Given an arbitrary message m with random value $r \in Z_p^*$, now tampering the content m to m', it is now desired to find a random number r' such that $H(m') = H(m)$. By the exponential property $g^a * g^b = g^{(a+b)}$, $(g^a)^b = g^{(ab)}$. The solution procedure for r' is as follows:

$$H(m) = g^m h^r \bmod p = g^m g^{xr} \bmod p = g^{(m+xr)} \bmod p,$$
$$H(m') = g^{m'} h^{r'} \bmod p = g^{m'} g^{xr'} \bmod p = g^{(m'+xr')} \bmod p. \tag{6}$$

Therefore, m, m', x, and r are known, $r' = (m + xr - m')/x \bmod p$.

3.4. Attribute Decision Tree. The attribute decision tree is modeled after the access control tree and is set up according to the needs of the data collector. The leaf nodes of the attribute decision tree represent various attributes, and the intermediate nodes and roots are replaced by AND and OR. When an attribute of the data satisfies the requirements of the attribute decision tree, it is passed and the next calculation is performed; if not, other calculations or steps are performed.

For example, Mr. Li is a professor in the school of computer science of a university, so his attribute set matches the attribute strategy, as shown in Figure 2. Miss Wang is a professor in the school of information security of a university. Her attribute set does not match the attribute policy, as shown in Figure 3.

4. Edge-Assisted Lightweight Power Data Aggregation Encryption Scheme

4.1. Edge Privacy Aggregation Model. The edge privacy aggregation model contains four subjects: the User's Smart Meter (USM), the Marginal Power Services Institutions (MPSI), the Cloud Power Distribution Center, and the trusted organization. First, the USM encrypts data and divides it into optimal regions according to the change of energy consumption at different moments using a simulated annealing region partitioning algorithm, and as the energy consumption of USM changes at different moments, the number and location of clustered meters also change, thus realizing distributed data storage, which is conducive to the privacy protection of user data. Secondly, MPSI aggregates data with user identity anonymized and without affecting the privacy of any party. Finally, CPDC performs secure decryption, and TO performs key generation and distributes the key to the system. The model is shown in Figure 4.

User's Smart Meter. Smart meters use TPM chips to securely store and encrypt data. The SARD algorithm is executed using the handheld unit (including the sensor). Divide the smart meters of all users to meet the power load balance of the meters. The cluster meter regularly sends the collected data to the edge server. Perform data encryption calculation and chameleon signature calculation.

Marginal Power Services Institutions(MPSI). It consists of edge servers. The edge server performs chameleon signature aggregation and verification calculations and data aggregation calculations.

Cloud Power Distribution Center. The cloud server receives the aggregated data and decrypts it.

Trusted Organizations. The real identities of all users are virtualized to form virtual names and distribute system parameters and all private keys, and the distribution channels are all secure channels. The three parties of cloud, edge, and smart meter collaborate with trusted organizations to generate all private keys, as shown in Figure 5. Compared with existing solutions, our private keys require only a one-time setup between the three parties, which is beneficial for resource-limited systems. In addition, the private keys owned by TO are involved in decrypting the ciphertext and verifying the ciphertext, confusing the attacker, and making it impossible to tamper with the ciphertext.

4.2. Scheme Construction. The scheme proposed in this paper realizes the security and integrity of real-time power consumption data transmission between the smart meter and power server. The steps are as follows.

4.2.1. Initialization. TO inputs safety parameter (1^λ) and gets related parameter $(q_{1,}G_1, G_2, G_r, g_1, g_2, \omega, e)$, where q_1 is a large prime, G_1 and G_2 are two additive cyclic groups, G_r is a multiplicative cyclic group, q_1 is the order of the cyclic group, g_1 and g_2 are the generators of groups G_1 and G_2, respectively, satisfying that $\omega(g_2) = g_1$ and ω is an isomorphic mapping, e: $g_1 \times g_2 \longrightarrow g_r$ is bilinear mapping, and the storage list is established. TO chooses a system master key $s \in Z_p^*$, Z_p^* is a multiplication cycle group, and $y = g_2^s$ is the system public key. Two hash functions $H_1(.)$: $\{0,1\}^* \longrightarrow G_1$ and $H_2(.)$: $\{0,1\}^* \longrightarrow G_2$.

TO publishes system parameters and functions, selects a security parameter for the Paillier encryption algorithm, and sends it to the smart meter for initialization of the Paillier encryption algorithm. TO generates other parameters of the Paillier encryption algorithm: select two large prime numbers p and q, where $|p| = |q| = k$. The smart table computes $n = pq$ and chooses $g \in Z_{n^2}^*$ as the generator to use (n, g) as the public key of the Paillier encryption algorithm. CPDC computes the private key of the Paillier encryption algorithm $\lambda = lcm(p-1, q-1)$.

For the initialization of the chameleon signature, TO selects an element g_3 of order q in Z_p^* and an arbitrary index

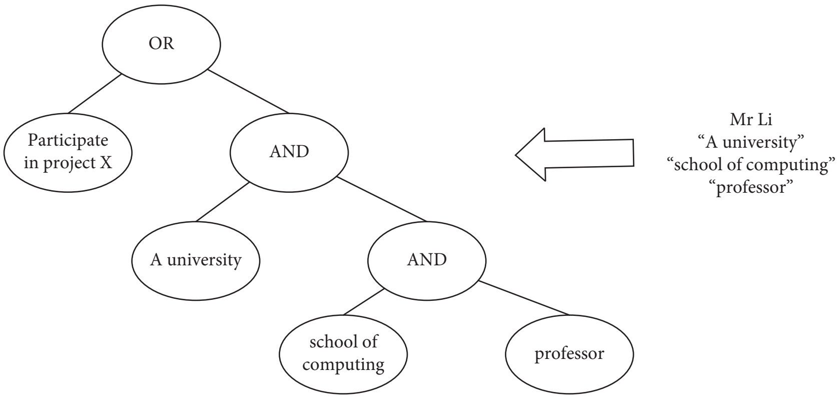


Figure 2: Schematic diagram of successful matching of policy and attribute collection.

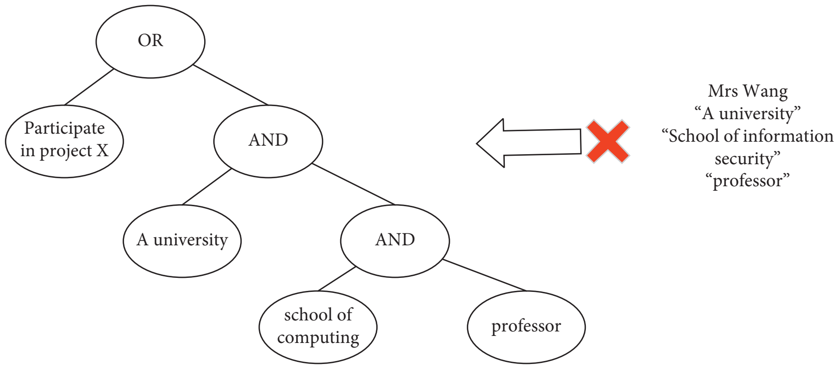


Figure 3: Schematic diagram of policy and attribute collection mismatch.

x , then the private key of the chameleon signature is $CK = x$, and the public key is $HK = g_3^x$.

TO sets the regularized attribute set F as a multiplicative cyclic group; then, any attribute f in the attribute set F is any element in the multiplicative cyclic group. The attribute set F is sent to the smart meter. Similarly, if TO sets the attribute set A of the attribute decision tree as a multiplicative cyclic group, then any attribute a in the attribute set A is any element in the multiplicative cyclic group, and the set attribute set A is sent to CPDC.

4.2.2. User Registration. Assuming a secure channel between TO and the user, in order to complete the user registration, the operation steps between the user and TO are as follows:

User i sends ID, serial number of smart meter to TO.

TO sends a Cert to user i after confirmation.

User i uses the Cert to get permission to request the parameters and key of the algorithm from TO.

TO sends the signature key etc. to the smart meter of user i.

TO calculation:

$$\begin{aligned} DP\,SI\,D &= H(I\,D, t)^{\mathrm{Cert}}, \\ pid_{i,0} &= H(DP\,SI\,D, 0), \\ pid_{i,1} &= H(DP\,SI\,D, 1). \end{aligned} \tag{7}$$

TO calculates the signature key of user i:

$$\begin{aligned} S_{i,0} &= pid_{i,0}^{s}, \\ S_{i,1} &= pid_{i,1}^{s}. \end{aligned} \tag{8}$$

TO sends the signature key $S_i = (S_{i,0} S_{i,1})$, the real-time virtual name $DP\,SI\,D$ to the smart meter of user i.

4.2.3. Data Processing. Within data acquisition time t, the smart meter of user i encrypts the data with the Paillier homomorphic encryption and signs the encrypted data with the chameleon hash function which is referred to as chameleon signature for short. The cluster meter j collects data within the divided area. Finally, the real-time encrypted data and signatures are sent to MPSI. The steps are as follows.

The smart meter of user i selects a random number $a \in Z_{n^2}^*$ and encrypts data m_i.

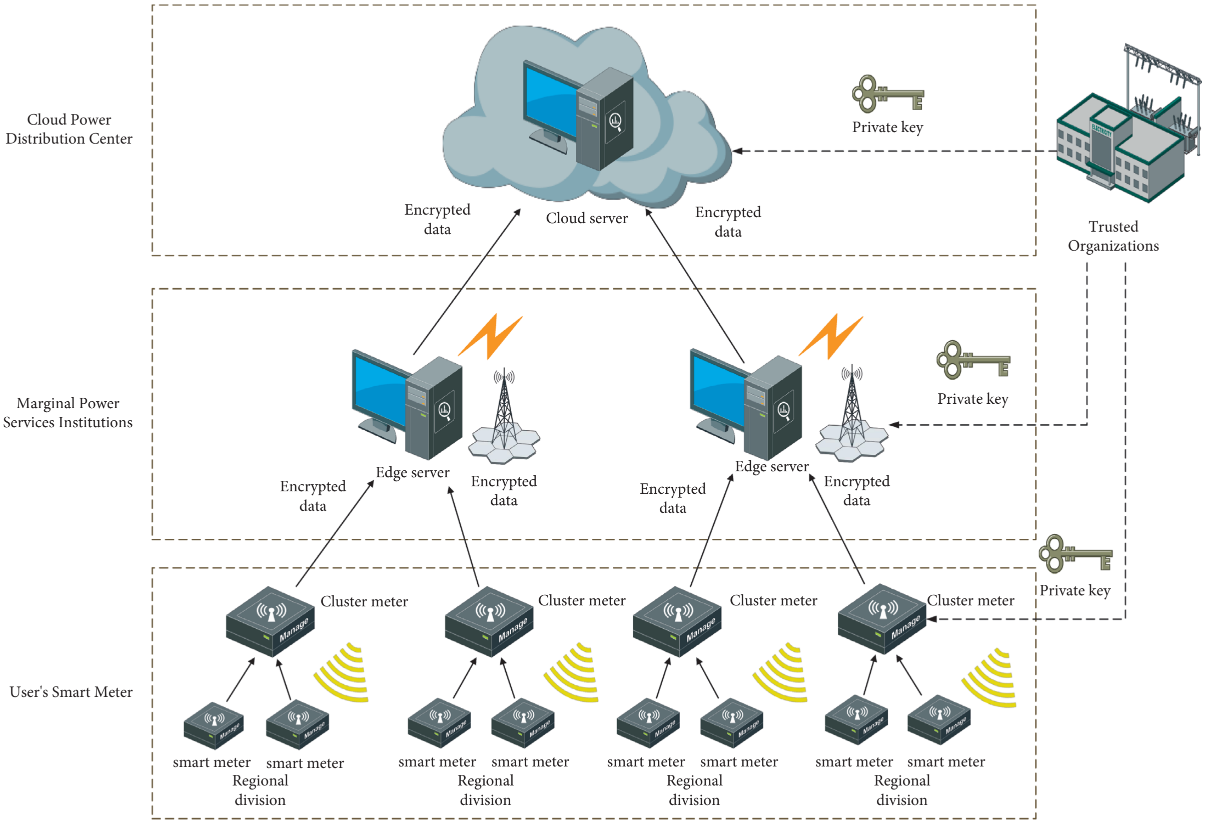


FIGURE 4: Edge privacy aggregation encryption model.

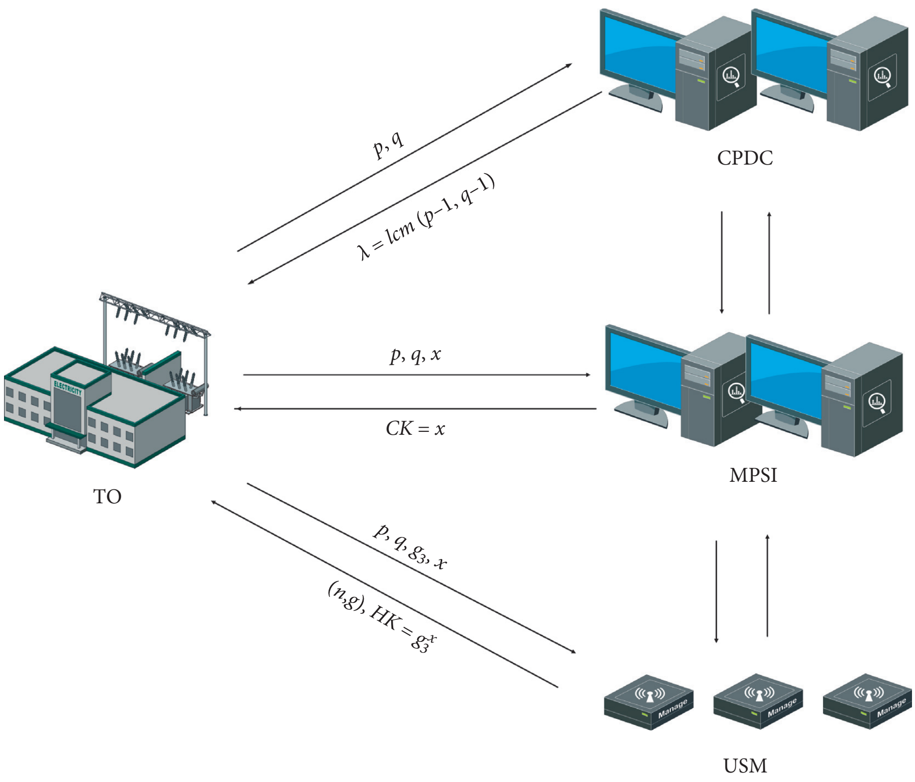


FIGURE 5: Key generation.

$$c_i = E\left(m_i\right) = g^{m_i} a^n \bmod n^2. \tag{9}$$

The smart meter of user i uses signature key $S_i = (S_{i,0} S_{i,1})$, virtual name, and attribute set to sign encrypted data by the chameleon hash function and finally send it to the cluster meter j.

$$\begin{aligned} h_i &= \text{Chamelelon}.H\left(c_i, HK, DP\ SI\ D, f\right), \\ \sigma_i &= s_{i,0} s_{i,1}^{h_i}. \end{aligned} \tag{10}$$

Cluster meter j sends $(c_i, \sigma_i, DP\ SI\ D)$ to MPSI.

MPSI receives the information and runs the virtual name-based verification algorithm as shown in Algorithm 1.

The algorithm first aggregates chameleon signatures. After verification, the attribute set f of the data is obtained, and the attribute set A of the data decision tree is matched in turn, and the data satisfying the data decision tree can be data aggregated with other data satisfying that decision tree for the data aggregation operation.

$$\begin{aligned} c &= \prod_{i=1}^{n} c_i \bmod n^2 \\ &= \prod_{i=1}^{n} g^{m_1} \ldots g^{mw} a^n \bmod n^2 \quad . \\ &= \prod_{i=1}^{n} g^{m_1+m_2+\ldots+m_w} a^n \bmod n^2 \end{aligned} \tag{11}$$

After aggregation, MPSI sends the aggregated data to the CPDC through the secure channel. Data decryption: CPDC decrypts the encrypted aggregate data.

$$m_i = \frac{L\left(c^{\lambda} \bmod n^2\right)}{L\left(g^{\lambda} \bmod n^2\right)} \bmod n. \tag{12}$$

$m_i = m_1 + m_2 + \ldots\ldots + m_w$, CPDC stores data for power grid operation and puts forward decisions.

4.2.4. Track. While making power consumption analysis and decision-making, CPDC may find that some power consumption values do not meet its predetermined range or abnormal conditions. At this time, CPDC will start the tracking process, and the steps are as follows:

CPDC sends the command to the edge server that submits the relevant abnormal power consumption: let MPSI send the stored power consumption and virtual name at that time to CPDC.

CPDC first decrypts each encrypted data received, detects and finds the abnormal power consumption, and locks its $DP\ SI\ D$.

CPDC sends the virtual name of the abnormal power consumption determined by it to TO and applies for identity tracking.

TO can query the real identity of the users who send out abnormal electricity consumption. TO sends the real identity to CPDC, and CPDC processes the user and his power consumption accordingly.

4.3. Safety Analysis

4.3.1. User Identity Privacy Protection. Before sending data to the CPDC, the USM registers with the TO to obtain a virtual name and signing key. The USM uses the virtual name as the identity of the data transfer in the architecture and performs encryption, signing, and other actions based on it. The USM has a tamper-proof storage device. This storage device can be thought of as a "black box" that can read and write data, but only by the USM; no other device can read or write information. According to the one-way and collision-free characteristics of the hash function, even if the attacker obtains the virtual name, it cannot crack the real identity. This scheme can effectively protect user identity and prevent illegal intrusion.

4.3.2. Security Analysis of Chameleon Signature. The chameleon signature is a preferable designated verifier signature. Compared to other signatures, the chameleon signatures are more suitable for lightweight aggregated encryption schemes due to their ability to transmit data efficiently and reduce computational overhead. Chameleon signatures are also nontransmissible, nonforgeable, and nonrepudiation, which also ensure data security and meet the security requirements of the system.

4.3.3. User Fine-Grained Data Privacy Protection. USM encrypts the electricity consumption data using the Paillier encryption algorithm, sends it to MPSI, which does not have the ciphertext decryption key, and sends the ciphertext to CPDC after successful verification. CPDC mainly receives aggregated numbers of electricity consumption data, so it protects the user's fine-grained data privacy, while CPDC can get the complete electricity consumption data.

5. Experimental Analysis

5.1. Simulated Annealing Region Division. Intraregional connectivity and participation: each region is divided into relatively balanced regions from one or several fully centralized regions based on the energy consumption of smart meters to achieve a balanced electrical load in each region. We define the intraregional connectivity, in order to measure whether the smart meter u is well connected to other smart meters in the region.

$$Z_u = \frac{k_u - \bar{k}_{s_u}}{\sigma_{k_{s_u}}}, \tag{13}$$

where k_u is the number of links from the smart meter u to other smart meters in zone s_u, $\overline{k}_{s_u}$ is the average number of links from all smart meters in the zone s_u, and $\sigma_{k_{s_u}}$ is the standard deviation of all links in the zone.

Of course, we also need to consider unexpected situations. For example, a smart meter u may not be connected to its own area. Therefore, we define the participation degree p_u of a smart meter u.

$$p_u = 1 - \sum_{s=1}^{s_M}\left(\frac{k_{us}}{k_u}\right)^2, \tag{14}$$

where k_{us} is the number of links from the smart meter u to smart meters in zone s , and k_u is the total number of degrees of the smart meter u. According to equation (14), if the connections of the smart meter u are evenly distributed in all areas, then the participation degree of the smart meter u is close to 1. If all its connections are in its own area, the participation degree is 0.

We use a MATLAB environment with a Dell laptop (i5-6200u, CPU 2.40 GHz, Windows 10 OS) for simulation experiments. Assuming that 100 smart meters are randomly distributed in a $1.0 * 1.0$ km smart grid, and each smart meter has a random electricity consumption $N(T_e)$, a zoning model is established. First, the 100 randomly distributed smart meters are generated as a subset of the neighborhood of electricity consumption $N(T_e)$ Download the open-source dataset from the website Open Energy Data Initiative (OEDI) and randomly select the electricity consumption information from 100 apartments with no missing points and a time granularity of 15 minutes. The average value is calculated based on the electricity load of 100 users at different times of the day, as shown in Figure 6. 14:00–20:00, the user's electricity load continues to grow, with 20:00 reaching the highest peak of the day; 20:00–24:00, the user's electricity load continues to fall to a stable value. After reasonable analysis, we divide the average value of the electricity load of 100 users in different time periods of a day into 6 electricity consumption states. A power consumption state of $S(k)$ is randomly selected for the regional division scheme, and the next power consumption state of S' is randomly selected as the candidate scheme for the next regional division scheme. Calculate $\Delta C = C(S') - C(T_e)$; if $\Delta C < 0$, accept S' for the next region division scheme; otherwise, we judge the random update probability $p = \exp(-\Delta C/cT) > \alpha$, $\alpha \in (0, 1)$; if true, accept S' for the next region division scheme; namely, $S(k+1) = S'$, $k = k+1$. Then, we check whether the connectivity and participation in the region satisfy equations (13) and (14). Finally, we use $S(k+1)$ for the region partition scheme and return the SARD algorithm.

Figures 7(a)–7(f) show the experimental process of the SARD algorithm. We performed six rounds of state calculation, divided the six power consumption states into different regions, and terminated the algorithm. Cluster meters in each area are used to collect data and process the data accordingly to realize power load balancing under different power consumption states.

First, the power consumption of smart meters increases with the increase of users in the smart grid. Since all the data eventually needs to be sent to the cloud server of CPDC for processing, the power consumption of the cloud server also increases with the increase of data, as shown in Figure 8. Then, we introduce edge computing into the smart grid, and the power consumption of MPSI increases with the increase of edge servers. This layer processes a large amount of data and then sends it to the CPDC. Since the CPDC does not need to process a large amount of data, the power consumption of the cloud server in the CPDC does not fluctuate much, as shown in Figure 9. Comparing the experiments in the two figures, the introduction of edge servers to process large amounts of data in the edge privacy aggregation model of the smart grid effectively reduces the power consumption of the CPDC and the total system power consumption.

5.2. Total Computing Overhead. The computational overheads of this scheme and the LPDA scheme mainly involve the following three operations: bilinear pair operation, exponential operation, and Paillier homomorphic encryption and a decryption operation, and other operations are neglected. The bilinear pair operation and the exponential operation are C_b and C_e, respectively, and the encryption and decryption of the Paillier algorithm are C_A and C_B, respectively, and the other computational overheads are neglected. The AMDM scheme is mainly multiple operations, C_{pe} is the multiplication operation in the cyclic group $Z^*_{N^2}$, C_{pm} and C_m are the multiplication operation in $Z^*_{p'}$, C_e is the exponential operation, and C_{gm} is the multiplication operation in the group G_1 because C_{pm} and C_m produce little effect negligible.

The scheme uses the MATLAB environment of a Dell laptop (i5-6200u, CPU 2.40 GHz, Windows 10 OS) for simulation experiments. The simulation measures the amount of time needed by the Dell laptop to perform basic operations in the experimental environment. It takes 1.1 ms to calculate a single C_e, 3.1 ms to calculate C_b, 4.5 ms to calculate C_{pe}, and 2.1 ms to calculate C_{gm}. Since all three scenarios in this paper have only one pair of encryption and decryption operations, we first disregard C_A and C_B .

The scenarios in this paper consider the computational overhead of each of the three participants, Smart Meter, MPSI, and CPDC, and compare them with other scenarios, as shown in Table 1. The total computational overhead of all the solutions is the total computational overhead of the three participants. As can be seen from the table, this paper is significantly more efficient than the other two schemes.

As shown in Figure 10, the computing energy consumption of the scheme in this paper is significantly lower than the aggregated encryption schemes of the remaining two schemes, where the AMDM scheme resists malicious attacks and requires more computing energy and is significantly higher than the LPDA scheme and the scheme in this chapter, while the scheme in this chapter does not cause additional computation while ensuring data security due to the use of the chameleon signature, so the total computation overhead is lower, and it can be said that the scheme in this chapter is better than the LPDA scheme and the AMDM scheme.

5.3. Total Communication Overhead. The total communication overhead of this scheme mainly refers to all the communication data that needs to be transmitted in the system. The output data length of the hash function is

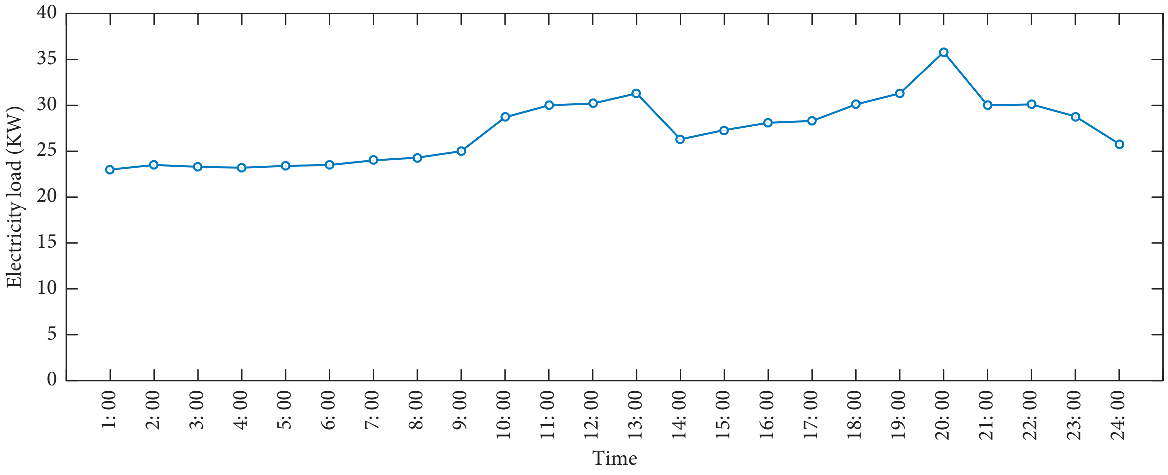
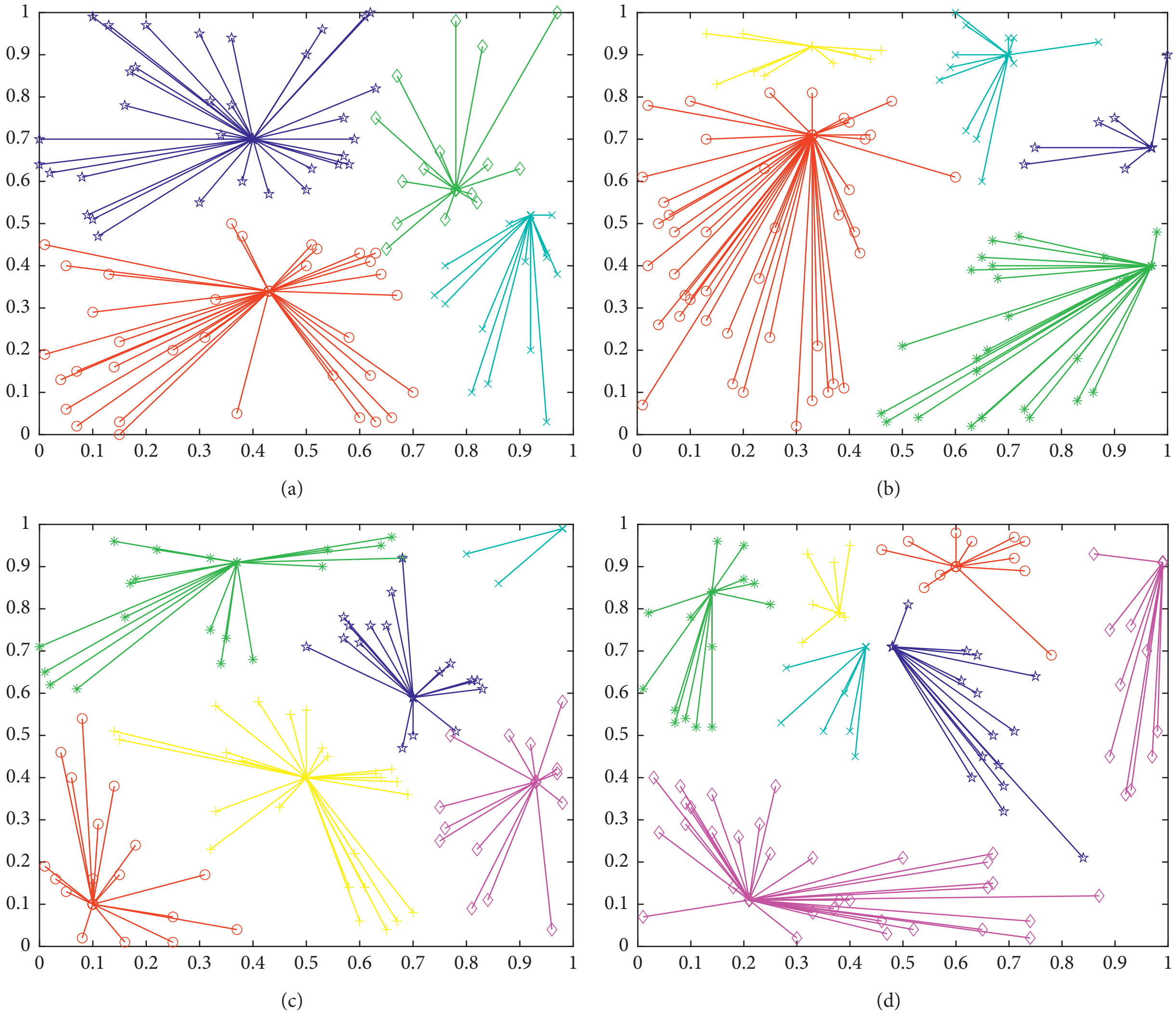


Figure 6: Edge privacy aggregation encryption model.



