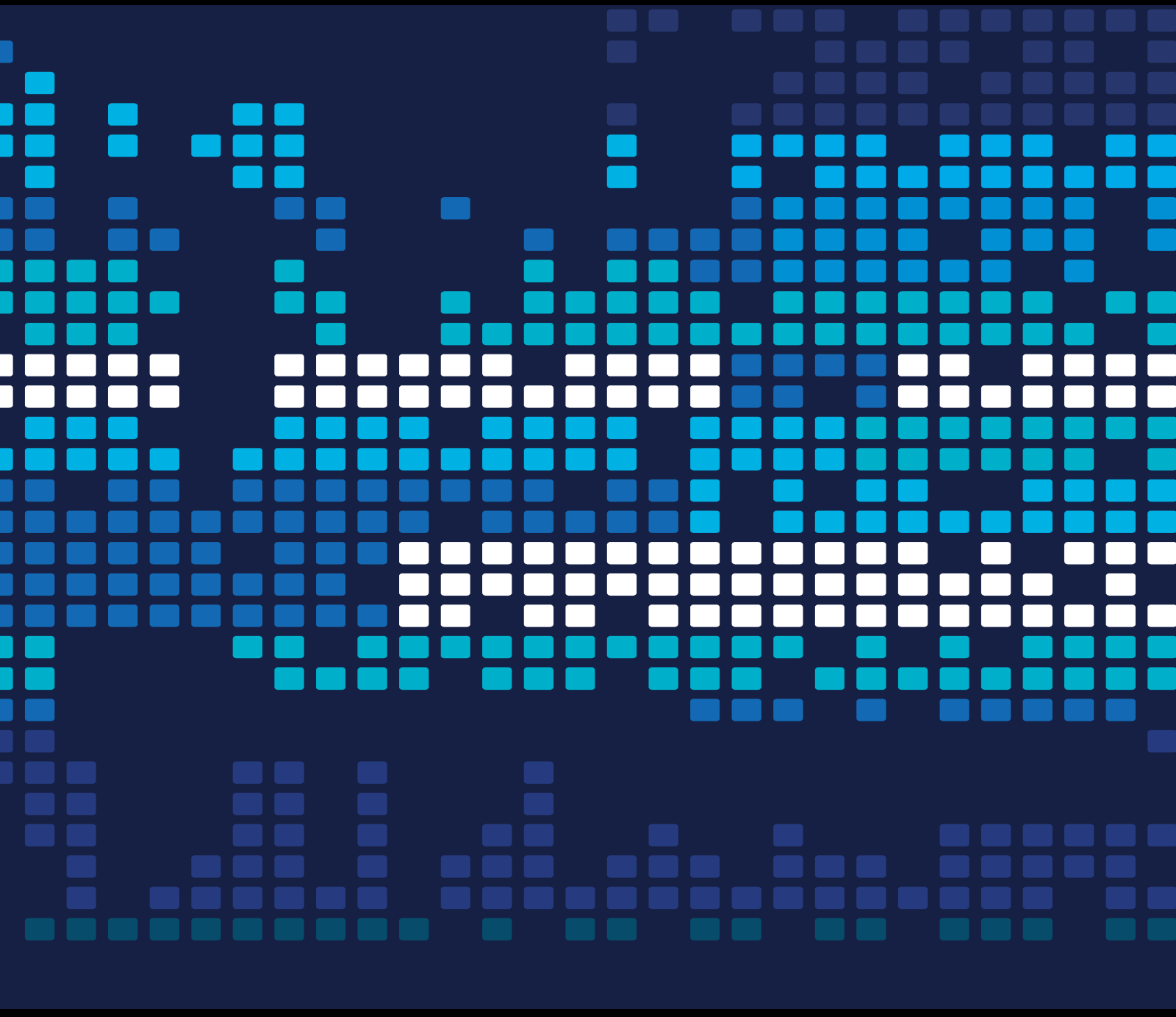
Scientific Programming

Scientific Programming Towards a Smart World 2020

Lead Guest Editor: Wenbing Zhao
Guest Editors: Chenxi Huang and Chao Huang



Scientific Programming Towards a Smart World 2020

Scientific Programming Towards a Smart World 2020

Lead Guest Editor: Wenbing Zhao
Guest Editors: Chenxi Huang and Chao Huang

Contents

Contents

Hindawi
Scientific Programming
Volume 2020, Article ID 8879086, 9 pages
https://doi.org/10.1155/2020/8879086

Research Article

Moving Window Differential Evolution Independent Component Analysis-Based Operational Modal Analysis for Slow Linear Time-Varying Structures

Yongshuo Zong,[1,2] **Jinling Chen**,[1] **Siyi Tao**,[1] **Cheng Wang**,[3] **and Jianbing Xiahou**[1]

[1]*School of Informatics, Xiamen University, Xiamen 361005, China*
[2]*Department of Computer Science and Technology, Tongji University, Shanghai 201804, China*
[3]*College of Computer Science and Technology, Huaqiao University, Xiamen, 361021, China*

Correspondence should be addressed to Cheng Wang; wangcheng@hqu.edu.cn and Jianbing Xiahou; jbxiahou@xmu.edu.cn

Received 20 April 2020; Revised 8 October 2020; Accepted 19 October 2020; Published 23 November 2020

Academic Editor: Chenxi Huang

In order to identify time-varying transient modal parameters only from nonstationary vibration response measurement signals for slow linear time-varying (SLTV) structures which are weakly damped, a moving window differential evolution (DE) independent component analysis- (ICA-) based operational modal analysis (OMA) method is proposed in this paper. Firstly, in order to overcome the problems in traditional ICA-based OMA, such as easy to go into local optima and difficult-to-identify high-order modal parameters, we combine DE with ICA and propose a differential evolution independent component analysis- (DEICA-) based OMA method for linear time invariant (LTI) structures. Secondly, we combine the moving widow technique with DEICA and propose a moving window differential evolution independent component analysis- (MWDEICA-) based OMA method for SLTV structures. The MWDEICA-based OMA method has high global searching ability, robustness, and complexity of time and space. The modal identification results in a three-degree-of-freedom structure with slow time-varying mass show that this MWDEICA-based OMA method can identify transient time-varying modal parameters effectively only from nonstationary vibration response measurement signals and has better performances than moving window traditional ICA-based OMA.

1. Introduction

It is hoped that the engineering structure will have a high degree of self-adaptive or self-control ability, and it can automatically change its shape and performance to adapt to changes in environmental disturbances and new usage requirements as the environment or use functions change. Operational Modal Analysis (OMA) can estimate modal parameters without input data, which is not easy to obtain in large-scale engineering structures [1]. The modal parameters of a linear time-varying structure can reflect the overall dynamics of the structure [2].

Blind Source Separation (BSS) is a technique that recovers source signals from mixed signals without prior knowledge [3]. To cite a few, independent component analysis (ICA) [4, 5], second-order blind identification (SOBI) [6], complexity pursuit (CP) [7, 8], and sparse component analysis (SCA) [9] are typical BSS methods. The ICA method requires that the source signals be independent of each other and the number of sensors is not less than the number of source signals [10]. Kerschen [11] is the first to recognize the mapping between the modal shapes and the columns of the mixing matrix in BSS and, then, apply the ICA method to OMA. Some methods have been proposed to improve the ICA's restrictions [12, 13]. However, because its optimization method is easy to go into local optima, traditional ICA-based OMA has low robustness, and it is difficult to identify high-order modal parameters using that [14]. Proposed by Strorn and Price [15, 16] firstly, differential evolution (DE) has much higher robustness and stronger global searching ability compared to the classical optimization method, such as the Gradient descent [17], stochastic

gradient descent [18], and quasi-Newton iteration method [19].

However, because the mixing matrix of BSS is a constant matrix, BSS-based algorithms can only be used for time invariant structures. In reality, most of the structural parameters change with time and are time-varying structures. The main research on the time-vary structure is based on the frozen-in coefficient method which is considering the time-varying system as time invariant over a short time. Ramnath defines the slow linearly time-varying (SLTV) structures as the change of the system coefficient is much slower than the change of the system solution [20]. The three main approaches of SLTV structures are time-frequency analysis [21], forgetting factor [22], and moving window [23, 24]. However, the time-frequency analysis method is not suitable for online identification. The performance of forgetting factor methods is closely related to prior knowledge, which is very troublesome.

Based on the moving window [24] and independent component analysis [15], this paper presents a moving window differential evolution independent component analysis- (MWDEICA-) based OMA method for weakly damped SLTV structures.

The primary contributions of this paper can be summarized as follows:

(1) In this paper, we propose a new DEICA-based OMA method to identify modal parameters only from stationary random response signals for LTI structures. Using DE algorithm to replace the traditional linear regression optimization algorithm to search the separation matrix, the DEICA-based OMA method has high global searching ability, robustness, and complexity of time and space. Compared with the traditional ICA method based on gradient optimization algorithm, OMA based on DEICA can identify higher-order modal parameters and has high recognition accuracy.

(2) A sliding window algorithm based on differential evolution independent analysis of sliding window is proposed by combining sliding window with the DEICA method. The instantaneous modal natural frequency and mode shape of the linear time-varying structure are identified by using the linear tracking characteristics of sliding window. This method can effectively identify the transient time-varying modal parameters only from the nonstationary random response measurement signals, which is better than the traditional moving window method. The modal identification method has better performance.

(3) We design a three-degree-of-freedom structure with slowly time-varying mass simulation to verify the validity and recognition accuracy of the MWDEICA-based OMA algorithm.

The remainder of this paper is organized as follows. In section 2, we describe MWDEICA in detail. The simulation verification is demonstrated in section 3. Finally, we conclude this paper in section 4.

2. Theoretical Inference of Algorithm

2.1. Differential Evolution Independent Component Analysis-Based OMA for the LTI Structure. The gradient-based optimization algorithm used by the traditional ICA is easy to fall into the local optimum. Therefore, there are problems such as low robustness and inability to recognize high-order modes when performing modal analysis on a continuum structure. Differential evolution algorithm has the ability to solve global optimization problems. The DEICA-based OMA method uses DE as an optimization algorithm which replaces the traditional linear regression method, and the core is to optimize the separation matrix $\mathbf{W}$. Therefore, DEICA-based OMA can identify high-order modes and has higher precision. The specific steps are as follows:

(1) The number of modes m to be extracted, the number of structural finite element nodes N, the number of response data collection points T, and the response matrix of the structure $\mathbf{X}(t) \in \mathbb{R}^{N\times T}$ are determined, and then, $\mathbf{Y}(t) = \mathbf{W}\mathbf{X}(t)$. $\mathbf{Y}(t) \in \mathbb{R}^{m\times T}$ and $\mathbf{W} \in \mathbb{R}^{m\times N}$.

(2) The steps of the differential evolution algorithm are adjusted. Since $\vec{w}_r$ $(r = 1, 2, \ldots, m)$ needs to be extracted one by one, each individual $\vec{p}_i$ in the population is a $1 \times N$ vector, which is stored in the population matrix $\mathbf{P}$. $\mathbf{P} = (\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_i, \ldots, \vec{p}_{\mathrm{NP}})^T \in \mathbb{R}^{\mathrm{NP}\times N}$ and $\vec{p}_i = (p_i(1), p_i(2), \ldots, p_i(j), \ldots, p_i(N))^T$, $j = 1, 2, \ldots, N; i = 1, 2, \ldots, \mathrm{NP}$.

(3) $\vec{p}_i$ is zero-centered by $\vec{p}_i = \vec{p}_i - \mathbf{P}\mathbf{P}^T\vec{p}_i$, $\vec{p}_i = (\vec{p}_i/|\vec{p}_i|)$ and starts loop. The kurtosis is used as a function, which is $\mathrm{kurt}(\vec{y}_i) = E\{(\vec{y}_i)^4\} - 3(E\{(\vec{y}_i)^2\})^2$, $\vec{y}_i = \vec{p}_i\mathbf{X}(t)$. Each offspring of results is also zero-centered. Convergence condition of the algorithm: the maximum number of iterations (4000 times) is exceeded or $|\vec{p}_{\mathrm{best}}^{(g)T} \cdot \vec{p}_{\mathrm{best}}^{(g+1)} - 1| < 10^{-9}$. The optimal solution is $\vec{p}_{\mathrm{best}}$, and let $\vec{w}_r = \vec{p}_{\mathrm{best}}$.

(4) $r = r + 1$; return to step (2) until $r > m$.

The flow diagram is shown in Figure 1.

2.2. Nonstationary Response Signals Decomposition in the Modal Coordinate and MWDEICA-Based OMA for SLTV. Based on the "time-freezing" theory [24], the frozen-in coefficient method, the "short time invariant," and "quasistationary" assumptions, the nonstationary random response signals of the SLTV structures can be approximated as the stationary random response time series of LTI structures in short time interval. Figure 2 shows the length selection of moving window.

The choice of moving window length is fixed (the length is L), and the displacement responses $\mathbf{X}_L^{(i)} \in \mathbb{R}^{N\times L}$ can be decomposed by the modal expansion:

$$\mathbf{X}_L^{(i)} \approx \mathbf{\Phi}_L^{(i)}\mathbf{Q}_L^{(i)}, \quad i = 1, 2, \ldots, T + 1 - L, \tag{1}$$

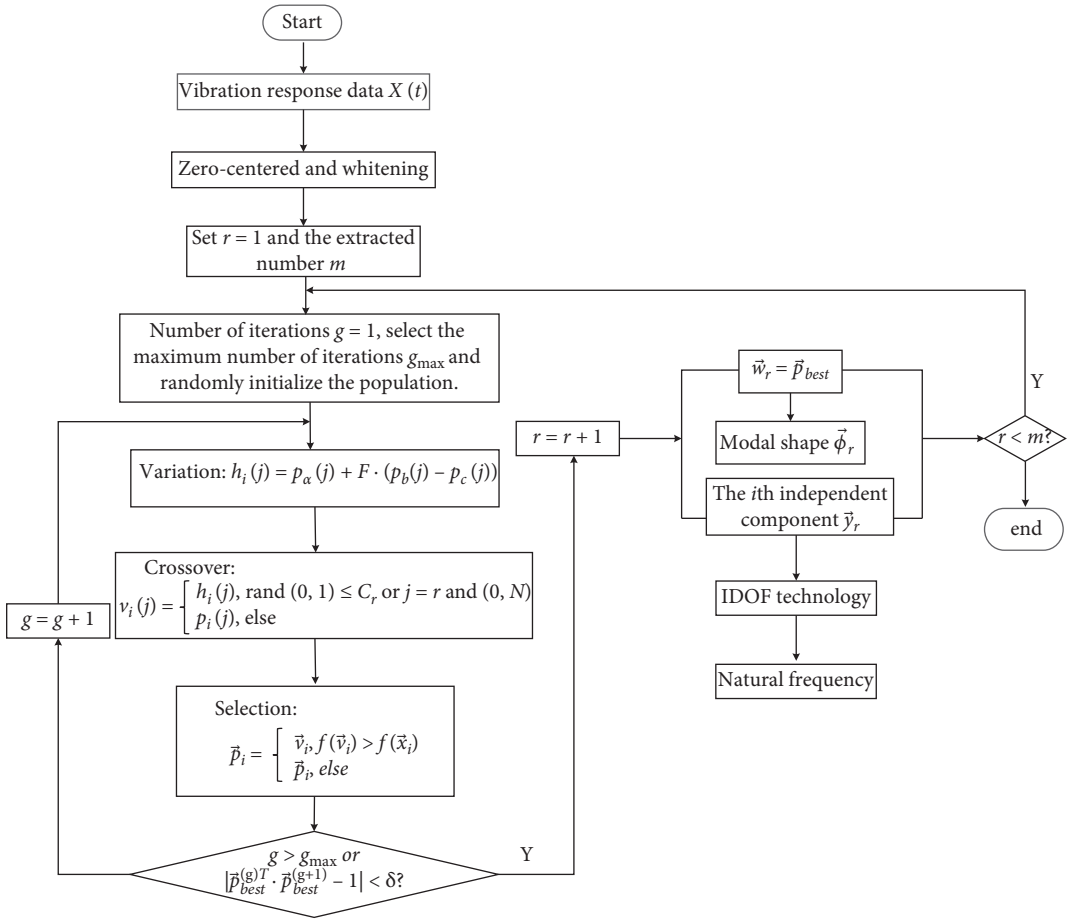


Figure 1: The flow diagram of DEICA-based OMA for the LTI structure.

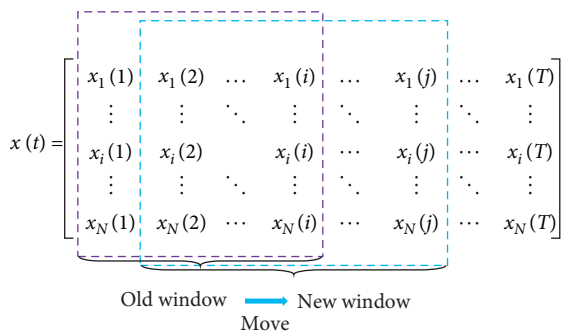


Figure 2: The description of fixed length moving window.

where $\Phi_L^{(i)} = [\vec{\phi}_1^{(i)}, \ldots, \vec{\phi}_{k'}^{(i)}, \ldots, \vec{\phi}_N^{(i)}] \in \mathbb{R}^{N \times N}$ is the modal shapes matrix and $\mathbf{Q}_L^{(i)} = [\vec{q}_1^{(i)}, \ldots, \vec{q}_{k'}^{(i)}, \ldots, \vec{q}_N^{(i)}] \in \mathbb{R}^{N \times L}$ is the modal coordinate response matrix. When the order of modal natural frequencies $f_{k'}^i$ is not equal, the modal coordinate response vectors $\vec{q}_{k'}^{(i)}$ are independent of each other.

From equation (1), we can see that $\Phi_L^{(i)}$ and $\mathbf{Q}_L^{(i)}$ represent, respectively, the LTV structure's statistical average modal shapes and modal response over the window time L. When time reaches $t_{i+(L-1)/2}$, the predicted modal shape vector $\vec{\phi}_{k'}(i + (L-1)/2)$ approximates to $\vec{\phi}_{k'}^{(i)}$. Similarly, after using the single-degree-of-freedom (DOF) technique, the nature frequency $f_{k'}(i + (L-1)/2)$ of k'th order at time $t_{i+(L-1)/2}$ can also be replaced by $f_{k'}^{(i)}$.

In view of equation (1), there is a one-to-one relationship between the modal response matrix $\Phi_L^{(i)}$ of weakly damped structures and the linear transformational matrix $\mathbf{W}_L^{(i)}$ in MWICA and a one-to-one mapping between the independent components $\mathbf{Y}_L^{(i)}$ of MWICA and the modal response matrix $\mathbf{Q}_L^{(i)}$ in equation (1). Through ICA decomposition, the uniqueness, certainty, and existence of the MWDEICA-based OMA method can be proved and shown in Figure 3.

2.3. Theoretical Analysis and Comparison between Traditional ICA and DEICA. As an optimization solution method, DE has much higher robustness and stronger global searching ability compared to the classical optimization method such as gradient descent, random gradient descent, and Newton iteration. Table 1 shows the characteristics of ICA and DEICA.

2.4. Application Scopes of the Proposed Method. The application scope of the OMA method based on moving window and DEICA is as follows:

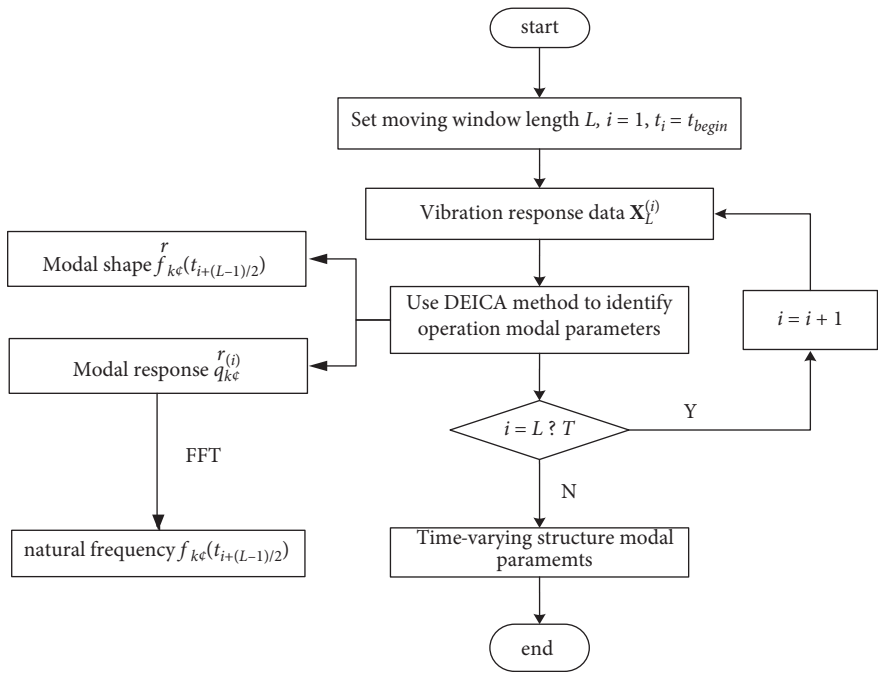


Figure 3: Process of MWDEICA-based OMA algorithm for SLTV structures.

Table 1: Theoretical analysis and comparison between traditional ICA and DEICA.

	Traditional ICA	DEICA
Fall into local optimum	Easy	Not easy
Global searching ability	Low	High
Robustness	Low	High
Complexity of time and space	Low	High

(1) The system should be a slow linear time-varying (SLTV) structure which is weakly damped.

(2) Because of the limitation of blind source separation and independent component analysis, the number of vibration response sensors should be greater than or equal to the identification order of identifiable operational modal parameters.

(3) Because of the limitation of sampling theorem, the sampling frequency of the vibration response signal should be greater than or equal to 2 times of the natural frequency of the identifiable modal natural frequency.

(4) Every order of modal shape is with different amplitude. In case of the ICA method, the energy of separation matrix is not unique, and independent component also loses amplitude information. Unlike the principal component analysis method [25], the ICA method cannot get the contribution ratio information of each modal. Modal shape is a relative quantity rather than an absolute value. So, in order to compare the modal shape with the real modal shape, the separation matrix and modal shape identified by the ICA method should be normalized.

(5) The order of modal parameter identified by ICA is uncertain. The modal identified by the ICA method is not in accordance with the order from small to large. In fact, the first separated output source and vector are the ones whose independence is the strongest rather than the first-order modal parameter. Therefore, in order to compare the natural frequencies with the real natural frequencies, the modal parameters identified by ICA need to be reordered by the modal frequencies.

3. Simulation Identification

3.1. Simulation Dataset of a Three-Degree-of-Freedom Structure with Slow Time-Varying Mass. To confirm MWICA's availability and identification accuracy, we have designed a simulation of an LTV three-DOF spring oscillator system in MATLAB/Simulink. This model is shown in Figure 4. The initial conditions for the three modal displacements and velocities are zero. The stiffness matrix is set to $k_1 = 1000$ (N/m), $k_2 = 1000$ (N/m), and $k_3 = 1000$ (N/m), and the damping matrix is set to $c_1 = 0.01$ (N · s/m), $c_2 = 0.01$ (N · s/m), and $c_3 = 0.01$ (N · s/m). The external force $F(t) = [F_1(t)\ 0\ 0]^T$ is white noise excitation with zero mean and unit variance. The mass matrix is set to $m_2 = 1$ kg and $m_3 = 1$ kg, and m_1 is a time-varying parameter as follows:

$$m_1(t) = \begin{cases} 1, & 0 \le t \le 50, \\ e^{-0.0005(t-50)}, & 50 < t \le 2000. \end{cases} \quad (2)$$

This kinetic equation becomes

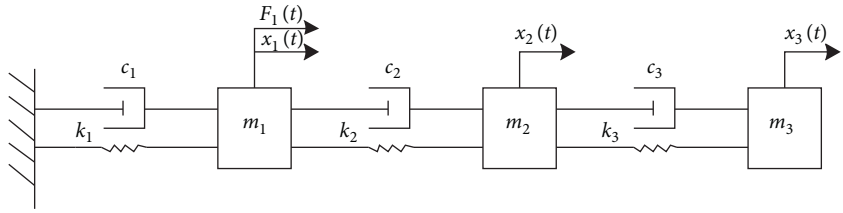


Figure 4: Linear time-varying three-degree-of-freedom spring vibrator model.

$$\begin{bmatrix} m_1(t) & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \begin{bmatrix} \ddot{x}_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} f(t) \\ 0 \\ 0 \end{bmatrix}. \tag{3}$$

In the Simulink model, we use the Runge–Kutta algorithm to obtain responses of 40 Hz sampling frequency and 2000 s recorded time [23]. In the time domain, 2.0% level Gauss measurement noise disturbances are added to displacement response signals.

3.2. Simulation Parameters Setting. According to the third theoretical frequency 12.2951 Hz at time1987.225 s in Table 2, the sampling frequency f is set to 40 Hz in simulation of the SLTV structure. The window length is set to $L = 1024$ in this simulation, and then, the frequency resolution of FFT $\Delta f = 0.04$ Hz.

$$\Delta f_L(1) \triangleq \frac{L}{f} \times \frac{f_{\text{end}}(1) - f_{\text{begin}}(1)}{t_{\text{end}} - t_{\text{begin}}} = \frac{1024}{40} \times \frac{2.31 - 2.24}{2000 - 50} \approx 9.25 \times 10^{-4} \text{ Hz}, \tag{4}$$

$$\Delta f_L(2) \triangleq \frac{L}{f} \times \frac{f_{\text{end}}(2) - f_{\text{begin}}(2)}{t_{\text{end}} - t_{\text{begin}}} = \frac{1024}{40} \times \frac{7.28 - 6.28}{2000 - 50} \approx 0.0131 \text{ Hz}, \tag{5}$$

$$\Delta f_L(3) \triangleq \frac{L}{f} \times \frac{f_{\text{end}}(3) - f_{\text{begin}}(3)}{t_{\text{end}} - t_{\text{begin}}} = \frac{1024}{40} \times \frac{12.33 - 9.07}{2000 - 50} \approx 0.0428 \text{ Hz}. \tag{6}$$

This objective function of ICA algorithm in this paper is the measurement of maximization non-Gaussian (kurtosis). The optimized method of class ICA algorithm is the quasi-Newton iteration method. The ICA method is used multiple times simultaneously, and the maximum iterations time is 4000. The modal parameters can be identified most of the time in the case of 60 runs.

The parameter setting of DE: the individual number is 80; maximum iterations time is 4000; and proportionality factor F and cross probability C_r are random transformation each loop repeats.

3.3. Modal Assurance Criterion. To determine the accuracy of the modal shapes, the modal assurance criterion (MAC) is introduced to verify the operational modal identification results. The MAC is given by [26]

$$\text{MAC}_{\overrightarrow{\varphi}_i, \overrightarrow{\varphi}_j} = \frac{\left(\overrightarrow{\varphi}_i^T \overrightarrow{\varphi}_j\right)^2}{\left(\overrightarrow{\varphi}_i^T \overrightarrow{\varphi}_j\right)\left(\overrightarrow{\varphi}_i^T \overrightarrow{\varphi}_j\right)}, \tag{7}$$

where $\overrightarrow{\varphi}_i$ is the i^{th} identified modal shape, $\overrightarrow{\varphi}_j$ is the j^{th} theoretical modal shape, and $\overrightarrow{\varphi}_i^T$ and $\overrightarrow{\varphi}_j^T$ are the transpositions of $\overrightarrow{\varphi}_i$ and $\overrightarrow{\varphi}_j$, respectively. The MAC values range between 0 (no coincidence) and 1 (complete coincidence). However, MAC only reveals information about the direction and shape of the modal results and does not contain any information about the modal amplitudes. The MAC value is always 1 if the two vectors are proportional.

3.4. Transient Operational Modal Parameters Identification Results. The two ICA algorithms were repeated 60 times, respectively. For the modal natural frequency, Figure 5 shows the theoretical transient natural frequencies and the identified natural frequencies by class MWICA and MWDEICA for the SLTV structure.

Figure 5 shows that certain modal parameters are not identified by class MWICA and MWDEICA at any time. In MWICA, the percentage of modal parameters of unidentified windows is more than 0.2% in the average of 60 trials.

We choose two moments $t = 50.25$ s and $t = 1350.00$ s to compare the theoretical modal shapes with the identified modal shapes. The recognition value of the frequency is similar in each simulation, while the modal shape recognition result fluctuates. We select the best value in 60 times.

Table 2: The comparison of transient natural frequencies at moments 50.025 s and 1987.225 s.

Modal order	50.025 s (Hz)	1987.225 s (Hz)	Change of transient frequency (%)
1	2.24	2.31	3.173
2	6.28	7.27	13.73
3	9.07	12.30	26.24

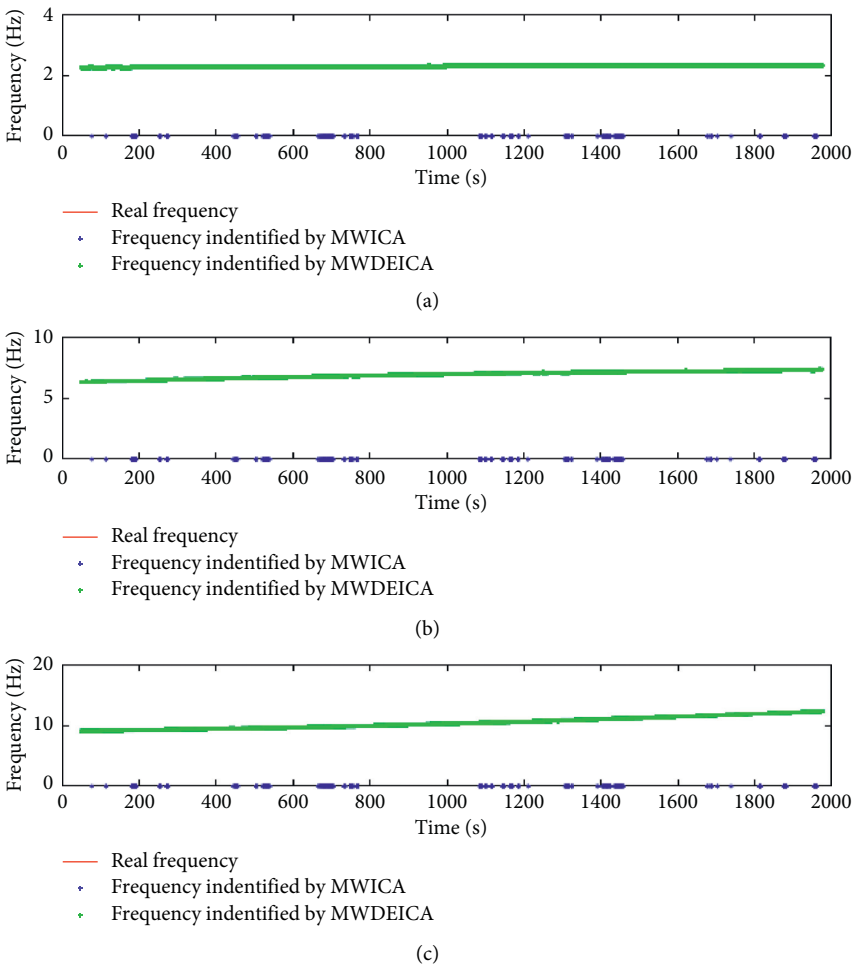


Figure 5: Comparison between the identified frequencies and theoretical frequencies. (a) Comparison of first-order modal frequencies, (b) Comparison of second-order modal frequencies, (c) Comparison of third-order modal frequencies.

3.5. Results Analysis

(1) Figures 5 and 6 and Tables 3 and 4 show that using only nonstationary random response signals, the MWICA-based and MWDEICA-based OMA method can well identify the transient natural frequency and modal shape of the SLTV structure.

(2) Figures 5, and 6 show that, in the case of adding 2.0% level Gauss measurement noise disturbances to displacement response signals, compared with the MWDEICA-based OMA method, the MVDEICA method can identify more modal parameters of windows, and the class MWICA-based OMA method has low robustness and it is difficult to identify modal parameters at some moment because its optimization method is easy to go into local optima [16]. For MWDEICA-based OMA, the recognition results are unstable.

(3) Table 5 shows that the MWDEICA-based OMA method has much higher time and space requirements than the MWICA-based OMA method.

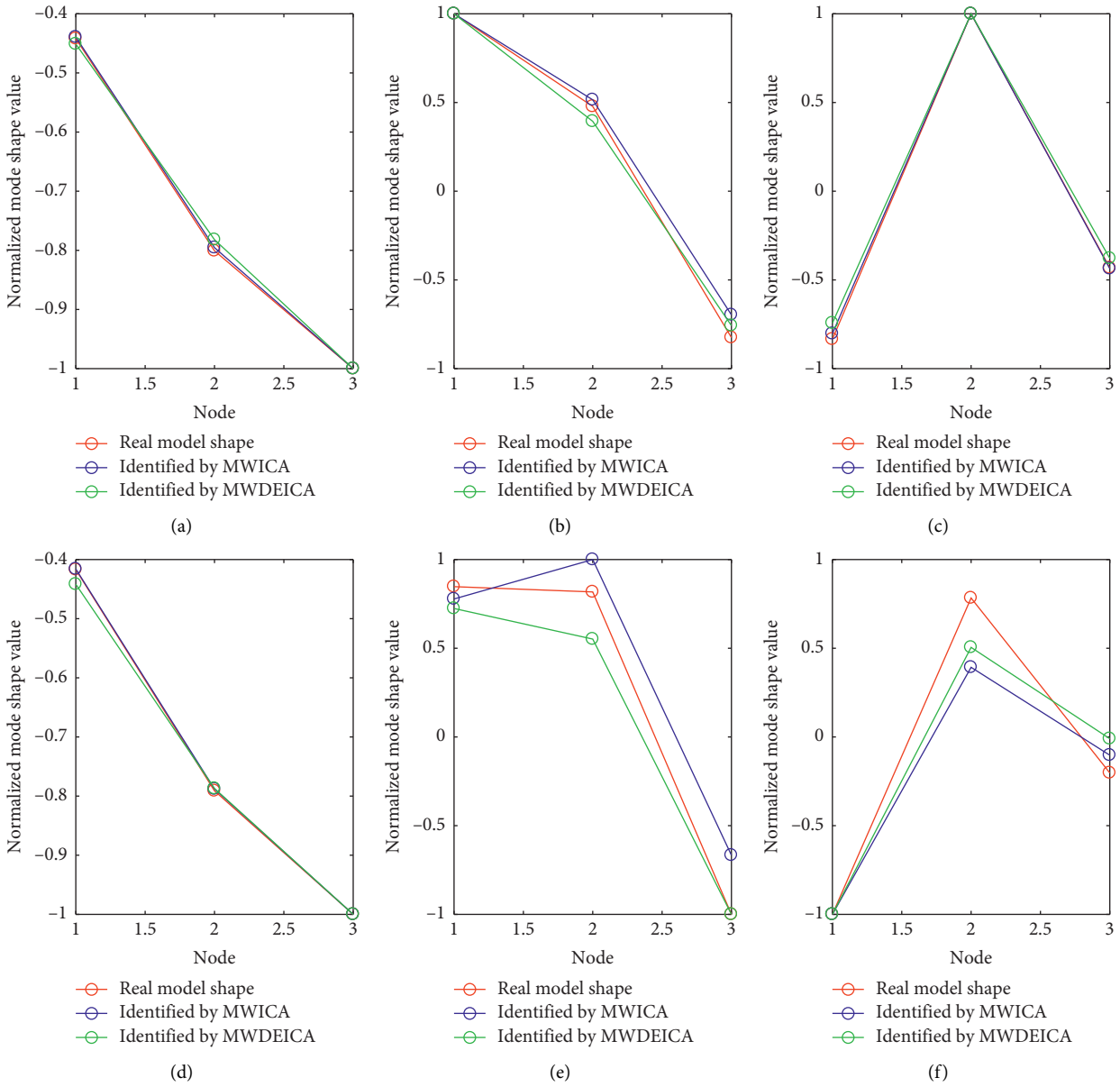


Figure 6: Comparison of the modal shape identified by MWICA and MWDEICA to the real modal shape at two moments. (a) When $t = 50.25$ s, the 1st modal shape, (b) when $t = 50.25$ s, the 2nd modal shape, (c) when $t = 50.25$ s, the 3rd modal shape, (d) when $t = 1350$ s the 1st modal shape, (e) when $t = 1350$ s the 2nd modal shape, and (f) when $t = 1350$ s the 3rd modal shape.

Table 3: The MAC values comparison at time 50.25 s.

Theory modal shape no.	MWICA identification order	MWDEICA identification order	MAC identified by MWICA	MAC identified by MWDEICA
1	2	1	1.0000	0.9998
3	1	3	0.9922	0.9981
2	3	2	0.9999	0.9979

Table 4: The MAC values comparison at time 1350 s.

Theory modal shape no.	MWICA identification order	MWDEICA identification order	MAC identified by MWICA	MAC identified by MWDEICA
1	1	1	1.0000	0.9998
2	3	2	0.9419	0.9734
3	2	3	0.9076	0.9326

Table 5: Time and space requirements of different methods.

Method	Time spent (second)	Space requirement (megabyte) (MB)
MWICA-based OMA	1700.3	103
MWDEICA-based OMA	6804.2	578

4. Conclusions

In order to solve the problem that the traditional ICA method is easy to fall into local optimum and difficult to identify high-order modal parameters, combining moving window technology and differential evolution independent component analysis, this paper proposes a transient operational modal parameters identification method only from nonstationary vibration response measurement signal s for slow linear time-varying structures which is weakly damped. Theoretical comparison and simulation identification results show that MWDEICA-based OMA can effectively identify time-varying modal parameters, while the parameters of some sample points are not recognized in MWICA.

Due to moving window technology and differential evolution algorithm, we are still unable to identify modal parameters in some time points. How to reduce the complexity of the algorithm and use MVDEICA to identify higher-order modals is the future work. It is also the future work to solve the problem and experimentally verify on more complex three-dimensional engineering structures.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by CERNET Innovation Project (grant no. NGII20160410).

References

[1] E. Reynders, "System identification methods for (operational) modal analysis: review and comparison," *Archives of Computational Methods in Engineering*, vol. 19, no. 1, pp. 51–124, 2012.

[2] K. Liu and M. R. Kujath, "Adaptation of the concept of modal analysis to time-varying structures," *Mechanical Systems and Signal Processing*, vol. 13, no. 3, pp. 413–422, 1999.

[3] W. Guan, L. L. Dong, J. M. Zhou, Y. Han, and J. Zhou, "Data-driven methods for operational modal parameters identification: a comparison and application," *Measurement*, vol. 132, pp. 238–251, 2019.

[4] J. Hang and Z. Yi, "Operational modal identification technique based on independent component analysis," in *Proceedings of the 2011 International Conference on Electric Technology and Civil Engineering (ICETCE)*, pp. 2430–2433, Lushan, China, April 2011.

[5] S. Chauhan, R. Martell, R. J. Allemang, and D. L. Brown, "Application of independent component analysis and blind source separation techniques to operational modal analysis," in *Proceedings of the 25th IMAC*, Orlando, FL, USA, February 2007.

[6] C. Rainieri, "Perspectives of second-order blind identification for operational modal analysis of civil structures," *Shock and Vibration*, vol. 2014, Article ID 845106, 2014.

[7] M. A. Eitner, B. G. Miller, J. Sirohi, and C. E. Tinney, "Operational modal analysis of a thin-walled rocket nozzle using phase-based image processing and complexity pursuit," in *Rotating Machinery, Optical Methods & Scanning LDV Methods*, vol. 6, pp. 19–29, Springer, Cham, Switzerland, 2019.

[8] J. Antoni, R. Castiglione, and L. Garibaldi, "Interpretation and generalization of complexity pursuit for the blind separation of modal contributions," *Mechanical Systems and Signal Processing*, vol. 85, pp. 773–788, 2017.

[9] H. Zhou, K. Yu, Y. Chen, R. Zhao, and Y. Wu, "Output-only modal estimation using sparse component analysis and density-based clustering algorithm," *Measurement*, vol. 126, pp. 120–133, 2018.

[10] M. D. Spiridonakos, E. N. Chatzi, and B. Sudret, "Polynomial chaos expansion models for the monitoring of structures under operational variability," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, vol. 2, no. 3, Article ID B4016003, 2016.

[11] G. Kerschen, F. Poncelet, and J.-C. Golinval, "Physical interpretation of independent component analysis in structural dynamics," *Mechanical Systems and Signal Processing*, vol. 21, no. 4, pp. 1561–1575, 2007.

[12] C. Wang, J. Wang, X. Lai et al., "Operation modal analysis following fast independent component analysis," *International Journal of Applied Electromagnetics and Mechanics*, vol. 52, no. 1-2, pp. 103–111, 2016.

[13] C. Wang, J. Wang, B. Zhong et al., "Negentropy and gradient iteration based fast independent component analysis for multiple random fault sources blind identification and separation," *International Journal of Applied Electromagnetics and Mechanics*, vol. 52, no. 1-2, pp. 711–719, 2016.

[14] J. Wang, C. Wang, T. Zhang, and B. Zhong, "Comparison of different independent component analysis algorithms for output-only modal analysis," *Shock and Vibration*, vol. 2016, pp. 1–25, 2016.

[15] R. Storn and K. Price, "Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[16] P. Civicioglu and E. Besdok, "A conceptual comparison of the cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms," *Artificial Intelligence Review*, vol. 39, no. 4, pp. 315–346, 2013.

[17] C. Burges, T. Shaked, E. Renshaw et al., "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning (ICML-05)*, pp. 89–96, Bonn, Germany, August 2005.

[18] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," *Advances in Neural Information Processing Systems*, vol. 23, no. 23, pp. 2595–2603, 2010.

[19] D. F. Shanno, "Conditioning of Quasi-Newton methods for function minimization," *Mathematics of Computation*, vol. 24, no. 111, p. 647, 1970.

[20] R. V. Ramnath, *Multiple Scales Theory and Aerospace Applications, Multiple Scales Theory and Aerospace Applications*, American Institute of Aeronautics and Astronautics, Reston, VA, USA, 2015.

[21] A. Klepka and T. Uh, "Identification of modal parameters of non-stationary systems with the use of wavelet based adaptive filtering," *Mechanical Systems and Signal Processing*, vol. 47, no. 1-2, pp. 21–34, 2014.

[22] W. Guan, C. Wang, D. S. Chen, X. Luo, and F. F. Su, "Recursive principal component analysis with forgetting factor for operational modal analysis of linear time-varying system," *International Journal of Applied Electromagnetics and Mechanics*, vol. 52, no. 3-4, pp. 999–1006, 2016.

[23] C. Wang, J. Wang, and T. Zhang, "Operational modal analysis for slow linear time-varying structures based on moving window second order blind identification," *Signal Processing*, vol. 133, pp. 169–186, 2017.

[24] C. Wang, H. Huang, J. Chen, W. Wei, and T. Wang, "An online and real-time adaptive operational modal parameter identification method based on fog computing in internet of things," *International Journal of Distributed Sensor Networks*, vol. 16, no. 1, pp. 1–12, 2020.

[25] C. Wang, J. Gou, J. Bai, and Y. Guirong, "Modal parameter identification with principal component analysis," *Journal of Xi'an Jiaotong University*, vol. 47, no. 11, pp. 97–104, 2013.

[26] J. Zou, K. Yu, and B. Yang, "Methods of time-varying structural parameter identification," *Advances in Mechanics*, vol. 30, no. 3, pp. 370–377, 2000.

Hindawi
Scientific Programming
Volume 2020, Article ID 4579291, 14 pages
https://doi.org/10.1155/2020/4579291

Research Article

An Intelligent, Secure, and Smart Home Automation System

Rizwan Majeed,[1] Nurul Azma Abdullah,[1] Imran Ashraf,[2] Yousaf Bin Zikria,[2] Muhammad Faheem Mushtaq,[1,3] and Muhammad Umer[4]

[1]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Parit Raja 86400, Johor, Malaysia*
[2]*Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea*
[3]*Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan*
[4]*Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan*

Correspondence should be addressed to Muhammad Faheem Mushtaq; faheem.mushtaq@kfueit.edu.pk

Received 18 February 2020; Revised 26 July 2020; Accepted 7 October 2020; Published 29 October 2020

Academic Editor: Chenxi Huang

The idea of a smart home is getting attention for the last few years. The key challenges in a smart home are intelligent decision making, secure identification, and authentication of the IoT devices, continuous connectivity, data security, and privacy issues. The existing systems are targeting one or two of these issues whereas a smart home automation system that is not only secure but also has intelligent decision making and analytical abilities is the need of time. In this paper, we present a novel idea of a smart home that uses a machine learning algorithm (Support Vector Machine) for intelligent decision making and also uses blockchain technology to ensure identification and authentication of the IoT devices. Emerging blockchain technology plays a vital role by providing a reliable, secure, and decentralized mechanism for identification and authentication of the IoT devices used in the proposed home automation system. Moreover, the SVM classifier is applied to classify the status of devices used in the proposed smart home automation system into one of the two categories, i.e., "ON" and "OFF." This system is based on Raspberry Pi, 5 V relay circuit, and some sensors. A mobile application is developed using the Android platform. Raspberry Pi acting as the server maintains the database of each appliance. The HTTP web interface and apache server are used for communication among the Android app and Raspberry Pi. The proposed idea is tested in the lab and real life to validate its effectiveness and usefulness. It is also ensured that the hardware and technology used in the proposed idea are cheap, easily available, and replicable. The experimental results highlight its significance and validate the proof of the concept.

1. Introduction

In this age of technology, the Internet of Things refers to the idiosyncratically identifiable objects and their virtual representation in cyberspace [1–4]. IoT is a pristine information processing and accretion method encompassing RFID [5] sensor machinery, smart technology, nanotechnology, and other technological advancements. IoT is not alone noble technology; instead, it circumvents considerable complementary technical development and provides capabilities that are appropriated together and aid to bridge the gap between the virtual and physical world [6]. Everyone is living a very busy life, and they want to get ease in every aspect of life. The Internet of Things (IoT) is a very vast field, and this research cannot cover up the whole thing to automate. But, the ease of people Home Automation is the first domain that comes in mind. Home Automation [7–10] is a term used to describe the working of all home appliances together, and we are controlling things by using an Android smartphone [11, 12], tablet, or computer having internet connectivity. The popularity of home automation is increased rapidly during the recent years. People like the comfort of maintaining and changing the status of appliances from any part of the world using remote access. It is eventually becoming the need of every person.

IoT provides more flexible and low-cost solutions for daily life problems which ultimately improve the user's life [13]. Although many previous researchers proposed many home automation systems by using different sensor combinations [14–18], as shown in Table 1, with the detailed description of previous studies provided, we may conclude that there are some reasons for the motivation to design an efficient home automation system.

(i) Previously proposed home automation systems are expensive and complex

(ii) Previously proposed Bluetooth home automation systems required intrusive installation

(iii) Previously proposed home automation systems use the connectivity of the Internet which may be not useable in those areas where the internet is not available

(iv) Previously proposed home automation systems have not focused on security and safety problems that may arise from their implementation

(v) The previously proposed home automation system has no intelligent decision-making ability

To assure home automation continues in a normal fashion, it is necessary to check all important parameters that may complicate the system, one of which is no GUI environment for the users. Users cannot understand the working of the system due to the lack of a GUI environment. Secondly, there is no device restoration which is harmful to home appliances. Thirdly, there is no prediction of the electricity bill for the users, and all the previous home automation solutions are very expensive. So, we provide an effective solution to overcome all these problems by undertaking the following objectives.

(i) The proposed solution is low cost because it did not use IP-based devices such as bulb and light.

(ii) The proposed solution provides electricity bill prediction to the users. The proposed solution uses GSM technology which means a Global System for Mobile Communication (GSM) modem to control home appliances such as light, conditional system, and security system via Short Message Service (SMS) text messages.

(iii) The proposed solution provides device restoration which means when you restore a computer or other electronic device, you return it to a previous state.

(iv) The proposed solution is a cloud-based solution which means it controls home appliances through the Internet.

(v) The proposed solution provides Arduino and Raspberry implementation. For those who love to tinker with electronics, the Raspberry Pi and Arduino have become invaluable tools. They are both extremely popular options but are also very budget-friendly. Arduino is best used for real-time applications of hardware and software, and Raspberry Pi easily connects to the Internet.

(vi) The proposed solution provides a prototype home design for the ease of the users.

(vii) The proposed solution has an intelligent decision-making ability to classify the status of IoT devices used in the smart home automation system.

(viii) The proposed solution ensures secure identification and authentication of users using blockchain technology.

The rest of the paper is structured as follows: Section 2 highlights the related work and key contributions in the area of smart home and smart buildings. Section 3 discusses the used approach, and Section 4 presents the implementation details of the used approach along with used hardware and software. Section 5 presents the results of the experiments along with the discussion. The paper is concluded in Section 6.

2. Related Work

The research gap identified in previous studies in the domain of smart homes and smart buildings is reported in Table 1.

Table 1 clearly mentions the research gap in the existing studies in the domain of smart homes, smart buildings, and smart environments. A few of the open challenges in smart environments are unsecured authentication of users and the device registration system that requires keeping track of signals between multiple transmitters. A few other systems are facing issues such as more time delay in turn on/off of an appliance, high cost, only showing temperature, lack of secured communication among connected devices, and lack of security and user authentication. Similarly, a few other issues are limited control of light and device on/off and limited GUI support for Symbian OS cell phones; limited range (50–100 m).

3. Used Approach

In this paper, a smart home automation system is presented that allows the user to check and change the status of electronic home appliances and the working state of sensors. Appliances include common lighting, heating, ventilation, and air-conditioning electronic devices. This product not only saves time but also provides energy-efficient solutions which let you know which device of your home consumes more power. This cost-efficient and the energy-saving system can also be installed in hotels and industrial/domestic places. Easy GUI icon-based interface and notifications of application let you be in touch with your home even if you are on a trip far away from your home. The concept of automating ordinary available installed devices of the home costs much less than special IP-based (RJ-45) [27] devices giving uniqueness and providing a solution that is affordable to people of middle-class families. The working of the proposed home automation system is shown in Figure 1.

In the admin side of the app, the user will able to draw the complete structure of his/her home using easy drag and drop interface. Firstly, the user adds floors of the house and, then, selects any added floor to add rooms inside it. After

Table 1: Previous studies and the research gap.

Work year	Sensor/technology used	Purpose	Limitation
[19] 2020	ArduinoTmega2560, IoT	Assist handicapped people to control home appliances	Unsecured authentication of users
[20] 2019	Elegoo Mega 2560, Web Server	Intimate owner about the door opening event in the home or office	A registration system requires keeping track of signals between multiple transmitters.
[21] 2018	Arduino UNO, ESP-8266, local Wi-Fi	Remotely control electrical and electronic devices in the home	More time delay in turn on/off of an appliance
[22] 2017	Wireless sensor network	Provide novel secured and energy efficient home automation	Expensive; only shows temperature
[23] 2016	Electrical switch, Internet	Regular monitoring of home appliances	Lack of secured communication among connected devices
[24] 2015	GPRS, speech recognition	Remote operating of electrical devices in home	Lack of security and user authentication
[25] 2013	Android ADK	Support handicapped and old-age people to control home appliances	Limited control of light and device on/off
[26] 2011	Bluetooth	Cell phone-based control of home appliances	Limited GUI support for Symbian OS cell phones; limited range (50–100 m)
[14] 2009	ZigBee microcontroller	Connectivity of devices within the home	Short range; low data speed
[15] 2002	Bluetooth	Control home appliances	Intrusiveness of the installation
[16] 2004	Personal computer (PC) based web server (Java-based technology)	Remotely connectivity of devices	Intrusive and expensive wired installation
[17] 1998	Phone-based controller	Remotely control home appliances	No graphical user interface
[18] 1993	Novel control network (hand gesture technology)	Remote control of objects using free-hand gestures	Inaccuracy of hand gestures; user fatigue if repetitive hand gestures are required

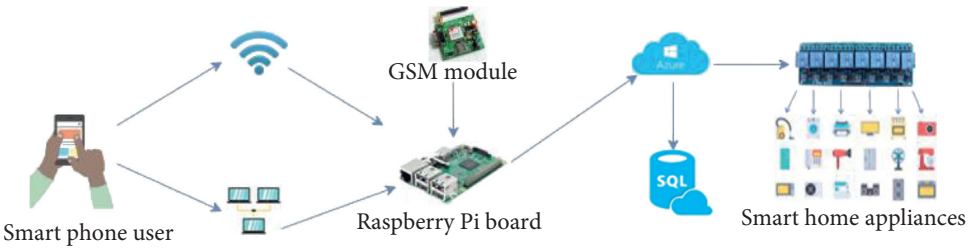


Figure 1: Complete working structure of home automation.

selecting the room, the user will be able to add appliances and can set the position of each device according to the real structure of a home. There is also an option available to add custom floors, rooms, and devices inside the home structure. After completing the setup of home from the admin application user sync database in the form of JSON [28] to our server, when a user logs in using credentials, he will able to see the complete structure of home the same as he made in the admin side as complete data are fetched from the server and get updated using VOLLEY [29, 30] service after 30 sec. At the main screen of the application, the user can see the number of total devices installed at home, and the floor details are shown. Three more tabs lie on the lower bottom of the application. The 2nd tab contains the complete information of sensor status. The third tab contains the history of each device status along with the name of the user who changes status and on which time. The fourth tab is logout from the application. Shared preferences are managed for credentials, and the information is saved.

Moreover, the SVM classifier is applied to classify the status of home appliances used in the smart home automation system. It classifies the status of home appliances into two categories, i.e., "ON" and "OFF"according to the use of these appliances. We have also compared the used prediction model with other models and choose the best one. Furthermore, blockchain technology is used to secure the whole communication and authentication between IoT devices and users that request to make "ON" or "OFF" any home appliances. Blockchain technology also ensures the security of data transfer between IoT devices, servers, applications, and users of the smart home automation system.

3.1. SVM-Based Classification. The proposed smart home automation system has ability to make an intelligent decision regarding the status of home appliances. To accomplish it, we used a linear SVM classifier. SVM was initially proposed by Vapnik [31]. SVM is a supervised machine learning

algorithm that can be used for both classification and regression challenges [32]. In the proposed system, we have employed it for classification purposes. Moreover, many techniques can be used for the classification of data such as decision tree, neural network, and K-nearest neighbor. We have preferred to use SVM for intelligent decision making in the proposed solution due to the following reasons.

(i) SVM is a suitable choice where binary output classes are required as in our case. Our proposed system needs to classify the status of devices into one of two classes, i.e., "ON" and "OFF."

(ii) It is required to find the best hyperplane using a kernel function to classify the data point of one class from another class in classification problems. SVM finds the optimal hyperplane by separating the data points of both classes.

(iii) When classification was performed using the SVM classifier, the resultant hyperplane has maximum margin between data points of target classes. However, the hyperplane generated by other classifiers does not achieve maximum margin [33].

To mathematically describe the SVM hyperplane and max hyperplane, we used equations (1) and (2) given by [34].

$$\left(\underset{x1}{\leftarrow}, y1\right)\cdots\left(\underset{xn}{\leftarrow}, yn\right), \tag{1}$$

where y has the value of either -1 or $+1$ and to represent whether the data points $\underset{x1}{\leftarrow}$ belong to which class, the value of y is used. Equation (2) is used to obtain the maximum margin hyperplane used by SVM.

$$\text{Max Hyperplane} = \frac{\sum_{i=1}^{n} xi \text{ where } yi = 1}{\sum_{j=1}^{n} xj \text{ where } yi = -1}. \tag{2}$$

In our problem, the input dataset contains the readings of sensors that include temperature, smoke, and light sensor. The output of the problem is status classes, i.e., "ON" or "OFF." The value of $y = 0$ represents that the device is "OFF," and $y = 1$ represents that the device is "ON."

SVM is a classic two-type classifier for segregating the two classes by determining an optimal hyperplane in a two-dimensional plane, as shown in Figure 2. The green color dots represent class 1 and the red color dots represent class 2, the optimal separating line denotes by H, the H1 line is parallel to H and linking samples in class 1, and the H2 line is also parallel to H and linking samples in class 2, as shown in Figure 2. w denotes that the distance between the H1 and H2 is called the margin, i.e., equal to $2/\|w\|$.

It is needed to separate the two classes in such a way that misclassification error becomes zero and maximize the margin. This can be achieved by using optimal separating line H that becomes the optimal separating hyperplane through spreading two-dimensional space into higher dimensional space. The equation of a separating hyperplane to separate the two classes using the optimal separating hyperplane is given as follows:

$$wx + b = 0. \tag{3}$$

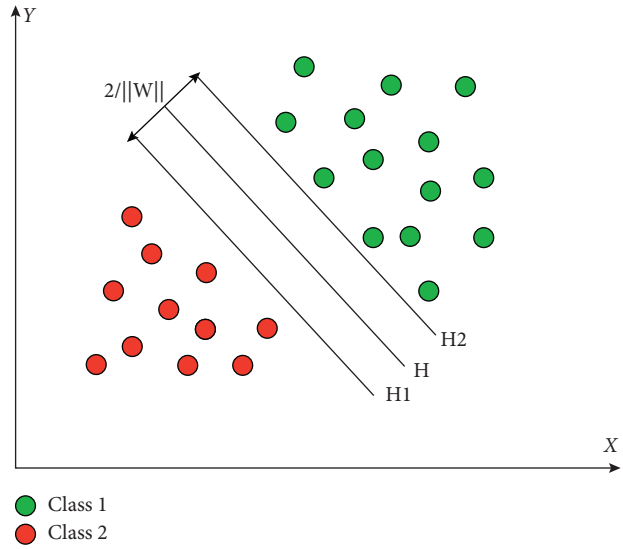


Figure 2: Schematic diagram of the hyperplane.

In equation (3), w is the weight vector and the offset of the separating hyperplane is b. A linearly separable hyperplane between class 1 and class 2 is shown in Figure 3.

3.2. Blockchain for Security of Smart Home. A key objective of the smart home automation system is to provide a trusted, secure identification and authentication of IoT devices. To ensure these objectives, we have used blockchain technology. Blockchain technology was introduced by Nakamoto in 2008 [35]. The main features of blockchain technology are security, anonymity, and decentralization [36]. These features can be fruitful to IoT by providing more security and reduced dependency on the central server. Moreover, blockchain technology also ensures tempered resistance data structure by utilizing data encryption and timestamp. In the proposed approach, the implementation of the blockchain module is accomplished in Java by defining the block's contents in a hash that is a unique identifier. Using each block, a block hash is computed, and a hash SHA-256 is, then, computed from it. On achieving a threshold, requests for connectivity are granted through managing blockchain, and then, a block is created. To validate the complete blockchain, a block of the chain is looped-over that verify a current block's hash to its corresponding previous block's hash.

Whenever a connectivity request is generated by a user, it will be authenticated using a sequence of steps described in the flowchart given in Figure 3.

We have also designed an algorithm to explain the working of blockchain technology in the proposed approach given as follows. The whole process of blockchain implementation is also depicted in a flowchart shown in Figure 3 and Algorithm 1.

First of all, a block is created using a block class implemented in Java calculating a hash using a previous hash, timestamp, and data string. After creating a block, a hash will be generated using the SHA256 algorithm

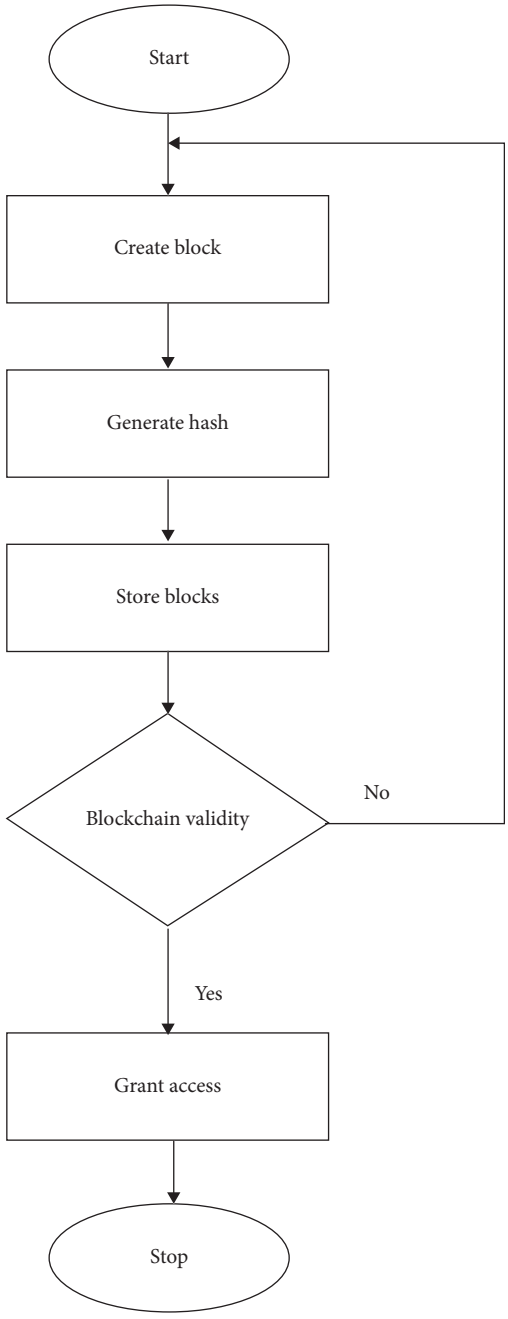


Figure 3: Flowchart of blockchain implementation in the proposed approach.

implemented in Java. When a hash is created, the next step is to store the generated blocks. At the final step, validation of the blockchain will be performed to check whether the hash is equal to the calculated hash. If it is true, then the block is valid and access is granted; otherwise, the whole process is repeated.

4. Implementation Details

To understand the working of the project, Figure 1 explains the complete working structure and integration of different devices with each other. The arrows show the flow of projects starting from user smartphones to changing the state of electronic devices. There are two network modes in which the user can interact with our Raspberry [37–41]. If the user is sitting inside the home (home local network), then the user will be able to use all of the IoT services at the local network without connecting to the internet cloud. This will also result in faster communication of devices with the user as everything is happening locally.

The second network mode is used if the user is residing outside the home anywhere in the world. Then, the user first connects to the internet. The processed request is sent to the Microsoft Azure Cloud [42, 43]. Based on credentials provided by the user matched with the Azure cloud database, the user request is sent to the respective Raspberry for processing. The account of each user is maintained individually at the Microsoft Azure Cloud Database. Services of each user are handled based on credentials from which the request is being generated. APIs are called from the cloud if the user is outside the home network. Then, the same APIs are also residing inside the Raspberry Pi server if the user is in the home network.

Data sharing between application and server database is performed in the form of JSON. APIs are secured using multiple hashing techniques. The change in the state of any device is performed using Raspberry Pi GPIO [44] pins. Raspberry Pi receives the request from the server. According to the request from the user, Raspberry Pi responds to devices. The database of each request generated by the individual user is maintained at cloud servers. The user can check the complete history of processed requests on his smartphone by setting the duration. Sensors installed inside the home update its state continuously after 30 sec and respond to the change to the Raspberry Pi server. In response, the Raspberry Pi server syncs all data to the cloud database, and the values of the mobile application are updated.

4.1. Hardware Components. Many electronic components, as well as sensors, are used in this project. Our working structure diagram Figure 1 shows a few of the components used in this project. A complete description of components along with description is shown in Table 2.

4.1.1. Raspberry Pi 2B. The Raspberry Pi is a series of small single-board computers to promote the teaching of basic computer science in schools and developing countries. It contains a 900 MHz quad-core ARM Cortex-A7 CPU processor along with 1 GB of RAM. It supports 100 MBPS Ethernet. It also contains 4 USB ports along with 40 GPIO pins, full HDMI support with camera, and card interface and supports a micro SD card. It contains a 3.5 mm audio jack and composite video support too.

4.1.2. 5 V Relay Circuit. A relay is an electrically operated device. It has a control system and (also called input circuit or input contactor) and a controlled system (also called output circuit or output contactor). It is frequently used in

(1) At the first step, a block is created with the help of the block class
(2) A hash is created using the SHA–256 algorithm on the successful creation of an already created block otherwise repeat step 1
(3) After the generation of hash, the next step is to store the blocks
(4) At the last step, the validity of the blockchain will be accomplished. If the block is valid, then request for connectivity is granted
(5) If the block is not valid, steps 1 to 4 are repeated

Algorithm 1: Working algorithm of blockchain technology.

Table 2: Electronic components with specification details used in the project.

Components	Specifications
Raspberry Pi 2B	40 GPIO pins, 1 GB RAM A 900 MHz quad-core ARM Cortex-A7 CPU, operational voltage 7–12 V
Relay circuit pack	The 5 V operational 8-relays circuit pack
L293D motor control shield [45]	Supply-voltage range: 4.5–36 V; output current: 600 mA/channel
Smartphone mobile	Android supported
DS18B20 temperature sensor [46]	Temperature range: −55 to 125°C (−67°F to +257°F)
LM393 LDR sensor [47]	Digital switching outputs (0 and 1), external 3.3 V–5 V vcc
MQ2 smoke sensor [48]	Combustible gas, smoke

the automatic control circuit. To put it simply, it is an automatic switch to controlling a high-current circuit with a low-current signal.

4.1.3. L293D Motor Shield. L293D [49] is a monolithic integrated, high-voltage, high-current, 4-channel driver. This means that using this chip, you can use DC motors and power supplies of up to 16 Volts, that is, some pretty big motors, and the chip can supply a maximum current of 600 mA per channel; the L293D chip is also what is known as a type of H-Bridge. The H-Bridge is typically an electrical circuit that enables a voltage to be applied across a load in either direction to an output, e.g., motor.

4.1.4. DS18B20 Temperature Sensor. The DS18B20 [50] temperature sensor is a 1-wire digital temperature sensor. This means that you can read the temperature with a very simple circuit setup. It communicates on a common bus, which means that you can connect several devices and read their values using just one GPIO pin of the Raspberry Pi.

4.1.5. LM393 LDR Sensor. The LM393 [51] is a simple photoresistor light sensor that has both analog and digital outputs. The digital output has a trim potentiometer that can be used to set a trigger light level. In the following example, we will use the analog output to measure the light level. If you are interested in reading the digital output, please refer to the pulse sensor example sketch.

4.1.6. MQ2 Smoke Sensor. MQ2 [52] is one of the commonly used gas sensors in the MQ2 sensor series. It is a Metal Oxide Semiconductor- (MOS-) type gas sensor also known as chemiresistor as the detection is based upon the change of resistance of the sensing material when the Gas comes in contact with the material. Using a simple voltage divider network, concentrations of the gas can be detected. An MQ2 Gas sensor works on 5 V DC and draws around 800 mW. It can detect LPG, smoke, alcohol, propane, hydrogen, methane, and carbon monoxide concentrations anywhere from 200 to 10000 ppm.

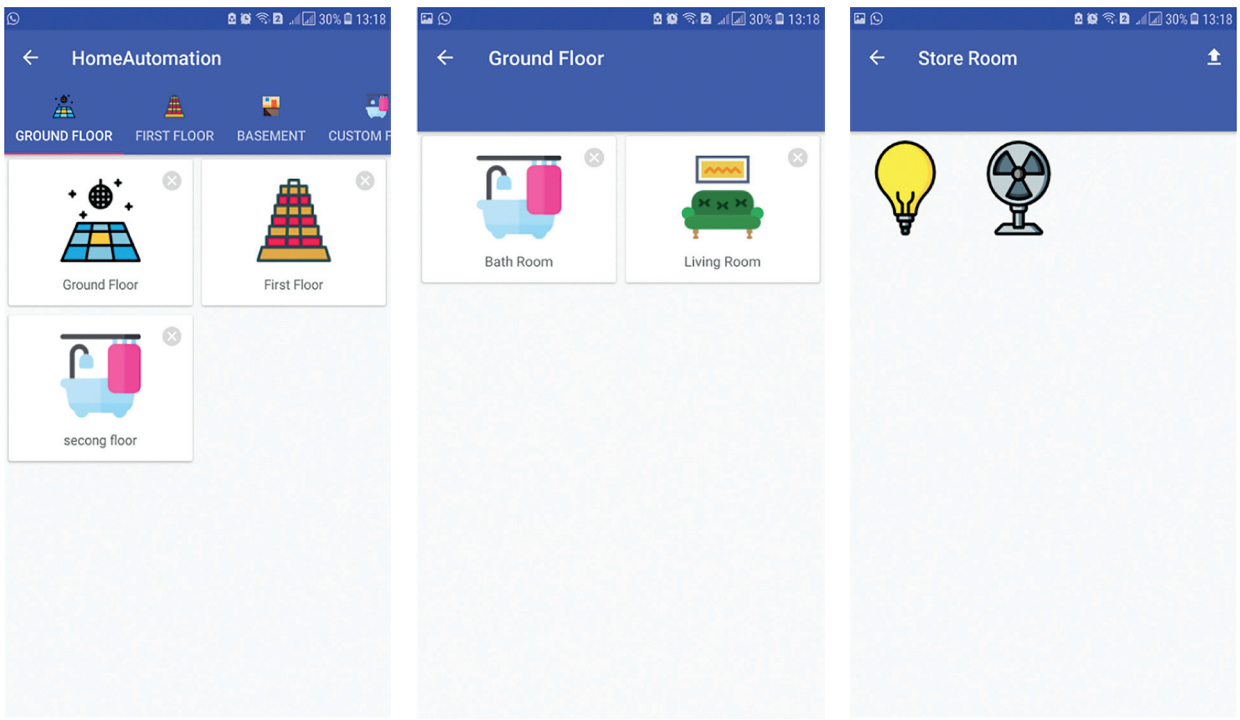
4.1.7. GSM Module. This module can be mostly used for developing IoT (Internet of Things) and embedded applications. It is a dual-mode module that works on frequencies 900 MHz and 1800 MHz. It has a feature of multislot class, i.e, class 10/class 8. The pins RXD and TXD are used to receive and transmit data. It works on a voltage of 3.4 V to 4.5 V, and any higher voltage may damage the module. The GSM module also requires low power consumption.

4.2. Software Components. For the development of the mobile application, there are many platforms available such as Symbian, Android, IOS, and Windows Mobile.

In our research paper, the Android platform is used for the development of the whole project. The main reason for the use of the Android platform is that it is the most widely used platform all over the world. The Android operating system is supported by almost every brand device. Java language with an Android software development kit (SDK) has been used for the development and implementation of home automation application.

4.2.1. Android Studio. Android Studio [53, 54] is used for the development of Android apk as it supports all development tools such as debuggers, libraries, and a handset emulator. The Volley library is used for all sensor services. Material design library is used to make the application more interactive.

4.2.2. Server-Side Scripting. For the development of the server-side on cloud and inside Raspberry Pi, LAMP (Linux, Apache, MySql, PHP) is used to provide all backend functionality.



(a) (b) (c)

Figure 4: Working flow of each appliance.

4.3. Mobile Applications for Smart Home. Mobile Application consists of two modes of operation. In the admin mode, the user able will be able to draw his/her complete home prototype using an easy drag and drop interface. Each device is assigned with a Raspberry Pi pin number which controls the actual working of electronic devices on the backend.

If we look at the user side of the mobile application from which the user will not only be able to see the complete prototype of the home which is designed but also can control the working of each appliance on basis of pin configuration that he did on the admin side, as shown in Figure 4. The main screen contains the number of devices installed in the house along with floor names and the number of rooms inside it. Tab layout is used for switching between app services.

A nice and interactive GUI-based interface with attractive icons lets the user know about its working easily, as shown in Figure 5. Icons changing are used to let the user know about the current state of an electronic device with a touch active button to change state. There is also an intensity bar to let the user maintain its fan speed and light brightness according to its choice. The status of devices is shown with the active state of working and active ago with the passive state.

In the second tab of the main screen, the state of each sensor is shown along with current values. The values of sensors get updated using backend services which refresh themselves after every 30 seconds, as shown in Figure 6. Currently, we are dealing with the live status of two sensors such as temperature and light, as shown in Figure 7. The light sensor value 0.0 shows that currently, daylight is active and the sensor is in off state.

The third tab on the main screen gives us complete details about the history of the device. The complete log with the details of the user who changes the status of appliances is maintained with time-stamped details. The user can check the on and off duration of appliances from history as well. Another interesting feature of this application is that the user will receive a notification on his smartphone if any device is working for more than two hours. It is similar to an alarm or reminder to the user to check and balance the power consumption of each appliance whether a person is sitting in a particular room or there is electricity wastage, as shown in Figure 8.

There is also a billing calculation system based on device electricity consumption and the time duration in which the device remains in the working state.

5. Results and Discussion

The functional behavior of this project can be further understood using Figure 9. The complete workflow of the project comprises two ways as we discussed earlier. In the first scenario, the functional behavior of the project assumes that the user is outside the home and accessing each device remotely. This scenario assumes the involvement of using the Microsoft Azure cloud database. Each request from the user is firstly received on the cloud, and according to the credentials provided by the user, APIs are called.

In the second scenario, we assume that the user is inside the home and connected to the same network to which Raspberry Pi is connected. All the requests generated by the user do not go to the internet or cloud rather than operateing

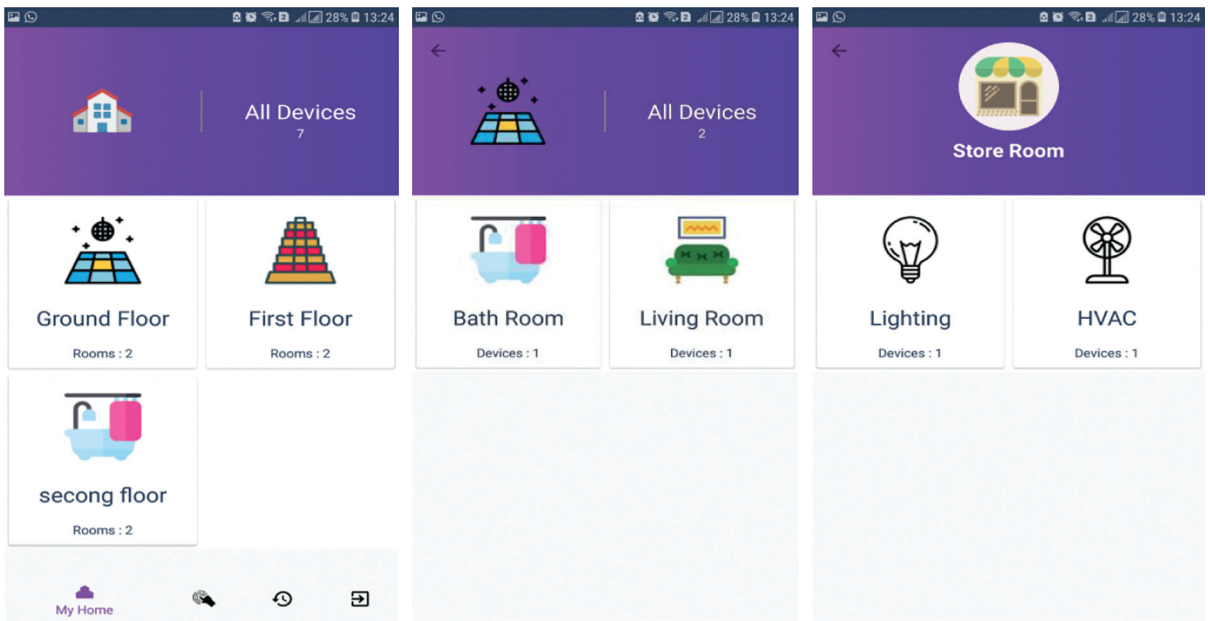


Figure 5: GUI of the user main screen.

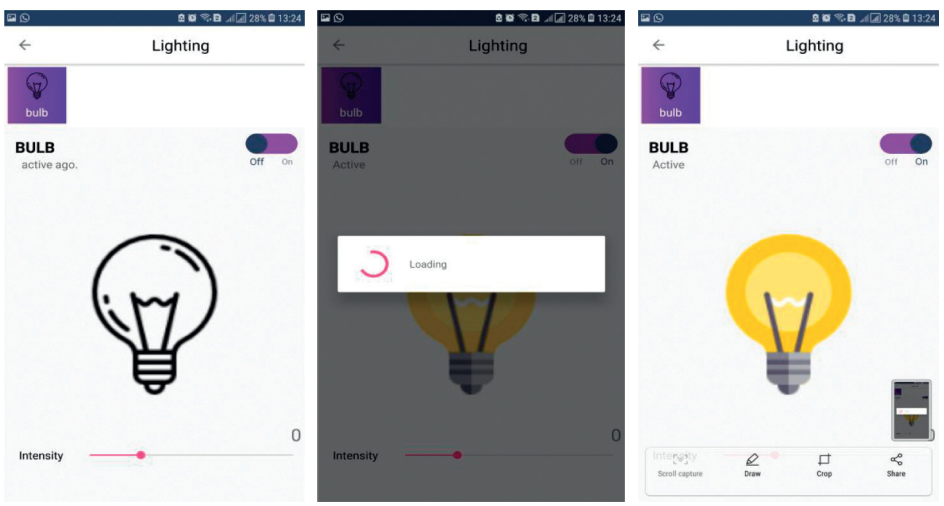


Figure 6: Changing the state of the appliance.

on Raspberry Pi server internally, as shown in Figure 10. This will lead to the processing of requests more quickly than the involvement of the cloud.

We have also stored data taken from the history tab of the developed App and stored it in a Microsoft Excel sheet (see Figure 11) and further used it to show the relationship between data that are discussed in this section. The excel sheet contains the values of temperature, smoke, and light. There is a status column where 0 represents the "OFF" status of smoke and light while 1 represents the "ONN" status of smoke and light. We have also drawn a joint plot between temperature and smoke, as shown in Figure 12 where each observation of temperature is shown along the x-axis and smoke values are shown along the y-axis. The histogram along the x-axis represents the different readings of temperature, and the histogram along the y-axis represents the different readings of smoke.

In Figure 11, the joint plot is drawn that shows the relationship between light and temperature. The values of light are shown along the x-axis, and temperature is shown along the y-axis. The histogram of light and temperature is also shown in Figure 13 along the x-axis and y-axis, respectively.

A scatter plot is also drawn to show the relationship between smoke and light (see Figure 14). The smoke values are shown along the x-axis, and light values are shown along the y-axis. The histogram of smoke is also represented along the x-axis, and the histogram of light is represented along the y-axis as shown in Figure 14.

We have shown the status of light from taken readings using the two-dimensional kernel density plot in Figure 15. It shows the status of light when it is "ON" and "OFF." The values of light fall between zero and six hundred, and the

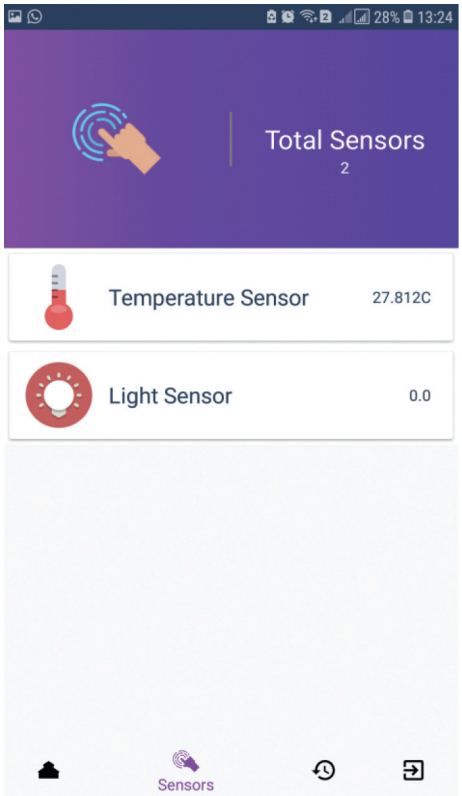


Figure 7: Sensor service data.

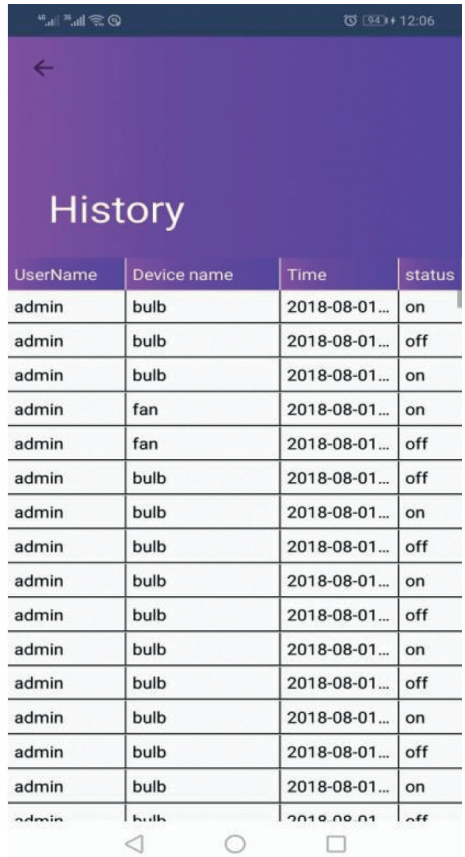


Figure 8: A history log of the device.

status of values lie between 0 and 1 that shows "OFF" and "ONN," respectively.

The kernel density plot for smoke is shown in Figure 15 that shows the status of smoke along the x-axis and its values along the y-axis. The presence of smoke is represented as zero and absence as one in Figure 16.

One of the most common types of graphics that are used to show the numeric values of variables is the box plot. It divides the dataset into three quartiles. The line that divides the box into two parts represents the median of the data, and the end of the corner shows the upper and lower quartiles. The extreme lines represent the highest and lowest values excluding outliers. It also hides the number of the values that lie behind the variable. We have also drawn a box plot of light, temperature, and smoke, as shown in Figures 17–19.

To make intelligent decisions about the status of home appliances, we have used a linear SVM classifier. First of all, we divided the recorded data into training and test data sets using the built-in function tain_test_split (·) of Scikit learn in Python. Figure 20 shows the confusion matrix of trained SVM classifiers that shows% of truly predicted and wrongly predicted classes.

In the confusion matrix, the instances of the true class are represented horizontally while instances of the predicted class are represented vertically. There were a total 49 predicted "OFF" classes out of which only one class was wrongly predicted as "ON" that was actually "OFF" while the total "ON" classes were 17 out of which only 2 classes were wrongly predicted as "OFF" that was actually "ON." We have also plotted the ROC curve of the proposed linear SVM classifier, as shown in Figure 21. In the ROC curve, the true positive rate of a linear SVM classifier was shown on the y-axis and the false positive rate was represented along the x-axis. The ROC of the linear SVM classifier showed an AUC = 0.957, as shown in Figure 21.

We have also compared linear SVM classifiers with other available classifiers such as KNN, Decision Tree, and Random Forest. Among all these classifiers, SVM outperforms, as shown in Figure 22. Therefore, the presented approach uses a linear SVM classifier for decision making about the status of home appliances.

In Figure 22, the accuracy of the SVM classifier is 95.73% that is greater than that of all other machine learning classifiers, i.e., K-Nearest Neighbour (KNN), Naïve Base (NB), Random Forest (RF), and Decision Tree (DT) having accuracy 94.32%, 93.80%, 93.24%, and 94.73%, respectively.

The implementation is performed in a way that if the Raspberry Pi restarts or any mishap happens in electricity, then all the devices are set again to the state before the restart. This involves the use of a database server that maintains the state of each device. The last state record of each device is being fetched and set again to each device accordingly. The application will get its details from the server too.

Sensors are installed to get the updates at regular intervals. There is a condition set out for user comforts that if

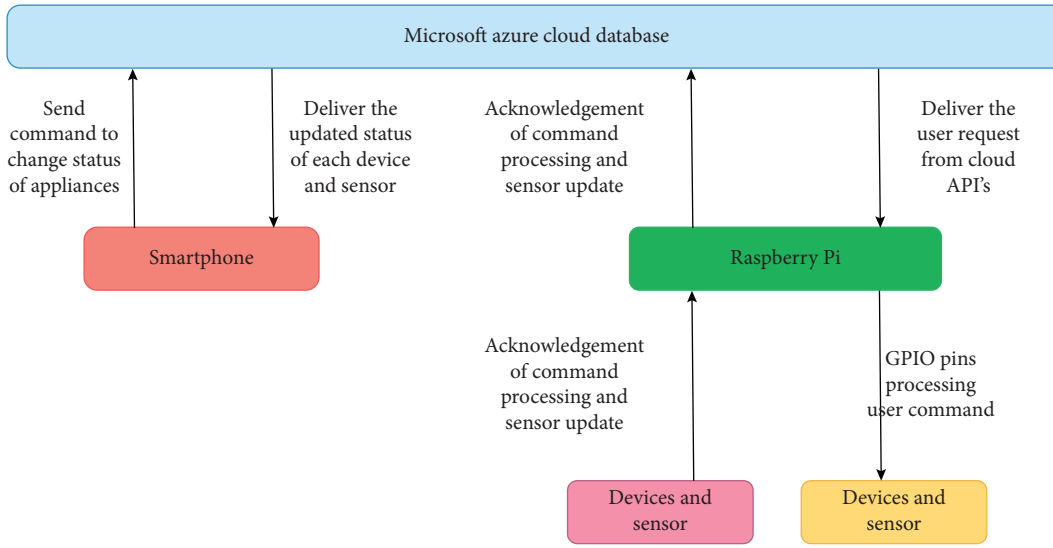


Figure 9: Microsoft Azure cloud database.

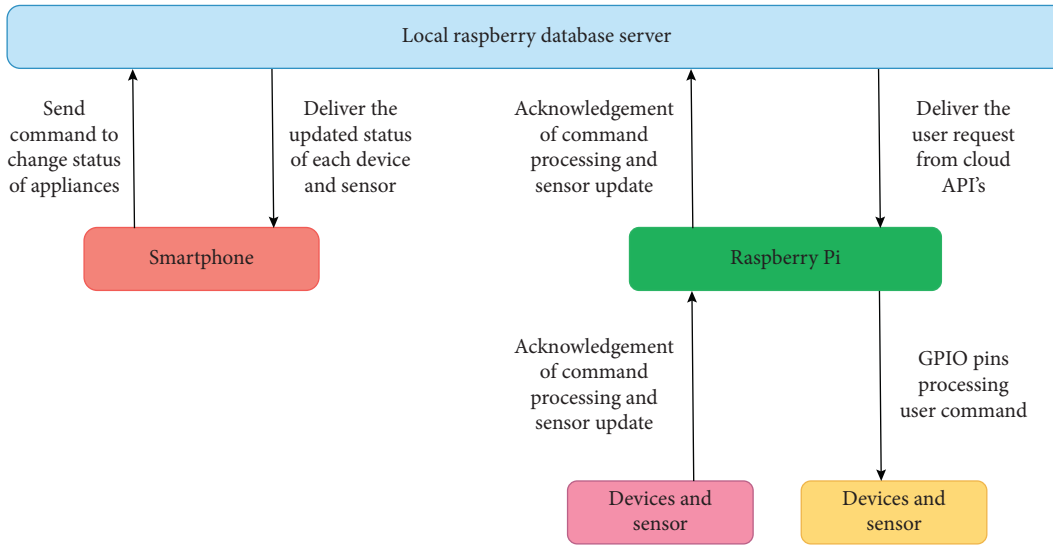


Figure 10: Local database server.

Temperature	Smoke	Light	Status
28.54	0	0	0
28.68	0	0	0
28.78	0.5	51	1
28.9	0	0	0
29.35	0	0	0
29.65	0.8	154	1
29.78	0.6	182	1
29.88	0.7	203	1
29.05	0.51	204	1
29.67	0.56	256	1
29.99	0.58	258	1
30.05	1.2	287	1
30.1	0.32	290	1
31.21	0	0	0
31.56	0.26	298	1
31.88	0	0	0
32.56	0	0	0

Figure 11: Recorded data.

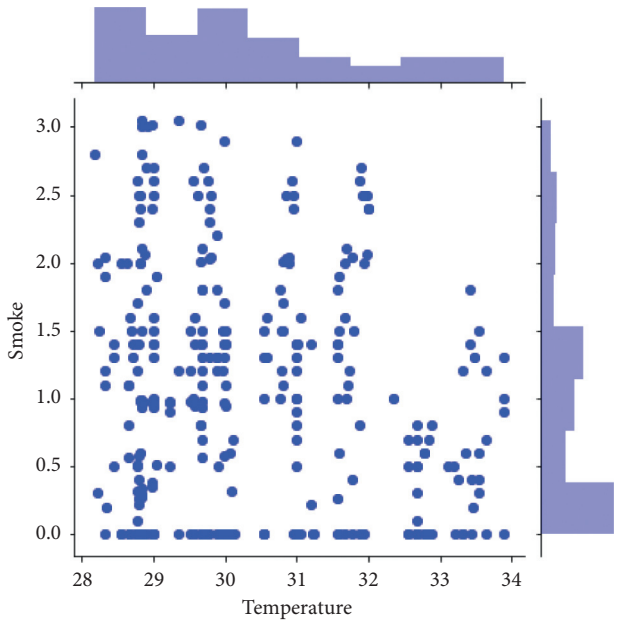


Figure 12: Scatter plot between temperature and smoke.

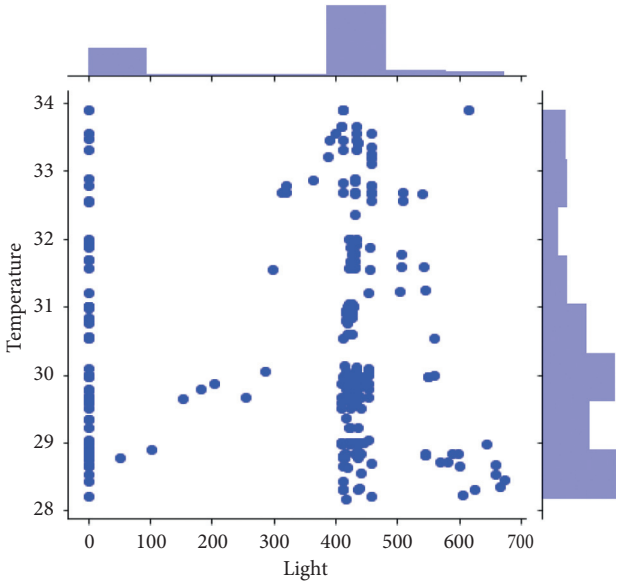


Figure 13: Scatter plot between light and temperature.

the temperature of the home rises up to the threshold set by the user, then all ventilation fans will start working automatically and vice-versa. Light sensors controlled the state of lights installed outside the home automatically in daylight and on the night. A comparison of the old and the proposed system is shown in Table 3.

All the features and functionality comparisons described in the table above will make this project unique from the rest. Designing the prototype of your own house and setting each device according to the room design will make it easier for a user to operate the electronic device with more ease.

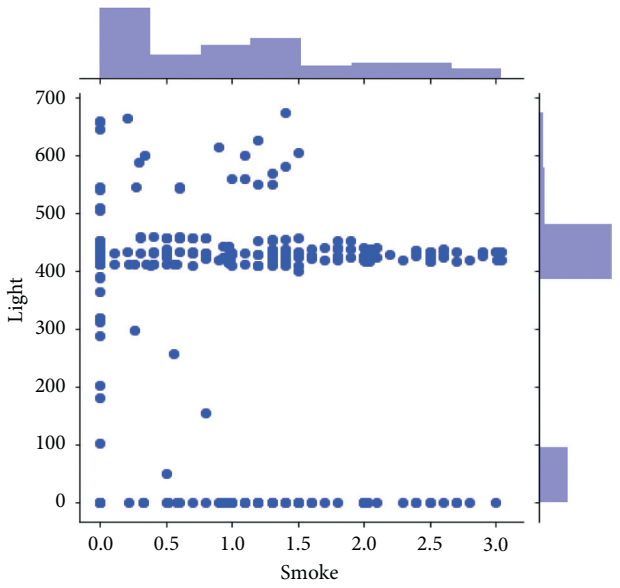


Figure 14: Scatter plot between smoke and light.

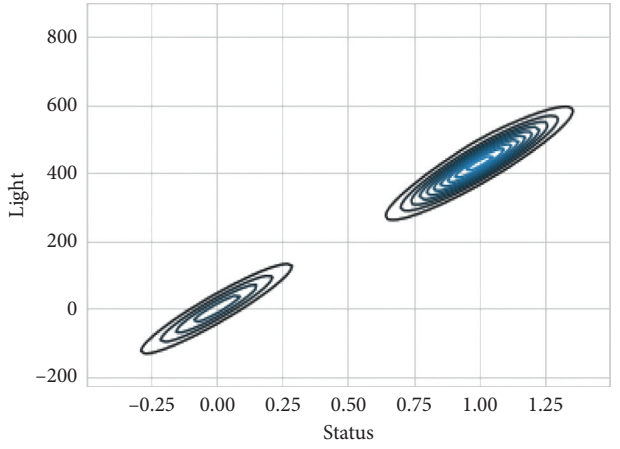


Figure 15: Kernel density plot of light.

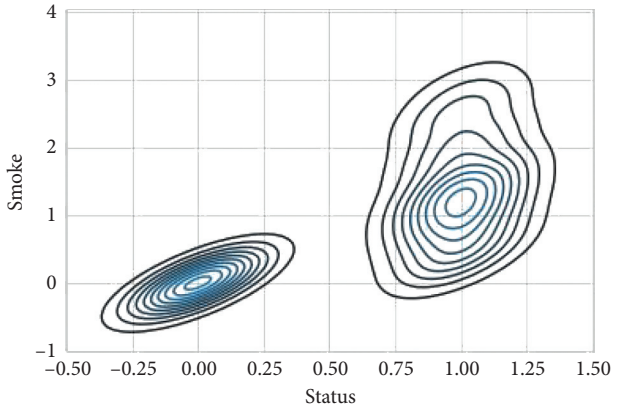


Figure 16: Kernel density plot of light.

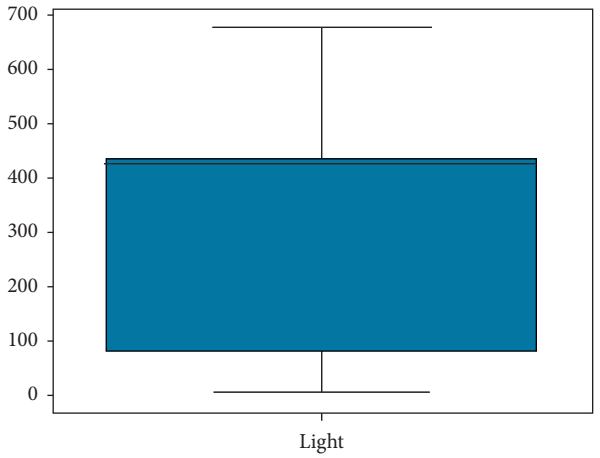


Figure 17: Box plot of light.

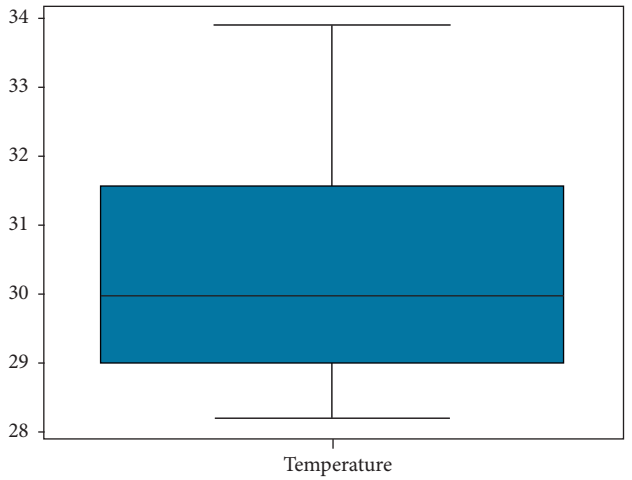


Figure 18: Box plot of temperature.

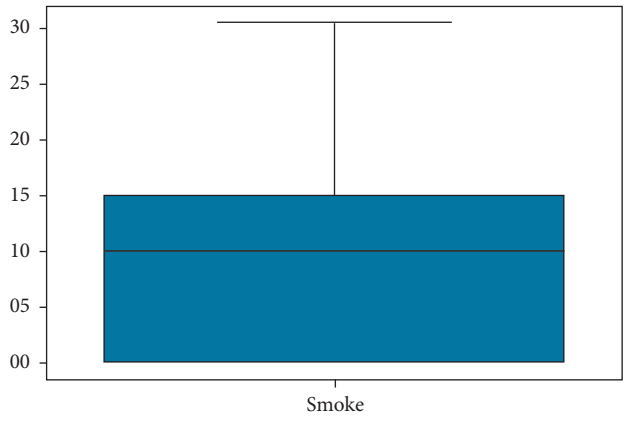


Figure 19: Box plot of smoke.

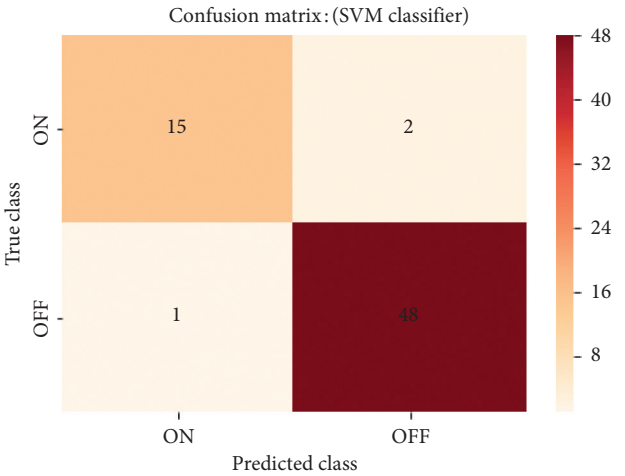


Figure 20: Confusion matrix.

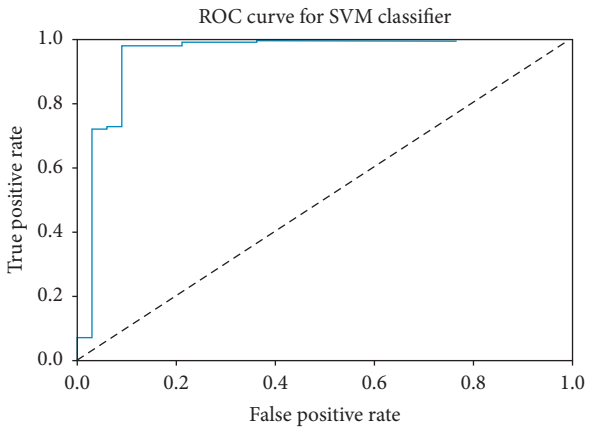


Figure 21: ROC curve of the SVM classifier.