(a)

(b)

(c)

(d)

Figure 7: Continued.

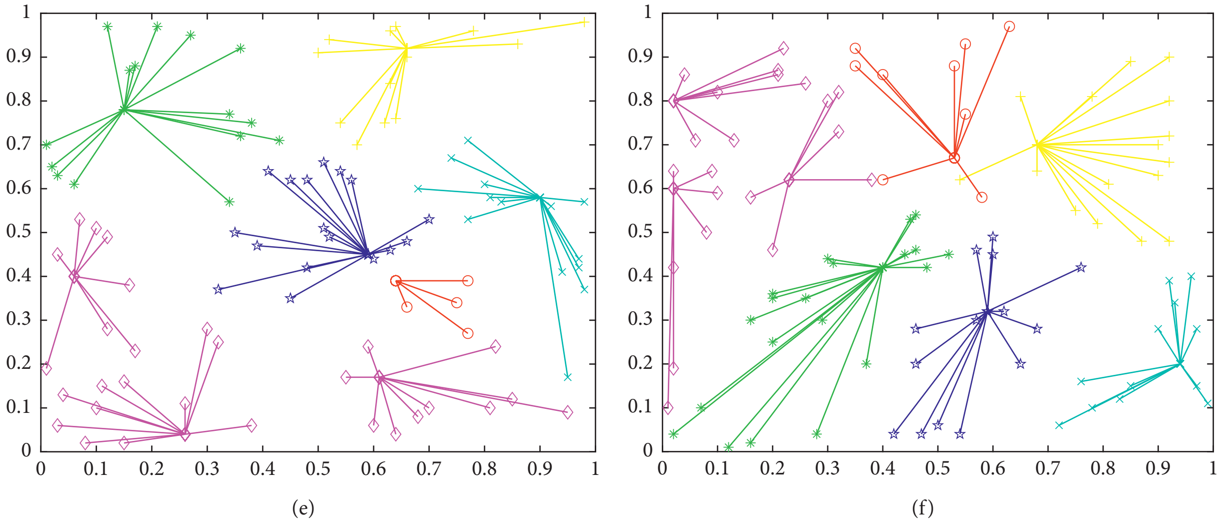


FIGURE 7: Regional division based on electricity consumption state. (a) Partition of 0:00–7:00. (b) Partition of 13:00-14:00. (c) Partition of 7:00–11:00. (d) Partition of 11:00–13:00. (e) Partition of 14:00–20:00. (f) Partition of 20:00–24:00.

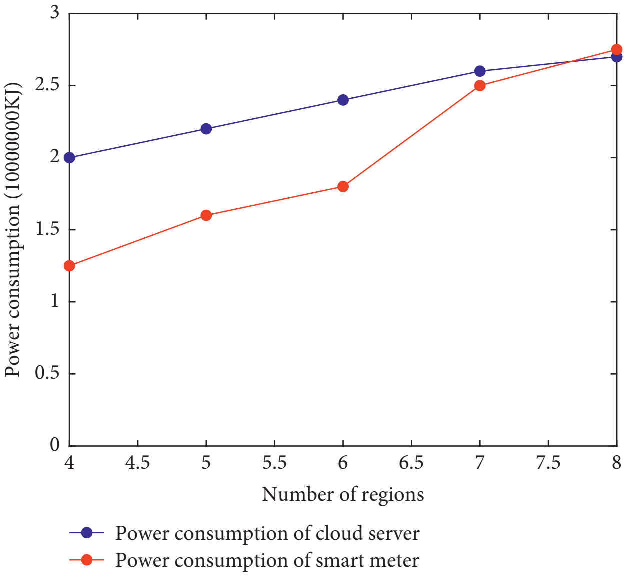


FIGURE 8: Energy consumption of cloud network.

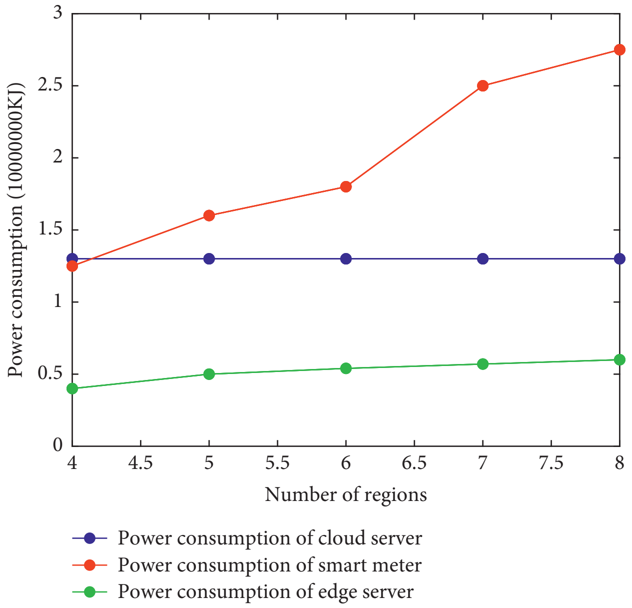


FIGURE 9: Energy consumption of edge computing network.

160 bits. Suppose the length of n in Paillier encryption algorithm is 512 bits, the length of group G_1 element is 161 bits, the length of $DP\ SI\ D$ is 32 bits, the length of attribute set f is 32 bits, and the length of σ_i is 32 bits. The total communication data volume of this scheme consists of two parts: the first part is from SM to MPSI, and the data transmitted is $(c_i, \sigma_i, DP\ SI\ D)$; the second part is from MPSI to CPDC, and the data transmitted is c . The total traffic of the LPDA scheme consists of two parts. The first part is from SM to ESP, which transmits 2048 bits through calculation, and the second part is from ESP to CC, which transmits 2048 bits through calculation. The total traffic of the AMDM scheme consists of two parts. The first part is SM to GW, which transmits 3264 bits through calculation, and the second part is GW to CC, which transmits 3264 bits through calculation. The comparison between this scheme and other schemes is shown in Table 2.The simulation experiment is carried out using MATLAB, and the results are shown in Figure 11.

We use the smart meter data of a year in London on the Open Energy Data Initiative (OEDI) website to simulate the total communication cost per day. As shown in Figure 12, different colors represent different communication situations; that is, when the number of edge servers and smart meters changes, the communication cost also changes. Based on the actual privacy requirements and cost requirements of the customer, we implement appropriate electricity usage data delivery mechanisms in the actual area.

```
Input: $c_i, \sigma_i$, $DP\ SI\ D$
Output: $c$
for $i = 1; i < n; i++$ do
    $\Omega = \prod_{i=1}^{n} \sigma_i$;
    $h_i$ = Chamelelon.$H(c_i', CK, DP\ SI\ D, f\prime), c_i', f\prime \in Z_p^*$;
    $f = f' - c_i - c_i'/x \bmod p$;
    $f$ match $A$;
    $pid_{i,0} = H(DP\ SI\ D, 0), pid_{i,1} = H(DP\ SI\ D, 1)$;
    $e(\Omega, g) = e(\prod_{i=1}^{n} pid_{i,0} pid_{i,1}^{h_i}, y)$;
    $c = \prod_{i=1}^{n} c_i \bmod n^2$;
end for
MPSI sends $c$ to CPDC;
```

Algorithm 1: Verification algorithm based on the virtual name.

Table 1: Analysis of computational complexity.

Scheme	SM	MPSI (ESP, DCP)	CPDC (CC)
Our scheme	$3C_e + C_A$	$NC_e + 2C_b$	C_B
LPDA	$C_e + C_A$	NC_e	$NC_e + C_B$
AMDM	$2C_{pe} + 2C_e + C_{gm} + C_A$	$(N + 2)C_b + C_{gm}$	$2C_b + C_{pe} + 2C_e + C_B$

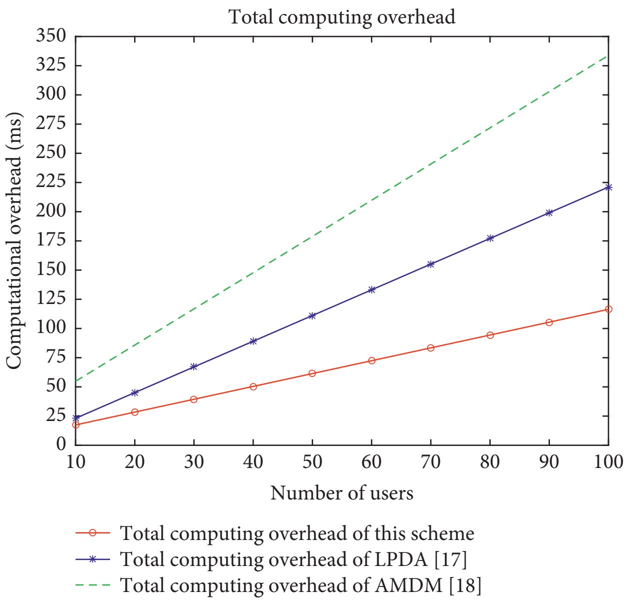


Figure 10: Total computing overhead.

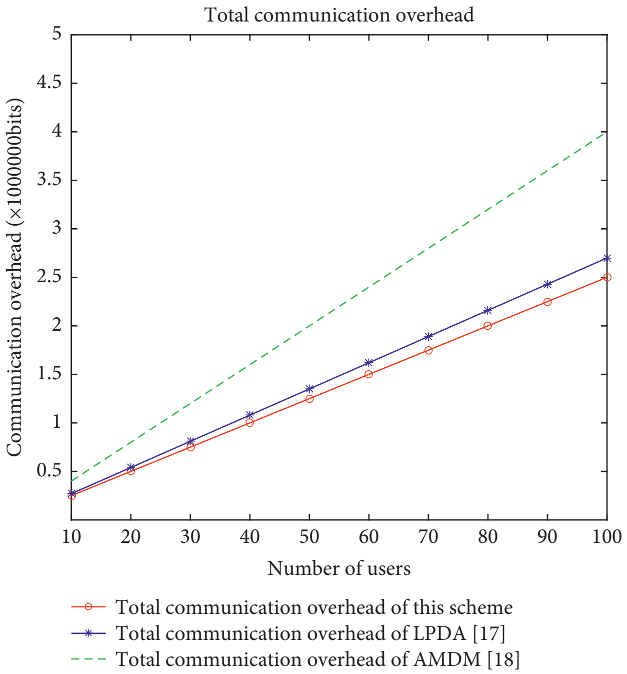


Figure 11: Total communication overhead.

Table 2: Analysis of communication complexity.

Scheme	SM (bit)	MPSI (ESP, DCP) (bit)
Our scheme	1409	1024
LPDA	2048	2048
AMDM	3264	3264

6. Conclusion

In this paper, we consider the actual smart grid, introduce edge computing, and propose an edge-assisted lightweight electricity consumption data aggregation and encryption

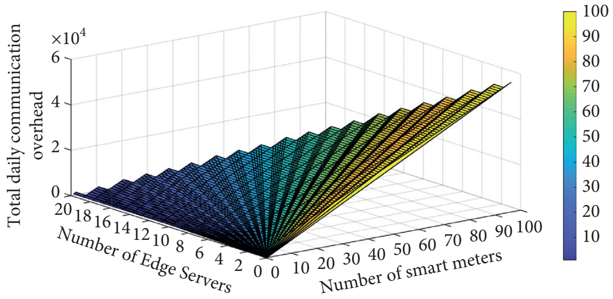


FIGURE 12: Total daily communication overhead.

scheme, which solves the problem of sending electricity consumption data to the cloud by users securely and efficiently. The scheme uses a simulated annealing zone partitioning algorithm to reasonably partition smart meters according to their electricity consumption energy consumption to achieve load balancing of smart grid systems; at each sending of data, licensed users apply for virtual names from trusted organizations to enable them to communicate with the grid as anonymous, which effectively protects the privacy of user identity security; in encrypting data, CPDC uses a virtual name-based verification algorithm which is used to Paillier encryption technology combined with chameleon signature to ensure authentication, integrity, and nonrepudiation of data, so that CPDC can only obtain encrypted data aggregated by MPSI, protecting the privacy of users' fine-grained data. Performance analysis shows that it is much better than the LPDA scheme and AMDM scheme in terms of communication overhead and computation overhead. In future work, we will evaluate our schemes in realistic smart grid scenarios with stronger adversarial models and study the impact of different signatures on system performance and security.

References

[1] A. Saleem, A. Khan, S. U. R. Malik et al., "FESDA: fog-enabled secure data aggregation in smart grid IoT network," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6132–6142, 2020.

[2] K. Wei, S. Jian, and Pandi, "A practical group blind signature scheme for privacy protection in smart grid," *Journal of Parallel and Distributed Computing*, vol. 136, pp. 29–39, 2020.

[3] J. Xiong, R. Bi, M. Zhao, J. Guo, and Q. Yang, "Edge-assisted privacy-preserving raw data sharing framework for connected autonomous vehicles," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 24–30, 2020.

[4] J. Xiong, J. Ren, L. Chen et al., "Enhancing privacy and availability for data clustering in intelligent electrical service of IoT," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1530–1540, 2019.

[5] F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *Proceedings of the First IEEE International Conference on Smart Grid Communications*, pp. 327–332, IEEE Press, Gaithersburg, MD, USA, October 2010.

[6] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Proceedings of the 6th International Conference on Security and Trust Management*, vol. 67, no. 10, pp. 226–238, Springer-Verlag, Berlin, Germany, 2011.

[7] R. Rongxing Lu, X. Xiaohui Liang, X. Xu Li, X. Xiaodong Lin, and X. Xuemin Shen, "EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, 2012.

[8] R. Petrlic, "A privacy-preserving concept for smart grids," *Sicherheit in Vernetzten Systemen*, pp. B1–B14, 2010.

[9] Q. Zhou, G. Yang, and L. He, "An efficient secure data aggregation based on homomorphic primitives in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 10, no. 1, pp. 2022–2037, 2014.

[10] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[11] Y. Yang, X. Wang, S. Zhu, and G. Cao, "Sdap," *ACM Transactions on Information and System Security*, vol. 11, no. 4, pp. 1–43, 2008.

[12] P. Paillier, "A public-key cryptosystem based on composite degree residuosity classes," *Advances in Cryptology - EUROCRYPT'99*, vol. 1592, pp. 223–238, Springer-Verlag, Berlin, 1999.

[13] E. Shi, T. H. Chan, E. G. Rieffel, R. Chow, and D. Song, "Privacy preserving aggregation of time-series data," *Network and Distributed System Security (NDSS)*, vol. 2, no. 4, pp. 1–17, 2011.

[14] C.-I. Fan, S.-Y. Huang, and Y.-L. Lai, "Privacy-enhanced data aggregation scheme Against internal attackers in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 666–675, 2014.

[15] R. Lu, K. Heung, A. H. Lashkari, and A. A. Ghorbani, "A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT," *IEEE Access*, vol. 5, pp. 3302–3312, 2017.

[16] Y. Dong, J. hen, S. Ji, Q. Rongxin, and L. Shuai, "A novel appliance-based secure data aggregation scheme for bill generation and demand management in smart grids," *Connection Science*, vol. 33, no. 4, pp. 1–22, 2021.

[17] L. Ren and L. Lin, "Simulated annealing algorithm coupled with a deterministic method for parameter extraction of energetic hysteresis model," *IEEE Transactions on Magnetics*, vol. 54, no. 11, pp. 1–5, 2018.

[18] S. A. Hua, A. Yl, X. D. Zhe, and Z. Mingwu, "An efficient aggregation scheme resisting on malicious data mining attacks for smart grid," *Information Sciences*, vol. 526, pp. 289–300, 2020.

[19] Y. Tian, T. Li, J. Xiong, M. Z. A. Bhuiyan, J. Ma, and C. Peng, "A blockchain-based machine learning framework for edge services in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1918–1929, 2022.

Enhanced Image Processing and Fuzzy Logic Approach for Optimizing Driver Drowsiness Detection

Sudhansu Sekhar Khuntia, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sskhuntia88@gmail.com*

Rakhi Jha, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, rakhijha91@yahoo.co.in*

Subrat Dash, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, subratdash43@gmail.com*

Niladri Bhusan Biswal, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, niladribiswal26@gmail.com*

Abstract

Driver drowsiness is a severe problem that usually causes traffic accidents, classified as more dangerous. The record of the National Safety Council reported that drowsy driving is caused by 9.5% of all crashes (100,000 cases). Therefore, preventing and minimizing driver fatigue is a significant research area. This study aims to design a nonintrusive real-time drowsiness system based on image processing and fuzzy logic techniques. It is an enhanced approach for Viola–Jones to examine different visual signs to detect the driver's drowsiness level. It extracted eye blink duration and mouth features to detect driver drowsiness based on the desired facial feature image in a specific driver video frame. The size and orientation of the captured features were tracked and handled for determining image features such as brightness, shadows, and clearness. Lastly, the fuzzy control system provides different alert sounds based on the tracked information from the face, eyes, and mouth in separate cases, such as race, wearing glasses or not, gender, and various illumination backgrounds. The experiments' results show that the proposed approach achieved high accuracy of 94.5% in detecting driver status compared with other studies. Also, the fuzzy logic controller efficiently issued the required alert signal of the drowsy driver status that helps to save the driver's life.

1. Introduction

According to Peura et al. [1], driver drowsiness is one of the significant factors that cause many traffic accidents in the world. Annually, upwards of 100,000 vehicles are crashed, and around 1,500 people die. Annually, almost 12.5 billion dollars are the total loss cost for these types of accidents. Preventing such driver fatigue accidents is a high focus effort on many current types of research [2, 3]. Driver fatigue increases the need to develop a monitoring system that analyses driver status and provides different alert sounds based on his facial features. Most researchers concentrated on detecting drowsy driving through analyzing the eyes' pupil's parameters [4–6]. Through this research, driver drowsiness investigations are based on capturing the driver's video and detecting the driver's face using some technique. After that, they analyze the eye blinking frequency and decide the driver fatigue status [7]. Some other researchers [8–10] include the mouth and features too. For some reason, the system may work inefficiently to detect driver drowsiness. Varying light conditions and vibration of the driver is one of the main challenges to identifying the driver status in real time [11]. However, using multiple visual features and cues is more efficient for detecting driver drowsy.

The proposed system detects driver fatigue in real time through observing various facial features and selecting the correct driver state. It provides different alert sounds based on a different level of drowsing for the driver. This has been carried out by using a fuzzy logic technique in addition to the face features' detection process. This study is organized as follows. Section 2 explores significant reviews of previous studies and related works in the same field of research systems. Section 3 describes the model of the system and how the experiments were conducted. Section 4 displays the results of the conducted experiments. Finally, present the conclusion and future work.

2. Background and Related Work

2.1. Driver Drowsiness. Drowsiness causes significant social and economic losses to the country about road accidents that often occur on highways. Drowsiness accidents happen when the driver delays responding to a specific situation and lose control of the vehicle. Moreover, it is difficult to determine the level of the driver's drowsiness because it cannot be measured after the accident. Drowsiness usually affects various attitudes such as vigilance, decision-making, and concentration for drivers [12].

2.2. Driver Fatigue Monitoring Techniques. Various techniques of monitoring vigilance and fatigue are used to measure driver performance, including the following.

2.2.1. Physiological Behaviors. Researchers have categorized this technique as the most accurate method to detect fatigue because it is based on physiological measures such as heart rate, eye movements, respiration, and brain waves. This technique is effective when the electrical activity of the brain and body muscles is recorded. These parameters are collected from various sensors placed on the driver's body or embedded in the car. The driver usually wears a wristband to measure heart rate and a helmet or special contact lenses to monitor eye or eye movements. Despite its effectiveness, the main drawback of this technique is intrusive because it requires attaching electrodes to the driver's body, which causes the driver discomfort.

2.2.2. Indirect Vehicle Behaviors. This technique requires a significant amount of time to analyze user behavior. It includes lateral position, steering wheel movements, and time crossings, which indicate the driver's vigilance and fatigue level. It is categorized as a nonintrusive technique, but it has several limitations: vehicle type, driver experience, engineering characteristics, and road condition.

2.2.3. Directly Observable Visual Behaviors. People with fatigue show many observable behaviors usually observed in facial features, such as eye movement, head movement, and facial expression. The technique is based on different typical visual characteristics that are detected from the captured image for a drive. The captured image includes parameters, such as a lower level of deflection, longer luminescence, slow eyelid movement, a smaller degree of eye-opening, eye closure, repetitive gestures, yawning, and slow motion [11].

2.3. Related Work. Various researchers use different methods and algorithms to measure driver fatigue. Anitha [13] proposed a novel twofold yawning detection system based on an expert system. In the first part of the system, they used the face detection algorithm's skin tone detection and defined the boundaries of the face; then, the blob dimensions for the mouth in the face containment are extracted. The system verified the yawning through a histogram of blobs whicj is taken from the vertical projection of the lower face part. If the histograms values are satisfying with the threshold values, then yawning is confirmed. The proposed system achieved 94% performance for yawning detection.

Kurylyak et al. [14] proposed an efficient approach to detect driver drowsy based on eye blinking. They used a web camera to acquire the driver image as input to the classifier using the Viola–Jones algorithm with Haar-like features to detect the driver's face and extract the eye region. A Kalman filter works with a set of discrete-time equations to compute and track the changes of the eye state. They calculate the frame's mutation for the thresholding value to detect the eyes' closure and opening. The frame processing algorithm is pointed out to distinguish the involuntary blinks from the voluntary ones. Experimental results of this proposed system presented 94% system accuracy to detect and determine the state of the eye.

In another work by Jo et al. [15], researchers used the same method of Kurylyak et al. [14] to detect driver face, while they proposed a new eye drowsy detection method that combined two methods. Principal component analysis (PCA) is used to detect the eyes status in the daytime and linear discriminant analysis (LDA) is used in the night. They applied support vector machine (SVM) to classify the eye states to open and close through a specific interval of 3 minutes. Experimental results of this proposed system showed that 99% of the design work accurately to detect eyes drowsiness and driver distraction. However, the systems fail to recognize the eye in various high illuminations.

Abtahi et al. [16] proposed a new method of yawning detection based on the changes in the mouth geometric features and eye movement. They used color statistics for detecting skin color and texture. Therefore, they improved detection efficiency by using bounding rules for different color spaces (RGB, YCbCr, and HSV). They experimented on more than 500 images with varying reflections of light, skin color, haircuts, beards, and eyeglasses.

Danisman et al. [17] proposed an automatic drowsy driver monitoring system to detect the eye blink duration. The proposed algorithm can catch the eye blinks' movement in real time using a webcam. Initially, they recognized the driver's face using the Viola–Jones algorithm, which is available in the OpenCV library of Python. Then, they discovered the positions of the pupils by using the symmetry property of the eye detector. The main drawbacks of that system are the presence of glasses and the various high illuminations, which affect the calculation and detection of the driver's drowsiness level. The proposed method achieved a 94% accuracy and a 1% false rate.

Bergasa et al. [18] proposed a nonintrusive computer vision system for tracking driver's vigilance in real time. The proposed method tried to test six parameters: face position and eye (eye rate, closure duration, blinks frequency, nodding frequency, and position of gaze). They used the fuzzy logic approach to combine this feature and determined the driver drowsiness level. The system was an experiment in different driving environments (night and day) with other users. The system achieved 100% accuracy at night; however, the system did not work with glasses and bright days.

Jie et al. [19] proposed new a spontaneous dataset of driver yawning in different simulated driving scenarios conditions. They present three labelling of different cases related to yawning, namely, speaking and mouth (covered or uncovered). HOGs and LBPs are popular algorithms for describing appearance in computer vision and image processing that has been used successfully in order to detect the driver yawning. These algorithms work based on intensity gradients or edge directions of the image, where it counts the pixel in the grayscale image number of oriented gradient occurrences in a dense grid of uniformly spaced cells. These occurrences are represented as a histogram for each cell normalized in a larger block area and show the mouth states.

Tipprasert et al. [20] proposed a method to detect the driver's eyes closure and yawn for drowsiness analysis by an infrared camera. The camera can work in low light condition processing by MATLAB R2015a. They obtain a 7.5% error in yawning detection because some driver opened their mouth too much, and the camera could not capture the entire driver's face.

Al-sudani et al. [21] proposed a yawning-based fatigue prediction method that monitors driver drowsiness levels. They used a camera inside the car to record driving scenarios (yawning or nonyawning driver). They built a deep CNN model to classify the drivers' fatigue into three levels, alert, early, and fatigue. Experiments are conducted using the YAWDD dataset, achieving 96.2% accuracy.

3. Proposed Approach

This study displays a nonintrusive real-time drowsiness system based on the webcam video analysis. This section shows the detection algorithm used to detect the driver's drowsiness level by investigating and analyzing the different visual cues of the driver. The main two parameters are eye blink duration and mouth state information. Figure 1 presents a block diagram of the proposed monitoring system of the drowsy driver. Develop a fuzzy controller that helps to determine the driver state and issue a suitable alert sound. The detection and monitoring approach consists of six stages as follows:

(1) Image acquisition
(2) Face detection and tracking
(3) Eye iris detection and tracking
(4) Mouth detection
(5) Mouth and eye information analysis
(6) Analysis driver state

3.1. Image Acquisition. It provides images of a driver's face from the recorded video to observe and gather the visual cues and then determine the fatigue level. The MatlabR2016a environment provides an image acquisition toolbox that enables the user to connect to the scientific cameras. The proposed approach used a webcam tool that installs from support hardware properties in MATLAB to create a webcam object and snapshot function to acquire images in-stream video and convert them to the frame. Then, we manipulated the webcam object properties to be efficiently correlated with the HP laptop webcam. The webcam object-specific properties shown in Table 1 are used for HP webcam.

3.2. Face Detection. Face detection is a computer technology that helps enhance human facial features in a digital image taken as input and used in various applications. This helps in processing the location and size of the human face and avoiding other objects [22]. There are many existing algorithms or methods for face detection technology, but the main difference is detection speed, accuracy, and purpose of use. Face detection algorithms work reasonably well with the detection of frontal and bright enough human faces images. It returns a sequence of analogous image coordinates where the human face is located and matched bit by bit. The proposal application assumes that the given input video detects only a single face (driver vehicle) in the camera view; otherwise, if there is more than one face, the system will detect the closest face to the center of the frame, as shown in Figure 2.

Implementing a face detection task is normalized to reduce and narrow down the domain of seeking the pupil and mouth detection. The eyes and mouth detector will not work if the desired face is not detected enough in the frame. However, pupil and mouth detection are located in the face area if the face is detected successfully by the face detector. After a comprehensive analysis, they suggest Viola and Jones's real-time detection technology, as presented in Figure 2. The Viola–Jones algorithm passes the image through 4 steps to detect the driver's face objects. We modified Viola and Jones's algorithm to detect the desired facial feature in a specific frame of the given video sequence of the vehicle driver instead of a static input image, as shown in Figure 3. After implementing the modified approach that extracted the driver image from the video and converted it to a binary bit, Figure 4 is obtained.

Step 1 (Haar feature method): initially, the desired face scanned with the Haar feature method, which contains scalar values representing two rectangles that can be horizontal or vertical in input image resolution 24 × 24 pixels with a possible number of rectangles' features 160,000.

For example, the area where the eyes and mouth are located then passed this feature as an argument to the calculating [23], where I is an image, P is a pattern, and $N \times N$ = size:

$$\begin{aligned}&\sum_{1\le i\le N}\ \sum_{1\le jN} I\left(i,\ j\right)1p\left(i,j\right)\text{ is white}\\&\quad-\sum_{1\le i\le N}\ \sum_{1\le jN} I\left(i,\ j\right)1p\left(i,j\right)\text{ is black.}\end{aligned}\tag{1}$$

Step 2 (integral image): the system calculates the Integral image and defines either the image contains the face or not at a very low computational cost using cumulative distribution functions:

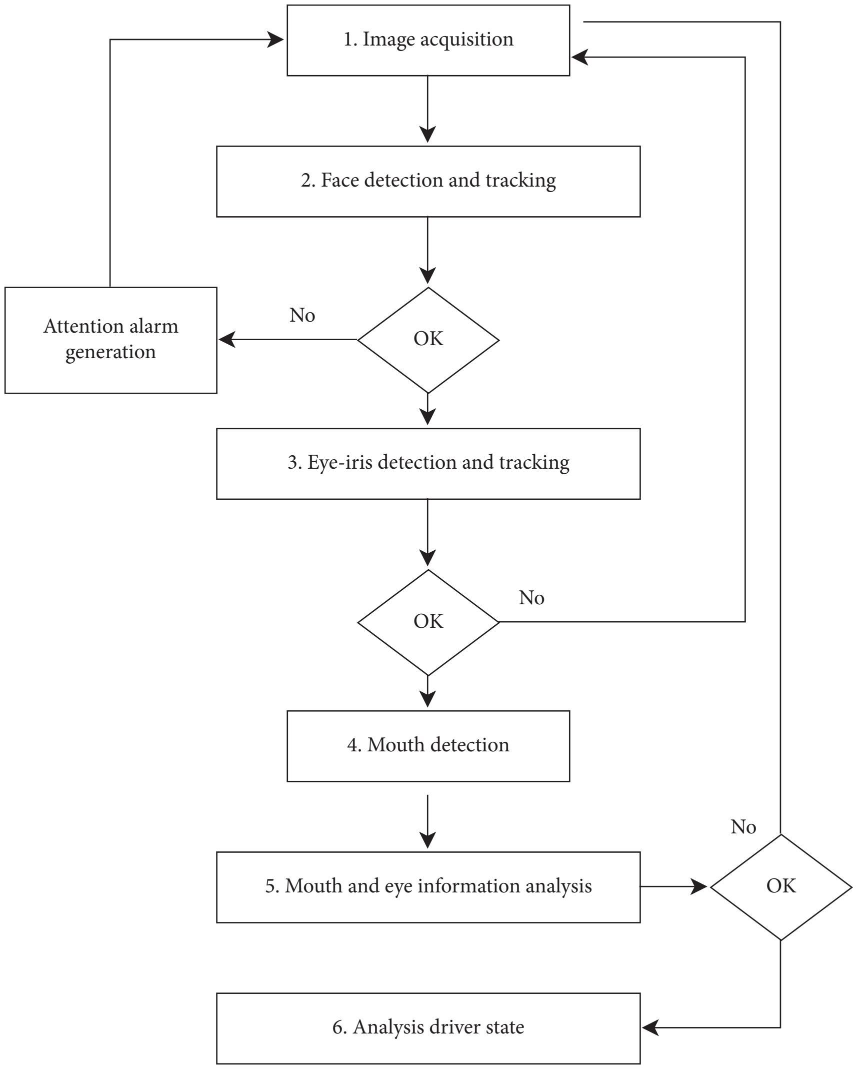

1. Image acquisition
2. Face detection and tracking
Attention alarm generation
No
OK
3. Eye-iris detection and tracking
OK
No
4. Mouth detection
No
5. Mouth and eye information analysis
OK
6. Analysis driver state

Figure 1: Block diagram of the proposed approach.

Table 1: Webcam object-specific properties.

Webcam Name:	HP Truevision HD
Quality rating	816
Frame rate	30 FPS
Stream type	Video
Image mode	rgb
Webcam megapixels	0.92 MP
Webcam resolution	1280 × 720
Video standard	HD
Number of colors	205668
Lightness	45.49%
Brightness	45.88%
Saturation	4.31%

Algorithm: Viola-Jones Face Detection Algorithm

```
Input: original test image
Output: image with face indicators as rectangles
for $i \leftarrow 1$ to num of scales in pyramid of images do
    Downsample image to create image$_i$
    Compute integral image, image$_{ii}$
    for $j \leftarrow 1$ to num of shift steps of sub-window do
        for $k \leftarrow 1$ to num of stages in cascade classifier do
            for $l \leftarrow 1$ to num of filters of stage $k$ do
                Filter detection sub-window
                Accumulate filter outputs
            end for
            if accumulation fails per-stage threshold then
                Reject sub-window as face
                Break this k for loop
            end if
        end for
        if sub-window passed all per-stage checks then
            Accept this sub-window as a face
        end if
    end for
end for
```

Figure 2: Viola–Jones detection algorithm.

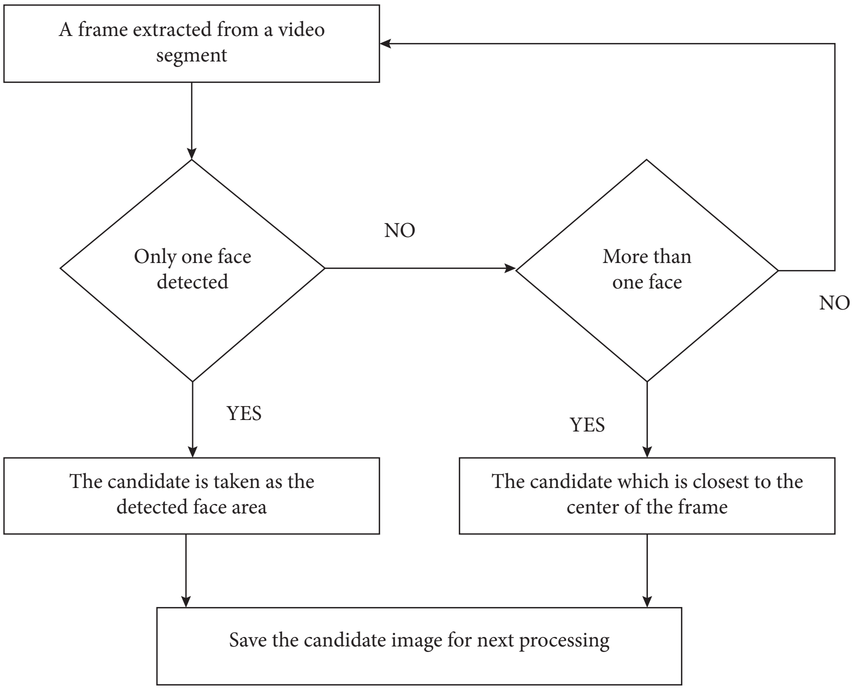
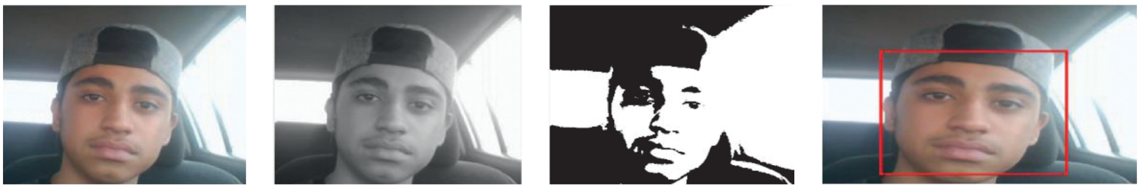

A frame extracted from a video segment
Only one face detected
NO
More than one face
NO
YES
YES
The candidate is taken as the detected face area
The candidate which is closest to the center of the frame
Save the candidate image for next processing

Figure 3: Flow diagram of the face detector.



(a) (b) (c) (d)

Figure 4: (a) Original image, (b) converted to gray image, (c) converted to binary image, and (d) face detection.

$$II(i,j) := \begin{cases} \sum_{1\le s\le i} \sum_{1\le t\le j} I(s,t), & 1 \le i \le N \text{ and } 1 \le j \le N, \\ 0, \quad \text{otherwise} & \end{cases} \tag{2}$$

$$\begin{aligned} \sum_{N_1\le i\le N_2} \sum_{N_3\le j\le N_4} I(i,j) &= II(N_2,N_4) - II(N_2,N_3-1) - II(N_1-1,N_3-1), \\ \sum_{N_1\le i\le N_2} \sum_{N_3\le j\le N_4} I(i,j) &= II(N_2,N_4) - II(N_2,N_3-1). \end{aligned} \tag{3}$$

Step 3 (feature's selection Adaboost): this technique will remove all irrelevant features and combine only relevant features with their weight to evaluate and deciding either the image contains a face (1) or not (−1), where (X, Y) is a training example to the probability P, weights $w_i(1)$, $1 \le i \le n$, and ht = decision stump. As the empirical loss goes to zero with T, so do both false positive $P(f^{T}(X) \ne 1 \mid Y = -1)$ and false negative rates $P(f^{T}(X) \ne 1 \mid Y = 1)$:

$$\sum_{i=1}^{n} w_i(1) 1_{y_i \sum_{t=1}^{T} \propto_t h_t(x_i) \le 0} := P\left(f^{T}(X) \ne Y\right). \tag{4}$$

Step 4 (cascade method): this method increases processing power features by distributed every 10 features among single stages and subwindows. Each subwindow will evaluate according to its feature by the cascading method. In Figure 5, the classifier triggers the evaluation and checks the characteristics of each subwindow. If the subwindow is classified as positive (face), it will be passing through the steps. In the other case, the negative subwindow (not face) will immediately reject. This method will increase the performance power of the detected face by removing nonface-related windows at the beginning. Equation (5) defines the cascade decision rule to obtain empirical loss:

$$f\text{cascade}(X) = 2\left(\prod_{p=1}^{i} 1_{f_{p,s_p}^{T_p}(X)=1} - \frac{1}{2}\right), \tag{5}$$

where $f(^{tp})$ is a classifier with false positive.

Face tracking is handling by the Kalman filter. The Kalman filter method is an efficient way to estimates the position of a moving object based on its historical values in the next time frame. It can predict the state of a dynamic system from a sequence of uncertainty measurements by using a recursive adaptive filter. In addition, the Kalman filter was implemented to predict the dynamic change rate of the moving object and reduce the location error [24]. The following equation is used to track and predict the face position, where x = target position, $x0$ = initial position, $v0$ = initial velocity, a = target acceleration, and Δ is the time interval (3 seconds in this example):

$$x = x_0 + v_0 \Delta t + \frac{1}{2} a \Delta t^2. \tag{6}$$

The face tracking method uses a particle filter based on face position, face speed, and location error. However, this method fails when the brightness of the face is not enough due to background illumination and sudden move of the head. In Figures 5(a) and 5(b), face tracking is trained to track the front face image and its maximum rotation between ±15 to ±50 degrees. However, in Figure 5(c), the detector fails when the rotation is more than this and the alarm will rise.

3.3. Eye Iris Detection and Tracking. The iris is a significant parameter that can be considered to assess the fatigue level of a driver. Tracking the eyelid and eye movement can reveal the size of the iris. Therefore, it is easy to determine the eye state through geometric calculations. As shown in Figure 4, in appropriate circumstances, the system first detects the driver's face and then segments the upper half of the face to identify the eyes. Once this region has been detected in a cropped image, it can help the system reduce the computing cost of the drowsiness levels. The cropped image is then converted into a binary image using the adaptive threshold technique to detect and replace all the white pixels in the input image with the value 1 and black pixels with the value 0. Following this, we use image processing, which provides morphological operations to process images based on shapes. The morphology technique only works with the relative ordering of pixel values; the value of each pixel in the picture is modified based on its neighboring pixel in the input image. To successfully perform a morphological operation, the size and shape of the neighborhood image need to be specified.

Morphological image processing or calling a flat structuring element allows using functions such as segmentation, skeletonization, thinning, erosion, dilation, external boundary, and internal boundary. The system darkens the eye region by identifying the steel object to create a flat structuring element. This step is essential to eliminate false pixels and compute valid pixels. The eye detector focuses on the threshold and the rotation of the eye. Figure 6 illustrates a flat structuring element to detect the eyes. The iris is tracked successfully using a Kalman filter. This method uses a particle filter based on the iris' position, speed, and location error. However, this method fails when the brightness of the eye iris is not enough due to background illumination and sudden move of the head or camera resolutions. The eye size is different from one person to another; the system assumed at the beginning that the user is awake. Then, the threshold will be calculated and compared with the current eye ratio in the video frame. The system cannot predicate eye iris positions correctly due to sudden movement of the head or camera resolutions.

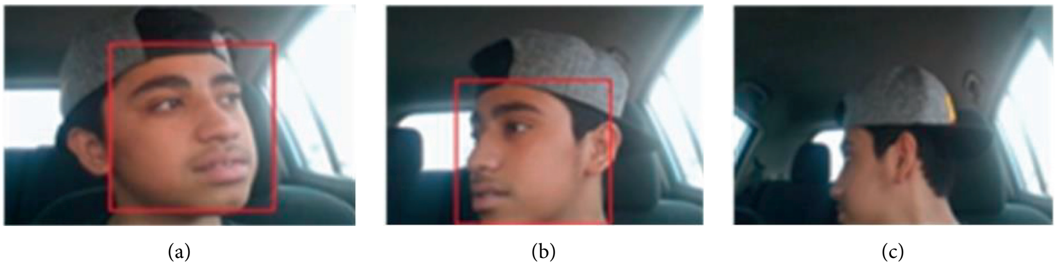
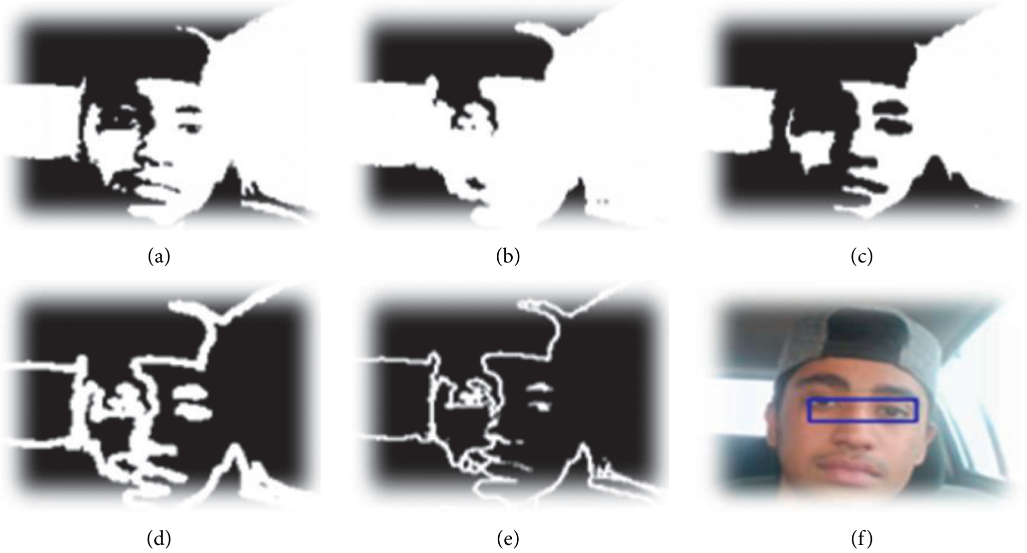

(a) (b) (c)

Figure 5: Face tracking. (a) Face detected in 30°. (b) Face detected in 45°. (c) Face undetected.


(a) (b) (c)

(d) (e) (f)

Figure 6: Flat structuring element to detect the eye region. (a) Original image. (b) Dilated image. (c) Eroded image. (d) Internal boundary. (e) External boundary. (f) Eyes detected.

3.4. Mouth Detection. Yawning is the most common sign of tiredness and drowsiness; hence, the system detects the driver's facial and eye movements. In the preprocessing stage, the system will crop the mouth region frame to determine the driver's state. Since the mouth is located in the lower part of the face, the mouth detectors will extract the mouth directly from the lower part of the face. After that, the system verifies mouth location using eye distance to detect the correct mouth segments. This is resolved by checking the boundaries of the mouth and eyes. Next, the system detects the mouth by calculating the connected object. The program will detect the upper and lower lips as two objects, assuming the mouth is open. When the mouth is closed, the program will calculate it as one object value; otherwise, the mouth would not be detected, and the program would return to zero value, as depicted in Figure 7. The monitored information will then pass to the fuzzy logic method to provide the driver's fatigue level.

3.5. Mouth and Eye Information Analysis. After successfully detecting and tracking the facial features of the mouth and eyes, the system will compute the mouth and eye states by defining the two input parameters, which are the total number of black pixel areas and the ratio of black pixels compared to the ratio of white pixels.

Number of white pixels = sum (binary image).

Number of black pixels = sum (binary image) – number of white pixels.

Ratio = number of black pixels/numbers of white pixels.

The system uses the rules base to define the states of the eye and mouth. First, starting with the eye state, the system will compare the eye detection frame with the eye threshold to calculate the correct eye size ratio. Then, it will check if the ratio of the left eye pixel is more than the threshold; the eye state is considered open or closed. If the eyes are detected to be closed for more than three seconds, an alarm will sound. Similarly, the mouth state will be defined by checking the detected lips frame against the threshold to determine if the mouth is open or closed, as shown in Figure 8.

3.6. Develop a Fuzzy Model for Analysis Driver Behavior. Fuzzy logic is a problem-solving algorithm that resembles human decision-making to provide a solution to a problem from vague or uncertain data. Fuzzy logic can describe fuzziness by representing a membership function and classifying the degree of truth of each element in a fuzzy set.

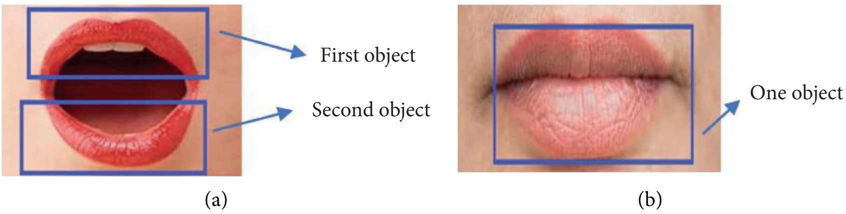


Figure 7: Mouth detection. (a) Open mouth. (b) Close mouth.

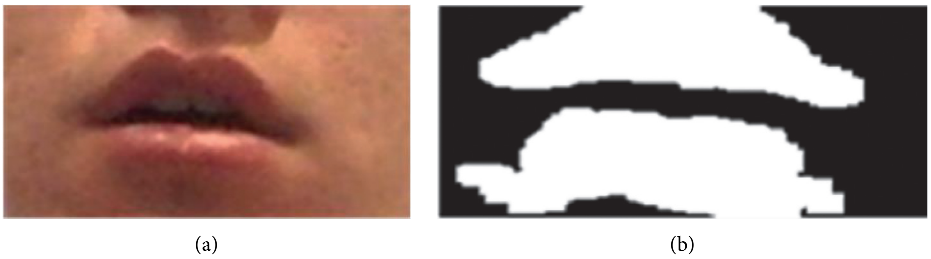


Figure 8: Mouth processing. (a) Open mouth sample. (b) Flat structuring element.

A membership function is used to present a graphical representation of the fuzzy set based on the principle of the fuzzy rule IF-THEN [25, 26]. Drowsiness is a type of fuzzy bodily state which is difficult to quantify objectively.

Therefore, developing a fuzzy model can be an easy way to analyze driver behavior and determine their level of fatigue. In this system, the fuzzy inference system involves three main steps:

Step 5 (fuzzification): define the two crisp input variables which are eye and mouth states with one output which is at the drowsy level, as shown in Figure 9.

Initially, we define the first input variable in which eye state is according to the ratio of eye closure as illustrated in Table 2. Similarly, for mouth state, we define the input variable according to the ratio of open, half open, and close. A drowsy level is defined by three terms: low, medium, and high. The inputs include eye state (open-half and open-close) and mouth state (open-half and open-close) values. The linguistic variables and terms for the inputs and outputs are given in Table 3.

Furthermore, researchers use fuzzy logic to describe the fuzziness of the variables by representing the values from 0 to 1 using a set of input membership functions of both inputs and the degree of truth of each element in a fuzzy set is classified as presented in Figures 10(a) and 10(b).

Step 6 (inference system): the if-then rules will be defined and evaluated in rule editor part of the fuzzy inference system. From the video frame sequence, the system can monitor the degree of the open/close eye and mouth frame. Then, the system uses knowledge rule-based fuzzy inference system and combines the two variables, which are eye and mouth state values. This process will help the system to classify the feeling of the driver, whether it has low, medium, or high drowsy, and proposed fuzzy rules based on the ratio and threshold of open and close eyes. This is a set of rules for the knowledge base using IF-THEN logic defining in rule editor, as presented in Table 4.

Step 7 (de-fuzzification)

It converts the fuzzy input to the crisp value by using the output membership functions to determine the drowsy output level, as shown in Table 4. Use the rule viewer of MATLAB to interpret the entire fuzzy inference process at once and show the diagram of membership functions which influences the overall result. Figure 11 shows the rules' viewers of nine rows of plots rule and three columns of the input and output variables.

The first two columns of plots in yellow show the membership functions of the two input variables, and the last column represents the output membership function. The last plot in the last column shows the aggregate weighted decision for the given fuzzy rule system, which depends on the input values for the system.

Figure 12 presents the surface viewer of MATLAB three-dimensional curve of the entire fuzzy inference system. It is equipped with X (eye state input), Y (mouth state input), and Z (drowsy level output) to allow the calculation time reasonable for complex problems. Also, the IF-THEN rules appear in 9 colors as displayed in the diagram of the fuzzy module; if the eye state is closed at 0.3 degrees and the mouth state is opened at 0.9 degrees, then the drowsy level will be high.

4. Experimental Results

4.1. Offline Data Analysis and Training for Eye and Mouth. One of the principal aims of the project was to conduct a nonintrusive real-time drowsiness system based on the webcam video analysis. The system is focused on different visual cues of the driver to collect data and detect the driver's

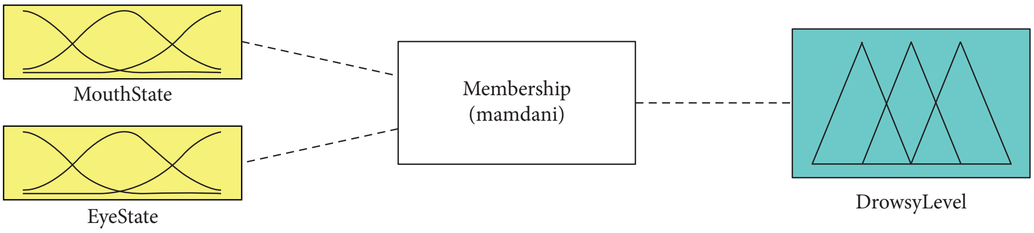


Figure 9: FIS editor showing the main simulation screen for driver drowsy.

Table 2: Eye closure.

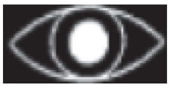
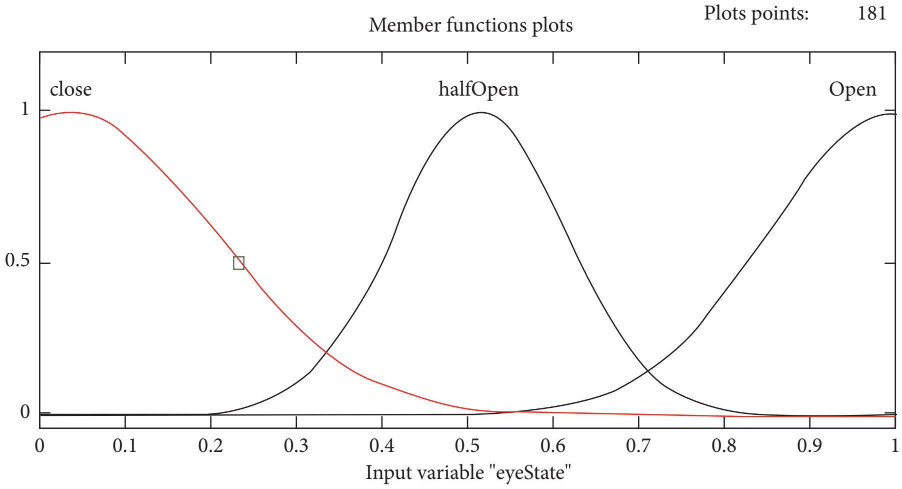
			
Ratio of eye closure	100–74%	73–35%	34–0%
Area of black pixels	Small	Medium	Large
Drowsy level	Low	Medium	High

Table 3: Linguistic variables and terms for the inputs and output.

Variable	Type	Fuzzy set	Membership function
Eye state	Input	Open, half open, and close	Trapezoidal
Moth state	Input	Open, half open, and close	Trapezoidal
Drowsy level	Output	Low, medium, and high	Trapezoidal



(a)

Figure 10: Continued.

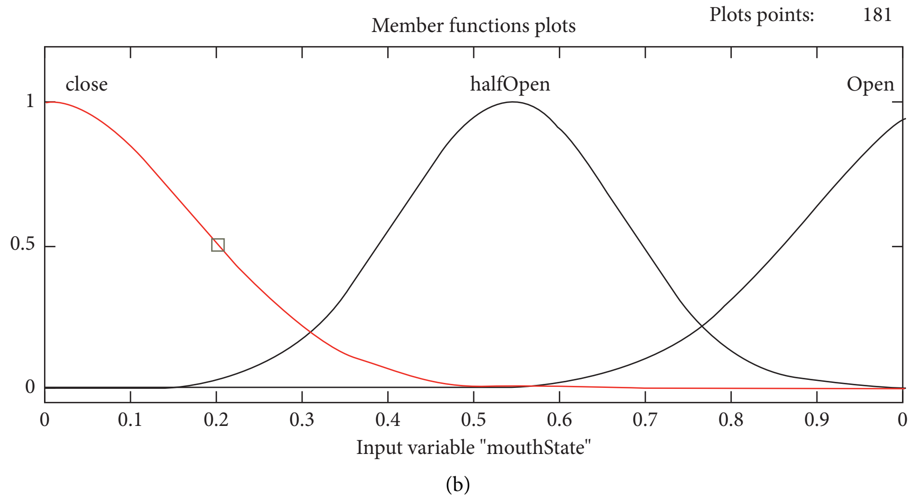


Figure 10: FIS editor showing the MF: (a) for eye state; (b) for mouth state.