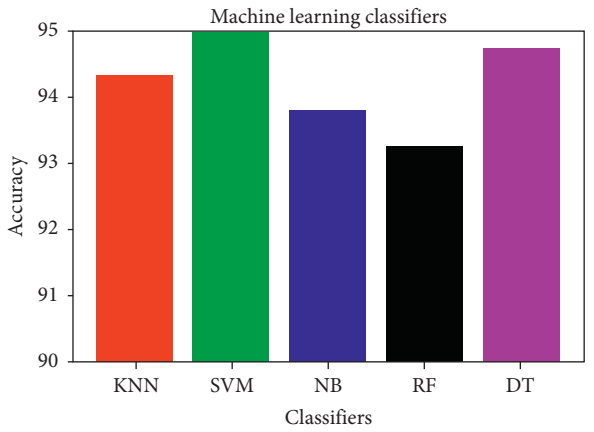


Figure 22: Comparison between machine learning classifiers.

Table 3: Comparison between the old and proposed system.

Functionality	Old system	Proposed system
Raspberry support	No	Yes
Data logging	No	Yes
Home prototype design	No	Yes
Sensor update	No	Yes
Intensity control	No	Yes
Billing system	No	Yes
Cloud database support	No	Yes
Machine learning classifier	No	Yes
Blockchain technology	No	Yes

6. Conclusions

In this research paper, the complete design scheme and working methodology of a home automation project are explained. The reason behind the paper is to get enhancement ideas to make it better and user friendly. This proposed system comprises two modes, the admin side in which the user able to design the complete prototype of the home and the user side in which the user able to control each device of the home with an easy GUI-based interface. The proposed system also has decision-making ability about the status of each device of the home. A machine learning algorithm SVM with a linear kernel is applied for decision making about the status of home appliances either ON or OFF. The presented approach also ensures secure identification and authentication of IoT devices using blockchain technology, while the existing systems which have not copped with intelligent decision making and analytical abilities are the need of time. The existing system also does not provide secure communication between IoT devices. Moreover, this project is user acceptance tested verified and operational in many homes. It provides simplicity, flexibility, reliability, and a low-cost system that is affordable to middle-class families too.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Disclosure

Rizwan Majeed, Imran Ashraf, and Yousaf Bin Zikria are primary co-authors.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] E. S. A. Ahmed, "Internet of things applications, challenges and related future technologies," *Internet of Things (IoT) Applications*, vol. 67, no. 2, pp. 126–148, 2017, https://www.researchgate.net/profile/Zeinab_Kamal2.

[2] R. Piyare, "Internet of things: ubiquitous home control and monitoring system using android based smart phone," *International Journal of Internet of Things*, vol. 2, no. 1, pp. 5–11, 2013.

[3] K. Mandula, R. Parupalli, C. A. Murty, E. Magesh, and R. Lunagariya, "Mobile based home automation using internet of things (IoT)," in *Proceedings of the 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 340–343, IEEE, Kumaracoil, India, December 2015.

[4] B. Kang, S. Park, T. Lee, and S. Park, "IoT-based monitoring system using tri-level context making model for smart home services," in *Proceedings of the 2015 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 198-199, IEEE, Las Vegas, NV, USA, January 2015.

[5] M. Darianian and M. P. Michael, "Smart home mobile RFID-based internet-of-things systems and services," in *Proceedings of the 2008 International Conference on Advanced Computer Theory and Engineering*, pp. 116–120, IEEE, Phuket, Thailand, December 2008.

[6] A. D. Plessis and B. Theron, "Virtual world—physical world: what is the real world?" *International Journal of Management Science and Business Administration*, vol. 2, no. 6, pp. 43–57, 2016.

[7] G. Lobaccaro, S. Carlucci, and E. Löfström, "A review of systems and technologies for smart homes and smart grids," *Energies*, vol. 9, no. 5, p. 348, 2016.

[8] O. Bingol, K. Tasdelen, Z. Keskin, and Y. E. Kocaturk, "Web-based smart home automation: PLC-controlled implementation," *Acta PolytechnicaHungarica*, vol. 11, no. 3, pp. 51–63, 2014.

[9] A. ElShafee and K. A. Hamed, "Design and implementation of a wifi based home automation system," *World Academy of Science, Engineering and Technology*, vol. 68, pp. 2177–2180, 2012.

[10] T. A. Abdulrahman, O. H. Isiwekpeni, N. T. Surajudeen-Bakinde, and A. O. Otuoze, "Design, specification and implementation of a distributed home automation system," *Procedia Computer Science*, vol. 94, pp. 473–478, 2016.

[11] M. Sarwar and T. Soomro, "Impact of smartphone's on society," *European Journal of Scientific Research*, vol. 98, 2013.

[12] M. Fahim, I. Fatima, S. Lee, and Y. K. Lee, "Daily life activity tracking application for smart homes using android smartphone," in *Proceedings of the 2012 14th International Conference on Advanced Communication Technology (ICACT)*, IEEE, PyeongChang, South Korea, pp. 241–245, February 2012.

[13] S. Otoum, B. Kantarci, and H. T. Mouftah, "On the feasibility of deep learning in sensor network intrusion detection," *IEEE Networking Letters*, vol. 1, no. 2, pp. 68–71, 2019.

[14] K. Gill, S.-H. Yang, F. Yao, and X. Lu, "A zigbee-based home automation system," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 422–430, 2009.

[15] N. Sriskanthan, F. Tan, and A. Karande, "Bluetooth based home automation system," *Microprocessors and Microsystems*, vol. 26, no. 6, pp. 281–289, 2002.

[16] A. R. Al-Ali and M. Al-Rousan, "Java-based home automation system," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 2, pp. 498–504, 2004.

[17] H. Ardam and I. Coskun, "A remote controller for home and office appliances by telephone," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 4, pp. 1291–1297, 1998.

[18] T. Baudel and M. Beaudouin-Lafon, "Charade," *Communications of the ACM*, vol. 36, no. 7, pp. 28–35, 1993.

[19] A. S. Abdulraheem, A. A. Salih, A. I. Abdulla et al., "Home automation system based on IoT," 2020.

[20] M. A. Hoque and C. Davidson, "Design and implementation of an IoT-based smart home security system," *International Journal of Networked and Distributed Computing*, vol. 7, no. 2, pp. 85–92, 2019.

[21] L. M. Satapathy, S. K. Bastia, and N. Mohanty, "Arduino based home automation using internet of things (IoT)," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 17, pp. 769–778, 2018.

[22] S. Pirbhulal, H. Zhang, M. E. E Alahi et al., "A novel secure IoT-based smart home automation system using a wireless sensor network," *Sensors*, vol. 17, no. 1, p. 69, 2017.

[23] P. Kumar and U. C. Pati, "Iot based monitoring and control of appliances for smart home," in *Proceedings of the 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 1145–1150, IEEE, Bangalore, India, May 2016.

[24] S. B. Sangeetha, "Intelligent interface based speech recognition for home automation using android application," in *Proceedings of the 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 1–11, IEEE, Coimbatore, India, March 2015.

[25] D. Javale, M. Mohsin, S. Nandanwar, and M. Shingate, "Home automation and security system using android ADK," *International Journal of Electronics Communication and Computer Technology (IJECCT)*, vol. 3, no. 2, pp. 382–385, 2013.

[26] R. Piyare, M. Tazil, Bluetooth based home automation system using cell phone," in *Proceedings of the 2011 IEEE 15th International Symposium on Consumer Electronics (ISCE)*, pp. 192–195, IEEE, Singapore, June 2011.

[27] G. Chong, L. Zhihao, and Y. Yifeng, "The research and implement of smart home system based on internet of things," in *Proceedings of the 2011 International Conference on Electronics, Communications and Control (ICECC)*, pp. 2944–2947, IEEE, Ningbo, China, September 2011.

[28] Z. U. Haq, G. F. Khan, and T. Hussain, "A comprehensive analysis of XML and JSON web technologies," *New Developments in Circuits, Systems, Signal Processing, Communications and Computers*, pp. 102–109, 2013.

[29] Y. J. Song, S. B. Ou, and J. W. Lee, *An Analysis of Existing Android Image Loading Libraries: Picasso, Glide, Fresco, AUIL and Volley*, DEStech Transactions on Engineering and Technology Research, Lancaster, PA, USA, 2016.

[30] L. Hang and D.-H. Kim, "Design and implementation of intelligent fire notification service using IP camera in smart home," *International Journal of Control and Automation*, vol. 11, no. 1, pp. 131–142, 2018.

[31] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA, 1995.

[32] S.-T. NelloCristianini, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 1st edition, 2000.

[33] V. Jakkula, *Tutorial on Support Vector Machine (SVM)*, p. 37, Washington State University, Pullman, WA, USA, 2006.

[34] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, Burlington, MA, USA, 2011.

[35] S. Nakamoto, "Bitcoin: apeer-to-peer electronic cash system," Triumph Books, Chicago, IL, USA, 2008.

[36] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[37] M. Maksimović, V. Vujović, N. Davidović, V. Milošević, and B. Perišić, "Raspberry pi as Internet of things hardware: performances and constraints," in *Proceedings of IcETRAN 2014*, Vrnjačka Banja, Serbia, June 2014.

[38] V. Patchava, H. B. Kandala, and P. R. Babu, "A smart home automation technique with raspberry pi using iot," in *Proceedings of the 2015 International Conference on Smart Sensors and Systems (IC-SSS)*, pp. 1–4, IEEE, Bangalore, India, December 2015.

[39] M. Richardson and S. Wallace, *Getting Started with Raspberry PI*, O'Reilly Media, Inc., Sebastopol, CA, USA, 2012.

[40] E. Upton and G. Halfacree, *Raspberry Pi User Guide*, John Wiley & Sons, NewYork, NY, USA, 2014.

[41] F. Leccese, M. Cagnetti, and D. Trinca, "A smart city application: a fully controlled street lighting isle based on raspberry-pi card, a zigbee sensor network and wimax," *Sensors*, vol. 14, no. 12, pp. 24408–24424, 2014.

[42] L. Qian, Z. Luo, Y. Du, and L. Guo, "Cloud computing: an overview," in *Proceedings of IEEE International Conference on Cloud Computing*, pp. 626–631, Beijing, China, December 2009.

[43] Z. Wei, S. Qin, D. Jia, and Y. Yang, "Research and design of cloud architecture for smart home," in *Proceedings of the 2010 IEEE International Conference on Software Engineering and Service Sciences*, pp. 86–89, IEEE, Beijing, China, July 2010.

[44] J. D. Brock, R. F. Bruce, and M. E. Cameron, "Changing the world with a raspberry pi," *Journal of Computing Sciences in Colleges*, vol. 29, no. 2, pp. 151–153, 2013.

[45] Sunfounder, "L293D Motor driver shield," 2019, http://wiki.sunfounder.cc/index.php?title= L293D_Motor_Driver_Shield.

[46] Sparkfun, "Temperature sensor-waterproof (DS18B20)," 2019, https://www.sparkfun.com/products/11050.

[47] Lm393 optical photosensitive LDR light sensor module, https://www.daraz.pk/products/lm393-optical-photosensitive-ldr-light-sensor-module-i100135168-s1118660259.html, 2019.

[48] Arduino and MQ2 gas sensor, https://www.hackster.io/karimmufte/arduino-and-mq2-gas-sensor-57f98c, 2019.

[49] S. A. I. Quadri and P. Sathish, "IoT based home automation and surveillance system," in *Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 861–866, IEEE, Madurai, India, June 2017.

[50] Y. Upadhyay, A. Borole, and D. Dileepan, "MQTT based secured home automation system," in *Proceedings of the 2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, pp. 1–4, IEEE, Indore, India, March 2016.

[51] S. S. Harsha, S. C. Reddy, and S. P. Mary, "Enhanced home automation system using internet of things," in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 89–93, IEEE, Palladam, India, February 2017.

[52] S. Chattoraj, "Smart home automation based on different sensors and arduino as the master controller," *International Journal of Scientific and Research Publications*, vol. 5, no. 10, pp. 1–4, 2015.

[53] H. Esmaeel, "Apply android studio (SDK) tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, 2015.

[54] K. K. Rout, S. Mishra, and A. Routray, "Development of an internet of things(iot) based introductary laboratory for under graduate engineering students," in *Proceedings of the 16th Inernational Conference on Information Technology (ICIT)*, Bhubaneswar, India, October 2017.

Hindawi
Scientific Programming
Volume 2020, Article ID 7832875, 13 pages
https://doi.org/10.1155/2020/7832875

Research Article

A User-Defined Location-Sharing Scheme with Efficiency and Privacy in Mobile Social Networks

Tao Peng,[1] Jierong Liu,[1] Guojun Wang,[1] Qin Liu,[2] Jianer Chen,[1] and Jiawei Zhu[1]

[1]*School of Computer Science, Guangzhou University, Guangzhou, Guangdong 510006, China*
[2]*School of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, China*

Correspondence should be addressed to Guojun Wang; csgjwang@163.com

Received 13 February 2020; Accepted 31 July 2020; Published 26 October 2020

Academic Editor: Chenxi Huang

The popularity of the modern smart devices and mobile social networks (MSNs) brings mobile users better experiences and services by taking advantage of location-aware capabilities. Location sharing, as an important function of MSNs, has attracted attention with growing popularity. While the users get great benefits and conveniences from MSNs, they also have high concerns about the privacy of location. However, in the existing solution, the privacy of users can hardly be guaranteed without the assumption of full trust in the service provider (SP), and few previous research studies have discussed the individual requirement of mobile users in MSNs. In this paper, we propose a user-defined location-sharing scheme (ULSS) to achieve enhanced privacy preservation under different contexts. We present a coarse-grained proximity detection method and a lightweight order-preserving encryption- (OPE-) based method to provide the users with flexible privacy preservation at different privacy levels. The proposed scheme preserves user's location privacy with respect to SP, friends, and other adversaries, getting rid of the introduction of fully trusted party (TTP). Extensive experiments were conducted to verify the effectiveness and efficiency of our proposed scheme.

1. Introduction

With the development of the smartphone and wireless communications technology, mobile social networks (MSNs) have seen a tremendous rise and are becoming a kind of important tool for our daily communication. MSNs such as Foursquare, Wechat, and Google Latitude bridge the users' physical and social worlds taking advantage of location-aware capabilities by mobile devices. Sharing and interaction around content and information is the key feature of the MSNs. In particular, *location sharing* [1], as an important function of MSNs, allows users to share their current locations to the friends on the social networks or to find the nearby friends within a certain physical distance. Based on it, MSNs bring mobile users better experiences and services by various location-based and personalized services, e.g., proximity-based detection [2] and friend locator [3], or just simple location sharing [1].

MSNs facilitate the mobile users and make it easier for social friends to connect in real life. While the users reap great benefit and convenience from MSNs, they also have high concerns about the privacy of location, since the users have to submit their exact location information to their friends or to the service provider (SP) when they enjoy the service. The centralized SP collects all user's information, e.g., the user identity, the exact location, and query content. If the private information is disclosed by potentially untrustworthy SP or is compromised by an adversary, it will put the users' sensitive information in jeopardy. Specifically, in MSNs, the user's location information is associated with the social network ID and the social relationships of users, which makes the user privacy preservation more significant and more complicated [4–7]. Therefore, there are both the development chance and challenge in MSNs, and the key is how to protect the privacy of users when providing high-quality services.

Over the past years, some approaches have been proposed to address location privacy in MSNs [8–10], and a few solutions [11–14] were presented to focus on user privacy preservation in case of location sharing. Li et al. [11] presented a secure location-sharing scheme in which the user information was stored in two kinds of servers, e.g., location server (LS) and social network server (SNS), which incurs high communication cost due to the multiple-round interaction between the user and the servers. Moreover, the privacy of users can hardly be guaranteed without the assumption of full trust in the SP. In our previous work [3], we proposed a lightweight privacy-aware friend locator (PAFL) to provide privacy guarantee for the user at low computational and communication cost. However, the coarse-grained friend locator method cannot meet the personalized needs of mobile users. The challenge is that the user may submit diverse requirements and wants to set different levels of privacy protection due to the varying contextual conditions. Here is a typical example of proximity-based location sharing: Alice travels to a city, she first wants to find out the friends who are located in the destination she is visiting by the MSNs, e.g., Facebook Connect. Next, according to the name list returned by server, Alice wants to query some specified friends whether they are in the nearby vicinity region defined by her, e.g., within a mile or so. In this application, Alice refuses to reveal her exact location to any third party (including SP and other users), whether she is the initiator of the query or the location provider (friend). While the existing schemes [11, 14] are inflexible for users to change the privacy setting, it is difficult to meet the individualized demand preferences for different users in the traditional system.

In this paper, to satisfy the users' individual requirements, we propose a user-defined location-sharing scheme (ULSS) to realize proximity query of users in different scales of vicinity regions. Our motivation is to provide the mobile users with flexible privacy controls under different contexts in an efficient and friendly way. We first propose a coarse-grained proximity detection method based on the Hilbert curve to determine the friends in a wide range of vicinity. Then, we provide a lightweight order-preserving encryption- (OPE-) based method to enable users to query the specular friends in the nearby specified vicinity, e.g., within 5 km. In the whole process, users can submit their proximity query according to their individual requirements and different personal privacy settings, without revealing the exact location information to any SP or to any other possibly malicious users.

Our main contributions are summarized as follows:

(1) We propose a flexible ULSS for private proximity detection, which allows each user to maintain his own privacy-preserving policy and provides the users with flexible privacy protection at different privacy levels.

(2) We propose two protocols to process the user queries in different phases, which preserve user's location privacy with respect to SP, friends, and other adversaries. We also prove that the proposed scheme is secure under the stronger security model with enhanced privacy.

(3) We evaluate the performance of our ULSS scheme by extensive experiments. Experiment results demonstrate that our scheme is extremely efficient for coarse-grained proximity detection and provides user-defined nearby friend locator at low computational and communication cost as well.

In our scheme, the fully-trusted third party is not required, the SP is assumed to be "honest but curious," and it honestly executes instructions in the system, like storing user data, handling user's queries, and returning the results to the issuer. However, it is curious about the collected information of users and may try to determine or locate a query user.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the technical preliminaries, including Hilbert curve and OPE. Then, the overview of the system is stated in Section 3. In Section 4, we describe the system design, and the security analysis of our scheme is provided in Section 5. We conduct a set of experiments to evaluate the effectiveness of our scheme in Section 6. The related work is presented in Section 7. Finally, the conclusion and future work are discussed in Section 8.

2. Preliminaries

2.1. Hilbert Curve. The Hilbert curve is a space-filling curve [15], which visits each point once and once only in a specified order by some algorithms in 2-dimensional (2-D) or multidimensional space. Given 2-D square space with the resolution of $n * n$ divided into $2^N * 2^N$, $n = 2^N$ equal-sized cells, the N-order Hilbert curve recursively resolves the space into four equal-sized blocks by a defined way. For the N-order Hilbert curve, the iterative process is described as four subblocks that are replicated and partitioned after rotation and reflection from the (N-1)-order curve blocks.

According to the order of the curve traverses, the sequence number for each cell is determined, which ranges from 0 to $2^{2N} - 1$. The integer value associated with each cell is denoted as the Hilbert value (H-value). For example, Figure 1 shows the 2-order Hilbert curve goes through $2^2 * 2^2$ cells in the given space, and each cell is given a sequence number (H-value). The Hilbert curve is often used as one of the space transformation tools, transforming a 2-D space point into 1-D H-value. Similar to our previous work [3], we present each cell in the given square space by a grid coordinate <X, Y>, and the corresponding Hilbert value of each cell can be determined by a Hilbert curve.

Definition 1. Given a Hilbert curve visiting each cell in a 2-D square space, the grid coordinate of a cell $c < X_c, Y_c >$ can be transformed to 1-D Hilbert value by a function $\dot{f}$ as follows:

$$H(c) = \dot{f}(\langle X_c, Y_c \rangle), \tag{1}$$

where $\dot{f}$ is the spatial transformation function and $0 \le X_c, Y_c \le 2^N - 1, 0 \le H(c) \le 2^{2N} - 1$. Once the Hilbert

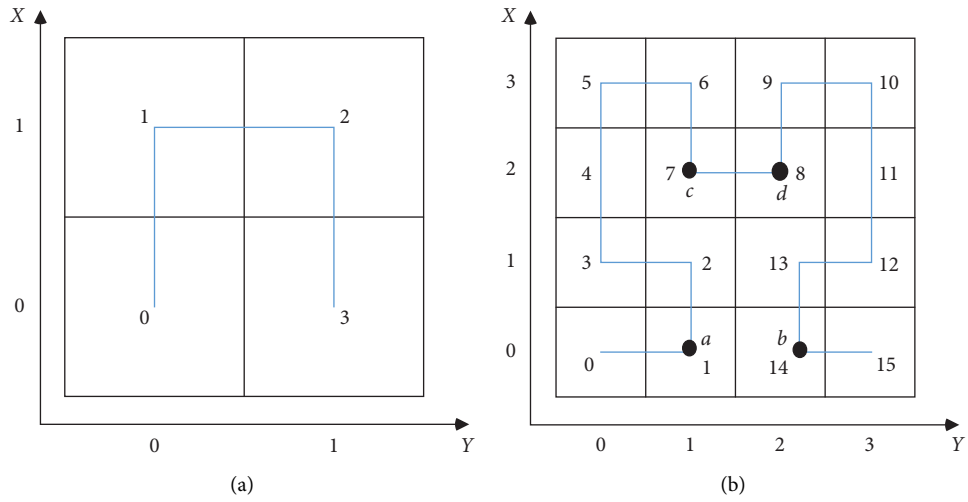


Figure 1: Hilbert curve: (a) 1-order; (b) 2-order.

curve setting parameters are determined, the generated H-values mapping to all cell are assigned. The setting parameters in the generating algorithm refer to the curve's starting point $<X_o, Y_o>$, curve order N, curve orientation Γ, and curve scale factor Θ. We term this parameter as spatial transformation parameter (STP), STP = $\{(X_o, Y_o), N, \Gamma, \Theta\}$. In our example, in Figure 1, users *a, b, c,* and *d* are with the grid coordinates of $\langle 1, 0\rangle$, $\langle 2, 0\rangle$, $\langle 1, 2\rangle$, and $\langle 2, 2\rangle$. As the 2-order Hilbert curve orderly visits each cell, generated Hilbert values of these users are 1, 14, 7, and 8, respectively.

In our study, we use the Hilbert curve to transform the user's real location for the purpose of privacy preservation. From the security aspect, the space transformation function f should be a one-way function, which is presented that it is easy to encode the elements of two-dimension space and generate one-dimension H-value but difficult to decode the one-dimension H-value back to the original space. The Hilbert curve has the above features and is suitable for our scheme. For an adversary without the STP, it is computationally impossible [15, 16] to reverse the function $f-1$ and determine the correct Hilbert curve by comparing the H-values for all original spaces.

2.2. Order-Preserving Encryption. Order-preserving encryption (OPE) [17–20] is a deterministic symmetric encryption scheme which preserves the sort order of plaintexts in the ciphertexts. This character makes OPE very useful when performing range queries on encrypted data. It allows a remote untrusted database server to index the encrypted data and directly performs range queries without decrypting the ciphertexts [21] (returning a result in the database whose decryptions fall within a specified range, e.g., $[a, b]$).

Definition 2. Order-preserving function. For any $A, B \subset N$, $|A| < |B|$,function $f: A \longrightarrow B$ is order preserving, if for all $i, j \in A$, $f(i) > f(j)$, and only if $i > j$.

Definition 3. Order-preserving encryption. A deterministic encryption scheme $S\varepsilon = \{\mathcal{K}, E(\cdot), D(\cdot)\}$, with associated plaintext-space $\mathcal{D}$, ciphertext-space $\mathcal{R}$, and encryption key-space $\mathcal{K}$, for all $\mathcal{K}$, if $E(\cdot)$ is an order-preserving function from $\mathcal{D}$ to $\mathcal{R}$, the deterministic encryption scheme $S\varepsilon = \{\mathcal{K}, E(\cdot), D(\cdot)\}$ is order preserving.

The OPE-based scheme in our system consists of the following algorithms, where *EncL* and *EncR* are interactive, *KeyGen* is probabilistic, and the rest are deterministic:

Key Generation: KeyGen $(1^k) \longrightarrow SK$. The KeyGen runs on the user side, it takes the security parameters k as input and outputs a secret key *SK*.

Location Encryption: EncL$\{l, SK\} \longrightarrow l'$. Given a location point l, the user outputs the ciphertext l' with the secret key *SK*.

Range Encryption: EncR$\{R, SK\} \longrightarrow R'$. Given a 2-element vector R, $R = [l_1, l_2]$, the user outputs the ciphertext R′with the secret key *SK*.

Decryption: Dec$\{l', SK\} \longrightarrow l$. Dec is a decryption algorithm taking the ciphertext l' and secret key SK as inputs and outputs the plaintext l.

Detection: Det$R, l' \longrightarrow \{0, 1\}$. It takes the ciphertext R' and l' as inputs, performs the vicinity detection, and outputs 1 if l falls into the range R, otherwise, it outputs 0.

3. Overview of the System

The main idea of our scheme is to enable users to share location information and to find their nearby friends in a flexible and privacy-preserving way. Each user u in the system determines his current location (e.g., longitude and latitude) by a GPS-equipped device, which is a point in Euclidean space, presented by a 2-tuple value $(x_u, y_u) \in R$. The distance between the user u and other mobile users, e.g., v is d_{uv}, $d_{uv} = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2}$. The user u in the system can specify a *vicinity region*, and we use the minimum bounding rectangle (MBR) to cover this region and determine it by the diagonal $\{(x_{V\min}, y_{V\min}), (x_{V\max}, y_{V\max})\}$.

The mobile user u can invite his friends v to join a *friend group* FG_u, and the invited user v may either accept or decline to join FG_u. If v accepts, the friendship relation F between u and v will be built as $\{(u, v)\} \in F$. In the same way, if user u is a friend of v, then users u and v are friends, and $\{(u, v), (v, u)\} \subseteq F$. The user u can manage the members of FG_u, and SP will maintain the friend list for him. The summary of notations used in the system is illustrated in Table 1.

3.1. System Model. The system architecture of the ULSS scheme is shown in Figure 2, which consists of the service provider, location-based social network server (LBSNS), and users. In the system model for our proposed protocol, the mobile user can be a request initiator or a participant (location provider). For ease of explanation, in the set of users, we assume Alice is the request initiator, and one of her friends Bob is the participant.

The mobile user Alice is an initiator who intends to share her real-time location in the MSNs and request the server for the proximity detection service. To preserve location privacy, Alice preprocesses the privacy information using the preloaded modules on her mobile device and then submits the encrypted location information to the server. Alice maintains different location policies according to the various contextual conditions, including the coarse granularity location privacy-preserving policy *Pa* and fine granularity location privacy-preserving policy *Pb*. With *Pa*, Alice can receive a user list from the server, and it shows the friends who are located over a large vicinity area (like a city) designated by her. Then, referring to some specified users on the received list, with *Pb*, Alice may query if they are in the nearby vicinity region (like around 2 km). The participant Bob is a friend of Alice. Bob follows the proposed protocol and the privacy policy of Alice. In the whole process, Bob only provides his encrypted location to the social network and preserves his real location information against to any other third party.

The service provider is a server in MSNs, which is an untrusted entity. Unlike the previous systems in [11, 22], which deploy the social network server and location server separately, in our architecture, the SP is an integrated server LBSNS. It provides users with two kinds of services, location-based services, and social network services. LBSNS achieves integrated service delivery, such as storing users' information and maintaining user friend list and offering dynamic user data update and adjustment. The LBSNS conducts proximity detection based on the encrypted locations of user and his friends and returns the results to mobile users based on their privacy-preserving policy.

3.2. Security Assumption. There are generally external attack and internal attack for the threat model [23]. The former is caused by unauthorized outsiders, and the latter is initiated by internal participant (for example, the users and LBSNS in our system). We assume the communication channels be secure by standard security schemes (such as SSL and SSH) or using cryptographic methods to resist internal attacks [24–26]. In the ULSS scheme, we consider only internal attack and assume conventional threat model in our system as follows: we consider the dishonest mobile users try to obtain other users' information outside the scope of the authorized access privileges. For example, Alice learns the exact location of Bob; Bob obtains the vicinity area or exact location of Alice. The untrusted mobile users may use some tools to forge their location, such attacks can be detected as presented in [27]. Besides, the LBSNS is assumed to be "honest but curious" by honestly following the proposed protocol in general but being curious about the content of queries, like the exact location and the real identity of users. In our system, LBSNS integrates location-based services and social network services on one server, which stores the information of all users, and social relationship network of them as well. In the ULSS scheme, all entities (including users and LBSNS) are assumed to be potential adversaries. The main security goal of our proposed scheme is to preserve user's location privacy against to all participants. In addition, the collusion between the malicious users and LBSNS is not allowed in our proposed ULSS scheme.

Table 1: Summary of notations.

Notation	Description
(x_u, y_u)	Point coordinate of user u
$\langle X_u, Y_u \rangle$	Grid coordinate of user u
(x_u^o, y_u^o)	Origin of the grid user u located
(x_u^f, y_u^f)	Offset to the ordinate origin of user u
G_u	Grid identification number of the user u
$H(u)$	Hilbert value of the cell user u located
(X_o, Y_o)	Starting point of Hilbert curve

4. System Design

In this section, we first present the user registration and then describe the user location representation and update. In our system, according to user-defined location privacy-preserving policies, we use different methods to perform proximity detection: coarse-grained proximity detection and user-defined nearby friend detection as described in this section.

4.1. Registration. Before each user enjoys the location-sharing services in the social network, he should first register with his ID_U on the LBSNS, who verifies and maintains the uniqueness of the ID. Generally, users can use pseudonyms in the registration for safety considerations. The initiator Alice builds a *friend group* FG_A by inviting her friends to join in. The LBSNS will generate a group ID (like a random 256-bit-string) for the FG_A, storing the user list of it. Alice can classify the users on the list into several subgroups, like colleagues, families, classmates, and so on, defined as $\mathrm{C}_1, \mathrm{C}_2, \ldots \mathrm{C}_n$, $\mathrm{FG}_A = \{\mathrm{C}_1, \mathrm{C}_2, \ldots \mathrm{C}_n\}$ The LBSNS creates a friend relationship network for each user and updates dynamically and maintains the information of friend list as requested by users.

Each user should set a distance threshold D_{su} to preserve the location privacy during the application's initialization

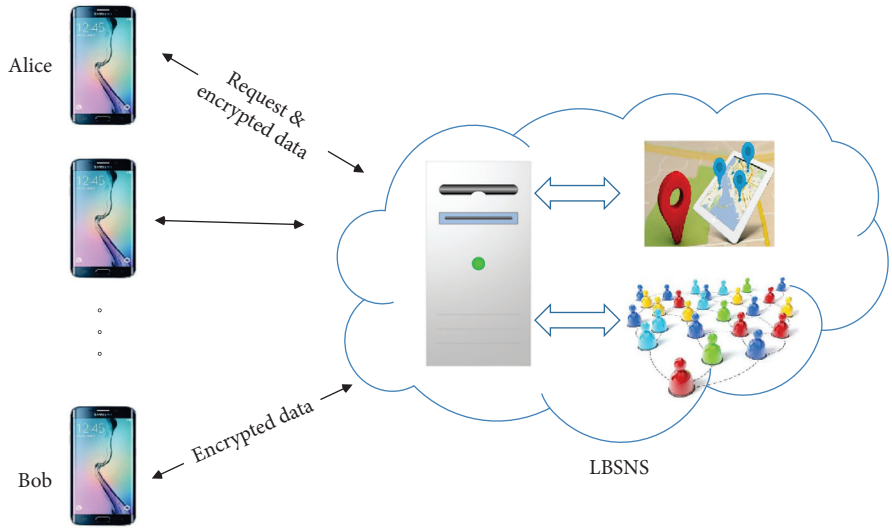


Figure 2: System architecture of ULSS scheme.

phase. D_{su} defines a minimum distance with which the user u allows his friends to search him. Otherwise, a user would be precisely located by the adversary continually reducing the range of vicinity detection. For example, Alice finds Bob somewhere around her by the vicinity detection service, and she can obtain the accurate location information of Bob according to continuous queries, such as by returning results when the range is set to 2 km, 1 km, 500 m, and so on. The SP will build a profile Pr for each user to keep these settings, e.g., for Alice, $Pr_A = (ID_A, PSW_A^*, FG_A, D_{SA}\}$, where ID_A is the ID of Alice, PSW_A^* is the ciphertext of the password, FG_A is the friend group of Alice, and D_{SA} is the distance threshold. The ULSS scheme allows users to customize and personalize their profiles to satisfy different users' needs.

4.2. Location Representation and Update. Each user u obtains his location in the forms of latitude-longitude by the GPS $(\text{lat}_u, \text{lon}_u)$. The ULSS scheme enables the user to transform his location by the Hilbert curve method as described in Section 2. Thus, the system first assigns a space range, where mapping the geolocation into a coordinate point denoted by $(x_u, y_u) \in \mathscr{R}$ in Euclidean space. For example, the system can define the square scale using the longitude and latitude of 4 points of east, west, north, and south. As shown in Figure 1, the given space range is divided into $2^N * 2^N$ equal-size cells and formed a grid system. Each grid in the space is numbered uniquely from $\mathbb{N}$. To protect the location privacy, each user in the ULSS scheme transforms his location coordinate (x_u, y_u) into two parts. The first part (x_u^o, y_u^o) is the origin coordinate of the grid where he is located in the grid system, the second part (x_u^f, y_u^f) is the offset coordinate in the grid, and $x_u = x_u^o + x_u^f$, $y_u = y_u^o + y_u^f$. We assume that the location coordinate range is separated into N scale levels, the coordinates of a user can be represented as $G_u \| (x_u^f, y_u^f)$. Here G_u is the grid identification number of the user u in the N-th scale level, and $0 \le G_u \le 2^{2N} - 1$, (x_u^f, y_u^f) is the offset coordinates in the corresponding grid. In the ULSS scheme, if the transformation parameter STP is defined, the H-values mapping to all cell are assigned, which means the G_u can be obtained by the H-value according to equation (1).

In general, if a user's motion or movement trajectory is in a small range (e.g., in a city), the origin coordinates of the grid in the given space are not changed, and the user can specify a time interval with which the user automatically updates his offset coordinate. While user traverses a large area, e.g., traverse different grids in the grid system, the changes in offset coordinates exceed a certain degree (e.g., greater than the unit length of a grid), and the user should recalculate his location $G_u \| (x_u^f, y_u^f)$ and submit related information to the server. In this way, we can reduce the computation and communication overheads for the user and improve the performance of the whole system.

4.3. Coarse-Grained Proximity Detection. In this phase, a user queries an LBSNS and gets the results showing friends in the coarse-grained vicinity area (such as in the same city). The main idea of coarse-grained proximity detection is to utilize the Hilbert curve method to transform user's location and the coordinates of vicinity. The scheme cloaks user and his vicinity by mapping 2-D points into identification numbers in corresponding scale level which can be presented by H-value under given STP. The LBSNS provides results by comparing the hash values of all values submitted by mobile users. The processes are shown in Figure 3, and the detailed steps are as follows:

Step 1. When the user Alice logs onto the system and wants to view the friends who are located in the coarse-grained vicinity area, she can submit a query $Q_A = \{ID_A, list_A, N\}$ to the SP, where ID_A is the login ID of Alice, which can be pseudonymous; $list_A$ is the user list in the friend group specified by Alice, which can be a classification of his friends. Actually, Alice can search

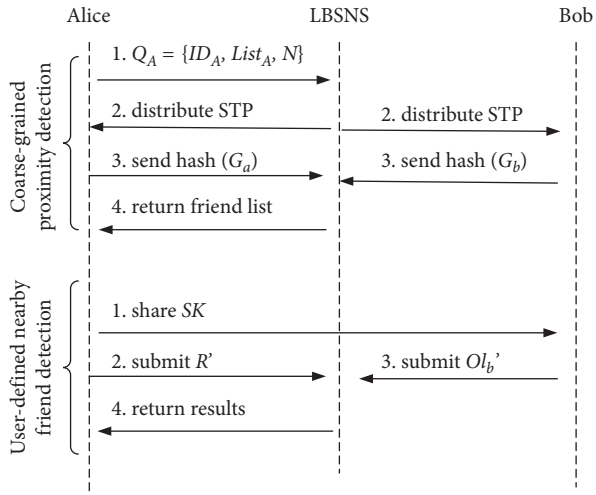


Figure 3: Processes in the ULSS scheme.

several classifications in his friend group FG_A, like $\{C_i, C_{i+1}, \ldots, C_j\}$, $1 < i, j < n$, for simplicity, and we denote it by $list_A$. According to the coarse granularity location privacy-preserving policy *Pa*, Alice sets the granularity *N*. For example, Alice wants to search the friends in the $list_A$ who are located over a large vicinity area, e.g., the same city she stays. In the given space range, Alice assigns the appropriate granularity N so that the vicinity can be presented by a grid number in the Hilbert transformation. In the Hilbert STP, N is the scale level $1 \le N \le 16$. Generally, the larger the N, the bigger the vicinity area.

Step 2. Receiving upon the request from an initiator Alice, the LBSNS generates the STP according to the Hilbert curve method for users. In the STP, the scale level N is defined by Alice. The system determines longitude and latitude coordinate ranges as curve scale factor Θ, mapping it to the points in the given space range. In this way, the given 2-D space can be transformed to be a grid system with the equal size of $2^N * 2^N$ grids. Then, the SP randomly selects the curve's starting point $<X_o, Y_o>$, and curve orientation Γ to build the STP and determine the Hilbert curve. Next, the LBSNS searches the friend list according to the $list_A$ submitted by Alice and distributes the generated STP to Alice and users on the list.

Step 3. When receiving STP, each user transforms his location point (x_u, y_u) to $G_u \| (x_u^f, y_u^f)$, and the process is as follows: Alice first maps her point location in Euclidean space (x_a, y_a) into the $2^N * 2^N$ grids system $<X, Y>$. For example, if the square space range Θ is determined by 4 points of east, west, north, and south, (x_E, y_E), (x_W, y_W), (x_N, y_N), and (x_S, y_S), the unit length $Unit$ of each grid can be presented by $(x_W — x_E)/2^N$ or $(y_S — y_N)/2^N$. With the site (x_a, y_a), Alice can locate her grid coordinates $<X_a, Y_a>$ in the grid system by calculating $X_a = [(x_W — x_a)/Unit]$ and $Y_a = [(y_a — y_S)/Unit]$ and also gets the offsets (x_a^f, y_a^f) in the corresponding grid. Then, she transforms the grid coordinates $<X_a, Y_a>$ to a Hilbert value $H(a)$ using equation (1) by the determined STP. Now Alice obtains the $G(a)$ and can transform her location (x_a, y_a) into $G_a \| (x_a^f, y_a^f)$. In order to prevent other entity getting real location information, Alice submits the hash value of it, $hash(G_a)$, to the LBSNS, where $hash(\cdot)$ is a collision-resistant hash function. In this way, Alice cloaks her exact location into a region, presented by $hash(G_a)$. The friends of Alice, e.g., Bob also transforms his locations and sends the $hash(G_b)$ to the LBSNS in the same way.

Step 4. The LBSNS collects the transformed location information from Alice and her friends. Then, it searches the users who have the same $hash(G_u)$ as Alice and returns the neighboring users located in the vicinity to Alice under the coarse-grained setting *Pa*.

4.4. User-Defined Nearby Friend Detection. When Alice receives the results of the coarse-grained proximity detection from LBSNS, she can further select some specified users she has interest to perform nearby friend detection under the fine granularity policy *Pb*. For example, referring to some particular users on the received list, presented by $list_B$, Alice may query if they are in the nearby vicinity region (like around 2 km). Note that the neighborhood range in this phase can be customized arbitrarily by Alice under the constrain conditions of security assumption. We intend to use lightweight OPE technology to implement the user-defined nearby friend detection, and the processes are as follows:

Step 1. Alice generates an encryption key SK according to the OPE key generation algorithm $\text{KeyGen}(1^k) \longrightarrow SK$ and shares the key through the secure channel to the friends to be queried.

Step 2. Alice customizes a neighborhood range to retrieve the nearby friends. For example, the range around her A meters or the range with A and B meters in the x-axis and y-axis directions. The scheme generates a rectangle R to cover this range, $R = [r_l, r_u]$, where $r_l = (x_a - A, y_a - B)$ and $r_u = (x_a + A, y_a + B)$ denote the lower-left corner and upper-right corner of the vicinity range R, respectively. Notice that, here we use the offset coordinates in the corresponding grid to present the vicinity R. Then, Alice encrypts R, by $\text{EncR}\{R, SK\} \longrightarrow R'$.

Step 3. When a specific user Bob receives the request, he encrypts his own location before submitting. The location point of Bob l_b is presented by $G_b \| (x_b^f, y_b^f)$. Here, Bob only needs to send his encrypted offset coordinates to further ensure the safety of his data. In this way, even some malicious attacker can decrypt the ciphertext, he has no way to obtain the real l_b without G_b. If we denote the offset coordinates of Bob by Ol_b, Bob inputs it to the encryption function $\text{EncL}\{Ol_b, SK\} \longrightarrow Ol_b'$ and submits Ol_b' to the LBSNS to avoid potential security risks from trusted third party (TTP) free server.

Step 4. LBSNS performs the range query on the ciphertexts directly and returns the results whether l' falls into the associated range by $\mathrm{Det}(R', l') \longrightarrow \{0, 1\}$. LBSNS collects the information from all users of $list_B$ conducting the proximity detection and returns those who is in the nearby vicinity R to Alice.

5. Security Analysis

The ULSS scheme uses the Hilbert curve method to perform coarse-grained proximity detection and utilizes the OPE method to conduct user-defined nearby friend detection. We give the security analysis for these two technologies. The security assumption and requirements are described in Section 3.2.

5.1. Hilbert Curve Transformation. In general, the privacy protection methods in most literatures are public, and the technology of Hilbert curve in the ULSS scheme may be obtained by some adversaries. Once an attacker learns some background knowledge of the users' original locations or their transformed locations, they may guess the secret key (e.g., STP in the ULSS scheme) with a certain probability. However, some researchers in [15, 16, 28] stated that it is computably impossible to get the correct STP by comparing H values of all locations, and it can resist brute force attacks.

Brute-Force Attack. In STP, we represent curve's orientation Γ in terms of q bits, and it will generate 2^q values in the entire 360° space. The malicious adversary needs to make 2^q attempts to determine the right Γ. Refer to the curve's starting point (X_0, Y_0), and we represent the value with q bit data, respectively. To search for the correct (X_0, Y_0) over $2^q * 2^q$ elements, it requires the adversary to try $2^q * 2^q$ candidate coordinate values on the X and Y axes. In our scheme, the square space range Θ is fixed value, and the order of the Hilbert curve N is determined by users. Therefore, we have $(N * 2^q * 2^q * 2^q)$ solutions for the entire space, where q is the number of bits presenting each parameter of STP. N is the order of the Hilbert curve, and $1 \leq N < 16$. If the value of q is big enough, the possibility of getting the correct STP by trying combinations of parameters is negligible. For example, if $q = 32$, the complexity of exhaustive search for the STP would be $O(2^{3*32})$.

Therefore, for anyone without the STP, it is computationally impossible to reverse the process of the spatial transformation and get the users' original locations. The one-way mapping of the Hilbert curve makes the transformation function $\dot{f}(\langle X_c, Y_c \rangle)$ be a secure encryption function [15, 16]. Any adversary has no way to learn the user's real location without the encryption key STP. We can say that the Hilbert curve transformation is a very appropriate approach to preserve user location privacy in our ULSS scheme.

5.2. Security of Order-Preserving Encryption. The researchers in [29, 30] formalized a security requirement for OPE and proposed an efficient blockcipher-based scheme provably meeting their security definition. The OPE scheme has been proven [29] that it is *security under distinct chosen-plaintext attack* (IND-DCPA). Popa et al. [31] presented that the ideal security guarantee for order-preserving encryption is to reveal no additional information about the plaintext values besides their order. They proposed an ideal security protocol for OPE, which can achieve *indistinguishability under ordered chosen-plaintext attack* (IND-OCPA) security. The ULSS scheme utilizes the OPE method to conduct user-defined nearby friend detection. Alice and her friends hold the secret key SK, derived from the security of the OPE scheme, and any other entities including the LBSNS without the key cannot get the real location of users. Therefore, the security of user-defined nearby friend detection in the ULSS scheme can be achieved from that of the OPE scheme.

6. Evaluation

In this section, the efficiency and effectiveness of our ULSS scheme were evaluated by the extensive experiments. During the phase of coarse-grained proximity detection, we focus on the location transformation using the Hilbert curve method, which is mainly affected by the parameter of Hilbert order N. During the phase of user-defined nearby friend detection, we focus on the location encryption by OPE, which is mainly affected by the size of data. We also conducted experiments to evaluate the computational and communication overheads of each entity under different system settings, like the size of vicinity s or number of friends n.

All experiments were conducted on Java Development Kit (JDK).1.7 with the Intel (R) Core (TM) i7-7500U 2.70-GHz CPU and Win 10 OS. We used the 256-bit OPE symmetric encryption algorithm and 256-bit SHA hash function to ensure data confidentiality and integrity. The cryptographic algorithms were performed with JPBC library. In the experiments, we used the Web API of Amap [32] to get the GPS positioning by the function *GetLongitudeAndLatitude* and initialized the location of each user by calling function *GetCurrentPosition*. The studies utilize uniform data sets to randomly generate batches of 10,000 user locations and then map them into the points in a $65536 * 65536$ area in 2-D space. We assume each mobile user stores his friend list on the LBSNS, and each user has a maximum of 2,000 friends. Each mobile user can specify an arbitrary vicinity to search for his nearby friends. For the security and privacy issue, the minimum size of vicinity is set to $1\,\mathrm{km}^2$, and we got the average data from 20 experiment results.

6.1. Hilbert Curve Order. In the phase of coarse-grained proximity detection, Alice and her friends use the Hilbert curve to transform their exact locations into a cloaked region, presented by an H-value. The LBSNS gives the results of coarse-grained detection according to the hash values submitted by all users, and the computational overhead of the LBSNS is almost negligible. In this set of experiments, we mainly evaluated the processing time on the user side. When performing the Hilbert curve transformation, curve's

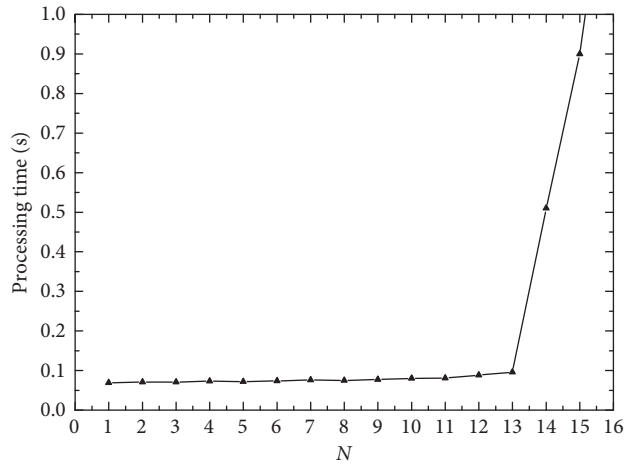


Figure 4: Computational overhead of location transformation for the user.

starting point is set to $\langle 0, 0\rangle$ and the curve orientation Γ is set to be clockwise, the square space range Θ is determined by 4 points of east, west, north, and south in China. Figure 4 shows the average computational cost during the location transformation with various Hilbert curve order N for the user. As shown by it, the computational processing time of the user slightly increases as $N \leq 13$, and the transformation time is no more than 0.2 second. In fact, in our ULSS scheme, if $N \geq 16$, the generated region is too small to cloak user's exact location. Generally, in the initialization phase, we assign the value of N is no bigger than 15. After transformation, each user submits the hash value of his transformed location to the LBSNS to prevent the information leakage on the third entity. Table 2 shows that this operation completes within several milliseconds. Thus, in practical application of the scheme, for the user, we can say the Hilbert curve transformation is computationally efficient.

6.2. Encryption and Decryption. In the phase of user-defined nearby friend detection, we used the 256-bit OPE symmetric encryption algorithm to ensure the data confidentiality for users. Table 2 shows the computational overheads of the encryption and decryption in various data sizes. In the ULSS scheme, the initiator Alice encrypts her vicinity range R, and her friend Bob encrypts his offset coordinates (x_b^f, y_b^f), the size of them is no more than 64 bytes. For each user, we can see, from Table 2, the average processing time for the encryption is about 13 ms. For the LBSNS, it conducts the nearby friend detection directly on the ciphertexts since the OPE preserves the sort order of plaintexts in the ciphertexts. It means that, without decrypting any ciphertext, the LBSNS can return a result in the list whose decryptions fall within a specified range R. This process can be completed by comparing the value of ciphertexts, which ensures minimal overhead on the LBSNS.

Table 2: Computational overhead of encryption and decryption.

Size (bytes)	32	64	128	1024	10 k	1 M
Hash (ms)	6	6	8	9	11	14
Encryption (ms)	11	13	21	122	136	2049
Decryption (ms)	10	11	29	130	522	43822

6.3. Number of Friends. We gave an experiment to evaluate the efficiency of the ULSS scheme in terms of various parameters, e.g., number of friends n and the size of vicinity s. We illustrated the experimental results by the computational time and traffic overheads from the Alice, Bob, and the LBSNS, respectively. We assumed Alice submits a coarse-grained proximity query to LBSNS, upon receiving the results, and she selected half of friends from the list to further put forward nearby friend query. In this process, we ignored the communication network transmission delay, as it was impacted by different network environments. Figure 5 shows the computational time of each entity in the scheme when the number of friends n changes from 20 to 2000, and we set the default vicinity was 2 km^2. The processing time is 92–98 ms for Alice and 80–85 ms for Bob. While it grows for LBSNS with number of friends, because the server should handle the proximity detection requirement from each participant, the greater the number of friends is, the longer the time it takes. Similarly, LBSNS needs to distribute STP to Alice and her friends, and its communication cost grows linearly with n (as shown in Figure 6). During the phase of user-defined nearby friend detection, Alice should share the encryption key to her friends, the traffic also grows, and the communication cost of Bob maintains at a very low level.

6.4. Size of Vicinity. In this set of experiments, the computational and communication cost of each entity in the ULSS scheme is evaluated when the size of vicinity s changes from 1 km^2 to 20 km^2. We assumed Alice first specified the Hilbert order to 5, when she got the results in the coarse-grained region, she defined a nearby vicinity (the size is s) to search for proximity friend. Figures 7 and 8 show that the computation and communication overheads for Alice, and SP slightly grow with increasing vicinity area, while it holds constant for Bob.

6.5. Comparison with Other Approaches and Discussion. We compared our proposed scheme with other similar approaches for a holistic evaluation of our system. Table 3 shows the results in terms of TTP, server, cryptography method, and efficiency characteristics when comparing with approaches like MLS in 2017 [11], VPPLS in 2017 [13], and UDPLS in 2017 [22].

From the table, we find that some approaches, e.g., the VPPLS scheme, need a fully trusted third party (TTP) to verify the user ID and run some cryptographic computation. In the TTP-based schemes, the third party stores user's sensitive information and some important information of the system, which may put the system in jeopardy if it is comprised. However, our scheme uses the location

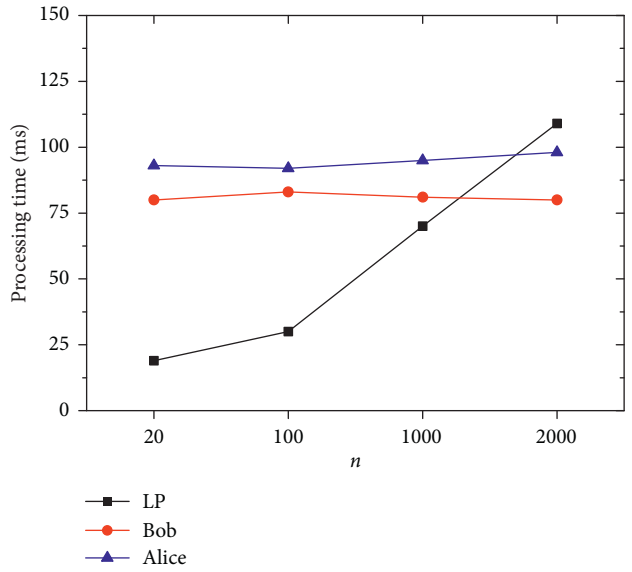


Figure 5: Computational overhead with various numbers of friends.

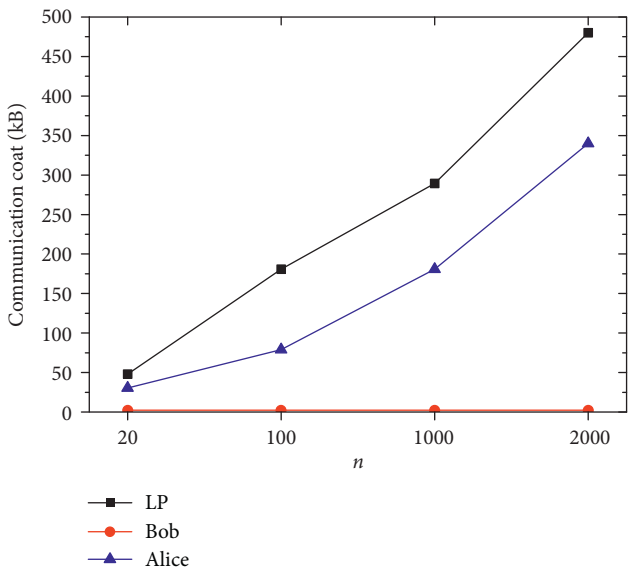


Figure 6: Communication cost with various number of friends.

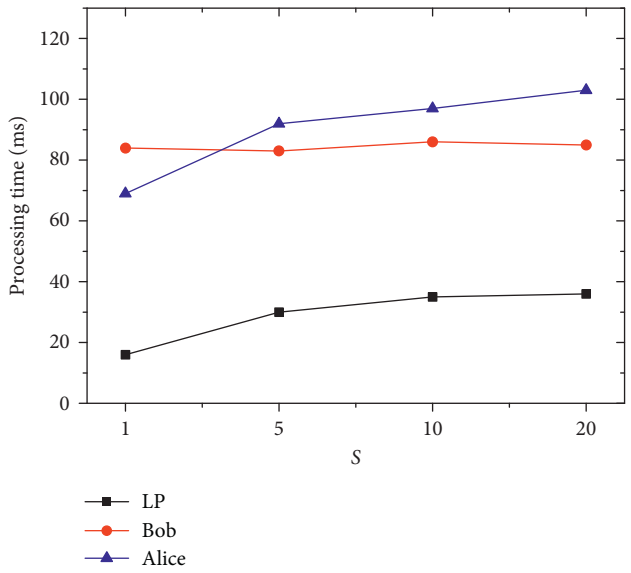


Figure 7: Computational cost with various sizes of vicinity.

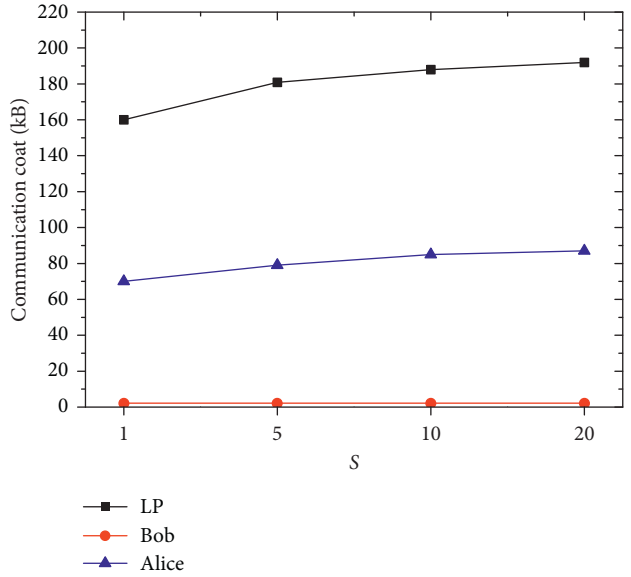


Figure 8: Communication cost with various sizes of vicinity.

transformation and OPE method to protect user privacy getting rid of the introduction of TTP.

Some exiting solutions, e.g., MLS and UPDLS, separately deploy two different servers including location server and social network server to store information and provide users two kinds of services, location-based services, and social network services. The main drawback of LBS/SNS separated architectures is the multiple-round interactions between the servers, which may incur higher communication cost and cause great computation and storage burden as well. In practice, mobile users can enjoy their online social network services by the wireless mobile network and, meanwhile, obtain their exact location information through the GPS equipped in the smart terminals. Nowadays, location sharing and services are important functions of MSNs, and integrating two kinds of services can bring more convenient and flexibility for users, which has become an indispensable tool of users' daily life. In security aspect, in the LBS/SNS separated architectures, the user's location information and user profiles are separately stored in two servers, and they are assumed to be no-collusion to guarantee the system safety. However, commonly, in some location-sharing scenarios, the location information can be easily captured by adversaries, if the server providers collude with them or collude with each other, the user privacy and system safety are insecure, which makes these distributed architectures unpractical in real life. In our model, these two kinds of servers

Table 3: Comparison of our proposed with other protocols.

Protocol	TTP	Server	Cryptographic methods
MLS [11]	No	Separately deployed LS and SNS	AES and RSA signature
VPPLS [13]	Yes	Integrated LBSNS	Homomorphic encryption
UDPLS [22]	No	Separately deployed LS and SNS	AES and RSA signature
Our scheme	No	Integrated LBSNS	OPE

are combined by one LBSNS, providing integrated service delivery.

In terms of cryptographic methods, the VPPLS protocol used the homomorphic encryption method to compute the distance between Alice and her friends without disclosing the exact position to other party, and MLS and UDPLS rely on symmetric encryption AES and asymmetric encryption RSA signature to ensure the data integrity and confidentiality. These privacy preservation approaches can provide user location sharing privately and securely. However, the main concern is the heavy computing burden both on user and server sides when running the encryption algorithms. Different with the traditional encryption method, our scheme utilizes efficient OPE algorithm to protect user data, and the ULSS only needs to encrypt very small amount of user data, in our evaluation result, Table 2 shows that user only takes several milliseconds to complete the encryption process. Besides, for the LBSNS, it can directly conduct the proximity detection by comparing the ciphertexts, without having to decrypt, which can greatly save the computing and communication resources. Thus, we can deduce that our scheme is more efficient.

Discussion. In our proposed scheme, we presented two methods, Hilbert curve-based proximity detection method and lightweight order-preserving encryption- (OPE-) based method, to provide the users with flexible privacy preservation in an efficient and friendly way. The evaluation results of Hilbert curve transformation show that the processing time on the user side and server side is very small. And from the experiments with respect to order-preserving encryption, we can also find that the average processing time of users is several milliseconds. Meanwhile, due to OPE, the server can directly perform range queries without decrypting the ciphertext, which minimizes the overhead on it. Compared with the homomorphic encryption- (HE-) based method [13] and AES&RSA-based method [11, 22], we can state that our ULSS is more efficient and practical. From the aspect of security, in some existing solution, the privacy of users is guaranteed based on the assumption of fully trust in the service provider or TTP. However, the TTP-based structure results in trust issues, like the single-point failure and the bottle-neck problem of communications, and if the TTP is comprised, the security of the entire system is at risk. Our scheme is TTP-free architecture, which can protect user location privacy with respect to SP, friends, and other adversaries. We also prove that the proposed scheme is secure under the stronger security model with enhanced privacy. Moreover, unlike the multiserver structure in [11, 22], in our scheme, the LBS and SNS are integrated by one server to provide entire and convenient services, and it also can reduce communication and computational overheads and security risks as well.

7. Related Work

7.1. Location-Based Services. With the popularity of LBS, the concern of privacy leakage in LBS raises, and many researches for privacy preservation have been proposed. K-anonymity [33] is one of the popular technologies to solve the privacy leakage issue in LBS, which employs a trusted third party (TTP) called *anonymizer* to replace the exact location of the user by a cloaked area including at least K users so that the user location is indistinguishable from $K-1$ other locations. Based on this fundamental idea, researchers [34–37] proposed efficient methods and models to construct K-anonymity spatial region (K-ASR) to protect user privacy. The authors in [36] proposed a location privacy-preserving K-anonymity method based on the credible chain, in which the optimal K value for the user is determined according to the user's environment and social attributes. The authors in [37] proposed a privacy scheme through caching and spatial K-anonymity (CSKA) and utilized the Markov model to predict the next query location according to the user mobility in continuous LBS. The TTP-based schemes, to some extent, solve the problem of privacy leakage in LBS. However, the introduction of the *anonymizer* in these schemes actually transfer users' trust from the SP to the intermediate entities. If it is compromised by the adversaries, it will pose a serious threat to users. Thus, these TTP-based schemes can only provide limited security assurance. In order to address the problem, in our previous work [38], we designed a privacy-preserving scheme based on location transformation getting rid of the fully trusted entities to provide enhanced security. The *anonymizer* can provide users accurate results without knowing any information about a user's real location.

The TTP-free schemes were adopted in the distributed peer to peer (P2P) environment to protect user privacy. Montazeri et al. in [39] introduced an information-theoretic notion for location privacy offering two models in both snapshot LBS and continuous LBS. Huang et al. [40] proposed a multimodal Bayesian embedding model (MMBE) for point-of-interest (POI) recommendation on location-based cyber-physical social networks. Sangaiah et al. used machine learning techniques to propose a method for conserving position confidentiality of roaming PBSs users in [41]. The authors in [42] considered the privacy and utility requirements of each user to propose an optimal user-centric location obfuscation mechanism. Sun et al. [43] introduced the location-label based (LLB) algorithm to distinguish locations of mobile users to sensitive and

ordinary locations and designed three protocols to protect user privacy for different user environments. Zhang et al. [9] proposed a deviation-based query exchange (DQE) scheme to obfuscate the users' query point to mitigate trajectory disclosure in MSNs. In our previous work [44], we presented a collaborative trajectory privacy-preserving (CTPP) scheme for continuous LBS queries, in which trajectory privacy is guaranteed by caching-aware collaboration between users. The main idea is to spatiotemporally break the correlations of continuous LBS queries to prevent the adversary from reconstructing a user's actual trajectory. The main drawback of TTP-free scheme is that multiple-round interaction between the servers and the user may cause a higher communication cost and may incur a higher computation overhead on the user side.

In addition, differential privacy technology [45–47], mix-zone method [48, 49], and encryption-based methods are also adopted to protect user privacy in LBS.

7.2. Location-Sharing Service. Our paper briefly focuses on a popular LBS location-sharing service, which enables users to share their current locations by the GPS-enabled devices to their friends on the MSNs or to find whether any friends are within a given vicinity area. It makes a large number of social network applications by the virtue of the smartphone and MSNs. Previous research studies [11, 12] discussed the issue of privacy preservation for the location-sharing service. In order to protect location privacy and social network privacy of mobile users, Li et al. [11] proposed a location-sharing construction with multiple location servers, in which the user's friend set was divided into multiple subsets randomly. Zheng et al. [2] presented a location tag construction method by environmental signals to provide an unforgeable location proof and used Bloom filters to efficiently represent users' location tags and vicinity regions. In the spatial-generalization-based method [50, 51], the user location space is divided into grids, and the precise position of user is replaced by the generalized grid prior to sending to SP. The authors in [50] proposed two protocols named "Hide and Seek" and "Hide and Crypt" based on spatial generalization to offer private proximity detection. Based on the "one degree" grid, Jing et al. in [51] presented a flexible and private proximity testing (FPODG) protocol. However, in the existing solution, the privacy of users can hardly be guaranteed without the assumption of fully trust in the service provider. Homomorphic encryption- (HE-) based privacy-preserving location-sharing scheme methods [13, 14] allow mobile users to compute distances between them and their friends without knowing the exact locations of each other. Based on HE, Tang and Cai proposed privacy-preserving location-sharing scheme (VPPLS) to build two models to enable users to query their friends' location in a secure way, which allowed user to make a classification of his friends and provide a verification for query result. The main drawback of these methods is the large computational and communication cost of the system, which makes the scheme less practical. Li and Jung [52] used the technology of ciphertext policy attribute-based encryption (CP-ABE) to design fine-grained privacy-preserving location query protocol (PLQP), and it satisfies different levels of location query and realize fine-grained and multileveled access control. These cryptographic-based methods can provide strong privacy guarantee for users at the high expense of computation and traffic. We proposed a lightweight privacy-aware friend locator (PAFL) in our previous work [3] to provide privacy guarantee for the user in an efficient way. However, the coarse-grained friend locator method cannot meet the personalized needs of mobile users. In this paper, we proposed a flexible ULSS scheme for private proximity detection, in which each user can maintain his own privacy-preserving control policy.

8. Conclusion

In this paper, we proposed user-defined location-sharing scheme in mobile social networks to protect user privacy in proximity detection service. We proposed two protocols, coarse-grained proximity detection and lightweight user-defined nearby friend detection to realize proximity query of users in different scales of vicinity regions. Our scheme is TTP-free architecture, in which the location-based services and social network services are integrated by one LBSNS server, making it more practical and convenient. Experimental results suggest that the ULSS scheme consumes low computational and communication overheads even for a large size of vicinity area and mount of friends. In our scheme, we used a minimum bounding rectangle to cover the user-specified vicinity region, which affects the accuracy of the results. In the future, we will propose an approach which can support the irregular shape vicinity to deliver services in more accurate ways.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61802076, 61632009, and 61872097), Guangdong Provincial Natural Science Foundation (Grant no. 2017A030308006), High-Level Talents Program of Higher Education in Guangdong Province (Grant no. 2016ZJ01), and CERNET Innovation Project (Grant no. NGII20190408), and CERNET Innovation Project (Grant no. NGII20190408).

References

[1] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: attacks and defense," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 646–660, 2018.

[2] Y. Zheng, M. Li, W. Lou, and T. Hou, "Location based handshake and private proximity test with location tags," *IEEE Transactions on Dependable & Secure Computing*, vol. 14, no. 4, pp. 406–419, 2017.

[3] T. Peng, Q. Liu, G. Wang, and J. Chen, "A lightweight privacy aware friend locator in mobile social networks," in *Proceedings of the 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pp. 17–23, Guangzhou, China, December 2017.

[4] Q. Liu, P. Hou, G. Wang, T. Peng, and S. Zhang, "Intelligent route planning on large road networks with efficiency and privacy," *Journal of Parallel and Distributed Computing*, vol. 133, pp. 93–106, 2019.

[5] Y. Wu, H. Huang, Q. Wu, A. Liu, and T. Wang, "A risk defense method based on microscopic state prediction with partial information observations in social networks," *Journal of Parallel and Distributed Computing*, vol. 131, pp. 189–199, 2019.

[6] Y. Wu, H. Huang, N. Wu, Y. Wang, M. Z. Alam Bhuiyan, and T. Wang, "An incentive-based protection and recovery strategy for secure big data in social networks," *Information Sciences*, vol. 508, pp. 79–91, 2020.

[7] D. V. Medhane, A. K. Sangaiah, M. S. Hossain, G. Muhammad, and J. Wang, "Blockchain-enabled distributed security framework for next generation IoT: an edge-cloud and software defined network integrated approach," *IEEE Internet of Things Journal*, vol. 7, pp. 1–8, 2020.

[8] Q. Liu, G. Wang, F. Li, S. Yang, and J. Wu, "Preserving privacy with probabilistic indistinguishability in weighted social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 5, pp. 1417–1429, 2017.

[9] S. Zhang, G. Wang, Q. Liu, and J. H. Abawajy, "A trajectory privacy-preserving scheme based on query exchange in mobile social networks," *Soft Computing*, vol. 22, no. 18, pp. 6121–6133, 2018.

[10] E. Luo, K. Guo, Y. Tang, X. Ying, and W. Huang, "Hidden the true identity and dating characteristics based on quick private matching in mobile social networks," *Future Generation Computer Systems*, vol. 109, pp. 633–641, 2020.

[11] J. Li, H. Yan, Z. Liu, X. Chen, X. Huang, and D. S. Wong, "Location-sharing systems with enhanced privacy in mobile online social networks," *IEEE Systems Journal*, vol. 11, no. 2, pp. 439–448, 2017.

[12] Z. Liu, D. Luo, L. Jin, X. Chen, and C. Jia, "N-mobishare: new privacypreserving location-sharing system for mobile online social networks," *International Journal of Computer Mathematics*, vol. 93, no. 2, p. 384400, 2014.

[13] C. Tang and C. Cai, "Verifiable mobile online social network privacy preserving location sharing scheme," *Concurrency & Computation Practice & Experience*, vol. 29, no. 1, pp. 1–10, 2017.

[14] E. Novak and Q. Li, "Near-pri: private, proximity based location sharing," in *Proceedings of the IEEE Infocom-IEEE Conference on Computer Communications*, pp. 31–45, Toronto, Canada, April 2014.

[15] A. Khoshgozaran and C. Shahabi, "Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy," in *Proceedings of the International Conference on Advances in Spatial & Temporal Databases*, Boston, MA, USA, July 2007.

[16] I. Kamel, A. M. Talha, and Z. A. Aghbari, "Dynamic spatial index for efficient query processing on the cloud," *Journal of Cloud Computing*, vol. 6, no. 1, pp. 1–16, 2017.

[17] P. Grubbs, K. Sekniqi, V. Bindschaedler, M. Naveed, and T. Ristenpart, "Leakage-abuse attacks against order-revealing encryption," in *Proceedings of the Security and Privacy (SP), 2017 IEEE Symposium on. IEEE*, pp. 655–672, San Jose, CA, USA, May 2017.

[18] D. S. Roche, D. Apon, S. G. Choi, and A. Yerukhimovich, "Pope: partial order preserving encoding," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM*, pp. 1131–1142, Vienna, Austri, October 2016.

[19] F. Kerschbaum, "Frequency-hiding order-preserving encryption," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM*, pp. 656–667, Denver, CO, USA, October 2015.

[20] F. Kerschbaum and A. Schropfer, "Optimal average-complexity idealsecurity order-preserving encryption," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM*, pp. 275–286, Scottsdale, AR, USA, November 2014.

[21] Q. Liu, Y. Tian, J. Wu, T. Peng, and G. Wang, "Enabling verifiable and dynamic ranked search over outsourced data," *Transactions on Services Computing*, 2019.

[22] S. Gang, Y. Xie, L. Dan, H. Yu, and V. Chang, "User-defined privacy location-sharing system in mobile online social networks," *Journal of Network & Computer Applications*, vol. 86, pp. 34–45, 2017.

[23] Y. Xu, Z. Quanrun, G. Wang, C. Zhang, J. Ren, and Y. Zhang, "An efficient privacy-enhanced attribute-based access control mechanism," *Concurrency and Computation: Practice and Experience*, vol. 32, pp. 1–10, 2020.

[24] Q. Liu, G. Wang, X. Liu, T. Peng, and J. Wu, "Achieving reliable and secure services in cloud computing environments," *Computers & Electrical Engineering*, vol. 59, pp. 153–164, 2017.

[25] T. Peng, Q. Liu, B. Hu, J. Liu, and J. Zhu, "Dynamic keyword search with hierarchical attributes in cloud computing," *IEEE Access*, vol. 6, 2018.

[26] A. K. Sangaiah, D. V. Medhane, G.-B. Bian, A. Ghoneim, M. Alrashoud, and M. S. Hossain, "Energy-aware green adversary model for cyber physical security in industrial system," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3322–3329, 2020.

[27] W. He, X. Liu, and M. Ren, "Location cheating: a security challenge to location-based social network services," in *Proceedings of the International Conference on Distributed Computing Systems*, San Jose, CA, USA, June 2011.

[28] T. Peng, Q. Liu, G. Wang, Y. Xiang, and S. Chen, "Multidimensional privacy preservation in location-based services," *Future Generation Computer Systems*, vol. 93, pp. 312–326, 2019.

[29] A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, "Order-preserving symmetric encryption," in *Proceedings of the 28th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Cologne, Germany, April 2009.

[30] A. Boldyreva, N. Chenette, and A. O'Neill, "Order-preserving encryption revisited: improved security analysis and alternative solutions," in *Advances in Cryptology-CRYPTO 2011*, pp. 578–595, Springer, Berlin, Germany, 2011.

[31] R. A. Popa, F. H. Li, and N. Zeldovich, "An ideal-security protocol for order-preserving encoding," in *Proceedings of the*

2013 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, May 2013.

[32] http://https://lbs.amap.com/.

[33] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services. ACM*, pp. 31–42, San Francisco, CA, USA, May 2003.

[34] X. Gong, X. Chen, K. Xing, D.-H. Shin, M. Zhang, and J. Zhang, "From social group utility maximization to personalized location privacy in mobile networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1703–1716, 2017.

[35] S. Zhang, G. Wang, M. Z. A. Bhuiyan, and Q. Liu, "A dual privacy preserving scheme in continuous location-based services," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 4191–4200, 2018.

[36] H. Wang, H. Huang, Y. Qin, Y. Wang, and M. Wu, "Efficient location privacy-preserving k-anonymity method based on the credible chain," *ISPRS International Journal of Geo-Information*, vol. 6, no. 6, p. 119, 2017.

[37] S. Zhang, X. Li, Z. Tan, T. Peng, and G. Wang, "A caching and spatial k-anonymity driven privacy enhancement scheme in continuous location based services," *Future Generation Computer Systems*, vol. 94, p. 4050, 2019.

[38] T. Peng, Q. Liu, and G. Wang, "Enhanced location privacy preserving scheme in location-based services," *IEEE Systems Journal*, vol. 11, no. 1, pp. 219–230, 2017.

[39] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving perfect location privacy in wireless devices using anonymization," *IEEE Transactions on Information Forensics & Security*, vol. 12, no. 11, Article ID 26832698, 2017.

[40] L. Huang, Y. Ma, Y. Liu, and A. K. Sangaiah, "Multi-modal bayesian embedding for point-of-interest recommendation on location-based cyber-physical-social networks," *Future Generation Computer Systems*, vol. 108, pp. 1–10, 2020.

[41] A. K. Sangaiah, D. V. Medhane, T. Han, M. S. Hossain, and G. Muhammad, "Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4189–4196, 2019.

[42] R. Shokri, G. Theodorakopoulos, and C. Troncoso, "Privacy games along location traces," *ACM Transactions on Privacy and Security*, vol. 19, no. 4, pp. 1–31, 2017.

[43] G. Sun, D. Liao, H. Li, H. Yu, and V. Chang, "L2P2: a location-label based approach for privacy preserving in LBS," *Future Generation Computer Systems*, vol. 74, pp. 375–384, 2017.

[44] T. Peng, Q. Liu, D. Meng, and G. Wang, "Collaborative trajectory privacy preserving scheme in location-based services," *Information Sciences*, vol. 387, pp. 165–179, 2017.

[45] H. To, G. Ghinita, L. Fan, and C. Shahabi, "Differentially private location protection for worker datasets in spatial crowdsourcing," *IEEE Transactions on Mobile Computing*, vol. 16, no. 4, pp. 934–949, 2016.

[46] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially private data publishing and analysis: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, 2017.

[47] T. Wang, Y. Mei, W. Jia, X. Zheng, G. Wang, and M. Xie, "Edge-based differential privacy computing for sensor-cloud systems," *Journal of Parallel and Distributed Computing*, vol. 136, pp. 75–85, 2020.

[48] X. Yi, R. Paulet, E. Bertino, and V. Varadharajan, "Practical approximate k nearest neighbor queries with location and query privacy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, pp. 1546–1559, 2016.

[49] I. Memon, Q. A. Arain, M. H. Memon, F. A. Mangi, and R. Akhtar, "Search me if you can: multiple mix zones with location privacy protection for mapping services," *International Journal of Communication Systems*, vol. 30, no. 16, pp. 1–23, 2017.

[50] S. Mascetti, C. Bettini, D. Freni, X. S. Wang, and S. Jajodia, "Privacy-aware proximity based services," in *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pp. 1–10, Taipei, Taiwan, May 2009.

[51] T. Jing, P. Lin, Y. Lu, C. Hu, and Y. Huo, "Fpodg: a flexible and private proximity testing based on "one degree" grid," *International Journal of Sensor Networks*, vol. 20, no. 3, pp. 199–207, 2016.

[52] X.-Y. Li and T. Jung, "Search me if you can: privacy-preserving location query service," in *2013 Proceedings of the IEEE INFOCOM*, pp. 2760–2768, Turin, Italy, April 2013.

Hindawi
Scientific Programming
Volume 2020, Article ID 8839720, 8 pages
https://doi.org/10.1155/2020/8839720

Research Article

Urban Traffic Signal Control Based on Multiobjective Joint Optimization

Yongrong Wu,[1] Yijie Zhou,[1] Yanming Feng,[1] Yutian Xiao,[2] Shaojie He,[2] Junsheng Zhou,[2] Tianhe Ren,[2] Jinling Chen,[2] Mingsong Chen ,[1] Jianbing Xiahou ,[2] and Fan Lin [2]

[1]*School of Information and Communication, Guilin University of Electronics Technology, Guilin 541004, China*
[2]*Xiamen University, Xiamen 361001, China*

Correspondence should be addressed to Mingsong Chen; cms@guet.edu.cn, Jianbing Xiahou; jbxiahou@xmu.edu.cn, and Fan Lin; iamafan@xmu.edu.cn

Received 7 April 2020; Revised 25 May 2020; Accepted 10 August 2020; Published 15 October 2020

Academic Editor: Chao Huang

This paper proposes two algorithms for signal timing optimization of single intersections, namely, microbial genetic algorithm and simulated annealing algorithm. The basis of the optimization of these two algorithms is the original timing scheme of the SCATS, and the optimized parameters are the average delay of vehicles and the capacity. Experiments verify that these two algorithms are, respectively, improved by 67.47% and 46.88%, based on the original timing scheme.

1. Introduction

In order to solve the current traffic problem more effectively, this paper proposes a single intersection signal control method. The traffic signal control of a single intersection is the smallest unit of the entire city traffic signal control system. It plays an important role in improving traffic order at intersections, reducing vehicle delays, and improving traffic capacity [1]. With the continuous deepening of the concept of sustainable development of urban transportation, the goal of traffic signal control is gradually changing from a single goal to multiple goals. Therefore, the multiobjective joint optimization of traffic signal control has attracted researchers' attention.

When optimizing traffic signal parameters, many factors need to be considered comprehensively, such as average delay, number of stops, and queue length [2]. Among them, the most commonly used criteria are delay and number of stops. These service quality measures are considered to be the most important factors for users to perceive the service level [3]. Webster established a steady-state random delay model for unsaturated traffic flow [4]. Heydecker and Wey used the average delay of motor vehicle flow at the intersection as the optimization target of signal control parameters [5, 6]. Robertson used a linear combination of optimization goals such as average delay and number of stops to establish an optimized model of signal control parameters [7, 8].

The multiobjective problem of traffic control has become a hotspot in current research, which has aroused extensive attention of researchers. Zeng et al. established a multiobjective optimization model of signal control parameters. The cell transmission model was used to simulate the operation law of motor vehicle flow at signal control intersections. The average delay and capacity were used as optimization goals [9]. Taking the average delay, average parking rate, and capacity as the optimization goals, Yao et al. established a multiobjective optimization model but did not give a specific solution algorithm [10]. Taking the average delay of motor vehicle flow, average parking rate, and pedestrian waiting time as optimization goals, Ma et al. established a multiobjective optimization model for the period duration of a single intersection with fixed period signal control [11]. Li and Lu took the average delay and average number of vehicle stops at intersections as the optimization goals, established a multiobjective optimization model of signal control parameters for a single

intersection, and used genetic algorithms for optimization [12]. Taking the average delay of motor vehicle flow, the average number of stops, and the total passing traffic as the optimization goals, Cao and Xu established a weighted combination optimization model of signal control parameters for single intersection with saturation constraints and solved the model using genetic algorithm [13].

Although there are many models for multiobjective joint optimization, there are no studies comparing the actual vehicle flow at the intersection with the average delay of vehicles and the capacity of the intersection as variables, compared with the original SCATS. Therefore, the optimization algorithm in this paper combines the actual intersection road conditions to analyze the night peak traffic information, which is an innovative point.

2. Signal Timing Method Based on Multiobjective Joint Optimization

The basic idea of signal distribution is to allocate time resources to traffic lights at intersections according to the traffic flow in each direction, so as to conduct traffic guidance. It is very important for signal synchronization [14]. In general, the best period should be the average period. If the period is too long to significantly improve the traffic capacity, it will lead to further delays of vehicles [15]. According to the traffic distribution, geometry, intersection type, area, and relevant conditions of each intersection, each intersection has a corresponding optimal time period, so the optimal time period changes in different time periods. Traffic function and control intersection signals are the best cycle, which can provide the best length for the whole vehicle [16].

In real life, urban intersections often adopt the timing scheme, which is simple and easy to operate, but it has many limitations. Through the transformation of the Webster formula, we can get accurate signal timing for real-time traffic and improve the efficiencies of road. This paper analyzes the traffic flow of the same intersection at the same time through two algorithms and then compares these two algorithms with the classic timing scheme of SCATS.

2.1. Notes on SCATS. The SCATS [17], researched and developed by the New South Wales Road Traffic Authority (RTA), is one of the few advanced urban signal traffic control systems in the world. When the system is selected through a real-time solution, the real-time selection of the signal period and green ratio is based on overall demand.

The corresponding green light time of the intersection shall be determined in accordance with the principle that the saturation degrees of all phases are equal or close to each other [18]. With the adjustment of the signal period, the green time of each phase also changes.

2.2. Basis of Algorithm. This part will be introduced separately from the signal phase of the intersection, the model of the intersection, and the formula for calculating the objective functions.

2.2.1. Objective Function Calculation Formula. TRRL method is also known as Webster method or F-B method. This signal timing method and corresponding calculation formula were first proposed by Corbusier and Webster in the 1990s [19]. The basic idea of this model is to take the vehicle delay time as the only evaluation index when the vehicle passes the intersection and to optimize the signal timing scheme through calculation. According to the theory of signal timing at the intersection [20], the two decisive factors are signal period and green time.

The formula for calculating the objective function PI is as follows:

$$\mathrm{PI} = \min\left(\alpha\frac{D}{D_0} - \beta\frac{B}{B_0}\right). \tag{1}$$

In (1), D is the vehicle delay at the intersection, B is the capacity of the intersection, and α and β are two evaluation indicators, respectively, representing the proportion of vehicle delay and capacity (they are considered equally important in this paper, both assigned the value 0.5).

The formulas for calculating the vehicle delay D and the capacity of the intersection B are as follows:

$$D = \frac{\sum D_i q_i}{\sum q_i}, \tag{2}$$

$$B = B_A + B_B + B_C + B_D. \tag{3}$$

In (2) and (3), D_i is the stop delay of each vehicle in phase i, q_i is the flow in phase i, and B_i is the capacity of the intersection in phase i.

The delay at the ith phase intersection is D_i, and the traffic capacity at the ith phase intersection is B_i. The calculation formulas for the two of them are as follows:

$$D_i = \frac{C(1-\lambda_i)^2}{2(1-y_i)^2} + \frac{y_i^2}{2q_i\lambda_i(\lambda_i - y_i)}, \tag{4}$$

$$B_i = k_i \times \frac{3600}{h_i} \times \lambda_i. \tag{5}$$

In (4) and (5), C is the cycle, λ_i is the green ratio in phase i, y_i is the traffic intensity in phase i, k_i is the number of lanes in phase i, and h_i is the lane saturation headway in phase i (the default value for this article is 2.5 s).

The formula for period C is as follows:

$$C = \sum g_i + \sum(l_i + R_i). \tag{6}$$

In (6), g_i is the decision variable and there are 4 decision variables in total (g_A, g_B, g_C, g_D), g_A is the effective green time of phase A, g_B is the effective green time of phase B, g_C is the effective green time of phase C, and g_D is the effective green time of phase D.

That is to say, $C = g_A + g_B + g_C + g_D +$ (delay_start + delay_redlight) × 4; in this paper, when the green light is on, delay-start is set to 3 seconds and delay-red light is set to 2 seconds.

λ_i is the green ratio in phase i, and the calculation formula is as follows:

$$\lambda_i = \frac{g_i}{C}. \tag{7}$$

In (7), y_i is the traffic intensity in phase i, and the calculation formula is as follows:

$$y_i = \frac{q_n h_i}{g_i k_i}. \tag{8}$$

In (8), q_i is the traffic flow in phase i.

The calculation of the average vehicle delay D_0 and the initial intersection capacity B_0 is similar to that of D_i and B_i, but there is a difference in calculating the period and the green ratio; C_0 and λ_{i0} use the calculation method of timing period, and the calculation formula is as follows:

$$C_0 = \frac{1.5L + 5}{1 - Y}, \tag{9}$$

$$\lambda_{i0} = \frac{(C_0 - L)y_i}{C_0 Y}. \tag{10}$$

In (9) and (10), L is the total delay time of four phases and Y is the sum of the four phases of y_i.

The above formula is the basis of optimizing traffic signals in this paper. The following are a detailed introduction of the process of each algorithm and the final results from different algorithm perspectives.

2.2.2. Signal Phase and Intersection Model. Before determining which signal phase is used at the intersection, it is important to fully understand the condition of the vehicle as it rushes from all directions to the intersection.

In this paper, the investigation and research intersection is a three-lane intersection; after taking full account of the actual traffic situation, a four-phase traffic signal is set up. Based on the road near schools, supermarkets, residential areas, vegetable markets, and many other crowded and traffic-intensive areas, we set the signal allocation time separate for the turn (left or right), so as to alleviate the traffic problems that may face the turning vehicles, reducing the traffic pressure at the intersection. The four-phase traffic signal is shown in Figure 1, and the urban intersection model is shown in Figure 2.

The numbers 1–12 in Figure 2 are the coil numbers, which are used to calculate the ratio of the actual traffic volume at the intersection to the saturated traffic volume. The directions of the arrows nos. 1–4 are the same as the four phase directions A–D in Figure 1.

2.2.3. Experimental Data Source. The data source of this article is the real data of one hour (17 : 28–18 : 30) of the night peak of a certain intersection. The set of data includes the effective green light duration of the four phases, the number of vehicles passing through the set green light cycle, and its coil number. Because of the large amount of recorded traffic information, this article randomly selects 20 lists as templates.

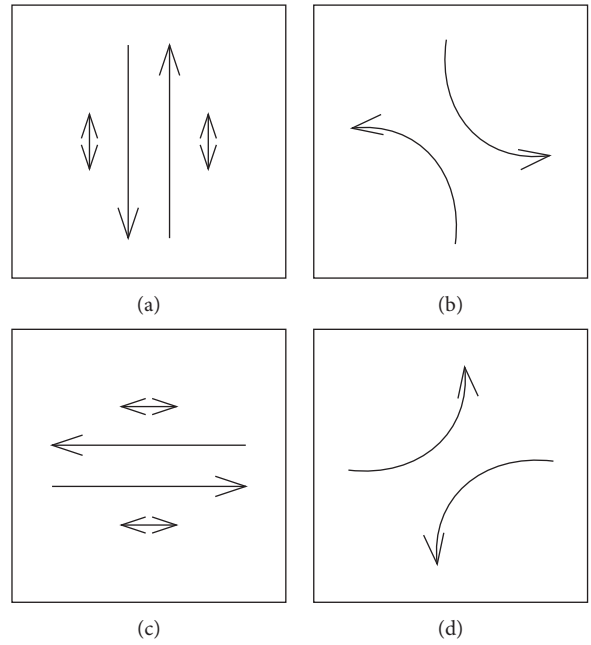


Figure 1: Phase of the traffic flow.

In the comparative experiment of this article, there are four sets of timing schemes for the SCATS of the intersection of the test subject. Taking the third set of plans as reference, the green light ratio information of the four sets of plans is shown in Table 1, and the data in Table 2 are records randomly extracted from traffic information.

2.3. Optimization of Traffic Signals Based on Microbial Genetic Algorithm

2.3.1. Experiment Procedure. Genetic algorithm is a global random search algorithm that simulates the evolutionary process and natural selection process of natural species. The algorithm was proposed by Professor J. Holland of Michigan University in 1975. This method was based on genetic theory and natural selection process [21] and used the natural selection rule and the cross mutation process of chromosomes and genes in biological groups as simulation object [22–24]. In this paper, the optimization process of microbial genetic algorithm is as follows:

(1) Initialize the values of 4 decision variables randomly (effective green light time for 4 phases).

(2) Binary coding is used to encode the individual values in this population to generate an initial population of about 300.

(3) The target value is $(\alpha D/D_0) - (\beta B/B_0)$; we hope the expected target value in this article is as small as possible; the ultimate goal is to get a value, so when the objective function can take the minimum value, what are the values of the corresponding 4 decision variables?

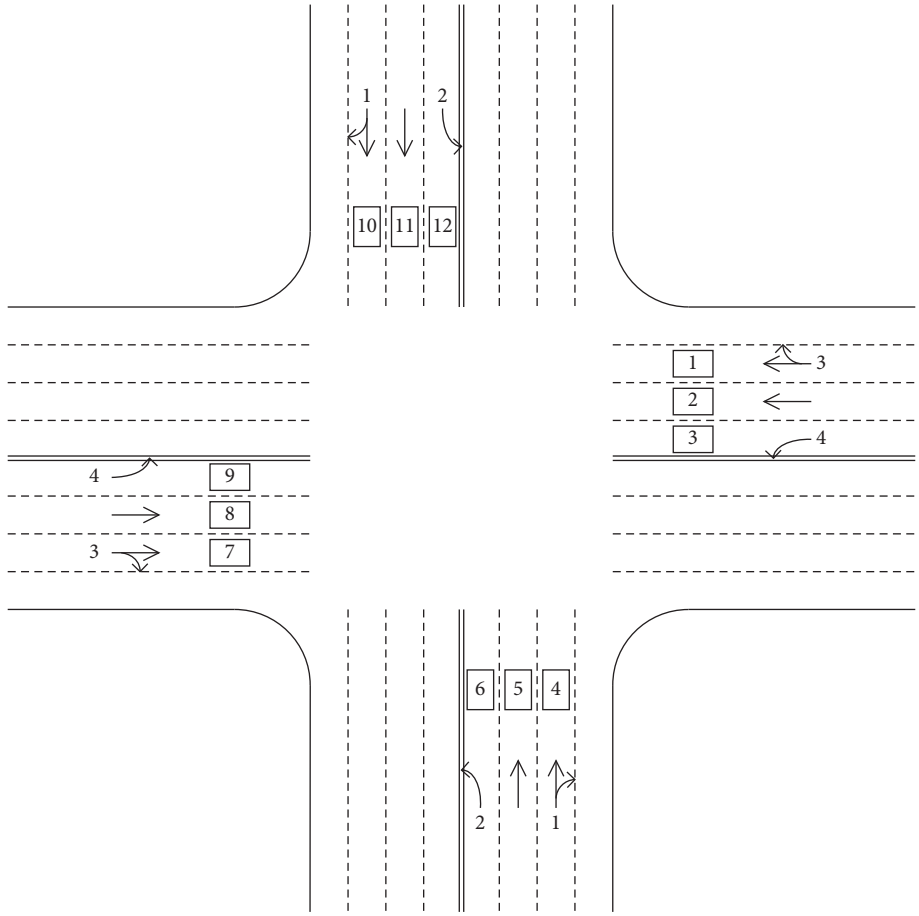


Figure 2: City intersection simulation map.

Table 1: Timing schemes for the SCATS.

Timing scheme	Phase *A* green light duration (%)	Phase *B* green light duration (%)	Phase *C* green light duration (%)	Phase *D* green light duration (%)
1	29	24	24	23
2	35	19	28	18
3	24	24	26	26
4	25	24	27	24

(4) Two individuals are randomly selected from the population pop, and the two individuals are calculated to obtain the objective function value; according to the objective value, the fitness of the two individuals can be calculated.

(5) According to their fitness, these two individuals are defined as winner and loser.

(6) For the winner, make sure it does not change; for the loser, make it a crossover; the object of the cross comes from the winner; put the loser back in the original pop.

(7) Through crossover, the loser has a certain chance to perform mutation operations, determine the position of chromosomal mutations, and perform corresponding mutations according to different coding strategies; then, a new generation of groups has emerged.

(8) Repeat steps (3)–(7) and update the 4 decision variables from generation to generation until the number of iterations reaches the original preset value.

2.3.2. Experimental Results. Time dependent objective function is shown in Figure 3.

In the figure, the abscissa represents the number of iterations, and the ordinate represents the value of the objective function at the end. The figure visualizes the entire optimization process of the microbial genetic algorithm. It is clear from the figure that the value of the corresponding objective function changes with the increase of iteration [25]. Before the objective function completes 24 iterations, the value of the objective function fluctuates with the accumulation of the iteration and the displayed state is unstable, but the value of the objective function tends to be stable, approaching 25 iterations [26]. Therefore, it is not

Table 2: Information statistics of traffic flow.

Phase	Coil number	Green light duration	Number of vehicles
C	1	36	6
C	2	36	7
D	3	24	3
C	7	36	2
C	8	36	2
D	9	24	2
A	4	44	0
A	5	44	0
A	10	44	4
A	11	44	7
C	1	32	3
C	2	32	3
B	6	26	0
C	7	32	0
C	8	32	2
B	12	26	0
A	4	51	0
A	5	51	0
A	10	51	6
A	11	51	10
⋮	⋮	⋮	⋮

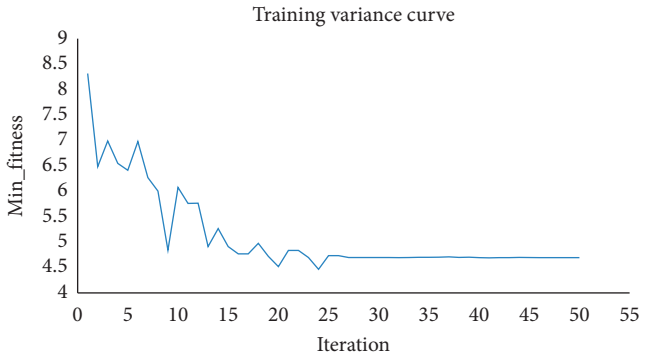


Figure 3: Objective function changes with the increase of iteration.

difficult to conclude from the relationship change diagram that, in terms of accuracy and training time, the microbial genetic algorithm performed better.

The traffic data is derived from the traffic condition information of an-hour-long evening rush in a certain domestic city. Recording the values of vehicle delay *D*, traffic capacity *B*, and objective function PI during this period and comparing them with the SCATS at the same time, we get the results in Table 3.

As the data in Table 3 show, we set the same external environmental factors, taking the same vehicle information in the same time period for comparison. The microbial genetic algorithm has improved by 67.465% in optimizing the objective function [27]. Obviously, the microbial genetic algorithm proposed in this paper is superior to SCATS in traffic signal optimization.

The SCATS is the basis of the algorithm optimization in this paper, with a period of 230 s. The time given in Table 4 is the effective green light time when the microbial genetic algorithm obtains the best solution.

Table 3: Comparison of performance parameters between SCATS and MGA.

	Vehicle delay (s)	Capacity (v/h)	PI
SCATS	4191.56723	3999.30131	14.56855
MGA	1430.77263	3189.67742	4.73989

Table 4: Green light duration of each phase obtained by GMA.

Phase number	Green light duration (s)
A	15.00
B	33.00
C	15.00
D	10.00

2.4. Signal Optimization Based on Simulated Annealing Algorithm

2.4.1. Algorithm Research History and Main Announcement. American physicist N. Metropolis and colleagues published an article in 1953 that studied complex systems and calculated the energy distribution in them. They used Monte Carlo simulation to calculate the energy distribution of molecules in multimolecular systems [28]. This is a term often referred to in the simulated annealing algorithm as the Metropolis criterion.

In 1983, physicists S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi of the IBM Corporation in the United States borrowed the method of Metropolis and others to explore a rotating glass state system [29]; they discovered the energy of the physical system and the cost of some optimal combination problems. The function is quite similar: seeking the lowest cost is like seeking the lowest energy. As a result, they developed an algorithm based on the metropolitan process to find the best solution. The optimal solution (the maximum value) of the problem is similar to the minimum energy of the system; when the system cools down, the energy also gradually decreases, and in the same sense, the solution to the problem falls down to its maximum. The calculation formula for the Metropolis criterion is as follows:

$$p = \begin{cases} 1, & \text{if } E(x_{\text{new}}) < E(x_{\text{old}}), \\ \exp\left(-\dfrac{E(x_{\text{new}}) - E(x_{\text{old}})}{\mathrm{T}}\right), & \text{if } E(x_{\text{new}}) \geq E(x_{\text{old}}), \end{cases} \tag{11}$$

$$p(dE) = \exp\left(\frac{dE}{KT}\right). \tag{12}$$

In (11) and (12), p is the probability of cooling, K is the constant, and exp is the natural index.

The cooling calculation formula of the simulated annealing algorithm is as follows:

$$T(t) = \frac{T_0}{1+t}. \tag{13}$$

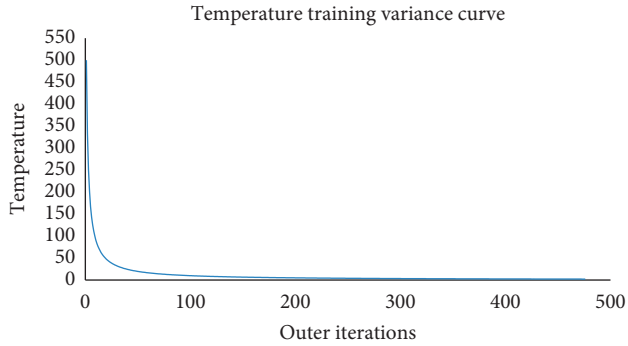


Figure 4: Relationship between the temperature and the number of outer iterations.

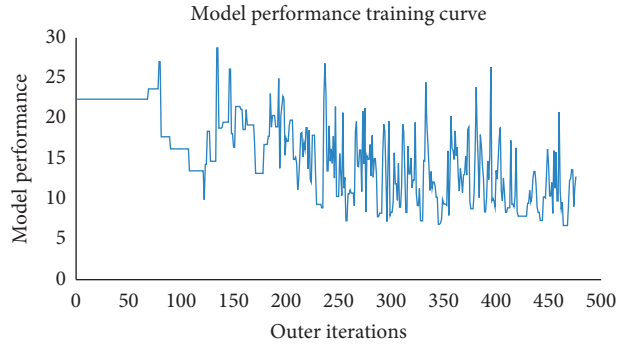


Figure 5: Relationship between the objective function and the number of outer iterations.