Table 4: Rule editor.

Rule no.	Rule
Rule 1	IF eye state is open OR mouth state is close, THEN the drowsy level is low
Rule 2	IF eye state is close OR mouth state is open, THEN the drowsy level is high
Rule 3	IF eye state is half open OR mouth state is half open, THEN the drowsy level is high
Rule 4	IF eye state is open AND mouth state is half open, THEN the drowsy level is low
Rule 5	IF eye state is close AND mouth state is close, THEN the drowsy level is high
Rule 6	IF eye state is close AND mouth state is half open, THEN the drowsy level is high
Rule 7	IF eye state is half open AND mouth state is open, THEN the drowsy level is high
Rule 8	IF eye state is half open AND mouth state is close, THEN the drowsy level is medium
Rule 9	IF eye state is open AND mouth state is open, THEN the drowsy level is medium

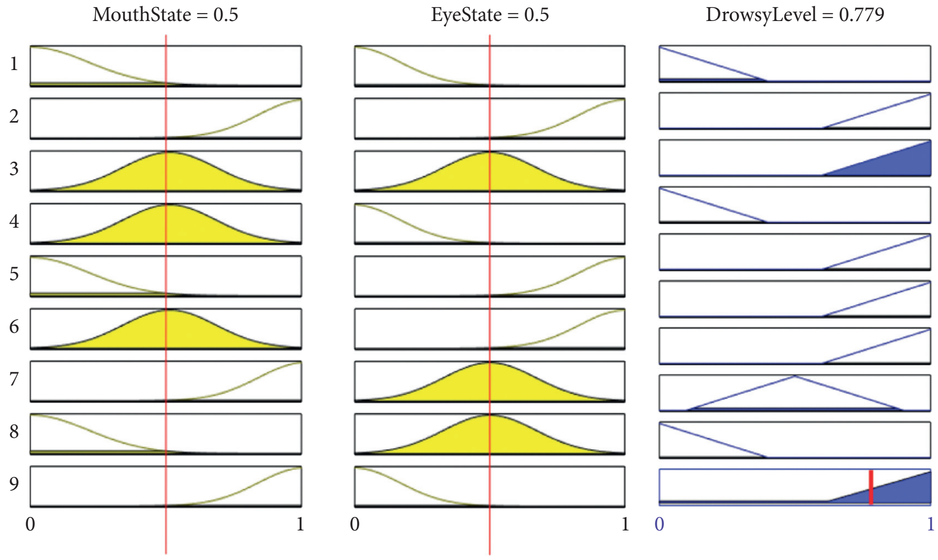


Figure 11: Rule's viewer.

level of alertness. The systems focused on two parameters used to identify the driver's status: eye status and mouth status information. Firstly, we conducted an offline test on the MRL eye dataset [27] of the open and closed eye, as presented in Figure 13. Similarly, for mouth training, the OuluVS2 dataset contains different people's pictures with different ages and gender and is captured from different angles. The main objective of using the dataset is to figure out the threshold of open and closed eyes and to test the system's capability for detecting the eye state on different people or

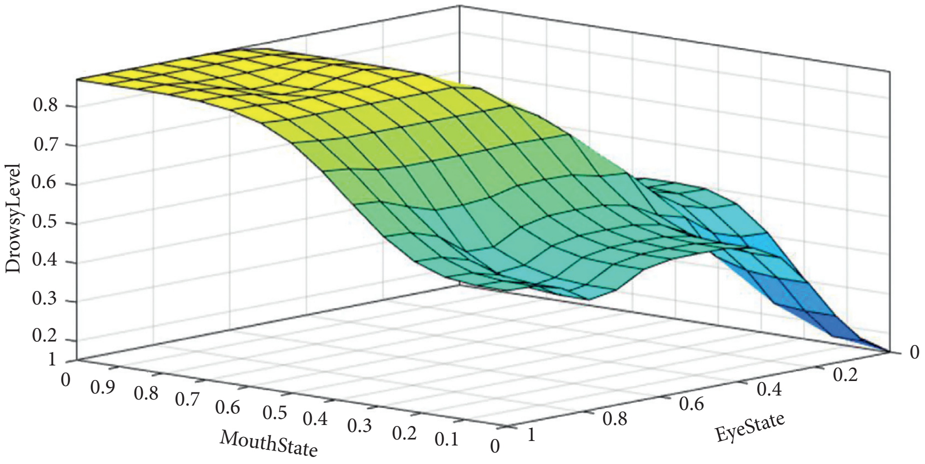


Figure 12: Surface viewer of fuzzy inference system.

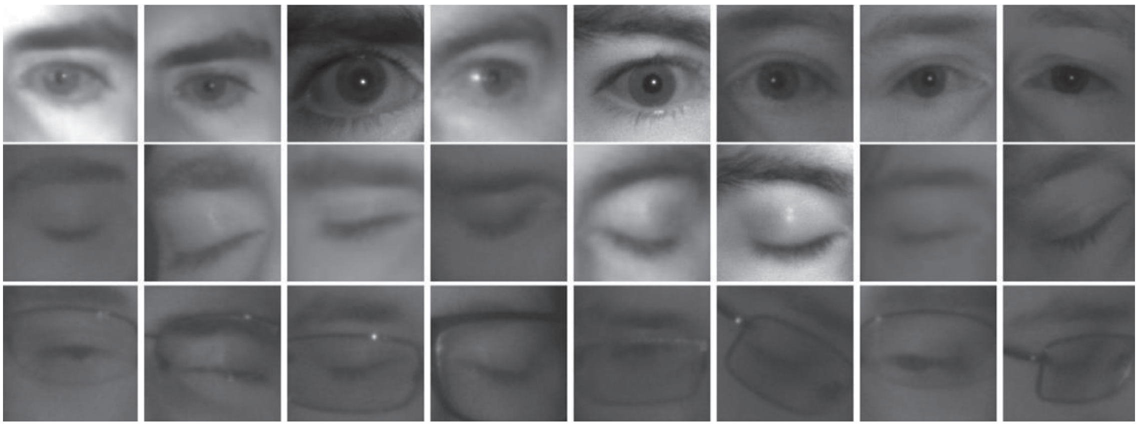


Figure 13: MRL eye dataset samples.

not. By testing different eye samples, we figured out that if the eye image ratio exceeds the threshold, the eye will be considered open, otherwise closed. Also, the results show that the program can detect all eyes in most pictures except the pictures with high reflection and lousy quality.

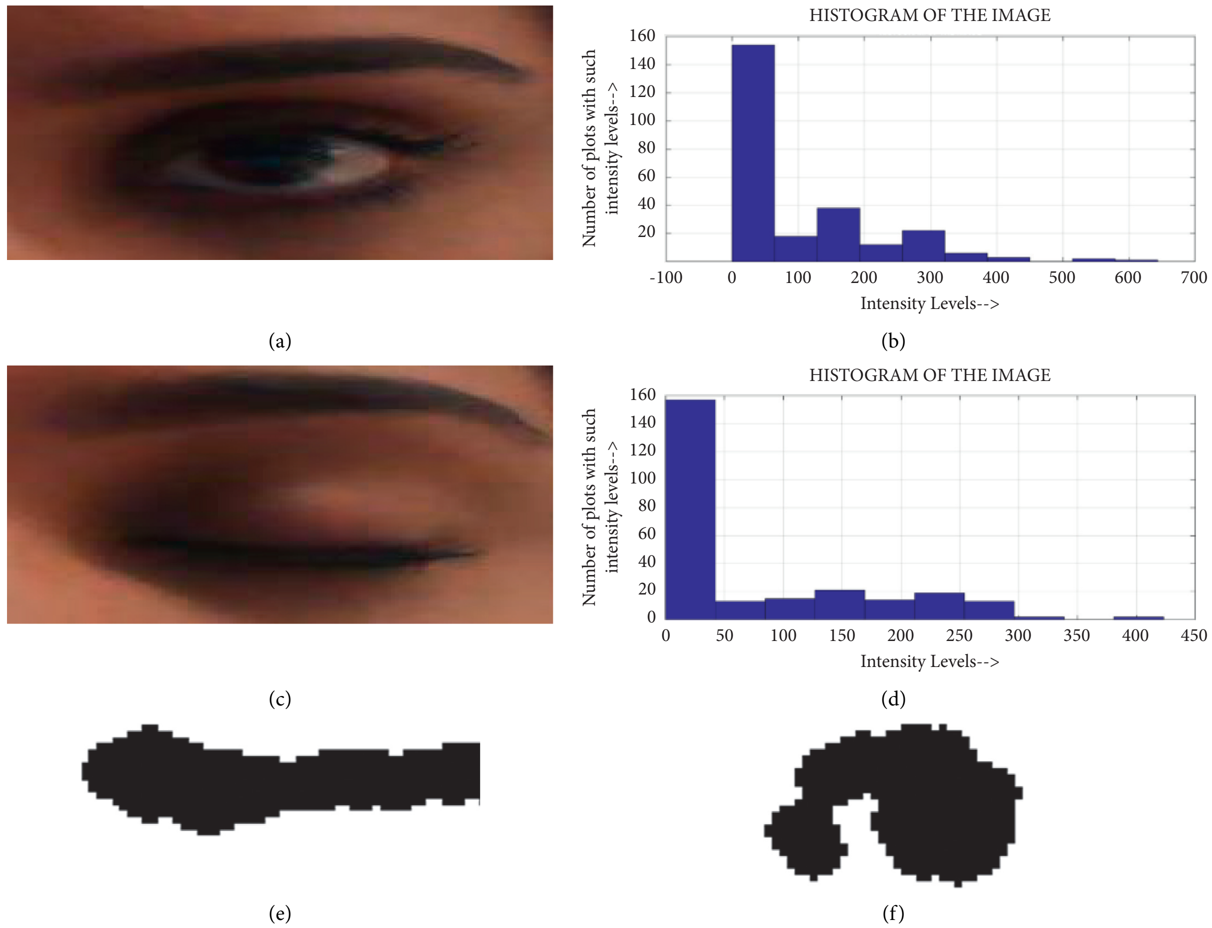
The system achieves promising results on detection eye status. Hence, it can conclude as results that the system can recognize the eye status successfully for both genders and wearing glasses or not. Besides, the system works effectively in case of offline detection due to the stability of the image. The error rate can be estimated due to the reflection and bad light of the pictures, see Table 5. However, the programs achieve 100% accuracy on offline detection mouth status. According to our results, the system can easily detect the mouth state from the pictures because no external obstacles affect how the system works, such as the sudden move, the distance between driver and camera, camera resolution, and background light.

Eye state visualization by using a histogram chart for open and close eye. MatlabR2016a environment provides an imhist function to represent an image histogram chart to show the distribution of information density in the grayscale image. It demonstrates the number of times where the density value in the image occurs. The digital image is a grayscale image that includes some pixels with one scalar value called intensity. Therefore, the number of intensity levels refers to encoded images with $28 = 256$ intensity values, where 0 displays the black pixel and 255 display the white. The cropped of the open and close eye image is converted to a binary image by using an adaptive threshold to detect and compute the black and white pixel of the picture. Then, identify the steel object to create a flat structuring element essential for dilation and erosion in the binary images to eliminate false pixels and compute valid pixels. First, the system will crop open eye region from the sample face, as displayed in Figure 14(a). Then, to visualize the open eye in the histogram chart, the system needs to binaries the image to white and black pixels. The larger eye is opened based on the pupil; the darker pixels are found in the picture, as shown in Figure 14(b). By using binary image and imhist function, the histogram chart is presented in Figure 14(c).

The histogram chart shows the distribution of the information density of the binary image and the threshold value, by calculating the ratio of black pixels (pupil) to white pixels (skin). Therefore, the system can determine whether the eye is close or open. According to the above histogram chart, the high distribution of the open eye image density is

Table 5: Offline MRL eye dataset analysis.

Subject	Gender	Glasses	Close	Open	No reflection	Low reflection	High reflection	Bad light	Good light	Accuracy of detection close eye (%)	Accuracy of detection open eye (%)
s0001	Male	555	406	361	267	30	470	523	244	82	80.7
s0029	Male	0	177	223	390	10	0	163	237	95.9	96.8
s0016	Female	0	455	173	417	183	28	120	508	89.3	90.1
s0036	Female	522	209	409	420	73	125	450	168	85	84



(a)

(b)

(c)

(d)

(e)

(f)

Figure 14: Results of implementation. (a) Sample of open eye. (b) Open eye sample flat structuring element. (c) Histogram of open eye. (d) Histogram of close eye. (e) Open eye sample flat structuring element. (f) Sample of close eye.

between 0 and 400, and then, the level of image density is gradually decreased. The histogram shows a peak of the dark pixels at around 155 when the density level is between 0 and 50. Similarly, for a close eye, the system will crop close eye area from the sample face, as displayed in Figure 14(d). Then, convert the cropped image to a binary image to visualize it in the histogram chart, as shown in Figure 14(e). According to the histogram of the close eye, shown density distribution of the close eye is between 0 and 300, and then, the level of image density is gradually decreased. The histogram shows a peak of the dark pixels at around 155 when the density level is between 0 and 50. The number of dark pixels slightly increased increasing to 20 in 150 and 250 intensity level with such intensity levels in Figure 14(f).

4.2. Online Data Analysis and Training for the Eye and the Mouth. In this section, the system focused on the analysis of real-time driver drowsiness detection. We conduct an experiment for training and testing on the 7 YaWDD videos' dataset [27] under different conditions such as gender, age, wearing glasses and, illumination conditions. Consecutively, Table 6 presents 7 video samples of monitoring drivers' behavior, where the acquired videos resulted in a resolution of 1280 × 720 of 30 frames per second. These videos were tested in 3 different situations: normal driving (without speaking) and speaking and yawning while driving. The experience indicates that the system could detect eye status for people who wear glasses and those who do not.

Table 6: Online YAWDD videos' dataset analysis.

7 YAWDD (videos)	Face detection (%)	Eye status (%)	Mouth status (%)	Yawning detection (%)
Detection accuracy	100	94	95	94.5

Table 7: Online YAWDD videos' dataset MRL eye image dataset analysis.

	YAWDD (videos)	MRL eye (images)
Total sample number	7	2413
Males	1	2
Females	2	2
Conditions' test	Normal driving (no talking) and talking and yawning while driving	Eye status (open/close) with and without glasses
Detection accuracy	95%	87.875%

On observation of the experiments on eye state detection, it was found that some cases of the eye state cannot be detected. However, these undetected eyes belong to the people who wear glasses. We can estimate the error due to the reflections of background light on the glasses' glass, which negatively affect the camera. Therefore, the eye detector is unable to determine the eye status. The programs achieve 94.56 for online data analysis and training for eye and mouth status, as shown in Table 7 [28].

5. Conclusions

This study proposes a real-time drowsy detection system that monitors the eye and mouth of the driver through driving and issue the suitable alert sound. The system was conducted under different experiments for detecting eye and mouth states with different people and conditions. It achieved high accuracy compared to other works, about 87.875% for eye detection and 100% for mouth detection. Also, the system was tested in real time with other videos recorded in a day and night, with different people wearing and not wearing glasses. The proposed method achieved 94.5% accuracy on real-time detection driver status. It also developed a fuzzy model for a mouth and eye variable input based on defined and evaluated IF-THEN rules to investigate driver fatigue levels. This control system is efficient in determining the driver status and issuing the needed alert. It is easy to modify and update based on new cases and factors. Finally, it visualizes the eye state using a histogram chart to check and examine the difference between open and closed eyes.

The contributions of this work are

(a) Optimize the current detection method (Viola–Jones) by adding new features that help fast and accurately detect the driver's face and mouth

(b) Develop a fuzzy logic controller that can determine the driver status quickly and efficiently issued the needed alert

(c) The proposed approach can work in offline and online systems and embed easily with any framework

(d) The proposed detecting approach includes several new features that enable it to detect the driver's face and mouth in different conditions and a light contrast

(e) The proposed approach effectively detects the driver images and captures the driver image from a video

6. Future Work

Driver drowsy is a major cause of road accidents and economic risk. Using a Webcam tool to detect fatigue is still not efficient enough because the face cannot be detected when the driver moves their head quickly or suddenly. Also, facial features cannot be discovered in the video frame under various lighting conditions or while wearing sunglasses, which gives inaccurate data in the fatigue warning system. Hence, in the future, the system should use infrared camera-sensitive camera to allow the system work robustly in any lighting conditions. The proposed approach cannot detect the driver's mouth because a typically yawning gesture covers the mouth while yawning. Therefore, in the future, the system should include other parameters which are steering wheel angles. The steering wheel angles' sensor is widely used to measure the driver behavior and determine the fatigue level in real time. The system should improve the response time to get accurate results and avoid any wrong alarms of the drowsy detection system.

References

[1] C. Peura, J. A. Kilch, and D. E. Clark, "Evaluating adverse rural crash outcomes using the NHTSA State Data System," *Accident Analysis & Prevention*, vol. 82, pp. 257–262, 2015.

[2] K. Anjali, A. K. Thampi, A. Vijayaraman, M. F. Francis, N. J. James, and B. K. Rajan, "Real-time nonintrusive monitoring and detection of eye blinking in view of accident prevention due to drowsiness," in *Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1–6, IEEE, Nagarcoil, India, March 2016.

[3] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, 2015.

[4] S. Mehta, S. Dadhich, S. Gumber, and A. Jadhav Bhatt, "Real-time driver drowsiness detection system using eye aspect ratio and eye closure ratio," in *Proceedings of the International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM)*, Amity University Rajasthan, Jaipur-India, June 2019.

[5] A. Madhan Kumar, M. Kabilan, S. Karthikeyan, and S. Sathiyapriya, "Drowsy driving warning and traffic collision information system using iot," *IJARIIT*, vol. 5, 2019.

[6] E. P. Herrera-Granda, J. A. Caraguay-Procel, P. D. Granda-Gudiño et al., "Drowsiness detection in drivers through real-time image processing of the human eye," in *Proceedings of the Asian Conference on Intelligent Information and Database Systems*, pp. 626–637, Springer, Yogyakarta, Indonesia, April 2019.

[7] A. B. Roig, M. Morales, J. Espinosa, J. Perez, D. Mas, and C. Illueca, "Pupil detection and tracking for analysis of fixational eye micromovements," *Optik*, vol. 123, no. 1, pp. 11–15, 2012.

[8] P. Shelke, S. Magar, and R. Jawkar, "Real time detection system of driver fatigue," *IJCERT*, vol. 3, 2016.

[9] T. Azim, M. A. Jaffar, and A. M. Mirza, "Automatic fatigue detection of drivers through pupil detection and yawning analysis," in *Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*, pp. 441–445, IEEE, Kaohsiung, Taiwan, December 2009.

[10] W. Liu, H. Sun, and W. Shen, "Driver fatigue detection through pupil detection and yawing analysis," in *Proceedings of the 2010 International Conference on Bioinformatics and Biomedical Technology*, pp. 404–407, IEEE, Chengdu, China, June 2010.

[11] T. Azim, M. A. Jaffar, and A. M. Mirza, "Fully automated real time fatigue detection of drivers through fuzzy expert systems," *Applied Soft Computing*, vol. 18, pp. 25–38, 2014.

[12] J. Berg, G. Neely, U. Wiklund, and U. Landstrom, "Heart rate variability during sedentary work and sleep in normal and sleep-deprived states," *Clinical Physiology and Functional Imaging*, vol. 25, no. 1, pp. 51–57, 2005.

[13] C. Anitha, M. K. Venkatesha, and B. S. Adiga, "A two fold expert system for yawning detection," *Procedia Computer Science*, vol. 92, pp. 63–71, 2016.

[14] Y. Kurylyak, F. Lamonaca, and G. Mirabelli, "Detection of the eye blinks for human's fatigue monitoring," in *Proceedings of the 2012 IEEE International Symposium on Medical Measurements and Applications*, pp. 1–4, IEEE, Lausanne, Switzerland, June 2012.

[15] J. Jo, S. J. Lee, J. Kim, H. G. Jung, and K. R. Park, "Vision-based method for detecting driver drowsiness and distraction in driver monitoring system," *Optical Engineering*, vol. 50, no. 12, Article ID 127202, 2011.

[16] S. Abtahi, B. Hariri, and S. Shirmohammadi, "Driver drowsiness monitoring based on yawning detection," in *Proceedings of the 2011 IEEE International Instrumentation and Measurement Technology Conference*, pp. 1–4, IEEE, Hangzhou, China, May 2011.

[17] T. Danisman, I. M. Bilasco, C. Djeraba, and N. Ihaddadene, "Drowsy driver detection system using eye blink patterns," in *Proceedings of the 2010 International Conference on Machine and Web Intelligence*, pp. 230–233, IEEE, Toronto, Canada, August 2010.

[18] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, 2006.

[19] W. Tipprasert, T. Charoenpong, C. Chianrabutra, and C. Sukjamsri, "A method of driver's eyes closure and yawning detection for drowsiness analysis by infrared camera," in *Proceedings of the 2019 First International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)*, pp. 61–64, IEEE, Bangkok, Thailand, 2019, January.

[20] A. R. Al-sudani, "Yawn based driver fatigue level prediction," *Proceedings of 35th International Confer*, vol. 69, pp. 372–382, 2020.

[21] Z. Jie, M. Mahmoud, Q. Stafford-Fraser, P. Robinson, E. Dias, and L. Skrypchuk, "Analysis of yawning behaviour in spontaneous expressions of drowsy drivers," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 571–576, IEEE, Xi'an, China, 2018, May.

[22] R. Boda, M. J. P. Priyadarsini, and J. Pemeena, "Face detection and tracking using KLT and Viola Jones," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, pp. 13472–13476, 2016.

[23] P. Gejguš and M. Šperka, "Face tracking for expressions simulations," in *Proceedings of the International Conference on Computer Systems and Technologies*, June 2003.

[24] R. R. Yager, "Expert systems using fuzzy logic," in *An Introduction to Fuzzy Logic Applications in Intelligent Systems*, pp. 27–44, Springer, Berlin, Germany, 1992.

[25] O. V. Komogortsev and J. I. Khan, "Kalman filtering in the design of eye-gaze-guided computer interfaces," in *Proceedings of the International Conference on Human-Computer Interaction*, pp. 679–689, Springer, Beijing, China, July 2007.

[26] J. Yousif and D. Saini, "Fuzzy and mathematical effort estimation models for web applications," *Applied computing Journal*, vol. 63, pp. 10–24, 2021.

[27] M. M. Rahman, M. S. Islam, M. K. A. Jannat et al., "EyeNet: an improved eye states classification system using convolutional neural network," in *Proceedings of the 2020 22nd International Conference on Advanced Communication Technology (ICACT).*, pp. 84–90, IEEE, PyeongChang, Korea, February 2020.

[28] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Y.D. D Hariri, "A yawning detection dataset," in *Proceedings of the 5th AC Multimedia Systems Conference*, pp. 24–28, Singapore, March 2014.

Theoretical Analysis of an Imprecise Prey-Predator Model with Harvesting and Optimal Control

Rashmita Panigrahi, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, rashmitapanigrahi116@gmail.com*

Arabinda Dash, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, arabindadash56@hotmail.com*

Nikunja Bihari Kar, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, nikunjabiharikar@gmail.com*

Priya Chandan Satpathy, *Department of Electrical and Communication Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar,priyachandan.satpathy57@outlook.com*

Abstract

In our present paper, we formulate and study a prey-predator system with imprecise values for the parameters. We also consider harvesting for both the prey and predator species. Then we describe the complex dynamics of the proposed model system including positivity and uniform boundedness of the system, and existence and stability criteria of various equilibrium points. Also the existence of bionomic equilibrium and optimal harvesting policy are thoroughly investigated. Some numerical simulations have been presented in support of theoretical works. Further the requirement of considering imprecise values for the set of model parameters is also highlighted.

1. Introduction

The eternal relationship between prey and predators is one of the major topics to be discussed in recent science. Scientists from various fields are currently engaged in finding out different interactions among prey populations and predator populations. With the help of mathematical modeling, one can describe the strong and competitive relationship between these two types of creatures. However the discussion on this fascinating topic, with the help of some mathematical tool, was started during the first quarter of the twentieth century, thanks to the age-breaking works of Lotka [1] and Volterra [2]. Influenced by those works, researchers are still engaged in theoretical study of the ecological system with the help of mathematical modeling. In their book, Kot [3], Britton (2003) described some ecological phenomena on various ecological interactions including prey-predator interactions. Smith [4] has demonstrated various aspects in theoretical ecology with the help of some basic mathematical models. May [5] has also considered and analyzed some other types of ecological systems including prey-predator dynamics, with the help of some sophisticated mathematical models, which are comparatively complex in nature. Some other research works on theoretical ecology including predator-prey dynamics can be found, like Cushing [6], Hadeler and Freedman [7], Chen et al. [8], Kar [9, 10], Kar et al. [11], Chakraborty et al. [12], and references therein.

Harvesting is however a common and quite natural phenomenon. In fishery harvesting is used frequently as the biological resources are mostly renewable resources. On an exploited fishery system with interacting prey and predator species, researchers are considering harvesting on either prey species or predator species or harvesting on both prey and predator species. Martin and Ruan [13] discussed the dynamics of harvesting of prey populations whereas, in his article, Kar [9] describes phenomenon on selective harvesting on a prey-predator system. Further harvesting in predator species or both of prey and predator species can be found in literature also (Kar and Pahari [14], Zhang and Zhang [15], Jana et al. [16–18], Pal et al. [19], Walters et al. [20], Liu and Zhang [21], etc.). From a bioeconomic point of view, harvesting on a species should be in a balance to both keep the resource live and keep fishermen in profitable mode. In his two books Clark [22, 23] describes different harvesting policies in some realistic ecological systems with optimal outcome.

In this regard, in our present article, we consider a prey-predator type ecological system with harvesting on both the species. However, till now most of the models are proposed by considering only precise set of parameters but the natural world may not be precise every time. In many situations at experimental field like birth and death rate of different individuals of the same species, interaction between two different species, etc., may be imprecise. For this purpose introduction of fuzzy sets (Zadeh [24]) is now considered as a revolutionary work. However, consideration of interval-valued parameters was due to the broad application of fuzzy sets. The imprecise parameters set may not always belong to the interval $[0, 1]$ but they may belong to any interval of positive number. Hence interval-valued parameter set of an imprecise mathematical model would be regardlessly better from a realistic point of view.

The rest of the paper is organized in the following manner: In Section 2, we put some preliminaries on interval value numbers. In Section 3, on the basis of some realistic assumptions we formulate our predator-prey system and then convert it to an imprecise parametric system whose dynamical behavior is thoroughly discussed in Section 4. Section 5 is devoted to discussing the existence of bionomic equilibrium, and optimal harvesting policies are studied in Section 6 keeping harvesting parameter as the control variable. In Section 7, we validate our theoretical results through some numerical simulation works, and in the last section we present some key findings.

2. Preliminaries

Here we give the definition of interval numbers with some operations. We use interval-valued function in lieu of interval number.