In (13), T_0 is the initial temperature.

In the process, the maximum tempering temperature gradually decreases with each annealing process, and the algorithm operation is stopped until the tempering end temperature is reached. The tempering process during the inner loop of the algorithm increases the algorithm's ability to accept nonoptimal solutions, which helps the algorithm to jump out of the local optimal solution and obtain the global optimal solution.

2.4.2. Experimental Results. In this simulation algorithm, the relationship between the temperature and the iterative algorithm in the driving process, and between the objective function and the external iterative algorithm is as follows.

Figure 4 shows the linear change in temperature as the number of iterations increases. Meanwhile, in another aspect, what is presented in Figure 5 is the change corresponding to the objective function when the number of iterations is accumulated. During the accumulation of the number of iterations, the temperature decreases accordingly. When the temperature reaches 0.4°C, the functional relationship is stable and convergent. In the graph of the relationship between the objective function and the number of outer iterations, the abscissa is each outer loop, and the ordinate is the objective function value. It can be seen that with the decrease of iteration depth and temperature, the value of the objective function gradually becomes smaller [30]. The optimization effect of the simulated annealing algorithm is getting better and better. Meanwhile, it can be seen from the figure that, at the 500th cycle, the SA model has approached convergence and the target value is around 7.7.

Similar to the microbial genetic algorithm, the values of vehicle delay D, traffic capacity B, and objective function PI are also recorded. The recorded data is shown in Table 5. Compared with the SCATS timing scheme, the objective function is improved by 46.848%.

By setting the same variables and taking the same dataset, the simulated annealing algorithm has been partially improved compared to the timing of the SCATS. The time given in Table 6 is the effective green light time when the simulated annealing algorithm obtains the best solution.

Table 5: Comparison of performance parameters between SCATS and SA.

	Vehicle delay (s)	Capacity (v/h)	PI
SCATS	4191.56723	3999.30131	14.56855
SA	2286.56378	3778.47053	7.74342

Table 6: Green light duration of each phase obtained by SA.

Phase number	Green light duration (s)
A	28.47
B	25.43
C	29.13
D	17.49

3. Conclusions

This paper uses two algorithms to optimize the signal timing scheme of intersections. The innovation of this paper lies in the use of mutation genetic algorithm and simulated annealing algorithm to optimize the timing scheme of SCATS. Through the real traffic flow data, the effectiveness of the two optimization algorithms in improving the objective function is demonstrated.

The advantage of the microbial genetic algorithm is that it will select the one with poor fitness from the two individuals as the mutation object; therefore, the quality of the optimal solution can be improved. At the same time, the microbial genetic algorithm is not highly dependent on the initial population, and two individuals can be randomly selected to calculate their objective function value and individual fitness. The global search ability of the algorithm is robust, and it will not fall into the trap of local optimal solution.

The advantage of the simulated annealing algorithm is that its entire operation process is very simple. The algorithm has strong robustness and applicability. Simultaneously, in order to ensure that the global optimal solution is not missed during the solution process, a higher initial temperature is set.

Both algorithms require a lot of iterations, so there are drawbacks in solving real-time problems. On the one hand, the signal optimization in this paper is only for a specific

intersection, which lacks general applicability. On the other hand, when selecting the objective function, other traffic factors are not fully considered, and there are certain errors here. Future research on signal control of intersections may consider more parameters, and the vehicle condition information of adjacent intersections will also affect that of the others. Incorporating more variables into the reference model, extending the research on individual intersections to intersection groups, and comparing the original SCATS timing scheme with actual traffic flow data can be the directions of hard research in the future.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by School of Information and Communication, Guilin University of Electronics Technology, "Research on Energy Efficiency Evaluation Mechanism of Next Generation Wireless Communication Service Transmission for Industry 4.0" (project number: NF150261), "High Precision Time-Frequency Transmission and Comparative Study" (project number: NF19002X) and CERNET Innovation Project (NGII20160410).

References

[1] L. I. U. Quan-Fu, B.-C. Lu, M. A. Qing-Lu, and J. Deng, "Traffic signal control multi-objective optimization at intersection," *Technology & Economy in Areas of Communications*, vol. 16, no. 1, pp. 47–50, 2014.

[2] Y.-F. Gao, H. Hu, H. Han, and X.-G. Yang, "Multi-objective optimization and simulation for urban road intersection group traffic signal control," *China Journal of Highway and Transport*, vol. 25, no. 6, pp. 129–135, 2012.

[3] A. Flannery, K. Wochinger, and A. Martin, "Driver assessment of service quality on urban streets," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1920, no. 1, pp. 25–31, 2005.

[4] F. V. Webster, *Traffic Signal Settings*, Road Research Laboratory, London, UK, 1958.

[5] B. G. Heydecker, "A decomposition approach for signal optimisation in road networks," *Transportation Research Part B: Methodological*, vol. 30, no. 2, pp. 99–114, 1996.

[6] W.-M. Wey, "Model formulation and solution algorithm of traffic signal control in an urban network," *Computers, Environment and Urban Systems*, vol. 24, no. 4, pp. 355–378, 2000.

[7] D. I. Robertson, *TRANSYT: A Traffic Network Study Tool*, Road Research Laboratory, Crowthorne, UK, 1969.

[8] S. C. Wong, W. T. Wong, C. M. Leung, and C. O. Tong, "Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control," *Transportation Research Part B: Methodological*, vol. 36, no. 4, pp. 291–312, 2002.

[9] J.-Q. Zeng, J.-J. Wang, K. Liu et al., "CTMMOGA based crossroad traffic signal control," *Journal of University of Science and Technology of China*, vol. 35, no. 2, pp. 284–290, 2005.

[10] X.-S. Yao, X. Luo, and D. U. Jin-You, "Traffic signal control based on satisfactory multi-optimization," *Computer Engineering and Applications*, vol. 36, no. 35, pp. 9-10, 2006.

[11] Y.-Y. Ma, X.-G. Yang, and Y. Zeng, "Multi objective cycle length optimization model and solution," *Journal of Tongji University: Natural Science*, vol. 37, no. 6, pp. 761–765, 2009.

[12] R.-M. Li and H.-P. Lu, "Traffic signal control multi-object optimization based on genetic algorithm," *Journal of Chang'an University: Natural Science Edition*, vol. 29, no. 3, pp. 85–88, 2009.

[13] C.-T. Cao and J.-M. Xu, "Multi-object traffic signal control method for single intersection," *Computer Engineering and Applications*, vol. 46, no. 16, pp. 20–22, 2010.

[14] K.-H. Gao, *Reach on Bus Priority Signal Coordination Control and Simulation in Urban Road Intersections*, Beijing Jiaotong University, Beijing, China, 2010.

[15] G. P. Liu, J. F. Whidborne, and J. B. Yang, *Multi-objective Optimisation and Control*, Research Studies Press Ltd., London, UK, 2003.

[16] L. F. C. Gonzalez, *Discharge Headway at Signalized Intersections in Monterrey*, Doetoral Dissertation of University of Texas, USA, Nuevo Leon, Mexico, 2006.

[17] M. Gentili and P. B. Mirchandani, "Locating sensors on traffic networks: models, challenges and research opportunities," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 227–255, 2012.

[18] A. Stevanovic, *Adaptive Traffic Control Systems: Domestic and Foreign State of Practice*, Transportation Research Board, Washington, DC, USA, 2010.

[19] F. V. Webster, "Traffic signal settings," Department of Scientific and Industrial Research, London, UK, Road Research Technical Paper No 39, 1958.

[20] J. Lu, S. Dissanayake, and N. Castillo, *Safety Evaluation of Right Turns Followed by U-Turns as an Alternative to Direct Left Turns Conflict Analysis*, Research Report, Department of Civil & Environmental Engineering, University of South Florida Institution, Tampa, FL, USA, 2001.

[21] H. Shen, "A study of welding robot path planning application based on genetic ant colony hybrid algorithm," in *Proceedings of the 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Xi'an, China, October 2016.

[22] O. Kramer, *Genetic Algorithm Essentials. Studies in Computational Intelligence*, Springer, Berlin, Germany, 2017.

[23] D. Hermawanto, "Genetic algorithm for solving simple mathematical equality problem," *Computer Science*, 2017.

[24] X. Li and L. Gao, "An effective hybrid genetic algorithm and tabu search for flexible job shop scheduling problem," *International Journal of Production Economics*, vol. 174, pp. 93–110, 2016.

[25] Y. Hou, "Research and implementation of hybrid clustering algorithm in big data processing," in *Proceedings of the 2018 International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC 2018)*, Chengdu, China, December 2018.

[26] Y. Meng, J. Zhang, and J. Cai, "Improved binary decision diagram based accelerated circuit evolutionary algorithm," in *Proceedings of the 2015 Joint International Mechanical, Electronic and Information Technology Conference*, Chongqing, China, December 2015.

[27] Q. Sun, T. Zhou, and H. Li, "A novel register allocation algorithm for testability," *Tsinghua Science & Technology*, vol. 12, no. S1, pp. 57–60, 2007.

[28] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 809–821, 2016.

[29] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[30] Z.-Q. Wang, N.-J. Zhou, J.-Q. Zhang, J. Guo, and X.-Y. Wang, "Parametric optimization and performance comparison of organic Rankine cycle with simulated annealing algorithm," *Journal of Central South University*, vol. 19, no. 9, pp. 2584–2590, 2012.

Hindawi
Scientific Programming
Volume 2020, Article ID 8814222, 12 pages
https://doi.org/10.1155/2020/8814222

Research Article

Short-Term Infectious Diarrhea Prediction Using Weather and Search Data in Xiamen, China

Zhijin Wang,[1] **Yaohui Huang**,[2] **Bingyan He**,[1] **Ting Luo**,[2] **Yongming Wang**,[3] **and Yonggang Fu**[1]

[1]*Computer Engineering College, Jimei University, Xiamen 361021, China*
[2]*Chengyi University College, Jimei University, Xiamen 361021, China*
[3]*China Electronics Technology Group Corporation, Shanghai 200001, China*

Correspondence should be addressed to Yonggang Fu; yonggangfu@jmu.edu.cn

Received 23 April 2020; Revised 13 August 2020; Accepted 27 August 2020; Published 30 September 2020

Academic Editor: Chenxi Huang

Infectious diarrhea has high morbidity and mortality around the world. For this reason, diarrhea prediction has emerged as an important problem to prevent and control outbreaks. Numerous studies have built disease prediction models using large-scale data. However, these methods perform poorly on diarrhea data. To address this issue, this paper proposes a parsimonious model (PM), which takes historical outpatient visit counts, meteorological factors (MFs) and Baidu search indices (BSIs) as inputs to perform prediction. An experimental evaluation was done to compare the short-term prediction performance of ten algorithms for four groups of inputs, using data collected in Xiamen, China. Results show that the proposed method is effective in improving the prediction accuracy.

1. Introduction

To keep up with the pace of income growth, urbanization, and globalization, risk management of infectious diseases in public has become a critical task [1]. Infectious diarrhea (ID) [2] is one of the most common infectious diseases in the world, which infects more than 1 billion persons. It became the 37th legally notifiable disease in the China's National Notifiable Disease Reporting and Surveillance System, and any new case must be reported within 24 hours of diagnosis [3].

Early warning techniques [4] have been developed to monitor the status of infectious diseases and the demand for healthcare and health services. These techniques can support decision making for medical intervention strategies [5], by preinforming people, health service providers, and the government.

The problem of predicting upcoming diarrhea outpatient visits can be viewed as time series prediction problem. In past decades, numerous studies have used autoregression (AR), autoregressive integrated moving average model (ARIMA), and machine learning methods to predict upcoming values based on past observations. The widely used machine learning methods are multiple linear regression (MLR), support vector regression (SVR), and random forest regression (RFR) [6–8]. A famous autoregression method is that of Box and Jenkins [9], which has been applied in many fields [10], such as for electricity load forecasting and stock price prediction. Another famous statistical method is spline interpolation [11], which learns and uses a cubic spline interpolation to predict future values. But the performance of these methods degrades when dealing with nonstationary and chaotic time series, such as those of diarrhea outpatients.

Recently, to alleviate the uncertainty of a time series, exogenous data have been collected and fused into machine learning methods to achieve better predictions [12–14]. Hereto, the methods using exogenous data are called NARX [15]. According to the structure of NARX, we categorize them into *wide models* and *deep models*.

A wide model commonly builds more than two components in a layer. To capture the sequential features of a

time series, a recurrent neural network (RNN) [16] is adopted to process exogenous data. To discriminatively process exogenous data and historical observations, the encoder-decoder framework [17] is introduced to undertake predictions. Meanwhile, the gated recursive unit (GRU) [18] is used to replace RNN in the framework, which captures long- and short-term memory (LSTM) [19]. Moreover, an attention mechanism is designed to adjust the values of exogenous inputs and historical inputs in the stage of encoding and decoding, respectively [20–22].

The deep model consecutively connects components from inputs to outputs using more than two neural network layers. For example, DL4Epi in [23] consists of a CNN layer, a RNN layer, and a residual link layer. DilatedRNN in [24] dilated three RNN layers from input to predict glucose incidence rate. TPA-LSTM in [25] uses a temporal pattern attention layer to deepen a model as well.

However, not only wide models but also deep models suffer from the problem of needing numerous samples. Given a time series of weekly or monthly diarrhea outpatients, the number of training samples is usually in hundreds. If a model has lots of parameters and the training samples are few, the learned model has a poor generalization performance. In the following, we call a model having many parameters a *heavy model*. In reality, heavy models remember almost the training samples and thus perform poorly for predicting unknown inputs.

How to build a parsimonious model, which has few parameters and learns from hundreds of samples? Observations about from previous models reveal that models assign weights to input elements, and then weights are passed and transformed into the target. Can the weights of elements be reduced? Can the reduced weights be passed to the output?

To address the issues, we propose a parsimonious model (PM). The proposed model first assigns a vector to each input dimension, and then the vectors are connected to the target.

The rest of this paper is organized as follows. Section 2 briefly introduces the study area and data sources. Section 3 illustrates the defined problem and presents the proposed method. Section 4 describes the experimental setting and related benchmarks. Section 5 compares different methods and analyses their sensitiveness. Finally, a conclusion is drawn in Section 6.

2. Study Materials

This section first introduces the study area, i.e., Xiamen city. Then, it briefly describes meteorological factors and search behavior in this city using basic descriptive statistics.

The basic statistical variables of inputs and the target are listed in Table 1.

The reasons for choosing meteorological factor (MF) data and Baidu search index (BSI) data as exogenous data are that those data are significantly correlated to the infectious diseases. MF data have been proved to be associated with the incidence of infectious diseases [26, 27], such as ID and "hand, foot, and mouth disease" (HFMD). As an important entrance for users to obtain digital information resources on the Internet, the search engine provides lots of useful information. The search index of specific keywords directly reflects the social attention on infectious diseases [28–30]. Hence, MF data and BSI data are considered as exogenous inputs to models.

In order to determine the relationships between exogenous variables and the target variable, the Person correlation coefficient (PCC) values are figured out in the last column of Table 1. All of the exogenous factors are significantly correlated to weekly outpaitent counts except weekly maximum temperature. $H_{\max}$ is used as an important input, since humidity is considered as an important factor [27].

2.1. Study Areas. Xiamen is a developing city, which has a population of around 4 million permanent residents and around 4 million temporary residents as of 2018. It is an important special economic zone in China since 1980 and is located in southeast China. It covered a land area of 1,699.39 km^2 and a sea area of over 390 km^2 in 2017. Xiamen has a monsoonal humid subtropical climate, characterised by long, hot, and humid summers (but moderate compared to much of the rest of the province) and short, mild, and dry winters, and the annual mean temperature is 20.7°C [31].

Herein, the detailed population from 2012 to 2016 is listed in Table 2 as a background for this study. The annual growth of its population is between 1.3% and 2.1%, which is relatively stable. Morbidity and mortality are influenced by disease trends within a period. Meanwhile, infectious diarrhea is the most common infectious disease. Hence, the number of infectious diarrhea outpatient visits is key to monitor the status of infectious diseases in a city.

2.2. Meteorological Factor (MF). The MF data from January 1, 2012, to December 31, 2016, were collected from Weather Underground (http://www.wunderground.com), which is a popular personal weather website. Those data from meteorological monitoring sites provide globally comprehensive, timely, and reliable meteorological data. At the moment, this website provides web-based application program interfaces for users to download data or develop third-party applications.

The provided weather information is formatted in days. We calculate 261 weekly weather information pairs by those daily records. The selected weather information in weeks consists of 12 factors, e.g., highest temperature (°C), average dew point (°C), lowest atmospheric pressure (hPa), and average relative humidity (%). Thus, we get 12 groups of exogenous data to describe weather conditions.

2.3. Baidu Search Index (BSI). Baidu is the most widely used search engine (http://index.baidu.com) in China, and it is also the largest Chinese search platform. People in Xiamen, China, are accustomed to using that search engine as well. Baidu search index records the search volume of many terms, which are queried by users since January 2011, and it is publicly available. BSI has been used to monitor the

Table 1: Basic statistical characteristics of exogenous time series and infectious outpatient visits (261 samples). p 0.01

Symbols	Parameter (unit)	Range		Mean	Median	STD	PCC
		Min	Max				
$T_{\max}$	Maximum temperature (°C)	14	39	29.21	31	5.64	−0.342**
$T_{\min}$	Minimum temperature (°C)	1	27	16.9	18	6.54	−0.304**
T_{avg}	Average temperature (°C)	9.29	31.29	22.32	23.14	6.03	−0.317**
$D_{\max}$	Maximum dew temperature (°C)	9	29	22.32	23	4.94	−0.385**
$D_{\min}$	Minimum dew temperature (°C)	−12	24	10.57	11	9.32	−0.345**
D_{avg}	Average dew temperature (°C)	3.57	26.71	17.47	18.14	6.54	−0.393**
$H_{\max}$	Maximum relative humidity (%)	69	100	98.33	100	4.25	−0.092
$H_{\min}$	Minimum relative humidity (%)	9	70	39.78	42	13.59	−0.372**
H_{avg}	Average relative humidity (%)	52.14	94.71	75.18	75.43	9.5	−0.339**
$A_{\max}$	Maximum atmospheric pressure (hPa)	1004	1037	1018.84	1019	7.12	0.381**
$A_{\min}$	Minimum atmospheric pressure (hPa)	964	1021	1007.3	1008	7.46	0.392**
A_{avg}	Average atmospheric pressure (hPa)	998.1	1026.6	1013.56	1013.7	6.63	0.342**
$B_{p,c}$	BSI of "腹泻" in PC, China (count)	1953	6000	3804.34	3584	998.43	−0.181**
$B_{m,c}$	BSI of "腹泻" in mobile, China (count)	5764	9121	6415.20	6328	447.32	−0.241**
$B_{t,c}$	BSI of "腹泻" in total, China (count)	7971	12500	10219.53	10046	978.45	−0.325**
$B_{p,x}$	BSI of "腹泻" in PC, Xiamen (count)	0	440	311.03	303	85.07	−0.323**
$B_{m,x}$	BSI of "腹泻" in mobile, Xiamen (count)	0	515	307.31	430	205.63	−0.348**
$B_{t,x}$	BSI of "腹泻" in total, Xiamen (count)	0	925	618.34	683	194.09	−0.324**
C_d	Infectious diarrhea outpatient visits (case)	15	349	69.7	52	50.03	—

"STD" denotes standard deviation. ** indicates that p value is less than 0.01, which means significant correlation.

Table 2: Summary of the population in Xiamen [32].

Year	Population (million)	Population growth rate (%)
2012	3.67	
2013	3.73	1.6
2014	3.81	2.1
2015	3.86	1.3
2016	3.92	1.6

incidence of influenza epidemics [33], analyse regional infectious diseases, and perform real-time monitoring and prevention [28].

A major concern to acquire query data (search indices) is to find proper query words [29], which reflects user behavior about searching for infectious diarrhea. The engine returns daily counts of a given keyword, which can be conditioned by a region and a platform. Six time series of each query word are downloaded by choosing a region from {Xiamen, countrywide} and a platform from {mobile, PC, total}, respectively. Those data pairs are calculated in weeks. We have six groups of search indices, and they are used as exogenous data for each keyword. A correlation coefficient analysis is carried out on those input indices and target output (i.e., the number of cases). We found that the indices of the keyword "腹泻" (i.e., ID) have significant correlation, while others do not. Hence, we adopt the six group indices as inputs of models.

3. Approach

This section first defines the problem of time series prediction using exogenous data and then introduces data preprocessing and postprocessing. Finally, the proposed method is presented.

3.1. Problem Formulation. The epidemic prediction problem can be viewed as a time series forecasting task [10]. A time series is defined as a sequence of observation values with consecutive identical interval lengths.

Let $y_t \in \mathbb{R}^1$ denote the observation measured at time t and $\mathbf{x}_t \in \mathbb{R}^n$ denote the exogenous data measured at time t, where n is the number of input series. Furthermore, let T be a time window size. The known exogenous series with window size T is symbolized as $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, and the historical target observations are denoted as $[y_1, y_2, \ldots, y_T]$.

The goal is to predict the value of a future time point $\hat{y}_{T+1}$, given historical observations and exogenous series. In detail, a nonlinear mapping $F(\cdot)$ is applied into the predictive formula:

$$\hat{y}_{T+1} = F(y_1, y_2, \ldots, y_T, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T). \tag{1}$$

Moreover, let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ denote the available training data in a time-span of size T and $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ be the recent historical outpatient visit counts.

3.2. One-Step Forward Split and Normalization

3.2.1. One-Step Forward Split. A time series cannot be directly used as input of regression models. Hence, we conduct one-step forward split to transform a time series to supervised data. The formula of the one-step split is shown as follows:

$$\begin{bmatrix} y_1 & y_2 & \cdots & y_T & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_T \\ y_2 & y_3 & \cdots & y_{T+1} & \mathbf{x}_2 & \mathbf{x}_3 & \cdots & \mathbf{x}_{T+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{M-T-1} & y_{M-T} & \cdots & y_{M-1} & \mathbf{x}_{M-T-1} & \mathbf{x}_{M-T} & \cdots & \mathbf{x}_{M-1} \end{bmatrix} \longrightarrow \begin{bmatrix} y_{T+1} \\ y_{T+2} \\ \vdots \\ y_M \end{bmatrix}, \tag{2}$$

where M is the length of inputted time series and T is the window size.

3.2.2. Normalization. It is a process of rescaling the data from the original range so that all values are within a given range [34], usually within 0 and 1 or 0.05 and 0.95. Both min-max normalization and standard (i.e., z-score) normalization methods are commonly applied to normalize time series data.

Min-max normalization scales data in the [0, 1] interval by using the bounds of the minimum and maximum values. Standardizing a dataset involves rescaling the distribution of values, so that the mean of observed values is 0 and the standard deviation is 1. The mean and standard deviation estimates of a dataset can be more robust to new data than the minimum and maximum.

We choose standard normalization to rescale inputs and the target, since we regard the outpatient visit counts as Gaussian distribution, and the standardization fits a Gaussian distribution well. The standard normalization of inputs is formulated and recovered as

$$\begin{aligned} \mathbf{d}' &= \frac{\mathbf{d} - \mu(\mathbf{d})}{\sigma(\mathbf{d})}, \\ \mathbf{d} &= \mathbf{d}' \cdot \sigma(\mathbf{d}) + \mu(\mathbf{d}), \end{aligned} \tag{3}$$

where $\mathbf{d} \in \mathrm{R}^M$ denotes a feature of observed samples, M is the number of observed samples, $\mathbf{d}'$ is the normalized data, $\mu(\mathbf{d})$ is the mean value of $\mathbf{d}$, and $\sigma(\mathbf{d})$ is the standard variation of $\mathbf{d}$. The recovered formula is applied on model outputs in the postprocessing stage. Because normalization is only applied on exogenous data, the recovery action can be skipped while generating predictions.

3.3. Parsimony Model (PM). The diagram of the proposed PM is shown in Figure 1. This diagram consists of three stages: data preprocessing, PM, and postprocessing. In the preprocessing stage, the exogenous time series is normalized and combined with target time series to split the supervised data using one-step forward. In the PM processing stage, a parameter layer is exploited to extract pattern features from heterogeneous inputs. The outputs of the model are generated by linear summarization of the intermediate state result.

The left bottom part of Figure 1 gives toy examples of the structure of the parsimonious method with respect to inputs and weights. The yellow solid circle denotes exogenous inputs, such as meteorological factor or Baidu search index. The blue solid rectangle denotes historical observed outpatient visits, and red solid rectangle is the target. The yellow solid rectangles are weights to be learned. The goal is to reduce these complex models and find a parsimonious model to learn and predict well on the diarrhea data collections.

Therefore, the weights of the first layer in a neural network are exploited to reconsider model structures. The left bottom part of Figure 1 gives inputs of exogenous data (i.e., MF, BSI, or MF + BSI) and historical cases, and the right part indicates the weights and connections among them (arrows between circles and rectangles). The two following definitions are introduced to provide clearer explanations.

Definition 1. A weight is assigned to each input element.

Definition 1 is formulated as follows:

$$F_1(\mathbf{W}, [\mathbf{X}; \mathbf{y}]) \longrightarrow \widehat{y}_{T+1}, \tag{4}$$

where $[\mathbf{X}; \mathbf{y}] \in \mathbb{R}^{(n+1)\times T}$ is the input of a given sample (i.e., input matrix), $\widehat{y}_{T+1}$ is the target of this sample, which means the outpatient visit counts of the upcoming week, $\mathbf{W} \in \mathbb{R}^{(n+1)\times T}$ is the weight corresponding to the inputs (i.e., weight matrix), and F_1 is a linear or nonlinear mapping from input and weight matrices to target.

Both wide models and deep models in the research field of epidemic prediction are based on this definition. A wide model processes $\mathbf{X}$ and $\mathbf{y}$ in different components. For example, the encoder-decoder structure encodes $\mathbf{X}$, passes a code to the decoder, and the decoder processes the code and y to make predictions. A deep model adds layers on the sequence of $\mathbf{W}$, such as CNN and RNN.

When learning with a small number of samples, the number of samples is usually in hundreds, and these models have a burden of parameter training. In reality, the number of parameters is greater than the number of samples times the number of exogenous factors. Hence, these trained models would remember all the training samples and have a value of loss close to 0, but they poorly predict upcoming events.

Therefore, the wide model and the deep model are condensed into a light neural network by setting F_1 as a linear function or nonlinear function, such as sigmoid function. But the condensed light neural network has many parameters as well. Thus, we try to use two vectors to restore weight $\mathbf{W}$ in equation (4). The aim is not only to have fewer parameters during the training stage but also to obtain better performance at the predicting stage.

Definition 2. A weight is assigned to each input dimension.

Definition 2 is formulated as follows:

$$F_2(\mathbf{w}_r, \mathbf{w}_c, [\mathbf{X}; \mathbf{y}]) \longrightarrow \widehat{y}_{T+1}, \tag{5}$$

where $[\mathbf{X}; \mathbf{y}] \in \mathbb{R}^{(n+1)\times T}$ is the input matrix, and $\widehat{y}_{T+1}$ is the prediction target (i.e., the upcoming week outpatient counts), $\mathbf{w}_r \in \mathbb{R}^{n+1}$ and $\mathbf{w}_c \in \mathbb{R}^T$ are the weights corresponding to two input dimensions (i.e., weight vectors), and F_2 is a linear function or a nonlinear mapping from input and weight matrices to the target.

There are limitations in accurately capturing the input matrix of a sample solely on the basis of using two vectors, whose lengths are the same as the size of input matrix. Hence, we exploit two ways to recover $\mathbf{W}$ in equation (4) using $\mathbf{w}_r$ and $\mathbf{w}_c$: the addition method and the multiplication method.

For each element w_{ij} in $\mathbf{W}$, it can be recovered using the addition method. The addition recovery formula is

$$\widehat{w}_{ij} = w_{ri} + w_{cj}, \tag{6}$$

where $\widehat{w}_{ij}$ is the recovered value, $w_{ri} \in \mathbf{w}_r$ denotes the i-th element of $\mathbf{w}_r$, and $w_{cj} \in \mathbf{w}_c$ denotes j-th element of vector

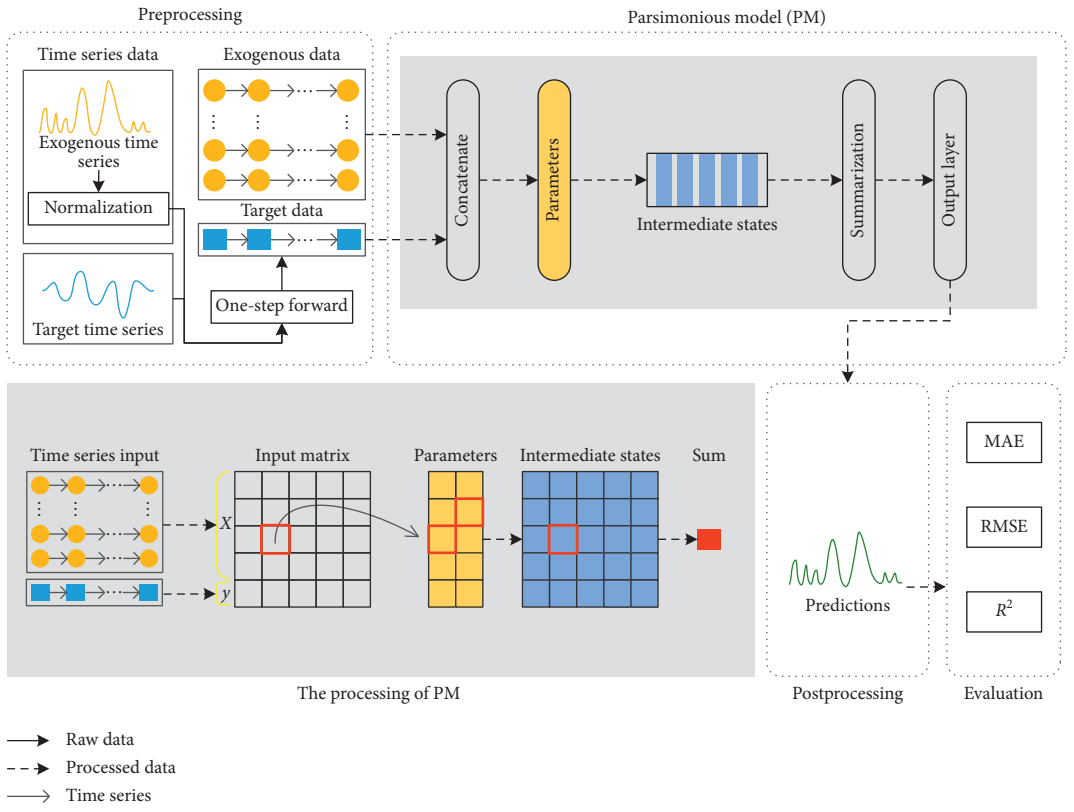


Figure 1: The diagram of the proposed PM.

$\mathbf{w}_c$, respectively. Equation (6) shows that a matrix can be recovered by two vectors, which would reduce the training parameters.

Another recovery method is the multiplication method. For each element w_{ij} in $\mathbf{W}$, it can be recovered using the multiplication method. The multiplication recovery formula is

$$\widehat{w}_{ij} = w_{ri} \cdot w_{cj}, \tag{7}$$

where $\widehat{w}_{ij}$ is the recovered value, $w_{ri} \in \mathbf{w}_r$ denotes the i-th element of $\mathbf{w}_r$, and $w_{cj} \in \mathbf{w}_c$ denotes j-th element of vector $\mathbf{w}_c$, respectively. Equation (7) shows that a matrix can be recovered by two vectors, which would reduce the training parameters as well.

To benefit the recovery effects from the both methods, we add them together and get

$$\widehat{w}_{ij} = \left(w_{ri} + w_{cj}\right) + w_{ri} \cdot w_{cj}. \tag{8}$$

According to equations (8) and (5), we have

$$F_2\left(\widehat{\mathbf{W}}, [\mathbf{X}; \mathbf{y}]\right) \longrightarrow \widehat{y}_{T+1}, \tag{9}$$

where $\widehat{w}_{ij} \in \widehat{\mathbf{W}}$. Lots of methods applied to equation (5) can be used to equation (6) as well, whereas the goal of this paper is trying to find a parsimony method to fit well with small-scale datasets. Hence, F_2 in equation (9) should be as simple as possible.

We adopt the summarization and sigmoid activate function to replace F_2 and make it be simple. The prediction function is formulated as

$$\widehat{y}_{T+1} \leftarrow \sum_{j=1}^{T} \sum_{i=1}^{n+1} \left(\left(w_{ri} + w_{cj}\right) + w_{ri} w_{cj}\right) s_{ij}, \tag{10}$$

where $s_{ij} \in [\mathbf{X}; \mathbf{y}]$ denotes the element in the i-th row and j-th column of input matrix $[\mathbf{X}; \mathbf{y}]$. Equation (10) is called the prediction function of the proposed PM.

4. Experimental Settings

This section first describes the exogenous feature selection, experimental settings, and relevant related work, which are compared with our proposed method. Finally, an analysis of the comparable results is provided.

4.1. Experimental Configuration. The disease data collection is divided into two subsets: the first part, from the 1st week of 2012 to the 52nd week of 2015, is used to build and train model; the remaining part, from the 1st to the 52nd week of 2016, is utilized to assess the learned models.

All neural models are trained using the Adam optimizer [35]. The batch size is set to 32. Their learning rate is set to 0.001, and mean squared error (MSE) is chosen as the loss function. For RNN and LSTM, the number of hidden neurons is set to {64, 128}.

We ran each method five times and reported the median. The optimal RMSE values of models at given T and input data are in bold types. DA-RNN [20] and DL4Epi [23] have been applied on the four groups of inputs, and they are not converged in training stage. Hence, their results are not listed and compared.

4.2. Evaluation Metrics. A number of performance evaluation criterion have been employed to evaluate and compare the performance of models, but there is a no uniform standard. Therefore, we evaluate models based on error metrics, which are commonly adopted in regression performance evaluation.

The evaluation criterion consists of the mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of determination (R^2). These metrics are expressed as the following mathematical expressions:

$$\begin{aligned} \text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}\left(\left|y_{T+1}^{i}-\widehat{y}_{T+1}^{i}\right|\right), \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{T+1}^{i}-\widehat{y}_{T+1}^{i}\right)^{2}}, \\ R^2 &= 1-\frac{\sum_{i=1}^{N}\left(y_{T+1}^{i}-\widehat{y}_{T+1}^{i}\right)^{2}}{\sum_{i=1}^{N}y_{i}^{2}}, \end{aligned} \tag{11}$$

where y_{T+1}^{i} is an actual value at week (relative time) i in the test set, $\widehat{y}_{T+1}^{i}$ is a predicted value at week (relative time) i, $\overline{y}_{T+1}$ is the mean value in the test set, and N is the number of weeks in the test period. The model providing the smallest MAE and RMSE and the largest R^2 is considered to have the best performance.

4.3. Comparable Methods. According to the type of input variables, the prediction methods are categorized into univariate models and multivariate models. We choose models which can work with multiple inputs and unstable series and keep ARIMA as a baseline.

4.3.1. Multiple Linear Regression (MLR). It is widely adopted to model the connection between several independent variables and dependent variables. The prediction model of MLR based on generalized MLR analysis formula is given below:

$$\begin{aligned} \widehat{y}_{T+1} &= \mathbf{w}_{y}^{\prime}\cdot\mathbf{y}+b+\varepsilon \\ &= \sum_{i=1}^{T}w_{y,i}\cdot y_{i}+b+\varepsilon, \end{aligned} \tag{12}$$

where $\mathbf{w}_{\mathbf{y}}\in\mathbb{R}^{T}$ and b are weights and bias to be learned, T is the size of the historical data, ε represents the random error term, and $E(\varepsilon)=0$.

We extend MLR to model exogenous inputs and target as follows:

$$\widehat{y}_{T+1}=\mathbf{w}_{xy}^{\prime}\cdot\left[x_1;x_2;\ldots;x_T;y\right]+b+\varepsilon, \tag{13}$$

where $\mathbf{w}_{xy}^{\prime}\in\mathbb{R}^{(n+1)T\times 1}$ denotes weights of exogenous inputs and historical data at past T time intervals.

4.3.2. Random Forest Regression (RFR). It is a CART [36] method. It has been used to predict the death characteristics of patients, postoperative prognosis of hepatocellular carcinoma, and other data [37]. Random forest regression (RFR) consists of a collection of unpruned regression trees using different bootstrap samples of the training data. In each bootstrap sample, a random sample with replacement and with the same length, some of the data are repeated, and the left-out samples are called out of bag (OOB). In practice, the number of trees n_{trees} and the size of the variable subset n_{subsets} should be optimized to reach the ideal forest by minimizing the OOB error. The values of the parameters n_{trees} and n_{subsets} were optimized simultaneously by using the grid-search method ranging from 10 to 1000 (with step size 10) and from 1 to 9 (with step size 1), respectively. The parameter values which give the lowest RMSE of the OOB data were selected as an indicator for the performance.

4.3.3. Support Vector Regression (SVR). It is a nonlinear kernel-based regression method, which tries to find the best regression hyperplane with smallest structural risk in a high dimensional feature space. It has been used in many medical related applications, such as diagnosis of the incidence of infectious diarrhea [26]. The relationship between input (or inputs) and output is formulated as

$$\begin{aligned} \widehat{y}_{T+1} &= \mathbf{w}_{xy}^{\prime}\cdot\varphi(\mathbf{y})+b, \\ \widehat{y}_{T+1} &= \mathbf{w}_{xy}^{\prime}\cdot\varphi([\mathbf{X};\mathbf{y}])+b, \end{aligned} \tag{14}$$

where $\phi(\cdot)$ denotes kernel function, $\mathbf{x}_t$ is a vector of inputs at time t, and b is a bias term. Radial basis function expressed as $K(a,b)=\exp(-|a-b|^2/\sigma^2)$ was used as a kernel function because of its advantages and simple implementation only one tuning parameter.

4.3.4. Gradient Boosting Regression (GBR). It is a powerful regression model that enhances the decision tree model using gradient boosting, which produces a regression model by combining a series of weak prediction models [38].

GBR iteratively constructs decision trees, and the newly added decision tree is trained according to the negative gradient information of the loss function from the current model. The goal of GBR is to learn an optimal model $F(x)$ that minimizes $\sum_{i=1}^{n}L(y_i,F(x_i))$ for a specified loss function $L(y,F(x))$. The optimal model $F^{*}(x)$ of GBR can be calculated as follows:

$$F^{*}([\mathbf{X};\mathbf{y}])=F_0([\mathbf{X};\mathbf{y}])+\nu\sum_{i=1}^{m}\rho_i h_i([\mathbf{X};\mathbf{y}]), \tag{15}$$

where $F_0(\cdot)$ is the first built decision tree, m is the number of iterations, ν represents the shrinkage parameter that controls the learning rate of GBR, $h_i(\cdot)$ denotes the tree trained in the

i-th iteration, and ρ_i is the weight of $h_i(\cdot)$. The generated optimal model is used for testing.

4.3.5. Extreme Gradient Boosting Regression (XGBoost). It is an improvement on the gradient boosting algorithm [39]. XGBoost adds the regularization terms during the decision tree construction phase. The loss function is

$$\mathscr{L}^{(t)} = \sum_{t}^{T} l(y_t, \widehat{y}_t) + \sum_{k} \Omega(f_k), \tag{16}$$

where $\widehat{y}_t$ represents the prediction of the t-th training sample and Ω is the regularization term. The regularization term is calculated as follows:

$$\Omega(f) = \gamma N + \frac{1}{2}\lambda|\mathbf{w}|^2, \tag{17}$$

where N is the number of leaf nodes, $\mathbf{w}$ is the vector formed by all leaf node values of the decision tree, and γ and λ are manually set parameters. Similar to GBR, the goal of XGBoost is also to minimize the loss function. Besides, XGBoost uses weight shrinkage and column sampling techniques to resist overfitting.

4.3.6. Convolutional Neural Networks (CNN1d). CNN [40] consists of convolutional layers that are based on the convolutional operation. Filtering with kernel window function gives an advantage of image or series processing to CNN architectures with fewer parameters, which are beneficial for computing and storage [41–43]. The basic convolution operation is shown as follows:

$$s = \text{relu}\left(\left[x_1; x_2; \ldots; x_T\right]^* \mathbf{W}\right), \tag{18}$$

where s denotes mapped space and $\mathbf{W}$ denotes kernel. CNN has been verified to have good accuracy results when applied to pattern recognition [44].

4.3.7. Neural Network Regression (NNR). On account of the neural network (NN) approximation and generalization property, NN-based prediction is widely used. Recurrent neural network is a type of deep neural network specially designed for sequence modelling [16]. The main idea of RNN is to provide a weighted feedback connection between layers of neurons and add time significance to the entire artificial NN. But RNN may face the problem of vanishing gradient or gradient explosion during backpropagation. LSTM was proposed by Hochreiter and Schmidhuber in [45].

We adopt a lite RNN model like NARX in [15], which consists of an input layer, an RNN layer, and an output layer, to test its performance. The nonlinear sequential layer can be a LSTM layer or a gated recurrent unit (GRU) [18] layer as well.

5. Experimental Results and Analyses

This section presents the results of intensive experiments to evaluate algorithms over four kinds of inputs and analyses them. Figure 2 shows the performance of PM with varying window size T. Figure 3 shows the evaluation results of all the methods over all the inputs for all metrics. Figure 4 gives the visual comparison of PM methods over four groups of inputs.

5.1. The Impact of Window Size T. Parameter sensitivity is a significant part of experimental analysis. In time series prediction study, T is the most interesting parameter of time delay. Different window sizes will affect the parameter complexity and the overall prediction performance of the methods. In addition, window size is an important indicator to reveal the incubation period of infectious diseases. Therefore, this study visualized the influence of different window sizes on the proposed model. The performance of PM with varying window size T in terms of MAE, RMSE, and R^2 is plotted in Figure 2.

According to the results, when we vary T and keep other parameters fixed, it is easily observed that the performance of PM degraded when T increases from 4 to 20, and the performance tends to get worse as the length of delay time steps continues to increase. A possible reason is that the longer lagged time delay leads to larger feature dimension of the inputs, which causes an increase in the complexity of the regression function in the training process and directly influences the weighted parameter representation for small-scale sample, resulting in the regression function unable to accurately fit the input. By setting $T = 4$, we notice that PM usually achieves the best performance, which possibly suggests that the diarrhea infection cycle pattern lags about four days in Xiamen. This is an important artificial parameter required by the predictive model.

5.2. The Performance of Time Series Prediction. The optimal performance of the benchmark and proposed methods is found at $T = 4$. Hence, the evaluation results of all the comparable methods are shown in Figure 2 when T is fixed at 4.

Several important observations are made about these results:

(1) By benefiting from MF and BSI data, the PM model obtains the best performance in terms of three metrics

(2) The performances of MLR and GBR are better than some complex RNN methods, and ARIMA cannot perform well under unstable time series

(3) CNN1d and RNN methods can achieve stable performance when processing each type of inputs

(4) Both MF and BSI data can improve the prediction accuracy of diarrhea outpatient counts in upcoming weeks.

To demonstrate the effectiveness of PM, we compare the performance of PM against some common methods for the same prediction tasks. The comparable experiments choose MLR, GBR, RFR, SVR, and XGBoost as the representative of the traditional methods and three other deep learning

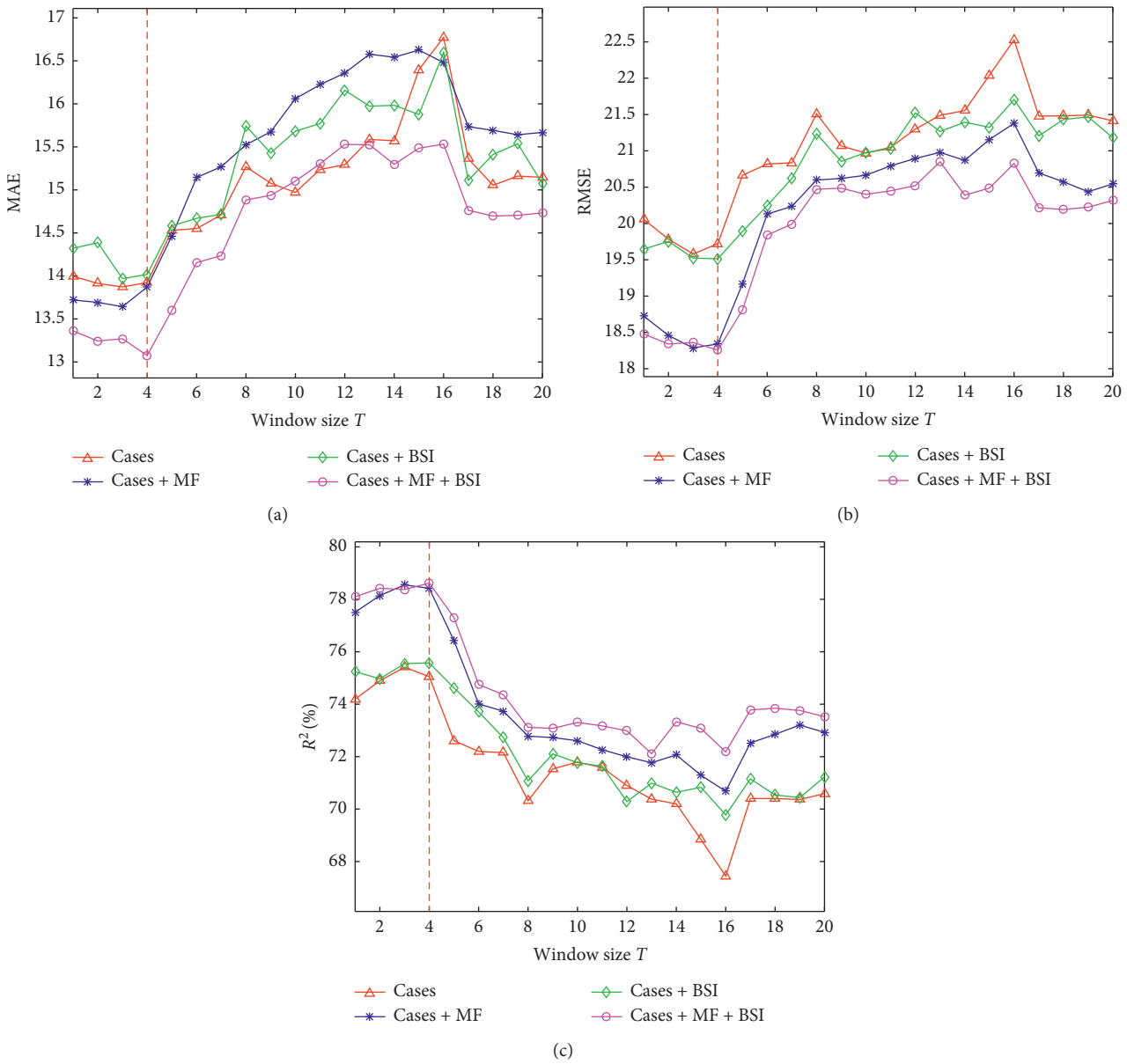


FIGURE 2: The performance of PM with varying window size T in terms of MAE, RMSE, and R^2. (a) The MAE of PM with varying T. The minimum value is found at $T = 4$ (see the red dashed line). (b) The RMSE of PM with varying T. The minimum value is found at $T = 4$ (see the red dashed line). (c) R^2 of PM with varying T. The maximum value is found at $T = 4$ (see the red dashed line).

methods. The deep learning methods include CNN, RNN and LSTM with {64, 128} hidden nodes.

Comparing the performances of all the methods over the four inputs, according to the results of predictions based on historical cases, we can observe that the PM method shows the best performance, and CNN1d, RNN, GBR, and MLR also have good results. The performance of MLR is better than the performance of deep learning methods when the inputs are solely on the basis of historical cases. A possible reason is that the small-scale data collections cannot train deep learning methods well, especially the LSTM method, either overfitting or underfitting. According to the results of predictions based on cases+BSI, we can observe that the traditional m`ts that common learning methods cannot extract sufficient features from high dimensional inputs to predict the upcoming values. According to the results of predictions based on historical cases and BSI, we can observe that the performance of all the methods has degraded, which suggests that too many missing values will greatly affect the prediction performance of the model. However, PM has ability to adapt to input and to deal with missing values, which is significantly stronger than other models. According to the the inputs are solely on the basis of predictions based on cases+MF+BSI, we can observe that PM obtains the best results. Compared to the CNN1d method, it is difficult for traditional learning models to give stable performance with exogenous inputs. The prediction accuracy of the LSTM method has a slight improvment when compared with MLR.

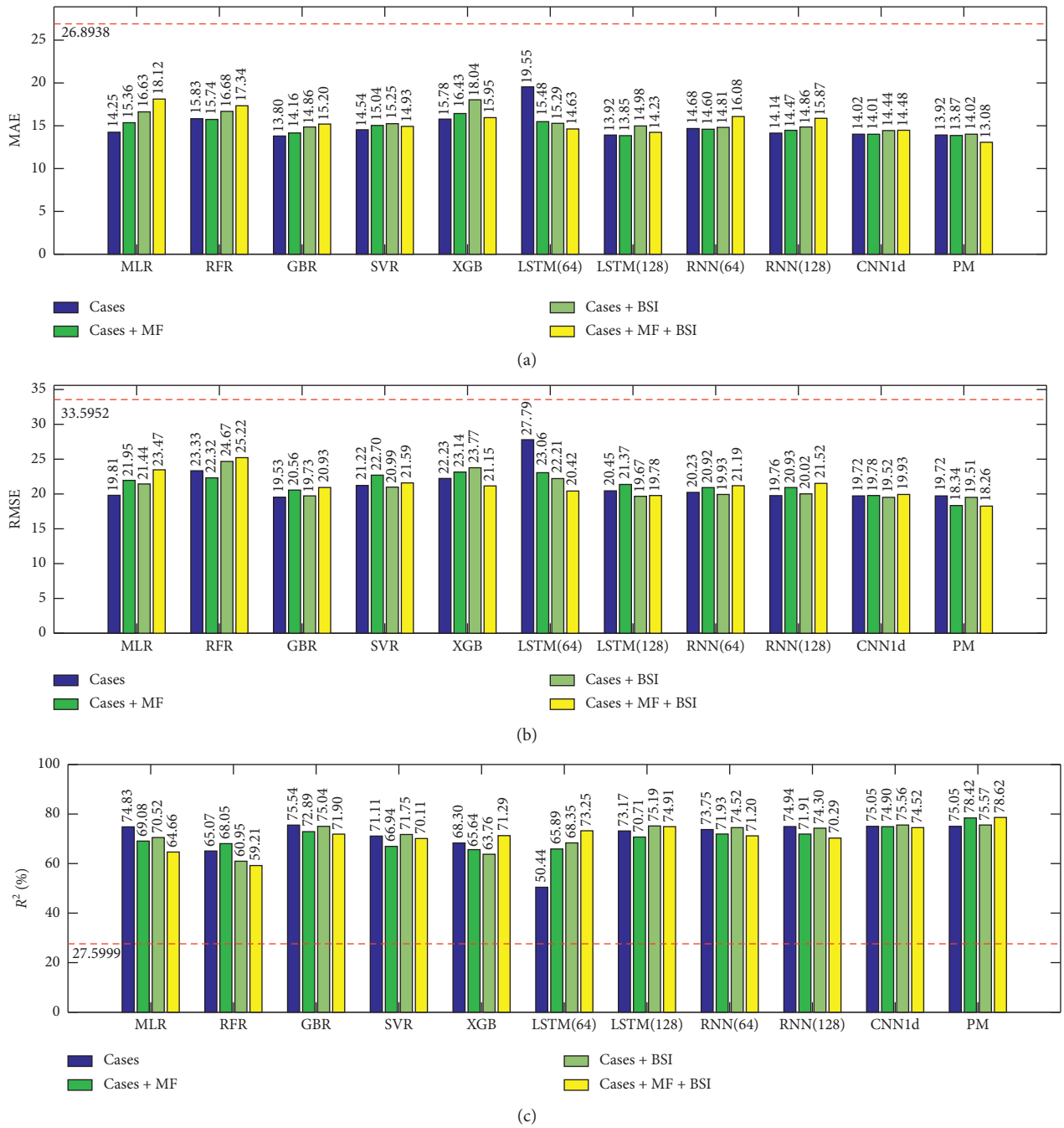


Figure 3: Comparisons of eleven methods on four inputs in terms of three metrics. T is fixed at 4. Lower MAE and RMSE values and higher R^2 value indicate better performance. (a) MAE comparison on methods over inputs (the red dashed line denotes ARIMA). (b) RMSE comparison on methods over inputs (the red dashed line denotes ARIMA). (c) R^2 comparison on methods over inputs (the red dashed line denotes ARIMA).

The prediction effect of the RNN method fluctuates significantly. It is suggested that exogenous inputs can effectively improve model performance, but it also related to the adaptability of the model's structure to exogenous inputs.

There are two potential reasons for the improvements in the PM method. On the one hand, the RNN models only consider the temporal dynamics of exogenous inputs. With small-scale samples, these models cannot extract sufficient features to predict the target. The traditional machine learning methods are easily converged and have worse performance in predicting the upcoming values with exogenous inputs. A possible reason is that these methods cannot capture the correlations among different components of the inputs. The proposed method has ability to learn the interactions of different exogenous factors by introducing the weighted layers, which can effectively improve model performance in small-scale data collections through more concise feature representation. On the other hand, and

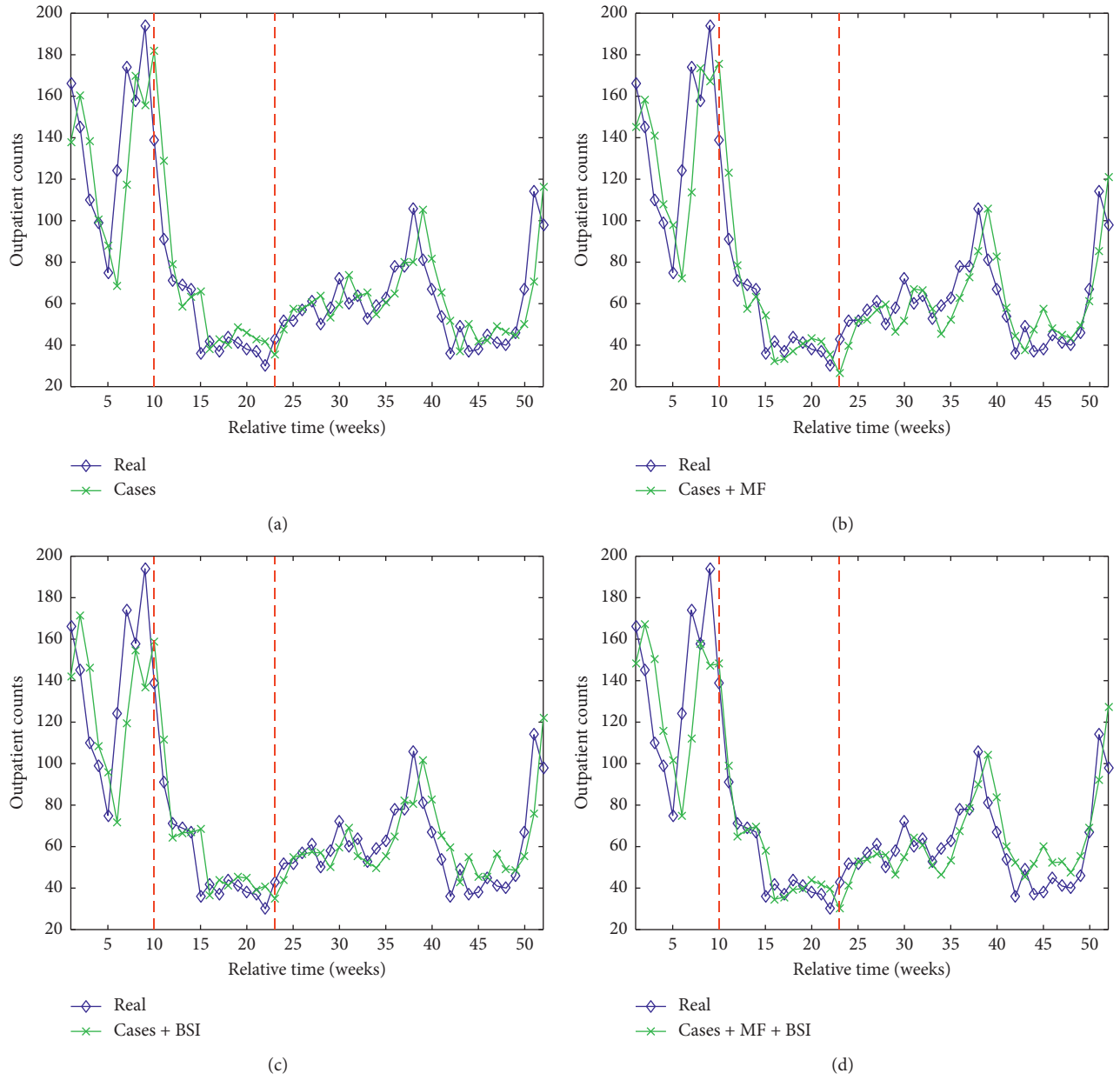


FIGURE 4: The visualizations of real values and PM predictions. The window size T is set to 4, where the best performance is found. The red dashed lines mark the significant differences, whose values follow the highest and lowest real values, respectively. (a) Real values vs. PM predictions based on cases. (b) Real values vs. PM predictions based on cases + MF. (c) Real values vs. PM predictions based on cases + BSI. (d) Real values vs. PM predictions based on cases + MF + BSI.

two weights for each input element greatly decreases the need of parameters, which can be updated by the training process, and the two weighted vectors are adopted to learn temporal weights and factor weights.

When compared with CNN1d, the proposed method exploits the multiplication and addition of two vectors to generate predictions instead of using a kernel matrix to calculate the inputted $2d$ data to get output values. When T is small, the convolution kernel of CNN has poor performance, since the kernel size should smaller than input size. Moreover, not only wide or deep models but also machine learning models have abilities in overcoming overfitting using decay weights (i.e., regularization items) in their loss functions. The key concern is that more parameters will lead to hard training on small-scale data collections. Therefore, PM is more concise and effective, which has better results and better adaption on broad inputs.

Another considerable result is that the performance differences in results among the methods are enlarged with complicated inputs. For example, when only considering the outpatient cases with $T = 4$, PM's performance is similar to other models. But the performance differences become larger when fusing exogenous inputs, since the mined

patterns from heterogeneous data are more complex than those found in only outpatient case data. PM has a better ability to handle small-scale heterogeneous input data. The results of correlation experiments also show that it is completely feasible to use exogenous factors to predict the incidence of infectious diarrhea in Xiamen.

5.3. The Visualization Analysis of Predictions. In time series prediction, it is sometimes more interesting to compare the performance on capturing the so-called extreme events, e.g., an oscillation after a stable growth or decay, or a huge sudden change during oscillations. Thus, a visualization comparing real values and PM predictions is presented in Figure 4. The red dashed lines mark the 10^{th} and 23^{rd} week, whose values follow the highest and lowest real values, respectively.

The prediction of time series extreme events has been proposed by previous studies [46]. We can see that the number of outpatients peaked at 9^{th} week and 10^{th} week, and their outpatient counts significantly dropped. In this situation, PM models with BSI + MF fit the real value better than other methods, which illustrates the ability of weighted layers in summarizing the interactions of exogenous inputs, and the heterogeneous data fusion is an effective method to improve prediction accuracy. The advantage is also retained in the method with MF and BSI. It also should be noted that at 9^{th} week, the number of cases reaches the peak values, and MF effectively improves the accuracy in this timestamp, which makes up for the loss of accuracy of fusion of BSI and historical data in the forecast at that time. Historical cases and MF have slightly fluctuated at 13^{th} week. By incorporating BSI data, the influence of random disturbance in this timestamp is obviously reduced for prediction results. When the value at 23^{rd} week followed the lowest real value, the performance of the PM model with all heterogeneous data fusion is not stable.

Moreover, the predicted value using MF or BSI of 45^{th} week is obviously higher, and the situation is reflected in the fusion of MF and BSI. A possible reason is that the internal random disturbance of exogenous data has not been completely eliminated in the preprocessing stage. Therefore, the extreme weather conditions or special events and degrade the prediction performance.

6. Conclusions

In this paper, we focus on predicting the number of infectious diarrhea outpatient visits in the upcoming week. A parsimonious model (PM) is proposed by condensing previous prediction models. The benchmarks of ten algorithms with four groups of inputs show the advantage of our method. It achieves better prediction performance by consolidating MF data and BSI data.

In the future, we will try to investigate the prediction of other cities and improve the robustness of PM in predicting those cities. Moreover, the simultaneous forecast of multiple cities would be our future research direction.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Xiamen Centre for Disease Control and Prevention for sharing the data. This study was supported in part by the Natural Science Foundation of Fujian Province of China (nos. 2018J01539 and 2019J01713) and the Science Project of Xiamen City (no. 2019SH400060).

References

[1] T. Wu, C. Perrings, A. Kinzig, J. P. Collins, B. A. Minteer, and P. Daszak, "Economic growth, urbanization, globalization, and the risks of emerging infectious diseases in China: a review," *Ambio*, vol. 46, no. 1, pp. 18–29, 2017.

[2] P. Kelly, "Infectious diarrhoea," *Medicine*, vol. 47, no. 6, pp. 341–346, 2019.

[3] L. Wang, L. Jin, and W. Xiong, "Chapter 2–nfectious diseasesurveillance in China," in *Early Warning for Infectious DiseaseOutbreak*, W. Yang, Ed., pp. 23–33, Academic Press, Cambridge, MA, USA, 2017.

[4] Y. Lan, D. Zhou, H. Zhang, and S. Lai, "Chapter 3-developmentof early warning models," in *Early Warning for Infectious Disease Outbreak*, W. Yang, Ed., pp. 35–74, Academic Press, Cambridge, MA, USA, 2017.

[5] I. N. Soyiri and D. D. Reidpath, "An overview of health forecasting," *Environmental Health and Preventive Medicine*, vol. 18, no. 1, pp. 1–9, 2013.

[6] L. Yu, L. Zhou, L. Tan et al., "Application of multiple seasonal arima model in forecasting incidence of hfmd in Wuhan, China," *International Journal of Infectious Diseases*, vol. 21, p. 271, 2014.

[7] P. Zhang, S. Shu, and M. Zhou, "An online fault detection model and strategies based on SVM-grid in clouds," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 2, pp. 445–456, 2018.

[8] J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, "Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest," *Applied Energy*, vol. 262, Article ID 114566, 2020.

[9] R. Oppenheim, "Forecasting via the box-jenkins method," *Journal of the Academy of Marketing Science*, vol. 6, no. 3, pp. 206–221, 1978.

[10] G. Mahalakshmi, S. Sridevi, and S. Rajaram, "A survey on forecasting of time series data," in *Proceedings of the 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, pp. 1–8, Kovilpatti, India, January 2016.

[11] S. Chen, X. Liu, Y. Wu et al., "The application of meteorological data and search index data in improving the prediction of HFMD: a study of two cities in Guangdong province, China," *Science of The Total Environment*, vol. 652, pp. 1013–1021, 2019.

[12] R. Adhikari and R. K. Agrawal, "An introductory study on time series modeling and forecasting," 2013, https://arxiv.org/abs/1302.6613.

[13] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 703–715, 2019.

[14] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 601–614, 2019.

[15] E. Diaconescu, "The use of NARX neural networks to predict chaotic time series," *WSEAS Transactions on Computer Research*, vol. 3, 2008.

[16] R. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Comput*, vol. 1, pp. 270–280, 1998.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014, https://arxiv.org/abs/1409.3215.

[18] K. Cho, B. Merrienboer, Y. Bengio et al., "On the properties of neural machine translation: encoder-decoder approaches," 2014, https://arxiv.org/abs/1409.1259.

[19] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," 2015, https://arxiv.org/abs/1506.00019.

[20] Y. Qin, D. Song, H. Chen et al., "A dual-stage attention-based recurrent neural network for time series prediction," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence IJCAI2017*, pp. 2627–2633, Melbourne, Australia, August 2017.

[21] G. Lai, W. Chang, and Y. Yang, "Modeling long- and short-term temporal patterns with deep neural networks," in *Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[22] Y. Chang, F. Sun, Y. H. Wu et al., "A memory-network based solution for multivariate time-series forecasting," 2018, https://arxiv.org/abs/1809.02105.

[23] Y. Wu, Y. Yang, H. Nishiura et al., "Deep learning for epidemiological predictions," in *Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval SIGIR2018*, pp. 1085–1088, Ann Arbor, MI, USA, June 2018.

[24] J. Chen, K. Li, P. Herrero et al., "Dilated recurrent neural network for short-time prediction of glucose concentration," in *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval SIGIR2018*, pp. 95–103, Ann Arbor, MI, USA, June 2018.

[25] S. Shih, F. Sun, and H. Lee, "Temporal pattern attention for multivariate time series forecasting," *Machine Learning*, vol. 108, no. 8-9, pp. 1421–1441, 2019.

[26] Y. Wang, J. Gu, Z. Zhou, and Z. Wang, "Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion," *Knowledge-Based Systems*, vol. 88, pp. 12–23, 2015.

[27] Y. Wang, J. Li, J. Gu, Z. Zhou, and Z. Wang, "Artificial neural networks for infectious diarrhea prediction using meteorological factors in shanghai (China)," *Applied Soft Computing*, vol. 35, pp. 280–290, 2015.

[28] S. Li, T. Chen, L. Wang, and C. Ming, "Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index," *Tourism Management*, vol. 68, pp. 116–126, 2018.

[29] Z. Li, T. Liu, G. Zhu et al., "Dengue Baidu search index data can improve the prediction of local dengue epidemic: a case study in Guangzhou, China," *PLoS Neglected Tropical Diseases*, vol. 11, no. 3, pp. 1–13, 2017.

[30] Z. Wang, Y. Huang, B. He et al., "TDDF.: HFMD outpatients prediction based on time series decomposition and heterogenous data fusion in Xiamen, China," in *Proceedings of the 15th International Conference, ADMA 2019*, pp. 658–667, Dalian, China, November 2019.

[31] "Xiamen" from Wikipedia, 2020, https://en.wikipedia.org/wiki/Xiamen.

[32] Xiamen Statistics Bureau, 2018, http://tjj.xm.gov.cn/tjnj/2018/.

[33] K. Liu, T. Wang, Z. Yang et al., "Using Baidu search index to predict dengue outbreak in China," *Scientific Reports*, vol. 6, p. 38040, 2016.

[34] J. Brownlee, "Long short-term memory networks with python: develop sequence prediction models with deep learning," Machine Learning Mastery, 2017, https://machinelearningmastery.com.

[35] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2015, https://arxiv.org/abs/1412.6980.

[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] C. C. R. Sady and A. L. P. Ribeiro, "Symbolic features and classification via support vector machine for predicting death in patients with Chagas disease," *Computers in Biology and Medicine*, vol. 70, pp. 220–227, 2016.

[38] H. F. Jerome, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[39] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system,"in *Proceedings of the Proceedings of the 22nd ACM-SIGKDD International Conference on Knowledge Discoveryand Data Mining SIGKDD2016,* San Francisco, CA, USA, August 2016.

[40] Y. Lecun, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, https://arxiv.org/abs/1803.01271.

[42] S. Barra, S. M. Carta, A. Corriga, A. S. Podda, and D. R. Recupero, "Deep learning and time series-to-image encoding for financial forecasting," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 3, pp. 683–692, 2020.

[43] M. M. Hassan, A. Gumaei, G. Aloi, G. Fortino, and M. Zhou, "A smartphone-enabled fall detection framework for elderly people in connected home healthcare," *IEEE Network*, vol. 33, no. 6, pp. 58–63, 2019.

[44] D. Freire-Obregón, F. Narducci, S. Barra, and M. Castrillón-Santana, "Deep learning for source camera identification on mobile devices," *Pattern Recognition Letters*, vol. 126, pp. 86–91, 2019.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[46] D. Ding, M. Zhang, X. Pan et al., "Modeling extreme events in time series prediction," in *Proceedings of the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1114–1122, July 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8884539, 13 pages
https://doi.org/10.1155/2020/8884539

Research Article

A Time-Critical Topic Model for Predicting the Survival Time of Sepsis Patients

Wenping Guo,[1,2] **Zhuoming Xu**,[2] **Xijian Ye**,[3] **Shiqing Zhang**,[1] **Xiaoming Zhao**,[1] **and Xue Li**[4,5]

[1]*Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China*
[2]*College of Computer and Information, Hohai University, Nanjing 210098, China*
[3]*Polytechnic Institute, Zhejiang University, Hangzhou 310015, China*
[4]*Neusoft Institute of Information, Dalian Neusoft University of Information, Dalian 116081, China*
[5]*School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia*

Correspondence should be addressed to Wenping Guo; guowp@tzc.edu.cn

Received 19 March 2020; Revised 18 August 2020; Accepted 29 August 2020; Published 15 September 2020

Academic Editor: Chenxi Huang

Sepsis is a leading cause of mortality in intensive care units and costs hospitals billions of dollars annually worldwide. Predicting survival time for sepsis patients is a time-critical prediction problem. Considering the useful sequential information for sepsis development, this paper proposes a time-critical topic model (TiCTM) inspired by the latent Dirichlet allocation (LDA) model. The proposed TiCTM approach takes into account the time dependency structure between notes, measurement, and survival time of a sepsis patient. Experimental results on the public MIMIC-III database show that, overall, our method outperforms the conventional LDA and linear regression model in terms of recall, precision, accuracy, and F1-measure. It is also found that our method achieves the best performance by using 5 topics when predicting the probability for 30-day survival time.

1. Introduction

Predicting the survival time of patients is an active research area for both clinicians and scientists [1–5]. It can significantly contribute to making decisions about clinical treatment, allocation of medical resources, and hospice care for patients [1]. Sepsis is a disease of life-threatening organ dysfunction caused by a dysregulated host response to infection [6]. Without timely treatment, sepsis can rapidly lead to tissue damage, organ failure, and death. Common signs and symptoms include fever, increased heart rate, increased breathing rate, and confusion. In US health systems, the cost for patients with sepsis accounted for more than $20 billion (5.2%) of total US hospital costs in 2011 [7]. In 2001–2010, one in twenty deaths in England was associated with sepsis based on information recorded on death certificates [8]. Although clinicians have made efforts to improve sepsis patient survival time, the mortality rate of sepsis is still very high [9, 10]. Thus, accurate prediction of survival time for sepsis patients could help clinicians conduct prevention, provide early warning and effective treatment, and reduce the mortality rate. Unfortunately, the pathogenesis of sepsis remains unclear. Predicting the survival time for specific diseases, such as sepsis, is still a challenging problem.

To help clinicians understand the overall body situations of a patient, ICUs have introduced many different mechanisms for describing the patient's body situation and progress in the ICU. Different severity of illness scores has been used to predict sepsis or mortality risk scores in the ICU. The most widely used score-based methods include the acute physiology and chronic health evaluation (APACHE III) [11], the simplified acute physiology score (SAPS II) [12], the modified early warning score (MEWS) [13], the sepsis-related organ failure assessment (SOFA) [14], and quick SOFA (qSOFA) [6]. In addition, Zhang and Hong [15] proposed a novel score for predicting hospital mortality for severe sepsis. These score-based methods utilize a set of easily obtainable measurements from various patients to

generate risk scores. Although they allow clinicians to make rapid diagnostics of a patient, the obtained results are not satisfactory [16, 17]. Additionally, these score-based approaches only evaluate the patient body situation at specific times and cannot predict the survival time of sepsis patients because the development of sepsis is a time-sensitive process.

Sepsis is a life-threatening condition that arises when the body's response to infection causes injury to its own tissues and organs [18, 19]. It is difficult to predict the development of sepsis based on a small number of measurements. Conventional topic models such as latent Dirichlet allocation (LDA) are unsupervised machine learning methods that can recognize latent topic information in massive document collections [20, 21]. Lehman et al. [22] proposed a novel approach for ICU patient risk stratification using a topic model. Ghassemi et al. [23] proposed a mortality model using a topic model to predict in-hospital mortality, 30-day postdischarge mortality, and 1-year postdischarge mortality. Vairamani [24] proposed an approach for mortality prediction in ICU patients based on LDA. Zhang et al. [25] proposed a novel survival topic model inspired by LDA for trauma patients. However, these topic models overlook the useful sequential information for disease development, thereby reducing the prediction accuracy for sepsis patient survival time. Unfortunately, the development and treatment of sepsis is a time-critical process that has a high correlation between the order of words in measurement and notes.

To address this issue, this paper proposes a time-critical topic model (TiCTM) inspired by the LDA model to predict the survival time of sepsis patients. The proposed TiCTM approach takes into account the time dependency structure between notes, measurement, and survival time of a sepsis patient. We consider the time-critical dynamic process of sepsis patients as an approximately linear variation under clinician treatment. Therefore, the linear change in the parameter of TiCTM can reflect the time-critical dynamic process, whereas the parameter of LDA is fixed. Our experimental results on the public MIMIC-III database show that, overall, TiCTM outperforms the conventional LDA and linear regression model in terms of recall, precision, accuracy, and F1-measure. In particular, TiCTM obtains the best performance when predicting the probability for 30-day survival time using 5 topics.

The remainder of this paper is organized as follows. In Section 2, we describe the proposed TiCTM for predicting survival time for adult sepsis patients in the ICU. The experiments and evaluations are discussed in Section 3. Finally, we conclude our research in Section 4.

2. Methodology

In this section, we first present a brief review of the classical LDA model and then describe our proposed TiCTM approach in detail.

2.1. Brief Review of LDA. LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [20]. The LDA model is represented as a probabilistic graphical model, as shown in Figure 1. The meaning of each notation for LDA is shown in Table 1.

The LDA model considers documents as the collection of words, which overlook the order of words. To address this issue, we propose a time-critical topic model inspired by LDA to predict the survival time of sepsis patients, as described as follows.

2.2. Our Proposed Method. Our proposed TiCTM approach considers the time dependency structure between notes, measurement, and survival time of a sepsis patient. The TiCTM model is depicted in Figure 2.

Assume that sequential notes and measurement submodel have several phases. For phase 1, the meaning of each notation is shown in Table 2.

For phase $i \in \{2, 3, \ldots, I-1\}$, the meaning of each notation is shown in Table 3.

For phase I, the meaning of each notation is shown in Table 4.

As shown in Figure 2, the proposed TiCTM model consists of two submodels: sequential notes and measurement and survival time prediction. This is because the development of sepsis is a time-critical process. The main idea of the TiCTM model is that we consider the time-critical dynamic process of sepsis patients as approximately linear variation under clinician treatment. Therefore, we employ the linear change in the parameter of TiCTM to reflect this time-critical dynamic process, whereas the parameter of LDA is fixed.

2.2.1. Sequential Notes and Measurement. For a sepsis patient m, the parameter μ_m represents the patient's initial body situation. In everyday clinician treatment, the patient's body situation changes $\Delta\mu$. Therefore, after i_m days of clinician treatment, the parameter μ_m of patient m will transit to $\mu_m + i_m\Delta\mu$, which is used to reflect the sequential changing process. The parameter $\mu_m + i_m\Delta\mu$ determines the disease probability distribution θ_i. The parameter θ_i determines the disease z_i for patient m. β (probability of bigrams appearing in the topics of notes dictionary) and z_i (topic variable) generate $N_{i,m}$ words in the notes of patient m. λ (probability of measurements appearing in the topics of measurement dictionary) and z_i generate $R_{i,m}$ measurements in the measurement result of patient m.

For each word $w_{i,n}$, $n = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, I$, we draw a topic assignment $z_{i,n}|\theta_i \sim \prod(A(\theta_i))$, where

$$A(\theta_i) = \left(\frac{1}{\sqrt{\pi}} e^{-\theta^2_{i,z_{i,n}}}\right)^{w^v_{i,n}}, \tag{1}$$

where $A(\theta_i)$ represents the probability of topic $z_{i,n}$ appearing, which is determined by parameter θ_i.

Draw a word $w_{i,n}|z_{i,n}, \beta \sim \prod(B(\beta))$, where

$$B(\beta) = \left(\frac{1}{\sqrt{\pi}} e^{-\beta^2_{z_{i,n},v}}\right)^{w^v_{i,n}}, \tag{2}$$

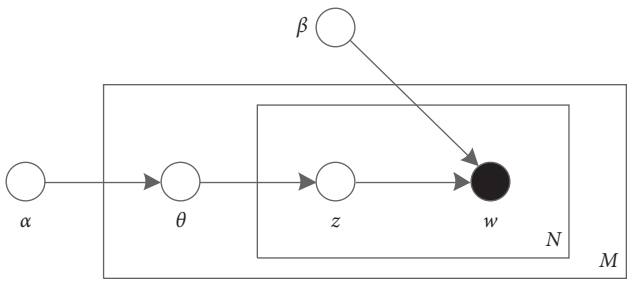


Figure 1: The graphical model of LDA.

Table 1: Notations for LDA.

Notation	Description
α	Parameter which is a K-vector
K	Total number of latent topics
θ	A K-dimensional Dirichlet random variable
z	Topic variable
β	The probability of words appearing in topics of dictionary
w	A set of N words in a document
N	A document which is a sequence of N words
M	A corpus which is a collection of M documents

where $B(\beta)$ represents the probability of word $w_{i,n}$ appearing, which is determined by topic $z_{i,n}$ and parameter β.

For each phase i, the measurement is shown in Figure 2. For each measurement $x_{i,r}$, $r = 1, 2, \ldots, R$ and $i = 1, 2, \ldots, I$.

Draw a topic assignment $z_{i,r}|\theta_i \sim \prod(A'(\theta_i))$, where

$$A'(\theta_i) = \left(\frac{1}{\sqrt{\pi}} e^{-\theta^2_{i,z_{i,r}}}\right)^{x^{v'}_{i,r}}, \tag{3}$$

where $A'(\theta_i)$ represents the probability of topic $z_{i,r}$ appearing, which is determined by parameter θ_i.

Draw a measurement $x_{i,r}|z_{i,r}, \lambda \sim \prod(B'(\lambda))$, where

$$B'(\lambda) = \left(\frac{1}{\sqrt{\pi}} e^{-\lambda^2_{z_{i,r},v'}}\right)^{x^{v'}_{i,r}}, \tag{4}$$

where $B'(\lambda)$ represents the probability of measurement $x_{i,r}$ appearing, which is determined by topic $z_{i,r}$ and parameter λ.

2.2.2. Survival Time Prediction. For each patient m, T_m denotes the real survival time, and $\widehat{T}_m$ represents the predicted survival time with TiCTM. Body situation $(\mu_m + I_m\Delta\mu)$ and regression coefficient (η) are used for predicting the survival time $(\widehat{T}_m)$ for patient m. The objective of the problem is to minimize the difference (Diff) between T_m and $\widehat{T}_m$. As shown in Figure 2, we define the time to death $\widehat{T}$ from formula (5). $\widehat{T}|\mu, \Delta\mu, i, h, \eta \sim C(\widehat{f})$, where

$$C(\widehat{f}) = \frac{\left(1.01/\left(1 + e^{\widehat{f}}\right)\right) - 0.005}{1.005 - \left(1.01/\left(1 + e^{\widehat{f}}\right)\right)}, \tag{5}$$

$$\widehat{f} = \sum_{k=1}^{K}\left(\frac{1}{1 + e^{-\mu_k - I\Delta\mu_k}}\right)\frac{\sum_{v=1}^{V} h_v \eta_{v,k} + \sum_{v'=1}^{v'} h_{v'}\eta_{v',k}}{\sum_{v=1}^{V} h_v + \sum_{v'=1}^{v'} h_{v'}}. \tag{6}$$

The details for predicting survival time for sepsis patients using the TiCTM model are further described in the following subsection.

2.3. Details of Survival Time Prediction Using TiCTM

2.3.1. Definition of Real Survival Time Function. Assume that there are M patients, f_m is defined as the one-variable function of T_m, whereas $\widehat{f}_m$ is defined as the one-variable function of $\widehat{T}_m$. These two functions are further described later. The objective of survival time prediction for sepsis patients is to minimize the difference (Diff) between f_m and $\widehat{f}_m$:

$$\min \text{Diff} = \min \sum_{m=1}^{M} \frac{1}{1 + e^{-f_m}} \times \left(f_m - \widehat{f}_m\right)^2. \tag{7}$$

We hypothesize T_m as the real survival time for patient m after the second phase. f_m is the function for real survival time analysis T_m.

Assume that patient m has the same probability p_m to survive every day; then, we can calculate the probability of death on the $(T_m + 1)$ day using

$$\begin{aligned} &p_m^{T_m}(1 - p_m), \\ &p_m \in [0, 1]. \end{aligned} \tag{8}$$

To maximize $p_m^{T_m}(1 - p_m)$, it must be satisfied with

$$p_m = \frac{T_m}{T_m + 1}, \tag{9}$$

where p_m is a form of sigmoid function as

$$p_m = \frac{1}{1 + e^{-x}}. \tag{10}$$

To make $p_m = 0$ (Death) and $p_m = 1$ (Survive), we modify p_m to

$$p_m = \frac{1.01}{1 + e^{-x}} - 0.005. \tag{11}$$

Then, we can obtain formula (12) for f_m and formula (13) for T_m:

$$f_m = \log\left(\frac{1.01}{\left(T_m/(T_m + 1)\right) + 0.005} - 1\right), \tag{12}$$

$$T_m = \frac{p_m}{1 - p_m} = \frac{\left(1.01/\left(1 + e^{f_m}\right) - 0.005\right)}{1.005 - \left(1.01/\left(1 + e^{f_m}\right)\right)}. \tag{13}$$

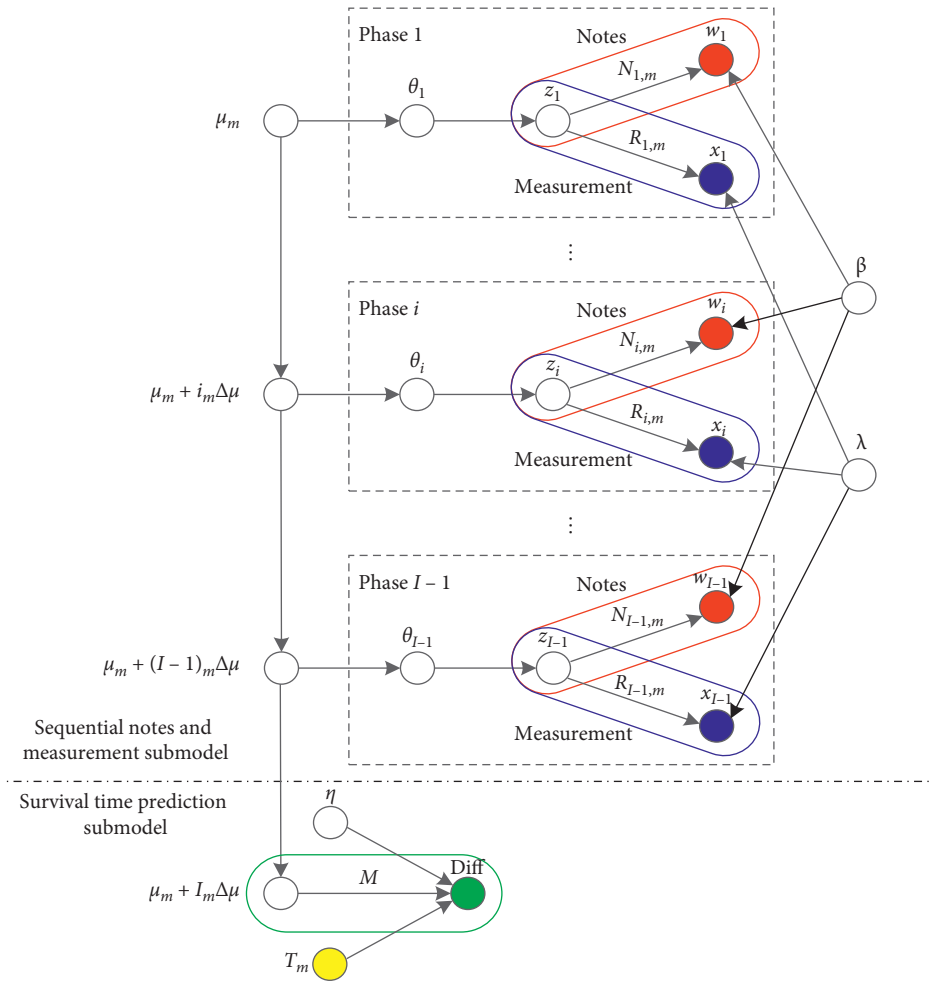


Figure 2: Our proposed TiCTM.

Table 2: Notations for phase 1.

Notation	Description
M	Patient m
μ_m	Parameter μ_m
θ_1	A K-dimensional normal random variable
z_1	Topic variable
k	Latent topic (latent disease) k
K	Total number of latent topics
w_1	A set of $N_{1,m}$ bigrams in a note
$N_{1,m}$	Number of bigram results by z_1
V	Size of notes dictionary
β	The probability of words appearing in topics of notes dictionary
$\beta_{k,v}$	The probability of v-th bigram appearing in k topic of notes dictionary, $\sum_{v=1}^{V}\beta_{k,v}=1$
x_1	Measurement x_1
$R_{1,m}$	Number of measurement results by z_1
v'	Size of measurement dictionary
λ	The probability of measurements appearing in the topics of measurement dictionary
$\lambda_{k,v'}$	The probability of the v'-th measurement appearing in k topics of measurement dictionary $\sum_{v'=1}^{v'}\lambda_{k,v'}=1$

2.3.2. Definition of the Survival Time Prediction Function. To verify the effectiveness of the proposed TiCTM model, this paper uses the two-phase survival model to analyze survival time analysis for adult sepsis patients. The measurement and notes in the first phase limit the time for the patient's admission within 24 hours. The second phase uses the patient's measurement and notes after 24 hours to the last.

To calculate the probability of measurement x_1 and note w_1 in the first phase by body condition μ_m, we use

Table 3: Notations for phase i.

Notation	Description
i	The phase i
i_m	Patient m treated by clinicians for i_m days
$\Delta\mu$	Parameter $\Delta\mu$
θ_i	A K-dimensional normal random variable
z_i	Topic variable
w_i	A set of $N_{i,m}$ bigrams in a note
$N_{i,m}$	Number of bigram results by z_i
x_i	Measurement x_i
$R_{i,m}$	Number of measurement results by z_i

Table 4: Notations for phase M.

Notation	Description
M	Total number of patients
I	Total number of phases
I_m	Patient m treated by clinicians for a total of I_m days
T_m	Real survival time for patient m after phase I-1
$\widehat{T}_m$	Predicted survival time for patient m after phase I-1
η	Regression coefficient
f_m	Function for computing real survival time
$\widehat{f}_m$	Function for computing predicted survival time
Diff	The function value of the difference between f_m and $\widehat{f}_m$

$$\int p(\theta_1|\mu_m)\left(\prod_{n=1}^{N_m}\sum_{z_{1,n}=1}^{K} p\left(z_{1,n}\,|\,\theta_1\right)p\left(w_{1,n}\,|\,z_{1,n},\beta\right)\right) \times\left(\prod_{r=1}^{R_m}\sum_{z_{1,r}=1}^{K} p\left(z_{1,r}\,|\,\theta_1\right)p\left(x_{1,r}\,|\,z_{1,r},\lambda\right)\right)\mathrm{d}\theta_1, \tag{14}$$

where

$$p\left(\theta_1\,|\,\mu_m\right)=\prod_{k=1}^{K}\frac{1}{\sqrt{\pi}}e^{-\left(\theta_{1,k}-\mu_{m,k}\right)^2},$$

$$p\left(z_{1,n}\,|\,\theta_1\right)p\left(w_{1,n}\,|\,z_{1,n},\beta\right)=\prod_{v=1}^{V}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{1,z_{1,n}}^2}\frac{1}{\sqrt{\pi}}e^{-\beta_{z_{1,n},v}^2}\right)^{w_{1,n}^v},$$

$$p\left(z_{1,r}\,|\,\theta_1\right)p\left(x_{1,r}\,|\,z_{1,r},\lambda\right)=\prod_{v'=1}^{v'}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{1,z_{1,r}}^2}\frac{1}{\sqrt{\pi}}e^{-\lambda_{z_{1,r},v'}^2}\right)^{x_{1,r}^{v'}}. \tag{15}$$

To calculate the probability of measurement x_2 and note w_2 in the second phase by body condition $(\mu_m + i_m\Delta\mu)$, we use

$$\int p\left(\theta_2\,|\,\mu_m+i_m\Delta\mu\right)\left(\prod_{n=1}^{N_m'}\sum_{z_{2,n}=1}^{K} p\left(z_{2,n}\,|\,\theta_2\right)p\left(w_{2,n}\,|\,z_{2,n},\beta\right)\right) \times\left(\prod_{r=1}^{R_m'}\sum_{z_{2,r}=1}^{K} p\left(z_{2,r}\,|\,\theta_2\right)p\left(x_{2,r}\,|\,z_{2,r},\lambda\right)\right)\mathrm{d}\theta_2, \tag{16}$$

where

$$p\left(\theta_2\,|\,\mu_m+i_m\Delta\mu\right)=\prod_{k=1}^{K}\frac{1}{\sqrt{\pi}}e^{-\left(\theta_{2,k}-\mu_{m,k}-i_m\Delta\mu_k\right)^2},$$

$$p\left(z_{2,n}\,|\,\theta_2\right)p\left(w_{2,n}\,|\,z_{2,n},\beta\right)=\prod_{v=1}^{V}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{2,z_{2,n}}^2}\frac{1}{\sqrt{\pi}}e^{-\beta_{z_{2,n},v}^2}\right)^{w_{2,n}^v},$$

$$p\left(z_{2,r}\,|\,\theta_2\right)p\left(x_{2,r}\,|\,z_{2,r},\lambda\right)=\prod_{v'=1}^{v'}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{2,z_{2,r}}^2}\frac{1}{\sqrt{\pi}}e^{-\lambda_{z_{2,r},v'}^2}\right)^{x_{2,r}^{v'}}. \tag{17}$$

The likelihood function of the two-phase of patient m can be obtained from

$$\begin{aligned}L\left(w_1,w_2,x_1,x_2|\mu_m,\Delta\mu,i_m,\beta,\lambda\right)&=\int p\left(\theta_1|\mu_m\right)\\&\cdot\left(\prod_{n=1}^{N_m}\sum_{z_{1,n}=1}^{K}p\left(z_{1,n}|\theta_1\right)p\left(w_{1,n}|z_{1,n},\beta\right)\right)\\&\times\left(\prod_{r=1}^{R_m}\sum_{z_{1,r}=1}^{K}p\left(z_{1,r}\,|\,\theta_1\right)p\left(x_{1,r}|z_{1,r},\lambda\right)\right)\mathrm{d}\theta_1\\&\times\int p\left(\theta_2|\mu_m+i_m\Delta\mu\right)\left(\prod_{n=1}^{N_m'}\sum_{z_{2,n}=1}^{K}p\left(z_{2,n}|\theta_2\right)p\left(w_{2,n}|z_{2,n},\beta\right)\right)\\&\times\left(\prod_{r=1}^{R_m'}\sum_{z_{2,r}=1}^{K}p\left(z_{2,r}\,|\,\theta_2\right)p\left(x_{2,r}\,|\,z_{2,r},\lambda\right)\right)\mathrm{d}\theta_2\\&=\int\left(\prod_{k=1}^{K}\frac{1}{\sqrt{\pi}}e^{-\left(\theta_{1,k}-\mu_{m,k}\right)^2}\right)\\&\cdot\left(\prod_{n=1}^{N_m}\sum_{z_{1,n}=1}^{K}\prod_{v=1}^{V}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{1,z_{1,n}}^2}\frac{1}{\sqrt{\pi}}e^{-\beta_{z_{1,n},v}^2}\right)^{w_{1,n}^v}\right)\\&\times\left(\prod_{r=1}^{R_m}\sum_{z_{1,r}=1}^{K}\prod_{v'=1}^{v'}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{1,z_{1,r}}^2}\frac{1}{\sqrt{\pi}}e^{-\lambda_{z_{1,r},v'}^2}\right)^{x_{1,r}^{v'}}\right)\mathrm{d}\theta_1\\&\times\int\left(\prod_{k=1}^{K}\frac{1}{\sqrt{\pi}}e^{-\left(\theta_{2,k}-\mu_{m,k}-i_m\Delta\mu_k\right)^2}\right)\\&\cdot\left(\prod_{n=1}^{N_m'}\sum_{z_{2,n}=1}^{K}\prod_{v=1}^{V}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{2,z_{2,n}}^2}\frac{1}{\sqrt{\pi}}e^{-\beta_{z_{2,n},v}^2}\right)^{w_{2,n}^v}\right)\\&\times\left(\prod_{r=1}^{R_m'}\sum_{z_{2,r}=1}^{K}\prod_{v'=1}^{v'}\left(\frac{1}{\sqrt{\pi}}e^{-\theta_{2,z_{2,r}}^2}\frac{1}{\sqrt{\pi}}e^{-\lambda_{z_{2,r},v'}^2}\right)^{x_{2,r}^{v'}}\right)\mathrm{d}\theta_2.\end{aligned} \tag{18}$$

The log of the likelihood function of the two-phase of patient m can be obtained from

$$\log\left(L\left(w_1, w_2, x_1, x_2 \mid \mu_m, \Delta\mu, i_m, \beta, \lambda\right)\right) \geq \left(-N_m - R_m - \frac{K}{2}\right)\log(\pi)$$

$$+ (N_m + R_m)\log K - \sum_{k=1}^{K} \mu_{m,k}^2 - \frac{1}{K}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\sum_{v=1}^{V} w_{1,n}^{v}\beta_{k,v}^2 - \frac{1}{K}\sum_{r=1}^{R_m}\sum_{k=1}^{K}\sum_{v'=1}^{v'} x_{1,r}^{v'}\lambda_{k,v'}^2 + \log(c_m)$$

$$+\left(-N'_m - R'_m - \frac{K}{2}\right)\log(\pi) + (N'_m + R'_m)\log K - \sum_{k=1}^{K}\left(\mu_{m,k} + i_m\Delta\mu_k\right)^2 - \frac{1}{K}\sum_{n=1}^{N'_m}\sum_{k=1}^{K}\sum_{v=1}^{V} w_{2,n}^{v}\beta_{k,v}^2 - \frac{1}{K}\sum_{r=1}^{R'_m}\sum_{k=1}^{K}\sum_{v'=1}^{v'} x_{2,r}^{v'}\lambda_{k,v'}^2 + \log(c'_m), \tag{19}$$

where

$$\log(c_m) = \frac{K}{2}\log(\pi) - \frac{K}{2}\log\left(1 + \frac{N_m + R_m}{K}\right) + \frac{K}{K + N_m + R_m}\sum_{k=1}^{K}\mu_{m,k}^2,$$
$$\log(c'_m) = \frac{K}{2}\log(\pi) - \frac{K}{2}\log\left(1 + \frac{N'_m + R'_m}{K}\right) + \frac{K}{K + N'_m + R'_m}\sum_{k=1}^{K}\left(\mu_{m,k} + i_m\Delta\mu_k\right)^2. \tag{20}$$

To simplify the calculation, we use formula (21) to replace the log of the likelihood function of formula (19):

$$\left(-N_m - R_m - \frac{K}{2}\right)\log(\pi) + (N_m + R_m)\log K - \sum_{k=1}^{K}\mu_{m,k}^2 - \frac{1}{K}\sum_{n=1}^{N_m}\sum_{k=1}^{K}\sum_{v=1}^{V} w_{1,n}^{v}\beta_{k,v}^2 - \frac{1}{K}\sum_{r=1}^{R_m}\sum_{k=1}^{K}\sum_{v'=1}^{V'} x_{1,r}^{v'}\lambda_{k,v'}^2 + \log(c_m)$$

$$+\left(-N'_m - R'_m - \frac{K}{2}\right)\log(\pi) + (N'_m + R'_m)\log K - \sum_{k=1}^{K}\left(\mu_{m,k} + i_m\Delta\mu_k\right)^2 - \frac{1}{K}\sum_{n=1}^{N'_m}\sum_{k=1}^{K}\sum_{v=1}^{V} w_{2,n}^{v}\beta_{k,v}^2 - \frac{1}{K}\sum_{r=1}^{R'_m}\sum_{k=1}^{K}\sum_{v'=1}^{v'} x_{2,r}^{v'}\lambda_{k,v'}^2 + \log(c'_m). \tag{21}$$

After removing the constant term in formula (21), the log of the total likelihood function of the two-phase for all patients can be obtained from the following formula:

$$\log\left(\prod_{m=1}^{M} L\left(w_1, w_2, x_1, x_2 \mid \mu_m, \Delta\mu, i_m, \beta, \lambda\right)\right) \sim -\sum_{m=1}^{M}\left(\frac{N_m + R_m}{K + N_m + R_m}\sum_{k=1}^{K}\mu_{m,k}^2 + \frac{N'_m + R'_m}{K + N'_m + R'_m}\sum_{k=1}^{K}\left(\mu_{m,k} + i_m\Delta\mu_k\right)^2\right). \tag{22}$$

To maximize the log of the total likelihood function, formula (23) is obtained as follows:

$$\mu_{m,k} = \frac{-\left(N'_m + R'_m\right) \times \left(K + N_m + R_m\right) \times i_m \times \Delta\mu_k}{\left(N_m + R_m\right) \times \left(K + N'_m + R'_m\right) + \left(N'_m + R'_m\right) \times \left(K + N_m + R_m\right)}. \tag{23}$$

Under the given constraints, $\sum_{k=1}^{K} \Delta\mu_k = 1$, the following formulas are obtained:

$$\Delta\mu_k = \frac{1}{\phi_k \times \sum_{k'=1}^{K} (1/\phi_{k'})}, \tag{24}$$

$$\phi_k = \sum_{m=1}^{M} i_m^2 \left(\frac{(N'_m + R'_m) \times (k + N_m + R_m)}{(N_m + R_m) \times (k + N'_m + R'_m) + (N'_m + R'_m) \times (k + N_m + R_m)} + \frac{N'_m + R'_m}{k + N'_m + R'_m} \right). \tag{25}$$

A body condition transition diagram is shown in Figure 3. Therefore, we can obtain the patient's body condition at discharge.

We hypothesize $\widehat{T}_m$ as the predicted survival time for patient m after the second phase. $\widehat{f}_m$ is a function for predicting survival time analysis $\widehat{T}_m$.

We predict the probability $(\widehat{p}_m)$ for patient m to survive with

$$\widehat{p}_m(\mu_m + (i_m + j_m)\Delta\mu, h_m, \eta) = \frac{1.01}{1 + e^{\sum_{k=1}^{K} \left(1/\left(1+e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}\right)\right)\left(\sum_{v=1}^{V} h_{m,v}\eta_{v,k} + \sum_{v'=1}^{v'} h_{m,v'}\eta_{v',k} \Big/ \left(\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}\right)\right)}} - 0.005, \tag{26}$$

where $h_{m,v}$ is the indicator function. When the notes of patient m contain the v-th bigram word of the notes dictionary, the $h_{m,v}$ value is 1; otherwise, the $h_{m,v}$ value is 0. $h_{m,v'}$ is the indicator function. When the measurement of patient m contains the v'-th measurement of the measurement dictionary, the $h_{m,v'}$ value is 1; otherwise, the $h_{m,v'}$ value is 0. $\eta_{v,k}$ is the regression coefficient. The v-th bigram word of the dictionary represents a danger factor when $\eta_{v,k} > 0$, which indicates a shorter survival time. Otherwise, the v-th bigram word of the notes dictionary represents a protective factor when $\eta_{v,k} < 0$, which indicates a longer survival time. $\eta_{v',k}$ is the regression coefficient. The v'-th measurement of the measurement dictionary represents a danger factor when $\eta_{v',k} > 0$, which indicates a shorter survival time. Otherwise, the v'-th measurement of the measurement dictionary represents a protective factor when $\eta_{v',k} < 0$, which indicates a longer survival time.

Then, we can obtain formula (27) for $\widehat{f}_m$ and formula (28) for $\widehat{T}_m$:

$$\widehat{f}_m = \sum_{k=1}^{K} \left(\frac{1}{1 + e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{\sum_{v=1}^{V} h_{m,v}\eta_{v,k} + \sum_{v'=1}^{v'} h_{m,v'}\eta_{v',k}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}}, \tag{27}$$

$$\widehat{T}_m = \frac{\widehat{p}_m}{1 - \widehat{p}_m} = \frac{\left(1.01/\left(1 + e^{\widehat{f}_m}\right)\right) - 0.005}{1.005 - \left(1.01/\left(1 + e^{\widehat{f}_m}\right)\right)}. \tag{28}$$

2.3.3. Object Optimization. To minimize the difference between f_m and $\widehat{f}_m$, we define Diff as follows:

$$\begin{aligned} \text{Diff} = \sum_{m=1}^{M} \frac{1}{1 + e^{-f_m}} \times \left(f_m - \widehat{f}_m\right)^2 = \sum_{m=1}^{M} \frac{1}{1 + e^{-f_m}} \\ \times \left(f_m - \sum_{k=1}^{K} \left(\frac{1}{1 + e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{\sum_{v=1}^{V} h_{m,v}\eta_{v,k} + \sum_{v'=1}^{v'} h_{m,v'}\eta_{v',k}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}} \right)^2. \end{aligned} \tag{29}$$

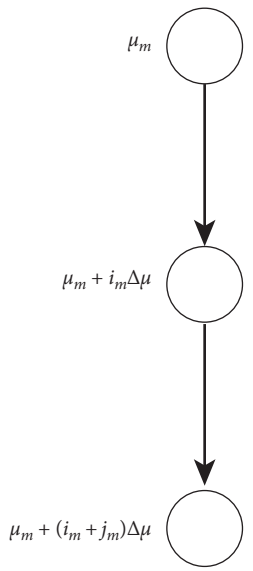


Figure 3: Body condition transition diagram.

To find the optimal η to minimize Diff, we obtain

$$\arg\min_{\eta} \sum_{m=1}^{M} \frac{1}{1+e^{-f_m}} \times \left(f_m - \sum_{k=1}^{K} \left(\frac{1}{1+e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{\sum_{v=1}^{V} h_{m,v}\eta_{v,k} + \sum_{v'=1}^{v'} h_{m,v'}\eta_{v',k}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}} \right)^2. \tag{30}$$

Then, we obtain the gradient update as the following formulas:

$$\begin{aligned} \frac{\partial \text{Diff}}{\eta_{v,k}} = \sum_{m=1}^{M} \frac{-2}{1+e^{-f_m}} \times & \left(f_m - \sum_{k=1}^{K} \left(\frac{1}{1+e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{\sum_{v=1}^{V} h_{m,v}\eta_{v,k} + \sum_{v'=1}^{v'} h_{m,v'}\eta_{v',k}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}} \right) \\ & \times \left(\sum_{k=1}^{K} \left(\frac{1}{1+e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{h_{m,v}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}} \right), \end{aligned} \tag{31}$$

$$\begin{aligned} \frac{\partial \text{Diff}}{\eta_{v',k}} = \sum_{m=1}^{M} \frac{-2}{1+e^{-f_m}} \times & \left(f_m - \sum_{k=1}^{K} \left(\frac{1}{1+e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{\sum_{v=1}^{V} h_{m,v}\eta_{v,k} + \sum_{v'=1}^{v'} h_{m,v'}\eta_{v',k}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}} \right) \\ & \times \left(\sum_{k=1}^{K} \left(\frac{1}{1+e^{-\mu_{m,k}-(i_m+j_m)\Delta\mu_k}} \right) \frac{h_{m,v'}}{\sum_{v=1}^{V} h_{m,v} + \sum_{v'=1}^{v'} h_{m,v'}} \right). \end{aligned} \tag{32}$$

3. Experiments

3.1. Experimental Design

3.1.1. Evaluation Tasks

Task 1: Evaluating the performance of the TiCTM model with different topics.

Task 2: Performance comparisons of TiCTM with LDA and linear regression models.

Task 3: F1 comparisons for different survival days with TiCTM ($K = 5$).

Task 4: Comprehensive performance comparisons of TiCTM (different topics) with LDA ($K = 5$, 10) and linear regression models using ROC.

3.1.2. Baselines. In this experiment, we use two methods as our baselines: LDA [14] and linear regression. For the linear regression model, the meaning of each notation is shown in Table 5.

$\widehat{T}_m^{\mathrm{LR}}$ represents the predicted survival time with linear regression. $\widehat{T}_m^{\mathrm{LR}}$ is defined as

$$\widehat{T}_m^{\mathrm{LR}} = \frac{\left(1.01/\left(1+e^{\widehat{f}_m^{\mathrm{LR}}}\right)\right) - 0.005}{1.005 - \left(1.01/\left(1+e^{\widehat{f}_m^{\mathrm{LR}}}\right)\right)}, \tag{33}$$

where $\widehat{f}_m^{\mathrm{LR}}$ is a function for predicting survival time analysis $\widehat{T}_m^{\mathrm{LR}}$. $\widehat{f}_m^{\mathrm{LR}}$ is defined as

$$\widehat{f}_m^{\mathrm{LR}} = \sum_{k=1}^{K}\left(\sum_{v=1}^{V} a_{v,k}h_{m,v} + \sum_{v'=1}^{v'} a_{v',k}h_{m,v'}\right) + b, \tag{34}$$

where a and b are regression coefficients.

The objective of survival time prediction for sepsis patients is to minimize the loss function between f_m and $\widehat{f}_m^{\mathrm{LR}}$:

$$\min \mathrm{Loss}_{\mathrm{LR}} = \min \sum_{m=1}^{M}\left(f_m - \widehat{f}_m^{\mathrm{LR}}\right)^2. \tag{35}$$

Then, we obtain the gradient update as formulas (36), (37), and (38):

$$\frac{\partial \mathrm{Loss}_{\mathrm{LR}}}{\partial a_{v,k}} = -2\sum_{m=1}^{M}\left(f_m - \widehat{f}_m^{\mathrm{LR}}\right)h_{m,v}, \tag{36}$$

$$\frac{\partial \mathrm{Loss}_{\mathrm{LR}}}{\partial a_{v',k}} = -2\sum_{m=1}^{M}\left(f_m - \widehat{f}_m^{\mathrm{LR}}\right)h_{m,v'}, \tag{37}$$

$$\frac{\partial \mathrm{Loss}_{\mathrm{LR}}}{\partial b} = -2\sum_{m=1}^{M}\left(f_m - \widehat{f}_m^{\mathrm{LR}}\right). \tag{38}$$

3.1.3. Evaluation Criteria. To evaluate the performance of the proposed method, we use a 3-fold cross-validation scheme. Evaluation metrics, such as recall, F1, and FPR, are adopted in this paper. They are defined as follows:

Recall (TPR) = TP/(TP + FN).

Precision = TP/(TP + FP).

Accuracy = (TP + TN)/(TP + TN + FP + FN).

F1 = precision × recall × 2/(precision + recall).

FPR = FP/(FP + TN), where TP indicates true positive, which means predicting a survival time less than or equal to a given time, while the true survival time is less than or equal to a given time; FP indicates false positive, which means predicting a survival time less than or equal to a given time but the true survival time is greater than a given time; TN indicates true negative, which means predicting a survival time greater than a given time, while the true survival time is greater than a given time; and FN indicates false negative, which means predicting a survival time greater than a given time but the true survival time is less than or equal to a given time.

3.1.4. Dataset. The dataset used in the experiments is the public Medical Information Mart for Intensive Care (MIMIC-III) [26]. MIMIC-III has been widely used by 845 publications as of the end of August 2019 [27]. The version of MIMIC is MIMIC-III v1.4, which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. The data spanned from June 2001 to October 2012. We used a dataset of 2,487 deceased adult (age > 14) sepsis patient records from MIMIC-III. The data processing flowchart is shown in Figure 4.

Patient features include text data, class data, and numerical data.

(1) Text data processing: clinical notes are also text data. This paper considers the clinical notes as a set of bigrams. We calculate the TF-IDF value of every bigram after removing the stopwords. Then, we sort every bigram according to TF-IDF and select the top 3,000 bigrams as the words for the dictionary.

(2) Numerical data processing: to calculate the mean and standard deviation (std) for numerical data, we divide it into five intervals: $(-\infty, \mathrm{mean} - 1.5 \cdot \mathrm{std})$, $[\mathrm{mean} - 1.5 \cdot \mathrm{std}, \mathrm{mean} - 0.5 \cdot \mathrm{std})$, $[\mathrm{mean} - 0.5 \cdot \mathrm{std}, \mathrm{mean} + 0.5 \cdot \mathrm{std})$, $[\mathrm{mean} + 0.5 \cdot \mathrm{std}, \mathrm{mean} + 1.5 \cdot \mathrm{std})$, and $[\mathrm{mean} + 1.5 \cdot \mathrm{std}, +\infty)$. Each interval is used as a word in the measurement dictionary.

(3) Class data processing: each class of data is used as a word in the measurement dictionary.

3.2. Experimental Results

3.2.1. Evaluating the Performance of the TiCTM Model with Different Topics. To evaluate the performance of our proposed TiCTM model, we choose different topics, that is, $K = 5$, 10, 15, 20, to predict the survival time for adult sepsis patients in the ICU. The results are shown in Tables 6–9. In Table 9, we can see that 5 topics ($K = 5$) achieve the best performance for predicting the probability for 30-day survival time. The number of topics is chosen by a 3-fold cross-validation scheme. We use the average value of 3-fold to select the best number of topics. It can be seen that the best number of topics for sepsis patients is 5 topics. The $F1$ score of 5 topics improved by 3.55% (in-hospital death), 3.88% (7 days), 2.56% (14 days), and 1.99% (30 days) compared to 20 topics. The possible reason is that there are many words in the dictionary, and the combination of topics and bigrams increases sharply when the number of topics increases. This situation will lead to model overfitting, and the $F1$ score will decrease.

Table 5: Notations for the linear regression model.

Notation	Description
M	Total number of patients
$h_{m,v}$	Indicator function for notes
h_{m,v_l}	Indicator function for measurements
T_m	Real survival time for patient m after phase *I-1*
$\widehat{T}_m^{LR}$	Predicted survival time for patient m for the linear regression model
a	Regression coefficient
b	Regression coefficient
f_m	Function for computing real survival time
$\widehat{f}_m^{\mathrm{LR}}$	Function for computing predicted survival time
$\mathrm{Loss}_{\mathrm{LR}}$	The loss function value between f_m and $\widehat{f}_m^{\mathrm{LR}}$

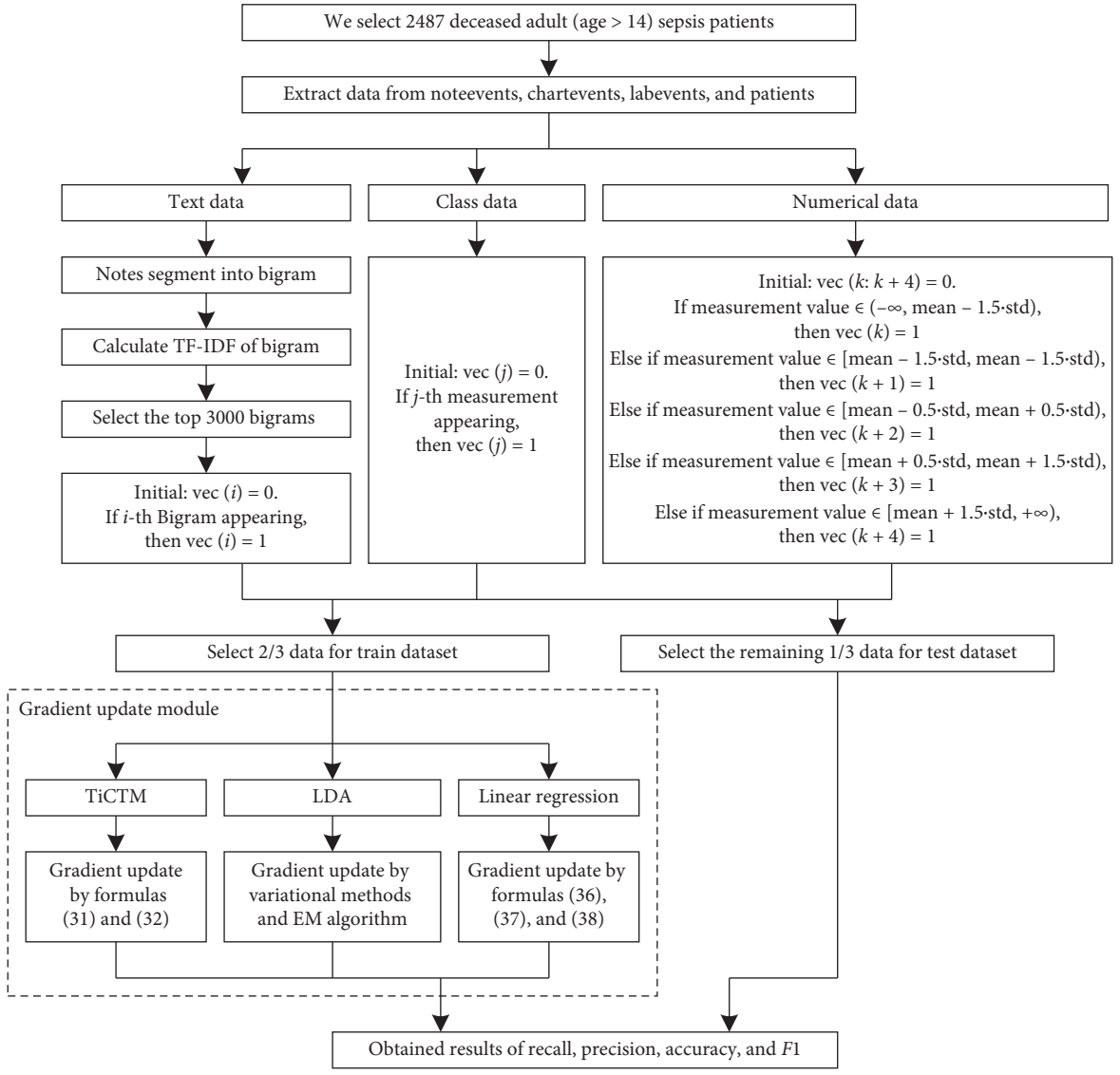


Figure 4: Data processing flowchart.

3.2.2. Performance Comparisons of TiCTM with LDA and Linear Regression Model. Performance comparisons of TiCTM ($K = 5$, 10) with LDA ($K = 5$, 10) and the linear regression model are shown in Tables 10–13. The performance of our proposed TiCTM outperforms LDA and the linear model overall in terms of recall, precision, accuracy, and $F1$-measure. When we consider sequential information for predicting the survival time of sepsis patients, the prediction accuracy increases.

Table 6: In-hospital death.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.6492**	**0.5773**	**0.7186**	**0.6097**
TiCTM ($K = 10$)	0.6190	0.5747	0.7138	0.5953
TiCTM ($K = 15$)	0.5901	0.5615	0.7027	0.5746
TiCTM ($K = 20$)	0.5895	0.5611	0.7019	0.5742

Table 7: Survival 7 days after discharge.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.8174**	**0.5360**	**0.6725**	**0.6447**
TiCTM ($K = 10$)	0.7743	0.5250	0.6582	0.6225
TiCTM ($K = 15$)	0.7537	0.5169	0.6502	0.6111
TiCTM ($K = 20$)	0.7421	0.5147	0.6478	0.6059

Table 8: Survival 14 days after discharge.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.8489**	**0.5263**	**0.6443**	**0.6472**
TiCTM ($K = 10$)	0.8179	0.5223	0.6379	0.6351
TiCTM ($K = 15$)	0.8032	0.5209	0.6355	0.6294
TiCTM ($K = 20$)	0.7857	0.5168	0.6311	0.6216

Table 9: Survival 30 days after discharge.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.8786**	0.5395	0.6276	**0.6655**
TiCTM ($K = 10$)	0.8562	**0.5419**	**0.6288**	0.6608
TiCTM ($K = 15$)	0.8414	0.5372	0.6228	0.6532
TiCTM ($K = 20$)	0.8275	0.5328	0.6164	0.6456

Table 10: In-hospital death.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.6492**	**0.5773**	**0.7186**	**0.6097**
TiCTM ($K = 10$)	0.6190	0.5747	0.7138	0.5953
LDA ($K = 5$)	0.3421	0.6753	0.7174	0.4528
LDA ($K = 10$)	0.4284	0.6414	0.7202	0.5128
Linear regression	0.4771	0.4723	0.6407	0.4736

Table 11: Survival 7 days after discharge.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.8174**	0.5360	0.6725	**0.6447**
TiCTM ($K = 10$)	0.7743	0.5250	0.6582	0.6225
LDA ($K = 5$)	0.5295	**0.6258**	**0.7098**	0.5728
LDA ($K = 10$)	0.5662	0.5641	0.6800	0.5645
Linear regression	0.5768	0.472	0.6085	0.5174

3.2.3. F1 Comparisons for Different Survival Days with TiCTM ($K = 5$). When predicting in-hospital death, 7-day survival, 14-day survival, and 30-day survival, the F1-score will increase, as shown in Table 14. For longer survival times, the patient's condition remains steady. Then, the clinician intervenes less. Therefore, the F1 score is more accurate for predicting longer survival times.

3.2.4. Comprehensive Performance Comparisons Using ROC. Figure 5 presents comprehensive performance comparisons of TiCTM (different topics) with LDA ($K = 5, 10$) and the linear regression model using ROC. In Figure 5, we can see that the performance of our proposed TiCTM is better than that of the classic LDA and linear regression models.

Table 12: Survival 14 days after discharge.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.8489**	0.5263	0.6443	**0.6472**
TiCTM ($K = 10$)	0.8179	0.5223	0.6379	0.6351
LDA ($K = 5$)	0.5791	**0.6049**	**0.6896**	0.5915
LDA ($K = 10$)	0.6060	0.5512	0.6558	0.5769
Linear regression	0.5916	0.4700	0.5826	0.5223

Table 13: Survival 30 days after discharge.

Methods	Recall	Precision	Accuracy	$F1$
TiCTM ($K = 5$)	**0.8786**	0.5395	0.6276	**0.6655**
TiCTM ($K = 10$)	0.8562	0.5419	0.6288	0.6608
LDA ($K = 5$)	0.6353	**0.6018**	**0.6661**	0.6179
LDA ($K = 10$)	0.6593	0.5594	0.6339	0.6045
Linear regression	0.6287	0.4968	0.5707	0.5530

Table 14: F1 score comparisons for TiCTM ($k = 5$).

Survival time prediction	$F1$
In-hospital death	0.6097
Survival 7 days after discharge	0.6447
Survival 14 days after discharge	0.6472
Survival 30 days after discharge	**0.6655**

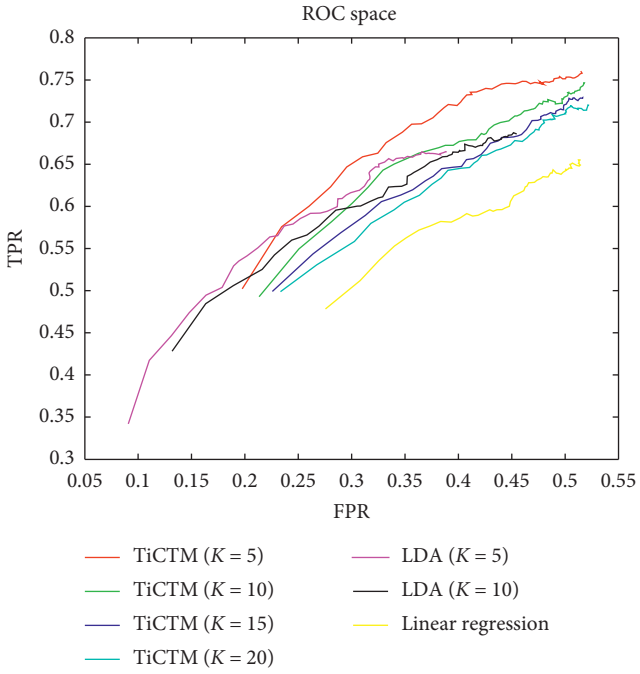


Figure 5: Comprehensive performance comparisons using ROC.

4. Conclusions

In this paper, we propose a time-critical topic model (TiCTM), which combines a patient's measurement with notes, to predict the survival time for adult sepsis patients in the ICU. We consider useful sequential information for predicting the survival time of sepsis patients, thereby increasing the prediction accuracy. Our experimental results show that the proposed TiCTM has the best performance when predicting the probability for 30-day survival time using 5 topics. In addition, the performance of our proposed TiCTM is better than that of the classic LDA and linear regression models. In the future, our study will focus on the explainable machine learning model for predicting survival time in the ICU [28–31].

Data Availability

The data used to support the findings of this study are available from https://mimic.physionet.org.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are grateful to the Institute of Big Data and Artificial Intelligence in Medicine, Taizhou University, for providing computing resources. This research was funded by the Key Projects of Zhejiang Province's Educational Science Planning (2017SB068), the Science and Technology Program of Taizhou (1803gy07), the Humanities and Social Science Project of the Chinese Ministry of Education (20YJAZH033), and the National Natural Science Foundation of China (61976149).

References

[1] L. Zhou, J. Cui, J. Lu, B. Wee, and J. Zhao, "Prediction of survival time in advanced cancer: a prognostic scale for Chinese patients," *Journal of Pain and Symptom Management*, vol. 38, pp. 578–586, 2009.

[2] M. Zhou, L. O. Hall, D. B. Goldgof, R. A. Gatenby, and R. J. Gillies, "A texture feature ranking model for predicting

survival time of brain tumor patients," in *Proceedings of 2013 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE Computer Society, Manchester, United kingdom, October 2013.

[3] M. Zhou, L. O. Hall, D. B. Goldgof, R. A. Gatenby, and R. J. Gillies, "Exploring brain tumor heterogeneity for survival time prediction," in *Proceedings of 22nd International Conference on Pattern Recognition*, Institute of Electrical and Electronics Engineers Inc, Stockholm, Sweden, August 2014.

[4] K. B. Ahmed, L. O. Hall, R. Liu, R. A. Gatenby, and R. J. Gillies, "Neuroimaging based survival time prediction of gbm patients using cnns from small data," in *Proceedings of 2019 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE Computer Society, Bari, Italy, October 2019.

[5] A. N. V. Dehkordi, A. Kamali-Asl, N. Wen, T. Mikkelsen, I. J. Chetty, and H. Bagher-Ebadian, "DCE-MRI prediction of survival time for patients with glioblastoma multiforme: using an adaptive neuro-fuzzy-based model and nested model selection technique," *NMR in Biomedicine*, vol. 30, Article ID e3739, 2017.

[6] M. Singer, C. S. Deutschman, C. W. Seymour et al., "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[7] C. M. Torio and R. M. Andrews, "National inpatient hospital costs: the most expensive conditions by payer," 2011, https://www.ncbi.nlm.nih.gov/books/NBK169005/.

[8] D. McPherson, C. Griffiths, M. Williams et al., "Sepsis-associated mortality in England: an analysis of multiple cause of death data from 2001 to 2010," *BMJ Open*, vol. 3, pp. 1–7, 2013.

[9] L. Ou, J. Chen, K. Hillman et al., "The impact of post-operative sepsis on mortality after hospital discharge among elective surgical patients: a population-based cohort study," *Critical Care*, vol. 21, p. 34, 2017.

[10] J. C. Yébenes, J. C. Ruiz-Rodriguez, R. Ferrer et al., "SOCMIC (Catalonian critical care society) sepsis working group. epidemiology of sepsis in catalonia: analysis of incidence and outcomes in a European setting," *Annals of Intensive Care*, vol. 7, p. 19, 2017.

[11] W. A. Knaus, D. P. Wagner, E. A. Draper et al., "The Apache III prognostic system," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[12] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, 1993.

[13] C. P. Subbe, A. Slater, D. Menon, and L. Gemmell, "Validation of physiological scoring systems in the accident and emergency department," *Emergency Medicine Journal*, vol. 23, no. 11, pp. 841–845, 2006.

[14] J.-L. Vincent, R. Moreno, J. Takala et al., "The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[15] Z. Zhang and Y. Hong, "Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: the use of electronic healthcare records with LASSO regression," *Oncotarget*, vol. 8, no. 30, pp. 49637–49645, 2017.

[16] J.-L. Vincent and R. Moreno, "Clinical review: scoring systems in the critically ill," *Critical Care*, vol. 14, no. 2, p. 207, 2010.

[17] A. B. Nielsen, H. C. Thorsen-Meyer, K. Belling et al., "Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the danish national patient registry and electronic patient records," *The Lancet Digital Health*, vol. 1, no. 2, pp. e78–89, 2019.

[18] J.-L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, "Sepsis definitions: time for change," *The Lancet*, vol. 381, no. 9868, pp. 774-775, 2013.

[19] T. J. Iwashyna, E. W. Ely, D. M. Smith, and K. M. Langa, "Long-term cognitive impairment and functional disability among survivors of severe sepsis," *JAMA*, vol. 304, no. 16, pp. 1787–1794, 2010.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[21] Z. Gou, Z. Huo, Y. Liu, and Y. Yang, "A method for constructing supervised topic model based on term frequency-inverse topic frequency," *Symmetry*, vol. 11, no. 12, p. 1486, 2019.

[22] L. W. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," in *Proceedings of the American Medical Informatics Association Annual Symposium*, pp. 505–511, Chicago, IL, USA, November 2012.

[23] M. Ghassemi, T. Naumann, F. Doshi-Velez et al., "Unfolding physiological state: mortality modelling in intensive care units," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Publisher: Association for computing Machinery, New York, NY, USA, August 2014.

[24] P. Vairamani, "Mortality prediction in ICU patients. Report of Project for CSE8803 Big data analytics for healthcare," *Spring*, 2016.

[25] Y. Zhang, R. Jiang, and L. Petzold, "Survival topic models for predicting outcomes for trauma patients," in *Proceedings of 33rd IEEE International Conference on Data Engineering*, IEEE Computer Society, San Diego, CA, United states, April 2017.

[26] A. E. Johnson, T. J. Pollard, L. Shen et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, Article ID 160035, 2016.

[27] Z. Shi, W. Zuo, S. Liang, X. Zuo, L. Yue, and X. Li, "IDDSAM: an integrated disease diagnosis and severity assessment model for intensive care units," *IEEE Access*, vol. 8, pp. 15423–15435, 2020.

[28] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: an overview of interpretability of machine learning," in *Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics*, Institute of Electrical and Electronics Engineers Inc, Turin, Italy, October 2018.

[29] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, vol. 9, no. 4, Article ID e1312, 2019.

[30] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): towards medical XAI," Article ID 07374, 2019, https://arxiv.org/abs/1907.07374.

[31] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

Hindawi
Scientific Programming
Volume 2020, Article ID 3878153, 13 pages
https://doi.org/10.1155/2020/3878153

Research Article

Multiview Active Learning for Scene Classification with High-Level Semantic-Based Hypothesis Generation

Tuozhong Yao,[1] Wenfeng Wang ,[1,2] Yuhong Gu,[3] and Qiuguo Zhu[4]

[1]*School of Electronic and Information Engineering, Ningbo University of Technology, Ningbo 315211, China*
[2]*School of Electronic and Electrical Engineering, Shanghai Institute of Technology, Shanghai 200235, China*
[3]*Shihezi Medical School, Shihezi 832000, China*
[4]*State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310007, China*

Correspondence should be addressed to Wenfeng Wang; wangwenfeng@nimte.ac.cn

Received 7 February 2020; Revised 23 April 2020; Accepted 18 August 2020; Published 1 September 2020

Academic Editor: Chenxi Huang

Multiview active learning (MVAL) is a technique which can result in a large decrease in the size of the version space than traditional active learning and has great potential applications in large-scale data analysis. This paper made research on MVAL-based scene classification for helping the computer accurately understand diverse and complex environments macroscopically, which has been widely used in many fields such as image retrieval and autonomous driving. The main contribution of this paper is that different high-level image semantics are used for replacing the traditional low-level features to generate more independent and diverse hypotheses in MVAL. First, our algorithm uses different object detectors to achieve local object responses in the scenes. Furthermore, we design a cascaded online LDA model for mining the theme semantic of an image. The experimental results demonstrate that our proposed theme modeling strategy fits the large-scale data learning, and our MVAL algorithm with both high-level semantic views can achieve significant improvement in the scene classification than traditional active learning-based algorithms.

1. Introduction

Scene classification is defined as using a computer to understand the class of an image scene. The related research studies can be roughly divided into two branches: some focus on fast holistic scene perception based on visual psychology and physiology [1, 2], while others build the statistical models through local image analysis to understand the scene, which is also the main developing tendency [3–5]. There have been many methods for image representation in the past two decades, which is a key step for scene classification. Low-level features such as color, texture, and edge have been widely used to represent the local regions of an image. Some researchers trained object detectors to achieve high-level semantics such as object's class, size, and shape for more accurate image representation [6, 7]. Prevailing statistical models are bag-of-words (BoW) and related theme statistical models. These models reduce the gap between the low-level features and high-level semantics by mining the hidden themes from local image regions such as pLSA [8] and LDA [9]. Other new scene statistical models [10–12] were proposed for more accurate object recognition in the scene. However, these mentioned models above mainly focus on the occurrence of the image semantics, and the spatial semantic correlations between different image regions are usually ignored.

For mining the spatial context information from an image, some researchers considered the information interaction between different spatial pyramid levels [13–15], and how to build reasonable attention mechanisms also can lead to significant improvement for scene classification. These methods used deep neural networks, and their large-scale network parameter estimation tasks usually lead to much higher computational complexity than nondeep learning based methods.

Active learning ranks the unlabeled samples iteratively and only selects the samples with high uncertainty or which cause great ambiguity for the classifier. In PAC learning theory, compared with traditional passive learning, it can exponentially reduce its sample complexity to $O(\log(1/\varepsilon))$ in the feature space for learning a classifier with expectation classification error ε [16–18], which has good potential of wide application in large-scale data leaning. However, most of the traditional active learning algorithms' lack of diversity of the hypotheses is generated usually by low-level image features, which affects their performances. This paper proposed a MVAL-based scene classification algorithm, which uses different high-level semantics as its views and can realize a decrease in more than a half size of the version space, and it is more efficient than both single-hypothesis-based and committee-based active learning [19].

2. Materials and Methods

2.1. Proposed Algorithm. The flowchart of our proposed algorithm is illustrated in Figure 1. Our algorithm uses different high-level semantics as its views to generate the corresponding hypotheses. First, object detectors are trained to achieve the responses of different object classes in image regions. Furthermore, we design a cascaded online LDA (CO-LDA) as a secondary view for achieving more accurate image representation. Finally, a fine-tuned MVAL algorithm is utilized with both two high-level image semantics as its views for classifying the scene of an image.

2.2. Object Semantic-Based Image Representation. Our object semantic-based image representation is illustrated in Figure 2.

First, multiple object objectors are used to achieve the local object response maps. Second, these maps are decomposed into three spatial pyramid levels, and the maximal object responses are computed in image blocks in each spatial level, which is annotated as red blocks in Figure 2. Finally, an object response histogram is computed, which can effectively reduce the influence of object response error in the whole image. For generating the object response, a latent SVM-based detector [7] is applied for recognizing the object classes with bulk type such as car and pedestrian. Another geometric context-based detector [6] is utilized for recognizing the object classes with different textures such as tree, sky, and building.

2.3. Theme Semantic-Based Image Representation. For satisfying the dynamic update of an active learning training set, an online LDA model [20] based on stochastic gradient descent strategy is used. It adds new samples sequentially, and old samples have been no longer stored, which can achieve efficient and accurate parameter estimation in large-scale data training.

Online LDA computes the posterior probability distribution $p(\theta, z, w, \beta \mid \alpha, \eta)$ of the hidden nodes based on observed samples. It actually uses variational inference to estimate the maximum likelihood of $p(w \mid \alpha, \eta)$ based on α and η. Three variational parameters ϕ, γ, and λ follow the distributions: $\phi \sim \text{multinomial}(\varepsilon)$, $\gamma \sim \text{Dirichlet}(\varepsilon)$, and $\lambda \sim \text{Dirichlet}(\varepsilon)$. The variational distribution follows

$$q\left(\beta_{1:k}, \theta_{1:M}, z_{1:M} \mid \lambda, \gamma, \phi\right) = \prod_{i=1}^{k} \text{Dir}\left(\beta_i \mid \lambda_i\right) \cdot \prod_{d=1}^{M} q_d\left(\theta_d, z_d \mid \phi_d, \gamma_d\right). \tag{1}$$

The optimal (γ, ϕ) is solved by maximizing the lower bound in the following equation:

$$\log p(w \mid \alpha, \eta) \geq L(w, \phi, \gamma, \lambda) = E_q[\log p(\theta, z, w, \beta \mid \alpha, \eta)] - E_q[\log q(\theta, z, \beta)], \tag{2}$$

where E_q denotes the conditional mathematical expectation. Maximizing the lower bound $L(w, \phi, \gamma, \lambda)$ is equivalent to minimizing KL divergence of $q(\theta, z, \beta \mid \gamma, \phi)$ and $p(\theta, z, \beta \mid w, \alpha, \eta)$:

$$\log p(w \mid \alpha, \eta) = L(w, \phi, \gamma, \lambda) + KL(q(\theta, z, \beta) \| p(\theta, z, \beta \mid w, \alpha, \eta)), \tag{3}$$

where $L(w, \phi, \gamma, \lambda)$ is factorized as follows:

$$L(w, \phi, \gamma, \lambda) = \sum_{d} \left\{ E_q\left[\log p\left(w_d \mid \theta_d, z_d, \beta\right)\right] + E_q\left[\log p\left(z_d \mid \theta_d\right)\right] - E_q[\log q(z_d)] + E_q\left[\log p\left(\theta_d \mid \alpha\right)\right] - E_q[\log q(\theta_d)] + \frac{\left(E_q[\log p(\beta \mid \eta)] - E_q[\log q(\beta)]\right)}{M} \right\}. \tag{4}$$

Equation (4) can be transformed into formula (5). In equation (5), n_{dw} denotes the frequency that word w occurs in text d. $l(n_d, \phi_d, \gamma_d, \lambda)$ reflects the contribution of d for the lower bound, which is iteratively optimized by a coordinate ascent algorithm:

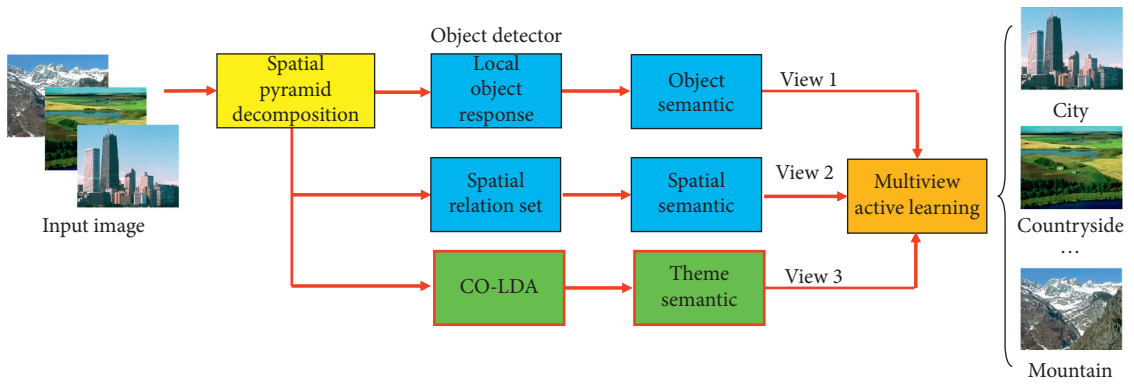


Figure 1: The flowchart of our scene classification algorithm.

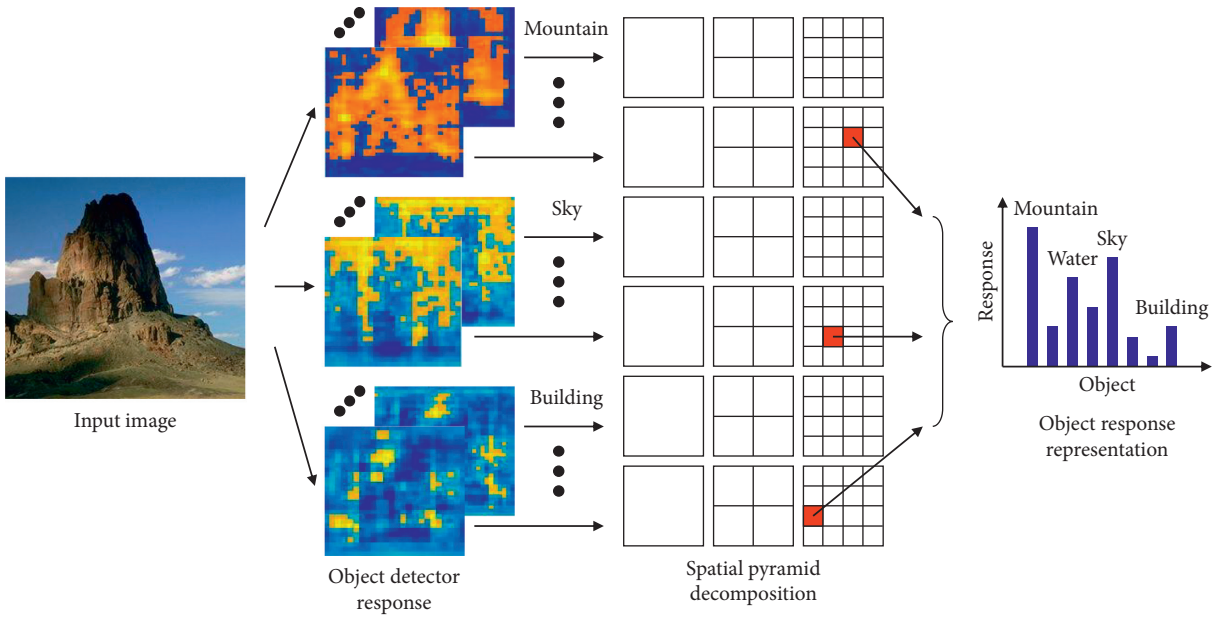


Figure 2: The flowchart of object semantic-based image representation.

$$
\begin{aligned}
L(w,\phi,\gamma,\lambda) = {} & \sum_{d}\sum_{w} n_{dw} \sum_{k} \phi_{dwk} E_q[\log \theta_{dk}] + E_q[\log \beta_{kw}] - \log \phi_{dwk} \\
& - \log \Gamma\left(\sum_{k} \gamma_{dk}\right) + \sum_{k} (\alpha - \gamma_{dk}) E_q[\log \theta_{dk}] + \log \Gamma(\gamma_{dk}) \\
& + \frac{\left(\sum_k - \log \Gamma\left(\sum_k \lambda_{kw}\right) + \sum_w (\eta - \lambda_{kw}) E_q[\log \beta_{kw}] + \log \Gamma(\lambda_{kw})\right)}{M} \\
& + \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \frac{(\log \Gamma(N\eta) - W \log \Gamma(\eta))}{M}, \\
= {} & \sum_{d} l(n_d, \phi_d, \gamma_d, \lambda).
\end{aligned}
\tag{5}
$$

ϕ_{dwk}in equation (5) is iteratively solved:

$$\begin{aligned}\phi_{dwk} &\propto \exp\left(E_q[\log(\theta_{dk})] + E_q[\log(\beta_{kw})]\right),\\ E_q[\log(\theta_{dk})] &= \Psi(\gamma_{dk}) - \Psi\left(\sum_{i=1}^{K}\gamma_{di}\right),\\ E_q[\log(\beta_{kw})] &= \Psi(\lambda_{kw}) - \Psi\left(\sum_{i=1}^{N}\lambda_{ki}\right),\end{aligned} \tag{6}$$

where digamma functionΨ is the first-order derivative of function Γ. γ_{dk} and λ_{kw} are iteratively solved in the following way: $\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}$,$\lambda_{kw} = \eta + \sum_d n_{dw}\phi_{dwk}$.

Whent^{th} vector of word frequency n_t is observed, we keep λ unchanged and update the local optimal solution of γ_t and ϕ_t in E step. In M step, ϕ_t and λ from last iteration are both used to update λ:

$$\begin{aligned}\lambda &= (1-\rho_t)\lambda + \rho_t\tilde{\lambda},\\ \rho_t &= (\tau + t)^{-k}.\end{aligned} \tag{7}$$

$\tilde{\lambda}$ in formula (7) is solved as follows:

$$\tilde{\lambda}_{kw} = \eta + \frac{M}{S}\sum_s n_{\text{ts}k}\phi_{\text{ts}kw}, \tag{8}$$

where n_{ts} is s^{th} text in each batch text set, M is the number of the training text set, and $S$is the size of each batch text set. Hyperparameters α and η are updated by the Newton–Rapson method: $\alpha \longleftarrow \alpha - \rho_t\tilde{\alpha}(\gamma_t)$ and $\eta \longleftarrow \eta - \rho_t$ $\tilde{\eta}(\lambda)$. Here, $\tilde{\alpha}(\gamma_t)$ is the product of Hessian matrix and gradient $\nabla_\alpha l$ of the objective function $l(n_d, \phi_d, \gamma_d, \lambda)$. $\tilde{\eta}(\lambda)$ is the product of Hessian matrix and gradient $\nabla_\eta L$ of the objective function $L(\omega, \phi, \gamma, \lambda)$.

Based on online LDA, we proposed the CO-LDA theme model, which is similar with the classic SP-pLSA model in structure for enhancing the spatial correlation between different image regions. The framework of CO-LDA is illustrated in Figure 3. The main difference between CO-LDA and SP-pLSA is that different online LDAs (LDA1, LDA2, and LDA3) are applied in different spatial levels to jointly mine the theme of an image. The main advantage of CO-LDA is that it integrates the spatial correlation of objects in different image resolutions, which further improves the holistic scene understanding. The visual histogram computation in online LDA is the same as the way of object response histogram in Section 2.2, and the theme feature of each spatial block is represented by variational parameter γ of the online LDA model.

Finally, the theme feature γ of the whole image is achieved by concatenating the theme features of different blocks of different spatial pyramid levels:

$$\gamma = \left(w_1 \cdot \gamma_{L_0}\right) \oplus \left(w_2 \cdot \gamma_{L_1}\right) \oplus \left(w_3 \cdot \gamma_{L_2}\right), \tag{9}$$

where γ_{L_i} denotes the theme feature of the corresponding block in L_i^{th} pyramid level, $\oplus$ denotes the linear concatenation between feature vectors, and the weights of different spatial levels are configured as follows: $w_1 = (1/2)$, $w_2 = (1/2)$, and $w_3 = (1/4)$.

2.4. Multiview Active Learning. The MVAL referred in this paper is our previous work [21], which has two improvements in both hypothesis generation and selective sampling. First, boosting-like technique is integrated into MVAL, which uses a similar way of iterative weak classifier optimization, and the current hypothesis is boosted by weighted voting of all the hypotheses from the past queries. Furthermore, an adaptive hierarchical competition sampling is presented. In this sampling strategy, if the number of the contention samples is large, an unsupervised spectral clustering is activated to obtain the coarse spatial distribution of these contention samples in the high-dimensional feature space, and then, a multiview-based batch mode selective sampling is run based on two measures: sample uncertainty and redundancy by solving quadratic programming to determine the queried samples in each cluster.

2.4.1. Hypothesis Generation. If an active learning can select enough number of contention samples, which could improve the hypothesis in each query, the number of unlabeled samples, which are incorrectly classified, will decrease. It is quite similar with boosting technique in weak classifier optimization. The MVAL incorporates the AdaBoost algorithm into our framework to boost the generated hypothesis in each query, and the main flowchart is described in Figure 4.

In Figure 4, a support vector machine (SVM) is used as a base classifier to construct a multiview classifier, which replaces the single-view classifier in AdaBoost, and this multiview classifier in each query can be considered as a weak classifier in each iteration in AdaBoost. The hypothesis of multiview classifier $h^i(x)$ is computed by weighted voting of n SVM base classifiers $v_1, v_2, \ldots, v_n$ whose weights are $\omega_1, \omega_2, \ldots, \omega_n$. Unlike traditional query by boosting, we update the weight of each base classifier in each query and obtain the boosted hypothesis $H^i(x)$ by weighting all the hypotheses from the past queries and not from the current query only.

The detailed process of the MAVL's hypothesis generation based on AdaBoost is as follows:

(a) In iteration $h^t(x_j) = \sum_{f\in\{f_1^t, f_2^t, \ldots f_n^t\}}\omega_i^t f_i^t(x_j)$, weighted voting is used to generate the initial multiview-based hypothesis:

$$h^t\left(x_j\right) = \sum_{f\in\left\{f_1^t, f_2^t, \ldots, f_n^t\right\}} \omega_i^t f_i^t\left(x_j\right), \tag{10}$$

where $f_i^t(x_j)$ is the classification confidence of sample x_j by view i, and ω_i^t denotes the contribution of view i for classification which is determined by the soft classification error rate ε_i^t, which defines how correctly a sample is classified:

$$\varepsilon_i^t = \frac{1}{\left(\sum_{x\in L, y=1} f_i^t(x) - \sum_{x\in L, y=-1} f_i^t(x)\right)}, \tag{11}$$

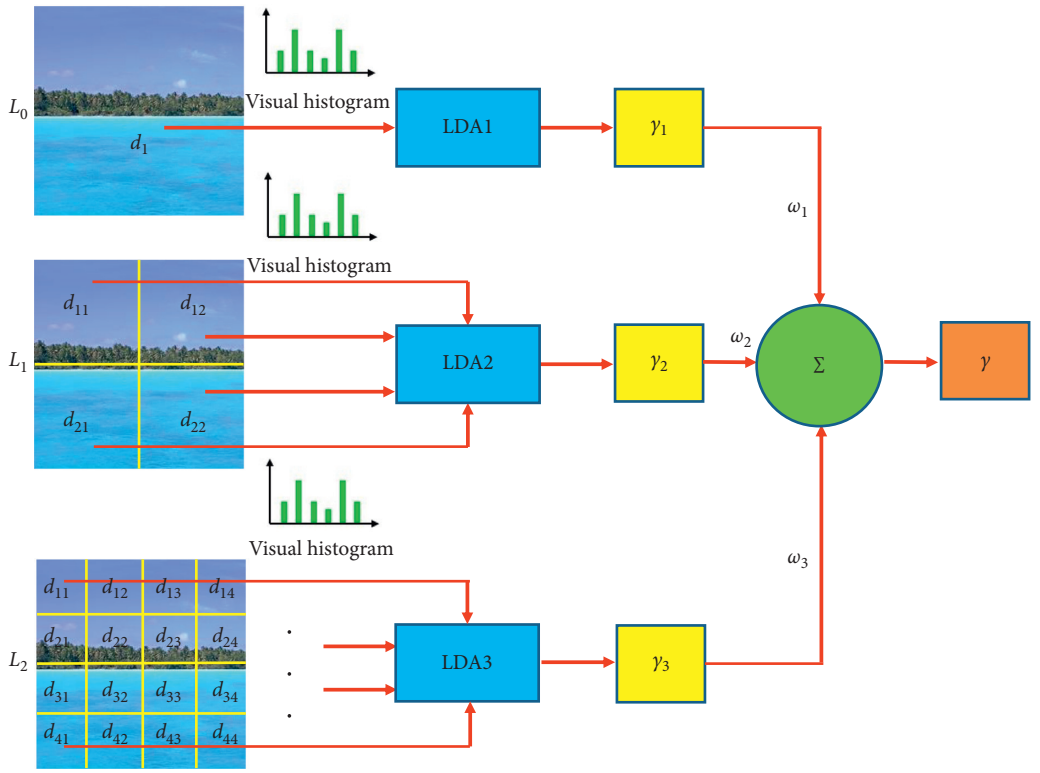


Figure 3: The framework of the CO-LDA theme model.

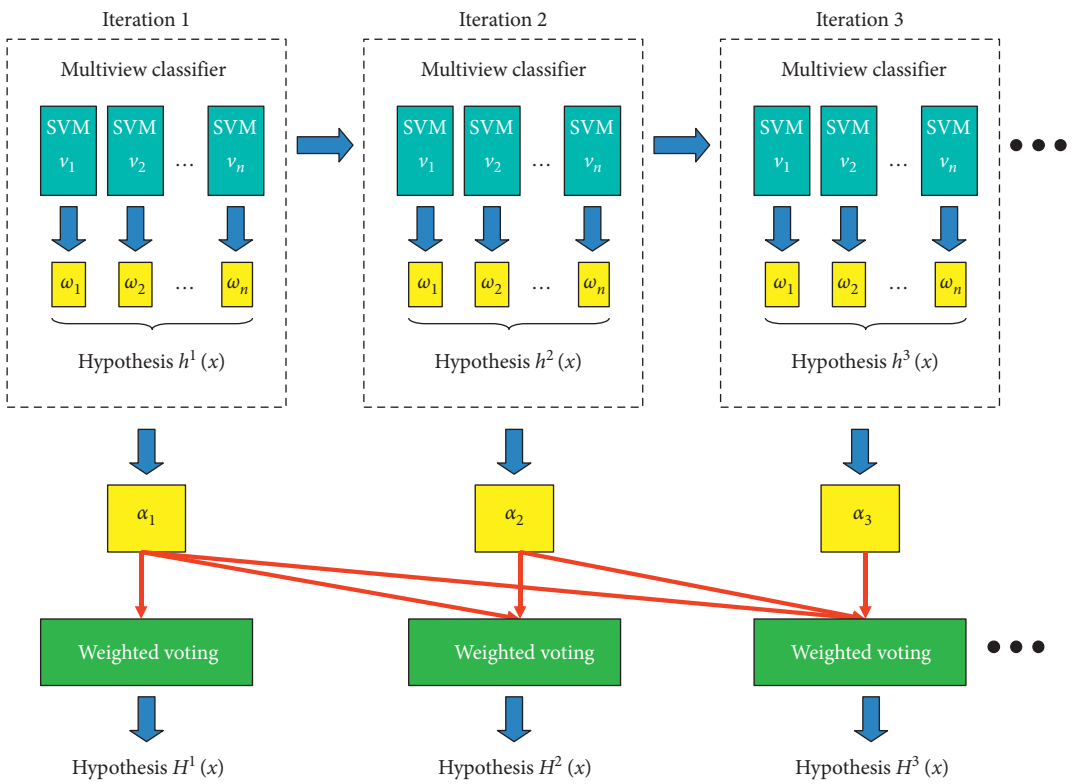


Figure 4: The framework of hypothesis generation in our MAVL [21].

where $\sum_{x\in L, y=1} f_i^t(x)$ and $\sum_{x\in L, y=-1} f_i^t(x)$ denote the sum of classification confidence of unlabeled samples, which are labeled as $y = 1$ and $y = -1$, respectively. For a "positive/negative" sample, the distance of it to the decision boundary in the "positive/negative" side reflects the degree of how correctly it is classified, and this information is utilized to calculate the error degree ε_i^t here instead of the traditional classification error calculated by the decision hypothesis in AdaBoost. Also, ω_i^t is updated through the following way: $\omega_i^t = (1/Z_1^t)\ln(1 - \varepsilon_i^t/\varepsilon_i^t)$, where Z_1^t is the normalized weight. Then, the classification confidence δ^t of the multiview classifier can be computed by the following equation:

$$\delta^t = \sum_{j=1}^{N} \beta_j^t \left| h^t(x_j) - y_j \right|. \tag{12}$$

(b) After iteration t, the size of the labeled sample set is increased as follows: $J^t = J^{t-1} \cup L^t$. J^t denotes the labeled sample set in iteration t, and L^t denotes the newly added samples after query. As we know, the size of the labeled samples set $|J^t|$ is increased during iteration in active learning. Thus, if the size of the initial labeled training set is small, the influence of $|J^t|$ should be considered when updating the weight η_t of the multiview classifier, which is illustrated by the following equation:

$$\eta_t = \frac{1}{Z_2^t}\left(\ln\left(\frac{1-\delta^t}{\delta^t}\right) + \lambda\left|J^t\right|\right). \tag{13}$$

Then, the weight of each sample is updated through the following way: $\omega_j^{t+1} = \omega_j^t \beta_t^{1-e^t}$, where $\beta_t = (\delta_t/(1-\delta_t))$, if x_j is correctly classified, $e_j = 0$, otherwise, $e_j = 1$.

(c) The final boosted hypothesis $H^t(x)$ of the queried sample x_i is equivalent to the weighted sum of all the hypotheses from the past K queries, which is defined by

$$H^t(x) = \begin{cases} 1, & \sum_{t=1}^{K} \eta_t h^t(x) \geq \frac{1}{2}\sum_{t=1}^{K} \eta_t, \\ 0, & \text{else.} \end{cases} \tag{14}$$

2.4.2. Sampling Strategy. The MVAL uses a new hierarchical competition-based sampling strategy in order to query the contention samples with high probability in different sample distributions, which is illustrated in Figure 5.

(1) *Intercluster Sampling Competition.* In the MVAL, a fast approximate spectral clustering algorithm is designed to reduce the computational complexity significantly to $O(\mathrm{KNT}) + O(K^3)$, where T is the iteration number of K mean clustering, and N is the total number of contention samples. The detailed process is illustrated as follows: (a) perform traditional K mean clustering on the contention unlabeled samples $x_1, x_2, \ldots, x_N$, compute the centroid of each cluster $y_1, y_2, \ldots, y_K$ as K representative points, and build a correspondence table to associate each x_i with the nearest cluster centroid $y_i y_i$; (b) run the normalized cut algorithm on $y_1, y_2, \ldots, y_K$ to obtain a m-way cluster membership for each of y_i; and (c) recover the cluster membership for each x_i by looking up the cluster membership of the corresponding centroid y_i in the corresponding table.

After fast spectral clustering, two intercluster sampling measures are defined: the number of samples in the cluster and its information entropy. Both measures are weighted to obtain the number of selected samples N_C^S in cluster C in the following equation:

$$N_C^S = \frac{N_T}{Z}\left[\gamma \mathrm{Num}(C) + (1-\gamma)\mathrm{Ent}(C)\right], \quad C = 1, 2, \ldots, K, \tag{15}$$

where $\mathrm{Num}(C)$ is proportional to the total number of samples N_C in cluster C, and computing $\mathrm{Ent}(C)$ is equivalent to kernel density estimation of x in cluster C. Weight $\gamma = 0.5$ reflects the impact of both measures in intercluster sampling competition, Z is the normalized factor, N_T is the total number of selected samples in the current query, and $[\cdot]$ is rounding operation.

(2) *Intracluster Sampling Competition.* In the MVAL, an efficient quadratic programming based-method [22] is utilized, which dynamically estimates the weights of the redundancy and uncertainty of an unlabeled sample in each query. It is used for intracluster selective sampling and solved by minimizing the following object function:

$$\begin{aligned} &\min_{p\in R^{n-l}} p^T \widetilde{f}_v + \frac{1}{2} p^T K_{u,u} p \\ &\text{s.t.} \quad p^T u = k,\ 0 \leq p \leq 1. \end{aligned} \tag{16}$$

Equation (16) aims to estimate the normalized parameter $p_i \in [0, 1]$, which reflects how probable the unlabeled sample is selected. $\widetilde{f}_v = (|f_v(x_1)|, \ldots, |f_v(x_l)|)^T$ is the classification confidence of sample x in v^{th} view. $x_1, \ldots, x_l$ are the queried unlabeled samples, u is a unit vector, and $k = N_C^S$ is the number of unlabeled samples in batch mode. The first part denotes the sample uncertainty in v^{th} view, and the sampling strategy tends to select the contention sample near the classification hyperplane of v^{th} view by minimizing $p^T \widetilde{f}_v$. The second part denotes the sample redundancy in v^{th} view, and the similar samples are selected by minimizing $p^T K_{u,u} p$. The sampling probability p is calculated by a convex quadratic programming, and finally, $\{p_1^v, p_2^v, \ldots, p_l^v\}$, which corresponds to $x_1, \ldots, x_l$ in v^{th} view, is obtained. For selective sampling in each cluster, the conservative sampling strategy is utilized in a classic co-testing algorithm [23].

3. Results and Discussion

In our experiment, two classic image sets (OT image set from MIT [9] and UIUC sports event image set from UIUC [24])

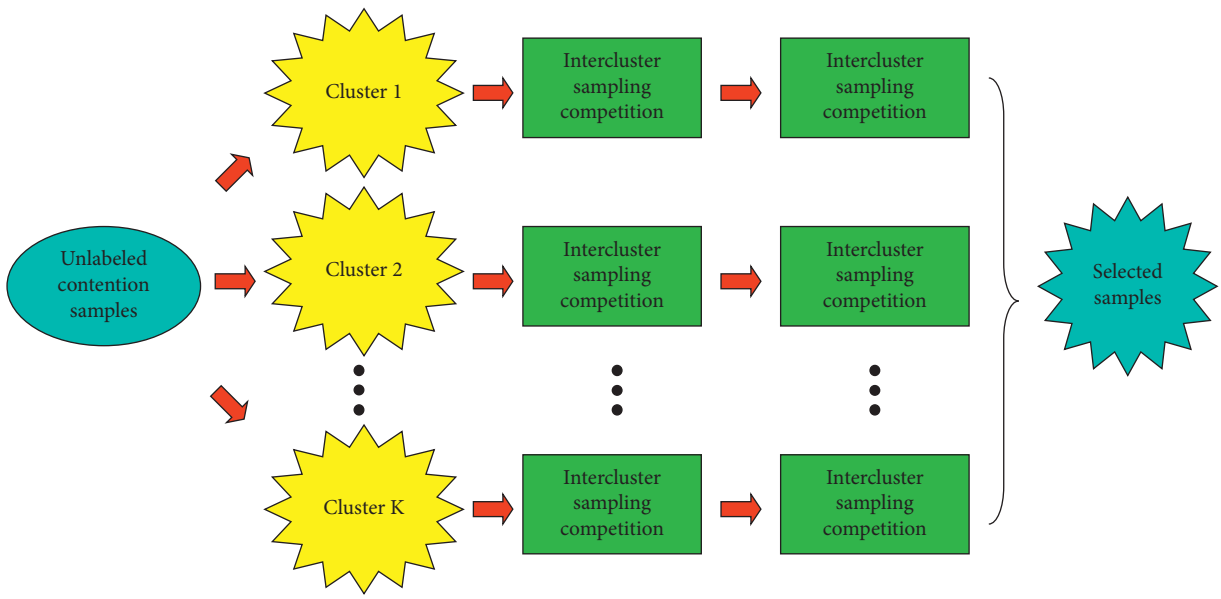


Figure 5: The framework of sampling strategy in our MAVL [21].

are used for algorithm comparison. Average classification precision (ACP) and mean of average classification precision (MACP) are both used for evaluating the performance of both CO-LDA models and multiview active learning algorithms.

3.1. Evaluation of Theme Semantic. The first experiment is designed for evaluating the performance of our proposed theme semantic. In OT and UIUC Sports datasets, the parameter configuration of the CO-LDA model is as follows: (1) $k_{\mathrm{OT}} = 0.5$, $\tau_{\mathrm{OT}} = 256$ and $k_{\mathrm{UIUC}} = 0.8$, $\tau_{\mathrm{UIUC}} = 1024$ in formula (7). (2) The batch sizes of sampled images in MVAL are $S_{\mathrm{OT}} = S_{\mathrm{UIUC}} = 512$.

We observe MACP variation of the CO-LDA model by changing the numbers of both theme and visual word: $T = 20, 30, 40, 50$ and $W = 200, 500, 800, 1200, 1500$, and a total of twenty groups of (T, W) are obtained. In Figure 6, we find that (T, W) curves for both datasets show the similar trends that MASP increase first and then decrease. Thus, in our CO-LDA model, we set $T_{\mathrm{OT}} = 30, W_{\mathrm{OT}} = 500$ and $T_{\mathrm{UIUC}} = 40, W_{\mathrm{UIUC}} = 1200$.

Figures 7(a)-7(b) and 8(a)-8(b) show the probability distributions of different themes by CO-LDA in OT and UIUC Sports image datasets.

In the OT image set, we can see that there are significant differences between four scene classes "Highway," "Forest," "Mountain," and "Tall building" in theme probability distributions, and multiview SVM classifier works well in scene classification. In the UIUC Sports image set, the theme probability distributions are very similar in four scene classes "Bocce," "Croquet," "Polo," and "Snowboarding," which significantly increases the difficulty of scene classification.

Furthermore, we compare the CO-LDA model with traditional LDA [9] and SP-pLSA [8] models, and the performance comparison of three theme models is shown in

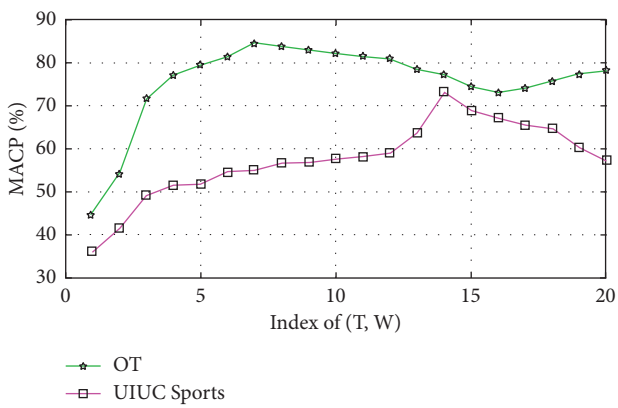


Figure 6: MACP curves with different groups of the theme and visual word in OT and UIUC Sports datasets.

Table 1. N1 ~ N8 denote the following eight natural scene classes: "Coast," "Forest," "Mountain," "Open country," "Highway," "Inside city," "Tall building," and "Street." S1 ~ S8 denote the following eight event scene classes: "Badminton," "Bocce," "Croquet," "Polo," "Rock Climbing," "Rowing," "Sailing," and "Snowboarding."

In the LDA model, each image is divided into 11×11 blocks, and 5 pixels are overlapped between neighbored blocks. For feature representation, gray-scale SIFT descriptors are sparsely sampled, and means of three color channels are calculated. The numbers of the theme and visual word are $T_{\mathrm{OT}} = 30, W_{\mathrm{OT}} = 200$ and $T_{\mathrm{OT}} = 50, W_{\mathrm{OT}} = 800$ by cross validation. In the SP-pLSA model, the ways of image division and feature representation are the same as the LDA model. The numbers of the theme and visual word are $T_{\mathrm{OT}} = 25, W_{\mathrm{OT}} = 1200$ and $T_{\mathrm{OT}} = 50, W_{\mathrm{OT}} = 1500$ by cross validation.

In the OT image set, CO-LDA achieves both higher ACP and MACP than SP-pLSA in six scene classes except "Mountain" and "Inside city." LDA performs the worst in all

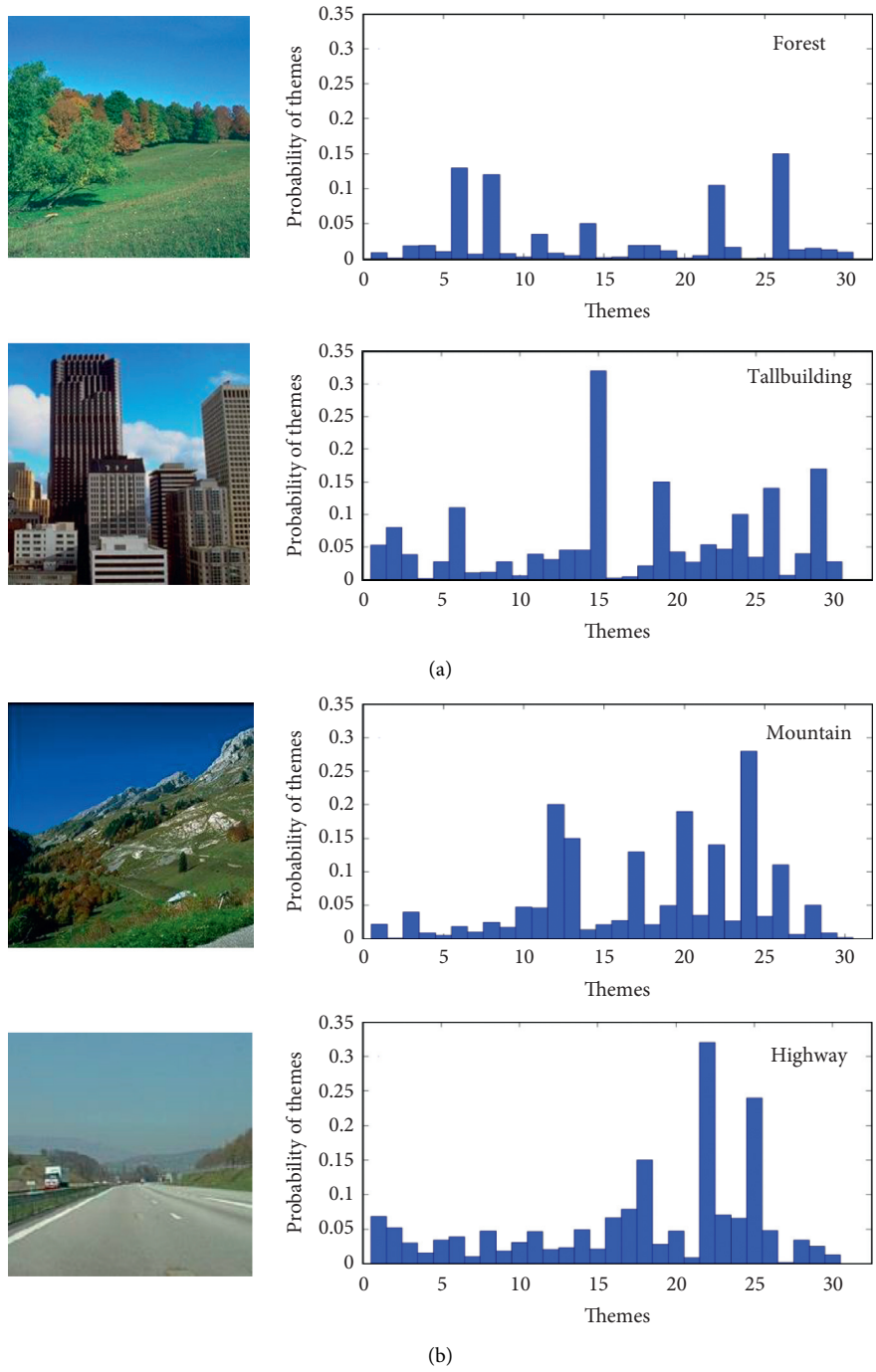


Figure 7: The theme probability distribution in different scene classes (OT image dataset).

of scene classes except "Street." It is easy to conclude that CO-LDA can achieve more accurate scene semantics than other two classic methods. In the UIUC Sports image set, CO-LDA achieves the highest ACP in the following three event classes: "Croquet," "Polo," and "Rowing," and SP-pLSA achieves the highest ACP in the following three event classes: "Bocce," "Rock Climbing," and "Snowboarding." But in the event classes "Badminton" and "Sailing," in which LDA has the highest ACP, CO-LDA still performs better than SP-pSLA. Thus, we can conclude that our proposed CO-LDA also have slightly better performance in theme mining than the two classic image representation methods.

3.2. Evaluation of MVAL. In the second experiment, we compare our algorithm with other single-view active learning algorithm with both high-level semantics and low-level features for scene classification. In our initial labeled training set, label size = 150, batch size = 20, and iteration = 10.

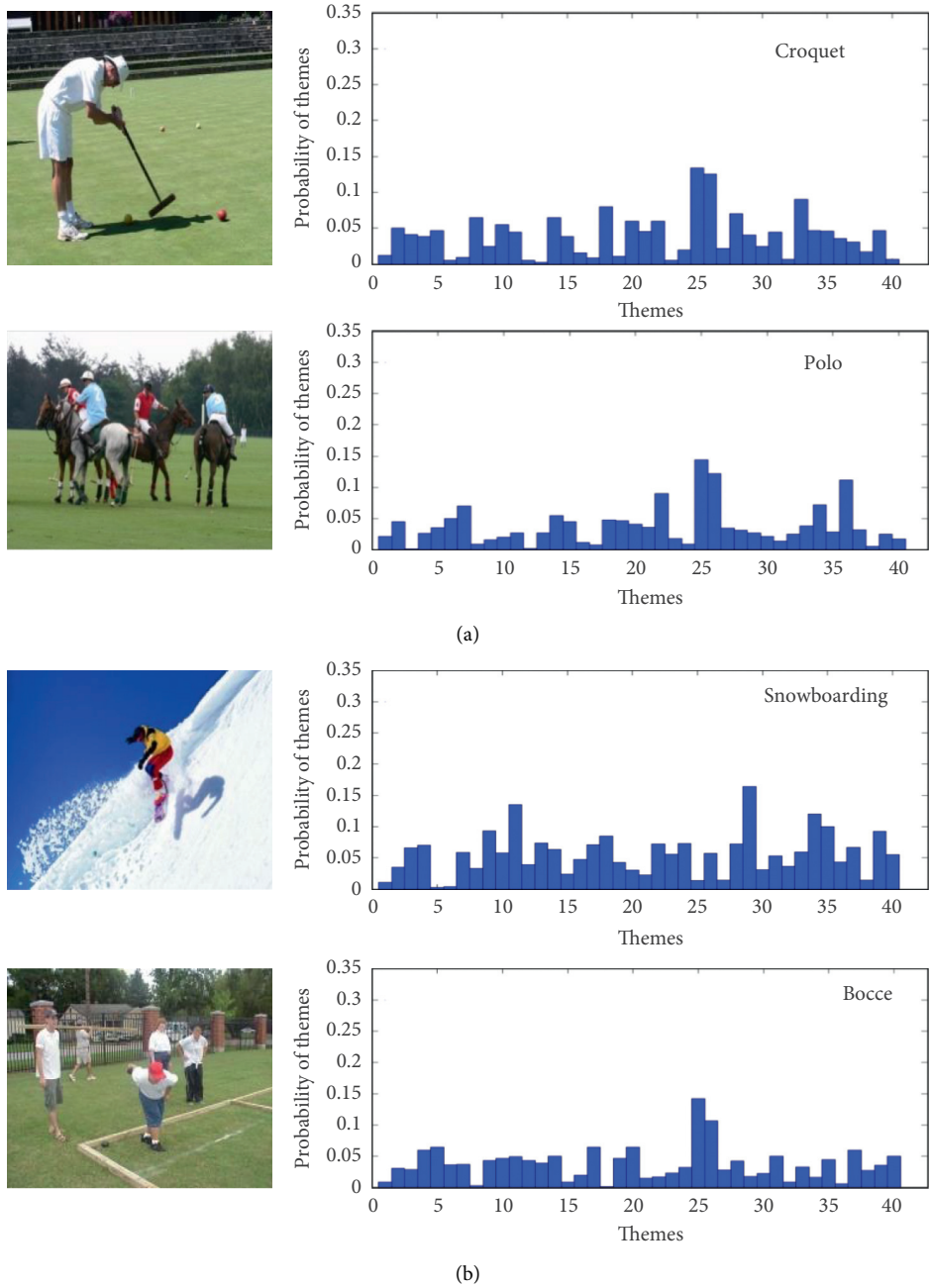


Figure 8: The theme probability distribution in different scene classes (UIUC image dataset).

Our proposed algorithm MVAL^{HS} (MVAL and HS denote MVAL [21] and two proposed high-level semantics, respectively) is compared with the following four algorithms: (1) MVAL^{LS} (LS denotes low-level image features): in MVAL, both means of three color channels and densely sampled color-SIFT descriptors are concatenated as a feature vector for image representation. (2) AL^{QP} [22]: a single-view SVM active learning by QP-based selective sampling, which relies on the sample uncertainty and redundancy. (3) Diff^{WS} [25–30]: a disagreement-based active learning from weak and strong labelers. (4) Graph^{GP} [23]: a graphical model-based active learning with robust Gaussian process. The feature representations of AL^{QP}, Diff^{WS}, and Graph^{GP} are the same as MVAL^{LS}. The performance comparison of the five active learning algorithms is shown in Table 2.

From Table 2, it is easily found that our algorithm MVAL^{HS} has the highest MACP in almost all scene classes than the other four algorithms in both image sets, which demonstrates that high-level semantics can achieve more significant improvement in holistic scene understanding than

Table 1: Performance comparison of different theme models.

ACP	N1	N2	N3	N4	N5	N6	N7	N8	*MACP*
(a) OT image dataset									
LDA	73.24 ± 1.16	86.69 ± 0.55	74.83 ± 0.72	64.42 ± 1.47	68.13 ± 1.85	80.75 ± 1.60	82.10 ± 0.50	84.74 ± 1.31	76.56 ± 1.15
SP-pLSA	82.40 ± 0.58	91.59 ± 0.34	87.74 ± 1.79	73.05 ± 1.04	79.66 ± 0.98	89.27 ± 0.59	89.68 ± 1.42	79.06 ± 1.55	83.80 ± 1.04
CO-LDA	86.32 ± 1.39	94.22 ± 1.53	85.17 ± 0.21	73.70 ± 0.30	83.53 ± 0.52	87.39 ± 1.26	90.74 ± 1.38	82.47 ± 0.41	86.26 ± 0.88
ACP	S1	S2	S3	S4	S5	S6	S7	S8	*MACP*
(b) UIUC Sports image dataset									
LDA	79.84 ± 2.28	63.86 ± 0.94	59.74 ± 1.53	65.49 ± 0.59	73.01 ± 0.29	61.65 ± 0.98	84.92 ± 1.38	58.22 ± 0.03	68.30 ± 1.00
SP-pLSA	75.25 ± 1.44	69.14 ± 0.51	67.19 ± 1.11	62.70 ± 1.73	86.32 ± 1.50	72.54 ± 0.13	80.75 ± 0.76	75.34 ± 1.02	73.64 ± 1.13
CO-LDA	77.62 ± 0.16	67.48 ± 0.74	68.73 ± 0.86	71.48 ± 2.13	82.75 ± 1.04	77.81 ± 0.07	83.43 ± 0.42	72.89 ± 0.55	75.27 ± 0.75

Note. Bold values represent the best performance of the algorithms corresponding to each class.

Table 2: Performance comparison of different active learning algorithms.

DS	Algorithm	N1	N2	N3	N4	N5	N6	N7	N8
OT image dataset									
	AL^{QP}	70.52 ± 1.80	85.65 ± 1.50	74.94 ± 1.33	53.52 ± 2.14	77.34 ± 074	80.04 ± 017	79.17 ± 1.34	75.00 ± 0.85
	$Diff^{WS}$	72.43 ± 1.67	84.19 ± 1.59	76.53 ± 1.73	55.14 ± 0.74	72.88 ± 0.98	78.58 ± 1.12	81.40 ± 0.88	74.30 ± 1.00
	$Graph^{GP}$	71.52 ± 1.44	86.32 ± 0.99	74.30 ± 0.15	52.05 ± 1.54	75.85 ± 1.17	79.53 ± 2.03	78.76 ± 1.89	76.75 ± 0.44
	$MVAL^{LS}$	73.23 ± 0.55	85.57 ± 1.76	77.53 ± 1.59	53.85 ± 0.95	77.34 ± 1.22	81.40 ± 2.00	81.87 ± 1.36	77.52 ± 0.53
	$MVAL^{HS}$	77.25 ± 0.88	90.41 ± 1.01	80.37 ± 1.55	60.49 ± 0.29	82.14 ± 0.72	87.38 ± 1.13	85.91 ± 0.14	81.73 ± 0.66
DS	Algorithm	S1	S2	S3	S4	S5	S6	S7	S8
UIUC-Sports image dataset									
	AL^{QP}	68.54 ± 1.41	49.42 ± 0.32	45.90 ± 1.40	55.17 ± 0.62	67.19 ± 0.88	60.20 ± 1.52	75.19 ± 1.07	51.09 ± 1.21
	$Diff^{WS}$	72.22 ± 1.50	50.98 ± 1.22	48.79 ± 0.59	57.33 ± 1.31	67.00 ± 0.55	57.53 ± 0.78	71.01 ± 0.16	54.58 ± 1.50
	$Graph^{GP}$	69.18 ± 1.37	54.10 ± 0.76	47.52 ± 2.13	53.75 ± 0.89	69.04 ± 0.34	62.55 ± 1.04	74.10 ± 1.00	55.22 ± 0.28
	$MVAL^{LS}$	71.18 ± 1.39	53.88 ± 0.55	49.62 ± 0.54	57.93 ± 1.65	71.05 ± 1.75	62.98 ± 1.70	75.59 ± 1.04	57.18 ± 1.24
	$MVAL^{HS}$	76.75 ± 1.39	55.44 ± 1.60	53.96 ± 1.23	59.78 ± 0.77	70.66 ± 0.52	66.40 ± 1.08	78.05 ± 1.51	60.68 ± 0.41

Note. Bold values represent the best performance of the algorithms corresponding to each class.

traditional low-level image features. Furthermore, we can see that $MVAL^{LS}$ performs better in most cases than other three single-view algorithms, which also means that multiple view setting can successfully result in larger decrease in the size of the version space than traditional single-view active learnings due to its independent and diverse views.

4. Conclusion

This paper proposed a MVAL-based scene classification algorithm, which applies two different high-level image semantics to generate the corresponding hypotheses. Different object detectors are first trained to achieve the responses of different object classes as object semantic. Furthermore, a CO-LDA model is proposed for achieving more accurate theme semantic by integrating the spatial correlation of objects in different image resolutions, which improves the holistic scene understanding. With the help of both two independent views, our MVAL algorithm has potential to not only handle large-scale data training but also improve the performance of scene classification.

Data Availability

All data utilized in our research can be accessed from the following website: https://archive.ics.uci.edu/ml/datasets/Corel+Image+Features and http://vision.stanford.edu/lijiali/event_dataset/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the Natural Science Foundation of China (Grant nos. 41571299 and 11601339), Key Research and Development Plan of Zhejiang Province (Grant no. 2018C01086), Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (no. ICT20047), Zhejiang Provincial Natural Science Foundation of China (Grant no. LY18F020025), and National Thousand Talents Program (Grant no. Y474161).

References

[1] P. F. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 32, no. 9, pp. 145–175, 2001.

[2] A. Oliva and A. Torralba, "Chapter 2 Building the gist of a scene: the role of global image features in recognition," *Progress in Brain Research*, vol. 30, no. 4, pp. 23–36, 2006.

[3] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the conference of computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[5] H. Guo, K. Zhang, X. C. Fan, H. K. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.

[6] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from a single image," in *IEEE 11th International conference on computer vision*, pp. 1–8, Rio de Janeiro, Brazil, 2007.

[7] P. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[8] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 712–727, 2008.

[9] L. Fei-Fei and P. Peronma, "A bayesian hierarchical model for learning natural scene categories," in *IEEE computer society conference on computer vision and pattern recognition*, pp. 524–531, San Diego, CA, USA, 2005.

[10] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *Proceedings of the conference of computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018.

[11] Z. M. Chen, X. S. Wei, P. Wang, and Y. W. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.

[12] X. Y. Zhang, S. H. Du, and Y. Zhang, "Semantic and spatial co-occurrence analysis on object pairs for urban scene classification," *IEEE Journal of Selected Ttopics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 2630–2643, 2018.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 2015.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: single shot multibox detector," in *European conference on computer vision*, Amsterdam, Netherlands, 2016.

[15] T. Y. Lin, P. Dollar, and R. Girshick, "Feature pyramid networks for object detection," in *Proceedings of the conference of computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[16] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 712–724, 2017.

[17] K. Wang, D. Y. Zhang, Y. Li, R. M. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[18] L. Yang, Y. Z. Zhang, J. X. Chen, S. Y. Zhang, and D. Z. Chen, "Suggestive annotation: a deep active learning framework for biomedical image segmentation," in *Proceedings of the conference of computer vision and pattern recognition*, Honolulu, HI, USA, 2017.

[19] W. H. Yang, G. Q. Liu, L. Zhang, and E. H. Chen, "Multi-view learning with batch mode active selection for image retrieval," in *Proceedings of the 21st international conference on pattern recognition*, pp. 979–982, Tsukuba, Japan, 2012.

[20] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *Proceedings of Neural Information Processing Systems*, pp. 1–9, 2010.
[21] T. Z. Yao, P. An, and J. T. Song, "Multi-view active learning based on weighted hypothesis boosting and hierarchical competition sampling," *Acta Electronica Sinica*, vol. 45, no. 1, pp. 46–53, 2017.
[22] S. C. H. Hoi, R. Jin, J. K. Zhu, and M. R. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *IEEE conference on computer vision and pattern recognition*, pp. 1–7, Anchorage, AK, USA, 2008.
[23] C. C. Long and G. Hua, "Multi-class multi-annotator active learning with robust Gaussian process for visual recognition," in *IEEE international conference on computer vision*, pp. 2839–2847, Santiago, Chile, 2015.
[24] L. J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *IEEE 11th International conference on computer vison*, pp. 1–8, Rio de Janeiro, Brazil, 2007.
[25] C. C. Zhang and K. Chaudhuri, "Active learning from weak and strong labelers," *Advances in Neural Information Processing Systems*, 2015.
[26] X. M. Zhang, T. T. Wang, J. Q. Qi, H. C. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the conference of computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018.
[27] T. Zhao and X. Q. Wu, "Pyramid Feature attention network for saliency detection," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.
[28] H. L. Zheng, J. L. Fu, Z. J. Zha, and J. B. Luo, "Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the conference of computer vision and pattern recognition*, Long Beach, CA, USA, 2019.
[29] L. J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: a high-level image representation for scene classification & semantic feature sparsification," *Proceedings of Neural Information Processing Systems*, pp. 1–9, 2010.
[30] I. Muslea, S. Minton, and C. A. Knoblock, "Active learning with multiple views," *Journal of Artificial Intelligence Research*, vol. 27, no. 1, pp. 203–233, 2006.

Hindawi
Scientific Programming
Volume 2020, Article ID 8877128, 7 pages
https://doi.org/10.1155/2020/8877128

Review Article

The Literature Review of Platform Economy

Chen Xue,[1] **Wuxu Tian**,[1] **and Xiaotao Zhao**[2]

[1]*School of Business, Moutai Institute, Renhuai 564507, China*
[2]*School of Economics and Management, Southeast University, Nanjing 211189, China*

Correspondence should be addressed to Chen Xue; 120068672@qq.com

Received 19 April 2020; Revised 24 May 2020; Accepted 6 July 2020; Published 1 September 2020

Academic Editor: Chenxi Huang

Since the 1990s, the increasing development of digital-driven technologies such as the Internet, cloud computing, big data, and the Internet of Things and the popularization of computers and mobile electronic devices have accelerated the evolution of global business organizations, thus making a new form of business organization, platform economy. As the most important form of industrial organization in the new economic era, the development of the platform has received extensive attention from the academia. Through literature analysis and inductive deduction, this paper reviews the connotation of platform economy, the historical context of development, the competition and monopoly (differentiation) of multilateral platforms, the evaluation mechanism of platform, antimonopoly governance, and research methods, and provides theoretical references and new ideas for future research directions.

1. Introduction

What is the "platform economy?" Evans [1] defines platform economy as a study of the unique economic phenomena of specific two-sided markets in traditional market economics. The platform economy studied in this paper refers to a series of digital technologies driven by the Internet, cloud computing, big data, and the Internet of Things, with a large number of platform enterprises as the lead, designing and implementing a complete set of platforms, consumers, and service providers, and influencing upstream and downstream enterprises, to reduce transaction costs of organizational rules and services and to achieve a new type of economic integration in which resources are highly integrated with traditional industries. Since the 1990s, mass platforms applying digital-driven business models show around the world where Google, eBay, Alibaba, Baidu, Tencent, JD.com, and other enterprises have stood out on the Internet, leading the increasing development of various industries and gradually forming the platform economy model, which has profoundly affected all aspects of the national economy and reshaped the market structure and competitive behavior of different industries. Meanwhile, it brings cross-country and cross-region business models and accelerates the global economic integration. This rapid growth of platform enterprises has inspired vast in-depth research studies in academia.

2. The Historical Development of Platform Economy

McAfee and Brynjolfsson [2] regard the rise of the platform as one of the three iconic events of the "digital revolution." Its rise has changed people's production and life and has also changed the way of human thinking. Most successful enterprises now have platform attributes [3]. Platform is the intermediary to realize the exchange between other participants, and most major technology companies can be regarded as platform-based enterprises [4].

Liebowitz and Margolis [5] put forward the conjecture of network externality and call the "market regulation effect" as "(indirect) network effect." Their conjecture is considered by academia to be the oldest openness issue in network economics. Subsequently, Rochet and Tirole [6], Armstrong [7], and Caillaud and Jullien [8] have pioneered platform research and guided the economics community's interest in platform research.

At the beginning of the 21st century, with the rapid development of information technology and the further refinement of the social division of labor, the market

operation model based on buyers and sellers has become increasingly mature. The platform enterprises based on buyers and sellers have formed a new market relationship, that is, the two-sided markets in the production and operation market. Caillaud and Jullien [8], Rochet and Tirole [6], and Amstrong [9] believe that the increase of users on one side of the platform will cause the increase of users on the other side. The two-sided "cross-group network externality" is called two-sided markets. Armstrong [9], Evans [1], Evans and Schmalensee [10], and Filistrucchi et al. [11] believe that two-sided market should embody at least one side with "cross-group network externality." Rochet and Tirole [12] believe that the definition of cross-group network externality in two-sided markets lacks inclusiveness and should be defined from the price structure. Parker and van Alstyne [13] study two-sided markets earlier. They believe that the matching market is two-sided because the matching "platform" (such as dating services) is more important. Hagiu and Wright [14] impose restrictions on two-sided markets: one is the direct trade between the sides of the market, and the other is that each side of the market is "affiliated" to the platform, with higher cost of leaving the platform. Evans [1, 15] divides two-sided markets into three types: market maker, audience maker, and demand coordinator.

There are two main reasons to trace the emergence of the platform. Firstly, the platform helps to match. In the sharing economy, the platform provides a new structure to quickly and effectively match with low search cost [16] and acts as an intermediary between the buyer and the seller [17]. In view of the matching background, many scholars have conducted a lot of in-depth research on the competition and pricing strategies in the platform business and gradually focused on the importance of indirect network effects [8, 18–20] and [21]. Secondly, the platform has improved trade efficiency. The platform improves transaction frequency and efficiency by reducing search cost, low replication, and verification cost. Through zero cost replication, the platform enables application providers to quickly provide services for a large number of customers, with interoperability. Simcoe [22] emphasized platform interoperability and the strategic nature of standard decision-making [6, 23], Hanna and Yehezkel [24] tested whether market participants would "multiple" and use multiple platforms through empirical data.

Regarding the research of platform economy, from the perspective of two-sided markets, many scholars focus on market intermediary behavior, especially market pricing; from the perspective of network effects, scholars focus on user adoption and optimal network scale; from the perspective of industry focuses, the media, payment system, and matching market are highly paid attention to the research and literature of two-sided markets, focusing more on high-tech and telecommunication market about network effects.

3. The Competition Effect of Platform Economy

The theoretical economics literature on multisided platforms focuses on competition (differentiation) between antitrust and multisided platforms serving the same customer group. According to Evans and Noel [25], various platforms face a more complex competitive environment. The existence of the "cross-group network externality" of the platform has led to mutual reciprocity on both sides, and platform enterprises have grown up at an extremely rapid rate. Platform enterprises with "intragroup network externality" will be provided with entrance barriers as their scale grows, and it will be difficult for new enterprises to reenter, often causing winner-take-all issues [1, 8]. Meanwhile, after the platform gains market power, it will become a "modern antitrust" to a great extent [26]. Evans [27] conducts a research on the operating status of the world's top platforms and finds that their industry rankings have changed greatly in recent years, and antitrust platforms are also facing various challenges.

Platform competition theory has been one of the most active areas in industrial organization research for the past decade [6, 8, 9, 12, 20, 28]. Jacqueline and Jonathan [29] believe that the network effect model is conducive to the competition between platforms. Spiegler [30] studies that there are positive externalities between the two agents. The platform extracts these externalities by using exclusive interactive contracts. If another agent signs a contract with the platform, the payment to one of the agents is accordingly reduced. Caillaud and Jullien [31] believe that there is a market for price competition between the two platforms, and there is no difference between the platforms. When the existing platform has the market power, another new entry platform is difficult to develop. If the platform does not charge transaction fees, even if there is no product differentiation, the platform that has already occupied the market can still obtain profits; with the transaction fee, both platforms can get profits. Rochet and Tirole [6] study the issues of "single destination" and "multiple destination" for multisided platforms. Armstrong [9] shows the importance of "multiple destination" for competition. Armstrong and Wright [32] put forward that if the multisided platforms of competition is regarded as homogeneous by members of one group but differentiated by members of another group, then the "competitive bottleneck" will be endogenous. Economides and Katsamakas [33] study the competition between common platforms and open source platforms and find that the property platform dominates open source platforms by having greater market share and higher profitability. Jullien [34] conducts multifaceted background research studies based on the market of vertical differentiated platforms and sequential games and finds the pricing strategy of competitive platforms. Weyl [35] analyzes the pricing strategy of the platform from the perspective of multifunctional platform and from the perspective of user heterogeneity and platform monopoly, establishes a general theory of network antitrust pricing, and lays the foundation for the platform economic theory. Tiwana [36] and Mukhopadhyay et al. [37] believe that the platform is a dynamic, purposeful, or internal interdependent network, and participants can jointly create value [38] and add complementary products, services, and technologies [39]; Annabelle and Cusumano [40]. Mcintyre and Srinivasan [41] believe that the value creation in the platform system is jointly participated by platform

owners, suppliers, and final consumers. This is due to internal competition and cooperation between participants, which occurs in the interaction between independent participants and the evolutionary process. Reiley, Hall [42, 43] believe that the reduction of transaction costs will lead to more flexible platform pricing.

Brynjolfsson et al. [2] found that online prices are much lower than offline prices by comparing the products of four pure Internet retailers, four offline retailers, and four "hybrid" retailers with both online and offline stores. Orlov [44] found that the platform will increase the price dispersion within the enterprise, but it does not have a great impact on the price dispersion among enterprises. Instead, the search cost reduces the price dispersion [45–47].

Regarding information asymmetry in platform competition, Damiano and Li [48], Ambrus and Argenziano [49], Peitz et al. [50], Weyl [20], and White and Weyl [51] focus on the research of ex-ante asymmetric information. Based on this issue in the two-sided markets, Hanna and Yaron [24] used multiple destinations to solve the market failure caused by information asymmetry and empirically studied the influence of ex-ante uncertainty of new technology value and ex-post asymmetric information on platform strategies and results.

The focus of the platform economy is to solve the problem of how platforms can price both sides of the market at the same time. Rochet and Tirole [12] focus on the price structure when defining two-sided platforms, which has the characteristics of using externality and member externality. Evans and Schmalensee [10] define two-sided platforms, grasping the key features of the platform business, namely, (a) having two or more customer groups, (b) to some extent need each other, (c) unable to obtain value from mutual attraction, and (d) the platform creates more value. Evans [52] believe that the pricing of one side of the market depends not only on the demand and costs brought by consumers but also on how their participation affects the other party and the profits derived from the participation. In one-sided market, the elasticity of demand and marginal cost can be used to describe price cost increase, but in two-sided markets, the pricing decision also needs to consider the flexibility of the other party's response and the price increase charged to the other party. Rochet and Tirole [53] point out that if the services provided to consumers and merchants are completely competitive, then the exchange fee does not affect the profits of members but affects the terms and total transactions faced by merchants and consumers. Rochet and Tirole [6, 12] and Weyl [54] believe that prices on both sides of the market depend on the elasticity of demand and the marginal cost of each party. A platform acting as an intermediary is regarded as antitrust that has access to members who do not use other platforms. Enterprises using a single network compete fiercely in order to charge antitrust price to the other party trying to reach it [9].

An important issue of platform research is price discrimination in the case of heterogeneous demand. Weyl [35] has empirically found that, by manipulating the prices of participation and use, the platform can obtain more profits. Discrimination increases the value of one party, which causes the decreasing price of the other party. Caillaud and Jullien [8] empirically study how the new platform uses price discrimination to achieve success when market participants expect new entrants to fail.

Regarding the research on platform antitrust competition, Rochet and Tirole [6] and Armstrong [9] study antitrust pricing and price competition of platforms. Rochet and Tirole believe that the platform is priced based on transaction fees. Jullien [55] believes that there are more than two consumer subgroups and intragroup network externality on the platform, constructing a duopoly model. Evans [1] points out that, in many industries, enterprises act as catalysts to set prices below marginal costs, sometimes even zero, such as on some software platforms, advertising-supported media, exchanges, and payment systems.

Regarding the research on heterogeneity in platform competition, Ellison et al. [56] conduct a study on the competition between two auction websites and find that even if there is a lack of heterogeneity of products and agents, multiple platforms coexist. Damiano et al. [48, 57] [61] believe that consumers are heterogeneous in the platform economy, and different types of consumers can be distinguished through registration fees. Caillaud and Jullien impose monotonicity on the consumer demand, assuming that the full market coverage under equilibrium conditions is selected among equilibriums. Ambrus and Argenziano [58] study the conditions for the coexistence of multiple asymmetric networks in a bidirectional market with network externality from the perspective of consumer heterogeneity and find that one side of a network platform is cheaper and larger, the other side is even cheaper and larger. Weyl [54] believes that, in the case of heterogeneous demand, the platform can obtain more profits by operating participants and prices. Rochet and Tirole [6] put out that customers with prestige and influence will have significant direct or indirect externality.

4. The Governance Issues of Platform Economy

Platform governance is jointly constructed by the platform, its participants, and the government and revolves around three issues: "who sets the rules," "how to allocate rights and obligations," and "how to resolve disputes." The idea of platform governance includes spontaneous organization of platform subjects, participation in governance, government-led regulation, and consumer supervision. Jin and Kato [59], through empirical research on eBay's rating system, find that credibility is an effective means of identifying integrity platforms. Avery et al. [60] started the research on the recommendation system earlier and devoted to building the evaluation system. Jacqueline and Jonathan [29] conduct an empirical study on eBay, establishing a communication mechanism to encourage other users to evaluate the quality of sellers or products by setting up a communication mechanism that encourages and aggregates user feedback, thus building the evaluation system of a platform reputation. Dellarocas [61] has empirically found that sellers with higher feedback scores enjoy some benefits when prices and sales rates are higher.

Cabral and Hortacsu [62] and Saeedi et al. [63] find that the evaluation mechanism set up by the platform to increase the evaluation rate will breed strategic behavior of evaluation. The existence of strategic behavior will distort the reputation mechanism and affect its effectiveness. Rieder and Sire [64] have empirically found that if the platform's search accepts advertising sponsorship, it may affect the objectivity of the platform's search.

In the fields of economics, management, and strategic management, a large amount of literature on platform antitrust governance has appeared, especially because of the research on pricing issues in the field of platform economics, which provides an important reference for antitrust governance. Rochet and Tirole [53] find that if the antitrust law is not violated when setting the exchange fee, but the equilibrium result deviates from the social optimal value, the reasonable solution is price regulation. Evans [65] believes that, with the increasing revolution of the Internet, mobile devices, and information technology and the increasing emergence of global large-scale multisided platforms, antitrust will become a key link in governance issues.

Evans [52] studies how platforms develop governance systems to reduce platform participants' undesirable behavior that may reduce the value of the platform. Hagiu [66] discusses the use of platform regulations to increase positive externality and reduce negative externality. Ruhmer [67] raises the issue of whether multisided platforms can increase profits by colluding to charge only part of the price. Evans and Schmalensee [15] believe that the fierce competition between the two platforms may eliminate the profitability of colluding with each other's prices. Rysman [28] believes that predatory pricing and overpricing will lead to anticompetitive platforms.

5. The Method Research of Platform Economy

Rochet and Tirole [53], Jullien [55], Armstrong and Adendorff [68], and Parker and Van Alstyne [69] conduct empirical research on the entry, pricing, and other strategies of the platform industry, providing a research background for the emerging economic theory of platform. Rysman et al. [70] provides the empirical and policy [1] research for multisided platforms market. Rochet and Tirole [6] take the lead in constructing one of the two most basic models of two-sided platforms pricing. It is assumed that two-sided antitrust platforms have no member externality. Only by using externality and charging usage fees for each transaction while without charging, member fees can the cost bonus be lowered with the higher demand elasticity. Armstrong [9] proposes the second model, two-sided antitrust platforms do not use externality, only the externality of members, no user fees, only member fees, and the price of maximizing profits. Hagiu [66] modifies the Armstrong [9] model, and the size of the profit share of the antitrust platform comes from the consumer's preference for varieties. Jacqueline and Jonathan [29] conduct empirical research on platform and platform competition and construct a platform competition model for analyzing platform pricing, competition, and market tilt issues. Rochet and Tirole [6] carry out empirical research on both sides of the platform, which finds that, from the perspective of maximizing profits or maximizing social welfare, the optimal price may cause the pricing to be lower than the marginal supply cost of one party and higher than the marginal supply cost of the other party.