Definition 1 (interval number). We denote interval number A as $[\underline{a}, \overline{a}]$ and define it as $A = [\underline{a}, \overline{a}] = \{x : \underline{a} \le x \le \overline{a}, x \in \mathfrak{R}\}$ where $\mathfrak{R}$ is called the set of all real numbers and $\underline{a}, \overline{a}$ are the lower and upper limits of the interval number, respectively.

A real number a can also be used in form of interval number as $[a, a]$.

The basic operations between any two interval numbers are as follows:

(i) $[\underline{a_1}, \overline{a_1}] + [\underline{a_2}, \overline{a_2}] = [\underline{a_1} + \underline{a_2}, \overline{a_1} + \overline{a_2}]$.

(ii) $[\underline{a_1}, \overline{a_1}] - [\underline{a_2}, \overline{a_2}] = [\underline{a_1} - \underline{a_2}, \overline{a_1} - \overline{a_2}]$.

(iii) $c[\underline{a_1}, \overline{a_1}] = [c\underline{a_2}, c\overline{a_2}]$ where c is a real number.

(iv) $[\underline{a_1}, \overline{a_1}].[\underline{a_2}, \overline{a_2}] = [\min\{\underline{a_1}\underline{a_2}, \overline{a_1}\underline{a_2}, \underline{a_1}\overline{a_2}, \overline{a_1 a_2}\}, \max\{\underline{a_1}\underline{a_2}, \overline{a_1}\underline{a_2}, \underline{a_1}\overline{a_2}, \overline{a_1 a_2}\}]$.

(v) $[\underline{a_1}, \overline{a_1}]/[\underline{a_2}, \overline{a_2}] = [\underline{a_1}, \overline{a_1}].[1/\overline{a_2}, 1/\underline{a_2}]$.

Definition 2 (interval-valued function). For an interval $[a, b]$ the interval-valued function can be created as $f(p) = a^{(1-p)} b^{p}$ for $p \in [0, 1]$.

3. Predator-Prey Model with Harvesting with Imprecise Parameter

3.1. Crisp Model. In this article, we consider only two species, namely, prey species and predator species. Let $x(t)$ denote the prey biomass and $y(t)$ the predator class at any time t. Let the prey population grow logistically with intrinsic growth rate r and environmental carrying capacity k. Also let the predator attack the prey in the predation rate $\alpha(> 0)$ following the mass action law. Thus the differential equation for the prey population becomes

$$\frac{dx(t)}{dt} = rx\left(1 - \frac{x}{k}\right) - \alpha xy. \tag{1}$$

Let $m(> 0)$ be the conversion factor from the prey population to the matured predator population, d be the natural death rate, and δ be the intraspecific competition rate for the predator populations (due to Ruan et al. [25]). Then the differential equation of the predator population $y(t)$ reduces to

$$\frac{dy(t)}{dt} = m\alpha xy - dy - \delta y^2. \tag{2}$$

Here $m, r, k, \alpha, \delta, d$ are all positive parameters.

Next if we consider that both the species are harvested, this is carried out on assuming the demand in the market of both species (prey and predator). Taking E as the harvesting effort for both species and q_1 & q_2 as the catchability coefficient of the prey species and predator species, respectively, then our system reduces to

$$\begin{aligned} \frac{dx(t)}{dt} &= rx\left(1 - \frac{x}{k}\right) - \alpha xy - q_1 Ex \\ \frac{dy(t)}{dt} &= m\alpha xy - dy - \delta y^2 - q_2 Ey \end{aligned} \tag{3}$$

subject to the initial conditions

$$\begin{aligned} x(0) &\ge 0, \\ y(0) &\ge 0. \end{aligned} \tag{4}$$

3.2. Fuzzy Model. The environment and other factors including temperature and food habits caused the parameters to be imprecise. So they should be taken as interval number rather than a single value. Let $\hat{r}, \hat{k}, \hat{\alpha}, \hat{m}, \hat{d}, \hat{\delta}$ be the corresponding interval numbers for $r, k, \alpha, m, d, \delta$, respectively. Then the prey-predator model with combined harvesting effort E becomes

$$\begin{aligned} \frac{dx}{dt} &= \hat{r}x\left(1 - \frac{x}{\hat{k}}\right) - \hat{\alpha}xy - q_1 Ex \\ \frac{dy}{dt} &= \hat{m}\hat{\alpha}xy - \hat{d}y - \hat{\delta}y^2 - q_2 Ey \end{aligned} \tag{5}$$

where $\hat{r} = [\underline{r}, \overline{r}], \hat{k} = [\underline{k}, \overline{k}], \hat{\alpha} = [\underline{\alpha}, \overline{\alpha}], \hat{m} = [\underline{m}, \overline{m}], \hat{d} = [\underline{d}, \overline{d}]$, $\hat{\delta} = [\underline{\delta}, \overline{\delta}]$.

For the interval number $[\underline{r}, \overline{r}]$ we consider the interval-valued function $r_p = (\underline{r})^{(1-p)}(\overline{r})^p$ for $p \in [0,1]$. Similarly taking the other interval numbers in the same way as function form, we get the model as

$$\begin{aligned}\frac{dx}{dt} &= (\underline{r})^{(1-p)}(\overline{r})^p x\left(1 - \frac{x}{(\underline{k})^{(1-p)}(\overline{k})^p}\right)\\ &\quad - (\underline{\alpha})^{(1-p)}(\overline{\alpha})^p xy - q_1 Ex\\ \frac{dy}{dt} &= (\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p xy - (\underline{d})^{(1-p)}(\overline{d})^p y\\ &\quad - (\underline{\delta})^{(1-p)}(\overline{\delta})^p y^2 - q_2 Ey\end{aligned} \tag{6}$$

subject to the initial conditions

$$\begin{aligned}x(0) &\geq 0,\\ y(0) &\geq 0.\end{aligned} \tag{7}$$

Here $p \in [0,1]$, where the value of p depends on the underlying environment.

4. Dynamical Behavior

In this section, we describe a thorough dynamical behavior of the proposed model system. To do so, we first check the positivity of the solutions of crisp system and uniform boundedness of the solution of the same system. Now, it can also be concluded that uniform boundedness and positivity in the solutions also hold for the corresponding fuzzy systems, if these things hold in crisp system.

4.1. Positivity. First we consider the corresponding crisp system in following form.

$$\begin{aligned}\frac{dx}{x} &= \left[r\left(1 - \frac{x}{k}\right) - \alpha y - q_1 E\right]dt\\ \frac{dy}{y} &= \left[\widehat{m}\widehat{\alpha}x - \widehat{d} - \widehat{\delta}y - q_2 E\right]dt\end{aligned} \tag{8}$$

Now, on integration, we have, from above system of equations,

$$x(t) = x(0)\exp\left(\left[\widehat{r}\left(1 - \frac{x}{\widehat{k}}\right) - \widehat{\alpha}y - q_1\right]E\right) \quad (\geq 0) \tag{9}$$

and

$$y(t) = y(0)\exp\left(\left[\widehat{m}\widehat{\alpha}x - \widehat{d} - \widehat{\delta}y - q_2 E\right]\right) \quad (\geq 0). \tag{10}$$

Hence from above, two expressions related to two state variables will always be positive. Thus the solution of corresponding crisp problem will be nonnegative and so the solution of the corresponding fuzzy system will also be nonnegative.

4.2. Uniform Boundedness. In this section we now study the uniform boundedness of the proposed imprecise system. Now from the first expression of system (5), we have

$$\frac{dx}{dt} + q_1 Ex \leq \widehat{r}x\left(1 - \frac{x}{\widehat{k}}\right). \tag{11}$$

Now, by simple mathematics, it can be concluded that $\widehat{r}x(1 - x/\widehat{k})$ has maximum value $\widehat{r}\widehat{k}/4$, which is obtained for $x = \widehat{k}/2$. Thus from above, we have

$$\frac{dx}{dt} + q_1 Ex \leq \frac{\widehat{r}\widehat{k}}{4}. \tag{12}$$

Now Integrating both sides of the above inequality and then applying the theory of differential inequality due to (see Birkhoff and Rota [26]), we have

$$0 < x(t) \leq \frac{\widehat{r}\widehat{k}}{4q_1 E}\left(1 - e^{-q_1 Et}\right) + x(0). \tag{13}$$

Now on letting $t \longrightarrow \infty$, we have

$$0 < x(t) \leq \frac{\widehat{r}\widehat{k}}{4q_1 E} + \epsilon_1. \tag{14}$$

Hence the biomass density of prey population $x(t)$ is uniformly bounded with an upper and lower limit $\widehat{r}\widehat{k}/4q_1E + \epsilon_1$ and 0, respectively.

Next we are targeting to show that the biomass of predator population $y(t)$ is uniformly bounded. In this regard from (5), we have

$$\frac{dy}{dt} + q_2 Ey \leq \widehat{m}\widehat{\alpha}\left(\frac{\widehat{r}\widehat{k}}{4q_1 E} + \epsilon_1\right)y - \widehat{\delta}y^2. \tag{15}$$

Thus similarly to the above, it is to be claimed that the right hand side of the above expression has maximum value at $y = (\widehat{m}\widehat{\alpha}/2\widehat{\delta})(\widehat{r}\widehat{k}/4q_1E + \epsilon_1)$ and this maximum value is $((\widehat{m}\widehat{\alpha})^2/4\widehat{\delta})(\widehat{r}\widehat{k}/4q_1E + \epsilon_1)^2$.

Thus proceeding in the same way as prey populations and with the help of Birkhoff and Rota [26], we can write

$$\begin{aligned}0 &< y(t)\\ &\leq \frac{(\widehat{m}\widehat{\alpha})^2}{4q_2 E\widehat{\delta}}\left(\frac{\widehat{r}\widehat{k}}{4q_1 E} + \epsilon_1\right)^2\left(1 - e^{-q_2 Et}\right) + y(0).\end{aligned} \tag{16}$$

Now on letting $t \longrightarrow \infty$, we have

$$0 < y(t) \leq \frac{(\widehat{m}\widehat{\alpha})^2}{4q_2 E\widehat{\delta}}\left(\frac{\widehat{r}\widehat{k}}{4q_1 E} + \epsilon_1\right)^2 + \epsilon_2. \tag{17}$$

So the biomass density of predator populations y is also uniformly bounded with lower and upper bound, respectively, 0 and $((\widehat{m}\widehat{\alpha})^2/4q_2E\widehat{\delta})(\widehat{r}\widehat{k}/4q_1E + \epsilon_1)^2 + \epsilon_2$.

Hence the biomass density of both the population species is uniformly bounded.

4.3. Existence of Equilibria. The equilibrium points of this system are given below.

(1) Trivial equilibrium: $E_T(0,0)$.

(2) Axial equilibrium: $E_A(x_1,0)$ [where $x_1 = (\underline{k})^{(1-p)}(\overline{k})^p(1-q_1E/(\underline{r})^{(1-p)}(\overline{r})^p)$] exists if $(\underline{r})^{(1-p)}(\overline{r})^p > q_1E$.

(3) Interior equilibrium: $E_I(x^*, y^*)$ where

$$x^* = \frac{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p\left\{(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p(\underline{d})^{(1-p)}\left(\overline{d}\right)^p + (\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p + (\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_2E - (\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p q_1E\right\}}{(\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p + (\underline{m})^{(1-p)}(\overline{m})^p(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{\alpha})^{2(1-p)}(\overline{\alpha})^{2p}}$$

$$y^* = \frac{(\underline{m})^{(1-p)}(\overline{m})^p(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p\left\{(\underline{r})^{(1-p)}(\overline{r})^p - q_1E\right\} - (\underline{r})^{(1-p)}(\overline{r})^p\left\{(\underline{d})^{(1-p)}\left(\overline{d}\right)^p + q_2E\right\}}{(\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p + (\underline{m})^{(1-p)}(\overline{m})^p(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{\alpha})^{2(1-p)}(\overline{\alpha})^{2p}}. \tag{18}$$

The interior equilibrium exists if

$$\begin{aligned}&(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p(\underline{d})^{(1-p)}\left(\overline{d}\right)^p \\ &\quad + (\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p + (\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_2E \\ &\quad > (\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p q_1E\end{aligned} \tag{19}$$

and

$$\begin{aligned}&(\underline{r})^{(1-p)}(\overline{r})^p(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p \\ &\quad > r\left\{(\underline{d})^{(1-p)}\left(\overline{d}\right)^p + q_2E\right\} \\ &\quad + (\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_1E\end{aligned} \tag{20}$$

hold if

$$E < \min\left(E_1, E_2\right) \tag{21}$$

where

$$E_1 = \frac{(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p(\underline{d})^{(1-p)}\left(\overline{d}\right)^p + (\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p}{(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p q_1 - (\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_2} \tag{22}$$

and

$$E_2 = \frac{(\underline{r})^{(1-p)}(\overline{r})^p\left\{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p - (\underline{d})^{(1-p)}\left(\overline{d}\right)^p\right\}}{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_1 + (\underline{r})^{(1-p)}(\overline{r})^p q_2}. \tag{23}$$

4.4. Local Asymptotic Stability. In this section we state and prove the local asymptotic stability criteria at different equilibrium points. Also the corresponding conditions for which the system is stable at different equilibria are given below.

Case 1. For trivial equilibrium the variational matrix at $E_T(0,0)$ is given by the following.

$$V\left(E_T\right) = \begin{pmatrix}(\underline{r})^{(1-p)}(\overline{r})^p - q_1E & 0 \\ 0 & -(\underline{d})^{(1-p)}\left(\overline{d}\right)^p - q_2E\end{pmatrix} \tag{24}$$

Therefore, the eigenvalues are given by $\lambda_1 = (\underline{r})^{(1-p)}(\overline{r})^p - q_1E$, $\lambda_2 = -(\underline{d})^{(1-p)}(\overline{d})^p - q_2E$.

Here $\lambda_2 < 0$; then $E_T(0,0)$ is asymptotically stable if $\lambda_1 < 0$, i.e., if $(\underline{r})^{(1-p)}(\overline{r})^p - q_1E < 0$ which implies $E > (1/q_1)(\underline{r})^{(1-p)}(\overline{r})^p$.

In the next theorem, we state the stability criteria of trivial equilibrium point

Theorem 3. *Trivial equilibrium point $E_T(0,0)$ of the system is locally asymptotically stable if $E > (1/q_1)(\underline{r})^{(1-p)}(\overline{r})^p$ holds.*

Case 2. At axial equilibrium $E_A(x_1,0)$ the variational matrix is

$$V\left(E_A\right) = \begin{pmatrix}E_{A11} & E_{A12} \\ E_{A21} & E_{A22}\end{pmatrix} \tag{25}$$

where

$$\begin{aligned}E_{A11} &= q_1E - (\underline{r})^{(1-p)}(\overline{r})^p, \\ E_{A12} &= -(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p x_1, \\ E_{A21} &= 0,\end{aligned} \tag{26}$$

and $E_{A22} = (\underline{k})^{(1-p)}(\overline{k})^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p\{1 - q_1E/(\underline{r})^{(1-p)}(\overline{r})^p\} - (\underline{d})^{(1-p)}(\overline{d})^p - q_2E$. Then the eigenvalues of the characteristic equation of $V(E_A)$ are $q_1E - (\underline{r})^{(1-p)}(\overline{r})^p$ and $(\underline{k})^{(1-p)}(\overline{k})^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p\{1 - q_1E/(\underline{r})^{(1-p)}(\overline{r})^p\} - (\underline{d})^{(1-p)}(\overline{d})^p - q_2E$. The first one of them is negative since $(r)^{(1-p)}(\overline{r})^p > q_1E$. Now E_A is asymptotically stable if the second one is negative, i.e.,

$$(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p \cdot\left\{1-\frac{q_1E}{(\underline{r})^{(1-p)}(\overline{r})^p}\right\}-(\underline{d})^{(1-p)}\left(\overline{d}\right)^p-q_2E<0 \tag{27}$$

which implies

$$E>\frac{(\underline{r})^{(1-p)}(\overline{r})^p\left\{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p-(\underline{d})^{(1-p)}\left(\overline{d}\right)^p\right\}}{q_1(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p+q_2(\underline{r})^{(1-p)}(\overline{r})^p}. \tag{28}$$

In the next theorem we will state the local asymptotic stability criteria of the axial equilibrium or the predator free equilibrium $E_A(x_1, 0)$.

Theorem 4. *The axial equilibrium $E_A(x_1, 0)$ is locally asymptotically stable if*

$$\frac{1}{q_1}(\underline{r})^{(1-p)}(\overline{r})^p>E>\frac{(\underline{r})^{(1-p)}(\overline{r})^p\left\{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p-(\underline{d})^{(1-p)}\left(\overline{d}\right)^p\right\}}{q_1(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p+q_2(\underline{r})^{(1-p)}(\overline{r})^p}. \tag{29}$$

In this condition the trivial equilibrium becomes unstable.

Case 3. The variational matrix for interior equilibrium $E_I(x^*, y^*)$ is written below.

$$V(E_I) = \begin{pmatrix} \dfrac{-(\underline{r})^{(1-p)}(\overline{r})^p}{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p}x^* & -(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p x^* \\ (\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p y^* & -(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p y^* \end{pmatrix} \tag{30}$$

The characteristic equation of $V(E_I)$ is given by

$$\lambda^2+S\lambda+M=0 \tag{31}$$

where

$$S=\frac{(\underline{r})^{(1-p)}(\overline{r})^p}{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p}x^*+(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p y^* \tag{32}$$

and

$$M=\left\{\frac{(\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p}{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p} + (\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{2(1-p)}(\overline{\alpha})^{2p}\right\}x^*y^*. \tag{33}$$

Here $S > 0$ and $M > 0$ since $x^* > 0$ and $y^* > 0$. Then the values of λ are negative.

Therefore, The system is locally asymptotically stable at (x^*, y^*) and we state this criteria in the following theorem.

Theorem 5. *The interior equilibrium $E_I(x^*, y^*)$ of the system exists and is locally asymptotically stable if*

$$E<\min\left(E_1,E_2\right), \tag{34}$$

where

$$E_1=\frac{(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p(\underline{d})^{(1-p)}\left(\overline{d}\right)^p+(\underline{r})^{(1-p)}(\overline{r})^p(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p}{(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p q_1-(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_2} \tag{35}$$

and

$$E_2=\frac{(\underline{r})^{(1-p)}(\overline{r})^p\left\{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p-(\underline{d})^{(1-p)}\left(\overline{d}\right)^p\right\}}{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p q_1+(\underline{r})^{(1-p)}(\overline{r})^p q_2}. \tag{36}$$

4.5. Global Stability. Here we will discuss the global asymptotic stability criteria of the system around its interior equilibrium point. In next theorem we study the criteria.

Theorem 6. *The interior equilibrium $E_I(x^*, y^*)$ of the system is globally asymptotically stable provided it is locally asymptotically stable there.*

Proof. A Lyapunov function is constructed here as follows

$$V(x,y) = \int_{x^*}^{x} \frac{x - x^*}{x} dx + P \int_{y^*}^{y} \frac{y - y^*}{y} dy \qquad (37)$$

where P is suitable positive constant to be determined in the subsequent steps.

Taking derivative with respect to t along the solutions of the system, we have

$$\frac{dV}{dt} = \frac{x - x^*}{x}\frac{dx}{dt} + P\frac{y - y^*}{y}\frac{dy}{dt}. \qquad (38)$$

Now

$$\begin{aligned}
\frac{1}{x}\frac{dx}{dt} &= (\underline{r})^{(1-p)} (\overline{r})^p \left(1 - \frac{x}{(\underline{k})^{(1-p)} (\overline{k})^p}\right) \\
&\quad - (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p y - q_1 E \\
\frac{1}{x}\frac{dx^*}{dt} &= (\underline{r})^{(1-p)} (\overline{r})^p \left(1 - \frac{x^*}{(\underline{k})^{(1-p)} (\overline{k})^p}\right) \\
&\quad - (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p y^* - q_1 E \\
\frac{1}{y}\frac{dy}{dt} &= (\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p x \\
&\quad - (\underline{d})^{(1-p)} (\overline{d})^p - (\underline{\delta})^{(1-p)} (\overline{\delta})^p y - q_2 E.
\end{aligned} \qquad (39)$$

Then

$$\begin{aligned}
\frac{dV}{dt} &= (x - x^*) \left[-\frac{(\underline{r})^{(1-p)} (\overline{r})^p}{(\underline{k})^{(1-p)} (\overline{k})^p} (x - x^*) \right. \\
&\quad \left. - (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p (y - y^*) \right] + P (y - y^*) \\
&\quad \cdot \left[(\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p (x - x^*) \right. \\
&\quad \left. - (\underline{\delta})^{(1-p)} (\overline{\delta})^p (y - y^*) \right] = -\frac{(\underline{r})^{(1-p)} (\overline{r})^p}{(\underline{k})^{(1-p)} (\overline{k})^p} (x \\
&\quad - x^*)^2 + \left(P (\underline{m})^{(1-p)} (\overline{m})^p - 1\right) (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p (x \\
&\quad - x^*) (y - y^*) - P (\underline{\delta})^{(1-p)} (\overline{\delta})^p (y - y^*)^2 .
\end{aligned} \qquad (40)$$

If we consider $P = 1/m$, then dV/dt reduces to the following.

$$\begin{aligned}
\frac{dV}{dt} &= -\frac{(\underline{r})^{(1-p)} (\overline{r})^p}{(\underline{k})^{(1-p)} (\overline{k})^p} (x - x^*)^2 \\
&\quad - \frac{(\underline{\delta})^{(1-p)} (\overline{\delta})^p}{(\underline{m})^{(1-p)} (\overline{m})^p} (y - y^*)^2
\end{aligned} \qquad (41)$$

It is seen from the above that $dV/dt \le 0$.

That is, the system is globally stable around its interior equilibrium $E_I(x^*, y^*)$. □

5. Bionomic Equilibrium

In this section we study the bionomic equilibrium of the competitive predator-prey model. Here we consider the following parameters: (1) c: fishing cost per unit effort, (2) p_1: price per unit biomass of the prey, (3) p_2: price per unit biomass of the predator. The net revenue at any time is given by

$$R = (p_1 q_1 x + p_2 q_2 y - c) E. \qquad (42)$$

The interior equilibrium point of the system is on the line given below

$$\begin{aligned}
& x \left((\underline{k})^{(1-p)} (\overline{k})^p q_1 (\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p \right. \\
&\quad \left. + (\underline{r})^{(1-p)} (\overline{r})^p q_2 \right) \\
&\quad + y \left((\underline{k})^{(1-p)} (\overline{k})^p q_2 (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p \right. \\
&\quad \left. - (\underline{k})^{(1-p)} (\overline{k})^p q_1 (\underline{\delta})^{(1-p)} (\overline{\delta})^p \right) = (\underline{k})^{(1-p)} (\overline{k})^p \\
&\quad \cdot \left((\underline{r})^{(1-p)} (\overline{r})^p q_2 + d q_1 \right).
\end{aligned} \qquad (43)$$

This biological equilibrium line meets x-axis at $(x_R, 0)$ and y-axis at $(0, y_R)$, where

$$\begin{aligned}
& \hat{x} \\
&= \frac{(\underline{k})^{(1-p)} (\overline{k})^p \left((\underline{r})^{(1-p)} (\overline{r})^p q_2 + (\underline{d})^{(1-p)} (\overline{d})^p q_1 \right)}{(\underline{k})^{(1-p)} (\overline{k})^p q_1 (\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p + (\underline{r})^{(1-p)} (\overline{r})^p q_2}
\end{aligned} \qquad (44)$$

and

$$\begin{aligned}
& \hat{y} \\
&= \frac{(\underline{k})^{(1-p)} (\overline{k})^p \left((\underline{r})^{(1-p)} (\overline{r})^p q_2 + (\underline{d})^{(1-p)} (\overline{d})^p q_1 \right)}{(\underline{k})^{(1-p)} (\overline{k})^p \left(q_2 (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p - q_1 (\underline{\delta})^{(1-p)} (\overline{\delta})^p \right)}.
\end{aligned} \qquad (45)$$

It is seen that always $\hat{x} > 0$ but $\hat{y}$ is feasible if $q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p > q_1(\underline{\delta})^{(1-p)}(\overline{\delta})^p$.

The 'zero-profit line' is given by

$$R = (p_1 q_1 x + p_2 q_2 y - c) E = 0. \qquad (46)$$

Equation (6) together with the above condition represents the bionomic equilibrium of prey-predator harvesting system.

For the points on the equilibrium line where $(p_1q_1x + p_2q_2y - c) < 0$, the fishery becomes useless. Because it cannot produce any positive economic revenue.

These three cases may arise in bionomic equilibrium.

Case 1. When fishing or harvesting of predator species is not possible, then $x_R = c/p_1q_1$ gives that $c = (p_1q_1(\underline{k})^{(1-p)}(\overline{k})^p((\underline{r})^{(1-p)}(\overline{r})^pq_2 + (\underline{d})^{(1-p)}(\overline{d})^pq_1))/((\underline{k})^{(1-p)}(\overline{k})^pq_1(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p + (\underline{r})^{(1-p)}(\overline{r})^pq_2)$.

Case 2. When harvesting of prey is not possible, then $y_R = c/p_2\ q_2$ gives that $c = (p_2\ q_2(k)^{(1-p)}(k)^p((r)^{(1-p)}(r)^pq_2 + (\underline{d})^{(1-p)}(\overline{d})^pq_1))/((\underline{k})^{(1-p)}(\overline{k})^p(q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p - q_1(\underline{\delta})^{(1-p)}(\overline{\delta})^p))$ with $q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p > q_1(\underline{\delta})^{(1-p)}(\overline{\delta})^p$.

Case 3. When the bionomic equilibrium is at a point (x_R, y_R) where both $x_R > 0$ and $y_R > 0$, then the fishing of prey and predator is possible. Here

$$x_R = \frac{x_{11}}{x_{12}}, \qquad y_R = \frac{y_{11}}{x_{12}}, \tag{47}$$

where

$$\begin{aligned} x_{11} &= (\underline{k})^{(1-p)}\left(\overline{k}\right)^p p_2q_2\left((\underline{r})^{(1-p)}(\overline{r})^p q_2 + (\underline{d})^{(1-p)}\right. \\ &\quad \cdot \left.\left(\overline{d}\right)^p q_1\right) - c(\underline{k})^{(1-p)}\left(\overline{k}\right)^p\left(q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p\right. \\ &\quad \left. - q_1\delta\right), \\ x_{12} &= p_2q_2\left((\underline{k})^{(1-p)}\left(\overline{k}\right)^p q_1(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p\right. \\ &\quad \left. + (\underline{r})^{(1-p)}(\overline{r})^p q_2\right) - (\underline{k})^{(1-p)}\left(\overline{k}\right)^p p_1q_1\left(q_2(\underline{\alpha})^{(1-p)}\right. \\ &\quad \cdot \left.(\overline{\alpha})^p - q_1(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p\right), \\ y_{11} &= c\left((\underline{k})^{(1-p)}\left(\overline{k}\right)^p q_1(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p\right. \\ &\quad \left. + (\underline{r})^{(1-p)}(\overline{r})^p q_2\right) - (\underline{k})^{(1-p)}\left(\overline{k}\right)^p p_2q_2\left(q_2\alpha\right. \\ &\quad \left. - q_1(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p\right). \end{aligned} \tag{48}$$

Since $x_R > 0$ and $y_R > 0$, then the following two conditions hold.

$$\text{(i)}\ \ c > \frac{c_{11}}{c_{12}} \tag{49}$$

$$\text{or } c < \frac{c_{11}}{c_{12}}, \tag{50}$$

where

$$\begin{aligned} c_{11} &= (\underline{k})^{(1-p)}\left(\overline{k}\right)^p(p_2)^2(q_2)^2\left((\underline{r})^{(1-p)}(\overline{r})^p q_2 + (\underline{d})^{(1-p)}\left(\overline{d}\right)^p q_1\right) \\ &\quad \cdot\left((\underline{k})^{(1-p)}\left(\overline{k}\right)^p q_1(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p + (\underline{r})^{(1-p)}(\overline{r})^p q_2\right), \end{aligned} \tag{51}$$

$$c_{12} = \left\{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p\right\}^2 p_1q_1\left(q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p - q_1(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p\right)^2$$

$$\text{(ii)}\ \ c > \frac{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p p_1p_2q_1q_2\left((\underline{r})^{(1-p)}(\overline{r})^p q_2 + (\underline{d})^{(1-p)}\left(\overline{d}\right)^p q_1\right)\left(q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p - q_1(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p\right)}{p_2q_2\left((\underline{k})^{(1-p)}\left(\overline{k}\right)^p q_1(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p + (\underline{r})^{(1-p)}(\overline{r})^p q_2\right)^2} \tag{52}$$

$$\text{or } c < \frac{(\underline{k})^{(1-p)}\left(\overline{k}\right)^p p_1p_2q_1q_2\left((\underline{r})^{(1-p)}(\overline{r})^p q_2 + (\underline{d})^{(1-p)}\left(\overline{d}\right)^p q_1\right)\left(q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p - q_1(\underline{\delta})^{(1-p)}\left(\overline{\delta}\right)^p\right)}{p_2q_2\left(kq_1(\underline{m})^{(1-p)}(\overline{m})^p(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p + (\underline{r})^{(1-p)}(\overline{r})^p q_2\right)^2}. \tag{53}$$

Then we conclude the bionomic equilibrium shorty in the following theorem.