Economists have developed various models to help analyze whether certain business practices may harm consumers by excluding competitors from the market, or benefit consumers by lowering prices or improving quality. Segal and Whinston [71] demonstrate that, under the condition of economies of scale, monopolistic enterprises can effectively prevent competitors from entering by signing exclusive transaction contracts. Armstrong and Wright [32] build a platform competition model, which is regarded as a differentiated platform by one customer group and a homogenized platform by another customer group. Exclusive transactions can be used to prevent the latter 's multihoming and exclude competitors, but antitrust equilibrium may be effective. Doganoglu and Wright [72] demonstrate the effectiveness of this strategy without economies of scale but with network effects. Leung and Lee [73] conduct empirical research on the video game industry, finding that exclusive contracts can facilitate entry rather than prevent entry. Rochet and Tirole [74] simulate the situation that tying increases social welfare and found that tying in a simple model helps to enhance welfare, but in a more complex model, the net effect of tying on welfare is ambiguous. Choi [75] proposes a theoretical model based on the influence of the combination of Microsoft Windows media player and Windows. Bundling as a means of price discrimination will reduce costs, and the network effect will make the price optimal. Then, Chao and Derdenger adopt video games to carry out empirical analysis of pricing and reach a conclusion consistent with the theoretical model. Amelio and Bruno [76] conduct a case study. Assuming the profit-maximizing price of a party is negative, it is not feasible to actually charge a negative price. It is found that tying can both make profits and increase welfare. Ruhmer [67] constructs a single-destination two-sided model. The indirect network effect increases the benefits of price reduction, making collusion difficult to maintain.

Hagiu [77] provides a duopoly model of two-sided platforms. Armstrong and Wright [32] establish a bottleneck model of two-sided competition, The empirical platform can charge the producers of the market for user fees. Spulber [78] proves the choice of buyers and sellers and finds that they should either search in a decentralized market or search through an intermediary. Hagiu [66] constructs a competitive model for platform pricing. By analyzing the economic and strategic factors in the optimal access pricing structure of the two-sided platforms connecting consumers and producers, the competition between producers is introduced.

Farrell and Klemperer [79] construct a network effect model based on three factors: the degree of substitution of competitive platforms, the intensity of positive network effects, and the degree to which production is characterized by economies of scale. Rysman [28] believes that the single

destination preference, the scale and intensity of network effects, and the price elasticity system are all key parameters of the platform competition network effect model. Brown and Morgan [80] demonstrate the competition between eBay and Yahoo through auctions. Edelman [81] and Hal [42] give the modeling formula of auction market. Levin [82] studies a related model and show that there is a near-effective equilibrium of mixed strategies. Gawer and Cusumano [83] demonstrate the pricing strategies of Microsoft, Apple, IBM, Palm, and other operating system enterprises.

Baye and Morgan build a price competition model where consumers can search by price. Einav et al. [84] empirically study auction pricing and additional prices and find that fees and prices have no significant impact on consumers. Gentzkow and Shapiro [85] empirically study the personal consumption of online media and discover the long tail demand theory of platform economy. Roberts and Sweeting [86] construct a platform dynamic pricing model. Rysman [28] constructs a differentiated platform competition model to solve the platform antitrust issue.

Regarding the issue of dynamic competition, Doganoglu [87] and Mitchell and Skrzypacz [88] derive the Markov perfect equilibrium of the indefinite game where the consumer utility is a growth function of market share in the past. Markovich and Moenius [89] develop an industry computing model including "hardware" and "software" components, which assumes that consumers live for two periods and benefit from indirect network effects through the quality of existing products. Chen et al. [90] develop a computational dynamic model in which it is assumed that consumer benefits are a growing function of the size of the network at the time of purchase (consumers are not forward-looking). These assumptions assume that consumer behavior is relatively simple. Driskill [91] constructs a deterministic and continuous-time model where consumers are forward-looking. Cabral [92] builds a dynamic model of price competition based on balanced symmetry and network effects.

6. Research Outlook

As one of the three landmark events of the digital economic revolution [93], the platform economy will become the most important form of economic organization in the foreseeable future when the academia will not only be limited to the research of platform economy such as economics, management, econometrics, law, and sociology, but also possible to conduct cross-field and multifield research. One of the main focuses of two-sided markets economy is to solve the issue of how the platform can price both sides of the market at the same time. Platform opening, market access, pricing analysis, governance supervision, and antitrust are all significant directions for future research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. S. Evans, "The antitrust economics of multi-sided platforms markets," *Yale Journal on Regulation*, vol. 20, pp. 325–382, 2003.

[2] Brynjolfsson, Erik, Smith et al., "Frictionless commerce? A comparison of internet and conventional retailers," *Management Science*, vol. 46, no. 4, pp. 563–585, 2000.

[3] P. Evans and A. Gawer, *The Rise of the Platform Enterprise: A Global Survey*, The Center for Global Enterprise, New York, NY, USA, 2015.

[4] A. Cusumano and C. Tucker, "Digital economics," *Journal of Economic Literature*, vol. 57, no. 1, pp. 3–43, 2019.

[5] S. J. Liebowitz and S. E. Margolis, "Network externality: an uncommon tragedy," *Journal of Economic Perspectives*, vol. 8, no. 2, pp. 133–150, 1994.

[6] J.-C. Rochet and J. Tirole, "Platform competition in two-sided markets," *Journal of the European Economic Association*, vol. 1, no. 4, pp. 990–1029, 2003.

[7] M. Armstrong, "Network interconnection with asymmetric networks and heterogeneous calling patterns," *Information Economics and Policy*, vol. 16, no. 3, pp. 375–390, 2004.

[8] B. Caillaud and B. Jullien, "Chicken & egg: competition among intermediation service 49providers," *The RAND Journal of Economics*, vol. 34, no. 2, pp. 309–328, 2003.

[9] M. Armstrong, "Competition in two-sided markets," *The RAND Journal of Economics*, vol. 37, no. 3, pp. 668–691, 2006.

[10] D. S. Evans and R. Schmalensee, "The industrial organization of markets with two-sided platforms," National Bureau of Economic Research, Cambridge, MA, USA, Working Paper No. 11603, 2007.

[11] L. Filistrucchi, D. Geradin, and E. van Damme, "Market definition in two-sided marketss: theory and practice," *Journal of Competition Law and Economics*, vol. 10, no. 2, pp. 293–339, 2013.

[12] J.-C. Rochet and J. Tirole, "Two-sided markets: a progress report," *The RAND Journal of Economics*, vol. 37, no. 3, pp. 645–667, 2006.

[13] G. Parker and M. G. van Alstyne, "Information compliments, substitutes, and strategic product design," *SSRN Electronic Journal*, 2000.

[14] A. Hagiu and J. Wright., "Multi-sided platforms," *International Journal of Industrial Organization*, vol. 43, pp. 162–174, 2015.

[15] D. S. Evans and R. Schmalensee, *Catalyst Code: The Strategies behind the World's Most Dynamic Companies*, Harvard Business School Press Books, Boston, MA, USA, 2007.

[16] B. Jullien, *Two-sided B to B Platforms*, Oxford University Press, New York, NY, USA, 2012.

[17] V. Nocke, M. Peitz, and K. Stahl, "Platform ownership," *Journal of the European Economic Association*, vol. 5, no. 6, pp. 1130–1160, 2007.

[18] A. Hagiu and J. Bruno, "Why do intermediaries divert search?" *The RAND Journal of Economics*, vol. 42, no. 2, pp. 337–362, 2011.

[19] M. R. Baye, J. Morgan, and P. Scholten, "Temporal price dispersion: evidence from an online consumer electronics market," *Journal of Interactive Marketing*, vol. 18, no. 4, pp. 101–115, 2010.

[20] E. G. Weyl, "A price theory of multi-sided platforms," *American Economic Review*, vol. 100, no. 4, pp. 1642–1672, 2010.

[21] D. A. Corniere, "Search advertising," *American Economic Journal: Microeconomics*, vol. 8, no. 3, pp. 156–188, 2016.

[22] T. Simcoe, "Standard setting committees: consensus governance for shared technology platforms," *American Economic Review*, vol. 102, no. 1, pp. 305–336, 2012.
[23] M. Rysman, "An empirical analysis of payment card usage," *Journal of Industrial Economics*, vol. 55, no. 1, pp. 1–36, 2007.
[24] H. Hanna and Y. Yehezkel, "Platform competition under asymmetric information," *American Economic Journal Microeconomics*, vol. 5, no. 3, pp. 22–68, 2013.
[25] D. S. Evans and M. D. Noel, "Analyzing market definition and power in multi-sided platform markets," *SSRN Electronic Journal*, 2005.
[26] A. Moazed and L. Nicholas, *Johnson. Modern Monopolies: What it Takes to Dominate the 21st Century Economy*, St. Martin's Press, New York, NY, USA, 2016.
[27] D. S. Evans, "Why the dynamics of competition for online platforms leads to sleepless nights but not sleepy monopolies," *Social Science Electronic Publishing*, 2017.
[28] M. Rysman, "The economics of two-sided markets," *Journal of Economic Perspectives*, vol. 23, no. 3, pp. 125–143, 2009.
[29] H. Jacqueline and D. Jonathan, "Marketing in context -- the marketing authenticity of owner/entrepreneurs of small firms: case evidence from Welsh [uk] sme food and drink producers and retailers," *Small Enterprise Research*, vol. 18, no. 1, pp. 33–50, 2011.
[30] I. Spiegler, "Knowledge management: a new idea or a recycled concept?" *Communications of the Ais*, vol. 3, 2000.
[31] B. Caillaud and B. Jullien, "Competing cybermediaries," *European Economic Review*, vol. 45, no. 4–6, pp. 797–808, 2001.
[32] M. Armstrong and J. Wright, "Two-sided markets, competitive bottlenecks and exclusive contracts," *Economic Theory*, vol. 32, no. 2, pp. 353–380, 2007.
[33] N. Economides and E. Katsamakas, "Linux vs. Windows: a comparison of application and platform innovation incentives for open source and proprietary software platforms," in *The Economics of Open Source Software Development*, pp. 207–218, Elsevier, Amsterdam, Netherlands, 2006.
[34] B. Jullien, "Competition in multi-sided markets: divide and conquer," *American Economic Journal, Microeconomics*, vol. 3, no. 4, pp. 186–220, 2011.
[35] E. G. Weyl, "Slutsky meets marschak: the first-order identification of multi-product production," *SSRN Electronic Journal*, vol. 16, 2009.
[36] A. Tiwana, *Platform Ecosystems: Aligning Architecture, Governance, and Strategy*, Morgan Kaufmann Publishers Inc., Burlington, MA, USA, 2014.
[37] S. Mukhopadhyay, M. de Reuver, and H. Bouwman, "Effectiveness of control mechanisms in mobile platform ecosystem," *Telematics and Informatics*, vol. 33, no. 3, pp. 848–859, 2016.
[38] R. Adner and R. Kapoor, "Value creation in innovation ecosystems: how the structure of technological interdependence affects firm performance in new technology generations," *Strategic Management Journal*, vol. 31, no. 3, pp. 306–333, 2009.
[39] M. G. Jacobides, "Who does what and who takes what: benefiting from innovation," 2006.
[40] G. Annabelle and M. A. Cusumano, "Industry platforms and ecosystem innovation," *Journal of Product Innovation Management*, vol. 31, no. 3, pp. 417–433, 2014.
[41] D. P. Mcintyre and A. Srinivasan, "Networks, platforms, and strategy: emerging views and next steps," *Strategic Management Journal*, vol. 38, no. 1, pp. 141–160, 2017.
[42] R. V. Hal, "Position auctions," *International Journal of Industrial Organization*, vol. 25, no. 6, pp. 1163–1178, 2007.
[43] R. E. Hall, *Digital Dealing: How E-Markets Are Transforming the Economy*, W. W. Norton, New York, NY, USA, 2002.
[44] E. Orlov, "How does the internet influence price dispersion? Evidence from the airline industry," *The Journal of Industrial Economics*, vol. 59, no. 1, pp. 21–37, 2011.
[45] R. Jensen, "The digital provide: information (technology), market performance, and welfare in the South Indian fisheries sector," *The Quarterly Journal of Economics*, vol. 122, no. 1, pp. 879–924, 2007.
[46] J. C. Aker, "Information from markets near and far: mobile phones and agricultural markets in Niger," *American Economic Journal: Applied Economics*, vol. 2, no. 3, pp. 46–59, 2010.
[47] C. Parke, K. Ramdas, and N. Savva, "Is IT enough? Evidence from a natural experiment in India]s agriculture markets," *Management Science*, vol. 62, no. 9, pp. 2481–2503, 2016.
[48] E. Damiano and L. Hao, "Competing matchmaking," *Journal of the European Economic Association*, vol. 6, no. 4, pp. 789–818, 2008.
[49] A. Ambrus and R. Argenziano, "Asymmetric networks in two-sided markets," *American Economic Journal: Microeconomics*, vol. 1, no. 1, pp. 17–52, 2009.
[50] M. Peitz, S. Rady, and P. Trepper, "Experimentation in two-sided markets," *Social Science Electronic Publishing*, 2010.
[51] A. White and E. G. Weyl, "Imperfect platform competition: a general framework," 2010.
[52] D. S. Evans, "The Antitrust Analysis of Multi-Sided Platform Businesses," National Bureau of Economic Research, Cambridge, MA, USA, 2013.
[53] J.-C. Rochet and J. Tirole, "Cooperation among competitors: some economics of payment card associations," *The RAND Journal of Economics*, vol. 33, no. 4, pp. 549–570, 2002.
[54] E. G. Weyl, "Monopoly, ramsey and lindahl in Rochet and Tirole (2003)," *Economics Letters*, vol. 103, no. 2, pp. 99-100, 2009.
[55] B. Jullien, "Competing with network externalities and price discrimination," 2001.
[56] G. Ellison, D. Fudenberg, and M. Möbius, "Competing auctions," *Journal of the European Economic Association*, vol. 2, no. 1, pp. 30–66, 2004.
[57] E. Damiano, H. Li, and W. Suen, "Unravelling of dynamic sorting," *Review of Economic Studies*, vol. 72, no. 4, pp. 1057–1076, 2005.
[58] A. Ambrus and R. Argenziano, "Network markets and consumers coordination," *Social Science Electronic Publishing*, vol. 10, 2004.
[59] G. Z. Jin and A. Kato, "Price, quality, and reputation: evidence from an online field experiment," *The RAND Journal of Economics*, vol. 37, no. 4, pp. 983–1005, 2006.
[60] C. Avery, P. Resnick, and R. Zeckhauser, "The market for evaluations", American Economics Review, vol. 89, no. 3, pp. 564–584, 1999.
[61] C. Dellarocas, "Strategic manipulation of internet opinion forums: implications for consumers and firms," *Management Science*, vol. 52, no. 10, pp. 1577–1593, 2006.
[62] L. Cabral and A. Hortaçsu, "The dynamics of seller reputation: evidence from Ebay," *The Journal of Industrial Economics*, vol. 58, no. 1, pp. 54–78, 2010.
[63] M. Saeedi, Z. Shen, and N. Sundaresan, "The value of feedback: an analysis of reputation system," *Social Science Electronic Publishing*, vol. 16, p. 2005, 2014.

[64] B. Rieder and G. Sire, "Conflicts of interest and incentives to bias: a microeconomic critique of Google's tangled position on the Web," *New Media & Society*, vol. 16, no. 2, pp. 195–211, 2013.
[65] D. S. Evans, "The economics of the online advertising industry," *Review of Network Economics*, vol. 7, no. 3, 2008.
[66] A. Hagiu, "Quantity vs. Quality and Exclusion by Two-Sided Platforms," *SSRN Electronic Journal*, Harvard Business School Strategy Unit, Boston, MA, USA, 2009.
[67] I. Ruhmer, *Platform Collusion in Two-Sided Markets*, University of Mannheim, Mannheim, Germany, 2011.
[68] M. Armstrong and M. Adendorff: "Data warehouse system" U.S. Patent Application 09/987,905, 2002.
[69] G. G. Parker and M. Van Alstyne, *Unbundling in the Presence of Network Externalities*, Mimeo, New York, NY, USA, 2002.
[70] M. Rysman, S. Engerman, L. Gallman et al., "The economics of network industries," *Journal of Economic Literature*, vol. 40, no. 2, p. 556, 2002.
[71] I. R. Segal and M. D. Whinston, "Exclusive contracts and protection of investments," *SSRN Electronic Journal*, vol. 31, no. 4, pp. 603–633, 2000.
[72] T. Doganoglu and J. Wright, "Exclusive dealing with network effects," *International Journal of Industrial Organization*, vol. 28, no. 2, pp. 145–154, 2010.
[73] L. Leung and P. S. N. Lee, "The influences of information literacy, internet addiction and parenting styles on internet risks," *New Media & Society*, vol. 14, no. 1, pp. 117–136, 2012.
[74] J.-C. Rochet and T. Jean, "Competition policy in two-sided markets," *Handbook of Antitrust Economics*, pp. 543–582, 2008.
[75] J. P. Choi, "Tying in two-sided markets with multi-homing," *The Journal of Industrial Economics*, vol. 58, no. 3, pp. 607–626, 2010.
[76] A. Amelio and B. Jullien, "Tying and freebies in two-sided markets," *International Journal of Industrial Organization*, vol. 30, no. 5, pp. 436–446, 2012.
[77] A. Hagiu, "Pricing and commitment by two-sided platforms," *The RAND Journal of Economics*, vol. 37, no. 3, pp. 720–737, 2006.
[78] D. L. Reiley and D. F. Spulber, "Business-to-business electronic commerce," 2000, https://www.vanderbilt.edu/econ.
[79] J. Farrell and P. Klemperer, "Chapter 31 coordination and lock-in: competition with switching costs and network effects," *Handbook of Industrial Organization*, vol. 3, no. 6, pp. 1967–2072, 2007.
[80] J. Brown and J. Morgan, "How much is a dollar worth? Tipping versus equilibrium coexistence on competing online auction sites," *Journal of Political Economy*, vol. 117, no. 4, pp. 668–700, 2009.
[81] B. O. Edelman, "Internet advertising and the generalized second-price auction: selling billions of dollars worth of keywords," *The American Economic Review*, vol. 97, no. 1, pp. 242–259, 2007.
[82] B. J. Levin, "Matching and price competition matching," *American Economic Review*, vol. 96, no. 3, pp. 652–668, 2006.
[83] M. A. Cusumano and A. Gawer, "The elements of platform leadership," *IEEE Engineering Management Review*, vol. 31, no. 1, p. 8, 2003.
[84] L. Einav, T. Kuchler, J. Levin et al., "Learning from Seller Experiements in Online Markets," National Bureau of Economic Research, Cambridge, MA, USA, 2011.
[85] M. Gentzkow and J. M. Shapiro, "What drives media slant? Evidence from U.S. daily newspapers," *Econometrica*, vol. 78, no. 1, pp. 35–71, 2010.
[86] J. W. Roberts and A. Sweeting, "Competition versus auction design," 2010.
[87] T. Doganoglu, "Dynamic price competition with consumption externalities," *Netnomics*, vol. 5, no. 1, pp. 43–69, 2003.
[88] M. F. Mitchell and A. Skrzypacz, "Network externalities and long-run market shares," *Economic Theory*, vol. 29, no. 3, pp. 621–648, 2006.
[89] S. Markovich and J. Moenius, "Winning while losing: competition dynamics in the presence of indirect network effects," *International Journal of Industrial Organization*, vol. 27, no. 3, pp. 346–357, 2009.
[90] J. Chen, U. Doraszelski, and J. E. Harrington Jr., "Avoiding market dominance: product compatibility in markets with network effects," *The RAND Journal of Economics*, vol. 40, no. 3, pp. 455–485, 2007.
[91] R. Driskill, *Monopoly and Oligopoly Supply of a Good with Dynamic Network Externalities*, vol. 47, Vanderbilt University,Nashville, TN, USA, 2007.
[92] L. Cabral, "Dynamic price competition with network effects," *The Review of Economic Studies*, vol. 78, no. 1, pp. 83–111, 2011.
[93] A. McAfee and E. Brynjolfsson, "Machine, platform, crowd: Harnessing our digital future," *WW Norton & Company*, New York, NY, USA, 2017.

Hindawi
Scientific Programming
Volume 2020, Article ID 8856640, 15 pages
https://doi.org/10.1155/2020/8856640

Research Article

Underwater No-Reference Image Quality Assessment for Display Module of ROV

Di Wu, Fei Yuan, and En Cheng

Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Xiamen University, Xiamen 361001, China

Correspondence should be addressed to En Cheng; chengen@xmu.edu.cn

Received 24 April 2020; Revised 22 May 2020; Accepted 3 August 2020; Published 28 August 2020

Academic Editor: Chao Huang

The optical images collected by remotely operated vehicles (ROV) contain a lot of information about underwater (such as distributions of underwater creatures and minerals), which plays an important role in ocean exploration. However, due to the absorption and scattering characteristics of the water medium, some of the images suffer from serious color distortion. These distorted color images usually need to be enhanced so that we can analyze them further. However, at present, no image enhancement algorithm performs well in any scene. Therefore, in order to monitor image quality in the display module of ROV, a no-reference image quality predictor (NIPQ) is proposed in this paper. A unique property that differentiates the proposed NIPQ metric from existing works is the consideration of the viewing behavior of the human visual system and imaging characteristics of the underwater image in different water types. The experimental results based on the underwater optical image quality database (UOQ) show that the proposed metric can provide an accurate prediction for the quality of the enhanced image.

1. Introduction

In recent years, there have been a growing number of ocean-related activities, such as aquaculture, hydrological exploration, and underwater archaeology. The optical images collected by the observational remotely operated vehicle (ROV) provide very convenient conditions for these activities, and high-quality underwater images play an essential role in these activities. However, due to the absorption and scattering effects of water limiting the visibility of the underwater objects, the images captured by an optical sensor of ROV often suffer from diminished color (color distortion), which affects our understanding of underwater conditions, so poor quality underwater images need to be enhanced. It is worth noting that not all underwater images need to be enhanced, as shown in Figure 1, because bodies of water exhibit extreme differences in their optical properties. Some lakes are as clear as distilled water, and some change colors several times a year, among white, blue, green, and brown [1]. In the ocean, coastal harbors are often murky, while offshore waters are blue and clear. Simply put, whether an image needs to be enhanced depends on whether the visibility of the underwater objects in the image is good. However, the existing display module of observational ROV either displays the captured image directly in the terminal or integrates an enhancement algorithm in the system to display the captured image after enhancement. However, the existing display module of observational ROV either displays the captured image directly in the terminal or integrates an enhancement algorithm in the system to display the captured image after enhancement. None of them determines whether the image needs to be enhanced or not. Also, currently, there is no underwater image enhancement algorithm suitable for any scene, so we need reliable underwater image quality metrics to help us preassess whether the captured image needs to be enhanced or not and to monitor the quality of the enhanced image.

The most accurate methods of image quality estimation are subjective image quality assessment (IQA). However, subjective IQA is expensive, time-consuming, and impractical for real-time implementation and system integration. In order to automatically estimate image quality and

(a) (b)

Figure 1: (a) Images that do not need to be enhanced. (b) Images that need to be enhanced.

save workforce and resources, a reliable underwater objective image quality metric needs to be designed. In underwater image processing scenarios, an ideal reference image is usually not available, so no-reference (NR) IQA is the best choice for evaluating underwater image quality.

Some generic image quality measures, such as histogram analysis [2], the Variance [3], Image Entropy [4], and color image quality measure (CIQI) [5], have been developed. Popular BRISQUE [6] and LPSI [7] have also been proposed, which summarize the statistical rules of natural images and calculate the degree of deviation of distorted images. However, these objective measures are not designed specifically for underwater images. They fail to consider the strong absorption and scattering effects of the water, and they are not applicable to underwater images. There are also NR IQA based on in-depth learning, such as DIQA [8], Deep IQA [9], and RankIQA [10], which perform well in air images. The deep learning method has a strong learning ability and can automatically extract image features. However, the method requires a large amount of data (usually more than 5000 images) for training, and the acquisition of subjective scores (as the ground truth during training) is expensive and time-consuming. Currently, there is no relevant dataset available for training in the underwater image field, so it is temporarily impossible to design NR-IQA based on deep learning for underwater images.

Some paper [11] pointed out that the overall quality of an image can be effectively obtained by combinations of image attribute measures. At present, the most commonly used underwater image quality measures, UCIQE [12], UIQM [11], and CCF [13], are designed based on this principle. The UCIQE [12] proposed by Yang Miao is a linear combination of the standard deviation of chroma, the contrast of luminance, and the average of saturation. Karen Panetta's UIQM [11] is a linear combination of colorfulness, sharpness, and contrast. The CCF proposed by Yan Wang et al. [13] starts from the imaging analysis of underwater absorption and scattering characteristics, calculates the fog density index, and evaluates the underwater image quality by combining the color index, contrast index, and fog density index. They all determine the weighting coefficients by multivariate linear regression from the training image set. However, regardless of the performance of the training set, the generalization ability of these methods is largely limited by the training image samples. At the same time, the attention mechanism of the human visual system (HVS) [14] has not been paid enough attention in the underwater image evaluation. In underwater scenes, the image quality of the target object has higher research value and practical significance than that of the ocean background, which does not belong to the region of interest (ROI). Moreover, the three commonly used underwater metrics measure the image quality from the perspective of image statistics, and the robustness is not high. This results in an overemphasis on color richness. This paper holds that, in addition to the color fidelity in the statistical sense, the color fidelity of objects is also very important from the perspective of pixels. It is worth noting that the color fidelity here refers to whether the image color is reasonable, not to say the difference between the object color in the image and the real object color. Most of the underwater natural images are blue-green due to color selective attenuation, and the color is single, and the color richness is not ideal (as shown in Figures 2(a)–2(c)). After the enhancement algorithm processing, the underwater image can generally eliminate the color attenuation from the vision, and the color richness is greatly improved, as shown in Figures 2(d)–2(f), but the color fidelity of the enhanced image is questionable, and the color of the fish in Figures 2(d) and 2(e) is not reasonable, and the artificial facilities in Figure 2(f) are obviously different from what we know. That is, Figures 2(a)–2(c) have high color fidelity (because they are real natural images, although the color of objects in them is different from that of real objects, the pixel color is reasonable), but the color richness is very low; Figures 2(d)–2(f) have low color fidelity and high color richness. That is, the underwater absorption and scattering characteristics cause image color distortion (where the type of distortion is the large difference between the object color

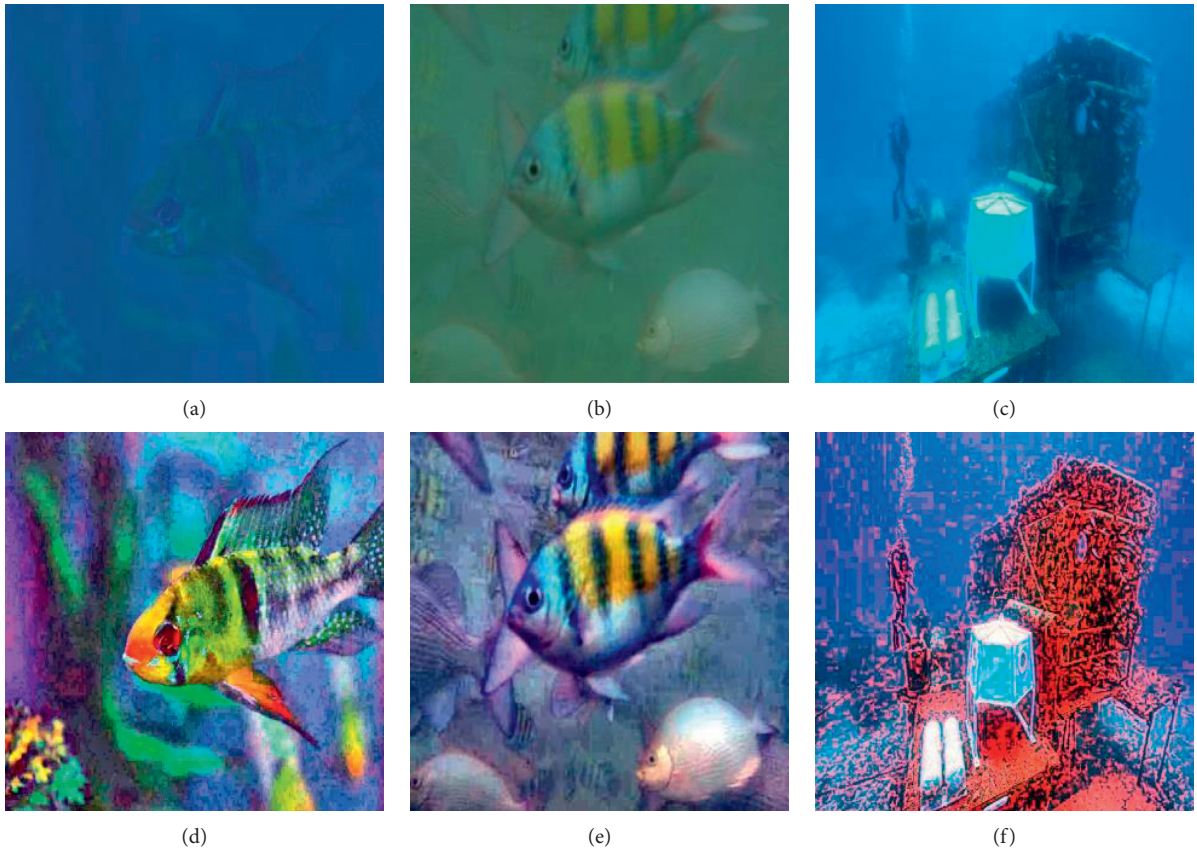


Figure 2: (a), (b), and (c) are original images. (d), (e), and (f) are images after enhancement algorithm. Although the color attenuation is eliminated visually, the color is still distorted and it is oversaturated.

in the image and the color of the real object), the image tends to be blue-green, and the color richness of the image is reduced. Overemphasis on color richness can also result in color distortion (in this case, the kind of distortion refers to the unreasonable color in the image), which affects the viewing effect of the image and subsequent use of the image.

In view of the shortcomings of existing metrics, this paper proposes a no-reference image quality predictor NIPQ. NIPQ is designed with a three-stage framework. The first stage focuses on the attention mechanism of the human visual system (HVS), which can also be interpreted as ROI. Because in the IQA field our ROIs are not fixed, some are task-driven; some are data-driven; some such as fish, corals, divers, or even artifacts of unknown shape may be of interest to us. For more applications, we interpret the foreground area (nonocean background) as ROI. This paper extracts ROI based on background/foreground and focuses on the image quality of ROI. The second stage considers the impact of color richness on image quality. As the distance between the camera and the object increases (horizontally), the color of the object in the underwater image will keep approaching blue and green [1]. At the same time, as the position of the optical sensor gets deeper, the object will be farther away from the sunlight source, and the color of the object will be darker, and the contrast will be lower. It can be understood that if the ROI of a natural underwater image has good color richness, its image quality will be significantly better than that of a low color richness image. The third stage considers the fidelity of the color. As mentioned earlier, if NR IQA overemphasizes color richness, it will cause the enhanced underwater image to become oversaturated, which is also a form of color distortion. Inspired by the underwater image formation model, we distinguish the water types (yellow water, green water, and blue water) in the image by the ocean background area of the image and estimate the reasonable range of pixel intensity of ROI in the enhanced underwater image from the perspective of pixels. In this stage, the difference between the reasonable range of pixel intensity and the ROI pixel intensity of the actual enhanced image is used to represent the rationality of the enhanced image, that is, color fidelity. Finally, in the fourth stage, color richness and color fidelity are systematically integrated for quality prediction.

In order to measure the performance of NIPQ, a underwater optical image quality database (UOQ) is established. The database contains some typical underwater images and their mean opinion scores (MOS). Based on the comprehensive analysis of all experimental results, the contribution of NIPQ proposed in this paper is summarized as follows:

(a) It is a kind of NR IQA inspired by underwater imaging characteristics. By considering the color attenuation of images in different water bodies, the color fidelity and color richness metrics are proposed.

(b) By adopting a suitable ROI extraction method for the underwater IQA field, ROI and IQA are effectively combined due to the block strategy in the ROI extraction method.

(c) It is superior to many commonly used IQA metrics and can effectively evaluate the performance of the image enhancement algorithm and can be used as quality supervision.

(d) We propose NR IQA-based underwater smart image display module, which embodies the role of our IQA in application.

We arrange the reminder of this paper as follows. Section 2 describes the NR IQA-based underwater smart image display module, which is the application background of NR IQA. Section 3 describes the detail of our NIPQ metric explicitly. Section 4 describes the establishment of our database UOQ for evaluating IQA performance, which consists of underwater optical images and their enhanced images. In Section 5, performance comparisons of the NIPQ metric with selected existing NR IQA methods are performed. We conclude this paper in Section 6.

2. NR IQA-Based Underwater Smart Image Display Module

Most of the underwater images captured by optical sensors have practical applications. For underwater images with severe color distortion, image enhancement is often needed before the terminal display. However, not all underwater images need to be enhanced. We believe that whether the image needs to be enhanced or not depends on the visibility of the underwater objects. Besides, because no image enhancement algorithm can achieve good results in all scenes, NR IQA can be used as a guide for image enhancement, so that the system can automatically select more appropriate image enhancement algorithm in real time. From the application point of view, the framework of the NR IQA based display module is shown in Figure 3. The traditional image display module only provides a single image enhancement scheme or displays the image directly, which cannot flexibly cope with the different water environment. The display module proposed in this paper builds various image enhancement algorithms into the Image Enhancement Algorithm Database. The system can choose a more appropriate enhancement algorithm according to the results of NR IQA. Firstly, the input underwater image is preassessed, and the natural image with less color distortion is directly displayed. And for the natural image with severe color distortion, the default enhancement algorithm in the Image Enhancement Algorithm Database is used to enhance the image. The Selector automatically determines whether to enable the alternative image enhancement scheme and which alternative scheme to choose according to the results of NR IQA.

According to the above analysis of NR IQA-based underwater smart image display module and the consideration of the characteristics of underwater image in Section 1, this paper uses the color richness of ROI and color fidelity of ROI to estimate image quality. In the display module proposed in our paper, the color richness of ROI is used as the metric of pre-NR IQA in the display module, and the NIPQ, which combines ROI, color richness, and color fidelity, will be used as the metric of NR IQA in the display module.

The ROI extraction method is based on background/foreground. The block strategy in the extraction method helps ROI and IQA better combine. The ROI extraction method is introduced in Section 3.1 in detail. The color richness represents the distribution of image color in a statistical sense, which is described by the spatial characteristics of image in CIE XYZ space and detailed in Section 3.2. The color fidelity is based on the underwater image formation model in the sense of pixel, which is used to describe whether the pixel intensity is within a reasonable range. It is introduced in Section 3.3 in detail.

3. Proposed NIPQ Metric

3.1. ROI Extraction Based on Background/Foreground. Considering that the final receiver of display module of ROV is often human, it is particularly important that IQA can reflect the feeling of human eyes well. The mechanism of human visual attention is an important feature of HVS. The mechanism of human visual attention enables the brain to quickly understand the overall information of the image and obtain the regions that need attention. Then the brain begins to focus on the target information and suppress other background information. Therefore, the human eye is usually sensitive to the damage of the area of concern. At the same time, compared with the ocean background, high-quality ROI has better practical significance and value. Therefore, it is necessary to introduce ROI into image quality assessment.

Researchers usually get ROI by saliency detection or object detection [15, 16]. Different from the image in the air, the contrast of most underwater images is low, and the traditional significance detection method in the air is not applicable in the underwater. At present, there is no robust saliency detection algorithm in underwater image field. Some researchers combine image enhancement with saliency detection [17]. Some researchers combine fish localization and saliency detection [18]. In IQA, the purpose of the metric is to evaluate the enhancement algorithm, and the ROI of the underwater image is not always one or several fixed categories of targets with predictable shapes. Therefore, the above method is not applicable. Considering that, compared with the target, the underwater background features are easier to be recognized, this paper extracts ROI based on background/foreground, and the process is shown in Figure 4. In order to better combine ROI and IQA in the next steps, the preprocessed underwater image is divided into $m \times n$ image blocks (we call it the block strategy).

Then, we map the boundary connectivity $\mathrm{BndCon}(p_i)$ (definition in [19]) of block i region p_i by (1) and obtain the background region probability w_i^{bg}:

$$w_i^{bg} = 1 - \exp\left(1 - \frac{\mathrm{BndCon}^2(p_i)}{2\sigma^2_{\mathrm{BndCon}}}\right), \tag{1}$$

where $\sigma^2_{\mathrm{BndCon}} = 1$. The background block, the target block, and the uncertain block are initially divided by using w_i^{bg},

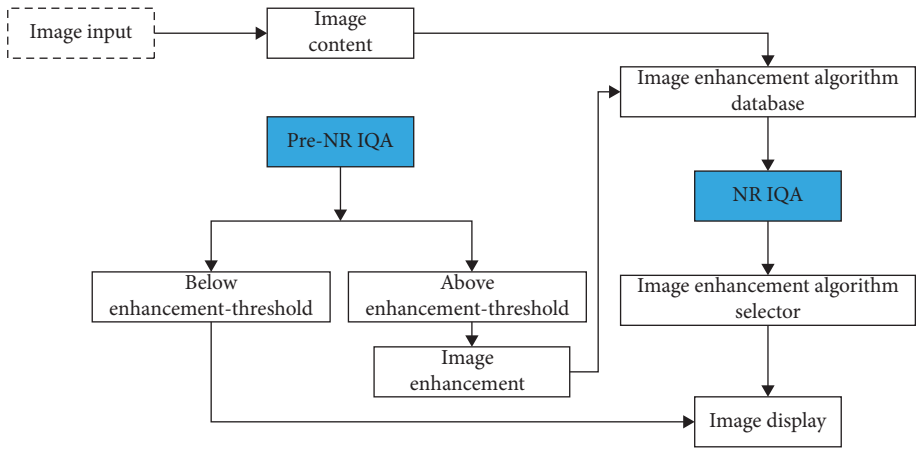


Figure 3: The framework of the NR IQA-based underwater smart display module.

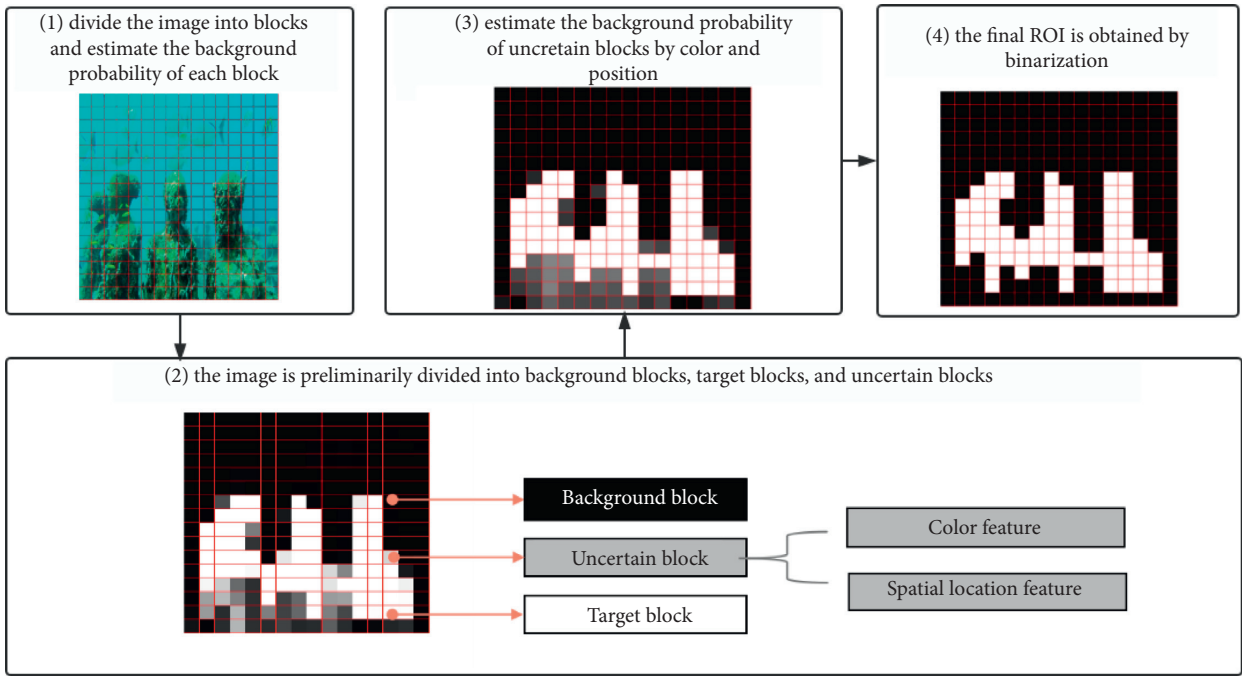


Figure 4: ROI extraction process.

threshold$_{bg}$, and threshold$_{\text{roi}}$. Then the color feature and spatial position feature of the block are used as the correction of the uncertain block to help judge the background probability of the uncertain block, which is expressed by a mathematical formula as shown in the following equation:

$$W_{BG(p)} = \begin{cases} 1, & w_i^{bg} > \text{threshold}_{\text{bg}}, \\ 1 - \sum_{i-1}^{N} d_{\text{app}}(p, p_i) w_{\text{spa}}(p, p_i) w_i^{bg}, & \text{threshold}_{\text{roi}} < w_i^{bg} < \text{threshold}_{\text{bg}}, 0, w_i^{bg} < \text{threshold}_{\text{roi}}, \end{cases} \tag{2}$$

$$w_{\text{spa}}(p, p_i) = \exp\left(-\frac{d_{\text{spa}}(p, p_i)}{2\sigma^2_{\text{spa}}}\right), \tag{3}$$

where $d_{\text{app}}(p,q)$ is the color similarity between blocks p and q, which is calculated by the Euclidean distance between the average colors of blocks p and q. $d_{\text{spa}}(p,q)$ is the Euclidean distance between blocks p and q.$w_{\text{spa}}(p,q)$ is obtained after mapping according to (3), among which $\sigma^2{}_{\text{spa}} = 0.25$. Finally, we use the method of maximum variance between classes to get the final ROI.

3.2. Color Richness. With the aggravation of the phenomenon of color attenuation, the color of the natural underwater image will become less and less, and the visibility of the object will become worse. Therefore, the color richness of ROI is a simple and fast metric to measure whether the color distortion of natural underwater image is serious, which is suitable for the evaluation of image quality.

In this paper, the richness of color is measured by the spatial characteristics of color in CIE *XYZ* color space. The color richness should not only include color diversity but also consider the lightness distribution, so *XYZ* color space is a good choice. CIE *XYZ* color space can represent all colors, and the *Y* parameter is the measurement of color lightness. According to the *XYZ* color space distribution of the two images shown in Figure 5, the wide distribution of image color, respectively, in the three dimensions of *X*, *Y*, and *Z* does not mean that the color richness is good. That is because the three components of *X*, *Y*, and *Z*, have a certain correlation. So the spatial characteristics of color can better represent the distribution of color. According to (4), the image color divergence in *XYZ* color space is defined to determine the color richness of the image:

$$C_d = \sum_{mn} \operatorname{dis}\left(P_{mn}\min, P_{mn}\max\right) \times \max\left(\operatorname{dis}\left[P_{mn}(i,j), P_{mn}\min, P_{mn}\max\right]\right) \times \frac{1}{2}, \tag{4}$$

where dis represents the shortest distance between two points or between points and lines, and mn belongs to *X*-*Y*, *Y*-*Z*, and *X*-*Z* sections. P_{mn} min and P_{mn} max represent the closest and farthest points from the origin, respectively.

3.3. Color Fidelity. As mentioned in Section 1, the enhanced image may be oversaturated/pseudobright (as shown in Figure 2). If too much attention is paid to the color richness, the color of ROI in the image will deviate from the color of real objects. Therefore, we should not only consider the color richness of the enhanced underwater image but also consider the color fidelity of ROI, that is, whether the intensity of pixels is within a reasonable range.

It is necessary to understand the formation and degradation of underwater images if we want to estimate a reasonable range of intensity of pixels. The formation of the underwater image is dominated by the following factors [1, 20, 21]:

$$I_c = J_C e^{-\beta(D/C)\times z} + B_c^{\infty}\left(1 - e^{-\beta(B/C)\times z}\right), \tag{5}$$

where $C = R$, G, B is the color channel, I_c is the underwater image captured by the camera, and $J_C e^{-\beta(D/C)\times z}$ is the direct signal, recorded as D_c. $B_c^{\infty}(1 - e^{-\beta(B/C)\times z})$ is a backscattered signal, which is recorded as B_c. z is the distance between the camera and the photographed object; B_c^{∞} is the obscured light. J_C is the unattenuated scene, that is, the RGB intensity of the surface captured by the sensor with the spectral response $S_c(\lambda)$ at the distance $z0 = 0$ (generally $z0$ is regarded as 0):

$$J_C = \frac{1}{k_c}\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\rho(\lambda)E(d,\lambda)d\lambda. \tag{6}$$

k_c is the camera's scaling constant. β_c^D and β_c^B have a certain dependence on the spectrum of distance z, reflectivity $\rho(\lambda)$, ambient light $E(\mathrm{d},\lambda)$, camera's spectral response $S_c(\lambda)$, scattering coefficient $b(\lambda)$, and beam attenuation coefficient $\beta(\lambda)$, as shown in (7) and (8). $z0$ and $(z0 + z)$ are the starting and ending points along the line of sight.

$$\beta_c^D = \frac{\left[\ln\left(\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\rho(\lambda)E(d,\lambda)e^{-\beta(\lambda)z0}d\lambda \Big/ \int_{\lambda_1}^{\lambda_2} S_c(\lambda)\rho(\lambda)E(d,\lambda)e^{-\beta(\lambda)(z0+z)}d\lambda\right)\right]}{z}, \tag{7}$$

$$\beta_c^B = -\frac{\left[\ln\left(\int_{\lambda_1}^{\lambda_2} S_c(\lambda)B^{\infty}(\lambda)e^{-\beta(\lambda)z}d\lambda \Big/ \int_{\lambda_1}^{\lambda_2} S_c(\lambda)B^{\infty}(\lambda)d\lambda\right)\right]}{z}. \tag{8}$$

So, we can calculate the unattenuated scene J_C as

$$J_C = D_c e^{\beta(D/C)\times z} = \left[I_c - B_c^{\infty}\left(1 - e^{-\beta(B/C)\times z}\right)\right] \times e^{\beta(D/C)\times z}. \tag{9}$$

We need to estimate the reasonable range of values range $[J'_{c,\min}, J'_{c,\max}]$ of RGB intensity of each pixel in the foreground (that is, ROI). The color fidelity metric defined by (10) is calculated by the out-of-range part of the

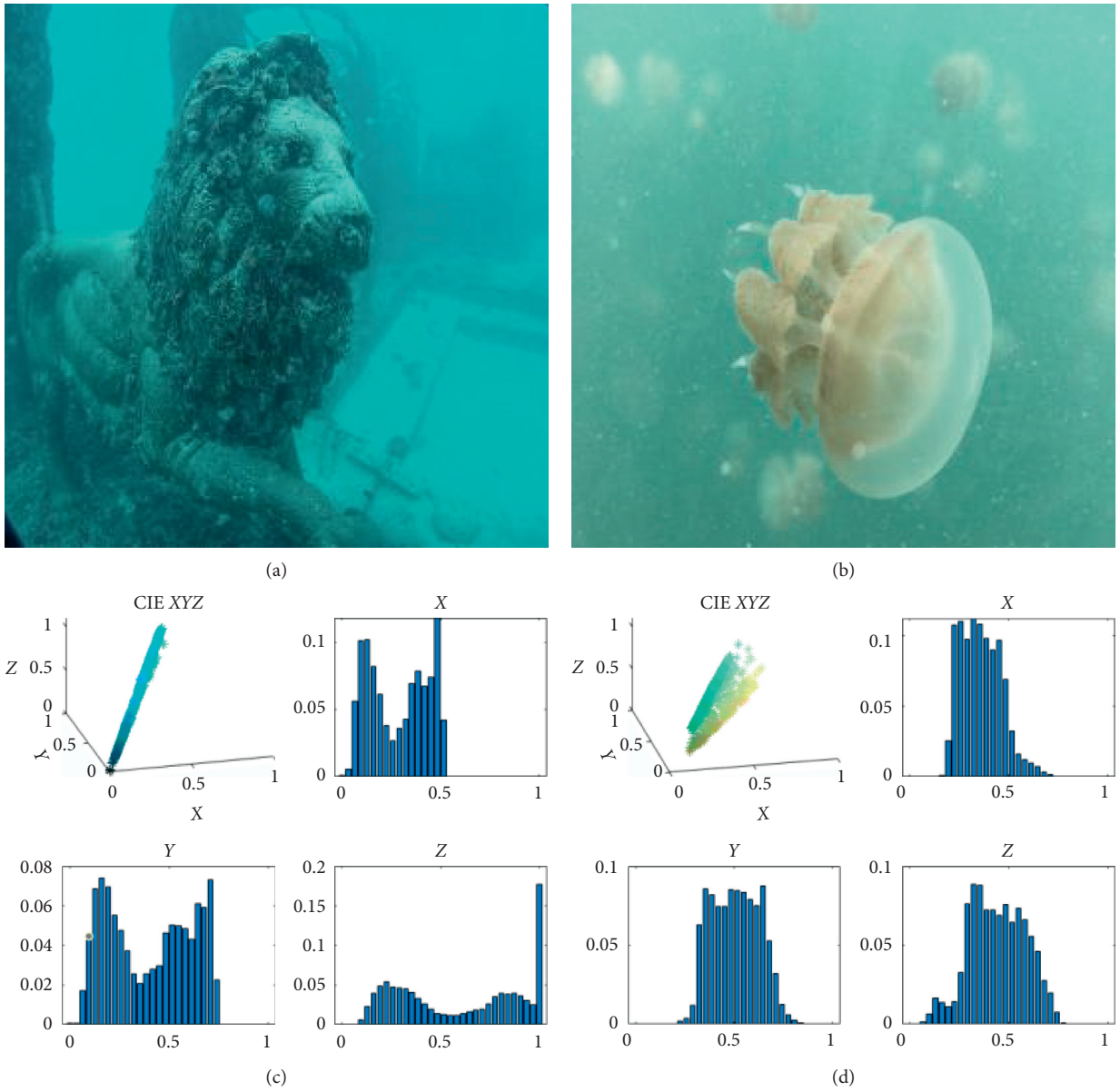


Figure 5: Color distribution in CIE *XYZ* space.

enhanced underwater image $\widehat{J}_c$. The process is shown in Figure 6.

Firstly, the veiling light B_c^∞ of the underwater image I_c is estimated. Backscatter increases exponentially with z and eventually is saturated [1]. In other words, at infinity, $I_c = B_c^\infty$. Referring to [22], we assume an area without objects is visible in the image, in which the pixels' color is determined by the veiling light alone. Such areas are smooth and have no texture. This assumption often holds in the application scenarios of our IQA. First, the edge graph of the image is generated. Then, the threshold value is set, and the largest connected pixel area is found. The veiling light is the average color of the pixels in these areas.

Next, we estimate the type of water body in the underwater image I_c by veiling light B_c^∞. The reason for estimating the type of water is that the common notion that water attenuates red colors faster than blue/green only holds for oceanic water types [1]. We simplified Jerlov water types [23] into blue water (Jerlov I–III), green water (Jerlov 1c–3c), and yellow water (Jerlov 5c–9c) and simulated the average RGB value of perfect white surface appearance under these three water types (data from [23], using D65 light source, Canon camera 5D Mark2, $\rho(\lambda) = 1$). We calculate Euclidean distance between the veiling light B_c^∞ and the average RGB value and estimate water body type based on Euclidean distance.

Then, we calculate the reasonable intensity range $[J'_{c,\min}, J'_{c,\max}]$ of each pixel after the enhancement of the underwater image. β_c^D varies most strongly with range z [1]. So, the most important thing to calculate the range is to estimate the distance z in addition to the water body type. Due to the limitation of real conditions, the distance z of the

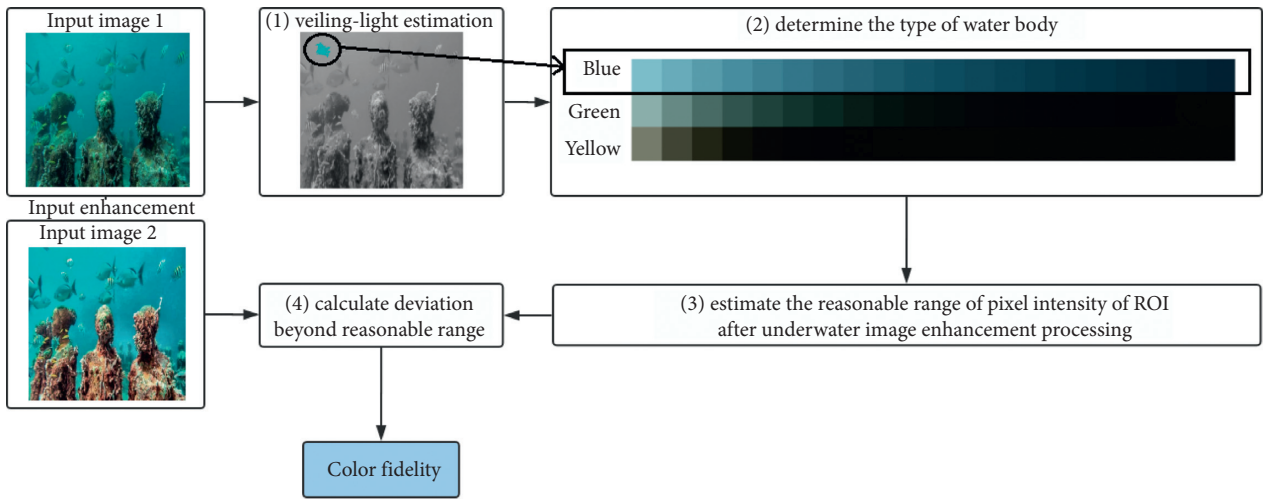


Figure 6: The estimation process of color fidelity.

object in the image cannot be obtained, so it is necessary to roughly estimate the possible range of the distance z. For the foreground, the distance z from the camera is approximately the same; for the background, the distance z from the camera tends to be infinite. We assume that the distance z from the camera is the same at each part of the foreground, and there may be white objects. Therefore, the distance z, which makes the J_C of the foreground pixels under the three RGB channels not greater than 255 and not less than 0, is considered as the possible distance. In order to simplify the calculation, the attenuation coefficient β_c^D of white in three color channels $C = R$, G, B is adopted for all colors (using $\rho(\lambda)$ in Macbeth ColorChecker).

Finally, the color fidelity defined by (10) is calculated:

$$C\_f = \left[1 - \frac{(\mathrm{Sum}_{\mathrm{oor}}/255)}{\mathrm{Num}_{\mathrm{oor}} \times 3}\right]^2. \tag{10}$$

$\mathrm{Num}_{\mathrm{oor}}$ represents the number of pixels in ROI block, and $\mathrm{Sum}_{\mathrm{oor}}$ represents the total number of pixel intensity deviations that are not within a reasonable range.

We make some qualitative analysis on the influence of simplification on $J'_{c,\min}$ and $J'_{c,\max}$ during the calculation. As shown in Figure 7(b), (8) is used to calculate the broadband (RGB) attenuation coefficient β_c^D (using $\rho(\lambda)$ of the color block in Macbeth ColorChecker, depth $d = 1$m, distance $z = 1$m) of seven common colors of red, orange, yellow, green, green, blue, and purple (Figure 7(a)) under all Jerlov water types. It can be seen that the β_c^D difference of each color is not large in the same scene. Figure 7(c) shows the influence of different camera types on β_c^D in three types of water bodies. The influence of camera parameters on the attenuation coefficient β_c^D is not significant. The experimental results in [1] also prove this view.

3.4. NIPQ Metric. Section 3.1, Section 3.2, and Section 3.3 above, respectively, introduce the ROI extraction method, color richness in statistical sense, and color fidelity in pixel sense. In this paper, the color richness of ROI and color fidelity of ROI are combined by the multiplication model to get our NIPQ. The common underwater image evaluation models UIQM [11], UCIQE [12], and CCF [13] with multiparameters use linear weighting to measure the comprehensive quality of the image. We consider that if a submetric points to a very low value (indicating low quality), the subjective feeling of the whole image will be very poor regardless of other metrics. Therefore, this paper uses the multiplication model to generate the overall underwater image quality assessment, as follows:

$$C_R = \left|C_d(\mathrm{ROI})\right| \times C_f(\mathrm{ROI}). \tag{11}$$

represents normalization, C_R represents color quality of ROI block, $C_R \epsilon (0, 1)$, and the larger the value is, the higher the image quality is.

The overall process of NIPQ is shown in Figure 8, which is divided into four step. Firstly, the ROI of the original image (not enhanced) is extracted based on background/foreground. Then, the color richness of ROI of the enhanced underwater image is estimated. Then the ocean background information is extracted from the original image, from which the water body type is estimated, and the reasonable range of pixel intensity distribution is estimated. According to the estimated range, the ROI color fidelity of the enhanced underwater image is estimated. Finally, the two metrics of color richness and color fidelity are integrated to obtain the comprehensive NIPQ metric for the whole underwater image.

4. UOQ Database

In order to better evaluate the performance of NIPQ metric, we built an underwater optical image quality database UOQ.

Image Selection. In order to fully consider various underwater scenes, we selected 36 typical underwater optical images with a size of 512×512. These images include blue water, green water, yellow water, dark light, single object, multiobject, simple texture and complex texture, serious

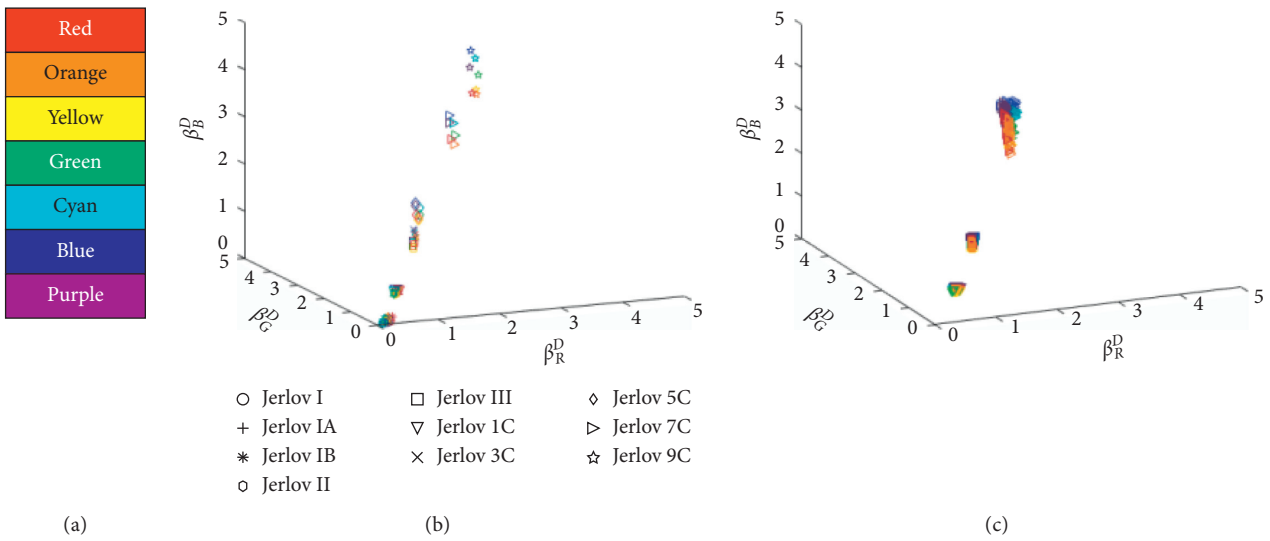


FIGURE 7: Qualitative analysis on the influence of simplification on β_c^D. (a) Seven common colors. (b) β_c^D of different colors in different water bodies. (c) β_c^D of different colors under different cameras. We simplify the types of water into three. It can be seen from (b) that the calculation error of yellow water caused by simplification is larger than that of blue and green water body caused by simplification. The yellow water body in the underwater image is not common, so the simplification of water body type is applicable to most occasions.

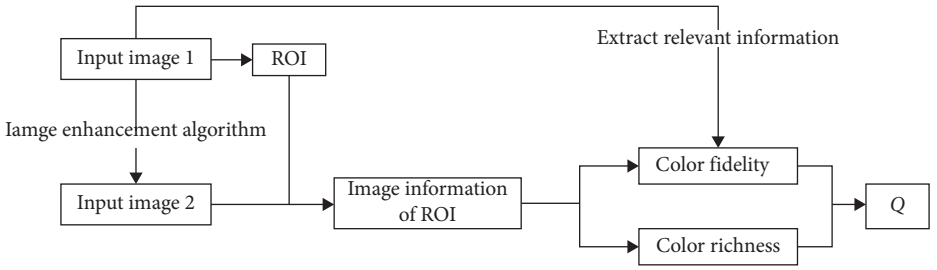


FIGURE 8: Overall process of NIPQ.

color distortion, and a little color distortion. Considering that there is no general ROI related dataset in the field of underwater image, we label their foreground region (ROI) pixel by pixel to prove the reliability of ROI in this paper. And, we use five popular image enhancement algorithms (white balance algorithms [24], Fu's algorithm [25], multifusion algorithm [26], histogram equalization [27], and Retinex [28]) to process these 36 natural images. 180 enhanced images were obtained. Some images and their enhanced images processed by the white balance algorithm [24] are shown in Figure 9.

Evaluation Methods and Evaluation Protocols. In this database, the single incentive evaluation method is used. Volunteers only watch one image to be evaluated each time, and each image only appears once in a round of evaluation. After each image was displayed, volunteers gave subjective quality scores to the corresponding images. Underwater optical images usually have practical applications, so volunteers will not be affected by any aesthetic factors in the process of subjective quality assessment, and the evaluation protocols are shown in Table 1.

Choosing Volunteers. In order to avoid the evaluation bias caused by prior knowledge, none of the volunteers had the experience of image quality assessment. We consider the strong application background of underwater images, so all volunteers selected are graduate students with relevant work experience in underwater acoustic communication, underwater detection, and so on.

All the obtained subjective scores are used to calculate the mean opinion scores (MOS). Note $S_{i,j}$ as the subjective score of the image j by the i-th volunteer and N_j as the number of subjective scores obtained by image j. MOS is calculated as follows:

$$\mathrm{MOS}_j = \frac{I}{N_j} \sum_i S_{ij}. \tag{12}$$

We draw a histogram about MOS of all images in the database, as shown in Figure 10. It can be seen that our image covers a wide range of quality, which is conducive to the design of IQA. And there are many images with scores in the middle score segment because the volunteer will try to avoid giving extreme scores when scoring images. It also can be

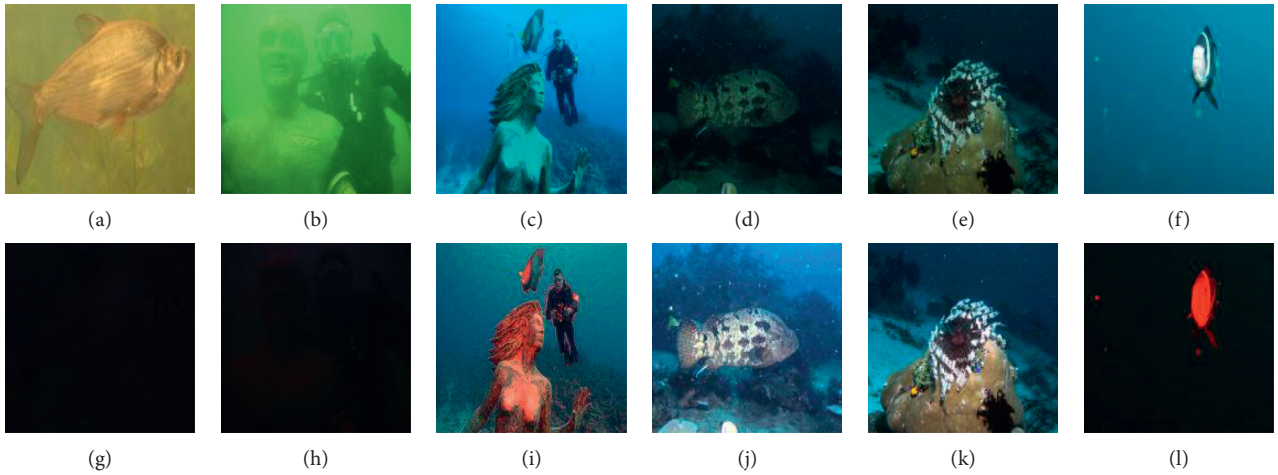


Figure 9: Underwater image processed by white balance algorithm [24]. (a)–(f) are the original images. Their MOS are 2.40, 1.70, 3.00, 1.30, 2.55, and 4.05, respectively. (g)–(l) are enhanced images. Their MOS are 1.05, 1.15, 2.05, 2.80, 4.55, and 1.15, respectively.

Table 1: Evaluation protocols.

Score	Comprehensive feelings
5	The subjective feeling is excellent, foreground information is recognizable, and no color distortion is felt
4	The subjective feeling is good, the foreground information is visible and recognizable, there is a small amount of perceptual distortion, but it does not affect the extraction of important information
3	The subjective feeling is general, part of the information in the foreground is damaged, and a small amount of important information is lost due to distortion
2	The subjective perception is poor, and only the general outline of the foreground content can be distinguished; the distortion leads to the loss of some important information
1	The subjective feeling is very poor, it is difficult to recognize the foreground content, and it is almost impossible to extract any effective information from the image

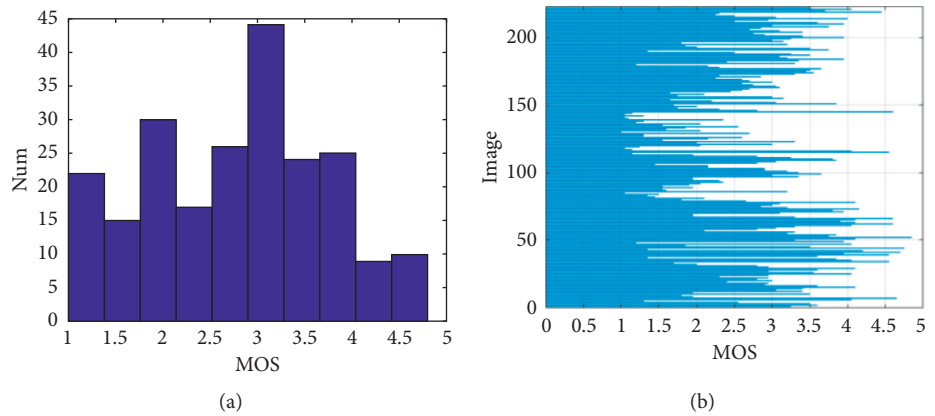


Figure 10: (a) Frequency histogram about MOS and (b) MOS of all images.

seen that the lower quality image is slightly more than the higher quality underwater image. This is because most underwater images have the characteristics of blue-green and poor contrast, and sometimes the quality of the enhanced image is still not ideal. In the practical applications, more robust enhancement algorithms will be built into the underwater image enhancement algorithm database of the display module mentioned in Section 2.

5. Experiment

In combination with the UOQ database, we mainly evaluate the performance of IQA through five criteria. The prediction monotonicity of IQA is measured by the Spearman rank order correlation coefficient (SROCC) and Kendall's rank order correlation coefficient (KROCC). The prediction accuracy of IQA is measured by the Pearson linear correlation

coefficient (PLCC). Root mean square error (RMSE) is used to measure the prediction consistency of IQA. The mean absolute error (MAE) is also used to evaluate the performance of IQA. The high values (close to 1) of SROCC, PLCC, and KROCC and the low values (close to 0) of RMSE and MAE indicate that IQA has a better correlation with subjective scores.

The selected IQA metrics for performance comparison include the following:

(1) The popular no-reference metrics underwater: UIQM [11], UCIQE [12], and CCF [13]

(2) The popular no-reference metrics in the air: BRISQUE [6] and LPSI [7]

(3) Common color metrics for underwater images: UICM [11] and variance of chromaticity (Var Chr) [29]

For the BRISQUE, a low score means high quality, and other metrics are that the higher the score, the better the quality.

5.1. Effect Analysis of Introducing ROI into IQA. In order to observe the influence of the introduction of ROI on the quality evaluation of underwater images, we need to combine ROI with the popular underwater no-reference IQA. The block strategy mentioned in Section 3.1 is necessary because it helps us combine ROI with IQA better. According to the block fusion strategy represented by (13), we combine image block with IQA and get comprehensive quality score. We can observe the change of correlation between objective metrics and MOS before and after combining with ROI.

$$\mathrm{ROI}_Q = \frac{\sum_{m\times n} WR(i) \times Q(i)}{\sum_{m\times n} WR(i)}. \tag{13}$$

$WR(i)$ represents the weight of the i-th image block, and $Q(i)$ represents the objective quality score under the metric. $WR(i)$ belongs to 0 or 1. That is to say, the difference between before and after IQA combined with ROI is that the original metric calculates the quality of the whole image, while the metric combined with ROI only calculates the image quality of ROI. The results are shown in the first six lines of Table 2. The results show that the correlation between the metric combined with ROI and MOS is higher than the original metric. This shows that the combination of ROI and IQA is helpful for IQA.

5.2. Performance Analysis of Proposed NIPQ. We calculated the correlation between various metrics and MOS in the database, and the results are shown in Table 2. It can be seen that the correlation between NIPQ metric and the subjective is significantly higher than other metrics.

In order to compare various NR IQAs intuitively, the scatter diagram between MOS and the estimated objective score is drawn, including six selected NR IQA and the NIPQ proposed in this paper, as shown in Figure 11. On this basis, the experimental data were regressed by the least square method, and the straight line is also drawn. The better the fitting effect of scatter point is, the better the correlation between the metric and MOS is. The regression line shows that the correlation between NIPQ and MOS is obviously better than other metrics. It validates the results of Table 2. It can be seen that LPSI and BRISQUE are the metrics designed for images in the air, which are not applicable to underwater images. As a whole, UIQM, UCIQE, and CCF are specially designed for underwater images, and their performance is better than that for images in the air. Performance of UICM, as a submetric indicating chromaticity in UIQM, is slightly worse than that of UIQM. Compared with the scatter plots of other NR IQA metrics, it can be seen that the performance of our NIPQ shows the best correlation with MOS. Although there are still some aberrant data points, generally speaking, the proposed NIPQ has better robustness to a variety of typical representative underwater images contained in the database. Further analysis shows that some of these aberrant points are caused by the fact that the submetric C_f of the original image (without enhancement) is directly taken as 1 in our experiment.

Table 2: Correlation between MOS and quality scores of objective evaluation metric before and after integration with ROI.

	PLCC	SROCC	KROCC	MAE	RMSE
UIQM	−0.173	−0.199	−0.132	0.751	0.903
ROI_UIQM	0.277	0.280	0.196	0.739	0.897
UCIQE	0.294	0.207	0.145	0.707	0.868
ROI_UCIQE	0.374	0.274	0.192	0.683	0.840
CCF	0.069	0.075	0.050	0.791	0.946
ROI_CCF	0.393	0.358	0.254	0.722	0.872
Var_Chr	0.158	0.180	0.125	0.674	0.841
UICM	−0.283	−0.338	−0.225	0.714	0.854
BRISQUE	−0.309	−0.265	−0.185	0.747	0.902
LPSI	0.323	0.245	0.169	0.734	0.898
C_d	0.481	0.465	0.335	0.635	0.789
C_f	0.478	0.432	0.303	0.658	0.806
Proposed	0.641	0.623	0.452	0.576	0.713

As shown in Figures 12 and 13, there are two natural underwater images and their enhanced images in the UOQ database. Table 3 shows the corresponding MOS and objective scores of these images. Figure 14 shows the color distribution of their ROI. From these images, the ROI of the original image of (1) is dark and that of (2) is blue. The image enhanced by the histogram algorithm is reddish, and the color distribution of ROI is wider, but the color of ROI is obviously oversaturated/pseudobright. There is no significant difference between the image processed by the Retinex algorithm and the original image. The color of the image processed by Fu's algorithm is not vibrant. For Figure 12, the overall difference between the white balance and the multifusion algorithm is small. The local graph (Figure 15) shows that the brightness distribution of the image processed by the multifusion algorithm is uneven, slightly oversaturated, and the image enhanced by the white balance algorithm has a better visual effect. For Figure 13, the image processed by the white balance algorithm is too dark and has

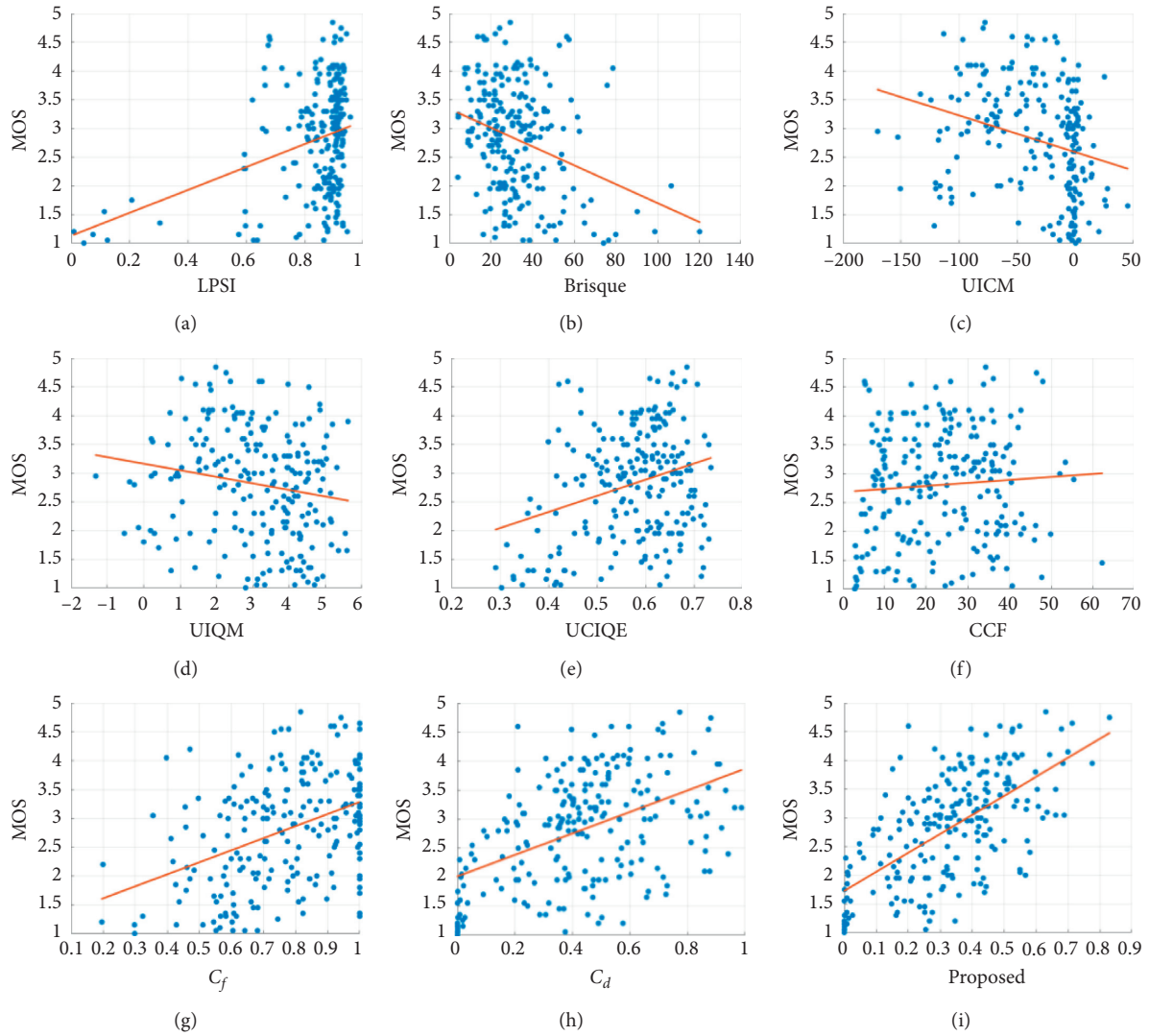


Figure 11: Scatter diagram between MOS and estimated objective score. (a) LPSI, (b) BRISQUE, (c) UICM, (d) UIQM, (e) UCIQE, (f) CCF, (g) C f, (h) C d, and (i) proposed.

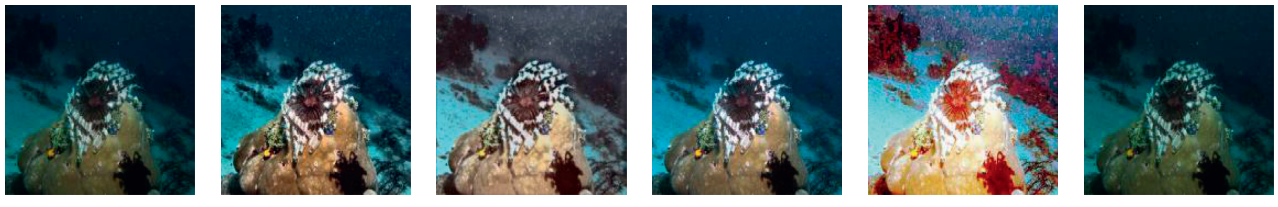


Figure 12: (a) Ori, (b) multifusion, (c) Fu, (d) white balance, (e) histogram equalization, and (f) Retinex.

Figure 13: (a) Ori, (b) multifusion, (c) Fu, (d) white balance, (e) histogram equalization, and (f) Retinex.

Table 3: The corresponding MOS and objective scores of Figures 12 and 13.

		Ori	Multifusion	Fu	White balance	Histogram equalization	Retinex
Figure 12	MOS	2.550	4.500	4.100	4.550	3.050	2.500
	CCF	13.265	22.292	23.887	16.437	30.794	13.379
	ROICCF	20.821	33.389	31.274	29.108	26.409	21.604
	UCIQE	0.554	0.664	0.591	0.652	0.684	0.569
	ROIUCIQE	0.560	0.647	0.573	0.627	0.580	0.575
	UIQM	3.983	4.543	4.850	3.969	4.780	4.085
	ROIUIQM	5.585	5.589	5.495	5.672	5.055	5.620
	BRISQUE	16.303	26.934	31.708	17.824	36.762	16.744
	LPSI	0.926	0.901	0.910	0.923	0.912	0.926
	C_d	0.243	0.715	0.578	0.698	0.846	0.324
	C_f	1.000	0.802	0.637	0.827	0.464	0.994
	Proposed	0.243	0.574	0.368	0.577	0.392	0.322
Figure 13	MOS	3.200	3.800	1.550	2.150	2.700	3.250
	CCF	31.443	31.465	37.069	18.688	36.928	29.029
	ROICCF	22.582	35.995	32.468	13.265	38.366	23.097
	UCIQE	0.519	0.628	0.623	0.476	0.693	0.541
	ROIUCIQE	0.541	0.620	0.588	0.447	0.676	0.564
	UIQM	1.504	3.337	4.325	3.840	4.100	2.182
	ROIUIQM	6.658	5.235	5.249	5.349	4.789	5.160
	Brisque	4.330	14.749	17.319	4.153	20.596	4.441
	LPSI	0.923	0.887	0.911	0.904	0.906	0.926
	C_d	0.475	0.730	0.317	0.029	0.640	0.401
	C_f	1.000	0.847	0.632	0.581	0.601	0.972
	Proposed	0.475	0.619	0.200	0.017	0.384	0.390

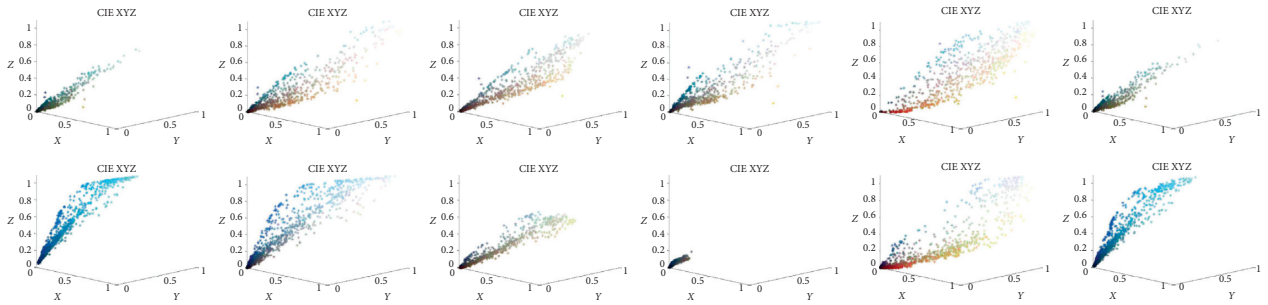


Figure 14: ROI color distribution of Figures 12 and 13.

(a) (b)

Figure 15: Local graph of Figures 12(b) and 12(d).

a single color. The image processed by the multifusion algorithm has a better visual effect.

Tables 3 shows that the selected IQAs do not perform well in the quality assessment of images in the UOQ database. They generally have higher objective scores for images enhanced by the histogram equalization algorithm because the color distribution of the images is wider. This is a disadvantage of quality evaluation based on statistics: color fidelity is not taken into account. It can be seen that if the performance of the original metric is not ideal, the metric combined with ROI will not necessarily improve this situation, because this is the limitation of the original metric itself.

6. Conclusion

Because of the characteristics of water medium, color has become one of the important concerns in underwater image quality assessment. Color contains important information. Severe color selective attenuation/pseudo-vividness can make it difficult to identify foreground content and extract key and effective information from images. In this paper, a new underwater image evaluation metric, NIPQ, is proposed based on the underwater environment characteristics and HVS. The NIPQ is designed in a three-stage framework. The first stage focuses on the attention mechanism of HVS. The second stage considers the influence of color richness in a statistical sense. The third stage is inspired by underwater image formation models and considers color fidelity from a pixel perspective. Finally, in the fourth phase, color richness and color fidelity are systematically integrated for real-time quality monitoring. At the same time, the relevant underwater image database UOQ with MOS is built to measure IQA performance. Experimental results show that, compared with other commonly used underwater metrics, NIPQ in this paper has better correlation with MOS, which shows better performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61571377, 61771412, and 61871336) and the Fundamental Research Funds for the Central Universities (20720180068).

References

[1] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6723–6732, Salt Lake City, UT, USA, March 2018.

[2] S. Bazeille, I. Quidu, L. Jaulin, and J.-P. Malkasse, "*Automatic Underwater image Pre-processing*," CMM'06, Brest, France, 2006.

[3] I. Avcibas, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 206–223, 2002.

[4] D.-Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," *Journal of Digital Imaging*, vol. 21, no. 3, pp. 338–347, 2008.

[5] Y. Y. Fu, "*Color image Quality Measures and Retrieval*," New Jersey Institute of Technology, Newark, NJ, USA, 2006.

[6] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] Q. Wu, Z. Wang, and H. Li, "A highly effificient method for blind image quality assessment," in *Proceedings of the 2015, IEEE International Conference on Image Processing (ICIP)*, pp. 339–343, IEEE, Quebec City, Canada, September 2015.

[8] J. Kim, A.-D. Nguyen, and S. Lee, "Deep cnn-based blind image quality predictor," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 11–24, 2018.

[9] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[10] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1040–1049, Venice, Italy, October 2017.

[11] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2016.

[12] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.

[13] Y. Wang, N. Li, Z. Li et al., "An imaging-inspired no-reference underwater color image quality assessment metric," *Computers & Electrical Engineering*, vol. 70, pp. 904–913, 2018.

[14] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 315–341, 2000.

[15] L. Zhang, J. Chen, and B. Qiu, "Region-of-interest coding based on saliency detection and directional wavelet for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 1, pp. 23–27, 2016.

[16] C. Zhu, K. Huang, and G. Li, "An innovative saliency guided roi selection model for panoramic images compression," in *Proceedings of the 2018 Data Compression Conference*, p. 436, IEEE, Snowbird, UT, USA, March 2018.

[17] Z. Cui, J. Wu, H. Yu, Y. Zhou, and L. Liang, "Underwater image saliency detection based on improved histogram equalization," in *Proceedings of the International Conference of Pioneering Computer Scientists, Engineers and Educators, Engineers and Educators*, pp. 157–165, Springer, Singapore, 2019.

[18] L. Xiu, H. Jing, S. Min, and Z. Yang, "Saliency segmentation and foreground extraction of underwater image based on localization," in *Proceedings of the OCEANS 2016*, Shanghai, China, 2016.

[19] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 2814–2821, Columbus, OH, USA, June 2014.

[20] D. Akkaynak and T. Treibitz, "Sea-thru: a method for removing water from underwater images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1682–1691, Long Beach, CA, USA, April 2019.

[21] D. Akkaynak, T. Treibitz, T. Shlesinger, Y. Loya, R. Tamir, and D. Iluz, "What is the space of attenuation coeffificients in underwater computer vision?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4931–4940, Honolulu, HI, USA, 2017.

[22] D. Berman, T. Treibitz, and S. Avidan, "Diving into haze-lines: color restoration of underwater images," in *Proceedings of the British Machine Vision Conference (BMVC)*, vol. 1, London, UK, September 2017.

[23] M. G. Solonenko and C. D. Mobley, "Inherent optical properties of jerlov water types," *Applied Optics*, vol. 54, no. 17, pp. 5392–5401, 2015.

[24] E. Y. Lam, "Combining gray world and retinex theory for automatic white balance in digital photography," in *Proceedings of the Ninth International Symposium on Consumer Electronics 2005*, IEEE, Melbourne, Australia, pp. 134–139, July 2005.

[25] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X.-P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," in *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4572–4576, IEEE, Paris, France, October 2014.

[26] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, RI, USA, pp. 81–88, June 2012.

[27] B. Zhang, "Image enhancement based on equal area dualistic sub-image histogram equalization method," *IEEE Transactions on Consumer Electronics*, vol. 45, no. 1, 75 pages.

[28] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, 1977.

[29] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human Vision and Electronic Imaging VIII*, vol. 5007, pp. 87–95, International Society for Optics and Photonics, Bellingham, WA, USA, 2003.

Hindawi
Scientific Programming
Volume 2020, Article ID 3180628, 15 pages
https://doi.org/10.1155/2020/3180628

Research Article

Short-Term Passenger Flow Forecast of Rail Transit Station Based on MIC Feature Selection and ST-LightGBM considering Transfer Passenger Flow

Zhe Zhang,[1] **Cheng Wang**,[1,2] **Yueer Gao**,[3] **Jianwei Chen**,[4] **and Yiwen Zhang**[1]

[1]*College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[2]*State Key Laboratory for Strength and Vibration of Mechanical Structures, Xi'an Jiaotong University, Xi'an 710049, China*
[3]*College of Architecture, Huaqiao University, Xiamen 361021, China*
[4]*Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA*

Correspondence should be addressed to Cheng Wang; wangcheng@hqu.edu.cn

Received 21 February 2020; Revised 15 July 2020; Accepted 19 July 2020; Published 25 August 2020

Academic Editor: Chenxi Huang

To solve the problems of current short-term forecasting methods for metro passenger flow, such as unclear influencing factors, low accuracy, and high time-space complexity, a method for metro passenger flow based on ST-LightGBM after considering transfer passenger flow is proposed. Firstly, using historical data as the training set to transform the problem into a data-driven multi-input single-output regression prediction problem, the problem of the short-term prediction of metro passenger flow is formalized and the difficulties of the problem are identified. Secondly, we extract the candidate temporal and spatial features that may affect passenger flow at a metro station from passenger travel data based on the spatial transfer and spatial similarity of passenger flow. Thirdly, we use a maximal information coefficient (MIC) feature selection algorithm to select the significant impact features as the input. Finally, a short-term forecasting model for metro passenger flow based on the light gradient boosting machine (LightGBM) model is established. Taking transfer passenger flow into account, this method has a low space-time cost and high accuracy. The experimental results on the dataset of Lianban metro station in Xiamen city show that the proposed method obtains higher prediction accuracy than SARIMA, SVR, and BP network.

1. Introduction

In recent years, China's economy has developed rapidly, and the process of urbanization has gradually accelerated. The country has continuously increased its efforts to build public transportation. Among them, urban rail transit is particularly noticeable as a new direction in the field of public transportation. Urban rail transit has the advantages of strong carrying capacity, a high punctuality rate, energy conservation, and environmental protection [1]. The development of urban rail transit is considered an effective way to alleviate the urban traffic congestion. Hence, it is the future trend of China's urban transportation development to establish a comprehensive transportation system with urban rail transit as the backbone, public transport as the main body, and various modes of transportation interconnected. By the end of 2019, rail transit has been built in 40 cities in China, and the total mileage of metro construction has reached 6736.2 km [2]. Passenger flow prediction not only plays a guiding role in the planning and design of rail transit but also plays an irreplaceable role in the operation of rail transit. The most commonly used passenger flow prediction method is the four-stage method [3, 4], which consists of four parts: travel generation, travel distribution, travel mode split, and travel assignment. It is a macrolevel prediction method. The first city, which actually used this method for traffic prediction in 1962, was Chicago. It is very suitable for the long-term prediction of passenger flow and is of great significance for the planning of rail transit networks, the construction of engineering

projects, and the selection of station equipment. However, long-term passenger flow forecasting cannot solve the problems arising from the daily operation of the rail transit. With the development of rail transit, most people choose rail transit as their main travel mode, which has directly led to the rapid growth of rail transit passenger flow. This has led to problems such as passenger congestion, low operating efficiency, unbalanced capacity and demand, and poor driving safety [5, 6]. Therefore, we must adopt more accurate short-term passenger flow forecasting method to scientifically forecast short-term passenger flow. Through short-term passenger flow forecasting, we can obtain passenger travel data for a short period of time in the future so as to grasp the accurate passenger flow change trend and provide the basis for the organization and management of the operation department (e.g., it can help the operation department to realize the dynamic adjustment of the rail transit capacity in the peak hours, the reasonable scheduling of service personnel, and the timely treatment of emergencies). In addition, short-term passenger flow forecasting can improve the operation efficiency of rail transit, reduce the time cost of passengers, and improve passengers' satisfaction, thus improving the level of public service of rail transit and increasing its competitiveness. However, the influencing factors of short-term passenger flow at metro station are intricate. And short-term passenger flow has the characteristics of nonlinearity, nonstationarity, randomness, and suddenness, which makes the prediction more difficult. Using the data-driven method to solve short-term forecasting problems is proven to be to be an effective way [7, 8]. LightGBM is a new boosting framework model that was proposed by Microsoft in 2015 [9]. It has a fast training speed, low memory consumption, can process massive data quickly, and has better model accuracy, which are suitable for solving the short-term passenger flow forecast problem of rail transit.

The research purpose of this study is to forecast short-term passenger flow of a metro station. The main contributions and novelty of this paper are as follows:

(1) In order to supplement the lack of scientific analysis of short-term metro passenger flow prediction problem, we formally describe the problem based on the data-driven model and analyze the difficulties of the problem to better describe the complexity of short-term metro passenger flow prediction.

(2) In order to overcome the problems of feature incompleteness and high cost of feature acquisition in traditional methods, we use temporal features, spatial similarity features, and spatial transfer features extracted from IC card data as the candidate influence features, which are more comprehensive and easy to obtain.

(3) In order to solve the problem of heavy computational burden caused by excessive input features, the candidate features are further selected by using a maximal information coefficient (MIC) feature selection algorithm to extract the significant features, which reduces the dimension of the features and reduces the computational cost.

(4) In order to solve the problems that the existing methods cannot reflect the uncertainty of short-term passenger flow and the prediction accuracy is not high enough, we use the integrated learning algorithm LightGBM as a prediction model to describe the nonlinear characteristics of short-term passenger flow and improve the prediction accuracy.

(5) The experimental results on the dataset of Lianban metro station in Xiamen city show that the proposed method obtains a higher prediction accuracy than SARIMA, SVR, and BP network.

2. Related Work

At present, many scholars have conducted a great deal of research on the prediction of short-term passenger flow. The historical average model was the first method applied to traffic flow prediction [10]. However, it is difficult for the historical average regression model to reflect the randomness of passenger flow. It requires strong stability and periodicity of data, which leads to its harsh application conditions. Thus, the performance of the historical average model in the research of El Esawey [11] and Yang et al. [12] was not good. The Kalman filtering [13] model is also one of the commonly used passenger flow prediction methods. Jiao et al. [14] proposed three improved Kalman filter models for the short-term prediction of rail transit passenger flow and achieved good prediction results. The time series model is a classic model for passenger flow prediction [15]. Milenković et al. [16] predicted railway passenger flow using the autoregressive integrated moving average model (ARIMA), which achieved good prediction results. Anvari et al. [17] constructed a time series prediction framework for a public transport system based on the Box–Jenkins method which included the ARIMA model. Li et al. [18] proposed a hybrid model that combined a symbolic regression model and ARIMA model to predict the passenger flow of *Xian* rail line 1. The prediction results showed that the hybrid model has better prediction accuracy than the simple ARIMA model. With the rise of machine learning, a nonparametric regression model based on data was applied to the study of short-term passenger flow prediction. Regarding the support vector machine (SVM) model [19, 20], Sun et al. [21] forecasted transfer passenger flow for Beijing rail transit by setting a wavelet SVM model. For the K-nearest neighbor (K-NN) regression model [22], Habtemichael and Cetin [23] proposed a nonparametric and data-driven methodology for short-term traffic forecasting based on identifying similar traffic patterns using an enhanced K-NN algorithm. Regarding the Bayesian network model, Roos et al. [24] proposed a method based on a dynamic Bayesian network to predict the short-term passenger flow of the Paris Metro,

which can work normally even when the data are incomplete. For the neural network model [25, 26], Zhu et al. [27] constructed a three-layer neural network to predict the outbound and inbound passenger flow of a metro station by analyzing the main dynamic factors that affect passenger flow in a rail transit station. The prediction accuracy was higher than the traditional linear regression method. Liu and Chen [28] used SAE to extract the nonlinear characteristics of the input and constructed a hybrid model (stacked autoencoder-deep neural network, SAE-DNN) to predict passenger flow in BRT stations. Chen et al. [29] constructed a long short-term memory network prediction model for rail transit passenger flow based on empirical mode decomposition. Liu et al. [30] used deep learning architecture to predict the outbound passenger flow of the research station according to the arrival schedule of the rail train and the inbound passenger flow of other stations. Han et al. [31] used the graph convolution to mine the temporal and spatial dependence of each station and proposed a short-term passenger flow prediction model for rail transit based on spatial-temporal graph convolutional neural networks. Both methods only take into account the spatial correlation of stations within the rail transit system and ignore the impact of transfer effects between other public transport modes (i.e., conventional bus transit and bus rapid transit (BRT)) and rail transit [32]. The historical average model cannot reflect the uncertainty caused by the change of passenger flow very well, so the prediction result error is relatively large. The Kalman filtering model requires many parameter vector calculations, which makes its operation complicated. When passenger flow fluctuates greatly, the time series model ARIMA cannot effectively capture the trend of passenger flow. The SVM and K-NN models have a high time complexity and cannot adapt to large-scale training data. The network construction process of the Bayesian network model is complex. The neural network model convergence speed is slow, it falls easily into the local solution, and it has a high demand for training data.

Recently, an integrated learning algorithm was also applied to the prediction of rail transit passenger flow and achieved a good effect [33]. LightGBM is an open-source, fast, and efficient boosting framework based on a decision tree algorithm, which is based on the idea of gradient boosting. LightGBM supports efficient parallel training and achieves good results in regression and classification problems [34–37], which is very suitable for this field. In this study, a spatial-temporal feature extraction method that considers transfer passenger flow is proposed, and a metro station passenger flow prediction model based on LightGBM is constructed. The remainder of this article is structured as follows. In Section 3, a formal description of the problem of metro passenger flow prediction is presented. In Section 4, a spatial-temporal feature extraction method and passenger flow prediction model are introduced. In Section 5, experimental research based on Xiamen (a city at the southeast end of Fujian Province, China) public transport data is introduced, and the experimental results and model performance are evaluated.

3. Formal Description of the Metro Station Passenger Flow Prediction Problem Based on the Data-Driven and Multiple Regression Model

3.1. Related Definitions

(1) *Rail Transit*. The general term of fast and large volume public transportation with electric energy as power and wheel-rail as transportation system (this study refers to the metro).

(2) *Metro Station*. A place to provide a stop for metro trains to carry goods or passengers.

(3) *Yitong Card*. A kind of intelligent card which can be used in public transportation payment system.

(4) *BRT QR Code*. A kind of QR code which can be used in BRT payment system.

(5) *Metro QR Code*. A kind of QR code which can be used in metro payment system.

(6) *BRT One-Way Ticket*. A kind of anonymous BRT ticket sold by automatic ticket vending machine, which is swiped once before entering the station and needs to be put into the recycling hole before leaving the station.

(7) *Metro One-Way Ticket*. A kind of anonymous metro ticket sold by automatic ticket vending machine, which is swiped once before entering the station and needs to be put into the recycling hole before leaving the station.

(8) *Data-Driven Model*. Without prior knowledge, the model is trained based on massive historical data.

(9) *Short-Term Inbound Passenger Flow Forecast of a Metro Station*. Forecast of the total number of passengers entering the station for a short period of time (several hours or less).

(10) *Transfer Passenger Flow*. The total number of passengers transferring between different modes of transportation in a unit of time.

3.2. Introduction to the Composition of the Data Dictionary. Having sufficient data is the basis for forecasting. With the rapid development of passenger data acquisition technology, sufficient data can be obtained for the short-term prediction of passenger flow. The Xiamen public transport system is considered as an example. During the study period, there were six main types of passenger payment in Xiamen: "Yitong card," "Coin payment," "BRT QR code," "BRT one-way ticket," "Metro QR code," and "Metro one-way ticket." Conventional bus transit supports the two payment methods of "Yitong card" and "Coin payment." BRT supports the three payment methods of "Yitong card," "BRT QR code," and "BRT one-way ticket." Rail transit supports the three payment methods of "Yitong card," "Metro QR code," and "Metro one-way ticket." Hence, we counted the rail transit passenger flow using the data of the "Yitong card," "Metro QR code," and "Metro one-way ticket." From the above

description, "Coin payment" can only be used for conventional bus transit; "BRT QR code" and "BRT QR code" can only be used for BRT; "Metro QR code" and "Metro one-way ticket" can only be used for the metro; and the "Yitong card" is the only universal payment method for the three modes of transportation (i.e., conventional bus transit, BRT, and rail transit). Additionally, the "Yitong card" has the property of a unique physical card number that corresponds to a unique passenger. Therefore, we can only use the "Yitong card" to identify transfer passenger flow. Additionally, we regard transfer passenger flow as one of the influencing factors in the subsequent section.