Theorem 7. *The bionomic equilibrium $(x_R, 0)$ always exists, $(0, y_R)$ exists when $q_2(\underline{\alpha})^{(1-p)}(\overline{\alpha})^p > q_1(\underline{\delta})^{(1-p)}(\overline{\delta})^p$, and (x_R, y_R) exists when conditions (49), (50) and (52), (53) hold simultaneously.*

6. Optimal Harvesting Policy

Here both prey and predator populations are considered as fish populations. The optimal net profit is obtained from fishing. We discuss in this section the optimal harvesting policy. We consider the profit gained from harvesting taking the cost as a quadratic function and focusing on the conservation of fish population. The price assumed here is inversely proportional to the available biomass of fish (prey and predator); i.e., if the biomass increases, the price decreases (see Chakraborty et al. (2011)). Let $\check{c}$ be the constant harvesting cost per unit effort and p_1 and p_2 be, respectively, the constant price per unit biomass of the prey and predator. Now our target is to get the maximum net revenues from fishery. Then the optimal control problem can be created in the following way:

$$J(E) = \int_{t_0}^{t_1} e^{-\sigma t} \left[(p_1 - v_1 q_1 Ex) q_1 Ex \right.$$
$$\left. + (p_2 - v_2 q_2 Ey) q_2 Ey - \check{c}E \right] dt, \tag{54}$$

subject to the system of differential equations (6) and the initial conditions (7). v_1 and v_2 are economic constants and σ is the instantaneous discount rate.

Here the control E is bounded in $0 \le E \le E_{max}$ and our object is to find an optimal control E_o such that

$$J(E_o) = \max_{E \in U} J(E) \tag{55}$$

where U is the control set defined by

$$U = \{E : E \text{ is measurable and } 0 \le E$$
$$\le E_{max}, \text{ for all } t\}. \tag{56}$$

Here the convexity of the objective functional with respect to the control variable E along with the compactness of the range values of the state variables can be combined to give the existence of the optimal control E_o. Now the optimal control can be found by using Pontryagin's maximum principle (Pontryagin et al. (1962)). To optimize the objective functional $J(E)$, we construct the Hamiltonian H of the system as follows:

$$H = (p_1 - v_1 q_1 Ex) q_1 Ex - (p_2 - v_2 q_2 Ey) q_2 Ey - \check{c}E$$
$$+ \lambda_1 \left((\underline{r})^{(1-p)} (\overline{r})^p x \left(1 - \frac{x}{(\underline{k})^{(1-p)} (\overline{k})^p} \right) \right.$$
$$\left. - (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p xy - q_1 Ex \right) \tag{57}$$
$$+ \lambda_2 \left((\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p xy \right.$$
$$\left. - (\underline{d})^{(1-p)} (\overline{d})^p y - (\underline{\delta})^{(1-p)} (\overline{\delta})^p y^2 - q_2 Ey \right).$$

Here the variables λ_1 and λ_2 are adjoint variables and the transversality conditions are as follows.

$$\lambda_i(t_1) = 0, \quad i = 1, 2 \tag{58}$$

First we use the optimality condition $\partial H / \partial E = 0$ to obtain the optimal effort which is as follows:

$$E_\sigma = \frac{p_1 q_1 x - p_2 q_2 x - \check{c} - q_1 \lambda_1 x - q_2 \lambda_2 y}{2 (v_1 q_1^2 x^2 + v_2 q_2^2 y^2)}. \tag{59}$$

The adjoint equations are

$$\frac{d\lambda_1}{dt} = \sigma \lambda_1 - \frac{\partial H}{\partial x} = -q_1 E \left(p_1 - 2 v_1 q_1^2 Ex \right)$$
$$+ \left((\underline{\alpha})^{(1-p)} (\overline{\alpha})^p y + q_1 E - (\underline{r})^{(1-p)} (\overline{r})^p \left(1 - \frac{2x}{K} \right) \right.$$
$$\left. + \sigma \right) \lambda_1 - (\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p y \lambda_2,$$
$$\frac{d\lambda_2}{dt} = \sigma \lambda_2 - \frac{\partial H}{\partial y} = -q_2 E \left(p_2 - 2 v_2 q_2^2 Ey \right)$$
$$+ x (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p \lambda_1$$
$$- \left((\underline{m})^{(1-p)} (\overline{m})^p (\underline{\alpha})^{(1-p)} (\overline{\alpha})^p x - (\underline{d})^{(1-p)} (\overline{d})^p \right.$$
$$\left. - 2 (\underline{\delta})^{(1-p)} (\overline{\delta})^p y - q_2 E - \sigma \right) \lambda_2, \tag{60}$$

and, therefore, we have the following theorem regarding the optimal value of the harvesting effort.

Theorem 8. *There exists an optimal control E_σ, corresponding to the optimal solutions for the state variables as x_σ and y_σ such that this control E_σ optimizes the objective functional J over the region U. Moreover, there exist adjoint variables λ_1 and λ_2 satisfying the first order differential equations given in (60) with the transversality conditions given in (58), where, at the optimal harvesting level, the values of the state variables x and y are, respectively, x_σ and y_σ.*

7. Numerical Simulation

In this section, we analyze our mathematical model through some simulation works. The main difference of our proposed model compared to other models of the same type is the consideration of interval-valued parameters instead of fixed-valued parameters. Inclusion of the parameter p assumes the value corresponding to the parameters of the system as an interval. In this regard we first analyze the importance of considering the parameter p in Figure 1. For simulation purpose we consider the following parametric values: $\hat{r} = [4, 6]$, $\hat{k} = [800, 1000]$, $\hat{\alpha} = [0.50, 0.75]$, $\hat{m} = [0.6, 0.8]$, $\hat{d} = [0.1, 0.2]$, $\hat{\delta} = [0.0001, 0.0020]$, $q_1 = 1.2, q_2 = 0.001$, $E = 1.4$. For different parametric values of p $(0 \le p \le 1)$, we have obtained various types of dynamical behavior of the proposed prey-predator system. From Figure 1, it can be said that lower values of p make the system unstable at the interior equilibrium point whereas the higher values of p gradually make the system locally asymptotically stable around the interior equilibrium point. As the numerical value of p increases, the instability solutions slowly become stable (unstable branches at $p = 0.7$ are less than the number of unstable branches at $p = 0.5$, but still at $p = 0.7$ the system is unstable but asymptotically stable at a higher value ($p = 0.9$)).

Next we describe optimal control theory to simulate the optimal control problem numerically. We consider the same parametric values as above and find the solution of optimal control problem numerically. For this purpose we solve the system of differential equations of the state variables (6) and corresponding initial conditions (7) with the help of forward Runge-Kutta forth order procedure. Also the differential equations of adjoint variables (50) and corresponding transversality conditions (52) are solved with the help of backward Runge-Kutta forth order procedure

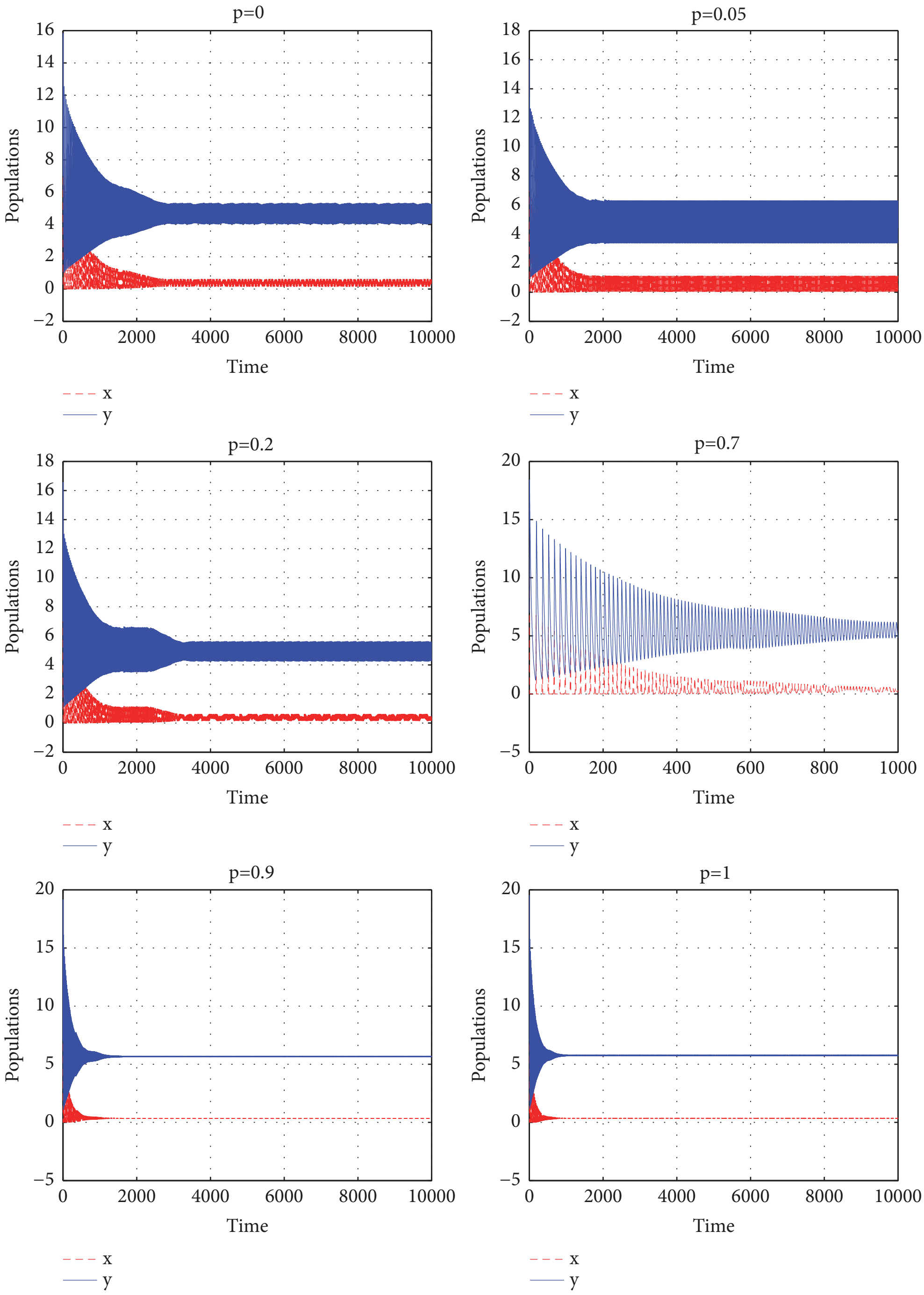


FIGURE 1: Stability of interior equilibrium depends on the numerical value of p. A higher value of p makes the interior equilibrium a stable equilibrium, whereas a lower value makes it unstable one.

for the time interval $[0, 100]$ (see, Jung et al. [27], Lenhart and Workman [28], etc.). Considering harvesting parameter E as the control variable, in Figure 2 and in Figure 3, we, respectively, plot the changes of prey biomass with respect to time and those of predator biomass with respect to time both in presence of control and in absence of control parameter. It is observed that when harvesting control is applied optimally, then the biomass of both prey species and predator species diminishes which is in accord with our expectations. Further in Figure 4, we plot the variation of control parameter (here harvesting effort is the control parameter), and in Figure 5, we plot variations of adjoint variables. It is also to be observed that the level of optimal harvesting effort always belongs to the range $[0.10, 0.45]$. Further according to the transversality conditions, both the adjoint variables λ_1 and λ_2 vanished at the final time (see Figure 5).

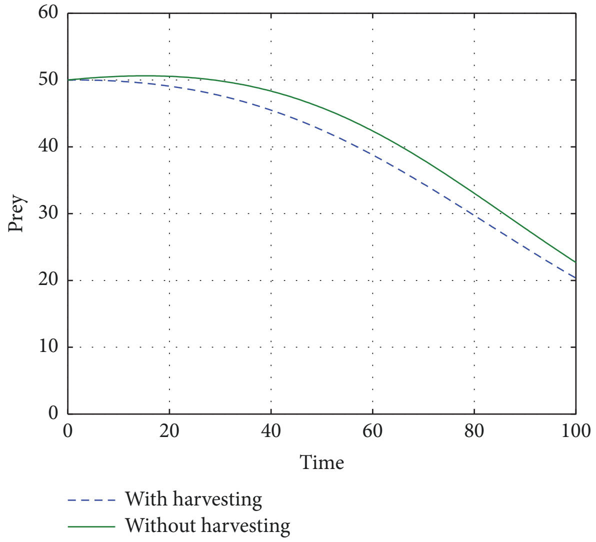


Figure 2: Variation of prey biomass both with control and without control cases.

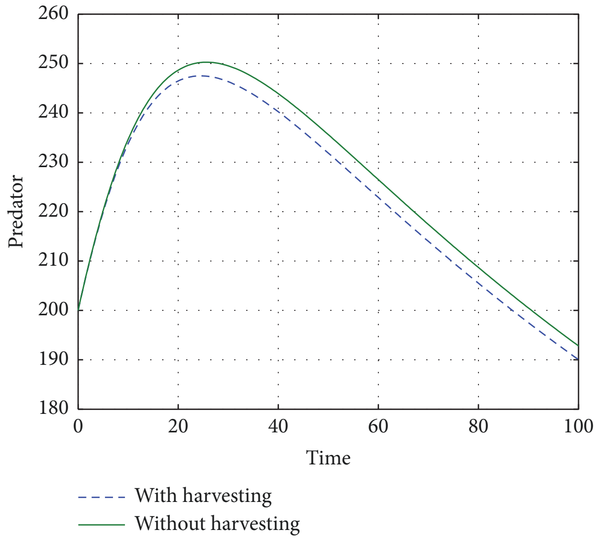


Figure 3: Variation of predator biomass both with control and without control cases.

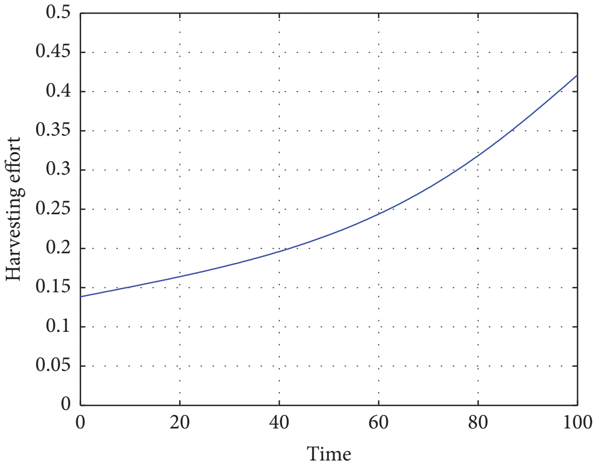


Figure 4: Variation of harvesting effort with time.

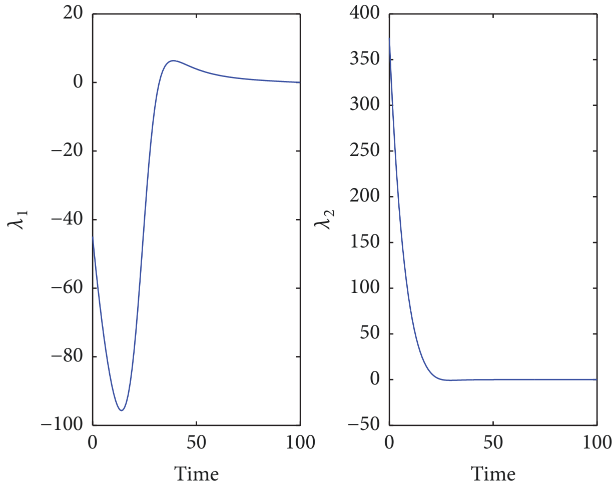


Figure 5: Variation of adjoint variables with time when harvesting control is applied optimally.

8. Discussions and Conclusions

The interactions between prey species and their predator species is an important topic to be analyzed. In present era, many experts are still analyzing the different aspects on this relationship. For this purpose in our present paper, we formulate and analyze a mathematical model on prey-predator system with harvesting on both prey and predator species. Further the model system is improved with the consideration of system parameters assuming an interval value instead of considering a single value. In reality due to various uncertainty aspects in nature, the parameters associated with a model system should not be considered a single value. But often this scenario has been neglected although some recent works considered these types of phenomena (see the works of Pal et al. [19, 29], Sharma and Samanta [30], Das and Pal [31], etc.). Influenced by those works, we also consider that all the parameters associated with our system are of interval value. Further harvesting of both prey and predator species is considered with catch per unit biomass in unit time with harvesting effort E.

The proposed model is analyzed for both crisp and interval-valued parametric cases. Different dynamical behavior of the system, including uniform boundedness, and existence and feasibility criteria of all the equilibria and both their local and global asymptotic stability criteria, has been described. It is found that the system may possess three equilibria, namely, the vanishing equilibrium point, the predator free equilibrium point, and the interior equilibrium point. Theoretical analysis shows that all of these three equilibria may be conditionally locally asymptotically stable depending on the numerical value of the harvesting parameter E. The classical prey-predator model with harvesting effort and without imprecise parametric space in general enables the vanishing equilibrium or trivial equilibrium point as an unstable equilibrium point, but the consideration of imprecise parametric space makes the trivial equilibrium a conditionally stable equilibrium. This phenomenon would surely describe the simultaneous extinction of a single species or both species although the crisp model failed to analyze it.

Next we study explicitly the existence criteria of bionomic equilibrium considering E as harvesting effort. Further considering harvesting effort as the control parameter, we form an optimal control problem with the objective of maximizing the profit due to harvesting in a finite horizon of time and solve that problem both theoretically and numerically. The objective functional considered in optimal control problem is also of both innovative and realistic type, as we consider here that the prices of biomass for both prey and predator species inversely depend upon their corresponding demands.

Consideration of imprecise parameters set makes the model more close to a realistic system which can be well explained with the help of Figure 1. It is shown that for different values of the parameter p, associated with the imprecise values, we obtain different nature of the coexisting equilibrium point. As the numeric value of the associated parameter p increases, the amount of unstable branches for both the species reduces and ultimately becomes a stable system for a higher value of $p(\geq 0.9)$. As the nature of the interior or coexisting equilibrium is one of the most important objects to study, we may claim that the different values of the imprecise parameter p are able to make the proposed prey-predator system understandable and reflect the real world problem.

However, in the present work we consider only a single prey species interacting with a single predator species which makes the model a quite simple one. For our future work we preserve the option of considering more than one type of prey species interacting with more than one type of predator species with imprecise set of parameters. Further due to unavailability of real world data, to simulate our theoretical works, we consider a hypothetical set for the parameters and obtain the result. However, we mainly aim to study the qualitative behavior of the system (not quantitative behavior) which would not be hampered at all due to the consideration of a simulated parametric set.

References

[1] A. J. Lotka, *Elements of Physical Biology*, Williams and Wilkins, Baltimore, New York, 1925.

[2] V. Volterra, "Variazioni e fluttuazioni del numero diâindividui in specie animali conviventi," *Mem. R. Accad. Naz. Dei Lincei*, vol. 2, 1926.

[3] M. Kot, *Elements of Mathematical Biology*, Cambridge University Press, Cambridge, UK, 2001.

[4] J. M. Smith, "Models in Ecology," *CUP Archive*, 1978.

[5] R. M. May, *Stability and Complexity in Model Ecosystems*, Princeton University Press, 2001.

[6] J. M. Cushing, "Periodic two-predator, one-prey interactions and the time sharing of a resource niche," *SIAM Journal on Applied Mathematics*, vol. 44, no. 2, pp. 392–410, 1984.

[7] K. P. Hadeler and H. I. Freedman, "Predator-prey populations with parasitic infection," *Journal of Mathematical Biology*, vol. 27, no. 6, pp. 609–631, 1989.

[8] F. Chen, L. Chen, and X. Xie, "On a Leslie-Gower predator-prey model incorporating a prey refuge," *Nonlinear Analysis: Real World Applications*, vol. 10, no. 5, pp. 2905–2908, 2009.

[9] T. K. Kar, "Selective harvesting in a prey-predator fishery with time delay," *Mathematical and Computer Modelling*, vol. 38, no. 3-4, pp. 449–458, 2003.

[10] T. K. Kar, "Modelling and analysis of a harvested prey-predator system incorporating a prey refuge," *Journal of Computational and Applied Mathematics*, vol. 185, no. 1, pp. 19–33, 2006.

[11] T. K. Kar, A. Ghorai, and S. Jana, "Dynamics of pest and its predator model with disease in the pest and optimal use of pesticide," *Journal of Theoretical Biology*, vol. 310, pp. 187–198, 2012.

[12] K. Chakraborty, S. Jana, and T. K. Kar, "Effort dynamics of a delay-induced prey-predator system with reserve," *Nonlinear Dynamics*, vol. 70, no. 3, pp. 1805–1829, 2012.

[13] A. Martin and S. Ruan, "Predator-prey models with delay and prey harvesting," *Journal of Mathematical Biology*, vol. 43, no. 3, pp. 247–267, 2001.

[14] T. K. Kar and U. K. Pahari, "Non-selective harvesting in prey-predator models with delay," *Communications in Nonlinear Science and Numerical Simulation*, vol. 11, no. 4, pp. 499–509, 2006.

[15] Y. Zhang and Q. Zhang, "Dynamical behavior in a delayed stage-structure population model with stochastic fluctuation and harvesting," *Nonlinear Dynamics*, vol. 66, no. 1-2, pp. 231–245, 2011.

[16] S. Jana, M. Chakraborty, K. Chakraborty, and T. K. Kar, "Global stability and bifurcation of time delayed prey-predator system incorporating prey refuge," *Mathematics and Computers in Simulation*, vol. 85, pp. 57–77, 2012.

[17] S. Jana, S. Guria, U. Das, T. K. Kar, and A. Ghorai, "Effect of harvesting and infection on predator in a prey-predator system," *Nonlinear Dynamics*, vol. 81, no. 1-2, pp. 917–930, 2015.

[18] S. Jana, A. Ghorai, S. Guria, and T. K. Kar, "Global dynamics of a predator, weaker prey and stronger prey system," *Applied Mathematics and Computation*, vol. 250, pp. 235–248, 2015.

[19] D. Pal, G. S. Mahapatra, and G. P. Samanta, "Stability and bionomic analysis of fuzzy prey-predator harvesting model in presence of toxicity: a dynamic approach," *Bulletin of Mathematical Biology*, vol. 78, no. 7, pp. 1493–1519, 2016.

[20] C. Walters, V. Christensen, B. Fulton, A. D. M. Smith, and R. Hilborn, "Predictions from simple predator-prey theory about impacts of harvesting forage fishes," *Ecological Modelling*, vol. 337, pp. 272–280, 2016.

[21] J. Liu and L. Zhang, "Bifurcation analysis in a prey-predator model with nonlinear predator harvesting," *Journal of The Franklin Institute*, vol. 353, no. 17, pp. 4701–4714, 2016.

[22] C. W. Clark, *Bioeconomic modelling and fisheries management*, John Wiley and Sons, New York, NY, USA, 1985.

[23] C. W. Clark, *Mathematical Bio-Economics: The Optimal Management of Renewable Resources*, Pure and Applied Mathematics (New York), John Wiley & Sons, New York, NY, USA, 2nd edition, 1990.

[24] L. A. Zadeh, “Fuzzy sets,” *Information and Computation*, vol. 8, pp. 338–353, 1965.

[25] S. Ruan, A. Ardito, P. Ricciardi, and D. L. DeAngelis, “Coexistence in competition models with density-dependent mortality,” *Comptes Rendus Biologies*, vol. 330, no. 12, pp. 845–854, 2007.

[26] G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, Ginn, Boston, 1982.

[27] E. Jung, S. Lenhart, and Z. Feng, “Optimal control of treatments in a two-strain tuberculosis model,” *Discrete and Continuous Dynamical Systems - Series B*, vol. 2, no. 4, pp. 473–482, 2002.

[28] S. M. Lenhart and J. T. Workman, *Optimal Control Applied to Biological Models*, Mathematical and Computational Biology Series Chapman & Hall/CRC, 2007.

[29] D. Pal, G. S. Mahaptra, and G. P. Samanta, “Optimal harvesting of prey-predator system with interval biological parameters: a bioeconomic model,” *Mathematical Biosciences*, vol. 241, no. 2, pp. 181–187, 2013.

[30] S. Sharma and G. P. Samanta, “Optimal harvesting of a two species competition model with imprecise biological parameters,” *Nonlinear Dynamics*, vol. 77, no. 4, pp. 1101–1119, 2014.

[31] A. Das and M. Pal, “A mathematical study of an imprecise SIR epidemic model with treatment control,” *Applied Mathematics and Computation*, vol. 56, no. 1-2, pp. 477–500, 2017.

Multisource Deep Transfer Learning Based on Balanced Distribution Adaptation

Sushree Sangita Jena, *Department of Computer Sciencel Engineering, Aryan Institute of Engineering & Technology, Bhubaneswar, sushreesangita665.com*

Subhalaxmi Nayak, *Department of Computer Scinece Engineering , NM Institute of Engineering & Technology, Bhubaneswar, subhalaxminayak715@outlook.com*

Anita Subudhi, *Department of Computer Scinece Engineering , Capital Engineering College, Bhubaneswar, anitasubudhi89@gmail.com*

Binayini Pradhan, *Department of Computer Scinece Engineering , Raajdhani Engineering College, Bhubaneswar, binayini.pradhan@gmail.com*

Abstract:

The current traditional unsupervised transfer learning assumes that the sample is collected from a single domain. From the aspect of practical application, the sample from a single-source domain is often not enough. In most cases, we usually collect labeled data from multiple domains. In recent years, multisource unsupervised transfer learning with deep learning has focused on aligning in the common feature space and then seeking to minimize the distribution difference between the source and target domains, such as marginal distribution, conditional distribution, or both. Moreover, conditional distribution and marginal distribution are often treated equally, which will lead to poor performance in practical applications. The existing algorithms that consider balanced distribution are often based on a single-source domain. To solve the above-mentioned problems, we propose a multisource transfer learning algorithm based on distribution adaptation. This algorithm considers adjusting the weights of two distributions to solve the problem of distribution adaptation in multisource transfer learning. A large number of experiments have shown that our method MTLBDA has achieved significant results in popular image classification datasets such as Office-31.