Table 1 is an introduction to the travel data records: ID, $otime$, $ostation$, $dtime$, $dstation$, $date$, $type$, and $public\ transport$ are the attributes that denote the card identification, origin time, origin station, destination time, destination station, date, payment type, and travel mode (rail transit, BRT, or conventional bus transit), respectively.

3.3. Formal Description of the Passenger Flow Prediction Problem in a Metro Station Based on the Data-Driven and Multiple Regression Model. Let j be the target metro station, Δt be the prediction time interval (e.g., 10, 20, or 30 minutes), and $x\_in_{j,t}$ be the inbound passenger flow of station j in target time period t. First, the feature set of the spatial-temporal influencing factors is determined and expressed as $Te = \{te_1, te_2, \ldots\ te_i, \ldots, te_n\}$, where te_i represents the i^{th} spatial-temporal influencing feature. It is used as the input to the model and $x\_in_{j,t}$ is the output of the model. Historical data are used as training data for the multi-input single-output regression model. The regression prediction model of metro station passenger flow is trained, with $Te^h = \{te^h{}_1, te^h{}_2, \ldots, te^h{}_i, \ldots, te^h{}_n\}$ as input and $x\_in^h_{j,t}$ as output. The trained model is used to obtain a prediction under the working condition. h is the training sample serial number, $h = 1, 2, \ldots, H$. H is the number of training samples. Under the working condition, with $Te^w = \{te^w{}_1, te^w{}_2, \ldots, te^w{}_i, \ldots, te^w{}_n\}$ as input, the output $\hat{x\_in}^w_{j,t}$ of the model can be used as the forecast value of real passenger flow $x\_in^w_{j,t}$. w is the testing sample serial number, $w = 1, 2, \ldots, W$. W is the number of testing samples. The accuracy of the model is evaluated by comparing $\hat{x\_in}^w_{j,t}$ with $x\_in^w_{j,t}$. The problem model is shown in Figure 1.

3.4. Difficulties of the Problem

(1) There are many factors that influence the short-term passenger flow of a metro station. Under the background of the integration of public transport, all types of public transport modes are bound together. Passenger flow in a metro is not only affected by its own system but also by other public transport modes. How to use existing data to extract and select the significant influencing factors from the space-time dimension is an important issue.

(2) The relationship between the influencing factors and short-term passenger flow is complex and nonlinear. To improve the prediction accuracy, it is also necessary to select a suitable model to express the nonlinear relationship between the influencing factors and passenger flow.

4. Short-Term Passenger Flow Forecast of Rail Transit Station Based on MIC Feature Selection and ST-LightGBM considering Transfer Passenger Flow

Metro station passenger flow forecasting is a complex problem in time and space. Thus, this section is divided into four parts: the first part is the extraction of the candidate temporal and spatial features that affect the inbound passenger flow of the metro station, the second part is the selection of candidate spatial-temporal features using the MIC algorithm, the third part is the introduction of the prediction model based on LightGBM, and the final part is the theoretical analysis and comparison of the proposed method and other methods.

4.1. Spatial-Temporal Feature Extraction

4.1.1. Temporal Feature Extraction. Let j be the target metro station, Δt be the prediction time interval (e.g., 10, 20, or 30 minutes), $x\_in_{j,t}$ be the inbound passenger flow of station j in target time period t, day_t be the "weekly information" (i.e., Monday, Tuesday, ..., Sunday) in target period t, and hour_t be the hour of the day that corresponds to target period t. Because passenger flow changes in a metro station during a week are different (e.g., working days and nonworking days) and passenger flow changes in a day are also different (e.g., peak hours and off-peak hours), $x\_in_{j,t}$ also changes with the changes of day_t and hour_t. Additionally, passenger flow has the property of time delay. Thus, historical inbound passenger flow is correlated with that of the current period. Therefore, the historical passenger flow set $His_{j,t} = \{x\_in_{j,t-k}, x\_in_{j,t-k+1}, \ldots, x\_in_{j,t-1}\}$ is another time feature that affects $x\_in_{j,t}$. Finally, three temporal features are extracted: day_t, hour_t, and $His_{j,t} = \{x\_in_{j,t-k}, x\_in_{j,t-k+1}, \ldots, x\_in_{j,t-1}\}$.

4.1.2. Spatial Feature Extraction

(1) Spatial Similarity Feature Extraction. Because the land function of the space in which adjacent stations are located is similar, the travel habits (i.e., departure time) of passengers in these adjacent stations are similar. Hence, there is spatial similarity between the passenger flow of a metro station and adjacent stations (i.e., adjacent conventional bus stations, BRT stations, and metro stations). Therefore, the current inbound passenger flow of a metro station is also related to the historical inbound passenger flow of adjacent stations. Suppose that the target metro station j has n adjacent metro stations and m adjacent bus stations (i.e., BRT and conventional bus stations). Then, the spatial similarity features of the passenger flow at the metro station can be represented by the adjacent station history inbound passenger flow

Table 1: Travel data records.

ID	otime	ostation	dtime	dstation	Date	Type	Public transport
178***521	08:05:26	Lianban	08:10:38	Wenzao	2018.11.01	Metro one-way ticket	Rail transit
261***345	09:15:23	Zhenhailu	09:25:46	Lianban	2018.11.05	Yitong card	Rail transit
174***020	10:41:00	Lianban	11:01:22	Zhenhailu	2018.11.14	Metro QR code	Rail transit
356***742	11:18:36	Jiageng	11:55:47	Lianban	2018.11.07	BRT QR code	BRT
194***942	16:30:38	Kaihelukou	16:42:39	Lianban	2018.11.03	Yitong card	BRT
285***865	13:44:29	Lianban	13:55:24	Ershi	2018.11.09	BRT one-way ticket	BRT
454***261	18:25:48	Huming	18:38:26	Lianjingerli	2018.11.02	Yitong card	Conventional bus

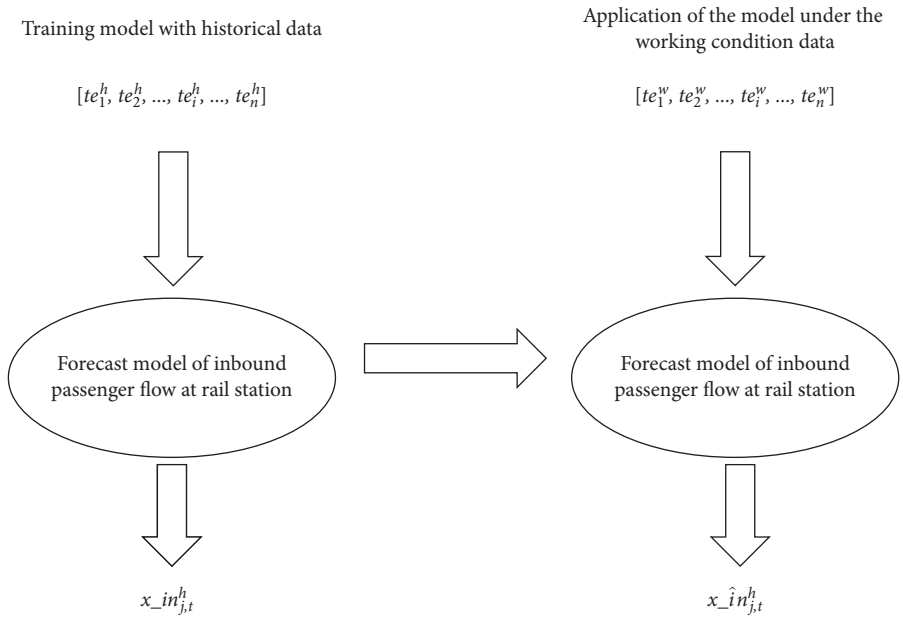


Figure 1: Data-driven model of metro station passenger flow prediction.

matrix (ASHIM). Select the historical inbound passenger flow in the past k periods. Then, the size of the ASHIM is $k \times (n+m)$, and it can be denoted by

$$\mathrm{ASHIM}_{j,t} = \begin{pmatrix} x\_in_{jr(1),t-1}, x\_in_{jr(2),t-1}, \ldots, x\_in_{jr(n),t-1}, x\_in_{jb(1),t-1}, x\_in_{jb(2),t-1}, \ldots, x\_in_{jb(m),t-1} \\ x\_in_{jr(1),t-2}, x\_in_{jr(2),t-2}, \ldots, x\_in_{jr(n),t-2}, x\_in_{jb(1),t-2}, x\_in_{jb(2),t-2}, \ldots, x\_in_{jb(m),t-2} \\ \cdots \\ x\_in_{jr(1),t-k}, x\_in_{jr(2),t-k}, \ldots, x\_in_{jr(n),t-k}, x\_in_{jb(1),t-k}, x\_in_{jb(2),t-k}, \ldots, x\_in_{jb(m),t-k} \end{pmatrix}, \quad (1)$$

where $x\_in_{jr(n),t-k}$ is the inbound passenger flow of the n^{th} adjacent metro station of target metro station j at time period $t-k$ and $x\_in_{jb(m),t-k}$ is the inbound passenger flow of the m^{th} adjacent bus station of the target metro station j at time period $t-k$.

(2) Spatial Transfer Feature Extraction. Passengers have transfer behavior in travel activities. Thus, some passengers may transfer to the rail system by other travel modes (BRT and conventional bus transit). Specifically, some passengers will transfer to an adjacent metro station after leaving the bus or BRT station and continue to travel by rail transit. Therefore, for metro station j, a proportion of passengers in the outbound passenger flow of the adjacent conventional bus and BRT stations in the several previous periods will transfer to metro station j at time period t and then continue to complete the travel activities by rail transit. Hence, the metro station's inbound passenger flow at the current period is also related to the transfer passenger flow from the historical outbound passenger flow of the adjacent BRT and conventional bus stations. According to the outbound historical passenger flow of m adjacent bus stations (i.e., BRT and conventional bus stations) in the past k periods, we can obtain the outbound passenger flow matrix of the adjacent bus stations, i.e., adjacent bus station history outbound passenger flow

(ABHOM). The size of the ABHOM is $k \times m$, and it can be expressed as

$$\mathrm{ABHOM}_{j,t} = \begin{pmatrix} x\_out_{jb(1),t-1}, x\_out_{jb(2),t-1}, \ldots, x\_out_{jb(m),t-1} \\ x\_out_{jb(1),t-2}, x\_out_{jb(2),t-2}, \ldots, x\_out_{jb(m),t-2} \\ \cdots \\ x\_out_{jb(1),t-k}, x\_out_{jb(2),t-k}, \ldots, x\_out_{jb(m),t-k} \end{pmatrix}, \tag{2}$$

where $x\_out_{jb(m),t-k}$ is the outbound passenger flow of the m^{th} adjacent bus station of target metro station j at time period $t-k$. In the analysis, we obtain the outbound passenger flow of each adjacent bus station in the previous period. However, as time period t has not yet occurred, for $x\_out_{jb(m),t-k}$, we do not know what proportion of the passenger flow will transfer to metro station j at time period t. To solve this problem, we set up the transfer ratio matrix (TRM) according to the historical average transfer ratio. The size of the TRM is $k \times m$, and it can be expressed as

$$\mathrm{TRM}_{j,t} = \begin{pmatrix} P_{jb(1),t-1}, P_{jb(2),t-1}, \ldots, P_{jb(m),t-1} \\ P_{jb(1),t-2}, P_{jb(2),t-2}, \ldots, P_{jb(m),t-2} \\ \cdots \\ P_{jb(1),t-k}, P_{jb(2),t-k}, \ldots, P_{jb(m),t-k} \end{pmatrix}, \tag{3}$$

$$p_{jb(m),t-k} = \left(\frac{\mathrm{transfer}_{jb(m),t-k}}{x\_out_{jb(m),t-k}}\right), \quad t \in T^*,$$

$$T^* = \{t + q\omega\}, \quad q = 0, 1, 2, 3, \ldots, l,$$

where $\mathrm{transfer}_{jb(m),t-k}$ represents the passenger flow of $x\_out_{jb(m),t-k}$, which transfers to metro station j in time period t; $p_{jb(m),t-k}$ represents the historical average proportion of $\mathrm{transfer}_{jb(m),t-k}$ to $x\_out_{jb(m),t-k}$; and T^* is the historical time series, which consists of time period t and the same period in several weeks earlier. Therefore, we obtain the transfer passenger flow matrix (TPM). The size of the TPM is $k \times m$, and it can be expressed as

$$\mathrm{TPM}_{j,t} = \begin{pmatrix} \mathrm{transfer}_{jb(1),t-1}, \mathrm{transfer}_{jb(2),t-1}, \ldots, \mathrm{transfer}_{jb(m),t-1} \\ \mathrm{transfer}_{jb(1),t-2}, \mathrm{transfer}_{jb(2),t-2}, \ldots, \mathrm{transfer}_{jb(m),t-2} \\ \cdots \\ \mathrm{transfer}_{jb(1),t-k}, \mathrm{transfer}_{jb(2),t-k}, \ldots, \mathrm{transfer}_{jb(m),t-k} \end{pmatrix}, \tag{4}$$

$$\mathrm{transfer}_{jb(m),t-k} = p_{jb(m),t-k} \times x\_out_{jb(m),t-k}.$$

By adding all the elements of TPM, we can obtain the total number of transfer passengers $\mathrm{All\_transfer}_{j,t}$ that is transferred from all the adjacent bus stations to metro station j in time period t. We obtain the spatial transfer feature $\mathrm{All\_transfer}_{j,t}$.

Finally, we extract the candidate temporal and spatial features that are composed of the candidate feature set Te: day_t, hour_t, $His_{j,t} = \{x\_in_{j,t-k}, x\_in_{j,t-k+1}, \ldots, x\_in_{j,t-1}\}$, $\mathrm{ASHIM}_{j,t}$, and $\mathrm{All\_transfer}_{j,t}$.

4.2. Feature Selection Based on the Maximal Information Coefficient (MIC). In the previous section, we constructed the candidate spatial-temporal features of passenger flow prediction and obtained a comprehensive set Te of candidate features. Feature selection can solve the problem of heavy computational burden caused by excessive input features [38]. To make passenger flow prediction more effective, we need to select more important features from set Te and obtain a simplified feature input so that the subsequent

learning process only needs to establish a model based on the important features. The performance of an embedded and wrapped feature selection algorithm is closely related to the learner. The algorithm is easy to overfit and has high time complexity and poor interpretability. Thus, we choose the filter feature selection algorithm MIC [39]. Compared with other filter feature selection methods, the MIC algorithm can widely measure dependence between variables, such as linear and nonlinear relations, even for nonfunctional dependence, which cannot be represented by a single function (e.g., dependence composed of multiple functions). Additionally, as a filtering feature selection algorithm, the execution efficiency is high, so we choose MIC as the feature selection method.

The MIC is mainly calculated using mutual information and grid division. Mutual information is an indicator that measures the correlation between variables. Given variables $A = \{a_i, i = 1, 2, \ldots, n\}$ and $B = \{b_i, i = 1, 2, \ldots, n\}$, n is the number of samples. Mutual information is defined as follows:

$$MI(A, B) = \sum_{a \in A} \sum_{b \in B} pro(a, b) \log \frac{pro(a, b)}{pro(a) pro(b)}, \tag{5}$$

where $pro(a, b)$ is the joint probability density of A and B and $pro(a)$ and $pro(b)$ are the edge probability densities of A and B, respectively. Histogram estimation is used to estimate the above probability density. Suppose $D = \{(a_i, b_i), i = 1, 2, \ldots, n\}$ is a finite set of ordered pairs. Define division G to divide the range of variable A into x segments and divide the range of B into y segments. Thus, G is an $x \times y$ grid. Calculate the mutual information $MI(A, B)$ in each grid division. There are many ways to divide the grid into $x \times y$, and the maximum value of $MI(A, B)$ in each way is taken as the mutual information value of G. Define the maximum mutual information formula of D under division G as follows:

$$MI^*(D, x, y) = \max MI(D \mid G), \tag{6}$$

where $D|G$ indicates that data D is divided by G. Use the maximum normalized MI values obtained under different divisions to form the feature matrix, which is defined as

$$M(D)_{x,y} = \frac{MI^*(D, x, y)}{\log \min\{x, y\}}. \tag{7}$$

Then, the MIC is defined as

$$MIC(D) = \max_{xy < B(n)} \left\{ M(D)_{x,y} \right\}, \tag{8}$$

where $B(n)$ is the upper limit value of grid division $x \times y$. Generally, Reshef et al. [39] suggested that $B(n) = n^{0.6}$ is best.

We use the MIC to define the correlation between the features and target value. The candidate feature set is $Te = \{te_1, te_2, \ldots, te_i, \ldots, te_n\}$. The correlation between any feature te_i and target *value* is defined as $MIC(te_i, target\ value)$. The value range is [0, 1]. The larger the $MIC(te_i, target\ value)$ value, the stronger the correlation between te_i and *target value*, and te_i is a strong correlation feature. The smaller the $MIC(te_i, target\ value)$ value, the weaker the correlation between te_i and *target value*, and te_i is a weak correlation feature.

A flowchart for feature selection is shown in Figure 2. Through the MIC feature selection algorithm, we obtain the significant feature set Te'.

4.3. ST-LightGBM Passenger Flow Prediction Model. LightGBM is an open-source, fast, and efficient lifting framework based on a decision tree algorithm, which supports efficient parallel training and can greatly shorten the training time. The idea of gradient boosting is to iterate variables once, increase the submodels individually in the process of iteration, and ensure that the loss function is constantly reduced. Let $f_i(X)$ be the submodel, $F_n(x) = \partial_0 f_0(x) + \partial_1 f_1(x) + \cdots + \partial_n f_n(x)$ be the composite model, and $Loss[F_n(x), Y]$ be the loss function. Every time a new submodel is added, the loss function decreases toward the gradient of the variable with the next highest information content $Loss[F_n(x), Y] < [F_{n-1}(x), Y]$. The gradient boosting decision tree (GBDT) is a classical model. GBDT has the functional characteristics of gradient boosting and decision tree and has the advantages of achieving good prediction results and is not easy to overfit. However, when calculating the information gain, it needs to scan all samples to determine the best partition point, which consumes a great deal of computing time. LightGBM is a type of GBDT that is used to solve the problems encountered by GBDT in massive data processing. It consists of two algorithms: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) to optimize GBDT. GOSS [9] was proposed to prove that the larger the gradient of samples, the more important the role they play in calculating information gain to obtain quite accurate information gain estimates from a small number of samples. The core idea of the GOSS algorithm is to select some samples with a large gradient from the total samples, select some samples randomly from the remaining samples, and combine them into new samples to learn a new classifier. This method makes the distribution of the new samples consistent with the total samples and trains the data of small gradient samples. Therefore, under the premise of not changing the distribution of samples, the accuracy of classifier learning is not lost and the speed of classifier learning is greatly reduced. EFB [9] is an algorithm that can reduce the number of features of high-dimensional data and minimize the loss. It binds nonzero features in sparse feature space together to form a feature and then establishes the same feature histogram as a single feature from the feature binding. Thus, the training of GBDT can be accelerated in the case of lossless accuracy.

Simultaneously, LightGBM adopts the method of leaf splitting, which has a low calculation cost. By controlling the depth of the tree and the minimum amount of data of each leaf node, it avoids the overfitting phenomenon. LightGBM chooses the decision tree algorithm based on a histogram, which can reduce the storage cost and calculation cost. Additionally, the processing of category features also improves LightGBM performance for specific data.

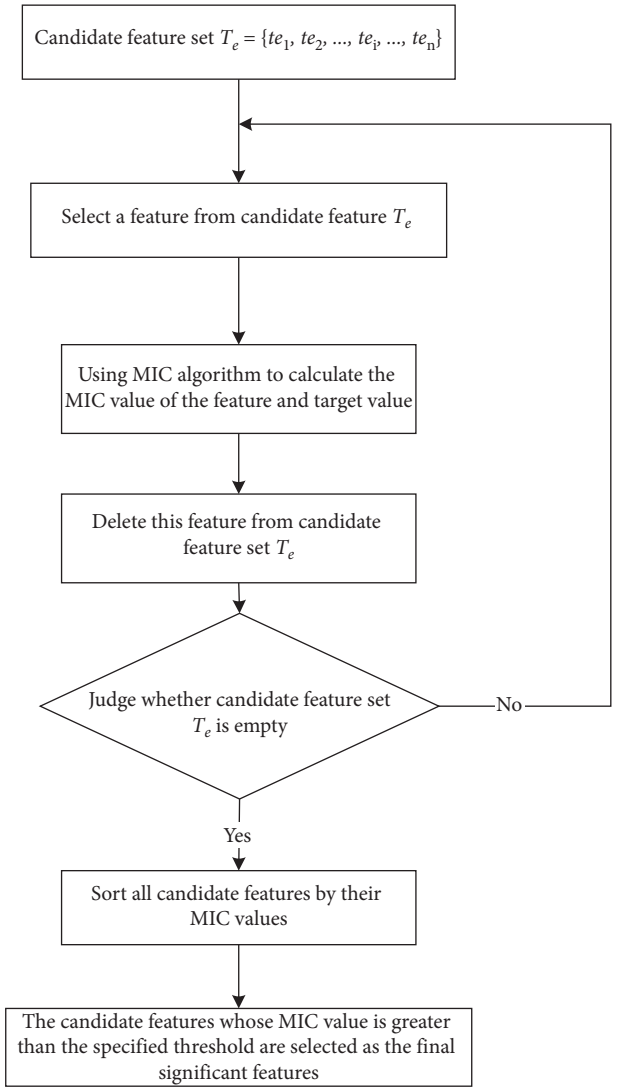


Figure 2: Flowchart for feature selection.

The framework of the proposed method is shown in Figure 3. As we can see from the plot, first, we extract temporal features from multisource traffic data. Second, according to the spatial location of the metro station, we extract spatial similarity features and spatial transfer features from the data. Third, we use the MIC algorithm to select the significant features. Finally, we establish an ST-LightGBM passenger flow prediction model to predict the inbound passenger flow of a metro station in a real-world scenario.

4.4. Scalability and Limitations of the Proposed Method

4.4.1. Scalability of the Proposed Method

(1) This method can be applied to the inbound passenger flow prediction of any metro station.

(2) This method can also be applied to the prediction of inbound passenger flow of conventional bus stations and BRT stations.

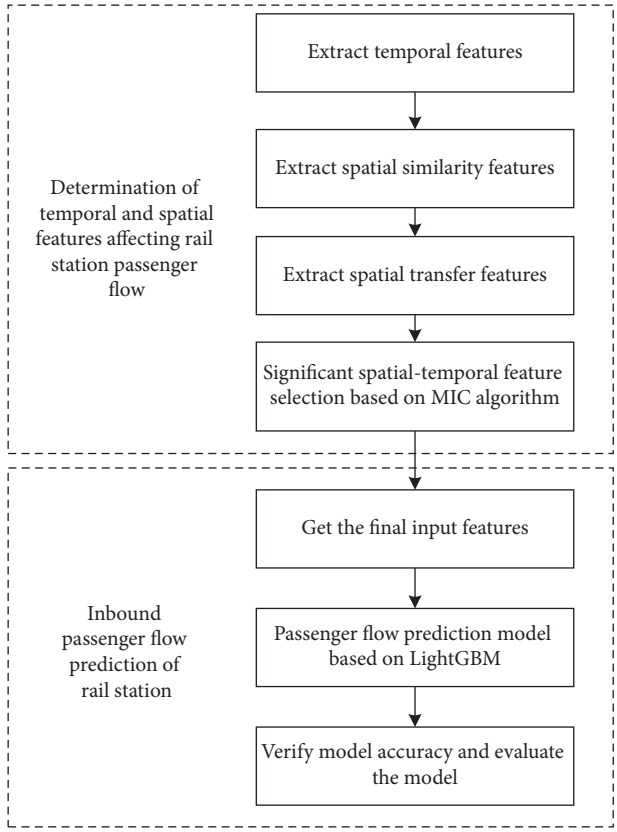


Figure 3: Framework of the proposed ST-LightGBM passenger flow prediction method.

(3) This method is not limited by the region and can also be applied to other cities.

(4) This method cannot be applied to the prediction of passenger flow at rail stations under the impact of emergencies, such as sudden bad weather (e.g., rainstorm, flood or typhoon, etc.), terrorist attacks, traffic accidents, and metro accidents.

(5) The application of this method is limited to station-level prediction, not applicable to line-level or city-level prediction.

(6) This method is only suitable for short-term prediction, not when the metro station surrounding environment changes.

4.4.2. Limitations of the Proposed Method

(1) The candidate features need to directly or indirectly reflect the factors that affect the passenger flow at rail stations. If there are important factors missing, such as transfer passenger flow, the accuracy of model prediction will be reduced.

(2) It is necessary to collect enough historical data as the training dataset to train the short-term prediction model. If the historical data are insufficient, inaccurate, or noisy, the accuracy of the prediction model will be reduced.

Table 2: Theoretical analysis and comparison of methods.

Studies	Temporal feature	Spatial feature	Transfer feature	Model	Applicable data size of model	Time complexity	Modeling difficulty	Parameter complexity	Accuracy
Feng and Cai [40]	Yes	No	No	ARIMA	Small-medium	Low	Low	Low	Low
Sun et al. [21]	Yes	No	No	Wavelet SVM	Small	Medium	Medium	Medium	Medium
Jin et al. [41]	Yes	No	No	BP network	Big	High	Medium	High	High
Tang et al. [32]	Yes	Yes	No	ST-LSTM	Big	High	High	High	High
Zhang et al. [42]	Yes	Yes	No	Multi-LSTM	Big	High	High	High	High
The proposed method	Yes	Yes	Yes	ST-LightGBM	Big	Low	Medium	Low	High

(3) The threshold of MIC algorithm and ST-LightGBM model superparameters will affect the accuracy of experimental results. It is necessary to adjust parameters in advance according to different objects. Improper selection of parameters will lead to a low accuracy.

(4) The process of feature extraction is complex, especially the feature extraction of spatial transfer features.

(5) When predicting passenger flow at rail stations, it is necessary to use the MIC algorithm to further select candidate features to determine the input of the model. This process is complicated.

4.5. Theoretical Analysis and Comparison of Methods. A comparison of various rail passenger flow prediction methods is shown in Table 2. Compared with other methods, the features extracted by the proposed method are more comprehensive. Particularly, it considers the impact of transfer passenger flow, which plays an important role in the prediction of metro station passenger flow. Furthermore, the proposed method has higher prediction accuracy and efficiency than other methods.

5. Experiment

5.1. Experimental Object and Dataset Description. Lianban metro station (as shown in Figure 4) is an important passenger flow point of Xiamen rail line 1, with a large and stable passenger flow. Therefore, we chose Lianban metro station as the research object. Taking 1,000 meters as the boundary condition, we selected 14 adjacent stations with a stable passenger flow. The adjacent metro stations of Lianban metro station are Hubin East Road metro station and Lianhualukou metro station; the adjacent BRT stations are BRT Lianban station and BRT Huoche station; and the adjacent conventional bus stations are Lianban Book City station, Lianjingerli station, Siming Court station, Lianbanguomao station, Lvjiayuanxiaoqu station, Lianbanbei station, Fengyulu station, Huoche station, Huming station, and Huminglijing station.

We considered Xiamen residents' travel data from November 1, 2018, to November 25, 2018, as the experimental data. The prediction time interval was $\Delta t = 10$ minutes. There are 144 pieces of data in one day. Hence, there are 3600 sample data.

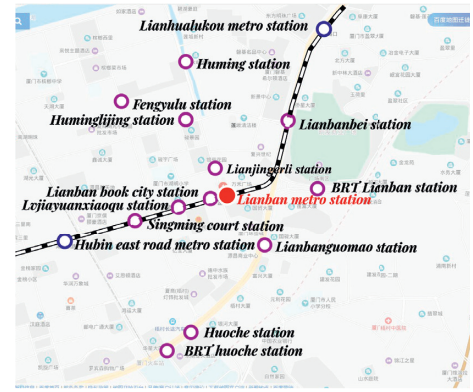


Figure 4: Spatial location of Lianban metro station.

5.2. Evaluation Methods and Indicators. To analyze and compare the prediction effect of each experiment, we use 5-fold cross-validation to get the average error. The number of training samples H is 2880, and the number of test samples W is 720.

We used two well-known error evaluation indices: mean absolute error (MAE) and mean square error (MSE). The calculation formulas are

$$\mathrm{MAE} = \frac{1}{W}\sum_{w=1}^{W}\left|\hat{x\_in}_{j,t}^{w} - x\_in_{j,t}^{w}\right|, \tag{9}$$

$$\mathrm{MSE} = \frac{1}{W}\sum_{w=1}^{W}\left(\hat{x\_in}_{j,t}^{w} - x\_in_{j,t}^{w}\right)^{2}. \tag{10}$$

The lower the values of MAE and MSE, the higher the prediction accuracy of the model.

5.3. Parameter Settings. There are 48 candidate spatial-temporal features in total. All candidate spatial-temporal features and their corresponding MIC values are shown in Table 3. According to Figure 5, the MIC threshold is 0.7. We selected the candidate spatial-temporal features

Table 3: MIC value of the candidate features.

Candidate spatial-temporal feature	MIC value
Inbound passenger flow of BRT Huoche station at time period $t-1$	0.880
Inbound passenger flow of Lianban metro station at time period $t-1$	0.849
Inbound passenger flow of Lianbanguomao station at time period $t-1$	0.833
Inbound passenger flow of Hubin East Road metro station at time period $t-2$	0.803
Inbound passenger flow of Siming Court station at time period $t-1$	0.771
Inbound passenger flow of Lvjiayuanxiaoqu station at time period $t-1$	0.738
Inbound passenger flow of Lianbanbei station at time period $t-2$	0.691
Inbound passenger flow of Fengyulu station at time period $t-3$	0.613
Inbound passenger flow of Lianban Book City station at time period $t-2$	0.454
Inbound passenger flow of Huminglijing station at time period $t-1$	0.337
hour_t	0.879
Inbound passenger flow of Lianhualukou metro station at time period $t-1$	0.840
Inbound passenger flow of BRT Lianban station at time period $t-3$	0.826
Inbound passenger flow of Lianbanguomao station at time period $t-2$	0.803
Inbound passenger flow of Lianbanguomao station at time period $t-3$	0.771
Inbound passenger flow of Siming Court station at time period $t-2$	0.723
Inbound passenger flow of Fengyulu station at time period $t-1$	0.658
Inbound passenger flow of Huming station at time period $t-1$	0.522
Inbound passenger flow of Lianban Book City station at time period $t-3$	0.448
Inbound passenger flow of Huminglijing station at time period $t-2$	0.323
Inbound passenger flow of BRT Lianban station at time period $t-1$	0.861
Inbound passenger flow of Hubin East Road metro station at time period $t-1$	0.838
Inbound passenger flow of BRT Huoche station at time period $t-3$	0.821
Inbound passenger flow of Huoche station at time period $t-1$	0.783
Inbound passenger flow of Lianhualukou metro station at time period $t-3$	0.770
Inbound passenger flow of Huoche station at time period $t-3$	0.706
Inbound passenger flow of Lvjiayuanxiaoqu station at time period $t-3$	0.648
Inbound passenger flow of Huming station at time period $t-2$	0.499
Inbound passenger flow of Lianjingerli station at time period $t-1$	0.430
Inbound passenger flow of Huminglijing station at time period $t-3$	0.313
Inbound passenger flow of BRT Lianban station at time period $t-2$	0.858
day_t	0.836
Inbound passenger flow of Lianban metro station at time period $t-2$	0.810
Inbound passenger flow of Lianban metro station at time period $t-3$	0.781
Inbound passenger flow of Huoche station at time period $t-2$	0.750
Inbound passenger flow of Lvjiayuanxiaoqu station at time period $t-2$	0.696
Inbound passenger flow of Fengyulu station at time period $t-2$	0.640
Inbound passenger flow of Huming station at time period $t-3$	0.487
Inbound passenger flow of Lianjingerli station at time period $t-2$	0.417
Inbound passenger flow of BRT Huoche station at time period $t-2$	0.857
All_transfer_t	0.835
Inbound passenger flow of Lianhualukou metro station at time period $t-2$	0.807
Inbound passenger flow of Hubin East Road metro station at time period $t-3$	0.773
Inbound passenger flow of Lianbanbei station at time period $t-1$	0.744
Inbound passenger flow of Siming Court station at time period $t-3$	0.695
Inbound passenger flow of Lianbanbei station at time period $t-3$	0.631
Inbound passenger flow of Lianban Book City station at time period $t-1$	0.455
Inbound passenger flow of Lianjingerli station at time period $t-3$	0.395

with an MIC value greater than 0.7 as the significant features, with a total of 23 significant spatial-temporal features.

Based on the Xiamen transit data, four models were used in the experiment: Seasonal ARIMA (SARIMA), SVR model, backpropagation neural network (BP network), and LightGBM. The target of the forecast was inbound passenger flow with a frequency of 10 minutes. The details of each model are as follows:

(1) SARMIA: the seasonal period "S" is 144 because there are 1440 minutes in a day (144×10 min). ARIMA $(2, 1, 0) \times (0, 1, 1)144$ is finally used.

(2) SVR: the time period is 144 (144×10 min per day).

(3) BP network: the network structure is composed of three layers and the number of units in each layer is 10.

(4) LightGBM (without temporal and spatial features, only using historical passenger flow): Max_depth is

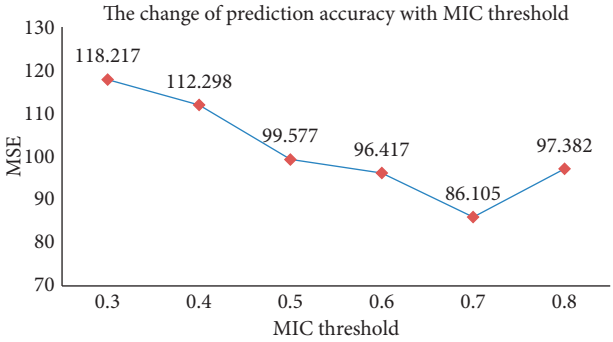


Figure 5: The change of prediction accuracy with MIC threshold.

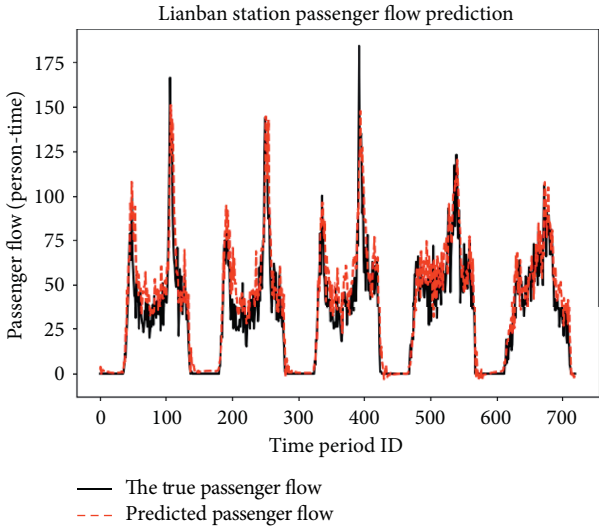


Figure 6: Prediction effect of SARIMA.

11 and Num_leaves is 512. To control overfitting, Min_data is 30.

(5) ST-LightGBM (with temporal and spatial features): Max_depth is 11 and Num_leaves is 1024. To control overfitting, Min_data is 12.

5.4. Experimental Results

Prediction Effect of the SARIMA Model. The prediction results of the model are shown in Figure 6, with the MAE value of 8.28 and MSE value of 164.06 (prediction results of a random fold).

Prediction Effect of the SVR Model. The prediction results of the model are shown in Figure 7. Without feature selection, as shown in Figure 7(a), the MAE value is 9.50 and MSE value is 170.67. With feature selection, as shown in Figure 7(b), the MAE value is 8.57 and MSE value is 155.94 (prediction results of a random fold).

Prediction Effect of the BP Network. The prediction results of the model are shown in Figure 8. Without feature selection, as shown in Figure 8(a), the MAE value is 9.40 and MSE value is 180.95. With feature selection, as shown in Figure 8(b), the MAE value is 8.34 and MSE value is 151.31 (prediction results of a random fold).

The Prediction Effect of ST-LightGBM (with Temporal and Spatial Features). The prediction results of the model are shown in Figure 9. Without feature selection, as shown in Figure 9(a), the MAE value is 6.95 and MSE value is 118.36. With feature selection, as shown in Figure 9(b), the MAE value is 5.77 and MSE value is 86.10 (prediction results of a random fold).

5.5. Analysis of the Experimental Results. The experiment results of the algorithms are shown in Table 4. The proposed ST-LightGBM achieved better performance than SARIMA, SVR, and BP network. Moreover, with feature selection, the proposed model achieved higher accuracy than the other models.

(1) As shown in Figure 10, we can see that the training time of ST-LightGBM is less than that of BP and SVR models, but longer than that of the SARIMA model, and so is the prediction time. This shows that the method has high computational efficiency and can be used in practical applications.

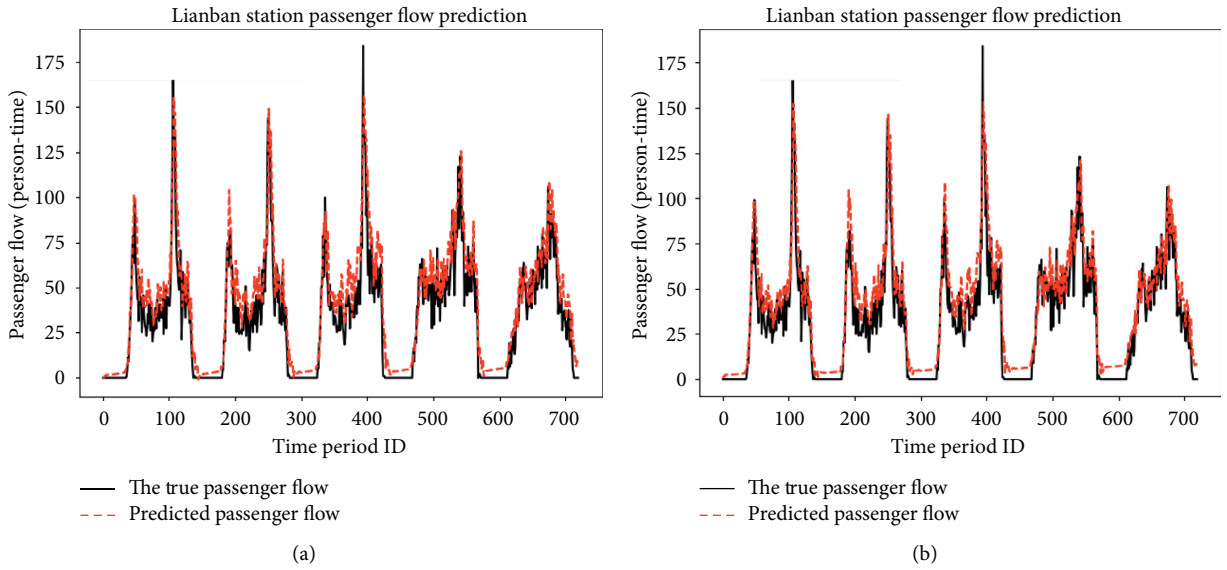


Figure 7: Prediction effect of SVR: (a) prediction results without feature selection and (b) prediction results with feature selection.

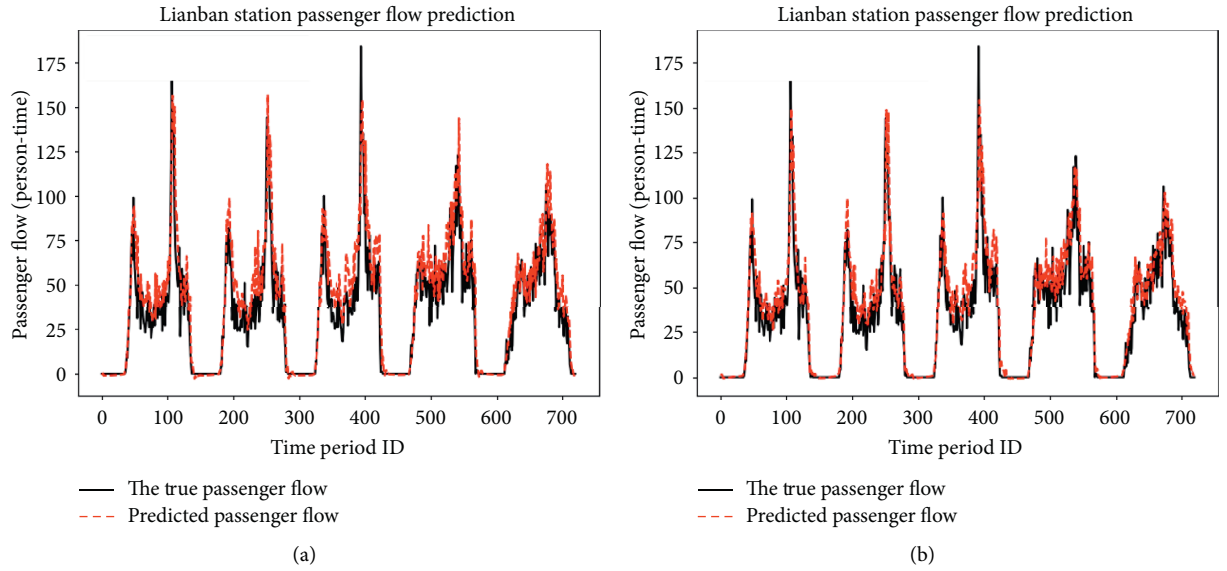


Figure 8: Prediction effect of the BP network: (a) prediction results without feature selection and (b) prediction results with feature selection.

(2) As shown in the second, third, and last rows of Table 4, with feature selection, the prediction accuracy of the models improved.

(3) From the viewpoint of MAE, the ST-LightGBM network was more accurate than the other models. The MAE error of the ST-LightGBM model was 30.41% less than that of the SARIMA model, 32.86% less than that of the SVR model, and 31.57% less than that of the BP network model.

(4) Moreover, in terms of MSE, the MSE error of ST-LightGBM model was 46.78% less than that of the SARIMA model, 43.77% less than that of the SVR model, and 44.39% less than that of the BP network model.

(5) According to the standard deviation, we can see that the ST-LightGBM method had better stability than the other models. Therefore, the proposed ST-LightGBM method is more suitable for short-term passenger flow forecasting for rail transit.

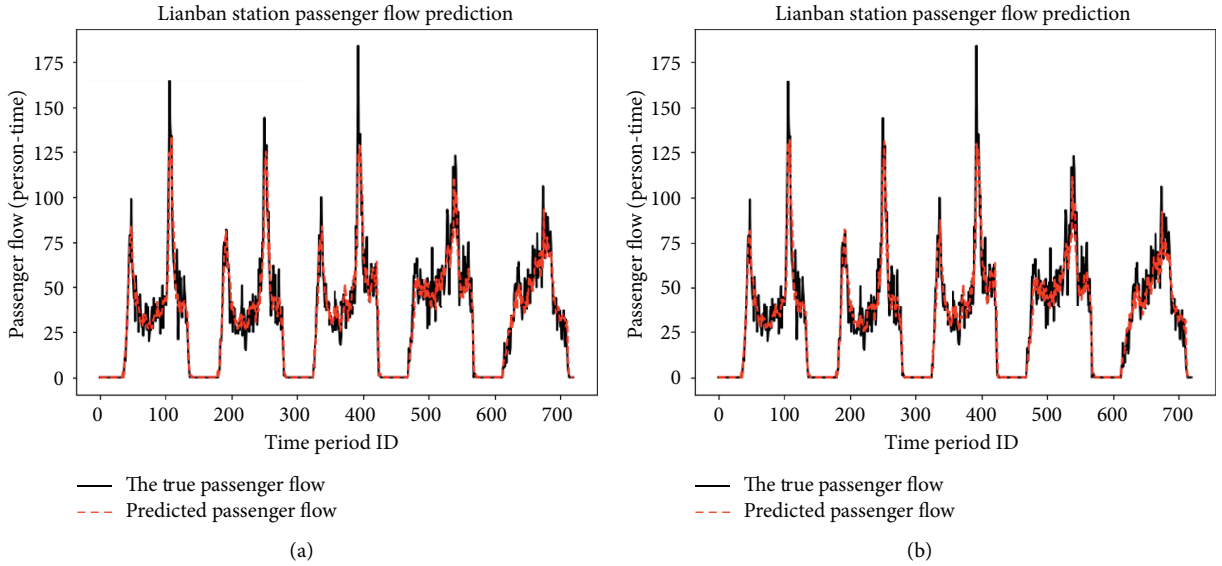


Figure 9: Prediction effect of ST-LightGBM: (a) prediction results without feature selection and (b) prediction results with feature selection.

Table 4: The MAE ($\overline{x} \pm \sigma$) and MSE ($\overline{x} \pm \sigma$) results of four experiments (listed mean $\overline{x}$ and standard deviation σ are averaged over 5 folds).

Experiment	Method	MAE	MSE
1	SARIMA	8.22 ± 0.51	161.16 ± 8.69
2	SVR (without feature selection)	9.65 ± 0.68	173.36 ± 9.62
	SVR (with feature selection)	8.52 ± 0.54	152.54 ± 7.92
3	BP network (without feature selection)	9.42 ± 0.49	181.35 ± 8.46
	BP network (with feature selection)	8.36 ± 0.42	154.22 ± 6.51
4	ST-LightGBM (without feature selection)	6.93 ± 0.13	118.27 ± 3.42
	ST-LightGBM (with feature selection)	5.72 ± 0.11	85.76 ± 2.42

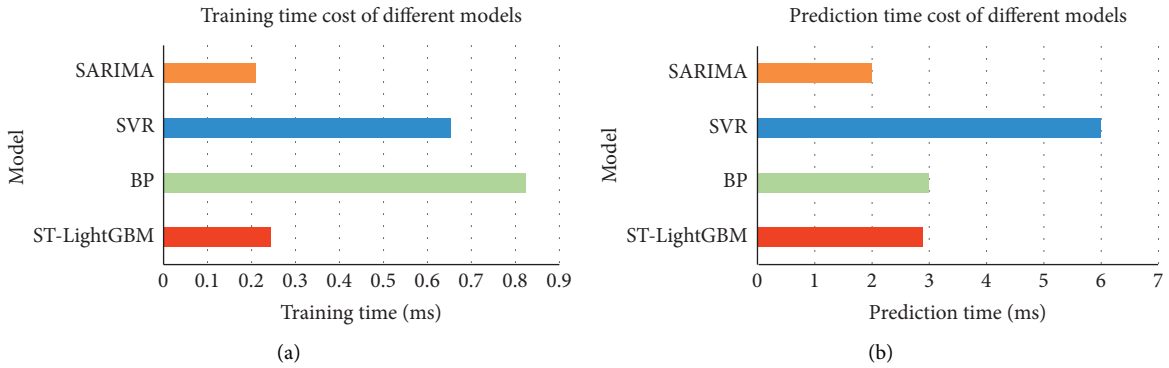


Figure 10: Average training time and prediction time cost of different models: (a) average training time cost and (b) average prediction time cost.

6. Conclusion and Future Work

We proposed a spatial-temporal LightGBM metro station passenger flow prediction model considering transfer passenger flow. Compared with previous research methods, this method considers the temporal and spatial features that affect inbound passenger flow in a metro station. Particularly, in terms of spatial features, we introduced the concept of spatial similarity and spatial transfer and established the ASHIM feature matrix and TPM feature matrix. Thus, the spatial influence factors were considered more comprehensively. Additionally, we used an MIC feature selection algorithm to obtain the important features; hence, the model input was simplified. Moreover, compared with other methods, the prediction accuracy of this method was also higher, so the proposed method has better applicability for the short-term prediction of metro station inbound passenger flow.

In future work, we will use scientific feature extraction methods [43] to further extract effective features from massive data. At present, it is difficult to further improve the prediction accuracy of the existing single model. In future work, we can further consider combining fast clustering algorithms [44–46] and other machine learning or deep learning models to establish a combined prediction model to further improve prediction accuracy. Moreover, we can combine distributed algorithms [47, 48] to improve the prediction efficiency of the model.

Data Availability

The data used to support the findings of this study are available in [42].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This study was supported by the National Natural Science Foundation of China Youth Fund (no. 1608209), the Project of Natural Science Foundation of Fujian Province of China (no. 2017J01090), the Project of Quanzhou City Science and Technology Program of China (no. 2018Z008), the 2018 Huaqiao University Research and Establishment of Postgraduate Education and Teaching Reform Project (no. 18YJG28), the Huaqiao University Postgraduate Research Innovation Ability Cultivation Program (no. 18014083027), the National Science Foundation (no. DMS-0907710), the National Natural Science Foundation of China (no. 18BTJ031), the Key Research and Development Program of Shaanxi Province (no. 2018ZDXM-GY-036), the Natural Science Foundation of Fujian Province of China (no. 2019J01080), the Fujian Province Science and Technology Plan (no. 2019H0017), and Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data (no. IPBED7).

References

[1] J. Cui, Y. Gao, J. Cheng, and L. Shi, "Study on the selection model of staying adjustment bus lines along rail transit," *Journal of Advanced Transportation*, vol. 2020, Article ID 6385359, 12 pages, 2020.

[2] China Association of Metros, "The statistics and analysis report of urban rail transit in 2019," 2020, http://www.camet.org.cn/tjxx/5133.

[3] K. N. Ali Safwat and T. L. Magnanti, "A combined trip generation, trip distribution, modal split, and trip assignment model," *Transportation Science*, vol. 22, no. 1, pp. 14–30, 1988.

[4] M. Florian and S. Nguyen, "A combined trip distribution modal split and trip assignment model," *Transportation Research*, vol. 12, no. 4, pp. 241–246, 1978.

[5] P. Shang, R. Li, Z. Liu, K. Xian, and J. Guo, "Timetable synchronization and optimization considering time-dependent passenger demand in an urban subway network," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 8, pp. 243–254, 2018.

[6] P. Shang, R. Li, Z. Liu, L. Yang, and Y. Wang, "Equity-oriented skip-stopping schedule optimization in an oversaturated urban rail transit network," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 321–343, 2018.

[7] 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*, p. 306, Jeju Island, South Korea, December 2009.

[8] Z. Guo, X. Zhao, Y. Chen, W. Wu, and J. Yang, "Short-term passenger flow forecast of urban rail transit based on GPR and KRR," *IET Intelligent Transport Systems*, vol. 13, no. 9, pp. 1374–1382, 2019.

[9] G. Ke, Q. Meng, T. Finley et al., "Lightgbm: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, pp. 3146–3154, MIT Press, Cambridge, MA, USA, 2017.

[10] D. Jeffery, K. Russam, and D. Robertson, "Electronic route guidance by AUTOGUIDE: the research background," *Traffic Engineering & Control*, vol. 28, 1987.

[11] M. El Esawey, "Daily bicycle traffic volume estimation: comparison of historical average and count models," *Journal of Urban Planning and Development*, vol. 144, Article ID 04018011, 2018.

[12] H. Yang, J. Yang, L. D. Han et al., "A Kriging based spatiotemporal approach for traffic volume data imputation," *PLoS One*, vol. 13, Article ID e0195957, 2018.

[13] J. Guo, W. Huang, and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[14] P. Jiao, R. Li, T. Sun, Z. Hou, and A. Ibrahim, "Three revised Kalman filtering models for short-term rail transit passenger flow prediction," *Mathematical Problems in Engineering*, vol. 2016, Article ID 9717582, 10 pages, 2016.

[15] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *European Transport Research Review*, vol. 7, p. 21, 2015.

[16] M. Milenković, L. Švadlenka, V. Melichar, N. Bojović, and Z. Avramović, "SARIMA modelling approach for railway passenger flow forecasting," *Transport*, vol. 33, pp. 1113–1120, 2016.

[17] S. Anvari, S. Tuna, M. Canci, and M. Turkay, "Automated Box-Jenkins forecasting tool with an application for passenger demand in urban rail systems," *Journal of Advanced Transportation*, vol. 50, no. 1, pp. 25–49, 2016.

[18] L. Li, Y. Wang, G. Zhong, J. Zhang, and B. Ran, "Short-to-medium term passenger flow forecasting for metro stations using a hybrid model," *KSCE Journal of Civil Engineering*, vol. 22, no. 5, pp. 1937–1945, 2018.

[19] L. Tang, Y. Zhao, J. Cabrera, J. Ma, and K. L. Tsui, "Forecasting short-term passenger flow: an empirical study on Shenzhen metro," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3613–3622, 2019.

[20] M. Gensuo, Z. Liqin, and L. Miao, "Subway station passenger flow forecast based on mixed kernel support vector machine optimized by golden section chaotic particle swarm optimization," *Computer Engineering and Applications*, vol. 2015, p. 44, 2015.

[21] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.

[22] G. A. Davis and N. L. Nihan, "Nonparametric regression and short-term freeway traffic forecasting," *Journal of Transportation Engineering*, vol. 117, no. 2, pp. 178–188, 1991.

[23] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transportation Research Part C: Emerging Technologies*, vol. 66, pp. 61–78, 2016.

[24] J. Roos, G. Gavin, and S. Bonnevay, "A dynamic Bayesian network approach to forecast short-term urban rail passenger flows with incomplete data," *Transportation Research Procedia*, vol. 26, pp. 53–61, 2017.

[25] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.

[26] Y. Bai, Z. Sun, B. Zeng, J. Deng, and C. Li, "A multi-pattern deep fusion model for short-term bus passenger flow forecasting," *Applied Soft Computing*, vol. 58, pp. 669–680, 2017.

[27] H. Zhu, X. Yang, and Y. Wang, "Prediction of daily entrance and exit passenger flow of rail transit stations by deep learning method," *Journal of Advanced Transportation*, vol. 2018, Article ID 6142724, 11 pages, 2018.

[28] L. Liu and R.-C. Chen, "A novel passenger flow prediction model using deep learning methods," *Transportation Research Part C: Emerging Technologies*, vol. 84, pp. 74–91, 2017.

[29] Q. Chen, X. L. Di Wen, D. Chen, H. Lv, J. Zhang, and P. Gao, "Empirical mode decomposition based long short-term memory neural network forecasting model for the short-term metro passenger flow," *PLoS One*, vol. 14, 2019.

[30] Y. Liu, Z. Liu, and R. Jia, "DeepPF: a deep learning based architecture for metro passenger flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 18–34, 2019.

[31] Y. Han, S. Wang, Y. Ren, C. Wang, P. Gao, and G. Chen, "Predicting station-level short-term passenger flow in a citywide metro network using spatiotemporal graph convolutional neural networks," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 243, 2019.

[32] Q. Tang, M. Yang, and Y. Yang, "ST-LSTM: a deep learning approach combined spatio-temporal features for short-term forecast in rail transit," *Journal of Advanced Transportation*, vol. 2019, Article ID 8392592, 8 pages, 2019.

[33] X. Dai, L. Sun, and Y. Xu, "Short-term origin-destination based metro flow prediction with probabilistic model selection approach," *Journal of Advanced Transportation*, vol. 2018, Article ID 5942763, 15 pages, 2018.

[34] L. Deng, J. Pan, X. Xu, W. Yang, C. Liu, and H. Liu, "PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine," *BMC Bioinformatics*, vol. 19, p. 522, 2018.

[35] G. Du, L. Ma, J.-S. Hu et al., "Prediction of 30-day readmission: an improved gradient boosting decision tree approach," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 3, pp. 620–627, 2019.

[36] A. Kadiyala and A. Kumar, "Applications of python to evaluate the performance of decision tree-based boosting algorithms," *Environmental Progress & Sustainable Energy*, vol. 37, no. 2, pp. 618–623, 2018.

[37] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," *Physica A: Statistical Mechanics and Its Applications*, vol. 534, Article ID 122370, 2019.

[38] X. Gu, Y. Li, and J. Jia, "Feature selection for transient stability assessment based on kernelized fuzzy rough sets and memetic algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 64, pp. 664–670, 2015.

[39] D. N. Reshef, Y. A. Reshef, H. K. Finucane et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[40] S. Feng and G. Cai, "Passenger flow forecast of metro station based on the ARIMA model," in *Proceedings of the 2015 International Conference on Electrical and Information Technologies for Rail Transportation*, pp. 463–470, Springer, Berlin, Germany, 2016.

[41] J. Jin, Y. H. Wang, and M. Li, "Prediction of the metro section passenger flow based on time-space characteristic," *Applied Mechanics and Materials*, vol. 397–400, pp. 1038–1044, 2013.

[42] Z. Zhang, C. Wang, Y. Gao, Y. Chen, and J. Chen, "Passenger flow forecast of rail station based on multi-source data and long short term memory network," *IEEE Access*, vol. 8, pp. 28475–28483, 2020.

[43] Q. Ke, J. Zhang, H. Song, and Y. Wan, "Big data analytics enabled by feature extraction based on partial independence," *Neurocomputing*, vol. 288, pp. 3–10, 2018.

[44] Y. Chen, L. Zhou, S. Pei et al., "KNN-BLOCK DBSCAN: fast clustering for large-scale data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–15, 2019.

[45] Y. Chen, X. Hu, W. Fan et al., "Fast density peak clustering for large scale data based on kNN," *Knowledge-Based Systems*, vol. 187, Article ID 104824, 2020.

[46] Y. Chen, S. Tang, N. Bouguila, C. Wang, J. Du, and H. Li, "A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data," *Pattern Recognition*, vol. 83, pp. 375–387, 2018.

[47] W. Wei, X. Xia, M. Wozniak, X. Fan, R. Damaševičius, and Y. Li, "Multi-sink distributed power control algorithm for Cyber-physical-systems in coal mine tunnels," *Computer Networks*, vol. 161, pp. 210–219, 2019.

[48] W. Wei, X. Fan, H. Song, X. Fan, and J. Yang, "Imperfect information dynamic Stackelberg game based resource allocation using hidden Markov for cloud computing," *IEEE Transactions on Services Computing*, vol. 11, no. 1, pp. 78–89, 2016.

Hindawi
Scientific Programming
Volume 2020, Article ID 8856944, 14 pages
https://doi.org/10.1155/2020/8856944

Research Article

Internet Penetration and Regional Financial Development in China: Empirical Evidence Based on Chinese Provincial Panel Data

Qingquan Jiang[1], **Xiaosan Zhang**[2], **Qiaozhen Lin**[1], **Guofu Chen**[3], **Rui Zhang**[1] and **Songxian Liu**[1]

[1]*School of Economics & Management, Xiamen University of Technology, Xiamen 361024, China*
[2]*Research Centre for Belt & Road Financial and Economic Development, Xiamen National Accounting Institute, Xiamen 361005, China*
[3]*School of Management, Xiamen University, Xiamen 361005, China*

Correspondence should be addressed to Xiaosan Zhang; zxs@xnai.edu.cn, Guofu Chen; 1160848498@qq.com, and Rui Zhang; r.zhang@ymail.com

Received 12 March 2020; Revised 18 June 2020; Accepted 14 July 2020; Published 3 August 2020

Academic Editor: Chenxi Huang

The Internet has revolutionized the patterns of financial development and economic growth. To assess the impacts of internet penetration on the financial industry, this paper analyzed ten-year Chinese provincial panel data and concluded that regional Internet penetration accelerates financial development. Furthermore, the efficiency of Internet investment in underdeveloped provinces is better than that in developed provinces. More meaningfully, Internet penetration promotes the transparency of the securities market and regional financial participation. This indicates that Internet technology facilitates the advancement of the finance industry and the securities market.

1. Introduction

The development of regional financial industry is affected by many factors, among which the level of regional economic development and the level of information technology are important to sustainable development of financial industry [1]. The influence of Internet on economy is various, from improving technical productivity to increasing foreign direct investment, from inflation to political and economic problems, and from corruption to democratic and free issues [2, 3].Governments of many developed and developing countries have recognized the huge growth potential of the Internet for their economy, and some countries have taken measures to greatly increase the construction of Internet infrastructure and increase the penetration rate of the Internet [4].

According to Moore's law (Moore), information technology represented by electronic information components and integrated circuits will continue to be updated [5]. Not only has the development of the Internet given birth to the development of a network economy, but also it has produced subversive changes in various industries [6]. In recent years, the extensive use of the Internet has produced a vast amount of data. Data are among the most useful tools in the Internet age. In fact, it has been noted that the Internet has entered the "Data Era" from the "Digital Era" [7]. Schumpeter pointed out that both endogenous R&D (Research and Development) investment and external technological shocks are driving forces of economic development [8]. Lucas argued that the accumulation of technological innovation, material objects, and human capital is an important driving force for economic growth [9]. The development of Internet technology can directly drive the development of network economy, promote the change of information circulation and payment mode, stimulate the development of other industries, and indirectly drive the economic development of other industries. Internet applications in the financial sector have a long history [10]. From the statistics in

Figures 1(a) and 1(b), it can be seen that, with the development of the network economy, financial services such as online payment and online banking have expanded to the Internet. However, the number of users and usage of online stocks have not gradually increased with the development of the Internet economy, indicating that the role of Internet in the development of the financial industry is different from that in traditional industry, and financial development under Internet environment has different qualities.

The Shanghai Stock Index (Figure 2) shows that since the emergence of the Internet in China in 1997, the proportion of online stocks reached 8% according to the 1999 survey. The unprecedented development of the Internet economy around 2000 brought about a short "Bull Market" and led to the first increase in online stocks. The collapse of the Internet bubble brought about a "Bear Market" for several years; at the same time, the proportion of online stocks' users fell. Subsequently, between 2005 and 2008, the proportion of online stocks' users and the Shanghai Stock Index had the same trend. In particular, around 2007, nearly one-fifth of the web APPs were online stocks. In general, the securities market is closely related to the development of the Internet. Therefore, the Internet has become an important factor affecting the development of the financial industry.

Scholars believe that financial intermediation has a positive relationship with economic growth [11], and the development of stock market has also been causally linked to economic growth [12]. However, from the perspective of financial development, the driving force comes from two aspects: one is the stimulation of the financial industry by economic development and the second is the internal innovations of the financial industry itself. However, whether it is economic development or innovation in financial instruments, this movement requires the conditions essential for scientific and technological progress. Not only does the Internet provide a new tool for the development of the financial industry, but also it provides a fast and convenient platform for financial innovation.

On the one hand, the Internet increases the efficiency of the financial industry, which, in turn, has stimulated economic growth. Studies have confirmed that information and communication technologies have a stimulating effect on economic growth, and telecommunications infrastructure has an impact on regional economic growth [13, 14]. Levine believes that a well-established financial business can improve information asymmetry, as well as identifying and investing in the most competitive and innovative companies and industries, thereby improving capital allocation efficiency [15]. The Internet can transmit information quickly and conveniently, which can greatly improve the asymmetry of information. For example, high-frequency transactions in the securities market require real-time information. The famous US network provider Spread Networks spent two years setting up the most direct fiber optic cable route possible between New York and Chicago, with a network round-trip speed 3 millisecond faster than the original [16]; therefore, many high-frequency traders have become its customers.

On the other hand, the development of the Internet and financial technology has provided a powerful platform for financial innovation, thus leading to "Internet finance" [17]. Hou et al. consider Internet finance an integration of Internet technologies and financial activities, playing an important role in payment and settlement, resource allocation, risk management, and networking channels [18]. Chang and Deng have argued that Internet financing is an advanced financial service that covers not only e-payment and e-banking, but also credit, guarantees, assurances, pledges, and collateralization [19]. Wang et al. define Internet finance as an emerging financial model that implements fund accommodation, payment, and infomediary services [20]. Therefore, the Internet plays a major role in promoting the innovations and development of the financial industry.

The structure of this paper is as follows: section 2 is literature review and proposes research hypotheses thereafter; section 3 mainly introduces the selection of indicators and data sources; section 4 conducts empirical analysis to verify the correlation between Internet development and regional financial development, financial development structure, and financial participation of investors; and section 5 is conclusions and recommendations with slight limitations and prospects of this research.

2. Literature Review

2.1. Internet Development Level and Regional Financial Development Level. Financial development is influenced by certain economic factors [21]. In the theory of exogenous economic growth, the Solow–Swan growth model identifies the role of technological growth in economic growth [22]. The development of the Internet belongs to both the infrastructure of economic development and the scope of technological progress. Internet infrastructure is closely connected to the economy [23] because ICT affects growth through fostering technology diffusion and innovation, enhancing the quality of decision-making by firms, increasing demand, and reducing production costs. Therefore, Vu argued that ICT is a source of economic growth [24]. Pradhan et al. concluded that ICT's role in GDP growth has been confirmed [25]; in addition, Binuyo and Aregbeshola explained the effect of ICT investment on the economy [26]. Tranos Granger tested the causality between the Internet infrastructure and the economic development of European city regions, with the results showing that ICT infrastructure is a necessary condition for economic development [27]. Pradhan et al. found that an embellishment of ICT infrastructure improved economic development in the form of per capital GDP in G20 countries [28]. Pradhan et al. confirmed that ICT infrastructure determines long-term economic growth in most Asian countries [28]. Using the statistical data of 1996–2007, Czernich et al. used an instrumental variable model and proved that when the broadband penetration rate increases by 10%, it contributes to a 0.9–1.5% gain in economic growth in OECD countries [29].

Chinese ICT is developing rapidly in the 21st century. One study indicated that the Chinese ICT infrastructure

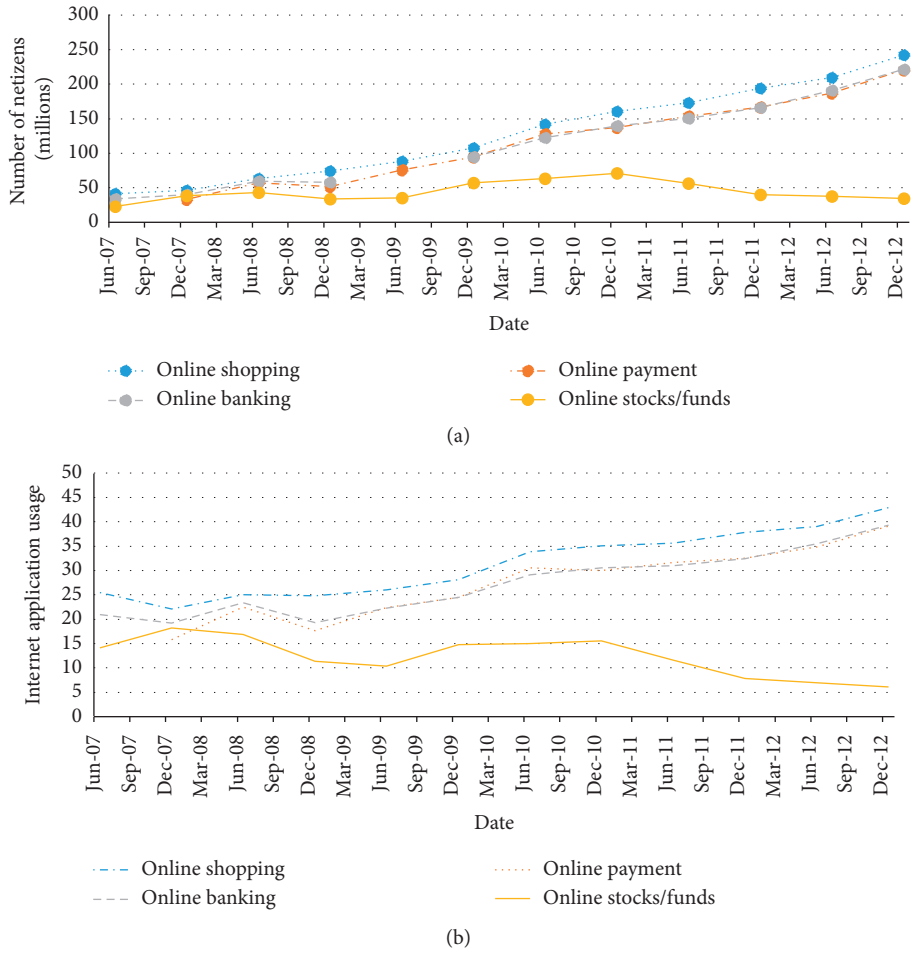


Figure 1: Statistics of Internet applications from 2007 to 2012 (data source: CNNIC China Internet Network Information Center).

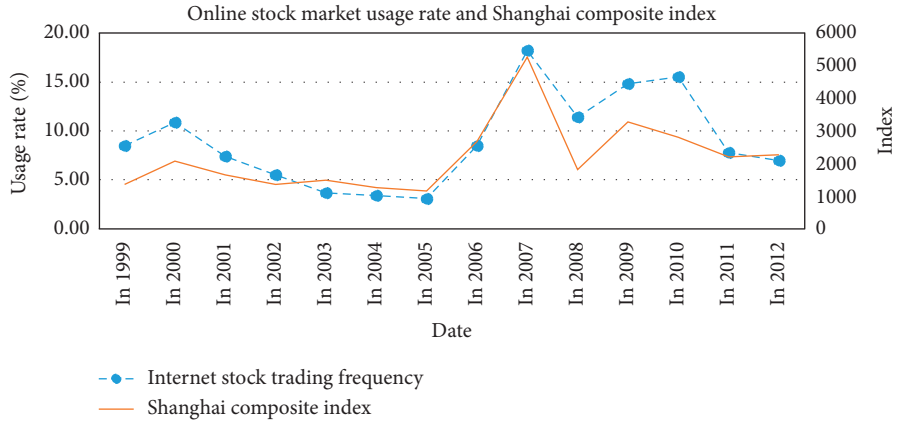


Figure 2: Online stock usage proportion and Shanghai Stock Index from 1999 to 2012 (data source: CNNIC China Internet Network Information Center and Wind database).

contribution rate has far exceeded that of the United States, Japan, and Germany since 2000. The results of Heshmati and Yang's research suggest that China has reaped the benefits of ICT investment [30]. Liu and Hu studied the spillover effects of transportation, energy, and information infrastructure on Chinese economic growth. The results show that transportation infrastructure and information infrastructure have a significant spillover effect on Chinese economic growth [31]. Regarding the promotion of technological progress, Li and Jing believe that the increase in the Internet penetration rate has different promoting effects on regional economic growth, and this stimulating effect reaches its

maximum after five years [32]. Luo et al. found that the marginal contribution of Chinese telecom investment to Chinese economic growth averaged 2.376%, which was 6.758% higher than other social infrastructure investments [33]. Zhong et al. concluded that a 10% increase in broadband penetration will boost the national economy by 0.444% [34]. Sharon [35] found that the Internet as a distribution medium was regarded by many enterprises as a legitimate way to reduce operating costs, and people were increasingly confident in using the Internet as a medium for transactions. Therefore, the author believes that the development of regional economy has promoted the development and penetration rate of the Internet to some extent, and the development of the Internet will further promote the improvement of financial development level. Based on the above analysis, this paper proposes the following hypotheses.

Hypothesis 1. Internet penetration promotes the development level of regional finance. That is, the higher the Internet penetration, the higher the level of regional financial development.

2.2. Internet Development Level and Financial Participation of Regional Investors. Lack of regional infrastructure construction may hinder the transmission of information, while afterwards information transmission or low efficiency may reduce people's understanding of things and decision-making ability. The development of Internet technology can speed up information transmission and alleviate various problems caused by information asymmetry, such as policy changes and financial investment. A few studies have argued that social interaction and trust promote the participation of citizens in the stock market [36–38]. The general investment losses caused by the stock market downturn will reduce the positive effect of social interaction, and the positive impact of social interaction on the participation of low-education residents in the stock market is more obvious [28]. Wang and Tian found that, in addition to the impact of economic development level and other factors on stock market participation, Chinese investors participation in the stock market will also change with that of financial market environment [39].

The popularization of the Internet enables the spread of scattered and asymmetric information and accelerates the utilization rate of information. According to the growth of modern endogenous theory, the Internet can promote the development of the industry by promoting the level of science and technology, big data analysis, information network circulation, and so on [40]. The spread of the Internet may affect the innovation capacity of the economy through the creation of knowledge spillover and the development of new products, new processes, and new business models to promote growth. The ease of access and dissemination of information facilitate the creation of new ideas and technologies [41]. The extensive use of the Internet fundamentally changes and improves the Internet information processing, leading to a significant increase in the productivity of companies using information technology [42]. The Internet makes it possible to exchange data across multiple locations and helps decentralize information processing. It may also help the emergence of new business enterprise cooperation model relying on the spatial exchange of a large amount of information which promotes the development of enterprise competition and innovation process. The Internet is likely to increase market transparency, thus enhancing competition and promoting industry reform and development [43]. Noh and Yoo analyzed the relationship between Internet usage rate, income inequality, and economic growth, concluding that for countries with high income inequality, Internet usage has a negative impact on economic growth [44].

The difference in stock market participation is caused by information asymmetry [45]. The popularity of the Internet provides investors with a broad information platform, thus greatly reducing the cost of financial information dissemination, so the Internet enables investment in financial activities to respond to market changes in a timely manner [46]. Another purpose of this research is to study the impact of Internet popularization on willingness to participate in the stock market. Therefore, this research believes that the development of the Internet will further affect the transmission and circulation of information. The cost and time for investors to obtain information will be reduced, and they can learn more reliable information in a timely manner, which will reduce financial participation to some extent and avoid blind investment. Based on the above analysis, this paper proposes the following hypothesis.

Hypothesis 2. Internet penetration inhibits the financial participation of regional investors. That is, the higher the Internet penetration, the smaller the participation of investors in finance.

Through the analysis of literature review, it can tell that the development and popularization of the Internet have greatly improved the flow and acquisition of information and numbers. As an important industry to measure the level of economic development, the level of financial development and the business of all financial industry are affected [47]. In addition, the improvement of information acquisition cost and efficiency alleviates the information asymmetry, and many investors will have a better understanding of the specific information of financial investment and make decisions [48], which may create new challenges for the development of the financial industry. Therefore, the research on the influence of Internet popularization and development level on the development of the financial industry and the influence of investors' financial participation can explore these issues.

There are different ways of selecting indicators for Internet penetration and financial development. Some scholars choose Telecommunications usage, such as telephone usage, for example, telephone penetration rate to measure the development of science and technology level and information flows, but these indicators can only measure part of the information infrastructure. The network port number, optical fiber length, and other data are not of the most direct practical significance to users due to regional and technical

differences, because the development of the Internet requires not only the investment of hardware equipment, but also the investment of technology. Later, the study of this paper also confirmed that although the information infrastructure investment in the region sent by the economy is more, the penetration rate is relatively high, but the efficiency of investment is low. Therefore, this paper updated the measurement indicators more comprehensively measuring the level of Internet development and financial development and further verified it through the robustness test.

3. Methodology

3.1. Variable Selection

3.1.1. Internet Development Level. The level of Internet development is mainly measured from two aspects. The first aspect is the investment in Internet infrastructure. Because investment in the Internet involves hardware and software in various fields such as telecommunications infrastructure, information transmission, the software industry, and digital products, the data are complicated, and this paper uses the capital stock of information transmission, computer service, and software industry fixed assets investment of China as substitute variables for Internet infrastructure investment, including industry infrastructure investment, technology investment, and asset investment. The second aspect is Internet popularization, including the Internet penetration rate, proportion of netizens, and number of netizens. The Internet penetration rate is calculated from the ratio of the number of netizens to the population of the region. The number of netizens is calculated based on the number of Internet users in the China Internet Network Information Center (CNNIC). The local population is based on the census data or the results of calculations published by the National Bureau of Statistics for that particular year.

The methods for estimating the capital stock of fixed assets investment in information transmission, computer services, and the software industry are as follows: first, this paper estimates the total fixed capital stock of the whole society by using the data of fixed assets investment of the whole society from 1985 to 2002. Second, based on the average ratio of fixed assets investment in information transmission, computer services, and the software industry from 2003 to 2006, the ratio data of information transmission, computer service, and software industry from 1985 to 2002 is collected. Finally, based on the perpetual inventory method (PIM), for estimating the capital stock of 2003, 2003 is made the base year and used for estimating capital stock from 2003 to 2012; the equation is as follows:

$$K_{it} = K_{it-1}(1-\varepsilon) + I_{it}. \tag{1}$$

This paper adopted the methods of Zhang et al. [49] and Jin [50]. 10% of the earliest investment as the capital stock of the base period is used, and the base period (1985) is the constant price, as shown in the following equation:

$$K_i = \frac{I_{i+1}}{g+\varepsilon}, \tag{2}$$

where K is the capital stock, I is the investment amount, ε is the depreciation rate, and g is the average growth rate during the observation period. This paper assumes that the depreciation rate of infrastructure investment in each region is equal. According to the calculation results of Jin [50], the depreciation rate of Chinese infrastructure is 9.2%. The ratio of the local investment amount to the national investment amount is applied as the weight to estimate the region's capital stock for the base year (2003). The calculation method is as follows:

$$K_{i,2003} = \frac{I_{i,2003}}{\sum I_{i,2003}} \cdot K_{2003}. \tag{3}$$

3.1.2. Financial Development Level. This paper uses two indicators to measure the level of regional financial development. First, in order to fully reflect the level of financial development, referring to Goldsmith [51], this paper uses the ratio of the contribution of the financial industry's production value to the GDP of each province as the financial development level indicator and adopts the per capital financial industry production value as another measure. Second, according to the research of King and Levine [52], the financial development indicators are refined and extended to the insurance and banking industries. In order to better measure the financial development structure of each region, this paper takes the local stock market value as a substitute variable for the securities market. In summary, this paper uses the insurance industry income, the year-end bank loan amount, and the stock market value as the main indicators of the financial development structure of the region.

(1) Regional Economic Development Level. Referring to the index system of economic growth created by King and Levine (1993), this paper uses per capital GDP as the main variable of the regional economic development level. Similarly, per capital consumption was used as an indicator of household consumption level.

(2) Residents' Income Level. This paper introduced the disposable income of residents as the control variable of the per capital disposable income of residents. The relevant literature suggests that property holdings may also be a significant factor in financial participation [53]. Therefore, this article introduces per capital property holdings as an alternative variable to residents' current disposable income.

(3) Education Level. The popularity of the Internet is related to the education level in each region. In this paper, sample survey data of the popularization rate of junior high school education, high school education, and college education in each province as the education level index of each region is applied. Using the years of education as the weight, weighted average is applied and the average level of education in the

area is obtained. The junior high school education is three years, so the weight is 3; high school is 6; and junior college is 10.

(4) Investor Participation in the Stock Market. This article uses the number of per capital accounts opened in each year as an indicator of stock market participation; in addition, (Financial Industry Output Value/Per Capita Disposable Income), (Securities Transaction Amount/Per capita Disposable Income), and (Stock Trading Volume/Per capita Disposable Income) were used as indicators of financial participation intention, indicating the financial production value of the per capital disposable income contribution and the per capital disposable income controlling the volume of securities and stocks traded.

3.2. Data Sources. On April 20, 1994, China was connected to the Internet through a 64 K international special line, and the Chinese Internet was born. The development of Internet in China has experienced three major waves: the first Internet wave (1994–2000), the second Internet wave (2001–2008), and the third Internet wave (2009–2014). Based on the development of Internet in China, this study chooses 10 years as the research period to analyze the relationship among Internet development, financial development, and regional investors. As the largest developing country, China has made remarkable progress in science, technology, and economic development. Therefore, it is valuable to study the relationship between Internet and economic development in China. The data adopted in this study are samples from more than 30 regions in China. Due to the imbalance of economic development level in various regions in China, the Internet development and economic development in different regions have their own characteristics. Therefore, the research on more than 30 regions in China can also help to understand the role of Internet technology level towards economic development in different regions.

The data on Internet penetration rate, proportion of netizens, number of netizens, and so forth mainly come from the China Internet Network Information Center (CNNIC); network speed data mainly come from the speed report released by 360 companies and broadband network alliances. Information industry investment data come from the China Statistical Yearbook and the official website of the National Bureau of Statistics. Macro data such as per capital GDP, education level, and financial industry data come from the national data website published by the National Bureau of Statistics. For property holdings, due to a lack of accurate, relevant data, this paper mainly used the annual data of residential sales across the country, as well as the data from the straight flush iFinD collected from the China Statistical Yearbook. Bank deposits, bank loans, and premium income come from the China Statistical Yearbook and the China Financial Statistics Yearbook; securities trading volume, stock trading quotas, and stock account opening data are from the Wind database and the China Securities and Futures Statistical Yearbook. Descriptive statistics for the main variables are shown in Table 1.

4. Empirical Test Results and Discussion

4.1. Internet Development Level and Regional Financial Development Level. In order to examine the impact of Internet development on regional financial development, the impact of Internet development levels on the overall level of regional financial development is examined. This paper represents the overall development level of regional finance using the financial industry GDP ratio and per capital financial production value. It expresses the development level of the Internet by the two indicators of capital transmission and the Internet penetration rate of information transmission, computer service, and software industry fixed assets investment in various regions. The capital stock represents the investment in the Internet in various regions, and the penetration rate indicates the output of the Internet in various regions. Based on the basic Cobb–Douglas model, the relationship between the capital stock of fixed asset investment and the Internet penetration rate of information transmission, computer service, and software industry in each region is estimated by constructing a basic Cobb–Douglas production function [53], as shown in the following formula:

$$\text{Networks} = \text{A} \times \text{Employment}^{\alpha} \times \text{Investment}^{1-\alpha}. \quad (4)$$

A logarithm formula (equation (4)) can be obtained:

$$Ln(\text{Networks}) = \ln\text{A} + \ln\ (\text{employment})^{\alpha} + \ln\ (\text{Investment})^{1-\alpha}, \quad (5)$$

where Networks is the Internet penetration rate that year, Employment represents the number of employees, and Investment represents the capital stock of fixed assets investment in information transmission, computer service, and software industry in various regions.

Table 2 shows the estimated capital stocks of fixed asset investment in information transmission, computer services, and software industries in various regions of China. In Table 3(panel A and panel B), the capital stock data in Table 2 is applied. The capital stock of fixed asset investment in information transmission, computer service, and software industry in each region is used as the capital investment for Internet development to test whether it has the same effect as the level of Internet popularization. Table 4 shows the impact of network popularity on financial development. In addition, the analysis of the individual fixed effect model in equation (1) found that the analysis results are not significant, and there is no time lag effect on the penetration rate of the Internet and the production value of the financial industry. Therefore, the subsequent analysis in this paper uses the current period data.

In terms of analysis methods, fixed effects model, random effects model, and mixed effects model are applied accordingly conducting the Hausman test. If the zero hypothesis is rejected, the fixed effect can be used; otherwise, if the zero hypothesis is accepted, the random effect can be

Table 1: Descriptive statistics of the main variables (2003–2012).

Parameter	Number of observations	Minimum value	Maximum value	Mean	Standard deviation
Internet penetration	300	0.02	0.68	0.20	0.15
Internet penetration in neighboring provinces	300	0.04	0.56	0.21	0.13
Information industry capital stock (100 million)	300	182.87	75,122.72	11,402.80	11,014.09
Education level (% junior middle school)	300	0.02	0.05	0.04	0.01
Education level (% senior high school)	300	0.00	0.03	0.01	0.00
Education level (% technical secondary school)	300	0.00	0.03	0.01	0.01
Stock market value (million)	300	75.84	204,030.20	6022.71	18,825.92
Bank deposit (year-end balance, 100 million yuan RMB RMB)	300	544.36	91,589.51	14,986.14	15,588.60
Bank loan (year-end balance, 100 million yuan RMB RMB)	300	566.98	58,611.22	10,371.00	10,427.83
Premium income (year-end balance, 100 million yuan RMB RMB)	300	7.37	1593.25	283.21	277.39
Securities transaction volume (millions)	300	84.02	255.09	226.96	411.01
Stock trading volume (millions)	300	5.98	1680.54	171.21	305.52
GDP per capital (yuan RMB)	300	3257.00	76,074.00	20,575.31	14,754.56
Consumption level (yuan RMB)	300	2502.00	35,439.00	8097.25	5238.15
Per capital disposable income (yuan RMB)	300	6530.48	36,230.48	13,638.90	5579.27
Proportion of finances in GDP	300	0.01	0.13	0.03	0.02
Securities account opening ratio (%)	240	0.29	15.85	3.15	3.19
Property holding level	240	9956.00	795,823.00	171,652.00	20,447.47

Table 2: Capital stock of fixed assets investment in information transmission, computer service, and the software industry in various regions of China.

District	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Beijing	4241.69	4638.63	6734.19	6670.15	10,861.47	12,228.68	17,777.49	24,146.84	25,044.59	34,136.06
Tianjin	1650.70	1785.16	2214.31	2386.66	3963.99	6578.45	6516.21	8037.71	15,588.12	15,864.54
Hebei	3300.41	4177.80	5722.15	7216.52	6709.46	11390.51	1571.11	6942.43	17,438.89	18,716.24
Shanxi	1747.39	2006.41	2609.04	3552.36	5299.97	8135.65	11,186.47	7026.68	5762.84	7505.47
Inner Mongolia	979.27	1275.02	1401.46	1748.46	6264.04	4296.40	6076.08	10,144.03	15,607.37	17,525.40
Liaoning	2834.21	3197.99	3460.55	8532.59	7976.93	13,819.78	16,831.64	24,686.06	23,386.82	28,059.61
Jilin	688.70	1094.23	2038.15	3602.46	4680.38	5421.52	7242.64	8357.87	8444.56	14,692.67
Heilongjiang	3765.63	5099.46	7219.75	9160.02	14,926.21	18,741.21	18,576.45	10,868.60	18,276.73	26,259.65
Shanghai	3774.01	3434.41	5421.24	11,161.60	12,347.69	13,495.30	15,969.20	19,647.74	18,227.63	25,574.65
Jiangsu	4407.45	4139.24	6760.51	7496.56	5987.25	7700.56	18,765.02	28,612.23	40,014.07	56,066.53
Zhejiang	4373.91	6699.41	8050.09	9053.48	13,950.17	16,061.02	20,332.72	26,741.82	30,686.09	23,399.52
Anhui	1708.91	1795.27	2831.85	3509.69	5134.30	6948.81	11,616.09	14,390.37	12,941.09	15,352.38
Fujian	3011.32	4844.71	5689.82	6354.37	8242.67	15,326.05	18,869.95	23,826.68	23,659.09	30,198.91
Jiangxi	2665.98	4101.32	5553.11	4798.71	7061.31	3863.46	6417.53	11,188.76	9878.10	10,660.67
Shandong	4444.95	3625.32	2502.17	3700.01	4421.66	6458.45	9033.33	9840.72	15,329.67	19,213.66
Henan	2494.30	3859.21	5636.76	11116.14	6594.47	6597.81	9658.11	9924.97	7077.50	9859.75
Hubei	2746.40	4552.03	4145.13	6759.41	6120.05	7476.12	8957.90	12,739.02	18,220.04	18,210.40
Hunan	2029.09	2818.07	4436.01	5228.04	7202.63	10,330.80	13,315.75	18,788.36	16,187.69	12,566.02
Guangdong	12,721.63	17,528.58	19,038.12	19,091.87	23,571.79	28,587.44	35,756.51	42,901.50	53,951.56	71,977.44
Guangxi	2301.90	3042.48	2972.91	3091.86	3580.46	4603.65	11,985.16	14,154.46	20,666.84	19,559.32
Hainan	251.11	773.74	1173.21	1339.45	1472.54	1913.14	2492.76	2965.70	6445.75	7360.04
Sichuan	5521.90	7945.98	9474.96	9718.24	11,486.24	18,109.42	24,972.60	31,763.29	28,310.04	30,964.00
Guizhou	1665.50	1586.67	2487.86	3876.47	4259.91	5608.91	7699.94	7869.21	3434.36	1344.70
Yunnan	2667.46	2940.08	4015.00	4184.72	6163.03	7288.46	8046.55	8728.58	10,913.08	12,582.88
Tibet	341.88	441.23	1041.42	944.75	1438.37	1181.19	1220.18	1870.41	1321.71	2685.19
Shangxi	1601.37	2845.89	3987.46	4264.07	7675.92	5199.23	7470.50	13,379.34	13,894.75	18,193.54
Gansu	870.24	1434.95	1361.19	1721.37	2009.78	1777.45	2482.08	3892.48	7205.24	6780.42
Qinghai	251.11	258.54	143.46	376.99	366.10	357.52	384.53	421.26	131.33	411.00
Ningxia	369.01	549.96	637.14	533.21	642.11	928.45	1662.70	2611.84	3615.04	2417.51
Xinjiang	2499.23	2294.66	3378.36	3472.63	4615.17	6283.54	5842.63	7077.23	10,559.14	9242.20

Note. Since Chongqing became a municipality in 1997, during most of the estimated period (1985–2002), Chongqing is attributed to Sichuan. Therefore, this paper follows the historical economic data and attributes Chongqing to Sichuan.

Table 3: Investment in Internet infrastructure and regional financial development level. Panel A: financial industry/GDP share and capital investment in Internet infrastructure. Panel B: per capital financial output and capital investment in Internet infrastructure.

	Mixed OLS model		Regional fixed effect model	
	1	2	3	4
Financial industry/GDP				
Intercept	−0.1746	−0.0846**	−0.1231***	−0.1587***
	0.0427	0.0366	0.0426	0.0327
Infrastructure capital stock	−0.0052***	−0.0026***	0.0021*	0.0021**
	0.0011	0.001	0.0012	0.0012
Education level (average)	0.0206***	—	0.009	—
	0.0069		0.0057	
Education level (above college)	—	0.0243***	—	0.002
		0.0031		0.0022
Per capital GDP	−0.0411***	−0.0439***	−0.0337***	−0.0338***
	0.0059	0.0052	0.0061	0.0061
Resident per capital disposable income	0.0144	0.0242	0.0354***	0.0344***
	0.0099	0.0079	0.0107	0.0107
Household consumption level	0.0689***	0.0525***	0.0212**	0.023**
	0.0091	0.0084	0.0095	0.0095
R^2	0.6407	0.7077	0.9568	0.9564
Observations	270	270	270	270
Per capital financial output				
Intercept	−0.9098	0.6033	−9.3739***	−8.8804***
	1.6503	1.442	2.0835	1.5919
Infrastructure capital stock	−0.3889***	−0.2675***	0.0973*	0.0972*
	0.0431	0.0413	0.0567	0.0567
Education level (average)	1.6017***	—	0.0142	
	0.2669		0.279	
Education level (above college)	—	1.1114***	—	0.0593
		0.1206		0.109
Per capital GDP	−1.0909***	−0.9737***	−1.256***	−1.2413***
	0.2282	0.1952	0.2973	0.2978
Resident per capital disposable income	0.6008	0.4868	2.8266***	2.8169***
	0.3824	0.3127	0.5228	0.5222
Household consumption level	1.8868***	1.4386***	−0.6482	−0.6824
	0.3521	0.3292	0.466	0.4609
R^2	0.6546	0.7090	0.9336	0.9337
Observations	270	270	270	270

Note. (1) *, **, and *** represent significance at the levels of 10%, 5%, and 1%, respectively; (2) the standard error is presented below the coefficient.

used. But this approach often backfires. Another method is to determine which model to use according to the nature of the data before establishing the model. For example, data are sampled from the population, and random effects can be used. If the data are aggregate data, such as GDP of 31 provinces and cities, there is no random sampling problem, so fixed effect is used. The choice between fixed effect model and mixed effect model is to regress the fixed effect model first, and then carry out the likelihood ratio test. If the test result is significant, it belongs to fixed effect model; otherwise, it belongs to mixed effect model. In this study, panel data of 31 provinces and cities in China are used, so fixed effect can be used for analysis. At the same time, in order to test data more comprehensively, fixed effect model and mixed effect model are chosen for comparison, to analyze the effect of Internet technology in different regions towards economic development.

Panel A of Table 4 indicates that the mixed OLS estimates indicate that the Internet penetration rate is strongly correlated with the financial industry GDP under the control of other factors, and the impact of educational level and the financial industry is positive; however, the results of the analysis after the regional fixed effects show that the level of education is not significant and is negative, indicating that the relationship between the development of local finances and the level of education is not significant, and the level of education has a weak correlation with local financial development. Panel B (per capital financial industry output) shows that the mixed OLS model indicates that the Internet penetration rate has a certain relationship with the per capital financial output value. Hypothesis 1 is correct. Similarly, the coefficient of the junior high school education level of 0.3850 relative to the high school education level is 1.1976, indicating that the role of per capital financial output will also become more pronounced with an improvement in the education level. From the panel regression analysis of regional fixed effects, it can be seen that the penetration rate of the Internet in each region has a positive impact on the per capital financial output. In addition, although Tan believes that the development of financial intermediation is strong

Table 4: Internet development level and regional financial development level. Panel A: financial industry/GDP ratio and Internet development level. Panel B: per capital financial output and Internet development level.

	Mixed OLS model			Regional fixed effect model		
	1	2	3	4	5	6
Financial industry/GDP (panel A)						
Intercept	0.0583 0.0421	0.0821** 0.0334	0.0066 0.0395	−0.1116** 0.0456	−0.1181*** 0.0348	−0.1204*** 0.0370
Internet penetration rate	0.0708*** 0.0092	0.0346*** 0.0091	0.0811*** 0.0084	0.0305*** 0.0073	0.0332*** 0.0074	0.0304*** 0.0070
Education level (average)	0.0119* 0.0069	—	—	−0.0030 0.0064	—	—
Education level (above college)	—	0.0241*** 0.0031	—	—	−0.0036 0.0023	—
Education level (high school or above)	—	—	−0.0053 0.0051	—	—	−0.0055 0.0043
Per capital GDP	−0.0385*** 0.0061	−0.0442*** 0.0052	−0.0329*** 0.0059	−0.0291*** 0.0061	−0.0274*** 0.0061	−0.0284*** 0.0060
Resident per capital disposable income	−0.0257*** 0.0098	−0.0002 0.0087	−0.0394*** 0.0098	0.0268** 0.0106	0.0255** 0.0105	0.0255** 0.0106
Household consumption level	0.0710*** 0.0096	0.0561*** 0.0087	0.0768*** 0.0098	0.0197** 0.0091	0.0194** 0.0090	0.0202** 0.0091
R^2	0.8932	0.8875	0.8645	0.9201	0.9322	0.9343
Observations	270	270	270	270	270	270
Per capital financial output (panel B)						
Intercept	7.4225*** 1.7959	6.9715*** 1.4045	5.1117*** 1.7200	−9.8075*** 2.1683	−7.4869*** 1.6563	−9.3250*** 1.7146
Internet penetration rate	1.8474*** 0.3946	0.3975*** 0.3837	2.4158*** 0.3662	1.7442*** 0.3495	1.8052*** 0.3508	1.6059*** 0.3255
Education level (average)	1.1347*** 0.2962	—	—	−0.8797*** 0.3063	—	—
Education level (above college)	—	1.1976*** 0.1294	—	—	−0.3539*** 0.1112	—
Education level (high school or above)	—	—	0.3850* 0.2204	—	—	−0.9029*** 0.1979
Per capital GDP	−0.8719*** 0.2583	−0.9853*** 0.2178	−0.6505** 0.2588	−0.8573*** 0.2896	−0.8126*** 0.2902	−0.8295*** 0.2786
Resident per capital disposable income	−1.4952*** 0.4190	−0.6199* 0.3659	−1.9383*** 0.4270	2.2381*** 0.5056	2.2571*** 0.5014	2.1082*** 0.4911
Household consumption level	2.2649*** 0.4101	1.6721*** 0.3664	2.3892*** 0.4256	−0.7806* 0.4345	−0.8772** 0.4309	−0.7420* 0.4215
R^2	0.9653	0.9443	0.9343	0.5604	0.6032	0.5843
Observations	270	270	270	270	270	270

Note. (1) *, **, and *** indicate significance at the levels of 10%, 5%, and 1%, respectively; (2) the standard error is presented below the coefficient.

for economic growth, the impact of education level, economic development level, and household consumption level on per capital financial production value was negative [54], indicating that economic development will not promote the development of the financial industry and the per capital financial output value, but also verifying that the relationship between the consumption of residents and financial investment is complementary.

Comparing the results of the mixed data model and the regional fixed model, it can be seen that the effects of disposable income and the consumption levels of residents in different regions are different, but the effect of the Internet on financial development is still significant; with an improvement in the education level, the effect of the Internet popularity rate is even more pronounced.

From the perspective of input and output, the development status of the Internet is the output of the network infrastructure investment. Therefore, on a deep level, investment in network infrastructure plays a fundamental role in promoting the development of regional finances. This paper uses the estimated value of the capital stock of the information transmission, computer service, and software industry fixed assets investment obtained in the previous sections as the capital investment of the Internet development and tests whether it has the same effect as the Internet popularization level. The results are shown in Tables 3 and 4. According to the results of columns 1 and 2 of panels A and B, the capital stock of network infrastructure and regional financial development have a significant negative relationship, but the impact coefficient is relatively small (0.0708 and

−0.0052, 0.0346 and −0.0026; 1.8474 and −0.3889, 0.3975 and −0.2675). This shows that the investment of the national economy has a competitive relationship between the industries. The result of the mixed OLS model does not take into account the regional economic development and industry characteristics. The test results of the fixed effect model in columns 3 and 4 of panels A and B indicate that the regional Internet capital stock has a significant impact on the development of the financial industry in each region. However, it is not obvious, and its coefficient is generally small relative to the impact of Internet penetration. In general, the capital stock of infrastructure has a much smaller effect on financial development than the Internet penetration rate. Therefore, it is further explained that when considering the investment in network infrastructure, one should consider the economic conditions, technical level, and network infrastructure investments of the region to obtain the maximum utility.

4.2. Internet Development Level and Regional Financial Development Structure. After nearly 20 years of development, the Chinese securities market has become the most active and representative market in the financial industry. As of 2012, the number of Chinese investors exceeded 100 million, and the number of Chinese investors has increased by nearly 5000 times in 20 years. The total number of accounts in the market has exceeded 150 million, so the securities market is also directly impacted by investors. Table 5 shows the impact of Internet penetration on various financial sectors. The securities transactions include the total amount of stock transactions, warrants, government bonds, futures, and other financial securities in the A-share and B-share stock markets.