1. Introduction

Machine learning can achieve good results in computer vision, and it is often based on the following assumptions: there are enough data samples in the training dataset and a high-precision classifier; the training data and testing data come from the same feature space and the same distribution. For a new domain, it is often difficult to obtain enough data labels. In this case, transfer learning [1] is a promising method that transfers knowledge from the source domain to the target domain. At the same time, the development of deep learning has accelerated the technical level of transfer learning models. Transfer learning usually assumes that training and testing data come from similar but different distributions [2]. For example, the object that takes a photo under different angles, backgrounds, and lighting may get different marginal condition distributions. The existing transfer learning methods mainly focus on distributed adaptation by observing and reducing the difference between each domain through joint distribution adaptation. For example, several unsupervised transfer learning methods [3–5] use maximum mean discrepancy in the neural network to reduce the domain difference; other models introduce different learning modes to align the source and target domains, including aligning second-order correlation [6, 7].

In recent years, most unsupervised transfer learning algorithms have focused on single-source unsupervised transfer learning problems, which are training samples that come from a single-source domain. In previous research, the work focused on estimating the sample's weight, which is the ratio of the source domain and the target domain [8–11]. In addition, the manifold learning method is used to sample a high-dimensional space and map it to a low-dimensional manifold space to make sure that the subspace of the source domain and the target domain comes closer. Some single-source transfer learning algorithms map the data of two domains to a common feature space and describe the invariant features of the source and target domains by minimizing the difference in domain distribution [6, 12–14]. Long [15], Hou [16], and Hashemi [17] had also proposed many joint distribution adaptive methods to solve the difference of distribution between the source domain and the target domain. In recent years, many deep transfer learning algorithms were proposed to solve the problem of data distribution adaptation. Tzeng et al. proposed DDC (deep domain confusion) [18], and Long [12] et al. proposed the DAN (deep adaptation network) to solve the problem of marginal distribution adaptation. Zhu [19] et al. proposed the DSAN (deep dynamic adaptation network), and Wang [20] proposed the DDAN (deep dynamic adaptation network) to solve the problem of jointly distributed adaptation.

However, in practical work, we often face multiple source domains, so it is more feasible to study the migration of multiple source domains, and it is also more meaningful in practice. For multisource transfer learning, a common simple idea is to combine all source domains into a new source domain and then use the single-source transfer learning algorithm to classify the target domain data. Due to the expansion of the dataset, these methods may yield better results. However, in practical applications, because of the large differences in the distribution of each domain, this type of method does not yield good results. Therefore, we need to find a better way to utilize data from multiple source domains.

With the rapid development of deep learning, there are many studies on transfer learning based on deep learning. Zhao [21] et al. proposed a multidomain adversarial network, which aligns the distribution of features in each source domain and target domain through multiple domain discriminators; Xu [22] et al. proposed a deep cocktail network. A separate domain discriminator and a classifier are designed for each source domain and target domain. The current deep multisource transfer learning algorithms often have the following two problems: 1. They first map the source domain samples and target domain samples to the same common feature space, but even for a single-source domain. It is also difficult to learn the same characteristics as those of the target domain. Moreover, in multiple source domains, their data samples are likely to cross, which leads to a reduced effect of feature alignment. 2. At present, the studies often consider only the marginal probabilities or conditional probabilities for the distribution of the source domain and the target domain. Current algorithms often adjust the marginal probability first and then adjust the conditional probability. The relationship between them is not fully utilized.

In this article, we combine the advantages of balancing distribution, convolutional neural networks, and multisource transfer learning, and then a new multisource transfer learning algorithm based on balanced distribution adaptation—MTLBDA—is proposed, which first maps multiple source domains and target domains to the same subspace and then aligns the features of multiple source domains and target domains. Then, according to the balanced distribution adaptation, the effect of the category in each source domain and target domain is decreased, and the difference between the marginal probability distribution and the conditional probability distribution in each source domain and target domain is reduced. Then the convolutional neural network is used as the classifier for each source domain and target domain to complete the task of classification. Finally, we generate a regularization term for the classifier of each source domain, which is weighted to prevent overfitting of the model.

Compared with the previous work, the contributions of this work include the following:

(1) A new multisource transfer learning algorithm named MTLBDA is proposed, which balances the difference between conditional probability distribution and marginal probability distribution to improve the classification effect. This method first maps all domains to the same feature space and then reduces the difference between the marginal probability distribution and conditional probability distribution with maximum mean discrepancy and adds a separate regularization term to the convolutional neural networks on this basis.

(2) For multisource domain samples, the conditional probability distribution and marginal probability distribution are considered. This method can adjust the category adaptation of the multisource domain and target domain.

(3) Multiple source domains provide more knowledge on the learning tasks of the target domain. Compared with the trained classifier set and the independent classifier, the trained classifier set system has a better prediction effect and more stability.

(4) Experiments on real datasets show that the proposed algorithm is superior to or at least comparable to advanced benchmark algorithms in classification accuracy.

The rest of the paper is arranged as follows: Section 2 reviews the work related to multisource transfer learning and joint distribution adaptation. Section 3 proposes multisource deep transfer learning based on balanced distribution adaptation. Section 4 verifies the effectiveness of the algorithm on the SVHN dataset, USPS dataset, MINIST dataset, Office-31 dataset, and DomainNet dataset. Section 5 summarizes the main work of this paper.

2. Problem definition

2.1. Joint Distribution Adaptation. A domain often has two probability distributions: one is marginal probability, and the other is conditional probability. Long [21] gave the hypothesis of joint distribution adaptation, whose purpose is to reduce the distance of joint probability distribution between the source domain and the target domain. Current research on joint distribution includes domain invariant clustering [16], increasing structural consistency [17], target optimization [23], and so on. Wang [20] proposed a dynamic balance adaptive algorithm, which pointed out that marginal distribution adaptation and conditional distribution adaptation are not equally important. However, these joint distribution adaptations are often used in the field of single-source transfer learning, and they have not played a role in the field of multisource transfer learning.

2.2. Multisource Transfer Learning. Multisource transfer learning (as shown in Figure 1) as a research direction of transfer learning has essential practical value. In the process of real life and practical application, there are often multiple source domains. Although each source domain has a different similarity to the target, these source domains can still be used for knowledge transfer. Moreover, multisource transfer learning contains more knowledge, which can make the effect of the model better. At the same time, transfer learning also has a theoretical basis. Crammer [24] first proposed the expected loss boundary condition of multisource transfer learning. Later, Mansour [25] proved that the distribution weighted combination rule can reduce the instantaneous function between the source domain and the target domain. Ben-David [26] gave two learning boundaries for minimizing empirical risk by introducing the distance between the target domain and the source domain.

In recent years, a lot of work was centered around multisource transfer learning and deep learning. Xu [22] proposed the deep cocktail network (DCTN), which uses a single domain discriminator and a classifier for each source domain and target domain. The domain discriminator is used to align the feature distribution, and the classifier outputs the predicted probability distribution. Based on the output of the domain discriminator, the DCTN designed a method of voting by multiple classifiers. Peng [27] proposed a moment matching multisource domain adaptation (M^{3}SDA) method, which not only considers the alignment between the source domain and the target domain but also aligns the feature distribution of different source domains. Zhu [28] et al. proposed a framework named aligning domain-specific distribution and classifier for cross-domain classification from multiple sources (MFSAN). However, the current deep multisource transfer learning algorithms often only consider the marginal probability distributions or consider the marginal probability distribution and the conditional probability distribution separately. In this paper, multisource transfer learning based on balanced distribution adaptation, which considers the joint probability distribution to improve the accuracy of the algorithm, is proposed.

Problem.

In multisource transfer learning, there are N source domains, and their labeled sample data can be represented as $X^{s_i} = \left\{\left(x_j^{s_i}, y_j^{s_i}\right)\right\}_{j=1\}}^{N_i}$, where $\left\{\left(x_j^{s_i}\right)\right\}_{j=1}^{N_i}$ represents the i-th sample data in the j-th source domain and $\left\{\left(y_j^{s_i}\right)\right\}_{j=1}^{N_i}$ represents the i-th source domain in the j-th source domain. The joint probability distribution of N different domains can be expressed as $\{P^{s_i}(x, y)\}_{i=1}^{N_i}$, where the marginal probability can be expressed as $\{P^{s_i}(x)\}_{i=1}^{N_i}$ and the conditional probability can be expressed as $\{P^{s_i}(y|x)\}_{i=1}^{N_i}$. Similarly, we give the definition of the target domain; the sample of the target domain can be expressed as $X^t = \left\{\left(x_j^t\right)\right\}_{j=1}^{N_t}$, and the probability distribution can be expressed as $P^t(x, y)$.

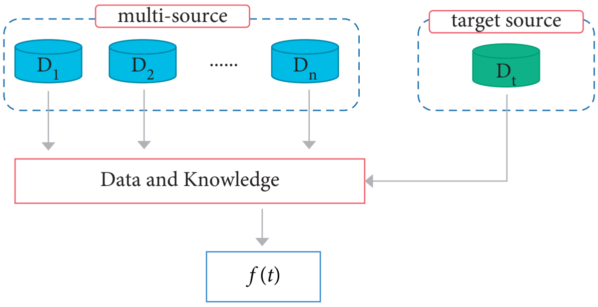


Figure 1: Multisource transfer learning.

In recent years, some papers had defined the objective function of multisource deep transfer learning. They first map all domains to the same target space and then use the common domain invariant representation in the common feature space for learning all domains. Zhu [28] et al. gave a definition of the loss function:

$$\min_{F,C} \sum_{i=1}^{N} \mathrm{E}_{x\sim X_{s_i}} J\left(C_i H_i\left(F\left(x_j^{s_i}\right)\right), y_j^{s_i}\right) + \lambda \sum_{i=1}^{N} \hat{D}\ H_i\left(F\left(X_{s_i}\right), F\left(X_t\right)\right) + \gamma L_{\text{disc}}. \tag{1}$$

The first term represents the loss of the classification function, the general classification loss is the cross-entropy loss, and the second term represents the statistical measurement of the source domain and the target domain. Nowadays, the commonly used metrics are MMD [15], reference loss [29], CORAL loss [12], and confusion loss [13, 14]. Zhu et al. defines CORAL loss as a specific difference loss. The common problem of these methods is that they only use MMD to calculate the marginal distribution difference between the source domain and the target domain without considering the influence of conditional probabilities on the model. Zhao [30]'s paper published on ICML2019 proved theoretically that reducing the marginal distribution difference between the source domain and the target domain is not enough. At the same time, Wang's paper also pointed out that equal consideration of marginal distribution and conditional probability distribution is not enough. Therefore, we propose a multisource deep transfer learning algorithm based on balanced distribution adaptation to solve these problems.

Similar to other multisource transfer learning algorithms, we first map multiple source domains and target domains to the same subspace, and then we align the marginal probability distribution and conditional probability distribution of each source domain and target domain. Of course, the best way is to tune the convolutional neural network for each pair of source and target domains. However, from a practical point of view, the amount of calculation in this method is very large, so we use shared weights to solve this problem. Finally, we add a specific regularization term to realize the problem of individual network tuning.

3. Multisource Deep Transfer Learning Based on Balanced Distribution Adaptation

To solve the impact of category imbalance on the existing multisource transfer learning algorithm, in this chapter, we introduce a multisource transfer learning algorithm based on distribution balance. We use the general regularization item proposed in [28] to replace the classification selector to output the final classification result.

Algorithm structure. Our algorithm structure contains three parts—a common feature selector, a distribution balancer, and a regularizer—as shown in the figure.

Common feature extractor: We propose a common subnet $f(\cdot)$ to extract the common representation of all domains, which maps images from the original feature space to a common feature space.

Domain-specific distribution balancer: We design a distribution balancer for each source domain and target domain data, given a set of images x_j^s from the source domain $X^s = \left\{(x_j^s, y_j^s)\right\}$ and a set of images x^t from the target domain $\left\{(x_j^t)\right\}$. The features of these specific fields are mapped to the same feature space through a common feature extractor, specifically expressed as the source domain mapping feature $f(x_j^s)$ and target domain mapping characteristics $f(x^t)$. Hence, we can get N independently distributed balancers $b(\cdot)$ corresponding to specific source domains $\left\{(x_j^s, y_j^s)\right\}$.

The class balancer we proposed is a domain-specific feature extractor. Generally, people use MMD, CORAL, adversarial, and other methods as feature extractors, but they often only consider one distribution. To balance the categories, we use the BDA algorithm proposed by Wang Jindong [20] while considering conditional distribution, marginal distribution, and multiclass balance as the distribution balancers. We use a convolutional neural network as our classifier, and we define C_i as the classifier of N source domains. Based on experience, our loss is classified as cross-entropy loss, and the loss function is denoted as $J(\cdot,\cdot)$.

Domain-specific regularization term: Based on the behavior regularization proposed in the literature, for the source domain i, we give the regularization term $\mathfrak{R}(w, w^*, x_j, y_j)$, where $\mathbf{w}$ is the d-dimensional parameter vector containing all d parameters of the target domain under the convolutional neural network. $\mathbf{w}^*$ is the parameter vector of the source domain. It is harder to calculate all parameters for each domain, so we share the parameters of the first $n-3$ layers.

Objective function:

According to Figure 2, we define the final objective function of the algorithm as

$$L = L_{cls} + Lb_{bda} + Lr_{reg}. \tag{2}$$

The classification loss Lb_{cls} is the loss caused by a specific domain classifier, and in Figure 2, we can see that the variable x_j in the source domain i undergoes a three-step transformation: first, $F(x_j^{s_i})$ is obtained through the public feature extractor; then $B_i(F(x_j^{s_i}))$ is obtained through the class balancer; finally, $C_i(B_i(F(x_j^{s_i})))$ is obtained through the CNN classification. The final classification loss is

$$L_{cls} = \sum_{i=1}^{N} \mathrm{E}_{x\sim X^{s_i}} J\left(C_i B_i\left(F\left(x_j^{s_i}\right)\right), y_j^{s_i}\right). \tag{3}$$

The balance loss Lb_{bda} is a specific domain balancer loss, and we follow the concept of single-source domain distribution balancers according to Wang et al. The algorithm considers the conditional probability and marginal probability distribution of the source domains and target domains at the same time. In particular, due to the inability to obtain the label of the target domain, we have no way of estimating the conditional probability distribution. Therefore, we use the proof given in [31]; when there are enough label samples, we can use the conditional distribution $P(x_t|y_t)$ of the class to approximately match the conditional distribution $P(y_t|x_t)$. In calculating the conditional distribution $P(x_t|y_t)$ of the class, we first use the specific domain classifier to label the target domain data samples to form the prelabels of the target domain samples.

$$D(\mathscr{D}_s, \mathscr{D}_t) \approx (1-\mu)D(P(x_s), P(x_t)) + \mu D(P(y_s \mid x_s), P(y_t \mid x_t)). \tag{4}$$

$\mu \in [0,1]$ is the balance factor. When $\mu \longrightarrow 0$, the marginal distribution is more important, and when $\mu \longrightarrow 1$, the conditional probability is more important. To calculate the marginal probability and conditional probability, according to MMD and TCA [32], we can define the specific domain balancer estimation empirically.

$$\hat{D}_H(p,q) = \left\| \frac{1}{n_s} \sum_{x_a \in X^s} \phi(x_a) - \frac{1}{n_t} \sum_{x_b \in X^t} \phi(x_b) \right\|_H^2,$$

$$\hat{D}^{\mathrm{cn}}\ (p_t, q_t) = \sum_{c=1}^{C} \left\| \frac{1}{n_c} \sum_{x_a \in X^s_{(c)}} \phi(x_a) - \frac{1}{m_c} \sum_{x_b \in X^t_{(c)}} \phi(x_b) \right\|_H^2. \tag{5}$$

For the i-th source domain, the squared distance between the empirical kernel average embeddings is obtained from the empirical estimation of MMD.

$$\wedge\ D^B = (1-\mu_i) \left\| \frac{1}{n_s} \sum_{x_a \in X^s} \phi(x_a) - \frac{1}{n_t} \sum_{x_b \in X^t} \phi(x_b) \right\|_H^2 + \mu_i \sum_{c=1}^{C} \left\| \frac{1}{n_c} \sum_{x_a \in X^s_{(c)}} \phi(x_a) - \frac{1}{m_c} \sum_{x_b \in X^t_{(c)}} \phi(x_b) \right\|_H^2. \tag{6}$$

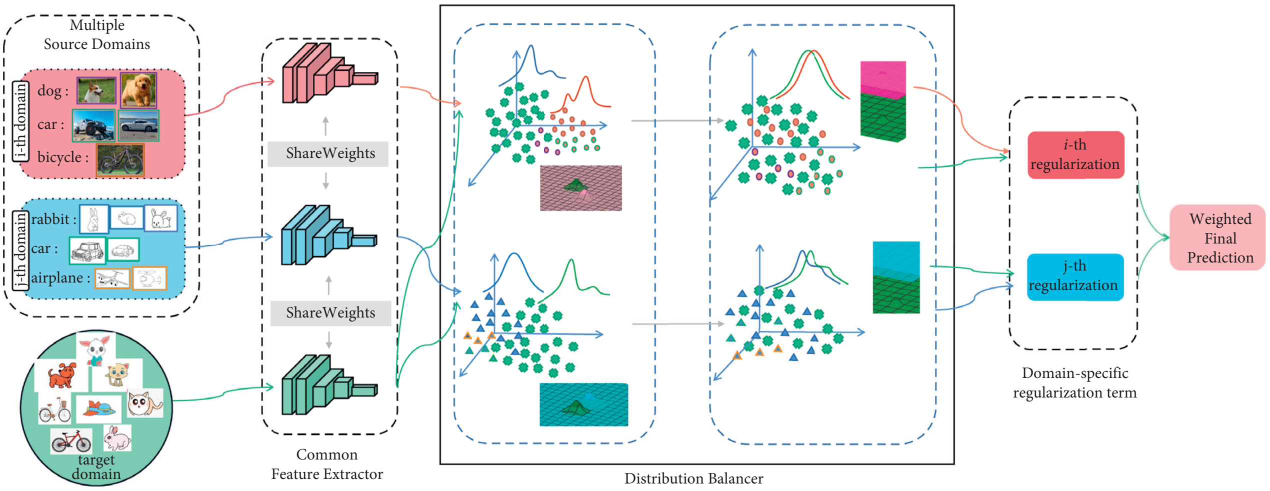


FIGURE 2: The framework of multisource deep transfer learning based on balanced distribution adaptation (MTLBDA). Our model consists of three components: (i) a common feature extractor, (ii) a domain-specific distribution balancer, and (iii) a domain-specific regularization term.

We define formula (6) as the estimation of the difference between the source domain and the target domain. Therefore, the balance loss is defined as follows:

$$L_{bda} = \sum_{i=1}^{N} \widehat{D}\left(B_i\left(F\left(X^{s_i}\right), F\left(X^{t}\right)\right)\right). \tag{7}$$

We define the regularization term for a specific domain *i* according to the behavior regularization term proposed in [33] as follows:

$$\begin{aligned}
\mathfrak{R}^{i}\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right) &= \sum_{j=1}^{n} \Omega\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right),\\
\Omega\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right) &= \alpha \cdot \Omega'\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right) + \beta \cdot L^{2}\left(\mathbf{w}, \mathbf{w}^{*}\right),\\
\Omega'\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right) &= \sum_{k=1}^{N} \mathrm{W}_k\left(\mathbf{w}^{*}, x_j, y_j\right) \cdot \left\| FM_k\left(\mathbf{w}, x_j\right) - FM_k\left(\mathbf{w}^{*}, x_j\right)\right\|_2^2.
\end{aligned} \tag{8}$$

$W_k(\mathbf{w}^*, x_j, y_j)$ refers to the weight assigned to the j-th image in the k-th layer of the network, $FM_k(\mathbf{w}, x_j) \cdot FM_k(\mathbf{w}^*, x_j)$ refers to the difference in the characteristics of the two images, $\|\cdot\|_2$ indicates their Euclidean distance, and $L^2(\mathbf{w}, \mathbf{w}^*)$ represents the L^2 regularization term of $\mathbf{w}$ and $\mathbf{w}^*$. In order to reduce calculation, $k = \{n, n-1, n-2\}$. Collecting the regularization terms of multiple source domains, we define the regularization loss as

$$L_{reg} = \mathfrak{R}\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right) = \sum_{i=1}^{N} \gamma_i \cdot \mathfrak{R}^{i}\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right). \tag{9}$$

$\gamma \in [0, 1]$ is the value of the regularization term ranging from 0 to 1, and its selected value is defined according to the subsequent selector.

Final objective function:

$$\begin{aligned}
L &= L_{cls} + Lb_{bda} + Lr_{reg},\\
L &= \sum_{i=1}^{N} \mathrm{E}_{x\sim X^{s_i}} JC_i\left(B_i\left(F\left(x_j^{s_i}\right), y_j^{s_i}\right)\right.\\
&\quad + \sum_{i=1}^{N} \widehat{D}\left(B_i\left(F\left(X^{s_i}\right), F\left(X^{t}\right)\right)\right)\\
&\quad + \sum_{i=1}^{N} \gamma_i \cdot \mathfrak{R}^{i}\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right),\\
L &= \sum_{i=1}^{N} \left(\mathrm{E}_{x\sim X^{s_i}} J\left(C_i\left(B_i\left(F\left(x_j^{s_i}\right)\right), y_j^{s_i}\right)\right.\right.\\
&\quad \left.+ \widehat{D}\left(B_i\left(F\left(X^{s_i}\right), F\left(X^{t}\right)\right)\right) + \gamma_i \cdot \mathfrak{R}^{i}\left(\mathbf{w}, \mathbf{w}^{*}, x_j, y_j\right)\right).
\end{aligned} \tag{10}$$

In summary, the specific process steps of the MTLBDA algorithm are shown in Table 1.

Table 1: MTLBDA algorithm steps.

MTLBDA algorithm training
Input: N source domains $X^s = \left\{X^{s_i} = \left\{(x_j^{S_i}, y_j^{S_i})\right\}_{j=1}^{N_{s_i}} \cdot i = 1, \ldots, N\right\}$. The number of label samples in each source domain is N_{S_i} ; the target domain is $X^t = \left\{(x_j^t)\right\}_{j=1}^{N_t}$.
Output: Loss function $f(x)$
1: Give the number of training iterations T
2: From 1 to T
3: Randomly take m samples from a certain source domain
4: Take m samples from the target domain
5: Send the source and target samples to a common feature extractor, and get a common expression as $f(\cdot)$
6: Input the common latent representation of the source sample into the domain-specific distribution balancer to obtain the domain-specific representation of the source sample $b(\cdot)$
7: The specific domain representation of the original sample is output to the specific domain classifier, and the calculation formula of the classifier is (3)
8: The general latent representation of the target sample is input to all domain-specific extractors to obtain the domain-specific representation of the target sample.
9: Use the formula to calculate balance loss (7)
10: Make all passes to minimize the total loss in formula (10), update public feature extractor $F(\cdot)$、 multiple domain distribution balancer $B_1 B_2 \cdots B_N$ and multiple classifiers $C_1 C_2 \cdots C_N$, multiple regularization terms $R_1 R_2 \cdots R_N$ ◦
11: Finish

4. Experimental Results

To test the effectiveness and generalization of the MTLBDA algorithm, we test it on two types of image datasets. The first type is a digital classification dataset including the SVHN [34] dataset, USPS [35] dataset, and MNIST [36] dataset. The second category is of image classification datasets including the Office-31 [37] dataset, Caltech [38] dataset, and DomainNet [24] dataset.

The experiment will compare single-source transfer learning algorithms DAN, DANN, BDA, and DDAN, and multisource transfer learning algorithms DCTN, MFSAN, and M^{3}SDA.

For fairness of the experiment, a 5-fold cross-validation strategy is selected for all experiments, and the experiments of this strategy are repeated twice to obtain the final comparison result. In the experiment, we use the average classification accuracy [39, 40] and recall rate of each algorithm after running it for 10 times as the evaluation criteria. The recall rate reflects how many positive examples in the sample are predicted correctly. The forms of expression of classification accuracy and recall are defined as follows:

Classification accuracy: Accuracy $= |x\colon x \in X \wedge f(x) = y(x)| / |x\colon x \in X|$.

Recall rate: $R = FP/TP + FN \times 100\%$.

Among them, TP represents the number of positive samples that are correctly classified as positive, FP represents the number of negative samples that are incorrectly classified as positive, TN represents the number of negative samples that are correctly classified as negative, and FN represents the number of positive samples that are incorrectly classified as negative.

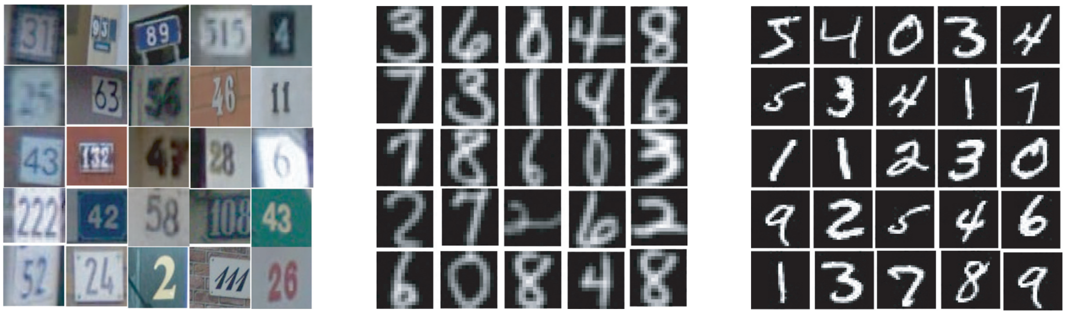
X represents the target domain number test dataset, $f(x)$ is the sample x-class label predicted by the classifier, and $y(x)$ is the reality-class label of sample x.

4.1. Digital Classification Dataset

4.1.1. Dataset Introduction. Both the USPS dataset and the MNIST dataset contain handwritten digits "0"–"9"; the former is composed of 9298 16 × 16 images, and the latter is composed of 70,000 28 × 28 images. The street view house number (SVHN) is obtained from Google. Each picture contains a group of Arabic numerals '0–9′, which contains 73257 digits, and the image pixel is 32 × 32. Figure 3 shows examples of USPS, MNIST, and SVHN. We can see that the distributions of USPS and MNIST are different but they contribute the same feature space. SVHN datasets are different in their distribution and feature space. We extract 9000 images from MNIST and SVHN as two domains. Since USPS has only 9298 pictures, we regard the whole dataset as a domain.

4.1.2. Experimental Data. In this part, we compare some single-source transfer learning algorithms and multisource transfer learning algorithms such as DCTN and MFSAN with our algorithm MTLBDA

It can be seen from Table 2 that among the three cross-domain tasks, the highest accuracy rates of the MTLBDA algorithm are 83.56%, 98.43%, and 96.14%, which are higher by 3.31%, 0.35%, and 0.02% than the algorithms of DCTN, MFSAN, and M^{3}SDA, respectively. Compared with the classification tasks with S as the target domain or source domain, the multisource transfer learning algorithm is clearly better than the single-source transfer learning algorithm. Due to the single-source transfer learning algorithm, in the tasks with S as the target domain, our algorithm MTLBDA is 13.04% more accurate than the best transfer learning algorithm DDAN. From Figure 4, we can see that the different types of S, U, and M pictures lead to their distribution differences, which also proves the accuracy of our algorithm.