The stock trading volume includes the stock trading in the Shanghai and Shenzhen stock markets. The data come from the China Securities and Futures Market Yearbook, published by the National Bureau of Statistics. The data are calculated using bidirectionally statistical calculation; the market value of stocks is by region. The market capitalization of listed companies, bank loans, bank deposits, and insurance income are all derived from the China Financial Yearbook.

Table 5 shows the impact of the Internet penetration rate (estimated by the two-way random effects model) on the number of securities transactions and the number of stock transactions. The impact of the Internet on the entire securities market is not significant, but it has a strong influence on the stock trading volume. The reason is that the trading volume of securities includes the total amount of securities, such as stocks, futures, options, treasury bonds, and funds, most of which are not securities but are generally traded by ordinary investors, so the penetration rate of the Internet has no significant impact on the total number of securities transactions. And stocks are the securities with the broadest securities group; therefore, the popularity of the Internet has a significant impact on its transaction volume. At the same time, compared with the coefficient of the control variable of the securities trading amount, the household consumption level has a significant, positive relationship with the resident stock trading and a negative relationship with the per capital disposable income; in addition, the levels of economic development and per capital disposable income have no significant impact on the stock trading quota. It is clear that the Internet has a significant impact on the daily financial behaviour of ordinary investors.

Regarding the specific content of other financial industries, because of its certain regional stability, the regional fixed effect model is adopted, and the dependent variable adopts one lag stage data. The results show that the impact coefficient of the Internet on the total production value of the major sectors of the financial industry is 1.0068 (<0.001 level), which is greater than the educational level and other economic factors. However, compared with the trading volume of the stock market, the influence of the Internet development level on the stock market value is not significant. The reason is that the popularity of the Internet plays a role in financial activities by broadening the financial participation channels of ordinary investors. The effects of other industries are not very effective over a short period of time.

4.3. Internet Development Level and Regional Investors. In order to better study the relationship between Internet development and investor participation, this paper selects two groups of samples for research. One group is the number of per capital accounts opened per year and the number of new per capital accounts opened per year, taking the proportion of accounts being opened in various places as the main indicator of financial participation; the other group is (Stock trading volume/Per capita disposable income) and (Securities transaction amount/Per capita disposable income), which indicates the number of stock transactions and the number of securities transactions contributed by the per capital disposable income of the unit.

The results in Table 6 show that, as the most representative indicator of investors participating in the stock market, the impact of the Internet on the number of new accounts per capital is more significant than the number of per capital accounts being opened, and the impact of the characteristics of other regions are also more significant than the number of new accounts per capital. The illustrated network popularity has a significant impact on the account opening situation in a year. However, the popularity of the network will have a negative impact on the stock market participation of general investors. A 1% increase in network popularity can reduce the number of new accounts opened by 1.47%. Hypothesis 2 is correct. Internet penetration inhibits the financial participation of regional investors. That is, the higher the level of Internet development, the smaller the participation of investors in finance.

Table 5 shows that the popularity of the Internet has a significant impact on the total volume of securities trading and stock trading. However, the last two columns of Table 6 indicate that, from the perspective of the contribution rate of per capital disposable income, the Internet penetration rate is increased by 1%, while the contribution of per capital disposable income to securities trading volume and stock

Table 5: Financial business and Internet development level.

Financial industry development	Stock market		Other financial industries				
	Securities transaction amount	Stock trading volume	Stock market value	Bank loan	Bank savings	Insurance	Gross output value
	Random two	Random two	Fixed effects	Fixed effects	Fixed effects	Fixed effects	Fixed effects
Intercept	3.0695	12.4553***	−2.0268	4.3139***	1.4567	−4.1735***	2.1978
	3.0951	2.3480	4.3409	0.7632	1.3703	0.9313	1.6412
Internet penetration rate	0.0803	0.6074***	0.8466	0.9769***	0.7468***	0.5345***	1.0068***
	0.3388	0.2273	0.7302	0.1284	0.2305	0.1567	0.2761
Education level (average)	0.6678***	0.5177***	1.3936**	0.3926***	0.2247	0.2741**	0.4893**
	0.2364	0.1556	0.5772	0.1015	0.1822	0.1238	0.2183
Per capital GDP	0.4828**	0.1384	0.9055	0.2099**	0.4597**	0.6815***	0.2971
	0.2289	0.1515	0.5662	0.0995	0.1787	0.1215	0.2141
Resident per capital disposable income	0.9134**	−0.1819	2.4646**	−0.1646	0.1207	0.5220**	0.7495*
	0.4539	0.3100	1.0343	0.1819	0.3265	0.2219	0.3910
Household consumption level	0.5110	0.8745***	−1.8895**	0.7103***	0.3886	−0.0932	−0.0617
	0.3565	0.2400	0.8802	0.1548	0.2778	0.1888	0.3328
R^2	0.2434	0.1859	0.8781	0.9926	0.9790	0.9916	0.9710
Observations	270	270	270	270	270	270	270

Note, (1) *, **, and *** represent significance at the levels of 10%, 5%, and 1%, respectively; (2) the standard error is presented below the coefficient.

Table 6: The degree of participation in the securities market and the level of Internet development.

Stock market participation willingness	Per capital account opening	Number of new accounts per capital	Securities trading volume/ per capital disposable income	Stock trading volume/ per capital disposable income
	Fixed Regional effect	Fixed Regional effect	Random two	Random two
Intercept	−3.4511*	−1.3004	1.6971	4.4866***
	2.0329	3.0475	1.3694	1.1683
Internet penetration rate	−0.7267**	−1.4707***	−0.4345**	−0.1520**
	0.3116	0.4671	0.2100	0.0760
Education level (average)	0.0832	1.6813***	0.8162***	0.6355***
	0.2509	0.3761	0.1839	0.1790
Per capital GDP	1.0243***	1.3385***	−0.3236**	−0.19053
	0.2306	0.3457	0.1549	0.1361
Per capital disposable income	0.6551	−1.7969***	—	—
	0.4444	0.6661		
Per capital property holding level	—	—	−0.4345	−0.3655
			0.2344	0.2443
Household consumption level	−0.6267*	0.6440	−0.4480**	−0.78902***
	0.3767	0.5647	0.2095	0.1652
R^2	0.9596	0.9061	0.1053	0.1298
Observations	232	232	232	232

Note. (1) *, **, and *** represent significance at levels of 10%, 5%, and 1%, respectively; (2) the standard error is presented below the coefficient.

trading volume decreased by 0.43% and 0.15%, respectively; in addition, the level of economic development, property holdings, and household consumption levels also had a negative impact. This shows that the more common the use of the Internet is and the higher the economic development level and the consumption level of residents are, the more the investors are reluctant to participate in the stock market. From the perspective of economic information dissemination, economic development has broadened investment channels, and the popularity of the network has improved the asymmetry of information; therefore, investors are no longer the only ones involved in investment in the securities market. In addition, consistent with Wang and Tian [55], education levels and property holdings have a negative impact on participation in the securities market, but the impact of property holdings is not significant.

4.4. Robustness Test. In order to further test the reliability of the Internet penetration rate, this paper conducts a robustness test. Lin et al. [56], Xianming et al. [57], and Wenqing et al. [58] all pointed out that the development of the financial industry at a provincial level has a spatially dependent effect. From the perspective of network transmission technology, the communication capability of each node of the Internet is determined not only by itself, but also

Table 7: Robustness test based on the spatial econometric model.

Robustness test	Financial industry	Per capital financial output	Per capital account opening	Securities trading/per capital disposable income	Stock trading volume/per capital disposable income
Intercept	0.0217	−5.1793**	7.5696***	−36.9791***	24.9292*
	0.0402	2.0890	1.8449	4.2087	14.1119
Internet penetration in neighboring provinces	0.0363***	1.3026***	−2.4712***	−3.8923**	−2.8614***
	0.0041	0.2118	0.6679	1.5236	0.9044
Education level (average)	−0.0075	−0.6679**	0.7729***	1.3074***	2.1835***
	0.0051	0.2640	0.2135	0.4872	0.7544
Per capital GDP	−0.0242***	−0.8442***	−0.2473	−0.7723**	2.7930***
	0.0050	0.2588	0.1635	0.3729	0.7510
Per capital disposable income	0.0200**	2.0762***	−1.8382***	—	—
	0.0092	0.4774	0.4105		
Per capital property holding level	—	—	—	4.0885*	−6.0615*
				0.9365	1.5541
Household consumption level	0.0037	−1.0500***	2.3904***	0.1964	2.2876*
	0.0077	0.4013	0.3063	0.6987	1.2026
R^2	0.9709	0.9830	0.8503	0.5155	0.9315
Observations	270	270	232	232	270

Note. (1) *, **, and *** represent significance at the levels of 10%, 5%, and 1%, respectively; (2) the standard error is presented below the coefficient.

by the communication capabilities of the entire network (Yuan and Li) [59]. At the same time, due to the popularity of the network brought about by the advancement of network technology, agglomeration effects can also occur across the network, and the level of communication in the network infrastructure is usually determined by the level of network infrastructure in the adjacent area. According to the model of spatial measurement, the penetration rate of neighboring provinces is calculated. Different from the economic indicators with a spillover nature, the Internet penetration rate of neighboring provinces is mainly determined by factors such as the province's own input and output. However, the network penetration rate of each province is partly reflected by the networks of neighboring provinces. Therefore, the neighboring network penetration rate is used as a robustness test. The adjacency relationship set up in equation (6) is used as the weight matrix of the provincial network penetration rate, where W_i is the weight matrix:

$$Y_{i,t} = \alpha_i + W_i X_{i,t} + X_{contr,t} + \varepsilon_{i,t}. \quad (6)$$

This paper selected some dependent variables and used the same estimation method as the previous one to test the robustness. The results are shown in Table 7. The influence of Internet penetration rate in neighboring provinces on the stock market participation of the financial industry and investors is generally significant. The relationship is consistent with the direction of Internet penetration in the provinces.

5. Conclusions and Recommendations

In this paper, the input and output data of network infrastructure and panel data of financial industry are used to analyze the input and output efficiency of network infrastructure and the influence of the Internet on the financial industry (the production value of the financial industry, various fields of the financial industry, and the financial participation of investors). This paper focuses on the impact of Internet investment efficiency and the Internet penetration rate of each province on the financial industry and finds that there is a significant difference between the Internet investment efficiency and the level of network development in each province. The popularization of the Internet can significantly improve the total output value and per capita output value of the financial industry. However, the popularity of the Internet will reduce the participation of investors in the securities market. The reason is that the popularity of the Internet provides a wide range of information channels for most investors, thus reducing the degree of information asymmetry. On the other hand, the development of the Internet provides investors with a more convenient platform, so it is easier for investors to access other areas of the financial industry.

The analysis of the above results shows that because of the different levels of economic development, the difference in consumption levels, and the income of residents, the starting point of network infrastructure construction is different, and the capital and technology that can be invested are different. However, by introducing advanced technologies in developed regions, with limited investment, areas that are more backward in terms of information and communication technologies can catch up with the informatization level of developed regions and directly realize leap-forward development; the advent of the Internet era has had a profound impact on the financial industry, achieved through the financial behaviour of investors and hundreds of millions of netizens. The demand of consumer groups and small and medium-sized investors has given rise to the impact of the "Internet finance" model of different forms on the traditional financial services industry. Therefore, the conclusions of this paper have significance for the integration of the traditional financial

industry, the networking of finance, and the development of Internet finance.

At present, network technologies represented by emerging technologies such as the Internet of things, cloud computing, social networks, and mobile payment have had huge impacts on the whole economy. Internet penetration, oriented to consumption preferences, may has the most significant impact on economic fluctuations. Therefore, the effect of the Internet on economic growth and the change in consumption modes caused by it are not considered in this paper, so these are potential further research directions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was funded by the Social Science Foundation of Fujian Province (FJ2018B062 and FJ2019B101); the Xiamen Science and Technology Plan Project (2018S2247); and Fujian Provincial Department of Science and Technology-Soft Science Research Plan Project (2019R0093 and 2019R0094).

References

[1] B. Zhu, S. Zhai, and J. He, "Is the development of China's financial inclusion sustainable? Evidence from a perspective of balance," *Sustainability*, vol. 10, no. 1200, 2018.

[2] C. Elgin, "Internet usage and the shadow economy: evidence from panel data," *Economic Systems*, vol. 37, no. 1, pp. 111–121, 2013.

[3] S. Sassi and M. Goaied, "Financial development, ICT diffusion and economic growth: lessons from MENA region," *Telecommunications Policy*, vol. 37, no. 4-5, pp. 252–261, 2013.

[4] H. Kim, E. Kang, E. A. Lee, and D. Broman, "A toolkit for construction of authorization service infrastructure for the internet of things," in *Proceedings of the 2nd ACM/IEEE International Conference on Internet-of-Things Design and Implementation*, vol. 3, p. 12, Pittsburgh, PA USA, April 2017.

[5] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics," *IEEE Solid-State Circuits Society Newsletter*, vol. 11, pp. 33–35, 2006.

[6] L. Zijun and C. Wenjun, "An empirical study on the relationship between internet development level and economic growth in China," *Economic Geography*, vol. 37, no. 8, pp. 108–113, 2017.

[7] S. Aithal, P. K. Paul, A. Bhuimali et al., *Emerging Internet Services Vis-à-Vis Development: A Theoretical Overview*, Social Science Electronic Publishing, Rochester, NY, USA, 2017.

[8] J. A. Schumpeter, "The creative response in economic history," *The Journal of Economic History*, vol. 7, no. 2, pp. 149–159, 1947.

[9] R. E. Lucas, "On the mechanics of economic development," *Journal of Monetary Economics*, vol. 22, no. 1, pp. 3–42, 1988.

[10] P. Spilling, "The Internet Development Process: Observations and Reflections," in *Procedings of the History of Nordic Computing 3-Third IFIP WG 9.7 Conference, HiNC*, vol. 3, Stockholm, Sweden, October 2010.

[11] Ü1 Seven and H. Yetkiner, "Financial intermediation and economic growth: does income matter?" *Economic Systems*, vol. 40, no. 1, pp. 39–58, 2016.

[12] G. M. Caporale, P. G. Howells, and A. M. Soliman, "Stock market development and economic growth: the causal linkage," *Journal of Economic Development*, vol. 29, pp. 33–50, 2004.

[13] R. P. Pradhan, M. B. Arvin, N. R. Norman, and S. K. Bele, "Economic growth and the development of telecommunications infrastructure in the G-20 countries: a panel-VAR approach," *Telecommunications Policy*, vol. 38, no. 7, pp. 634–649, 2014.

[14] M. E. Batuo, "The role of telecommunications infrastructure in the regional economic growth of Africa," *The Journal of Developing Areas*, vol. 49, no. 1, pp. 313–330, 2015.

[15] R. Levine, "Financial development and economic growth: views and agenda," *Journal of Economic Literature*, vol. 35, pp. 688–726, 1997.

[16] K. D. Frazer, *NSFNET:A Partnership for High-Speed Networking*, Merit Network, Ann Arbor, MI, USA, 1996.

[17] Z. Chen, Y. Li, Y. Wu, and J. Luo, "The transition from traditional banking to mobile internet finance: an organizational innovation perspective-a comparative study of Citibank and ICBC," *Financial Innovation*, vol. 3, no. 12, 2017.

[18] X. Hou, Z. Gao, and Q. Wang, "Internet finance development and banking market discipline: evidence from China," *Journal of Financial Stability*, vol. 22, pp. 88–100, 2016.

[19] X. Chang and S. Deng, "The development of supply chain finance in China," *International Journal of Management Excellence*, vol. 3, no. 3, pp. 475–479, 2014.

[20] K. Wang, S. Tsai, X. Du, and D. Bi, "Internet finance, green finance, and sustainability," *Multidisciplinary Digital Publishing Institute*, 2019.

[21] K. Liao, M. Yue, S. Sun et al., "An evaluation of coupling coordination between tourism and finance," *Sustainability-Basel*, vol. 102320 pages, 2018.

[22] S. Dowrick and M. Rogers, "Classical and technological convergence: beyond the Solow-Swan growth model," *Oxford Economic Papers*, vol. 54, no. 3, pp. 369–385, 2002.

[23] H. Priemus, "The network approach: Dutch spatial planning between substratum and infrastructure networks," *European Planning Studies*, vol. 15, no. 5, pp. 667–686, 2007.

[24] K. M. Vu, "ICT as a source of economic growth in the information age: empirical evidence from the 1996-2005 period," *Telecommunications Policy*, vol. 35, no. 4, pp. 357–372, 2011.

[25] R. P. Pradhan, G. Mallik, and T. P. Bagchi, "Information communication technology (ICT) infrastructure and economic growth: a causality evinced by cross-country panel data," *IIMB Management Review*, vol. 30, no. 1, pp. 91–103, 2018.

[26] A. O. Binuyo and R. A. Aregbeshola, "The impact of information and communication technology (ICT) on commercial bank performance: evidence from South Africa," *Problems and Perspectives in Management*, vol. 12, pp. 59–68, 2014.

[27] E. Tranos, "The causal effect of the internet infrastructure on the economic development of European city regions," *Spatial Economic Analysis*, vol. 7, no. 3, pp. 319–337, 2012.

[28] R. P. Pradhan, M. B. Arvin, and N. R. Norman, "The dynamics of information and communications technologies infrastructure, economic growth, and financial development:

evidence from Asian countries," *Technology in Society*, vol. 42, pp. 135–149, 2015.

[29] N. Czernich, O. Falck, T. Kretschmer, and L. Woessmann, "Broadband infrastructure and economic growth," *The Economic Journal*, vol. 121, no. 552, pp. 505–532, 2011.

[30] A. Heshmati and W. Yang, "Contribution of ICT to the Chinese economic growth," *Ratio Working Papers*, vol. 91, 2006.

[31] S. L. Liu and A. G. Hu, "Test on the externality of infrastructure in China: 1988–2007," *Economic Research Journal*, vol. 3, pp. 4–15, 2010.

[32] L. I. Li-wei and J. Feng, "Relationship between internet diffusion and economic growth: empirical research based on panel data of China's 31 provinces," *Journal of Beijing Technology and Business University (Social Sciences)*, vol. 28, pp. 120–126, 2013.

[33] Y. Z. Luo, M. J. Rui, L. J. Luo, and S. L. Zhu, "Telecommunications investment: infrastructure, network effect and regional disequilibrium," *Economic Research Journal*, vol. 6, pp. 61–72, 2008.

[34] H. Zhong, W. Zi-dong, C. Xia, and L. Ting-jie, "Impact of broadband on China's economic growth," *Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition)*, vol. 15, pp. 82–86, 2013.

[35] C. Sharon, "COBB, "redefining "offshore" in Latin America"," *Growth and Change*, vol. 40, no. 2, pp. 332–356, 2009.

[36] H. Hong, J. D. Kubik, and J. C. Stein, "Social interaction and stock-market participation," *The Journal of Finance*, vol. 59, no. 1, pp. 137–163, 2004.

[37] D. Georgarakos and G. Pasini, "Trust, sociability, and stock market participation*," *Review of Finance*, vol. 15, no. 4, pp. 693–725, 2011.

[38] C. Choi, D.-E. Rhee, and Y. Oh, "Information and capital flows revisited: the Internet as a determinant of transactions in financial assets," *Economic Modelling*, vol. 40, pp. 191–198, 2014.

[39] C. Wang and C. Tian, "Stock market participation, participation rate and determinants," *Economic Research Journal*, vol. 10, pp. 97–107, 2012.

[40] P. Aghion and P. Howitt, "*Endogenous Growth Theory*," MIT Press, Cambridge, MA, 1998.

[41] J. Benhabib and M. Spiegel, "Human capital and technology diffusion," 2005.

[42] D. W. Jorgenson, M. S. Ho, and K. J. Stiroh, "A retrospective look at the U.S. Productivity growth resurgence," *Journal of Economic Perspectives*, vol. 22, no. 1, pp. 3–24, 2008.

[43] M. H. Yi and C. Choi, "The effect of the internet on inflation: panel data evidence," *Journal of Policy Modeling*, vol. 27, no. 7, pp. 885–889, 2005.

[44] Y.-H. Noh and K. Yoo, "Internet, inequality and growth," *Journal of Policy Modeling*, vol. 30, no. 6, pp. 1005–1016, 2008.

[45] Y. F. Justin, L. X. Lin, and X. Colin, "*Some Recent Progresses on Financial Structure and Development*," Palgrave Macmillan UK, London. UK, 2012.

[46] F. Allen, "*Financial Structure*," Economic Growth and Development, Washington, DC, USA, 2018.

[47] T. T. T. Le and J. T. L. Ooi, "Financial structure of property companies and capital market development," *Journal of Property Investment & Finance*, vol. 30, no. 6, pp. 596–611, 2012.

[48] G. C. Liu, Y. Y. Liu, and C. S. Zhang, "Financial Development, Financial Structure and Income Inequality in China," *World Economy*, vol. 40, no. 9, 2017.

[49] J. Zhang, G. Y. Wu, and J. P. Zhang, "The estimation of China's provincial capital stock: 1952–2000," *Economic Research Journal*, vol. 10, pp. 35–44, 2004.

[50] G. Jin, "The estimation of China's infrastructure capital stock," *Economic Research Journal*, vol. 47, pp. 4–14, 2012.

[51] R. W. Goldsmith, "Financial structure and development as a subject for international comparative study," *Journal of Infection & Chemotherapy Official Journal of the Japan Society of Chemotherapy*, vol. 20, no. 3, pp. 143–145, 1959.

[52] R. G. King and R. Levine, "Finance and growth: Schumpeter might be right," *The Quarterly Journal of Economics*, vol. 108, no. 3, pp. 717–737, 1993.

[53] M. L. Cheng and Y. Han, "A modified Cobb-Douglas production function model and its application," *IMA Journal of Management Mathematics*, vol. 25, no. 3, pp. 353–365, 2013.

[54] R. Y. Tan, "An empirical study of the relationship between China's financial development and economic growth," *Economic Research Journal*, vol. 10, pp. 53–61, 1999.

[55] C. Wang and C. Z. Tian, "Stock market participation, participation rate and determinants," *Economic Research Journal*, vol. 10, pp. 97–107, 2012.

[56] Y. F. Lin, Z. Qi, and M. X. Liu, "Financial structure and economic growth: the case of manufacturing industry," *China Finance*, vol. 1, pp. 3–21, 2003.

[57] M. Fang, A. J. Sun, and Y. F. Cao, "Research on financial support and economic growth based on spatial model: evidence from Chinese provinces, 1998–2008," *Journal of Finance Research*, vol. 10, pp. 68–82, 2010.

[58] W. Q. Pan, Z. N. Li, and Q. Liu, "Inter-industry technology spillover effects in China: evidence from 35 industry sectors," *Economic Research Journal*, vol. 46, pp. 18–29, 2011.

[59] X. L. Yuan and S. P. Li, "Multimedia data flow transmission technology research in the network," in *Proceedings of the 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, IEEE, Nanchang, China, June 2015.

Hindawi
Scientific Programming
Volume 2020, Article ID 8814247, 13 pages
https://doi.org/10.1155/2020/8814247

Research Article

How to Evaluate the Productivity of Software Ecosystem: A Case Study in GitHub

Zhifang Liao,[1] Xiaofei Qi,[1] Yan Zhang,[2] Xiaoping Fan,[3] and Yun Zhou[3]

[1]*School of Computer Science and Engineering, Central South University, Changsha 410075, China*
[2]*School of Engineering and Built Environment, Glasgow Caledonian University, Glasgow G4 0BA, UK*
[3]*Department of Information Management, Hunan University of Finance and Economics, Changsha 410075, China*

Correspondence should be addressed to Xiaoping Fan; xpfan@csu.edu.cn

Received 23 March 2020; Accepted 4 May 2020; Published 3 August 2020

Academic Editor: Chenxi Huang

With the development of open source community, the software ecosystem has become a popular perspective in the research on software development process and environment. Software productivity is an important evaluation indicator of the software ecosystem health. A successful software ecosystem relies on long-term and stable production activities by the users, which ensures that the software ecosystem can continuously provide the value needed by users. Therefore, the measurement of software ecosystem productivity can help maintain the user development efficiency and the stability of the software ecosystem. However, there is still little literature on the productivity of open source software ecosystems. By analogy with the natural ecosystem, this paper gives the relevant definitions of software ecosystem productivity and analyzes the factors affecting the productivity of software ecosystem. Based on the factors of the ecosystem productivity and their interrelationships, this paper establishes a software ecosystem productivity model and takes the GitHub platform as an example for detailed analysis and explanation. The results show that the model can better explain the factors affecting the productivity of software ecosystems. It is helpful for the research on the measurement of the software ecosystem health and the software development efficiency.

1. Introduction

The software ecosystem (SECO) consists of a software platform, a set of internal and external developers, and a community of domain experts in service to a community of users that compose relevant solution elements to satisfy their needs [1]. With the fast development of open source software, the development model of an independent company has been replaced by the collective intelligence cooperative development model gradually. And it formed complex whole. More and more researchers are starting to study the complex whole generated from this collaborative model from the perspective of software ecosystems. At the same time, how to assess the health status of the software ecosystem has become a significant research content in the software ecosystem.

The concept of software ecosystem productivity mainly comes from the natural ecosystem, where ecosystem health refers to the stability and sustainability of an ecosystem. The health of an ecosystem can be defined by three characteristics: productivity, organizational structure, and resilience [2]. By analogy with the natural ecosystem, Manikas K et al. [3] define the software ecosystem health as the ability to maintain variables and productivity over time. They think that the actor's productivity and robustness influence the ecosystem. Jansen et al. [4] use productivity, robustness, and niche to assess the health of open source ecosystems. As an important indicator of the health of software ecosystems, the software ecosystem productivity provides evaluation criteria for the development efficiency, production activities, and health status of software ecosystems. Quantitative evaluation of the software ecosystem productivity can provide theoretical basis for managers of ecosystem-related organizations to make better decisions. Currently, the detailed concept of the software ecosystem productivity is not very clear, so it is difficult to clarify the meaning of software ecosystem

productivity and implement it on a specific open source platform. Secondly, there is a lack of general measurement models and methods for evaluating the software ecosystem productivity. It is difficult to quantify and explain the software ecosystem productivity specifically. Therefore, it is necessary to construct a general software ecosystem productivity assessment model. The purpose of the productivity assessment model is to provide reference basis for participants to choose appropriate software ecosystems for production activities.

In order to solve the above problems, this paper mainly makes three contributions. Firstly, this paper proposes the definition of the software ecosystem productivity. Based on the concept and evaluation method of the natural ecosystem productivity, the concept and influencing factors of software ecosystem productivity are determined. Then, according to the model of the natural ecosystem productivity, the quantitative model of software ecosystem productivity analysis is constructed by analyzing the influence of various factors on productivity. Finally, for detailed illustration, the typical open source software platform GitHub is selected for empirical study as an example.

The structure of this paper is as follows: Section 2 mainly introduces the related work of the software ecosystem productivity and health measurement. Section 3 introduces the research questions and dataset. Section 4 defines the productivity of software ecosystem and gives the software ecosystem productivity model. Section 5 is a case study of the productivity in GitHub software ecosystems. Section 6 verifies the productivity model. Finally, the paper is concluded in the last section.

2. Related Work

In 2003, Messerschmitt and Szyperski [5] proposed the concept of software ecosystems. They defined the software ecosystem as an online community organization that has a common interest in core software technologies. Later, many researchers defined the software ecosystem from different perspectives and backgrounds. Based on a detailed study of the existing literature in the field of the software ecosystem, in 2013, Manikas and Hansen [1] defined a software ecosystem as the interaction of a set of actors on top of a common technological platform that results in a number of software solutions or services.

In recent years, there are many different research directions in the field of software ecosystem. Researchers have studied software ecosystems from many aspects, such as software ecosystem health and software ecosystem architecture. Software ecosystem health is still a research topic that many people pay attention to. At present, there are many intelligent tools or models for assessing the health of personal [6] or natural ecosystems [2]. However, in the software ecosystem health, productivity is an important evaluation indicator. How to quantify the software ecosystem productivity is still a problem to be solved [7]. In terms of health measurement of software ecosystem, Manikas and Hansen [3] reviewed the existing literature on health status of software system, compared it with the concepts of natural ecosystem and commercial ecosystem, and defined the health status of software ecosystem as the ability of ecosystem to sustain development and maintain variability and efficiency. Gamalielsson et al. [8] used developer responsiveness as an open source ecosystem health indicator for a single project. Amorim et al. [9] proposed a conceptual framework to actively support SECO health participants in the public sphere. It is based on the business ecosystem health indicators including productivity, robustness, and niche creation defined by Iansiti and Levien [7]. The productivity in this framework mainly refers to the influence of architects on platform functions and management. In their work, Eclipse and KDE were used as examples to describe architecture practices that promote and improve ecosystem health. Compared with our work, it discussed the impact of different roles of software architects on ecosystem health. However, our work aims to measure productivity through the historical data of the software ecosystem. Berk et al. [10] proposed a SECO-SAM model for comprehensive evaluating software ecosystem strategy. This model involves the SECO biology, lifestyle, environment, and healthcare organizations. The productivity is a basic level of software ecosystem health in SECO biology, but they only discussed the SECO lifestyle and environment in the case study. Franco [11] proposed a QuESo model to measure the health of software ecosystem based on sustainability, maintenance, process maturity, network health, and resource health. This study mainly introduced the indicators about software ecosystems. These studies mostly use the productivity indicator but do not mention any approach. Jansen [4] provided a multilevel and comprehensive Open Source Ecosystem Health Operationalization (OSEHO) using three pillars, being productivity, robustness, and niche creation pillars, to assess the health of open source ecosystems. And the pillars are separated into three layers, being the theory level, the network level, and the project level. It provides insight into the indicators but does not provide a detailed method for their operation. Liao et al. [12] proposed to measure the sustainability of open source software ecosystem from the aspects of openness, stability, activity, and scalability and applied the evaluation method to Stack Overflow. And Liao et al. [13, 14] also defined the indicators affecting GitHub ecosystem health from the perspective of vitality, organizational structure, and elasticity and proposed the GitHub ecosystem health prediction method. They also forecasted the lifespan length of projects and proposed a prediction model to estimate the project lifespan in open source software ecosystems.

The software ecosystem currently lacks appropriate management theory, support tools, and solid experience. Although researchers have proposed a measurement framework for software ecosystem health, they have not quantified the model to a specific platform and have not had in-depth analysis of the specific conditions of various factors of software ecosystem health. The software ecosystem productivity plays a huge role in assessing the health of the software ecosystem and improving the productivity of the software ecosystem. This paper studies the relevant factors

affecting software ecosystem productivity and verifies it by constructing a software ecosystem productivity model.

3. Research Questions and Data Gathering

3.1. Research Questions. The purpose of this paper is to study the productivity of the open source software ecosystem. Through defining the representation of productivity in GitHub and analyzing related factors, we proposed a specific, concrete operational approach and analysis model to evaluate productivity. To achieve that, the research questions answered by the paper are as follows:

Q1 What is the definition of the productivity of open source software ecosystem? And what can explain the productivity on the GitHub platform? Based on the natural ecosystem, we defined the software ecosystem productivity including software primary productivity and software secondary productivity. Also, we compared the productivity of open source software ecosystem with the productivity of natural and business ecosystems.

Q2 What are the factors that affect productivity? What are the specific effects of these factors on primary productivity and secondary productivity? Based on the definition in *Q*1 and the calculation model of the natural ecosystem productivity, we found out the factors affecting the productivity of the ecosystem and proposed a hypothetical productivity model. It was analyzed and verified in the example study latter.

Q3 How should we measure the productivity in open source software ecosystems? Dose this hypothetical open source software ecosystem productivity model hold? To answer this question, we used an example study of GitHub to explain this approach and constructed the model in Section 5.

Q4 Can this evaluation method and productivity model be applied to ecosystems of GitHub and other platforms? In Section 6, the applicability of the productivity model on other platforms is verified with the same evaluation method.

3.2. Platform Introduction. As the largest and fastest-growing open source ecosystem in recent years, GitHub has attracted millions of developers to release open source projects. At the same time, it opened APIs to provide a convenient way for researchers to obtain the required data. Therefore, this paper focused on projects in the GitHub open source ecosystem. GitHub is an open source community that performs code changes based on pull requests. In open source projects, users can perform a variety of actions, including forking, adding stars, watching, creating issues, commenting, pulling requests, pushing, and making commits. Users can be divided into owners, core developers, and noncore developers. Owners can execute all activities and assign privileges to other developers. Core developers can directly perform submission activities on code by pushing their changes after assigning permissions. Noncore developers submit code changes by pulling requests and can only perform submission activities after core developers review. GitHub is a project hosting platform for open source and proprietary software projects. It is also the most popular open source library at present. GitHub platform saves a lot of historical data of development process. These data provide material for the study of project status. Project data is easy to obtain, and the API of GitHub website provided data crawling routes with high data integrity.

Therefore, this paper took GitHub ecosystem as an example to build a software ecosystem productivity model and verify the accuracy and universality of the model.

4. Definition and Methodology

In this section, the definition of software ecosystem productivity is provided firstly. Then, the method used in the experiment is explained in detail.

4.1. Productivity Definition. *Q*1 What is the definition of the productivity of open source software ecosystem? And what can explain the productivity on the GitHub platform?

In the natural ecosystem, biological productivity refers to the ability to produce substances at different life levels, such as individuals, groups, ecosystems, regions, and even biospheres. It determines the overall material cycle and energy flow and is also an important indicator of the health status of the system [15]. The concepts of biological productivity include primary productivity (GPP) and secondary productivity. Primary productivity in ecosystem refers to the fixed solar energy or organic matter manufactured by plants. Secondary productivity refers to the ability of consumers to metabolize the substances manufactured by primary production and stored energy and form their own substances and energy through assimilation. Additionally, in human ecology, many other energy sources, such as wind energy, are also used. There are many different affected factors and measurement methods for the different types of energy [16]. In business ecosystem, the productivity is defined as how the ecosystem effectively converts raw materials into living organisms. It is the capacity of the ecosystem to transform inputs into new products and functionalities with low cost [7]. And it emphasizes the delivery of innovations and the lower costs. The difference between the definition of productivity in natural ecosystems and business ecosystems is that natural ecosystems measure the ability to produce a product, mainly based on the total number, while business ecosystems pay more attention to efficiency and measure how to produce more products at a lower cost.

By drawing an analogy between natural ecosystem and business ecosystem, this paper gives the definition of software ecosystem productivity with more focus on the production and information. The definition of productivity in this paper differs from that in business ecosystem. It focuses on the ability of software ecosystem to produce ecological products and does not reflect the innovation of the software ecosystem. And it can be measured from activity data in the ecosystem.

Definition 1(software ecosystem total productivity). The total productivity of software ecosystem (SEP) refers to the amount of information generated by the interaction and collaboration of participants, platforms, and supporters in the ecosystem.

Taking the GitHub *platform* as an example, the user's main contributions include building code repository, committing codes, presenting issues, and commenting on different issues and commits. These behaviors produce a series of interactive information. This information brings vitality to the production activities of the ecosystem. It becomes the total production of the ecosystem.

Definition 2(software ecosystem primary productivity). The software ecosystem primary productivity (SEPP) refers to all information produced by participants that affects ecosystem products in a software ecosystem.

Different software platforms have different functions, and the information affecting ecosystem products is also different. For *example*, in the GitHub platform, ecosystem products are mainly codes. Therefore, the information affecting software ecosystem products produced by participants mainly includes committing codes, making an issue, and repairing defects.

Definition 3(software ecosystem secondary productivity). Software ecosystem secondary productivity (SESP) refers to the valuable information that the participants produce and the ecological products which directly affect the software eco-products in the software ecosystem.

This information is usually further processed by the previously generated information, such as reviewing and merging pull requests, closing issues and commits, or commenting on commits and issues in the GitHub. The commit generates a large amount of code information. After the code is approved and merged, it can be merged into the original code repository. It affects the project version and forming valuable information for the software ecosystem. Similarly, after the issue, commit, or pull requests behavior are generated, the user will comment on these issues and commit to exchange information. The information on the impact of these production activities on the production activities of users is the subproductivity of software ecosystem. In the platform-based open source software ecosystem, users are the main participants of the platform. A series of activities on the platform interact with the platform to generate information. This information promotes the normal development of the platform and becomes the energy to maintain the normal operation of the software ecosystem.

4.2. Productivity Model Hypothesis. Q2 What are the factors that affect productivity? What are the specific effects of these factors on primary productivity and secondary productivity?

The productivity index of software ecosystem can be used to evaluate the development efficiency of software projects and the health of software ecosystem. Participants carry out production activities in the ecosystem. The contribution of participants is also the main energy source of open source software ecosystems. Productivity in software ecosystem is mainly generated by the interaction between software participants and platforms. The main factors affecting the productivity of software ecosystem include platform factors and user factors. Platform factors are mainly external factors that have great impact on the software ecosystem, including project popularity and project development language. Participant factors mainly include the number of participants in the software ecosystem and the willingness of participants to contribute. This paper mainly studies whether there is a certain relationship between the activities of the participants and the productivity of the software ecosystem and whether this relationship can be quantified in a similar way to the quantitative model of natural ecosystem productivity.

In the measurement model of productivity in natural ecosystems, the main influencing factors of productivity are illumination radiation, that is, the input of energy. The calculation model of primary productivity is usually obtained by multiplying the coefficient of the factors affecting the conversion of illumination energy by the amount of effective illumination. In terrestrial ecosystems, geographic detection platforms usually use GLOPEM model algorithm [17] to retrieve primary productivity data from various satellite remote sensing data. Primary ecosystem productivity can be expressed as

$$\mathrm{GPP} = \mathrm{PAR} \times \mathrm{FPAR} \times \varepsilon. \tag{1}$$

In equation (1), *GPP* represents primary ecosystem productivity, PAR is photosynthetically active radiation, FPAR is the ratio of photosynthetically active radiation absorbed by vegetation, and ε is the actual light utilization rate based on the concept of GPP. The multiplication of PAR and FPAR is the photosynthetically effective radiation absorbed by vegetation, that is, the light utilization rate of vegetation.

This paper mainly analyzed the influence of factors of the number and activities of participants on productivity in software ecosystem. In software ecosystem, the contribution activities of participants are the main energy source of open source software ecosystem. Therefore, according to the abovementioned quantitative model of natural ecosystem productivity, the software ecosystem productivity model is similarly represented as a linear function of participants. We hypothesize a quantitative model of software ecosystem productivity as equation (2). Then, we verify it in the example study:

$$\mathrm{SEPP} = Ac * Pe + C. \tag{2}$$

In equation (2), SEPP represents the primary productivity of software ecosystems, and it is a linear function related to the number of users and the willingness to contribute. Pe is the number of participants, Ac is the introduced parameter, representing the activity factor, and C is a constant. In different natural ecosystems, the primary productivity of ecosystems will vary depending on the

effective radiation ratio of vegetation to absorb photosynthesis. Similarly, in different software ecosystems, different programming languages, project lifetimes, project followers, etc. are not exactly the same for the effective information production rate of the ecosystem. These different factors together constitute the Ac. And the default minimum productivity of a participant is 1. Thus, the value of SEPP is the number of participants when C is negative number and $Ac * Pe$ is less than C absolute value.

Similarly, according to the definition of secondary productivity, secondary productivity is transformed from primary productivity. And it is also a linear function related to the number of users and willingness to contribute. This hypothesis can be expressed as

$$\mathrm{SESP} = \mathrm{SEPP} * Cr + C1. \tag{3}$$

In equation (3), SESP is the secondary productivity of software ecosystem, Cr is the productivity conversion rate, and $C1$ is the constant. According to the primary productivity formula, the relationship model of secondary productivity can be transformed as follows:

$$\mathrm{SESP} = Ac' * Pe + C2. \tag{4}$$

In equation (4), Ac' is the conversion parameter of productivity, which is obtained by multiplying the active factor by the conversion rate. $C2$ is a constant; the value is equal to $C1$ multiplied by the conversion rate with a constant added. It can also be predicted that the secondary productivity of software ecosystem is linearly related to the number of users and their willingness to contribute.

In Sections 5 and 6, we specifically analyzed the influencing factors of software productivity for different platforms and made an empirical study on the feasibility and universality of the abovementioned hypothetical model to verify whether the models can express the impact of participants on the ecosystem productivity.

4.3. Methodology. $Q3$ How should we measure the productivity in open source software ecosystems? Dose this hypothetical open source software ecosystem productivity model hold?

The research method analyzed the factors affecting the productivity. Then, the productivity model of software ecosystem was verified and established. It was divided into three steps.

Step 1 (data collection): firstly, the appropriate data were selected and the ecosystem was divided. In order to verify the feasibility of the model, the typical open source software platform GitHub was selected for empirical research. We took GitHub platform as an example in Section 4. Because project popularity and development language are important platform factors affecting ecosystem, in order to eliminate the impact of these factors in the ecosystem, we divided the ecosystem with different types of platforms, mainly using development language as the index. For platforms that cannot divide ecosystem according to the development language, we adopt other dividing standards such as project type and project popularity.

Step 2 (correlation analysis): in order to analyze the impact of user activities on software ecosystem productivity, the relationship between productivity and ecosystem participants in software ecosystem was analyzed. The data of productivity and participants in the ecosystem were analyzed in a month-long observation period, and the relationship between different types of productivity in the ecosystem and those produced by people in a unit time were analyzed. The data used in this paper are spaced monthly and the data variables are equidistant, the person formula is applicable to measure the coefficient of linear relationship between the fixed distance variables, and the data scale is suitable for the calculation of the person correlation coefficient, so this paper used the person correlation coefficient to analyze the correlation between the participant data and the software productivity data:

$$r_{\mathrm{xy}} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\left(\sqrt{\sum_{i=0}^{n} (x_i - \overline{x})^2}\right)\left(\sqrt{\sum_{i=0}^{n} (y_i - \overline{y})^2}\right)}. \tag{5}$$

In equation (5), i is the first month of time, x is the number of software ecosystem participants, and y is productivity of this software ecosystem. r_{xy} represents the correlation coefficient between productivity and participants, $r_{\mathrm{xy}} \in (-1, 1)$; r_{xy} is larger, which indicates that the software ecosystem productivity is more positively correlated with the participants. The negative value means that the software ecosystem productivity is negatively correlated with the participants. And r_{xy} tends to 0, which means that there is no correlation between them.

Step 3 (model construction): the construction of this model was divided into two parts. The first is regression analysis. According to the prototype of the software ecosystem productivity model, the linear relationship between productivity and participants was judged. Under the condition of linear relationship, the initial regression equation between productivity and participants was obtained by the least square method [18]. Because the productivity created by different types of producers is different in the same ecosystem, the specific ecosystem presents different regression models because of the different activity of participants.

The second is the construction of the real model. The regression equation of each ecosystem is inconsistent. To determine a regression model applicable to most projects, this study used truth discovery methods [19]. Truth discovery is a method to measure the reliability of multisource information and estimate the real information. The flow of this algorithm was shown in Algorithm 1. By this method, we calculated the reliability of the regression equation based on each project and obtained a general regression equation with a greater accuracy.

Input: Data from n project: $\{Ac_1, \ldots\ldots, Ac_n\}$, $\{c_1, \ldots\ldots, c_n\}$
Output: Truths $Ac^{*(t)}$, $c^{*(t)}$
(1) Initialize the truths
$Ac^{*(1)} = \sum_{i=1}^{n} Ac_i/n$
$c^{*(1)} = \sum_{i=1}^{n} c_i/n$
(2) **repeat**
(3) for $i \leftarrow 1$ to N do
(4) for $i \leftarrow 1$ to T do
(5) $w_i = \log(\sum_{i=1}^{n}((Ac^{*(t-1)} - Ac_i)^2) + ((c^{*(t-1)} - c_i)^2)/(Ac^{*(t-1)} - Ac_i)^2 + (c^{*(t-1)} - c_i)^2)$
(6) $Ac^{*(t)} = \sum_{i=1}^{n}(w_i * lip_i)/\sum_{i=1}^{n} w_i$
(7) $c^{*(t)} = \sum_{i=1}^{n}(w_i * c_i)/\sum_{i=1}^{n} w_i$
(8) end for
(9) end for
(10) **until** Convergence criterion is satisfied;
(11) return $Ac^{*(t)}$, $c^{*(t)}$

Algorithm 1: Truth discovery.

The specific algorithm was described as follows. Firstly, the initial settings of $\{Ac_1, \ldots\ldots, Ac_n\}$, $\{c_1, \ldots\ldots, c_n\}$ are determined by the regression equation of each item and sorted. The results of the first iteration are the average value of $Ac^{*(1)}$ and $c^{*(1)}$. Secondly, the weight (w_i) of each item in the overall ecosystem is calculated, where i represents the ecosystem number i, c_i represents the activity of the ecosystem i obtained through the regression equation, $Ac^{*(t)}$ represents the activity result of the t-time iteration, and $c^{*(t)}$ represents the constant result of the t-time iteration. Finally, the results of the t-time iteration are calculated.

In this paper, we used the truth discovery algorithm to get the real software ecosystem productivity quantification model under a specific software ecological platform. The specific experimental process was described in detail in Section 5.

5. Case Study

This paper focused on the analysis of user factors influence on ecosystem productivity. This section verified the relationship between the number of users, activities, and productivity. In this section, we took GitHub as an example to quantify the software ecosystem productivity model and conduct empirical research.

5.1. Data Collection. The first task was data collection and preprocessing. The main external factors affecting ecosystem productivity include the function type, the popularity of the project, and the popularity of the development language. Based on the type of development language, this paper chose seven of the most popular languages and divided GitHub platform into several technological ecosystems to analyze the productivity model.

In order to scientifically compare the popular languages of each platform, we ranked the top 10 languages in the four platforms of Stack Overflow, GitHub, TIOBE, and IEEE. As shown in Table 1, we reversed the popularity of these languages. The first language in each platform had 10 points, the second had 9 points, and so on. The scores of the four platforms were added, and we selected the top 7 languages in terms of popularity. Then, technical ecosystems of the platform were divided by language as the main factor, and the ecosystem productivity was analyzed. The final scores were JavaScript, Java, C, C++, C#, PHP, and Python, the seven different development languages for analysis.

Table 1: Comparison of language popularity on different platforms.

Score	Stack Overflow	GitHub	TIOBE	IEEE
10	JavaScript	JavaScript	Java	Python
9	Java	Python	C	C++
8	C#	Java	C++	C
7	PHP	Ruby	Python	Java
6	Python	PHP	VB.net	C#
5	Html	C++	C#	PHP
4	C++	css	PHP	R
3	css	C#	JavaScript	JavaScript
2	SQL	Go	SQL	Go
1	ASP.NET	C	Swift	Ass

Using the seven popular languages mentioned above, the platform was divided into software ecosystems of different languages. According to the definition of software ecosystem productivity in this paper, the data of user contribution activities were used. GitHub platform saves a lot of historical data of development process, and the API of the website provides high data integrity for data crawling. As shown in Table 2, this paper used several projects with the highest attention in GitHub and collected 70 projects with the largest number of stars in JavaScript, Java, C, C++, C#, PHP, and Python.

5.2. Correlation Analysis. After data processing was completed, the correlation between productivity and participants was analyzed. This paper attempted to find out the relationship between the number of participants and productivity through the statistical analysis of average user activities.

Table 2: Projects of the largest number of stars in different languages on GitHub.

JavaScript	Java	C++	C#	C	Python	PHP
Vue	Java-design-patterns	TensorFlow	Shadowsocks-windows	Linux	Awesome-python	Laravel
React	RxJava	Electron	CodeHub	Netdata	System-design-primer	Symfony
D3	Elasticsearch	Swift	CoreFX	Redis	Public-apis	Faker
JavaScript	Spring-boot	Bitcoin	PowerShell	Git	Models	Composer
React-native	Retrofit	NW.js	Wox	Ijkplayer	Youtube-dl	CodeIgniter
Angular.js	Interviews	x64dbg	CoreCLR	Php-src	Flask	DesignPatternsPHP
Font-Awesome	OkHttp	Protobuf	Roslyn	Wrk	Thefuck	SecLists
Create-react-app	Guava	OpenCV	Dapper	How-to-Make-a-Computer-Operating-System	Httpie	Framework
Node	MPAndroidChart	Caffe	WaveFunctionCollapse	the_silver_searcher	Django	Guzzle

In a specific software ecosystem, it is necessary to select data that can represent the ecosystem productivity and make correlation analysis with the number of users. In GitHub ecosystem, the most important contribution of users usually comes from the pull requests behavior, so PR data is a good representation of the software ecosystem productivity, while valuable PR is usually merged into the project code base. So, the merged PR was used as the secondary productivity after transformation. The correlation between the productivity and the ecological participants in software ecosystem projects was analyzed.

In this paper, the productivity data of each project in unit time and the data of participants in statistical time threshold were counted according to the observation period, and the correlation coefficient of person was used to analyze the correlation between participant data and software productivity data. The results are shown in Figure 1.

As shown in Figure 1, in the GitHub platform, a highly positive correlation can be found between the total PR and the total number of participants. Among them, the lowest degree of association of a development language is 0.922, while the highest level is JavaScript ecology, up to 0.986. The number of PR that has been merged and the number of participants are also highly positively correlated, and the degree of association ranges from 0.811 to 0.967. Therefore, it can be concluded that, in GitHub ecosystem, there is a highly positive correlation between the productivity of software ecosystem and the number of participants in ecosystem, and the number of participants directly affects the productivity.

Based on the above analysis of the correlation between productivity and the number of users, it was concluded that there is a highly positive correlation between the productivity of software ecosystem and the number of participants in the ecosystem. The number of participants directly affects the value of productivity. Because productivity comes from the interaction of all participants in the ecosystem, the main influencing factors of software productivity are the number and activity of producers in the ecosystem. Therefore, this paper used the number of users and the activity of users to build a model of the impact of participants on productivity.

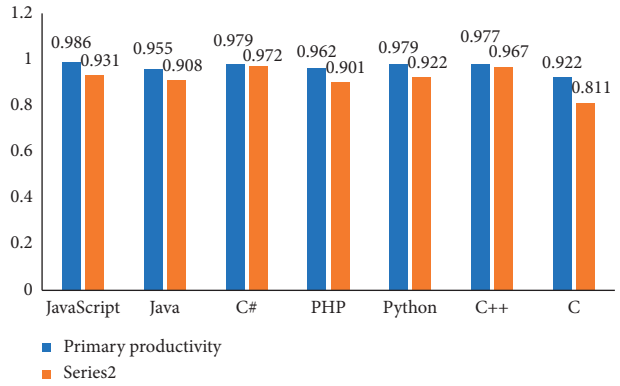


Figure 1: The correlation between the number of participants and productivity on the GitHub platform. The blue column represents the correlation between primary productivity and the number of participants, while the orange column represents the relationship between secondary productivity and the number of participants.

5.3. Model Construction. According to the above correlation analysis, it was found that there is a clear positive correlation between ecosystem productivity and the number of participants. Therefore, it was necessary to judge whether productivity and the number of participants have linear function relations and carry out regression analysis. On the GitHub platform, the impact model of participants on productivity was constructed by using seven representative languages. In Figure 2, we found that there is a linear relationship between productivity and the number of participants. Specific projects or ecosystems present different regression models depending on the activity of participants.

As shown in Figure 2, in every ecosystem, productivity is a linear function related to the number of participants. Primary and secondary productivity models can be met in both primary and secondary productivity. Therefore, the

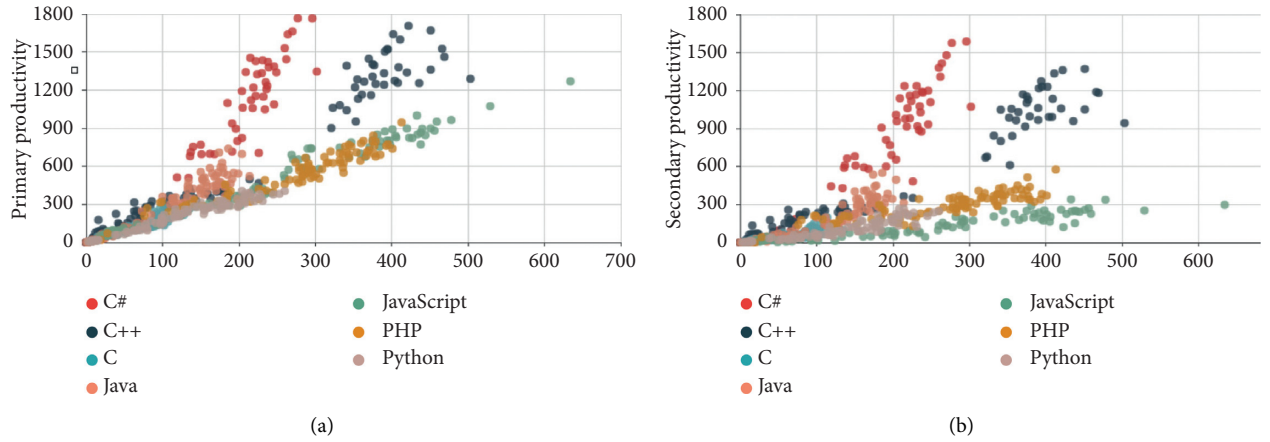


Figure 2: Correlation analysis between productivity and participant. (a) The correspondence between the number of participants and the primary productivity in each observation period on the GitHub platform. The dots of different colors represent the correspondence between the participants of different languages and productivity. (b) The correspondence between the number of participants and the secondary productivity.

Table 3: The linear function of productivity related to the number of participants.

Languages	Equation of primary productivity	Equation of secondary productivity	Ac	C1	Ac*Cr	C2
C#	$y = 5.46x - 31.985$	$y = 4.66x - 29.407$	5.46	31.985	4.665	29.407
C++	$y = 3.37x - 24.101$	$y = 2.69x - 31.796$	3.37	24.101	2.690	31.796
C	$y = 2.05x - 9.555$	$y = 0.97x - 6.834$	2.05	9.555	0.967	6.834
Java	$y = 2.94x - 13.341$	$y = 1.97x - 29.100$	2.94	13.341	1.974	29.100
JavaScript	$y = 2.05x - 5.639$	$y = 0.56x - 10.923$	2.05	5.639	0.557	10.923
PHP	$y = 1.93x + 11.439$	$y = 0.99x + 23.920$	1.93	11.439	0.987	23.920
Python	$y = 1.65x + 2.464$	$y = 0.94x - 19.103$	1.65	2.464	0.941	19.103

primary regression equation was obtained by the least square method, as shown in Table 3.

Since the regression equations of each project are inconsistent, the lowest activity factor, Ac, was 1.65 and the highest was 5.46. The constant C1 ranged from −31.99 to 11.44. Through analysis, it was found that the active factor can usually indicate the willingness of users in the ecosystem. To obtain a regression model for most projects, the participant's impact model on productivity was constructed using the real-discovery approach described in Section 3:

$$\text{SEPP} = 2.22 * Pe - 9.44, \tag{6}$$

$$\text{SESP} = 1.20 * Pe - 21.11. \tag{7}$$

In equation (6), SEPP is the primary productivity of GitHub ecosystem, and in equation (7), SESP is the secondary productivity. Pe represents the number of participants in the software ecosystem. The primary activity was 2.22, the secondary activity was 1.20, the constant C1 of primary productivity was −9.44, and the constant C2 of secondary productivity was −21.11. And in this paper, the default minimum productivity of a user was 1. The value of SEPP and SESP is the number of users when C is a negative number and Ac*Cr is less than C absolute value.

Through the analysis of the software ecosystem productivity of GitHub platform, it was found that the software productivity model proposed in Section 4 can be applied to multiple software ecosystems. Because of the different functions of the specific ecosystem and the different user groups, the user's active degree will be different. Therefore, when using the model, we need to analyze the productivity of the software ecosystem according to the user's active degree on different platforms. Usually, the average activity of users can be used to replace the active factors derived from the inversion of productivity and the number of participants.

6. Verification Analysis

Q4 Can this evaluation method and productivity model be applied to ecosystems of GitHub and other platforms?

To answer this question, we verified the model in other ecosystems in this section, which include three different class ecosystems in GitHub. Then, the ecosystems in Stack Overflow and Bugzilla were verified.

6.1. Verification in Other Ecosystems of GitHub. In this section, we verified the model in the three different class ecosystems in GitHub. In biology, the range of ecosystems can be large or small, and ecosystem productivity can represent the production capacity of individuals, groups,

ecosystems, regions, and even biosphere. Similarly, this paper selected three software ecosystems of different sizes and types in GitHub for verification. They are the single software product ecosystems, the software development team ecosystems, and the language ecosystems. The relationship between model productivity and actual productivity was verified by statistical calculation.

Three different ecosystems, Moby, GitBook, and Ruby, were selected to validate the model. Moby is an open source project dedicated to promoting the movement of software containerization. In the Moby project, users and software frameworks, components, and other software products gather to form a software ecosystem. The GitBook team is mainly a development team for text editors using Git technology. In the GitBook ecosystem, users and software development environments, software products, services, and others are condensed together through a team to form a software ecosystem. Ruby is a simple and fast object-oriented scripting language. It is a popular project development language in GitHub. In the Ruby ecosystem, users, software products, and development environment form a software ecosystem with the same development language.

Figure 3 shows that the software ecosystem primary productivity was calculated by the productivity model. It was consistent with the trend of the actual productivity of the software ecosystem, and the quantity is roughly the same. And through correlation analysis, the correlations between model productivity and measured productivity in the single software product ecosystem, the software team ecosystem, and the language ecosystem were higher than 90%. However, there was a difference between the predicted and the measured value. Because this model analyzed the characteristics of multiple ecosystems, the large dataset obscured the characteristics of specific ecosystems. It caused the predicted value to be different from the average productivity of the platform. The predicted value is greater than the measured value; this means that the productivity of this ecosystem is lower than the average productivity of this platform, and user activities should be improved. The predicted value is less than the measured value; this means that the users of this ecosystem are more active, and the productivity is higher than the average productivity of the platform. When the productivity of the ecosystem is continuously higher than the prediction model, the model should be adjusted according to the characteristics of this software ecosystem to ensure its accuracy. This fully proved that the abovementioned composition model of software ecosystem primary productivity is applicable to GitHub and other software ecosystems. It was also found that the number of users is limited by the environmental capacity of the ecosystem during the stable operation of the platform. In the software ecosystem, the transition from primary productivity to secondary productivity takes time. It takes some collaboration with other users to convert primary productivity to secondary productivity. Hence, the prediction of secondary productivity was not very good in the last few months. Also, the prediction of secondary productivity of these three ecosystems in GitHub is analyzed. The result is shown in Figure 4.

During the verification process, we found the impact of factors other than the number of users and activity on the software ecosystem productivity. On the GitHub platform, the software primary productivity problem had a high correlation with the number of participants, but when the ecosystem is small and the data volume is sparse, the accuracy of the model will be greatly reduced. During the productivity verification experiment of a single project team, it was found that each user participating in the PR submission would have one or two problems, but a few core developers would generate a large number of submissions during certain observation periods, resulting in partial errors in the model. The reason why the secondary productivity and the number of participants are lower than the primary productivity was that the large number of submissions generated by these few core developers is often incorporated into the code base, so the primary productivity depends more on the activities of the core developers. In the initial phase of the project, the contribution rate of core users is usually high. However, as the project progresses, the number of noncore developers participating in the project will increase, and the proportion of secondary productivity converted from primary productivity will gradually increase. Therefore, in the next step of the work, the impact of the core developer's user activities will be considered.

In the experiment, we also found that, in the process of stable operation of the platform, there is no sudden effect of external force. And after the number of users reaches a certain level, it will remain in a range for a long time. Therefore, the most important thing for primary productivity is to improve the participants and active level. And in these open source software ecosystems, the transition from primary productivity to secondary productivity takes time and needs to be discovered and collaboratively completed by other users to drive primary productivity into secondary productivity. Therefore, the amount of secondary productivity is not very good in the last few months. And because the ecological secondary productivity is converted from primary productivity, the secondary productivity is almost zero when the primary productivity is low, so the secondary productivity model is too dependent on the core user for the product of a small project team. Therefore, in order to increase secondary productivity, it is necessary to increase primary productivity, conversion factor, and number of core users.

6.2. Verification in Other Ecosystems of Other Platforms. To verify the universal applicability of this model, we also verified the model in the ecosystems of other ecosystems. Using the same method, the software ecosystem productivity of Stack Overflow platform and Bugzilla platform was obtained. Stack Overflow is a standard Q&A website on computer science and programming topics. On the Stack Overflow platform, users can perform a variety of different activities such as questions, answers, votes, and comments. Users participate in group intelligence collaborative activities such as questions and answers to form an open source ecosystem. It has been highly popular with software

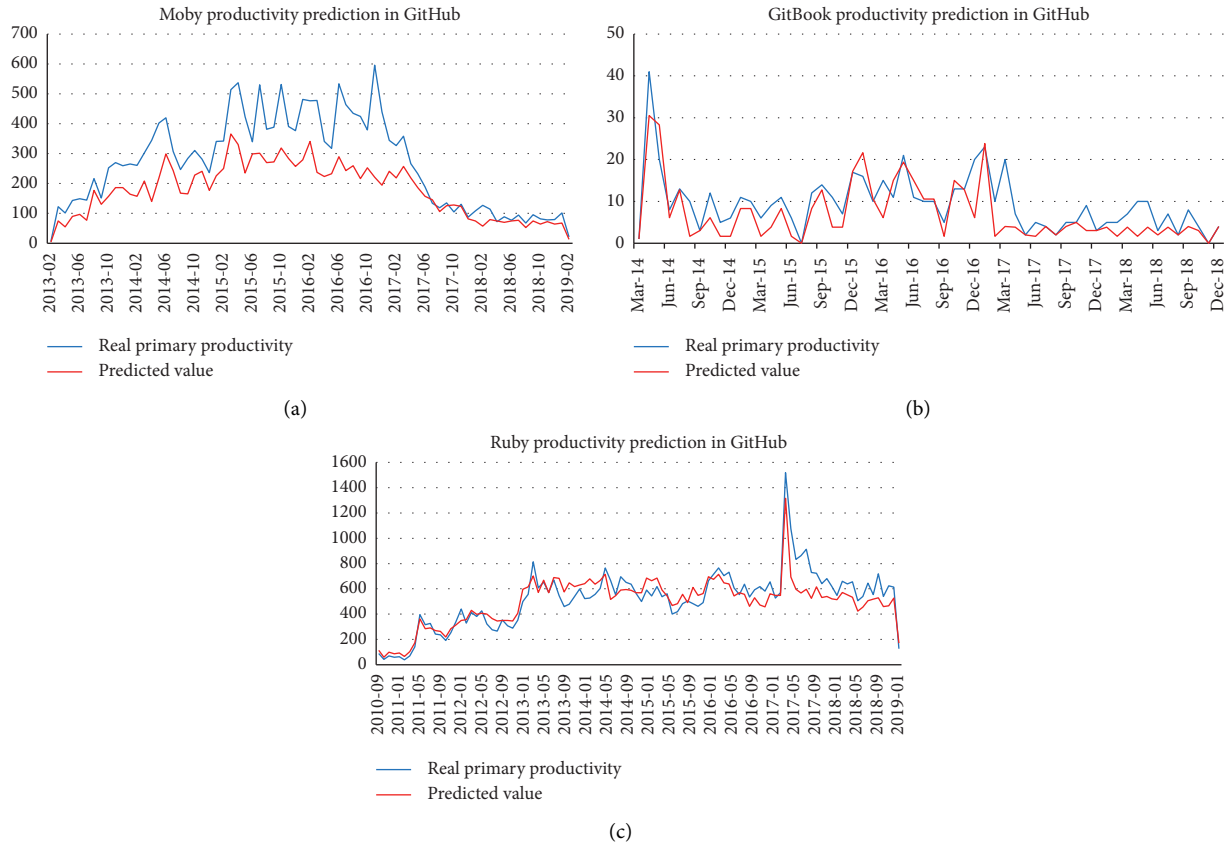


Figure 3: Model predictive value vs. real primary productivity. (a) The Moby project primary productivity predictions compared to real primary productivity. (b) The GitBook team's primary productivity predictions compared to real primary productivity. (c) The Ruby language ecosystem primary productivity predictions compared to real primary productivity. The red line indicates the predicted value. The blue line represents the real value. The abscissa axis is the different month and the ordinate axis is the productivity of the ecosystem.

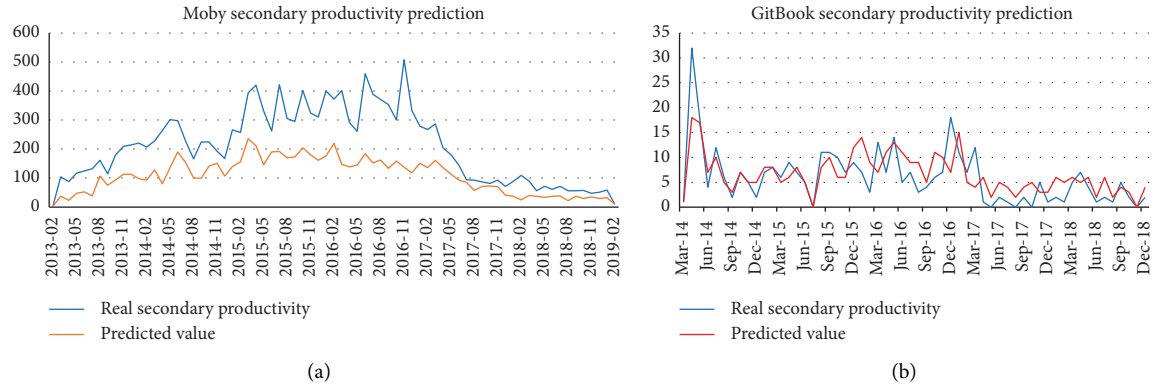
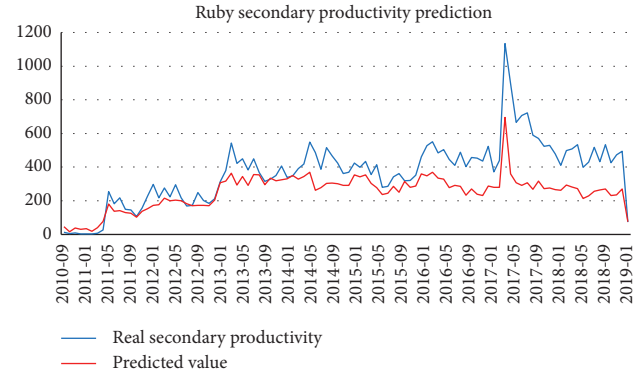


Figure 4: Continued.



(c)

Figure 4: Model predictive value vs. real secondary productivity. (a) The Moby project secondary productivity predictions compared to real secondary productivity. (b) The GitBook team's secondary productivity predictions compared to real secondary productivity. (c) The Ruby language ecosystem secondary productivity predictions compared to real secondary productivity.

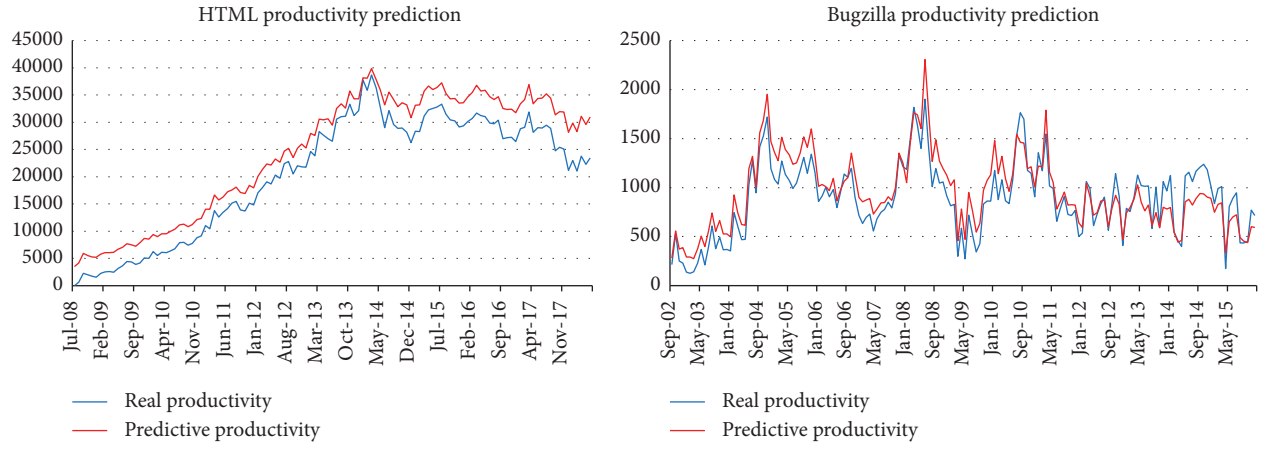


Figure 5: Model predictive value vs. real productivity. (a) The productivity predictions compared to real productivity on Stack Overflow. (b) The productivity predictions compared to real productivity on Bugzilla of Mozilla projects.

developers and is considered to be one of the most successful open source ecosystems. Bugzilla is a defect tracking system developed by the Mozilla Foundation (http://www.mozilla.org). Many users, including open source projects (Apache, Open Office for Linux), private projects, and public agencies (NASA, IBM) are using Bugzilla. In Bugzilla, the user's contribution mainly starts from the user's submission of a bug report, which is passed to the defect repair, inspection activity behavior, etc. Therefore, we selected these software ecosystems to verify the productivity model.

Before verification, the ecosystems and their productivity data in these platforms were determined. In the Stack Overflow, the information is in questions and answers. Therefore, the number of questions and answers represent productivity. In the Bugzilla, the user's contribution starts from the bug report to the bug repair. Therefore, the productivity is represented by the number of reports. Then, the ecosystems on the platform were divided. According to the characteristics of the platform, users in Stack Overflow usually gather with different technical fields. Therefore, Stack Overflow was divided according to the programming language. In Bugzilla, there are five types of projects, so the ecosystems were divided in terms of project types.

Then, we determined the activity factor of these ecosystem models according to the method in Section 4.3. First, the correlation between productivity and the number of participants was analyzed. It was found that, on the Stack Overflow, the productivity is highly positively correlated with the number of participants. The lowest correlation was 0.88 and the highest was 0.995. On the Bugzilla, there was also a positive correlation between productivity and participants with a maximum of 0.94 and a minimum of 0.65. Then, according to the method in 4.3, regression analysis and truth discovery were used to adjust the productivity model parameters of the corresponding platform. Finally, the model was verified. On the Stack Overflow platform, the ecosystem of HTML language was selected to verify the model. On the Bugzilla platform, we chose the ecosystem of

the single software product Firefox for verification. The result was shown in Figure 5.

The correlation between the model productivity and real productivity of the two platforms was well over 90%. This result verified that the proposed model of software ecosystem productivity is applicable to other platforms. It was proved that the model is able to represent the software ecosystem productivity. And it was also found that, in the software ecosystem, the transition from primary productivity to secondary productivity takes time. It took some collaboration with other users to convert primary productivity to secondary productivity. Therefore, the prediction of secondary productivity was not very good in the first and last few months.

7. Discussion and Conclusions

Based on the definition of natural ecosystem productivity, this paper proposed a definition of software ecosystem productivity. Analogous to the primary and secondary productivity of natural ecosystems, this paper decomposed software ecosystem productivity into primary and secondary productivity according to the different impacts of user activities on ecosystems. Also, the productivity was quantified in different software ecosystems. According to the source of productivity, this paper put forward the view that the number, scale, and activity of participants play the most direct and important role in ecosystem productivity. In the process of verifying the relationship between the number and activity of participants and the productivity of software ecosystem, we found that primary productivity is highly correlated with the number of participants, while secondary productivity was less correlated with the number of participants. This paper presented a model of software ecosystem productivity. The validity of the model was verified by experiments.

In this paper, the relevant factors of software ecosystem productivity and the composition model were studied, hoping to provide help for further software ecosystem research, software ecosystem health measurement, development efficiency measurement, and so on. This paper researched the impact of indirect factors such as user reviews and external environmental factors on software ecosystem productivity. However, this paper analyzed the number and activity of users as the factors influencing software ecosystem productivity, which is not comprehensive enough.

Some threats affect the validity of the measurement methods proposed in this paper, including external threats and internal threats. The main external threat is the validity of datasets. In this work, parameters of this model were obtained through data analysis. Therefore, the quality of this model is determined by the quality of the used dataset. In this paper, the verification datasets of Stack Overflow and Bugzilla were dumped from their platforms, and the GitHub datasets were obtained through the data service GitHub API V3. The small ecosystem and little data will lead to inaccurate acquisition of parameters in the model, which will threaten the validity of the conclusion. The internal threat is that the model only considers the number of users and the impact of user activities on productivity. However, there may be other factors affecting the accuracy of the model. Due to the characteristics of the large amount of data used in building the model, the characteristics of some specific projects in the ecosystem may have been ignored. And this model considers the productivity of the software ecosystem from a quantitative rather than qualitative perspective, ignoring the impact of different events on the ecosystem. For example, in the project-level ecosystem, some core developers frequently submit code during the observation period. It will threaten the validity of the conclusion.

In future work, we plan to validate the adaptability and reliability of the model in other open source websites and to analyze and verify other factors affecting software ecosystem production, such as the language of projects and life length of projects. It is clear that language affects the willingness of users to contribute to the ecosystem. Also, we consider the use of neural network to improve the model, especially robust multilayer extreme learning machine [20] and plan to use a visualization method [21] and develop a tool to show the impact of various factors in the ecosystem on its productivity and the evolution of the open source software ecosystem.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Ministry of Science and Technology: Key Research and Development Project (2018YFB003800), Hunan Provincial Key Laboratory of Finance and Economics Big Data Science and Technology (Hunan University of Finance and Economics) (2017TP1025), China NSF (61876190 and HNNSF 2018JJ2535), and Fundamental Research Funds for the Central Universities of Central South University (2019zzts150).

References

[1] K. Manikas and K. M. Hansen, "Software ecosystems - a systematic literature review," *Journal of Systems and Software*, vol. 86, no. 5, pp. 1294–1306, 2013.

[2] K. Ma, H. Kong, M. Guan, and B. Fu, "Ecosystem health assessment: methods and directions," *Acta Ecologica Sinica*, vol. 21, no. 12, pp. 2106–2116, 2001.

[3] K. Manikas and K. M. Hansen, "Reviewing the health of software ecosystems-a conceptual framework proposal," in *Proceedings of the Fifth International Workshop on Software*

Ecosystem (IWSECO-2013), vol. 987, pp. 26–37, Potsdam, Germany, May 2013.

[4] S. Jansen, "Measuring the health of open source software ecosystems: beyond the scope of project health," *Information and Software Technology*, vol. 56, no. 11, pp. 1508–1519, 2014.

[5] D. G. Messerschmitt and C. Szyperski, *Software Ecosystem: Understanding an Indispensable Technology and Industry*, MIT Press Books, Cambridge, MA, USA, 2003.

[6] M. Chen, Y. Li, X. Luo, W. Wang, L. Wang, and W. Zhao, "A novel human activity recognition scheme for smart health using multilayer extreme learning machine," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1410–1418, 2019.

[7] M. Iansiti and R. Levien, "Keystones and dominators: framing operating and technology strategy in a business ecosystem," *Harvard Business School Working Paper*, vol. 03-061, 2004.

[8] J. Gamalielsson, B. Lundell, and B. Lings, "Responsiveness as a measure for assessing the health of oss ecosystems," in *Proceedings Of the 2nd International Workshop On Building Sustainable Open Source Communities (OSCOMM 2010)*, IFIP, Notre Dame, IN, USA, June 2010.

[9] S. Amorim, J. D. McGregor, E. S. Almeida, C. Flach, and G. Chavez, "The architect's role in software ecosystems health," in *Proceedings Of the 2nd Workshop On Social, Human, and Economic Aspects Of Software (WASHES'17)*, DBLP, Salvador, Brazil, May 2017.

[10] V. Berk, I. V. Den, S. Jansen, and L. Luinenburg, "Software ecosystems: a software ecosystem strategy assessment model," in *Proceedings Of the 4th Software Architecture, European Conference (Ecsa), Companion Volume DBLP*, DBLP, Copenhagen, Denmark, August, 2010.

[11] O. Franco-Bedoya, D. Ameller, D. Costal, and X. Franch, "QuESo a quality model for open source software ecosystems," in *Proceedings Of the 9th International Conference On Software Engineering and Applications (ICSOFT-EA)*, pp. 209–221, Vienna, Austria, August 2014.

[12] Z. Liao, L. Deng, X. Fan et al., "Empirical research on the evaluation model and method of sustainability of the open source ecosystem," *Symmetry*, vol. 10, no. 12, 2018.

[13] Z. Liao, M. Yi, Y. Wang et al., "Healthy or not: a way to predict ecosystem health in GitHub," *Symmetry*, vol. 11, no. 2, p. 144, 2019.

[14] Z. Liao, B. Zhao, S. Liu et al., "A prediction model of the project life-span in open source software ecosystem," *Mobile Network Application*, vol. 24, no. 4, pp. 1382–1391, 2019.

[15] F. Jing and C. A. Ping, "Implications and estimations of four terrestrial productivity parameters," *Acta Phytoecologica Sinica*, vol. 25, no. 4, pp. 414–419, 2001.

[16] X. Luo, J. Sun, L. Wang et al., "Short-term wind speed forecasting via stacked extreme learning machine with generalized correntropy," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4963–4971, 2018.

[17] J. B. Wang, J. Y. Liu, and Q. Q. Shao, "Spatial-Temporal patterns of net primary productivity for 1988-2004 based on Glopem-Cevsa model in the 'Three-River Headwaters' region of Qinghai province, China," *Chinese Journal of Plant Ecology*, vol. 33, no. 2, pp. 254–269, 2009.

[18] X. Zhang, X. Pan, and M. Lu, "Meshless weighted least-square method," *Acta Mechanica Sinica*, vol. 17, no. 3, pp. 270–282, 2003.

[19] Q. Li, Y. Li, J. Gao et al., "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in *Proceedings Of the 2014 ACM SIGMOD International Conference On Management Of Data*, pp. 1187–1198, Snowbird, UT, USA, June 2014.

[20] X. Luo, Y. Li, W. Wang, X. Ban, J.-H. Wang, and W. Zhao, "A robust multilayer extreme learning machine using kernel risk-sensitive loss criterion," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 1, pp. 197–216, 2020.

[21] Z. Liao, D. He, Z. Chen, X. Fan, Y. Zhang, and S. Liu, "Exploring the characteristics of issue-related behaviors in GitHub using visualization techniques," *IEEE Access*, vol. 6, pp. 24003–24015, 2018.

Hindawi
Scientific Programming
Volume 2020, Article ID 8849011, 13 pages
https://doi.org/10.1155/2020/8849011

Research Article

Mushroom Toxicity Recognition Based on Multigrained Cascade Forest

Yingying Wang,[1,2,3] Jixiang Du,[1,2,3] Hongbo Zhang,[1,2,3] and Xiuhong Yang[1,2,3]

[1]*Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China*
[2]*Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361021, China*
[3]*Department of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*

Correspondence should be addressed to Jixiang Du; jxdu77@gmail.com

Received 24 April 2020; Revised 7 July 2020; Accepted 11 July 2020; Published 1 August 2020

Academic Editor: Chenxi Huang

Due to the tastiness of mushroom, this edible fungus often appears in people's daily meals. Nevertheless, there are still various mushroom species that have not been identified. Thus, the automatic identification of mushroom toxicity is of great value. A number of methods are commonly employed to recognize mushroom toxicity, such as folk experience, chemical testing, animal experiments, and fungal classification, all of which cannot produce quick, accurate results and have a complicated cycle. To solve these problems, in this paper, we proposed an automatic toxicity identification method based on visual features. The proposed method regards toxicity identification as a binary classification problem. First, intuitive and easily accessible appearance data, such as the cap shape and color of mushrooms, were taken as features. Second, the missing data in any of the features were handled in two ways. Finally, three pattern-recognition methods, including logistic regression, support vector machine, and multigrained cascade forest, were used to construct 3 different toxicity classifiers for mushrooms. Compared with the logistic regression and support vector machine classifiers, the multigrained cascade forest classifier had better performance with an accuracy of approximately 98%, enhancing the possibility of preventing food poisoning. These classifiers can recognize the toxicity of mushrooms—even that of some unknown species—according to their appearance features and important social and application value.

1. Introduction

Mushrooms are the fleshy fruiting bodies of certain fungus, some of which are edible, but a minority of them are toxic [1]. Every year, a large number of people die [2, 3] from eating poisonous mushrooms. It is useful to identify whether a mushroom is poisonous according to the appearance features of the mushroom. The automatic recognition of mushroom toxicity has important social and application value in effectively preventing food poisoning [4].

Current methods of recognizing poisonous mushrooms can be roughly divided into four categories: chemical determination, animal experimentation [5], fungal classification, and folk experience [6]. At present, the research of poisonous mushrooms based on these methods not only has been imperfect but also has left much to be desired [7].

The classification of poisonous mushrooms has evolved from the biological level to the molecular level [2]. Therefore, the application of chemical determination methods to detect poisonous mushrooms is becoming increasingly popular [8]. However, there are strict requirements for the experimental conditions, which are often limited to the laboratory. Due to cumbersome handling and the great number of unstable toxins, the method of toxic chemical detection cannot be used to distinguish edible mushrooms from poisonous ones [9]. This approach requires professional knowledge and is, therefore, not suitable for the average person.

Generally, mushrooms with intact cells, bright colors, and the lack of birds and insects interacting with them are likely to be poisonous, particularly if they are found in places where animals are foraging. To investigate the above situation empirically, the animal acute toxicity test is commonly

used to classify poisonous mushrooms [10]. Although the methods involved are simple, they carry some limitations, such as low efficiency, material and dosage concerns, and the varying sensitivities of different animals. Therefore, special institutions or facilities are needed to facilitate the application of these methods.

Fungal recognition includes three aspects: identification, classification, and phylogeny [11]. The development of fungal taxonomy has gone through two stages: traditional taxonomy and molecular biology. These methods have mainly been used to identify the mushrooms' species. The aim of these methods is subjective, however, because fungi contain many species and complex morphological features. These methods are limited to applications involving the artificial cultivation of fungi and are only suitable for professionals. Therefore, the identification of poisonous mushrooms is not straightforward.

For a long time, humans have recognized poisonous mushrooms by observing the shape, color, odor, and secretion features empirically [12]. This method is more intuitive, but is of low accuracy proven by the annual poisoning events. Thus, it is not a reliable method for identifying whether mushrooms are poisonous. However, this method relies on background knowledge acquired by humans. People get a lot of background knowledge and experiences so that the recognition accuracy rate is high. Otherwise, the accuracy rate is low. In this paper, automatic identification can break through the limitation to determine whether it is toxic. The machine learning methods not only do not require background knowledge but also can identify unknown species.

These mushroom toxicity recognition methods have some limitations, such as low accuracy, unqualified detection of unknown toxins, strict requirements for the experimental environments, sufficient professional knowledge, and complex experimental cycles. To solve these problems, an automatic model for mushroom recognition based on appearance features is constructed in this paper. According to the observed mushroom appearance data, a poisonous mushroom can be automatically and accurately identified by the proposed model.

With the advent of the data age, machine learning and deep learning have become the core of artificial intelligence [13]. In recent years, machine learning techniques have been used to identify the toxicity of mushrooms. Chaoqun [14] used machine learning models to identify poisonous mushrooms in an application. The android-based toadstool identification system can effectively classify toadstool in real time [15]. Zhifeng [16] proposed decision fusion based on the stacking algorithm to improve the accuracy of classification methods. The image database of mushroom, obtained from the Internet by Python Crawler, was constructed by Shuaichang et al. [17]. The model-based transfer learning and the Adam algorithm as the model optimization method were applied to construct the model structure of mushrooms' image recognition.

Deep neural networks require large-scale data volumes, making already complex models even more complicated. Machine learning has unique advantages for small-sample problems. For the identification of poisonous mushrooms, three different pattern recognition methods are discussed in this paper: logistic regression, support vector machine (SVM) [18], and multigrained cascade forest (gcForest) [19]. Mushroom toxicity recognition is regarded as a problem of binary classification. By observing the appearance features of mushrooms, these machine learning methods are used to determine whether a mushroom is toxic. gcForest has the following advantages: (1) feature-based learning and iterative classification through gcForest, which has the best performance; (2) no need for professional knowledge to use the system; (3) if the unknown mushroom varieties are toxic, they can be identified quickly; and (4) independence from the effects of the natural environment unlike other methods, expanding the scope of its use.

2. Methods

Driven by big data, deep neural networks (DNNs) show great potential [20]. DNNs have achieved remarkable success in various applications. However, deep neural networks have too many hyperparameters, and their learning performance critically depends on their careful tuning. At the same time, it has an impact on DNNs that is difficult to rein in [21].

In response to the above difficulties, Zhou and Feng proposed a multigrained cascade forest framework [19]. gcForest is a decision tree ensemble method. The gcForest method, which consists of multigrained scanning and a cascade forest, is explained in detail below.

As illustrated in Figure 1, after we had obtained the dataset, we checked the integrity of dataset. Firstly, we solved the problem of missing data by Process A and Process B. Secondly, the complete data was divided into test dataset and training dataset. Then, the training dataset was input to the constructed classifier model. After the model judgment, the result of classification depended on the category with high probability. Therefore, the experimental process is shown in Figure 1 in this paper.

In this paper, we use the cascade forest to discriminate the toxicity of mushrooms, which is regarded as a binary classification problem. The 22-dimensional features are easily obtained and used as data input. After running the gcForest model, the toxicity identification results of the corresponding samples can be obtained, as shown in Figure 1.

The gcForest classifier model can accurately judge whether a mushroom is poisonous in a timely manner; thus, it has strong practicality. The classifier also has the following features: (1) ease of trainability; (2) compatibility with various datasets; and (3) an adaptively adjustable hierarchy for the cascade structure depending on the desired the complexity of the model; such a small sample dataset can achieve good recognition performance.

2.1. Cascade Forest Structure. Feature learning by deep neural networks (DNNs) depends on the layer-by-layer processing of the original features [22]. Similar to DNNs,

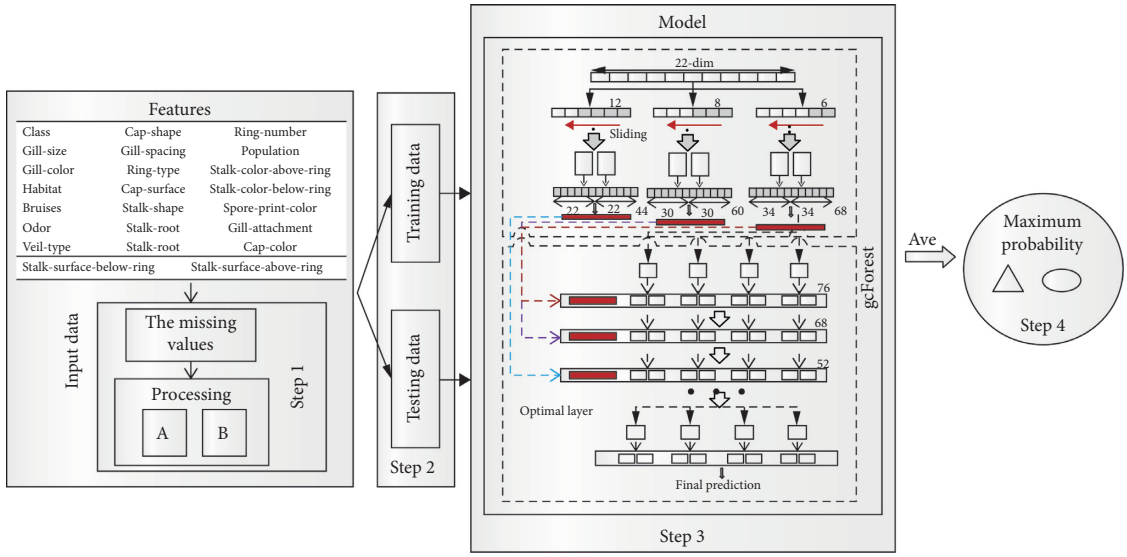


Figure 1: Flowchart of the classifier model. It consists of four main steps. (1) The missing values in the mushroom dataset are solved in two ways. (2) Seventy-five percent of the dataset is used as the training data. (3) The gcForest classifier is built on the mushroom dataset. (4) The maximum probability of whether the mushroom is toxic is determined according to the mushroom features.

gcForest uses a cascade forest structure where the information of each layer is processed in an upper layer and the result is delivered to the next layer.

As illustrated in Figure 2, suppose that the mushrooms can be divided into two categories: toxic and nontoxic. The leaf node of each forest will output a two-dimensional class as a vector, which is concatenated for the re-representation of the original input. Therefore, the next level of the cascade will receive $8 = 2 \times 4$ augmented features, and the vector dimension of the input feature will be $2 \times 4 + \text{length}(x)$. Namely, the feature dimension is equal to the number of enhanced features + the number of original (or transformed) features. Each forest will output two-dimensional class vectors, which are connected to the input features to produce the next original input. Additionally, each level contains several classifiers capable of ensemble learning [23].

For simplicity, we suppose that each layer of the cascade forest structure consists of two random forests and two completely random forests (CForests) [24]. Each forest is an aggregation of decision trees [25].

A completely random forest contains a number of completely random trees, generated by randomly selecting a feature for splitting at each node of the tree, and the tree is grown until each leaf node contains only the same class in Figure 3. Similarly, each random forest contains a number of trees, by randomly selecting $\sqrt{d}$ number of features as candidates (d is the number of input features) and choosing the one with the best Gini index (which refers to the index of optimal features when CART is used for classification problems) for splitting. The tree is grown tree until each leaf node contains only the same class of instances [26].

As illustrated in Figure 3, suppose that there are two classes, each forest will generate a two-dimensional class vector. Different symbols in the leaf nodes imply different classes. First, the red color highlights the paths along which the instance traverses to the leaf nodes, and each forest will generate the class distribution by counting the percentage of different classes of training examples at the leaf node where the concerned instance falls [19]. Then, the estimated class distribution forms a class vector. This vector is concatenated with the original feature vector as the input to the next level of the cascade. Finally, the cascade result vectors are averaged to two-dimensional vectors. The class of the maximum value is used to determine whether the mushroom is poisonous.

To reduce the occurrence of overfitting produced by each forest, k-fold cross-validation is used in this algorithm [27]. Each instance will be used as training data for $k - 1$ times to generate $k - 1$ vectors, which are finally averaged to produce the final class vector that represents the augmented features for the next level of the cascade. In other words, these augmented features are partially input in the new cascade layer, and the performance of the entire cascade on the validation set is evaluated. And if there is no significant performance improvement, the training process will automatically terminate. For that reason, the number of cascade layers can be adjusted automatically. Contrary to complex and fixed depth neural networks, gcForest can adaptively stop training to determine the number of cascade layers required.

2.2. Multigrained Scanning. The deep neural network has a powerful advantage in dealing with spatiotemporal and sequence features. Similarly, gcForest takes into account multigrained scanning prior to the cascade structure to enhance the feature learning ability of the overall structure. As shown in Figure 4, gcForest scans the original features using a multigrained sliding window. There are 22 raw features for the sample mushroom, and a window size of 8 features is used; this results in the production of 15 feature vectors. Assuming that there are two classes of mushroom

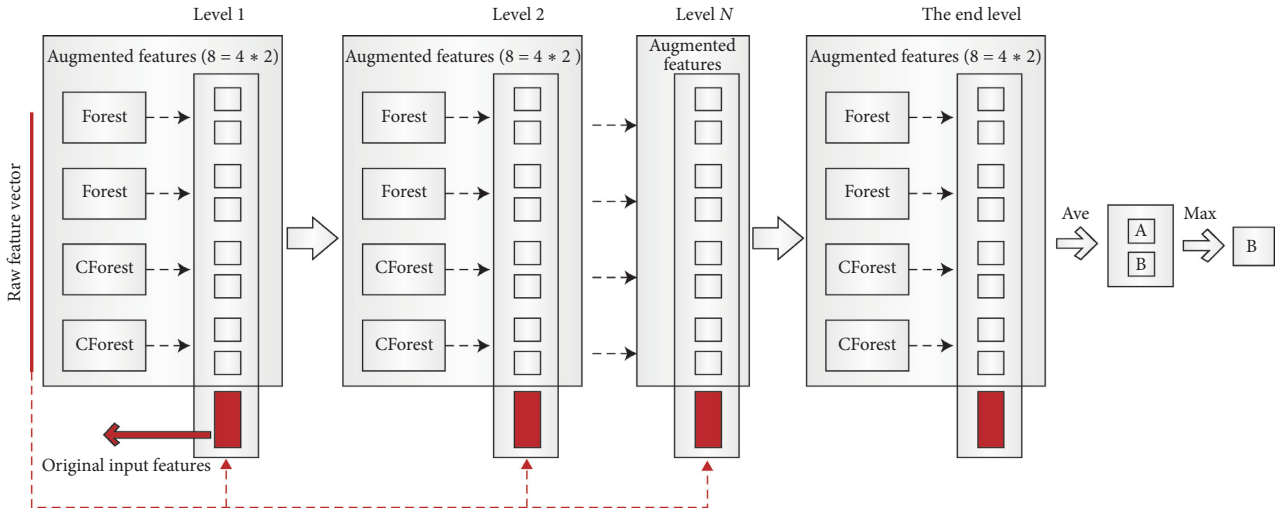


Figure 2: Illustration of the cascade forest structure [19].

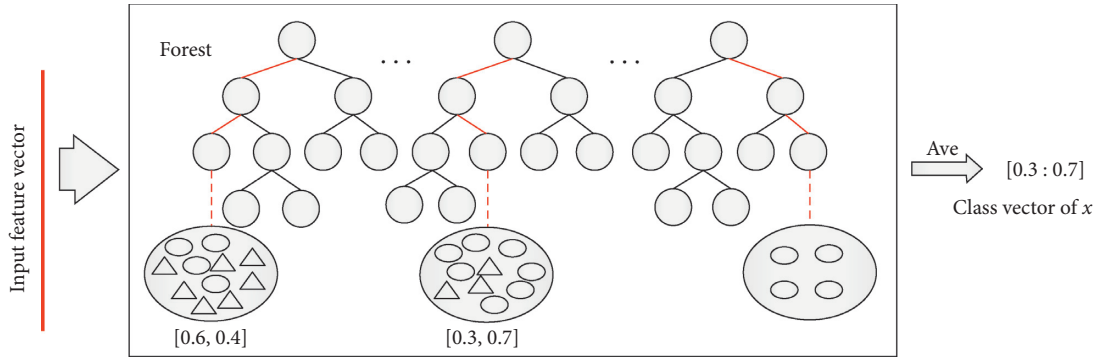


Figure 3: Illustration of class vector generation [19].

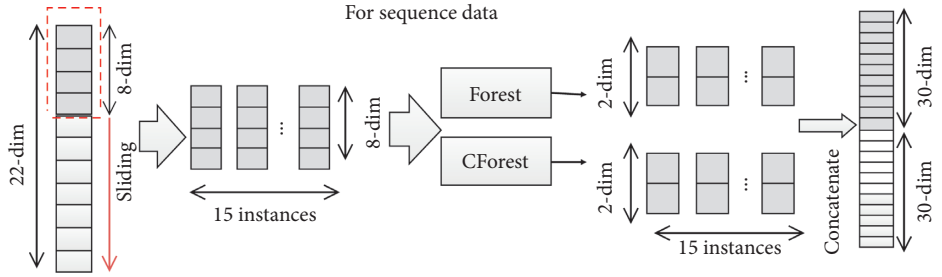


Figure 4: The procedure of sliding window scanning [19].

toxicity, the original 22-dimensional feature vector corresponds to 60 transformed dimensions.

Deep neural networks are effective in handling feature relationships. Inspired by this recognition, we enhance cascade forest with a procedure of multigrained scanning.

As illustrated in Figure 4, firstly, the complete appearance feature in the dataset is $i\,(i = 22)$ dimensions. A sliding sampling window with length $w\,(w = 8)$ is used to obtain $o\,(o = 15)$ subsample vectors with the size of v-dimensional feature. The process is similar to the sliding convolutional core $(o = \lfloor ((i - w + 2 * p)/s) \rfloor + 1 = 15)$, with the stride (s) is 1 and the padding (p) is 0. Secondly, each of the instances extracted is used to train a completely random tree forest and a random forest $(f = 1 * 2)$. Thirdly, each forest generates a length (c) probability vector, where c is the number of categories, here equal to two (corresponding to whether the mushroom is poisonous). Finally, the results of forests at each level are joined together to generate the output samples. A representation vector is generated in each forest, and these vectors can be concatenated together to obtain the final sample output in Figure 4. Therefore, there is output feature F-dimensions $(F = o * f * c = 60)$ [28].

Suppose the sliding windows with sizes of 8 and 12 features will generate 60-dimensional and 44-dimensional feature vector for each original training example, respectively. The transformed feature vectors, which contain augmented features by the previous grade, can be used to train the 2nd grade and 3rd grade of cascade forests, respectively. The procedure, in every three levels, will iterate until the boost of accuracy rate is less than the threshold, as illustrated in Figure 5. The repetition process of the training is completed.

Feature vectors of dataset will enter into the cascade forest structure in batches and connect with the upper output data to increase the disturbance of mushroom samples. The input feature vector of the cascaded forest structure will be connected with the output data of the first layer to form the input data of the second layer. Therefore, we have obtained the transforming features from the process. The final model is actually a cascade of cascade forests [19]. Each level in the cascade consists of multiple grades (cascade forests), and each corresponds to a grain of scanning, as shown in Figure 5.

The last layer classifies the upper-layer input data. Counting the percentage of different classes of training examples at the leaf node where the concerned instance falls, we then compute the average value in all forests to generate the maximum value of the class distribution to determine whether the mushroom is poisonous in Figure 5.

3. Results

3.1. Mushroom Dataset. The mushroom dataset provided by the University of California, Irvine was used to classify the toxicity of poisonous mushrooms [29]. The input features of the mushroom include class, cap shape, cap surface, cap color, bruises, odor, gill appendages, gill spacing, gill size, gill color, stalk shape, stalk root, stem-surface-above-ring, stem-surface-below-ring, stem-color-above-ring, stem-color-below-ring, veil type, veil color, ring number, ring type, spore print color, population, and habitat, for a total of 23 features (see Table 1). These features, which can be observed directly, are classified with the feature calculation. There are 8,124 recordings of mushroom data, which can be divided into two nearly balanced classes: poisonous (48%) and nonpoisonous (52%). We have created a new table (see Table 1) based on data attributes.