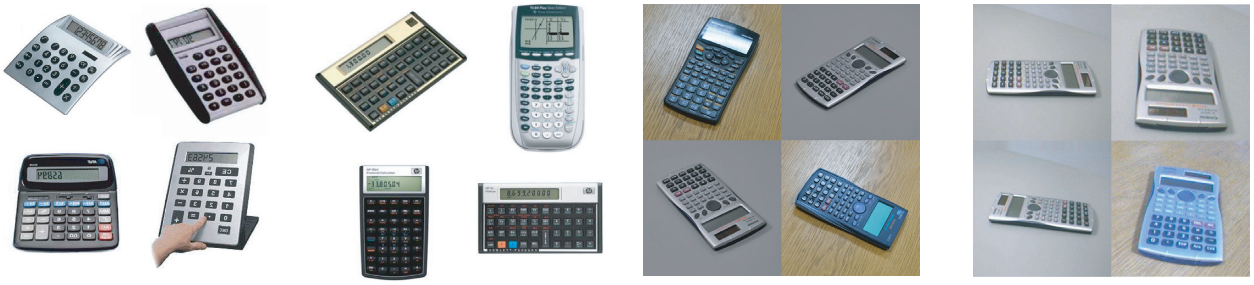


SVHN UPS MNIST

Figure 3: Example of USPS, MNIST, and SVHN pictures.

Table 2: Average classification accuracy (%) on the digital classification datasets.

Algorithms	U- > S	M- > S	U- > M	S- > M	M- > U	S- > U
DAN	68.23 (0.43)	67.84 (0.41)	97.5 (0.62)	66.91 (0.83)	93.49 (0.85)	65.33 (1.12)
DANN	68.65 (0.88)	68.14 (0.82)	97.92 (0.81)	67.23 (0.94)	93.47 (0.79)	66.25 (0.97)
BDA	68.36 (0.45)	67.72 (0.39)	98.13 (0.41)	67.54 (0.73)	93.62 (0.51)	65.13 (0.82)
DDAN	70.52 (0.61)	69.53 (0.59)	98.22 (0.46)	68.95 (0.39)	94.23 (0.41)	66.89 (0.55)
	U, M- > s		U, S- > m		M, S- > u	
DCTN	77.61 (0.41)	76.83 (0.39)	96.23 (0.82)	96.85 (0.73)	92.81 (0.47)	93.08 (0.56)
MFSAN	78.56 (0.95)	78.16 (1.12)	98.08 (0.97)	97.78 (1.17)	94.23 (0.79)	93.83 (0.92)
M3SDA	80.25 (0.82)	79.47 (0.75)	97.48 (0.85)	97.63 (0.91)	94.67 (0.93)	95.31 (0.96)
MTLBDA	83.56 (0.66)	82.82 (0.41)	97.91 (0.76)	98.43 (0.68)	96.14 (0.81)	94.81 (1.03)



Caltech-256 Amazon DSLR Webcam

Figure 4: Example of Office-31 and Caltech-256 pictures.

4.2. Image Classification Dataset

4.2.1. Dataset Introduction. The Office-31 dataset is a commonly used standard transfer learning dataset. It contains 4652 sample pictures collected from different areas, namely, Amazon (A), webcam (W), and DSLR (D), and these pictures can be divided into 31 categories. Among them, Amazon's samples are from https//www.amazon.com, and the samples in webcam and DSLR are obtained through web cameras and digital SLR cameras in different environments. Caltech-256 [38] is a standard database for object recognition. The database has 30607 images and 256 categories. In these experiments, we used the dataset as Office-31+Caltech published by Gong [37] et al., as shown in Figure 4. Specifically, we have four domains, H (Caltech-256), A (Amazon), W (webcam), and D (DSLR). We randomly select three domains as the source domain and the remaining one as the target domain, that is, (A,H,D- > W), (A,H,W- > D), (A,D,W- > H), and (H,D,W- > A).

Table 3: Average classification accuracy (%) on the image classification datasets.

Algorithms	A- > H	W- > H	D- > H	A- > D	W- > D	H- > D
DAN	81.73 (0.89)	70.87 (1.35)	77.96 (0.77)	95.71 (0.59)	98.25 (0.55)	97.12 (0.42)
DANN	85.23 (1.06)	75.31 (1.12)	83.15 (0.93)	96.12 (0.87)	98.12 (0.83)	97.35 (0.78)
BDA	87.95 (0.66)	80.47 (0.78)	85.73 (0.57)	95.42 (0.64)	98.46 (0.56)	97.52 (0.55)
DDAN	89.34 (0.94)	80.22 (0.97)	86.45 (1.04)	96.83 (0.98)	98.79 (0.93)	98.17 (1.03)
	A- > W	D- > W	H- > W	W- > A	D- > A	H- > A
DAN	93.47 (0.87)	96.31 (0.42)	93.69 (0.76)	88.84 (1.08)	90.27 (1.12)	90.82 (1.23)
DANN	95.31 (0.82)	96.24 (0.45)	95.75 (1.01)	89.25 (1.06)	91.76 (1.18)	90.41 (1.07)
BDA	96.24 (0.62)	96.14 (0.53)	95.43 (0.75)	90.55 (0.87)	91.43 (0.85)	90.73 (0.74)
DDAN	96.52 (1.01)	96.84 (0.99)	96.26 (0.87)	90.95 (0.82)	92.13 (0.86)	91.44 (0.99)
		A, W, D- > H			A, W, H- > D	
DCTN	89.51 (0.53)	90.24 (0.48)	88.65 (0.71)	98.25 (0.45)	99.06 (0.52)	98.76 (0.47)
MFSAN	91.43 (0.48)	90.54 (0.69)	91.17 (0.56)	99.27 (0.58)	98.03 (0.65)	98.77 (0.52)
M^3SDA	91.22 (0.52)	90.63 (0.49)	90.58 (0.54)	98.96 (0.63)	98.48 (0.46)	98.65 (0.43)
MTLBDA	92.24 (0.35)	91.89 (0.42)	92.03 (0.43)	99.01 (0.45)	99.28 (0.51)	98.68 (0.49)
		A, D, H- > W			W, D, H- > A	
DCTN	97.67 (0.65)	98.82 (0.57)	99.03 (0.55)	92.71 (0.67)	90.37 (0.85)	91.63 (0.76)
MFSAN	99.48 (0.38)	98.37 (0.69)	99.08 (0.43)	91.54 (0.75)	93.26 (0.82)	94.14 (0.65)
M^3SDA	99.31 (0.48)	99.15 (0.51)	98.78 (0.62)	93.72 (0.68)	92.63 (0.59)	94.26 (0.63)
MTLBDA	99.52 (0.47)	98.63 (0.54)	98.75 (0.61)	94.53 (0.55)	92.91 (0.72)	93.86 (0.53)

4.2.2. Experimental Data. In this part, we compare some single-source transfer learning algorithms and multisource transfer learning algorithms such as DCTN and MFSAN with our algorithm MTLBDA.

It can be seen from Table 3 that among the four cross-domain tasks, the highest accuracy of the MTLBDA algorithm is 93.03%, 99.28%, 99.52%, and 94.53%, which are higher than those of the comparative algorithms DCTN, MFSAN, and M^3SDA. At the same time, compared with the four cross-domain tasks, the single-source transfer learning algorithm shows higher accuracy than the optimal classification task by 4.25%, 0.82%, 3.16%, and 2.77%. In the A,W,D - > H task, MFSAN improved by 1.6% compared with the best transfer learning algorithm and by 3.25% compared with the best single-source learning algorithm. In contrast, the average accuracy is greatly improved, which proves the effectiveness of the proposed algorithm.

4.3. Influence of Category and μ.

To demonstrate the advantages of our algorithm category, we selected the network dataset domain proposed in Ref. [27] (as shown in Figure 5); the fields of the dataset are clipart, infographic, painting, quickdraw, real, and sketch, including 345 classes and 599859 data. The data distribution is shown in Table 4. Each domain contains 345 classes. We gradually increase the number of classes from 20 to 345 and show the impact of the number of iterations on the accuracy of the algorithm. Finally, we calculate the sensitivity of the algorithm to μ.

4.3.1. Experimental Data

(1) Category Influence. We plot how the performances of different models will change when the number of categories increases. The figure shows all multidomain combinations. Under DomainNet, it can be seen from Figure 6(a)) that the multisource transfer learning algorithm is very sensitive to the number of classes. At the same time, when there are many classes, our algorithm is clearly better than the algorithms DCTN and MFSAN. (b) When the number of classes is greater than 150, our algorithm's accuracy is generally higher than that of other algorithms. (c) Our

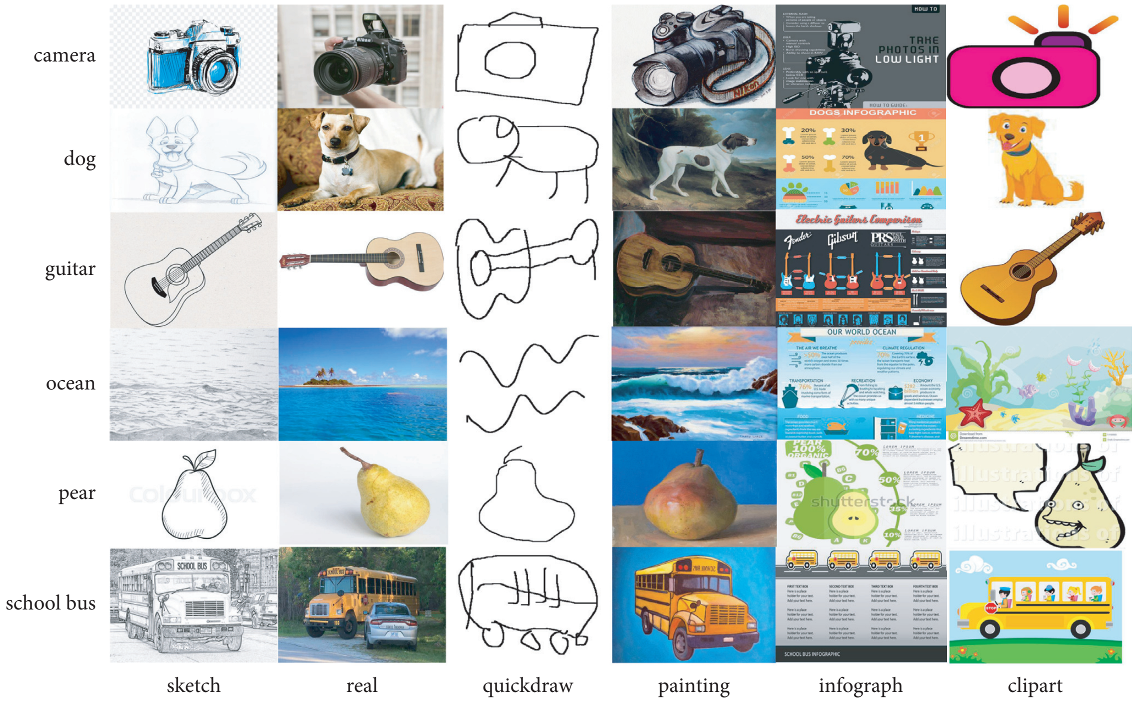


FIGURE 5: Example of DomainNet pictures.

TABLE 4: Data distribution example of DomainNet.

Domain	Category number	Total of data	Number of data of each subcategory (example)		
			Furniture	Mammal	Tool
Clipart	345	48921	5802	3437	3812
Infographic	345	53779	6513	3602	3096
Painting	345	76794	5002	8982	5124
Quickdraw	345	173500	17500	12500	14000
Real	345	70465	17104	15538	12938
Sketch	345	599859	7529	5151	4876

algorithm has a better effect on datasets with a very large difference between edge distribution and conditional probability distribution, such as DomainNet, which also proves that marginal probability and conditional probability have a great impact on classification in practical images.

(2) Influence of Iterations. Figure 7 shows the effect of the number of iterations on the accuracy. (a) When the number of iterations exceeds 1000, the accuracy of the algorithm tends to be stable. (b) At the same time, MTLBDA shows better results.

(3) Influence of μ. In this section, we will evaluate the effectiveness of the balance factor μ. We used running $\mu \in \{0, 0.1, \ldots, 0.9, 1.0\}$ in MTLBDA with $\mu = 0.5$ as the baseline on some tasks. Figure 8 shows the results. Clearly, the optimal μ is different in different tasks, indicating the importance of balancing the marginal distribution and conditional distribution between domains. In tasks C, I, P, Q, R- > S and C, I, P, S, R - > Q with optimal μ = 0.9, the marginal distribution is almost the same, so the performance of transfer learning mainly depends on the conditional distribution. In task U, M - > S with optimal μ = 0.4, the contribution of marginal distribution and conditional distribution is almost the same, but the marginal probability is more important. The observations were similar in other tasks. This shows that μ is essential for balancing marginal distribution and conditional distribution in cross-domain learning problems. Therefore, MTLBDA is more capable of obtaining good performance.

(4) Feature Visualization. From Figure 9, we can get the t-SNE plot of MTLBDA. Compared with the original state, we find that MTLBDA's clustering of features is very compact, which indicates MTLBDA-learned features have very ideal discrimination characteristics.

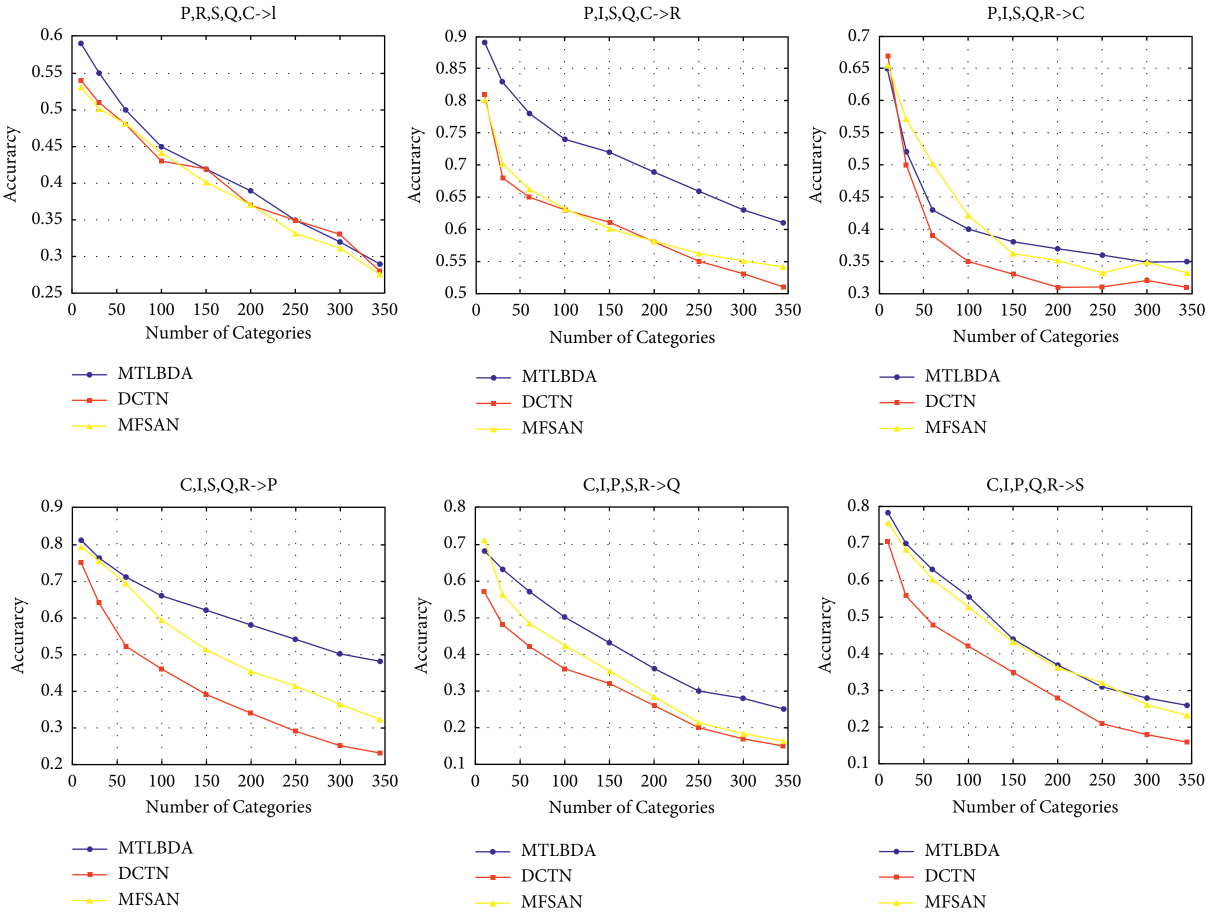


Figure 6: Comparison on DomainNet.

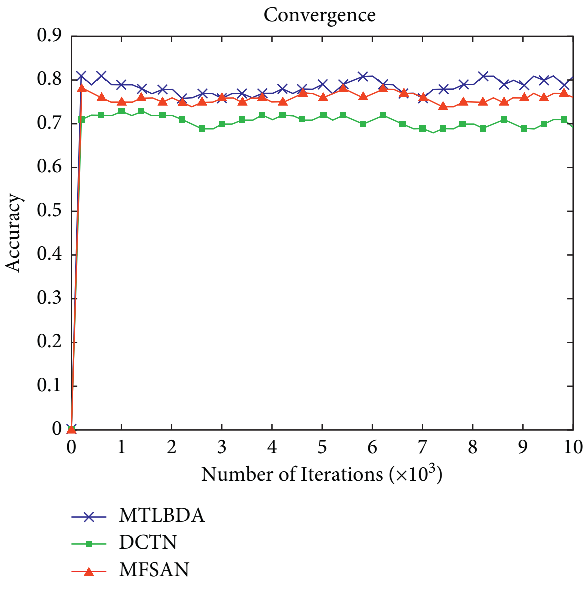


Figure 7: Effect of iteration time on accuracy.

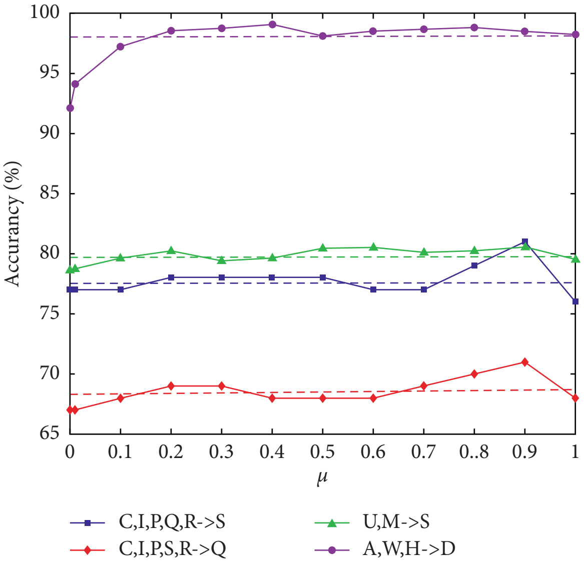
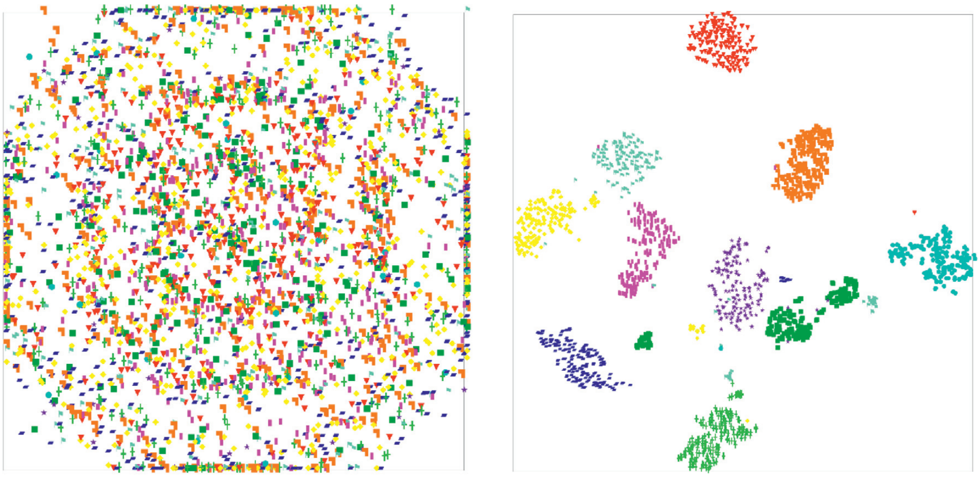


Figure 8: Effect of μ on accuracy.



origin

MTLBDA

Figure 9: T-SNE plot of Caltech.

5. Conclusion

In this article, to solve the problem of unbalanced categories in multiple source fields in transfer learning, a small sample data classification technique based on category adaptation balance and multisource transfer learning is proposed. Under the condition of unbalanced distribution, this method first maps multiple source domains and target domains to the same target space. Then, according to the balanced distribution adaption algorithm, the distribution in each source domain and target domain is balanced while adjusting its marginal distribution and conditional distribution. Then the convolutional neural network is used as the classifier for each source domain and target domain. Finally, the regularization term of each source domain is added to prevent overfitting of the model. The experimental results on the SVHN dataset, USPS dataset, MNIST dataset, Office-31 dataset, Caltech-256 dataset, and DomainNet dataset show that MTLBDA is superior to the benchmark algorithm in classification accuracy and training efficiency. Although the experimental results show that the MTLBDA algorithm is better than the benchmark algorithm, in the future, further research is still needed in the following area: the expansion of MTLBDA to the multiclassification problem; the accurate estimation of μ is also a challenge.

References

[1] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[3] M. Long, Z. Han, J. Wang, I. Michael, and Jordan, “Deep transfer learning with joint adaptation networks,” in *Proceedings of the 34th International Conferenceon Machine Learning, ICML 2017*, pp. 6–11, Sydney, NSW, Australia, August 2017.

[4] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and Trevor Darrell, “Deep domain confusion: maximizing for domain invariance,” 2014, arXiv preprint arXiv:1412.3474.

[5] M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *Proceedings of the 32nd International Conference on Machine Learning, Volume 37 of Proceedings of Machine Learning Research*, F. Bach and D. Blei, Eds., pp. 97–105, PMLR, Lille, France, 2015.

[6] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” *AAAI*, vol. 6, p. 8, 2016.

[7] X. Peng and K. Saenko, “Synthetic to real adaptation with generative correlation alignment networks,” in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*, Lake Tahoe, NV, USA, March 12-15, 2018.

[8] M. N. A. Khan and D. R. Heisterkamp, “Adapting instance weights for unsupervised dom ain adaptation using quadratic mutual information and subspace learning,” in *Proceedings of the A pattern Recognition(ICPR),2016 23rd International Conference on*, pp. 1560–1565.

[9] B. Zadrozny, “Learning and evalua ting classifiers under sample selection bias,” in *Proceedings of the 21st international*

conference on Machine learning, p. 114, ACM, Alberta, Canada, September 2004.

[10] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample selection bias correction theory," in *Lecture Notes in Computer Science, Ternational Conference on Algo-Rithmic Learning Theory*, pp. 38–53, Springer, NY, USA, 2008, Lecture Notes in Computer Science.

[11] W. Dai, Q. Yang, G.-R. Xue, and Y. Y u, "Boosting for transfer learning," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 193–200, Corvalis, OR, USA, June 2007.

[12] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 97–105, Lille, France, 2015.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research*, F. Bach and D. Blei, Eds., pp. 1180–1189, PMLR, Lille, France, 07–09 Jul 2015.

[14] E. Tzeng, J. Hoffman, K. Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Honolulu, HI, USA, July 2017.

[15] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer Feature Learning with Joint Distribution Adaptation," in *Proceedings of the International Conference on Computer Vision*, pp. 2200–2207, Sydney, NSW, Australia, December 2013.

[16] C.-A. Hou, Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, "Unsupervised domain adaptation with label and structural consistency," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5552–5562, 2016.

[17] J. Tahmoresnezhad and S. Hashemi, "Visual Domain Adaptation via Transfer Feature Learning," *Knowl. Inf. Syst*, 2016.

[18] E. Tzeng, J. Hoffman, N. Zhang, and S Kate, "Deep Domain Confusion: Maximizing for Domain Invariance," 2014, arXiv preprint arXiv:1412.3474.

[19] Y. Zhu, F. Zhuang, J. Wang et al., "Deep subdomain adaptation network for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 99, pp. 1–10, 2020.

[20] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, "Transfer learning with dynamic distribution adaptation," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 1, pp. 1–25, 2020.

[21] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domian adpation," *NeuIPS*, vol. 31, pp. 8559–8570, 2018.

[22] R. Xu, Z. Chen, W. Zuo, J Yan, and L Liang, "Deep cocktail network: multi-source unsupervised domain adaptation with category shift," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, Salt Lake City, UT, USA*, June 2018.

[23] C.-A. Hou, Y.-R. Yeh, and Y.-C. F. Wang, "An unsupervised domain adaptation approach for cross-domain visual classification," in *Proceedings of the 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, Karlsruhe, Germany, August 2015.

[24] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *Journal of Machine Learning Research*, vol. 9, pp. 1757–1774, 2008.

[25] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," *Advances in Neural Information Processing Systems*, pp. 1041–1048, Vancouver, BC, Canada, 2009.

[26] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[27] X. Peng, Q. Bai, and X. Xia, "Moment Matching for Multi-Source Domain Adaptation," in *2019 Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE Computer Society*, pp. 1406–1415, Seoul, South Korea, October 2019.

[28] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5989–5996, 2019.

[29] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Sch¨olkopf, and A. Smola, "A kernel two-sample test," *JMLR*, vol. 13, no. Mar, pp. 723–773, 2012.

[30] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *AAAI*, vol. 6, p. 8, 2016a.

[31] A. Gretton, K. M. Borgwardt, M. Rasch, B. Sch¨olkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems*, vol. 19, pp. 513–520, 2007.

[32] H. Zhao, R. T. d combes, K. Zhang, and G. J. Gordon, "On learning invariant representation for domain adaptation," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 12985–12999, Long Beach, CA, USA, June 2019.

[33] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[34] X. Li, H. Xiong, and H. Wang, "DELTA: DEEP LEARNING TRANSFER USING FEATURE MAP WITH ATTENTION FOR CONVOLUTIONAL NET- WORKS," in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 2019.

[35] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning," Nips Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain, 2011.

[36] X. Li, M. Fang, and J.-J. Zhang, "Projected Transfer Sparse Coding for cross domain image representation," *Journal of Visual Communication and Image Representation*, vol. 33, pp. 265–272, 2015.

[37] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: a general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2014.

[38] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proceedings of 2012 the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012.

[39] P. Gao, J. Li, and C. Ding, "Multisource mobile transfer learning algorithm based on dynamic model compression," *Security and Communication Networks*, vol. 2022, Article ID 3234078, 12 pages, 2022.

[40] P. Gao and J. Li, "Multi-source fast transfer learning algorithm base on support vector machine," *Applied Intelligence*, vol. 51, no. 11, pp. 8451–8465, 2021.

CITICON ENGINEERS LTD

PLOTTING

Citicon Engineers has now come up with this new township to give ownership of the land to aspiring and potential buyers.However, most of the buyers are sceptical about the authenticity of land they purchase;

INDEPENDENT HOUSES

A way of living that is kinder to the planet. Citicon Engineers belief has inspire to create a lush,tropical villa community where nature and innovation walk hand in hand. Where the gift of solar energy...

APARTMENT

Citicon Engineers focuses on building aesthetically and structurally remarkable projects. A full service landholding development company that offers the best commercial properties and residential properties .