Each recording of mushroom has 22 features and one class label; however, part of the recording in the mushrooms' dataset is missed in the stalk-root feature. There are two ways to solve the missing data problem in Table 2:

(1) Using a KNN to complete the missing data, which is called Process A in this paper. The value of parameter K is set to 12.

(2) Treating missing attribute values as special values [30]. A special value ("m") is used to fill in the missing values of the stalk-root, which is called Process B.

The first row shows an example of complete data. The second row shows that the original data of the stalk-root is missing. In Process A, the KNN algorithm predicts the value of stalk-root ("c"). In Process B, a special value ("m") is used as the new label for all of the missing data. And the range of the stalk-root $S\_R \in \{b, c, u, e, z, r, ?\}$ is changed to $S\_R \in \{b, c, u, e, z, r, m\}$.

The two processes were specifically verified and compared in the experiment to determine the feature's effect in identifying mushroom toxicity.

The missing values of the original data are processed to obtain a complete list of character data. In the experiment, the error in the numerical data is small relative to the result, and the data can be converted to the target type data by LabelEncoder [31].

When the support vector machine classifier is used, a toxicity category value of −1 or +1 is generally assigned. Therefore, the range of the feature data toxicity y is changed to $y \in \{-1, +1\}$, which is advantageous for obtaining the hyperplane of toxicity.

Three models were run in this paper: logistic regression, SVM, and gcForest.

3.2. Evaluation Standard. To evaluate the classification models, we adopt two measures.

The index for evaluating the performance of the classifier is the accuracy of classification generally. For the test dataset, it is the ratio of the correctly labeled samples to the whole pool of samples in (1).

Assuming that the model of classification is $Y = \widehat{f}(X)$, given by

$$\text{accuracy} = r_{\text{test}} = \frac{1}{N}\sum_{i=1}^{N} I\left(y_i = \widehat{f}\left(x_i\right)\right), \tag{1}$$

where N is the test sample size and I is the indicator function.

The second is a standard evaluation performance of the binary classification used, namely, recall and precision [32]. In addition, there are F1-score and ROC curve.

The mushroom class has two possible predicted classes: edible (e) and poisonous (p). And there are four situations in confusion matrix (see Table 3): true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

The precision and recall are given by the following relation:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

Precision is defined as the number of true positives TP over the number of true positives plus the number of false positives FP [32].

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

Recall is defined as the number of true positives TP over the number of true positives plus the number of false negatives FN [32].

These quantities are also related to the F1-score, which is defined as the harmonic mean of precision and recall [32].

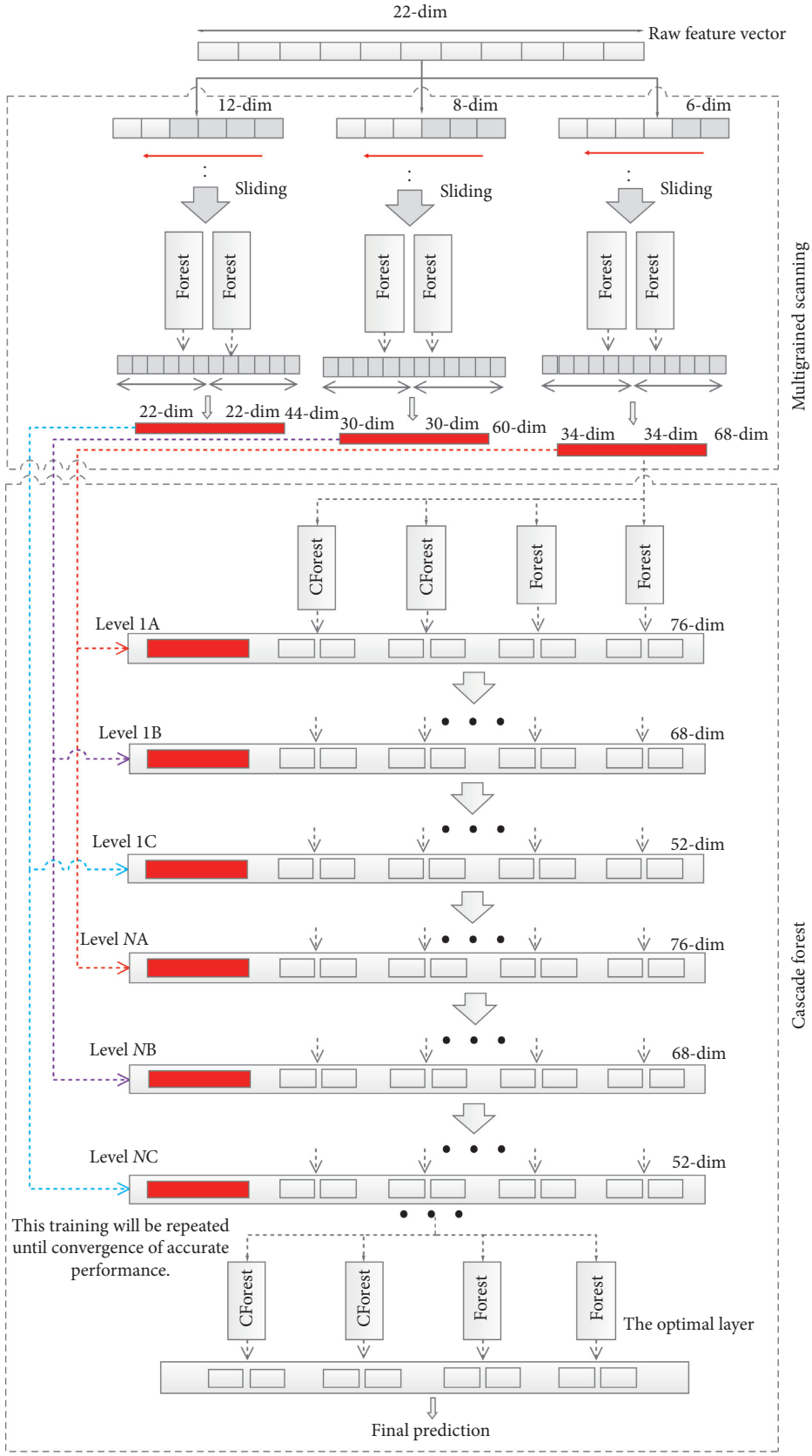


Figure 5: The overall procedure of gcForest [19].

Table 1: Attribute information.

Attribute	Range of value(s)
Class	{edible (*e*), poisonous (*p*)}
Cap-shape	{bell (*b*), conical (*c*), convex (*x*), flat (*f*), knobbed (*k*), sunken (*s*)}
Cap-surface	{fibrous (*f*), grooves (*g*), scaly (*y*), smooth (*s*)}
Cap-color	{brown (*n*), buff (*b*), cinnamon (*c*), gray (*g*), green (*r*), pink (*p*), purple (*u*), red (*e*), white (*w*), yellow (*y*)}
Bruises	{bruises (*t*), no (*f*)}
Odor	{almond (*a*), anise (*l*), creosote (*c*), fishy (*y*), foul (*f*), musty (*m*), none (*n*), pungent (*p*), spicy (*s*)}
Gill-attachment	{attached (*a*), descending (*d*), free (*f*), notched (*n*)}
Gill-spacing	{close (*c*), crowded (*w*), distant (*d*)}
Gill-size	{broad (*b*), narrow (*n*)}
Gill-color	{black (*k*), brown (*n*), buff (*b*), chocolate (*h*), gray (*g*), green (*r*), orange (*o*), pink (*p*), purple (*u*), red (*e*), white (*w*), yellow (*y*)}
Stalk-shape	{enlarging (*e*), tapering (*t*)}
Stalk-root	{bulbous (*b*), club (*c*), cup (*u*), equal (*e*), rhizomorphs (*z*), rooted (*r*), missing (?)}
Stalk-surface-above-ring	{fibrous (*f*), scaly (*y*), silky (*k*), smooth (*s*)}
Stalk-surface-below-ring	{fibrous (*f*), scaly (*y*), silky (*k*), smooth (*s*)}
Stalk-color-above-ring	{brown (*n*), buff (*b*), cinnamon (*c*), gray (*g*), orange (*o*), pink (*p*), red (*e*), white (*w*), yellow (*y*)}
Stalk-color-below-ring	{brown (*n*), buff (*b*), cinnamon (*c*), gray (*g*), orange (*o*), pink (*p*), red (*e*), white (*w*), yellow (*y*)}
Veil-type	{partial (*p*), universal (*u*)}
Veil-color	{brown (*n*), orange (*o*), white (*w*), yellow (*y*)}
Ring-number	{none (*n*), one (*o*), two (*t*)}
Ring-type	{cobwebby (*c*), evanescent (*e*), flaring (*f*), large (*l*), none (*n*), pendant (*p*), sheathing (*s*), zone (*z*)}
Spore-print-color	{black (*k*), brown (*n*), buff (*b*), chocolate (*h*), green (*r*), orange (*o*), purple (*u*), white (*w*), yellow (*y*)}
Population	{abundant (*a*), clustered (*c*), numerous (*n*), scattered (*s*), several (*v*), solitary (*y*)}
Habitat	{grasses (*g*), leaves (*l*), meadows (*m*), paths (*p*), urban (*u*), waste (*w*), woods (*d*)}

Table 2: Two ways to generate the missing data.

Class	Stalk-root		
	The original data	Process A	Process B
Edible (*e*)	Bulbous (*b*)	Bulbous (*b*)	Bulbous (*b*)
Poisonous (*p*)	?	Club (*c*)	Missing (*m*)

$$F_1 - \text{score} = \frac{2}{(1/\text{precision}) + (1/\text{recall})} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{4}$$

The ROC curve describes the change process of classifier performance changing with the change of thresholds settings, where the x-axis represents false positive rate and the y-axis represents true positive rate [33].

$$\text{FPR} = \frac{\text{FP}}{(\text{TP} + \text{TN})}, \quad \text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{5}$$

Ideally, we want the fraction of correct positive class predictions to be 1 (top of the plot) and the score of incorrect negative class prediction to be 0 (left of the plot) [33].

3.3. Experimental Analysis Using Logistic Regression. Assume that $x_0, x_1, \ldots, x_{22}$ is the feature set for mushrooms containing features such as cap shape, cap surface, cap color, swelling, and odor. According to the features, we calculate the probability of this mushroom as toxic, which is called a score. This score is used as the input of the sigmoid function:

$$s = \sum_{i=0}^{21} \omega_i x_i + \alpha. \tag{6}$$

At the same time, a sigmoid function is used to convert the scores into values within the interval of [0,1].

In the logistic regression function, the maximum likelihood function is applied to obtain the parameters of the model. Then, a logistic regression model is constructed, which is turned into an optimization problem using the log-likelihood function [34]. During the period of learning the logistic regression model, the gradient descent method or other improved scores is usually used [35].

Analysis of the cleaned features may result in excessively long training time or memory overflow, due to an excessively large feature matrix. Consequently, rescreening is the most direct and effective method to screen effective features with a random logistic regression model [36]. There are differences between the data obtained from Processes A and B, and so different effective features will be obtained.

As shown in Table 4, the effective feature stalk-color-below-ring is different across the different processes. The only variable is stalk-surface-above-ring. There may be a

Table 3: Confusion matrix.

		Predicted class	
		P	*N*
Actual class	*P*	TP	FN
	N	TP	TN

Table 4: Effective features under different treatments.

Features	Effective features by Process A	Effective features by Process B
1	Cap-surface	Cap-surface
2	Bruises	Bruises
3	Odor	Odor
4	Gill-spacing	Gill-spacing
5	Gill-size	Gill-size
6	Gill-color	Gill-color
7	Stalk-shape	Stalk-shape
8	Stalk-root	Stalk-root
9	Stalk-surface-above-ring	Stalk-surface-above-ring
10	Stalk-surface-below-ring	Stalk-surface-below-ring
11	Stalk-color-above-ring	Stalk-color-above-ring
12	Veil-color	Stalk-color-below-ring
13	Population	Veil-color
14	Habitat	Population
15		Habitat

correlation functional relationship between stalk-surface-above-ring and stalk-color-below-ring. Although the values of stalk-surface-above-ring are different processes, this feature is one of the reasons that affect accuracy.

First, the dataset is divided into 4 even parts; 3 of the parts are selected as the training set, while the remaining 3 are used as the testing set. Second, we establish a model for discriminant poisonous mushrooms on the basis of the training dataset and then judge whether the testing data indicate poisonousness. Then, statistics on the accuracy of prediction are calculated. Finally, the statistics for the effective feature analysis are compared with the statistics for the analysis using all the input features, and the results are described in Table 5.

It is easy to observe that, following Process A, the accuracy obtained with effective features and with all features differs by approximately 0.01% with maximum and minimum accuracy values of 0.955 and 0.940, respectively. Following Process B, the effects of the effective feature set and of the overall feature set on the accuracy vary greatly, with a minimum accuracy of 0.940 and a maximum accuracy of 0.955 (see Table 5). The effective feature set is an important factor in judging accuracy, but an incorrectly judged feature could reduce the accuracy of estimating mushroom toxicity.

3.4. Experimental Analysis Using SVM. The labels for the toxicity category are set to $\{-1, +1\}$ indicating whether a mushroom is toxic. The other features are converted to $0, 1, 2..., n$ using LabelEncoder to represent the samples. The dataset is divided into a training set with 75% of the data and a testing set with the remaining 25% by random selection.

Support vector machine (SVM) is essentially a binary classification model [37]. The basic idea is to find the optimal classification line (surface) from the feature space. The optimal demarcation line maximizes the distance in the binary classification of data.

The minimum accuracy under Process A is approximately 0.02% higher than that under Process B (see Table 6) on the test dataset. Further experiments with KNN constraints on the original data will produce partially correct data. Therefore, increasing the correct proportion of mushroom features can improve the accuracy of mushroom toxicity classification.

3.5. Experimental Analysis on gcForest. To reduce the contrast error between the three experiments, 25% of the dataset is used as the testing data. In this paper, the gcForest model requires two stages: multigrained scanning and the cascade forest [19]. Multigrained scanning generates the features, and the cascade forests use multiple forest cascades to derive prediction probability results [38]. The criterion for selecting each parameter is that the accuracy should be the lowest; this process is then iterated. Due to the inconsistent magnitude of the parameters, the weight assignment will cause analysis errors. The influencing factors of the parameters on the experiment will not be commented upon in detail (e.g., the number of trees in a random forest, the size of the sliding window, the sliding step, the number of cascading random forest, and the number of trees in a single cascade random forest). The maximum fluctuation in the experimental results is less than 8%.

The maximum average denotes the average of the sum of the maximum values of each parameter in the interval range. In the experimental results, when the original data are further tested under the KNN constraint, the processing will produce erroneous data with a higher error than the new category (Process B). Incorrectly judged features will reduce the accuracy of discriminating mushroom toxicity. Thus, the effect of predicting incorrect data in the experiment is greater than that from filling in special values. Table 7 shows that the average accuracy of the multigrained cascade forest classifier fluctuates between a maximum of 0.9835 and a minimum of 0.9260 on the test dataset.

3.6. Analysis of the Results. Three classifiers were built on a mushroom dataset to determine whether the mushrooms are poisonous according to their features.

In the logistic regression experiment, the accuracy of each feature in determining mushroom toxicity is calculated following steps such as dimension reduction and computing accuracy. At the same time, the results from the SVM and gcForest models are analyzed separately.

Table 5: Logistic regression results by different treatments.

Model results	Accuracy of results by Process A		Accuracy of results by Process B	
	Effective features	Overall features	Effective features	Overall features
Maximum value	0.9547021	0.953594	0.953964	0.954585
Minimum value	0.9401772	0.940423	0.945593	0.940431
Average value	0.9507167	0.950822	0.951318	0.950717

Table 6: SVM results with different treatments.

Model results	Accuracy of the results by Process A	Accuracy of the results by Process B
Maximum value	0.963873	0.962088
Minimum value	0.949419	0.937708
Average value	0.960022	0.959802

Table 7: gcForest results with different treatments.

Model results	Accuracy of the results by Process A	Accuracy of the results by Process B	Average value
Average of the maximum	0.981905	0.983506	0.982706
Average of the minimum	0.926004	0.931912	0.928958
Average value	0.953955	0.957709	0.955832

Based on the mushroom dataset, SVM has a slightly higher effect with Process A than that with Process B. Predicting erroneous data in the logistic regression and gcForest models will produce more errors than will adding a new class (Process B) from a certain feature; the data requirements are stricter, and incorrectly judged features will reduce the accuracy of determining mushroom toxicity. Table 8 shows that, among the three classifiers, the average accuracy of gcForest is 0.9835 and fluctuates less than 8%. The implementation of gcForest obviously improves the accuracy of the classification, but this improvement is not stable. Thus, it is necessary to improve the experiment to further improve the effect of the gcForest classification.

In addition, we used four indexes (precision, recall, F1-score, and ROC curve) to compare the performance of the three classification algorithms in this paper (see Table 9 and Figure 6). AUC stands for the area under the ROC curve.

From Table 9 and Figure 6, it is verified that Process B is better than Process A, and more features are beneficial to classification. In other words, the effective feature set is an important factor in judging accuracy, but an incorrectly judged feature could reduce the accuracy of estimating mushroom toxicity.

According to Table 10 and Figures 7 and 8 can be described as follows. It is verified that the result of partially correct data by KNN constraint is better than that by Process B.

The gcForest classifier obtains the higher precision rate and the lower recall rate than other classifiers on the balanced dataset. In addition, we intuitively get the information that ACU and ROC have reached the highest value on the gcForest classifier from Figures 7 and 8 and Table 10. The results further prove the outstanding classifier in the three classifiers and the applicability of gcForest classifier on this dataset.

4. Discussion

In this paper, we proposed an automatic mushroom toxicity identification method. The gcForest method proposed by Zhou has a recognition accuracy of more than 98% [19]. Based on the high accuracy requirements, we analyzed three pattern classification models. Logistic regression yielded classification results by analyzing the effective features necessary to identify toxicity. The accuracy of the SVM method is better than that of logistic regression. Compared to SVM and logistic regression, gcForest achieved better results in terms of identification accuracy. Therefore, gcForest is a good method for automatically identifying whether a mushroom is poisonous.

A number of common mushroom toxicity recognition methods are currently in use. These methods use different contributions for determining toxicity, but they have a number of limitations, such as low accuracy, unsatisfactory detection of unknown toxins, the need for a strict experimental environment, and sufficient professional knowledge and complex experimental testing techniques. To circumvent the limitations of these methods and apply them to small-sample data analysis, we used machine learning. In contrast to deep neural networks, which require great effort in hyperparameter tuning, gcForest is much easier to train and can be applied to different kinds of data in different domains. The gcForest algorithm has the following advantages: (1) it has a simple structure; (2) it can be applied to datasets of different sizes; (3) the testing techniques and handling are simple; and (4) for our experiments, mushroom toxicity is recognized quickly. Feature-based learning and iterative classifiers in the gcForest method have the best performance among the three methods of machine learning used proposed in this paper. This automatic identification

Table 8: Accuracy of the three classifiers with different treatments.

Model results	Accuracy of the results by Process A				Accuracy of the results by Process B			
	Logistic regression		SVM	gcForest	Logistic regression		SVM	gcForest
	Effective features	Overall features		Average value	Effective features	Overall features		Average value
Maximum value	0.9547	0.9536	0.9638	0.9819	0.9540	0.9560	0.9621	0.9835
Minimum value	0.9402	0.9404	0.9494	0.9260	0.9456	0.9404	0.9377	0.9319
Average value	0.9507	0.9508	0.9600	0.9540	0.9513	0.9507	0.9598	0.9577

Table 9: Indexes of precision, recall, F1-score, and AUC.

Model results	The dataset by Process A		The dataset by Process B	
	Effective features	Overall features	Effective features	Overall features
Precision	0.9623	0.9933	0.9644	0.9932
Recall	0.9596	0.9095	0.9595	0.9163
F1-score	0.9596	0.9534	0.9596	0.9573
AUC	0.9832	0.9943	0.9811	0.9964

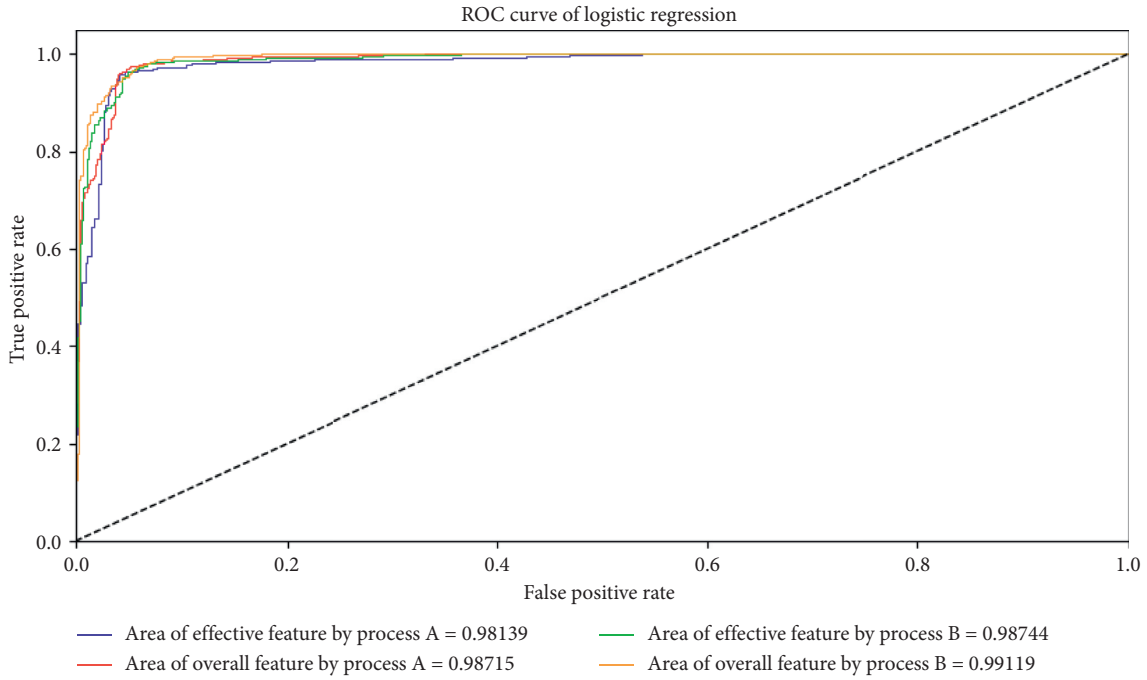


Figure 6: ROC curve of logistic regression.

Table 10: The second evaluation standard on the classifiers of SVM and gcForest.

Model results	The dataset by Process A		The dataset by Process B	
	SVM	gcForest	SVM	gcForest
Precision	0.9623	0.9846	0.9644	0.9816
Recall	0.9596	0.9189	0.9595	0.9292
F1-score	0.9596	0.95397	0.9596	0.9575
AUC	0.9832	0.9913	0.9811	0.9938

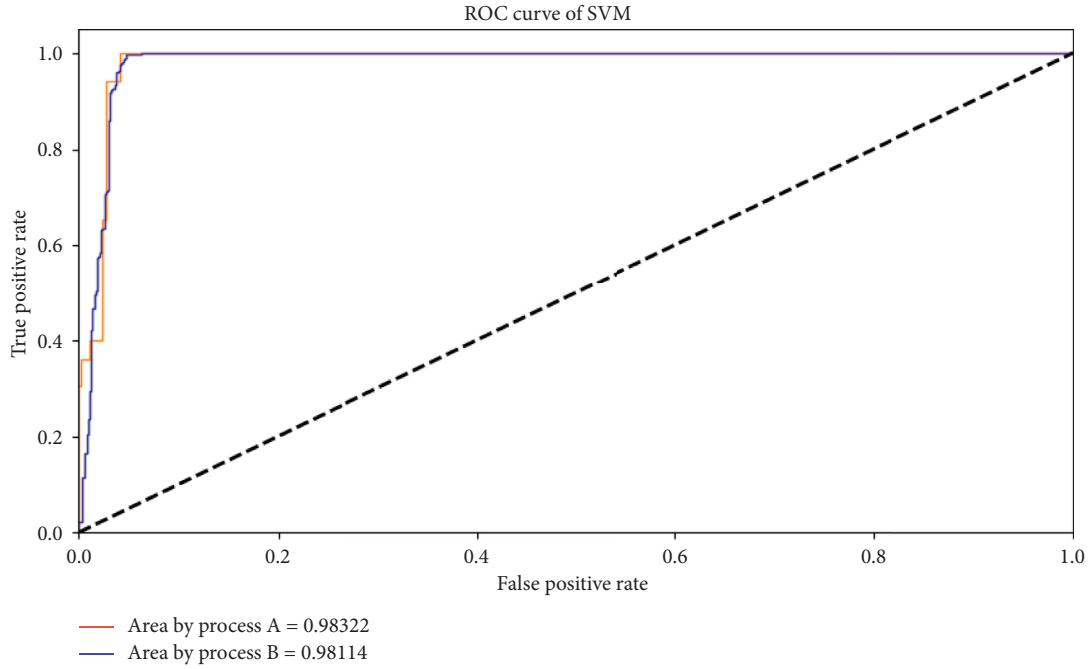


Figure 7: ROC curve of SVM.

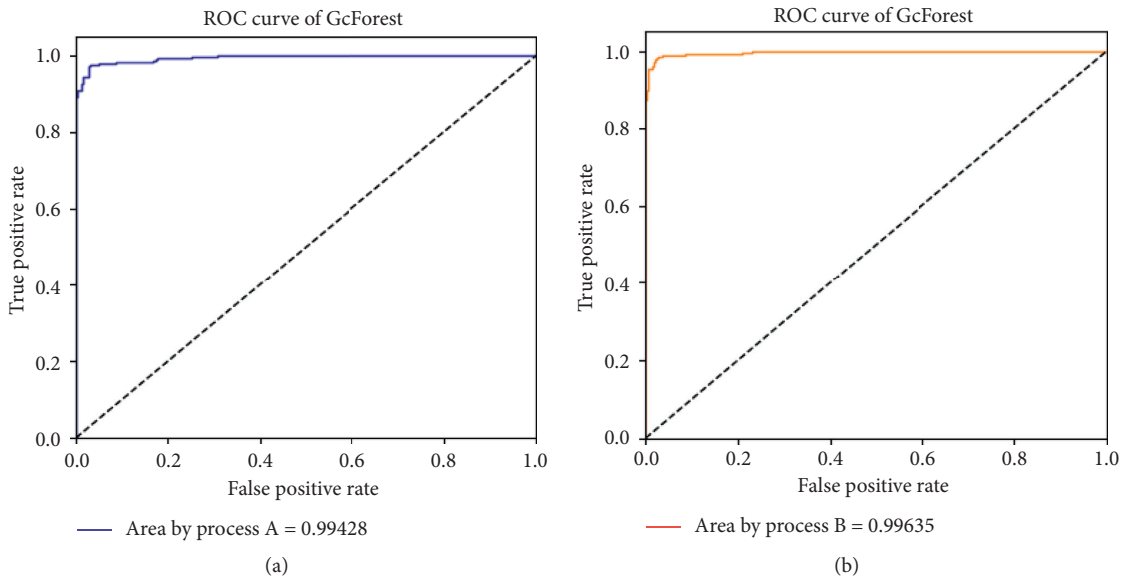


Figure 8: ROC curve of gcForest.

method is suitable for nonprofessional identification and for unknown mushroom varieties.

Among the three machine learning methods, gcForest yielded the best accuracy. However, the stability of the classifier, as shown in Table 8, needs to be improved. A reason for this error may come from the absence of features, the feature labeled stem-surface-above-ring, or it may come from the algorithm itself. Therefore, improving stability is a top priority when trying to improve the accuracy of the classifier. Since gcForest can be used for different types of datasets and recognition based on image features is more convenient than with the other classifiers [39], this method of identifying whether a mushroom is toxic can be extended to image recognition. Nonetheless, there is currently no dataset of mushroom images.

In this paper, we studied whether mushrooms were toxic, by comparing, analyzing, and summarizing four classic and traditional identification methods. According to their shortcomings, we adopted automatic identification methods to conduct the analysis based on machine learning. Based on the mushroom dataset, three pattern recognition analyses were performed. In contrast, gcForest has higher accuracy, but its stability needs to be improved. The used method identifies whether the unspecified mushroom

species is toxic in a timely manner. Because this automated identification method is not affected by the natural environment, it has important social and application value in effectively preventing food poisoning. Meanwhile, people also need to improve their safety awareness of mushrooms.

5. Conclusions

In this paper, multigrained cascade forest was used to determine whether a mushroom was poisonous based on its appearance features. LabelEncoder was used to encode the processed data to form numerical data for the mushroom dataset. According to the analysis of the dataset features, the accuracy of gcForest in data classification was approximately 98%. The maximum fluctuation of its accuracy was less than 8%, however, so the stability of the classifier needs to be improved. The gcForest structure can be used not only for large data but also for small-sized samples, and the adaptive selection of cascade layers can achieve the same accuracy as fixed patterns of deep neural networks in other datasets.

At present, research on toxins in poisonous mushrooms is still underway [40], and cases of mushroom poisoning still occur. Therefore, it is necessary to establish an automatic model for the appearance feature recognition of mushrooms' toxicity. Compared with other mushroom identification methods, the method proposed in this paper uses short cycles, is highly efficient, has low requirements in terms of the natural environment, and results in the timely identification of the toxicity of unknown species. Consequently, this method has important social and application value.

Data Availability

The mushroom data used to support the findings of this study have been deposited in the UCI (https://archive.ics.uci.edu/ml/datasets/mushroom) repository.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors have made significant contributions to this research. JXD conceived the research ideas. YYW designed the experiments, conducted the data analysis, and provided the writing of this paper. YYW performed the majority of the data processing, and the University of California Irvine [29] provided rapeseed mushrooms data. HBZ provided important insights and suggestions into this research from the perspective of a computer scientist. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2019YFC1604700), the National Natural Science Foundation of China (Nos. 61673186 and 61871196), the Natural Science Foundation of Fujian Province of China (No. 2019J01082), the Promotion Program for Young and Middle-Aged Teacher in Science and Technology Research of Huaqiao University (Nos. ZQN-YX601 and 18014083020), and the Subsidized Project for Postgraduates' Innovative Fund in Scientific Research of Huaqiao University.

References

[1] J. H. Tegzes and B. Puschner, ""Toxic mushrooms," the veterinary clinics of north America," *Veterinary Clinics of North America: Small Animal Practice*, vol. 32, no. 2, pp. 397–407, 2002.

[2] C. Lei, W. Tangkanakul, and L. Lu, "Mushroom poisoning surveillance analysis," *OSIR Journal*, vol. 1, no. 1, pp. 8–11, 2006.

[3] J. White, S. A. Weinstein, L. De Haro et al., "Mushroom poisoning: a proposed new clinical classification," *Toxicon*, vol. 157, pp. 53–65, 2019.

[4] J. H. Diaz, "Evolving global epidemiology, syndromic classification, general management, and prevention of unknown mushroom poisonings," *Critical Care Medicine*, vol. 33, no. 2, pp. 419–426, 2005.

[5] T. Fukuwatari, E. Sugimoto, K. Yokoyama, and K. Shibata, "Establishment of animal model for elucidating the mechanism of intoxication by the poisonous mushroom Clitocybe acromelalga," *Journal of the Food Hygienic Society of Japan (Shokuhin Eiseigaku Zasshi)*, vol. 42, no. 3, pp. 185–189, 2001.

[6] P. Wexler, B. D. Anderson, and S. C. Gad, *Encyclopedia of toxicology*, Vol. 1, Academic Press, Cambridge, MA, USA, 2005.

[7] M. Lu, "Present status and future prospects of the mushroom industry in China," *Acta Edulis Fungi*, vol. 13, no. 1, pp. 1–5, 2006.

[8] A. Salman, E. Shufan, and I. Lapidot, "Application of multivariate analysis and vibrational spectroscopy in classification of biological systems," in *Proceedings of the International Conference of Computational Methods in Science and Engineering*, Athens, Greece, March 2015.

[9] J. Brzezicha-Cirocka, M. Grembecka, I. Grochowska, J. Falandysz, and P. Szefer, "Elemental composition of selected species of mushrooms based on a chemometric evaluation," *Ecotoxicology and Environmental Safety*, vol. 173, pp. 353–365, 2109.

[10] J. Zhao, M. Cao, J. Zhang, Q. Sun, Q. Chen, and Z.-R. Yang, "Pathological effects of the mushroom toxin α-amanitin on BALB/c mice," *Peptides*, vol. 27, no. 12, pp. 3047–3052, 2006.

[11] J. Guarro, J. Gené, and A. M. Stchigel, "Developments in fungal taxonomy," *Clinical Microbiology Reviews*, vol. 12, no. 3, pp. 454–500, 1999.

[12] K. Tanaka, S. Miyasaka, and T. Inoue, "Histopathological effects of illudin S, a toxic substance of poisonous mushroom, in rat," *Human & Experimental Toxicology*, vol. 15, no. 4, pp. 289–293, 1996.

[13] W. A. Reynolds and F. H. Lowe, "Mushrooms and a toxic reaction to alcohol," *New England Journal of Medicine*, vol. 272, no. 12, pp. 630-631, 1965.

[14] Z. Chaoqun, *Recognition and Research of Poisonous Mushroom Based on Machine Learning*, Taigu: Shanxi Agricultural University, Jinzhong, China, 2019.

[15] P. FengLi, *Research and Design of Virus-Based Mushroom Identification System Based on Android*, Hohhot: Inner Mongolia University of Technology, Hohhot, China, 2019.
[16] Y. Zhifeng, *Application of Multi-Classifier Fusion Based on Stacking Algorithm in Identification of Poisonous Mushrooms*, Taigu: Shanxi Agricultural University, Jinzhong, China, 2019.
[17] F. Shuaichang, Y. Xiaomei, and L. Jian, "Toadstool image recognition based on deep residual network and transfer learning," *Journal of Transduction Technology*, vol. 33, no. 1, pp. 74–83, 2020.
[18] A. Kaur, K. Verma, A. P. Bhondekar, and K. Shashvat, "Implementation of bagged SVM ensemble model for classification of epileptic states using EEG," *Current Pharmaceutical Biotechnology*, vol. 20, no. 9, pp. 755–765, 2019.
[19] Z. H. Zhou and J. Feng, "Deep forest: towards an alternative to deep neural networks," in *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 3553–3559, Melbourne, Australia, August 2017.
[20] R. Giryes, G. Sapiro, and A. M. Bronstein, "Deep neural networks with random Gaussian weights: a universal classification strategy?" *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3444–3457, 2016.
[21] J. Wang, Y. Yang, and J. Mao, "Cnn-rnn: a unified framework for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Las Vegas, NV, USA, June 2016.
[22] P. Swietojanski and A. Ghoshal, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proceedings of the 2012 IEEE Spoken Language Technology Workshop*, IEEE, Miami, FL, USA, pp. 246–251, December 2012.
[23] M. P. Perrone and L. N. Cooper, "When networks disagree: ensemble methods for hybrid neural networks," in *Proceedings of the Brown University Providence Rhode Islad Institute For Brain And Neural Systems*, pp. 342–358, Providence, RI, USA, August 1992.
[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[25] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
[26] L. V. Utkin and M. A. Ryabinin, "A deep forest for transductive transfer learning by using a consensus measure," in *Proceedings of theConference on Artificial Intelligence and Natural Language*, pp. 194–208, Springer, Cham, Switzerland, November 2017.
[27] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the Fourteenth International. Joint Conference on Artificial Intelligence*, pp. 1137–1145, Montreal, Canada, August 1995.
[28] G. Hu, H. Li, Y. Xia, and L. Luo, "A deep Boltzmann machine and multi-grained scanning forest ensemble collaborative method and its application to industrial fault diagnosis," *Computers in Industry*, vol. 100, pp. 287–296, 2018.
[29] J. Schlimmer, *UCI Machine Learning Repository: Mushroom Data Set*, University of California, School of Information and Computer Science, Irvine, CA, USA, 1987.
[30] P. Balasubramaniam and R. Uthayakumar, "Mathematical modelling and scientific computation," in *Proceedings of the International Conference on Mathematical Moodelling and Scientific Computing*, Springer Science & Business Media, Gandhigram, Tamil Nadu, India, March 2012.
[31] A. Gogna, A. Majumdar, and R. Ward, "Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2196–2205, 2016.
[32] E. Alpaydin, *Introduction to Machine Learning*, MIT press, Cambridge, MA, USA, 2020.
[33] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "Machine Learning," *Neural and Statistical Classification*, vol. 13, no. 1994, pp. 1–298, 1994.
[34] K. Khamaru and R. Mazumder, "Computation of the maximum likelihood estimator in low-rank factor analysis," *Mathematical Programming*, vol. 176, no. 1-2, pp. 279–310, 2019.
[35] D. W. Hosmer and L. Stanley, *Applied logistic regression*, Vol. 398, John Wiley & Sons, Hoboken, NJ, USA, 2013.
[36] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting with discussion and a rejoinder by the authors," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.
[37] K. Peng, V. Leung, and L. Zheng, "Intrusion detection system based on decision tree over big data in fog environment," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 4680867, 10 pages, 2018.
[38] L. V. Utkin and M. A. Ryabinin, "A Siamese deep forest," *Knowledge-Based Systems*, vol. 139, pp. 13–22, 2018.
[39] B. Li, Z.-T. Fan, X.-L. Zhang, and D.-S. Huang, "Robust dimensionality reduction via feature space to feature space distance metric learning," *Neural Networks*, vol. 112, no. 4, pp. 1–14, 2019.
[40] R. Baselt, "Encyclopedia of toxicology," *Journal of Analytical Toxicology*, vol. 38, no. 7, p. 464, 2014.

Hindawi
Scientific Programming
Volume 2020, Article ID 7025403, 6 pages
https://doi.org/10.1155/2020/7025403

Research Article

Deep RetinaNet for Dynamic Left Ventricle Detection in Multiview Echocardiography Classification

Meijun Yang,[1] Xiaoyan Xiao,[2] Zhi Liu,[1,3] Longkun Sun,[4] Wei Guo,[3] Lizhen Cui,[3] Dianmin Sun,[5] Pengfei Zhang,[4] and Guang Yang[6]

[1]*School of Information Science and Engineering, Shandong University, Qingdao 266237, China*
[2]*Department of Nephrology, Qilu Hospital, Cheeloo College of Medicine, Shandong University, No. 107 Wenhuaxi Road, Jinan 250012, China*
[3]*Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan 250101, China*
[4]*Department of Cardiology, Qilu Hospital, Cheeloo College of Medicine, Shandong University, No. 107 Wenhuaxi Road, Jinan, Shandong Province, China*
[5]*Department of Thoracic Surgery, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan 250117, China*
[6]*National Heart and Lung Institute, Imperial College London, SW7 2AZ, London, UK*

Correspondence should be addressed to Zhi Liu; liuzhi@sdu.edu.cn and Pengfei Zhang; pengf-zhang@163.com

Received 18 February 2020; Revised 31 May 2020; Accepted 29 June 2020; Published 1 August 2020

Academic Editor: Chenxi Huang

Background. Currently, echocardiography has become an essential technology for the diagnosis of cardiovascular diseases. Accurate classification of apical two-chamber (A2C), apical three-chamber (A3C), and apical four-chamber (A4C) views and the precise detection of the left ventricle can significantly reduce the workload of clinicians and improve the reproducibility of left ventricle segmentation. In addition, left ventricle detection is significant for the three-dimensional reconstruction of the heart chambers. *Method*. RetinaNet is a one-stage object detection algorithm that can achieve high accuracy and efficiency at the same time. RetinaNet is mainly composed of the residual network (ResNet), the feature pyramid network (FPN), and two fully convolutional networks (FCNs); one FCN is for the classification task, and the other is for the border regression task. *Results*. In this paper, we use the classification subnetwork to classify A2C, A3C, and A4C images and use the regression subnetworks to detect the left ventricle simultaneously. We display not only the position of the left ventricle on the test image but also the view category on the image, which will facilitate the diagnosis. We used the mean intersection-over-union (mIOU) as an index to measure the performance of left ventricle detection and the accuracy as an index to measure the effect of the classification of the three different views. Our study shows that both classification and detection effects are noteworthy. The classification accuracy rates of A2C, A3C, and A4C are 1.000, 0.935, and 0.989, respectively. The mIOU values of A2C, A3C, and A4C are 0.858, 0.794, and 0.838, respectively.

1. Introduction

Heart disease is a common circulatory disease that not only seriously affects the function of the cardiovascular system but also causes certain damage to the respiratory system [1]. In severe cases, heart failure can endanger life [2]. Therefore, it is important to use more advanced methods to observe cardiac symptoms or exercise [3]. In the clinical diagnosis of heart disease, echocardiography is the most commonly used tool [4]. Echocardiography is a noninvasive technique for examining the anatomy and functional status of the heart and large blood vessels using ultrasound [5]. It uses pulsed ultrasound to measure the periodic activities of the underlying walls, ventricles, and valves through the chest wall and soft tissues [6].

The left ventricle is the focal part of the heart, and the symptoms of the left ventricle are an important basis for the diagnosis of heart disease [7]. Therefore, accurate information regarding the left ventricle extracted from echocardiography is crucial for further clinical procedures and

prognosis. To extract the information of the left ventricle, the first step is to accurately detect its position. Echocardiography contains multiple views, and each view contains multiple anatomical parts. For example, the apical two-chamber (A2C), apical three-chamber (A3C), and apical four-chamber (A4C) views of the echocardiography all contain the left ventricle, left atrium, or more anatomical parts. However, the morphology of the left ventricle is different in different views, so we need to classify these three views and detect the left ventricle. This detection will greatly reduce the time it takes for doctors to find useful information from numerous echocardiographic information.

Processing medical images through deep learning has become a popular topic. Based on the development of convolutional neural networks (CNNs), a variety of object detection algorithms have been proposed [8]. At present, the deep learning methods in the object detection field can be roughly divided into two categories: two-stage algorithms and one-stage algorithms. For example, two-stage algorithms include Region-CNN (R-CNN), Fast R-CNN, and Spatial Pyramid Pooling Convolutional Networks (SPP-Net). One-stage algorithms mainly include You Only Look Once (YOLO), Single Shot Multibox Detector (SSD), and other methods [9]. In general, two-stage algorithms are superior in detection accuracy and positioning accuracy, and one-stage algorithms are more efficient. To the best of our knowledge, there are no previous studies considering the use of deep learning-based object detection to identify the type of views and location of the left ventricle. RetinaNet [10] is a one-stage detector that can achieve the accuracy of a two-stage detector without affecting the algorithm efficiency. The innovation of RetinaNet is the introduction of focal loss, which overcomes the class imbalance problem of object detection. In this paper, to identify the A2C, A3C, and A4C views and locate the left ventricle accurately, we developed RetinaNet for multiview echocardiography. It is significant for the three-dimensional reconstruction of the heart chambers [11]. In addition, extracting the located left ventricle and then segmenting the left ventricle will improve the segmentation accuracy.

Currently, there are many classic object detection algorithms. Girshick et al. proposed the R-CNN [12], which is a milestone of deep learning-based object detection. Instead of selecting a sliding window to traverse over a picture, it used the strategy of selecting region candidate boxes, which might contain the objects to be detected. The R-CNN architecture consisted of 5 convolutional layers and 2 fully connected layers. In addition to the R-CNN, SPP-Net [13], Fast R-CNN [14], Faster R-CNN [15], R-FCN [16], and Mask R-CNN [17] improved the basis of the R-CNN to achieve better performance. The YOLO [18] algorithm proposed by Redom et al. in 2015 indicated that the object detection algorithm can be roughly divided into two categories: two-stage algorithms and one-stage algorithms. The major difference between the YOLO algorithm and the two-stage algorithms represented by the R-CNN series is that the YOLO algorithm discarded the candidate box extraction branch. The YOLO algorithm directly performed feature extraction, candidate frame regression, and classification in the same branchless convolution network, which made the network structure simple, and the detection speed was nearly 10 times faster than that of the Faster R-CNN. YOLOv2 [19], YOLOv3 [20], SSD [21], and RetinaNet are also one-stage algorithms.

There have been many studies to classify, detect, and segment cardiac images. Luo et al. [22] applied an eight-layer convolutional neural network to detect and locate the location of interest, which is a bounding box of the short-axis MRI image containing the left and right ventricles. This approach yielded better right ventricle segmentation. Vigneault et al. [23] designed so-called Ω-Net for multiview cardiac MR detection, orientation, and segmentation. Li et al. [24] applied an 11-layer convolutional neural network to automatically detect the bounding box of the myocardium from the myocardial echocardiography (MCE) images. Nizar et al. [25] proposed the use of a machine learning technique that automatically detects and localizes the aortic valve in echocardiography images. The detection used AlexNet, and the detection accuracy was 99.87% for the aortic valve.

Although previous studies have obtained promising results, these studies have only applied several layers of simple convolutional neural networks to detect a certain part of the heart and then performed segmentation based on this detection. The segmentation may not be still accurate or reproducible. Through applying the classic detection algorithms mentioned above, we can achieve more functions that have significant research value in the field of medical image analysis.

Khamis et al. [26] introduced a multistage classification algorithm that employed spatiotemporal feature extraction (cuboid detector) and supervised dictionary learning (LC-KSVD) approaches to classify the A2C, A4C, and apical long-axis (ALX) views. The recognition accuracies achieved were 97%, 91%, and 97% for the A2C, A4C, and ALX views, respectively. Madani et al. [27] trained a convolutional neural network to classify 15 standard views, and the accuracy achieved was 91.7%. Although the classification results shown in these articles were satisfactory, these studies simply identified types of views that were relatively simple and therefore have less clinical impact.

By analyzing the previous studies, we have found that the classification of multiview echocardiography has been studied, and the detection of a certain part of the heart has also been investigated. However, there is no previous study that directly detected the left ventricle with multiview classification. In this study, we developed a RetinaNet-based method for identifying A2C, A3C, and A4C views and detecting the left ventricle simultaneously from multiview echocardiography images. When a patient's echocardiographic image is put through our network, the network will automatically recognize the view types and detect the left ventricle. This detection will greatly reduce the time it takes for doctors to find useful information from numerous echocardiographic information. More importantly, this work is of great significance for the three-dimensional reconstruction of the heart chamber and left ventricular segmentation.

2. Materials and Methods

2.1. Datasets and Clinical Background. Our echocardiography images were collected from two hospitals in China, i.e., Shandong Qilu Hospital and Shengjing Hospital, with different devices by Philips and GE. The temporal rate is 65–70 Hz among frames. We extracted the A4C images of seven patients, with a total of 1238 images. We also extracted the A2C images of seven patients, with a total of 1011 images. We collected the A3C images of nine patients, with a total of 404 images. Each patient's image set contains at least one temporally cropped sequence that captures one complete cardiac cycle from end-systole (ES) to end-diastole (ED). In clinical practice, A4C is one of the main standard views of cardiac function analysis. However, when encountering a complex condition, the clinician needs to analyze heart function from multiple views, so A2C and A3C views are also necessary for research. To train and validate the results of the model, we asked a professional radiologist who is experienced in echocardiography to help us mark the correct position of the left ventricle on the A2C, A3C, and A4C images. Example datasets and left ventricle annotations are shown in Figure 1.

To have sufficient datasets and avoid overfitting, we performed data augmentation operations on the raw datasets. The data augmentation we used mainly includes the following three procedures:

(1) Random flip: flips the input images and the corresponding boxes with a probability of 0.5

(2) Random crop: crops the given images to a random size and aspect ratio

(3) Resize: resizes the input images to the given size

2.2. Network Architecture. RetinaNet mainly includes three subnetworks: a residual network (ResNet) [28], a feature pyramid network (FPN) [29], and two fully convolutional networks (FCNs) [30]. The RetinaNet network architecture is summarized in Figure 2.

ResNet: the main contribution of ResNet was the idea of residual learning, which allows the original input information to be directly transmitted to the following layer [31]. ResNet can use different network layers. The commonly used types of network layers are 50-layer, 101-layer, and 152-layer. We chose the 101-layer architecture with the best training performance [32]. We extracted the features of the echocardiography using ResNet and then put them forward to the next subnetwork.

FPN: FPN is a method for efficiently extracting the features of each dimension in a picture using a conventional CNN model. First, we used a single-dimensional image as the input to ResNet. Then, starting from the second layer of the convolutional network, the features of all the layers were selected by the FPN and then combined to form the final feature output combination.

FCN: the class subnet in the FCN performed the classification task. This subnet could identify which view the echocardiography image belongs to. The box subnet in the FCN performed the border regression task. Its role was to detect the position of the left ventricle in the echocardiography images and record the coordinates.

Focal loss: focal loss is an improved version of the cross-entropy loss, and the binary cross-entropy expression is as follows:

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p), & \text{if} \quad y = 1, \\ -\log(1-p), & \text{otherwise}, \end{cases} \tag{1}$$

where $y \in [\pm, 1]$ is the ground truth category and $p \in [0, 1]$ is the predicted probability of the model for category $y = 1$.

$$p_t = \begin{cases} p, & \text{if} \quad y = 1, \\ 1-p, & \text{otherwise}. \end{cases} \tag{2}$$

The above formula can be abbreviated as

$$\mathrm{CE}(p, y) = \mathrm{CE}(p_t) = -\log(p_t). \tag{3}$$

To solve the problem of the data imbalance between the positive and negative samples, the original form is changed into the following form:

$$\mathrm{CE}(p_t) = -\alpha_t \log(p_t), \tag{4}$$

among them,

$$\alpha_t = \begin{cases} \alpha, & \text{if} \quad y = 1, \\ 1-\alpha, & \text{otherwise}, \end{cases} \tag{5}$$

where $\alpha \in [0, 1]$ is the weight factor. To solve the problem of the difficult sample, the focusing parameter γ is introduced to obtain the final form of the focal loss:

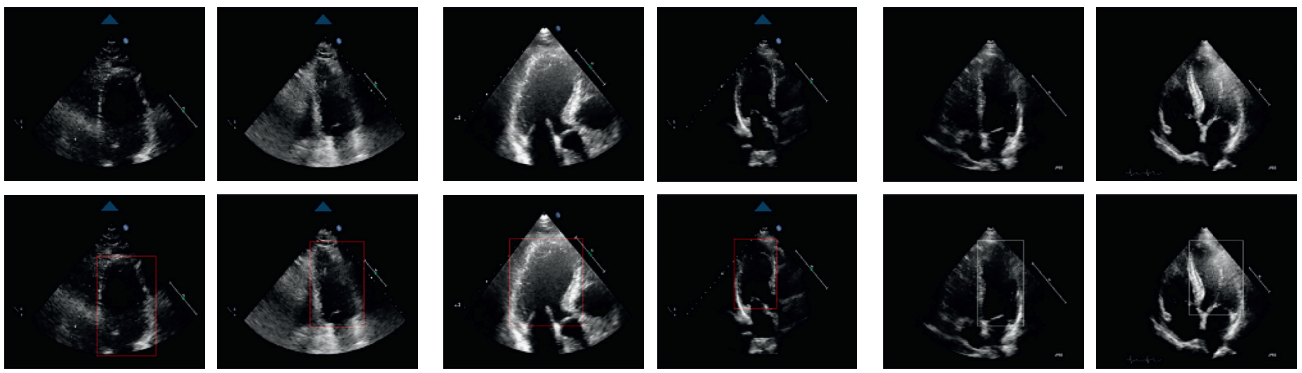
$$\mathrm{FL}(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t). \tag{6}$$

3. Results

We have a total of 1238 images of the A4C view, 1011 images of the A2C view, and 404 images of the A3C view. For each view, we divided the images into a training set, a validation set, and an independent testing set using the ratio of 7 : 1 : 2.

We used a Dell Inspiron 3670 workstation, which had an Intel Core i7-8700 CPU @ 3.20 GHz, 8 GB of memory, and an NVIDA GeForce 1050 Ti GPU for training and testing our developed model. Our work was implemented using the PyTorch platform.

The images are preprocessed so that only the cardiac images are selected and resized to 512 by 512 when put into the model. The gray value is normalized to [0, 1] by dividing by 256. We extract the features of the echocardiography using ResNet and then put them forward to the FPN to form the final features. The classifier subnet in the FCN identifies which view the echocardiography image belongs to. The box subnet in the FCN detects the position of the left ventricle in the echocardiography images and records the coordinates.



(a) (b) (c)

Figure 1: The first row shows the original echocardiography images, and the second row shows the images with left ventricle region annotations. (a) A2C images. (b) A3C images. (c) A4C images.

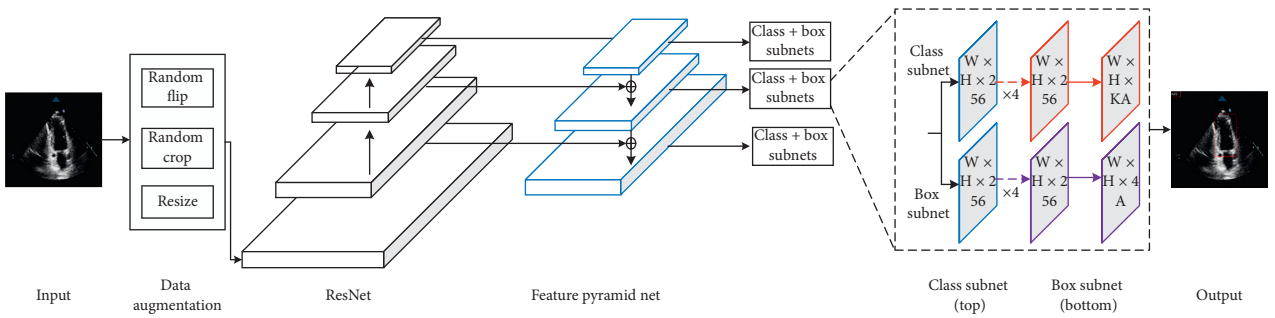
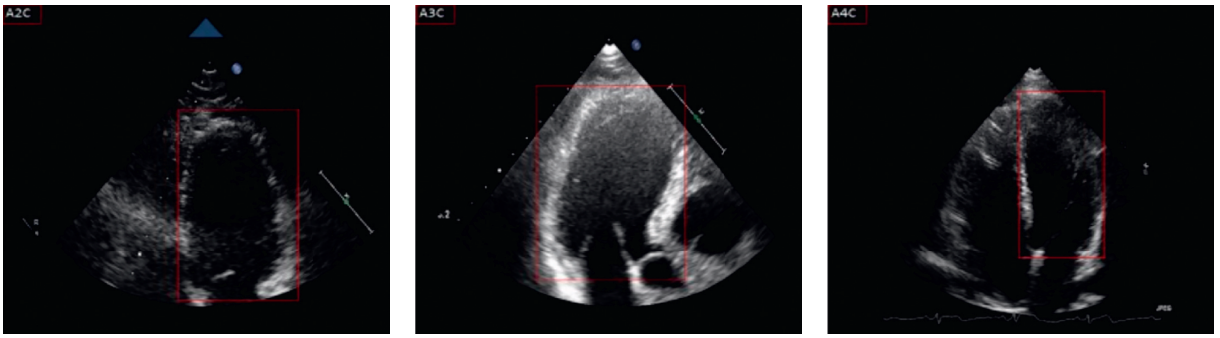


Figure 2: The RetinaNet network architecture consists of a ResNet, a FPN, and two FCN subnets.

Table 1: Results of the classification and mIOU of the three views.

	A2C	A3C	A4C
Accuracy	1.000	0.935	0.989
mIOU	0.858	0.794	0.838



(a) (b) (c)

Figure 3: We input test images into our trained RetinaNet for independent testing and automatically identified A2C, A3C, and A4C images and detected the localization of the left ventricle. The type of view the network predicts is shown in the upper left corner with white text. The red rectangle in the middle represents the left ventricle position detected by our model. Left ventricle detection results in (a) A2C, (b) A3C, and (c) A4C.

The kernel size is 3, and the batch size is 16. We use the ReLU activation in all constitutional layers of the deep networks and the sigmoid function in the prediction layers. We use the SGD optimizer with a learning rate of 1e-3. In addition to displaying the position of the left ventricle on the test image, we also display the category of the view in the upper left corner of the test image, which will facilitate diagnosis.

We use the accuracy and the mean intersection-over-union (mIOU) to evaluate the performance of our model. The accuracy was defined as the ratio of the number of correctly predicted samples to the total number of predicted samples. In our experiments, IOU was expressed as the overlap of the candidate box generated by the network with the ground truth box, that is, the ratio of its intersections to the unions. The higher the correlation, the larger the value. The ideal situation is a complete overlap, which is indicated by a ratio of 1. The mIOU and accuracy results are shown in Table 1. The types of views we predicted through our model and the detected left ventricles are shown in Figure 3.

4. Discussion

There are still a few improvements that can be applied to our work. For example, there are limited image samples in our dataset. In the next phase of the study, we will extract the detected left ventricle and, on this basis, try to achieve better left ventricle segmentation using an unsupervised method. At the same time, we will continue to study the application of this work in the three-dimensional reconstruction of heart chambers.

5. Conclusion

Left ventricle detection from multiview echocardiography images can help clinicians diagnose heart disease more comprehensively and accurately and, more importantly, is of great significance for the three-dimensional reconstruction of heart chambers. In addition, it can improve the accuracy of left ventricle segmentation. In this study, we propose to simultaneously use RetinaNet to identify A2C, A3C, and A4C images and detect the left ventricle in multiview echocardiography. The results have demonstrated that our proposed method can better classify the A2C, A3C, and A4C views and can also better detect the left ventricle from these views.

Data Availability

All the data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Meijun Yang contributed to the experiments and writing of the paper. Xiaoyan Xiao provided the fundamental medical analysis for the experimental results. And Meijun Yang and Xiaoyan Xiao contributed equally to this paper. Zhi Liu and Pengfei Zhang organized this study. Longkun Sun and Dianmin Sun were in charge of data collection and building the dataset. Wei Guo and Lizhen Cui gave computing supports. Guang Yang contributed to writing.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 1192780063 and 91846205, the Key Research and Development Plan of Shandong Province under Grant nos. 2018YFJH0506 and 2019JZZY011007, and the Major Fundamental Research of the Natural Science Foundation of Shandong Province under Grant no. ZR2019ZD05.

References

[1] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," in *Proceedings of 2nd International Conference on Computing for Sustainable Global Development (INDIA.Com)*, IEEE, pp. 704–706, New Delhi, India, March, 2015.

[2] C. Xu, L. Xu, Z. Gao et al., "Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture," *Medical Image Analysis*, vol. 50, pp. 82–94, 2018.

[3] Z. Gao, H. Zhang, D. Wang et al., "Robust recovery of myocardial kinematics using dual $\mathscr{H}\infty$ $\mathcal {H}_{\infty }$ criteria," *Multimedia Tools and Applications*, vol. 77, no. 17, pp. 23043–23071, 2018.

[4] V. Apte, L. Tam, A. Han et al., "Evaluation of circumferential and longitudinal strain in a rabbit fetal heart model using 4d echocardiography," in *Proceedings of 39th Annual Northeast Bioengineering Conference*, pp. 23-24, IEEE, Syracuse, NY, USA, April 2013.

[5] P. Joos, H. Liebgott, F. Varray et al., "High-frame-rate 3-D echocardiography based on motion compensation: an in vitro evaluation," in *Proceedings of IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, Washington, DC, USA, September 2017.

[6] R. Al Mukaddim, K. Samimi, A. Rodgers, T. A. Hacker, and T. Varghese, "Comparison of cardiac displacements in a murine model of myocardial ischemia using Cardiac Elastography and speckle tracking echocardiography," in *Proceedings of IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, Washington, DC, USA, September 2017.

[7] A. Goyal, D. Bathla, P. Sharma, M. Sahay, and S. Shagun, "MRI image based patient specific computational model reconstruction of the left ventricle cavity and myocardium," in *Proceedings of the International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1065–1068, IEEE, Noida, India, April 2016.

[8] M. Z. Alom, T. M. Taha, and C. Yakopcic, "The history began from alexnet: a comprehensive survey on deep learning approaches," 2018, https://arxiv.org/abs/1803.01164.

[9] H. Zhang and X. Hong, "Recent progresses on object detection: a brief review," *Multimedia Tools and Applications*, vol. 78, no. 19, pp. 27809–27847, 2019.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. He, and D. Piotr, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, IEEE, Venice, Italy, October 2017.

[11] L. Xu, X. Huang, J. Ma et al., "Value of three-dimensional strain parameters for predicting left ventricular remodeling after ST-elevation myocardial infarction," *The International Journal of Cardiovascular Imaging*, vol. 33, no. 5, pp. 663–673, 2017.

[12] R. Girshick, J. Donahue, and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, IEEE, Columbus, OH, USA, June 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[14] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, IEEE, Santiago, Chile, December 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal networks: Advances in Neural Information Processing Systems*, Neural Information Processing Systems (NIPS), Montreal, Quebec, Canada, 2015.

[16] J. Dai, Y. Li, K. He, and J. Sun, *Object Detection via Region-Based Fully Convolutional Networks: Advances in Neural Information Processing Systems*, Neural Information Processing Systems (NIPS), Barcelona, Spain, 2016.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, IEEE, Venice, Italy, October 2017.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, IEEE, Las Vegas, NV, USA, June 2016.

[19] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, IEEE, Honolulu, HI, USA, July 2017.

[20] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[21] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *European Conference on Computer Vision*, pp. 21–37, Springer, Cham, Switzerland, 2016.

[22] G. Luo, R. An, K. Wang, S. Dong, and D. Zhang, "A deep learning network for right ventricle segmentation in short-axis MRI," in *Proceedings of Computing in Cardiology Conference (CinC)*, pp. 485–488, IEEE, Vancouver, BC, Canada, September 2016.

[23] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "Ω-net (omega-net): fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks," *Medical Image Analysis*, vol. 48, pp. 95–106, 2018.

[24] Y. Li, C. P. Ho, M. Toulemonde, N. Chahal, R. Senior, and M.-X. Tang, "Fully automatic myocardial segmentation of contrast echocardiography sequence using random forests guided by shape model," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1081–1091, 2017.

[25] M. H. A. Nizar, A. Khalil, C. K. Chan, N. P. Utama, and K. W. Lai, "Pilot study on machine learning for aortic valve detection in echocardiography images," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 1, pp. 9–14, 2019.

[26] H. Khamis, G. Zurakhov, V. Azar, A. Raz, Z. Friedman, and D. Adam, "Automatic apical view classification of echocardiograms using a discriminative learning dictionary," *Medical Image Analysis*, vol. 36, pp. 15–21, 2017.

[27] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–8, 2018.

[28] 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*, pp. 770–778, IEEE, Las Vegas, NV, USA, June 2016.

[29] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, IEEE, Honolulu, HI, USA, July 2017.

[30] 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, IEEE, Boston, MA, USA, June 2015.

[31] Z. Gao, Y. Li, Y. Sun et al., "Motion tracking of the carotid artery wall from ultrasound image sequences: a nonlinear state-space approach," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 273–283, 2017.

[32] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "Automatic ship detection based on RetinaNet using multi-resolution gaofen-3 imagery," *Remote Sensing*, vol. 11, no. 5, p. 531, 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8855063, 12 pages
https://doi.org/10.1155/2020/8855063

Research Article

Optimal Utilization of Light Energy in Semiclosed Greenhouse Using Three-Dimensional Cucumber Model

Tingting Qian,[1,2] Xiuguo Zheng,[1,2] Juan Yang,[1,2] Yeying Xu,[1,2] Yan Wang,[1,2] Qiang Zhou,[1] Shenglian Lu,[3] and Xiaotao Ding[1]

[1]*Agricultural Information Institutes of Science and Technology, Shanghai Academy of Agriculture Sciences, Shanghai 201403, China*
[2]*Shanghai Engineering Research Center of Information Technology in Agriculture, Shanghai 201403, China*
[3]*College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China*

Correspondence should be addressed to Shenglian Lu; shll@126.com and Xiaotao Ding; xiaotao198108@163.com

Received 24 April 2020; Revised 16 June 2020; Accepted 9 July 2020; Published 1 August 2020

Academic Editor: Chenxi Huang

In the east of China, low temperature and light energy in winter are the main factors for the decline in cucumber yield, as well as in greenhouses without supplementary light. Optimal utilization of light energy is critical to increase cucumber yield. In this study, experimental measurements were conducted in two scenarios, April to May (Apr-May) and November to December (Nov-Dec) 2015, respectively, to analyze leaf development, dry matter accumulation, and yield of cultivated cucumber. Statistical analysis showed that leaves grew in Nov-Dec had larger leaf area and lower dry matter than leaves grew in Apr-May. This revealed that the dry matter accumulation rate per unit area was lower in winter. To be precise, the yield 0.174 kg/m^2 per day in Nov-Dec was 35.3% lower than the yield in Apr-May. Environmental monitoring data showed that there was no significant difference in the average temperature between two scenarios, but the light intensity in Nov-Dec was only 2/3 of that in Apr-May. Three-dimensional (3D) cucumber canopy models were used in this study to quantify the effects of weak light on dry matter production in Nov-Dec. Three 3D canopies of cucumber were reconstructed with 20, 25, and 30 leaves per plant, respectively, by using a parametric modeling method. Light interception of three canopies from 8:00 to 15:00 on 4 November 2015 was simulated by using the radiosity-graphic combined model (RGM) with an hourly time step. CO_2 assimilation per unit area was calculated using the FvCB photosynthetic model. As a result, the effects of light intensity and CO_2 concentration on the photosynthetic rate were considered. The leaf photosynthesis simulation result showed that during the daytime in winter, the RuBP regeneration-limited assimilation A_j was always less than the Rubisco-limited assimilation A_c. This means that the limiting factor affecting the photosynthesis rate in winter was rather light intensity. As the CO_2 concentration in the greenhouse was utmost in the morning, increasing the light intensity and therefrom increasing the canopy light interception at this time will be highly beneficial to increase the yield production. Through a comparative analysis of photosynthetic characteristics in these three virtual 3D canopies, the 25-leaf canopy was the best-performed canopy structure in photosynthetic production in winter. This study provides an insight into the light deficiency for yield production in winter and a solution to make optimal use of light in the greenhouse.

1. Introduction

Cucumber is one of the most popular crops grown in facility agriculture. It requires relatively high light intensity and temperature. Low temperature and deficient light energy in winter in the east and north of China are the main factors for the decline of cucumber yield. With the promotion and application of semiclosed greenhouses in China, the excess solar energy in summer was collected and stored, and this energy should be reused to heat the greenhouse in winter [1]. As a result, the impact of low temperature in winter on yield has been greatly improved [2]. Nevertheless, in greenhouses without supplementary light, low light intensity remains an important cause affecting winter production [1]. For most greenhouse vegetables, 1% light increment results in 0.7–1% increase in harvestable product [3]. Therefore, optimal

utilization of light energy should be critical for increasing the yield of cucumber in winter.

Optimization and control of the light environment in greenhouse have been widely studied. In order to increase the incident light intensity, supplementary light sources are often added on top of the canopy during the growing period as early as the 1980s [4]. For most greenhouse fruiting vegetables, such as cucumber and tomato, vegetative growth is concentrated on the top and middle of the canopy while reproductive growth occurs at the middle and bottom. In recent years, lighting strategies including light source type and distribution, spectral composition, and lighting duration have been widely studied. Hovi et al. [5] have shown that interlighting by using HPS (high-pressure sodium) can increase by 10% of cucumber fruit yield. However, these kinds of light sources are very difficult to be used as interlighting in narrow row arrangements, as they produce high heat. Hao et al. [6] studied the effect of HPS and LED (light-emitting diodes) light source and light position (on the top or inside of the canopy) on mini-cucumber growth rate and fruit yield. The result shows that the plants grown with LED interlighting achieved higher energy use efficiency than that with the top HPS. And, once canopy closure achieves, increasing total canopy light interception, especially increasing middle and lower light interception, is more conducive to yield [7].

In cucumber canopy, light interception is directly affected by canopy structure, which depends on plant growth and cultivation management. Accurate measurement of total light interception in the canopy has always been a great challenge. In previous studies, Beer–Lambert's law was widely used to calculate canopy light interception, but this law assumes that the canopy is a uniform medium, and ignored the influence of leaf distribution and inclination angle on light transmission [8]. For horticultural crops planted in rows, the canopy structure exhibits distinct heterogeneity. Since light distributes significantly different between and within crop rows, the orientation of the leaf directly determines its light interception [9]. The 3D structural model, which delivers a structure close to the real canopy, can accurately describe the distribution and location of organs in 3D space. Combing with ray casting, radiosity, and Monte Carlo ray-tracing approaches, local light conditions and light interception can be simulated at the leaf level and upscale to the canopy level [10]. Thus, the accuracy of the light interception estimation can be greatly improved.

Recent studies have shown that the 3D structural model is extremely useful for light interception-related calculation, for example, evaluating the effect of leaf area distribution and leaf angles on canopy light interception for "ideotype" selection [11, 12] and optimizing canopy structure pruning management [13]. In many FSPM (functional and structural plant models) research studies, 3D structural models were always combined with photosynthetic models and dry matter distribution models at the aim to design crops with higher yield potential and to enhance resource use efficiency [14–16].

In the present study, we combined 3D cucumber canopy model, ray-tracing method, and photosynthesis model together to analyze light conditions inside the canopy. By virtual experiments, we discussed a solution in greenhouse management to increase the photosynthetic production of cucumber crops in winter.

2. Materials and Methods

2.1. Field Experiments and Measurement on Plants. Field experiments were conducted in May and November 2015, respectively, in a semiclosed Venlo-type greenhouse at Shanghai Chongming Island (lat. 31°34′N, long. 121°41′E). The detailed information of the experimental greenhouse can be referenced in Ding et al. [2]. Mini-cucumbers (*Cucumis Sativus* L. cv. Deltastar) were planted on rockwool slabs along two parallel lines and each plant is located with a growing wire at a stem density of 2.8 stems/m^2. Air temperature, relative humidity, and CO_2 concentration in the greenhouse were recorded automatically at 5 min intervals using one climate sensor (Priva, De Lier, the Netherlands) per greenhouse, placed at the top of the canopy.

2.1.1. Measurement 1: Organ Developing Measurement. Two measurements were conducted in 2015 from 15th April to 14th May and from 4th November to 3 December, and for convenience, Apr-May and Nov-Dec will be used to refer the two experimental scenarios, respectively, hereinafter. Every day in the experiment, five newly appeared phytomers, including an internode with the directly attached leaves and petioles, were selected to be marked as the initial point of organ expansion. In total, 150 phytomers corresponding to 450 organs were marked in the thirty consecutive days. Morphological measurements were performed on all marked phytomers, respectively, on May 16, 2015, and December 3, 2015. The measurement contents include leaf length (LL), leaf width (LW), petiole length (PL), petiole diameter (PD), internode length (IL), and internode diameter (ID). Leaf length was measured from the lamina tip to the intersection end of petiole along the lamina midrib. Leaf width was measured from tip to tip between the widest lamina lobes, as referenced in Cho et al. [17]. The diameter of the petiole and internode was measured in the middle of the organ. The manual measurement error is ±1 mm. After the measurement, the leaves and petioles were destructively dried and weighted.

2.1.2. Measurement 2: Canopy Structure Measurement. Canopy structure measurements were conducted on 4 November 2015. A total of 8 cucumber plants from 2 (rows) ∗4 (plants) adjacent plants were selected, and all organs of the whole plant were measured manually. The measurement contents include leaf length, leaf width, petiole length, petiole diameter, internode length, and internode diameter. After the measurement, organs were segmented, dried, and weighed one by one.

2.2. 3D Reconstruction of the Canopy

2.2.1. Organ Expansion Modeling. Based on measurement 1, the least-squares method was used to fit statistical models for

organ development. We used logistics function for leaf length, leaf width, petiole diameter, and internode diameter and Boltzmann function for petiole length, internode length, leaf dry weight (LDW), and petiole dry weight (PDW). The parameters of descriptive models for organ development are listed in Tables 1 and 2. The sum of the thermal time (GDD) is calculated independently: $\mathrm{GDD} = \sum (T - T_{\mathrm{base}})/288$, where T is the temperature measured by the sensor, T_{base} is the base temperature, 10°C for mini-cucumber, and 288 is an invariant for the number of measurements per day. The duration is from the initial date till the end of measurement and valid temperature ranging from 10 to 40°C is applied for cucumber.

2.2.2. Parametric Construction of 3D Cucumber Model. Parameterized modeling method was used to construct the 3D cucumber canopies. The shape of each organ was determined by morphological parameters, which were generated by using the equations described above. Organ morphology was composed of skeletons and triangular mesh. B-splines curve was used to represent these skeletons in the computer. The control points on B-spline curve can determine the size of the organs (Figure 1), and the spatial distance of the control points was determined by the organ morphological parameters, which in turn determined the topology of the plant [18, 19].

2.2.3. 3D Canopy Construction for Cucumber. Three 3D canopies of cucumber were reconstructed with 20, 25, and 30 leaves per plant, respectively, by using the parameterized modeling method. Each virtual canopy was consisted of sixteen plants (4 rows∗4 plants). The interplant distance in one line was 40 cm, the line spacing was 70 cm (inside) and 110 cm (path), respectively (Figure 2).

2.3. Light Interception Modeling. Light interception of three canopies from 8:00 to 15:00 on 4 November 2015 was simulated by using the radiosity-graphic combined model (RGM) with an hourly time step. Distribution of 3D diffuse photosynthetically active radiation (DPAR) in the canopy was simulated discretely by dividing the sky hemisphere into a grid of fine cells that allowed for the anisotropic distribution of DPAR over the sky hemisphere. Diffuse radiation was calculated using the method of Wen et al. [10]. It was assumed that a leaf absorbs 87%, transmits 7%, and reflects 6% of the incident photosynthetically active radiation [15]. Total PAR and proportion of diffuse light intensity were measured using SunScan Canopy Analysis System (Delta Company, UK).

2.4. Leaf Photosynthesis Modeling. FvCB model (Farquhar–von Caemmerer–Berry biochemical photosynthesis model) was applied in leaf photosynthesis modeling. According to Farquhar et al. [20], assimilation rate A is the minimum of the Rubisco-limited (A_c, $\mu\,\mathrm{mol{\cdot}CO_2{\cdot}m^{-2}{\cdot}s^{-1}}$) or RuBP regeneration-limited (A_j, $\mu\,\mathrm{mol{\cdot}CO_2{\cdot}m^{-2}{\cdot}s^{-1}}$) photosynthesis rate:

$$A_c = \frac{V_{c\max} \cdot (C_c - \Gamma^*)}{[C_c + K_c \cdot (1 + O/K_0)] - R_d}, \tag{1}$$

$$A_j = \frac{J \cdot (C_c - \Gamma^*)}{(4C_c + 8\Gamma^*) - R_d}, \tag{2}$$

where C_c (chloroplastic CO_2 concentration, $\mu\,\mathrm{mol{\cdot}mol^{-1}}$) is calculated by using the following equation:

$$C_c = \frac{C_i - A}{g_m}. \tag{3}$$

When substituting equation (3) into equations (1) and (2), we get the quadratic equation of variable A_c and A_j, and the solutions are as follows:

$$A_c = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \tag{4}$$

$$\begin{aligned} a &= \frac{1}{g_m}, \\ b &= \frac{(R_d - V_{c\max})}{g_m - C_i - K_c(1 + O/K_0)}, \\ c &= V_{c\max}(C_i - \Gamma^*) - R_d\left[C_i + K_c\left(\frac{1+O}{K_0}\right)\right], \end{aligned}$$

$$A_j = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \tag{5}$$

$$\begin{aligned} a &= \frac{4}{g_m}, \\ b &= \frac{(4R_d - J)}{g_m - 4C_i - 8\Gamma^*}, \\ c &= J(C_i - \Gamma^*) - 4R_d(C_i + 2\Gamma^*). \end{aligned}$$

J (rate of electron transport, $\mu\,\mathrm{mol{\cdot}m^{-2}s^{-1}}$) was calculated by the following equation:

$$J = \frac{\left(K \cdot I + J_{\max} - \sqrt{(K \cdot I + J_{\max})^2 - 4\theta \cdot J_{\max} \cdot K \cdot I}\right)}{2\theta}, \tag{6}$$

where $V_{c\max}$ is the maximum rate of Rubisco carboxylation ($130.84\,\mu\,\mathrm{mol{\cdot}CO_2{\cdot}m^{-2}{\cdot}s^{-1}}$); Γ^* is the CO_2 compensation point in the absence of dark respiration ($43.02\,\mu\,\mathrm{mol{\cdot}mol^{-1}}$); K_c ($404\,\mu\,\mathrm{mol{\cdot}mol^{-1}}$) and K_o ($278\,\mathrm{mmol{\cdot}mol^{-1}}$) are Michaelis–Menten constants of Rubisco for CO_2 and O_2; O is the mole fraction of O_2 at the site of carboxylation ($210\,\mathrm{mmol{\cdot}mol^{-1}}$); R_d is the respiration rate ($1.08\,\mu\,\mathrm{mol{\cdot}CO_2{\cdot}m^{-2}{\cdot}s^{-1}}$); g_m is mesophyll conductance ($0.8\,\mathrm{mol{\cdot}m^{-2}{\cdot}s^{-1}}$); C_i is ambient CO_2 concentration measured by CO_2 sensor; $J_{\max}$ is the maximum electron transport rate ($125.66\,\mu\,\mathrm{mol{\cdot}m^{-2}{\cdot}s^{-1}}$); K is a constant describing the conversion efficiency of intercepted PAR (I) to J ($0.425\,\mathrm{mol{\cdot}mol^{-1}}$ photon); and θ is a constant convexity factor describing the response of J to I (0.7) [15].

2.5. Model Calibration and Validation. The structural parameters measured in measurement 2 were used for model

Table 1: Equations and parameter values of organ expansion in Apr-May.

Equation	Dependent variables (y)	Independent variables (x)	Parameter value				R^2
			a	b	k		
$y = a/1 + be^{-kx}$	LL	GDD	26.30	4.21	0.025		0.983
	LW	GDD	31.78	5.38	0.029		0.980
	PD	GDD	7.91	2.61	0.003		0.915
	ID	GDD	8.13	2.81	0.039		0.951
			A_1	A_2	x_0	dx	
$y = A_1 - A_2/1 + e^{(x-x_0)/\mathrm{d}x} + A_2$	PL	GDD	−8.01	32.17	67.08	47.69	0.989
	IL	GDD	−4.09	11.69	20.62	48.84	0.862
	LDW	GDD	−0.09	1.91	90.04	33.05	0.958
	PDW	GDD	−0.04	0.68	124.9	41.35	0.974

Table 2: Equations and parameters of organ expansion in Nov-Dec.

Equation	Dependent variables (y)	Independent variables (x)	Parameter value				R^2
			a	b	k		
$y = a/1 + be^{-kx}$	LL	GDD	29.43	4.21	0.030		0.986
	LW	GDD	33.49	5.40	0.035		0.977
	PD	GDD	7.65	2.28	0.022		0.981
	ID	GDD	7.69	1.98	0.094		0.898
			A_1	A_2	X_0	dx	
$y = A_1 - A_2/1 + e^{(x-x_0)/\mathrm{d}x} + A_2$	PL	GDD	−7.92	28.52	51.72	41.88	0.982
	IL	GDD	1.04	11.33	26.97	12.20	0.802
	LDW	GDD	−0.003	1.38	67.73	22.98	0.964
	PDW	GDD	−0.01	0.44	92.92	26.00	0.984

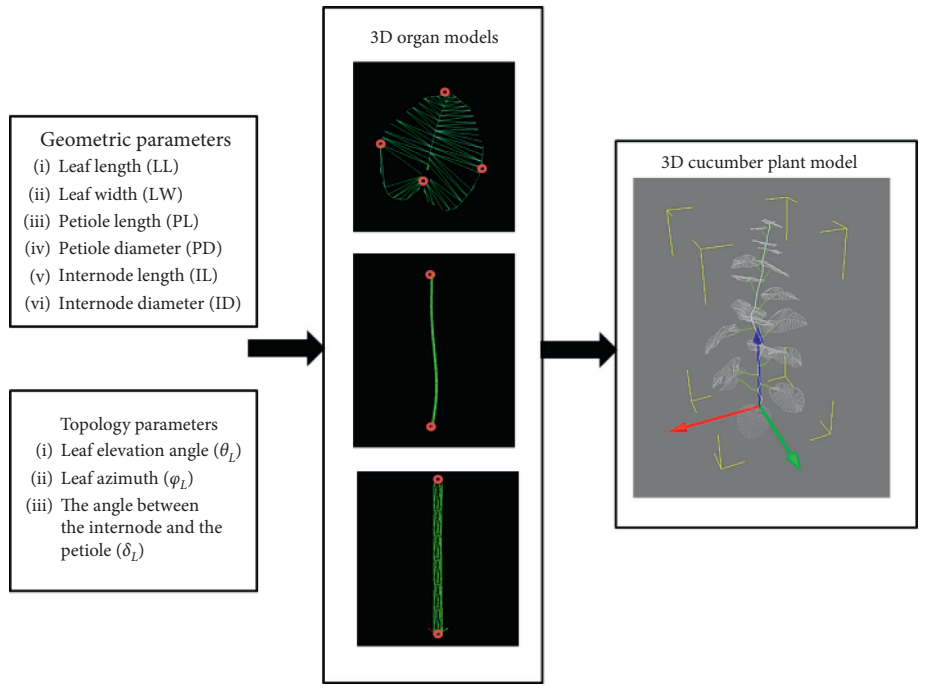


Figure 1: The workflow of reconstruction based on statistical models. The geometric parameters of blade, petiole, and internode came from equations of Tables 1 and 2, while the topology parameters of the angle between organs came from our former study of Qian et al. [18]. The spatial shape of the blade was determined by LL, LW, θ_L, and φ_L; the spatial shape of the petiole was determined by PL and PD and δ_L; the spatial shape of the internode was determined by IL and ID.

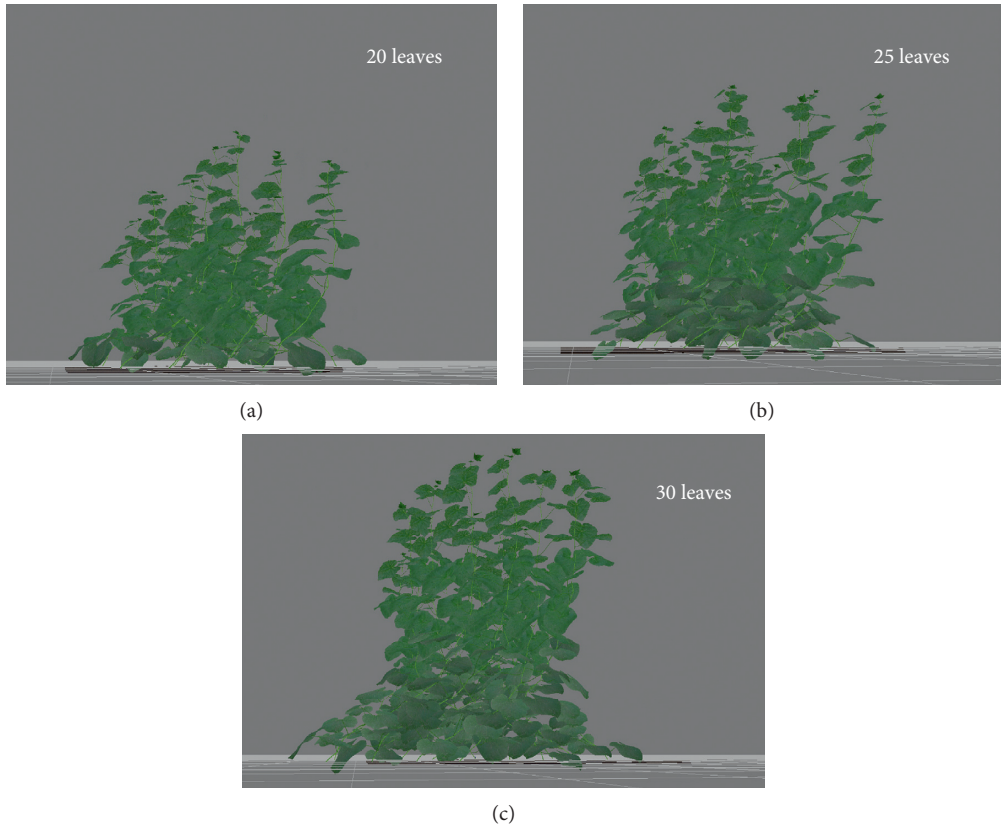


Figure 2: 3D canopies of sixteen cucumbers with 20, 25, and 30 leaves per plant, respectively.

validation. The root-mean-squared deviation (RMSD), bias, and accuracy (%) were determined following the approach described by Kahlen and Stützel [21]:

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{7}$$

$$\text{bias} = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i)^2, \tag{8}$$

$$\text{accuracy} = 1 - \frac{\text{RMSD}}{(1/n)\sum_{i=1}^{n} y_i}, \tag{9}$$

where x_i and y_i are the simulated and measured values, respectively. n is the number of simulated or measured values.

3. Results

3.1. Organ Development Analysis. The growth of cucumber organs conforms to the sigmoid growth curve. The maximum asymptote ($y = \max$), the maximum growth rate point (x_0, y_0), and the slope (y_0') at the inflection point can be used as characteristic values describing its growth process (Figure 3). The curve characteristic values based on the results of measurement 1 are shown in Table 3 and Figure 4. The y_0' of the leaf area expansion curves in Apr-May was 4.58, while the maximum expansion rate of leaf area occurs when the GDD reaches 92.75 Cd (Figure 4(a)). The maximum expansion rate of the petiole and internode length in Apr-May was 0.21 and 0.30, while the maximum expansion rate of the petiole appears at 67 Cd, and the maximum expansion rate of the internode appears at 41 Cd (Figure 4(c)). It indicated that internode reaches the maximum expansion rate before leaf and petiole reaching the maximum expansion rate, and the elongation of the internode is nearly completed. The dry matter accumulation of the petiole was less than that of the leaf, and the maximum values were 0.68 g and 1.91 g, respectively (Figure 4(b)). The dry matter of the leaf was 3 times the weight of the petiole.

The leaf expansion curve of cucumber in Nov-Dec was similar to that of Apr-May, and the leaf length expansion rate was lower than that of leaf width. Different from Apr-May, the maximum leaf expansion rate in Nov-Dec occurred when GDD reached 76 Cd, earlier than Apr-May. The maximum expansion rates of petiole length and internode length were both 0.21. The maximum expansion rate of the petiole appeared at 52 Cd, while the maximum expansion rate of the internode appeared at 27 Cd. In Nov-Dec, the maximum dry matter accumulation of petiole was 0.44 g and 1.38 g, and the maximum dry matter accumulation in winter leaf was significantly smaller than that in Apr-May (Table 3).

3.2. Analysis of Differences in Organ Development in Different Scenarios. The rate of leaf area expansion in Apr-May was

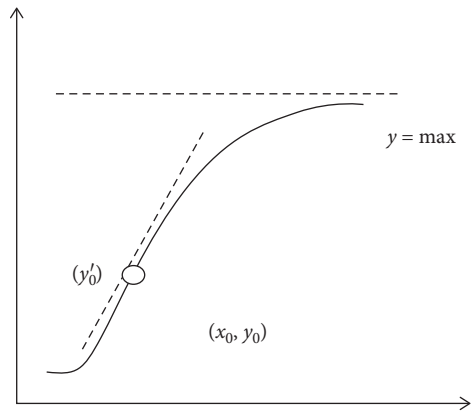


Figure 3: Sigmoid growth curve of organ expansion. $y = \max$ is the asymptote at the top of the curve, which represents the theoretical maximum value of organ expansion; (x_0, y_0) is the inflection point of the curve, which represents the point when the organ expansion reaches the maximum rate; y_0' is the slope at the inflection point and represents the maximum rate of organ expansion.

Table 3: Characteristic parameters of the organ expansion curve.

	Parameters	$y = \max$	y_0'	x_0, y_0
Apr-May	Leaf area	625.95	4.58	92.75, 312.97
	Leaf dry matter	1.91	0.02	90.05, 0.91
	Petiole length	32.17	0.21	67.08 12.08
	Internode length	10.75	0.30	41.92 6.07
	Petiole diameter	7.91	0.04	53.78, 3.95
	Internode diameter	8.13	0.08	26.70, 4.06
Nov-Dec	Leaf area	730.09	6.83	76.58, 365.05
	Leaf dry weight	1.38	0.01	67.73, 0.69
	Petiole length	28.52	0.21	51.72, 10.30
	Internode length	11.33	0.21	26.97, 6.19
	Petiole diameter	7.65	0.04	40.37, 3.82
	Internode diameter	7.69	0.18	7.13, 3.84

less than that in Nov-Dec, while the dry weight of leaves was greater in Apr-May than in Nov-Dec (Figure 4). That is, the leaf dry matter per unit area in Apr-May was greater than that in Nov-Dec. From the changing trend of the dry matter accumulation curve, there was unobvious difference between the dry matter accumulation in the early Apr-May and Nov-Dec, but the rapid growth period of the dry matter accumulation curve in Apr-May was obviously longer than that in Nov-Dec. Therefore, the accumulation ratio of leaf dry weight per unit area was greater in Apr-May than in Nov-Dec (Figure 5). The elongation of the petiole and internode had no significant difference in the two scenarios. Although the diameter change of the petiole and internode was slightly different during the expansion period, the final diameter was still relatively close, with no obvious difference.

3.3. Methodological Accuracy of Organ Expansion Models. The data obtained in experiment 2 was used for model verification. The relationship between the measured and simulated values of leaf length, leaf width, petiole length, and internode length is shown in Figure 6, while the bias, RMSD, and accuracy of the model simplifications are shown in Table 4. The simulated value was close to the measured value, and the accuracy of the four simulations was higher than 0.99.

3.4. Analysis of Leaf Light Interception and Photosynthetic Characteristics in 3D Canopies. Comparing the temperature and light environment conditions between Apr-May and Nov-Dec, the temperature difference was small, but the light difference was large. Since light directly participates in photosynthetic reactions during the accumulation of dry matter, there was a significant difference in yield in Apr-May and Nov-Dec (Figure 7). The yield was 0.174 kg/m^2 per day in Nov-Dec and it was 35.3% lower than the yield in Apr-May. To further analyze the canopy photosynthesis characteristics, it is necessary to combine the canopy radiation interception model and photosynthesis model.

In this study, light interception of three canopies from 8:00 to 15:00 on 4 November 2015 was simulated by using the radiosity-graphic combined model. Combining the 3D structure model and the radiosity-graphic combined model, the instantaneous light interception of each cucumber leaf in the canopy was calculated. The light interception data of the leaves of 4 plants inside the canopy were selected for analysis. The blocking effect of the surroundings was fully considered. The average light interception of a single leaf is shown in Figure 8. The light interception of the middle and lower canopy leaves was significantly lower than that of the upper part, and the light of the top leaves was relatively small due to the small leaf area. The lower leaves in the 20-leaf canopy have lower light interception than the other two canopies. The difference in light interception between the three upper canopy leaves was not obvious.

FvCB photosynthesis model was used in leaf assimilation rate modeling considering carbon dioxide assimilation mechanisms. The light interception of each blade calculated from the light interception simulation is input into the FvCB model as the incident light intensity (I). The FvCB model calculates the assimilation rate of Rubisco-limited (A_c, $\mu mol{\cdot}CO_2{\cdot}m^{-2}{\cdot}s^{-1}$) and RuBP regeneration-limited (A_j, $\mu mol{\cdot}CO_2{\cdot}m^{-2}{\cdot}s^{-1}$), respectively. The data showed that during the daytime in Nov-Dec, the A_j value was always less than the A_c value. It means that the limiting factor affecting the photosynthesis rate in Nov-Dec was light intensity. The high concentration of CO_2 in the morning can also promote photosynthesis (Figure 9). In the morning and afternoon, light intensity was only about half of the noon, while the photosynthesis rate was about 85% of noon. High CO_2 concentration can make up part of the effect of low light intensity on the photosynthetic rate in Nov-Dec. There was no significant difference in the photosynthetic rate of the upper leaves of the three canopies, but affected by the light interception of the leaves, the photosynthetic rate of the lower leaves of the 20-leaf canopy was significantly lower than that of the other two canopies.

4. Discussion

4.1. The Contribution of 3D Model in Simulation Research. In the light interception calculation, the consistency of the 3D model and the true structure of the canopy is a crucial

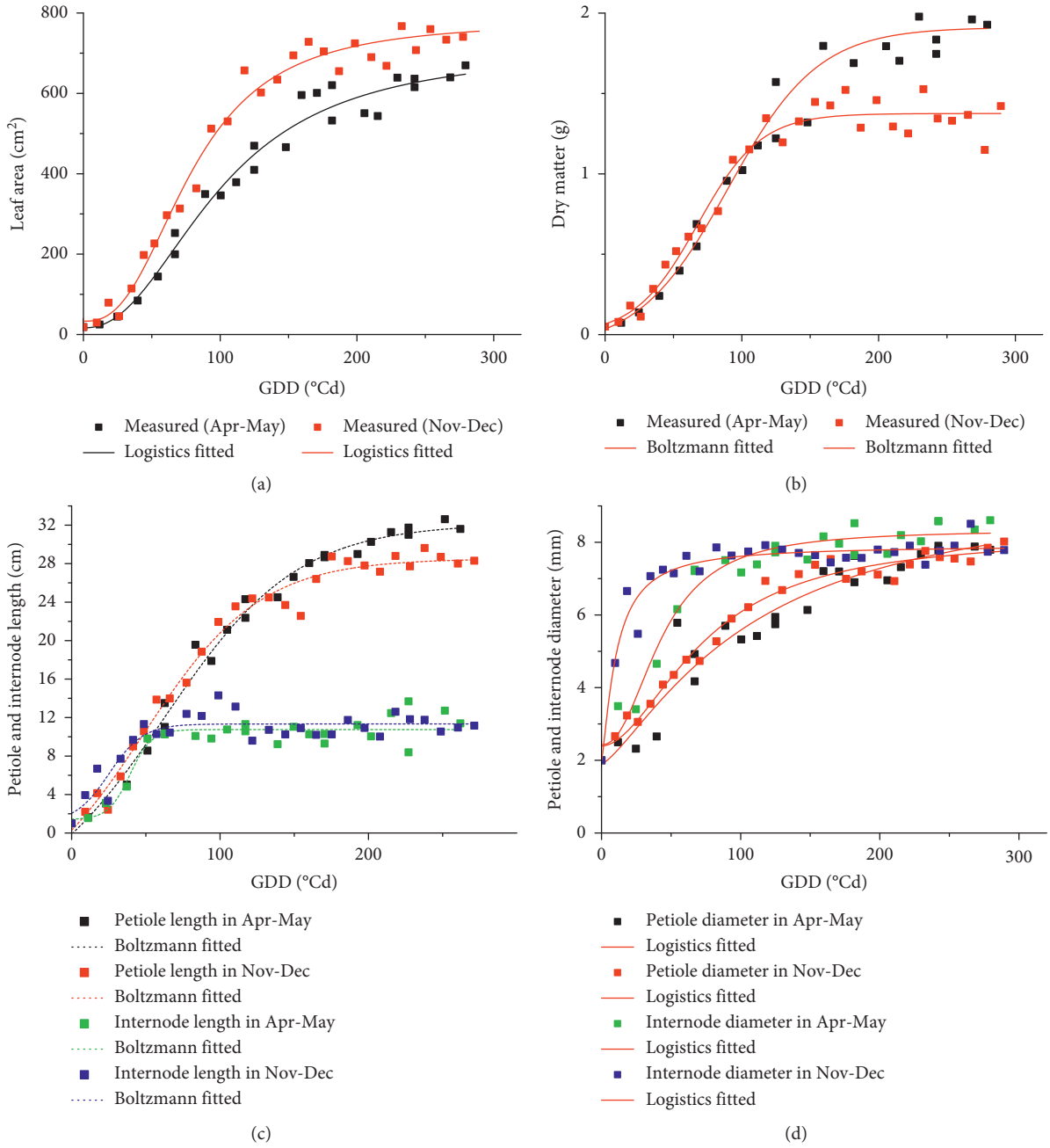


Figure 4: Leaf area expansion (a) and dry matter accumulation (b), petiole and internode length (c) and diameter increasing (d) following thermal time accumulation in Apr-May 2015 and Nov-Dec 2015. Measured data of leaf expansion and petiole and internode diameter increasing were fitted by logistic curves, while measured data of leaf dry matter accumulation and petiole and internode length increasing were fitted by Boltzmann curves.

factor that determines the reliability of the simulation results. At the field level, many research studies were conducted on 3D point cloud acquisition of field crop canopy based on UAV (unmanned aerial vehicle) and LiDAR [22, 23]. Based on these 3D point clouds, plant height, ground cover, and above-ground biomass were obtained [24, 25]. But these methodologies presented an underestimation of biomass and leaf area index and cannot extract the architecture parameters of individual plants from the canopy. In order to obtain more detailed 3D point clouds at the individual plant level, 3D laser scanner, 3D digitizing, and multiview imaging were used in plant and organ 3D point cloud acquisition [26, 27]. Using these methods, many specific morphological parameters of various crops have been achieved, such as leaf azimuthal angle, leaf length, and leaf area of maize [28]. Since the last few years, many 3D point cloud acquisition methods were used in plant 3D structural model construction at different scales, such as LIDAR, 3D laser scanner, 3D digitizing, and multiview imaging [29]. However, there are still many challenges in

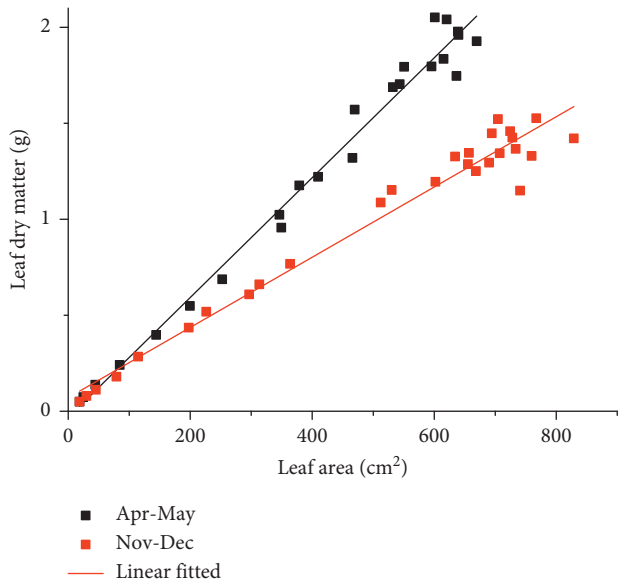


Figure 5: Correlation between leaf dry matter accumulation and leaf area expansion in different scenarios.

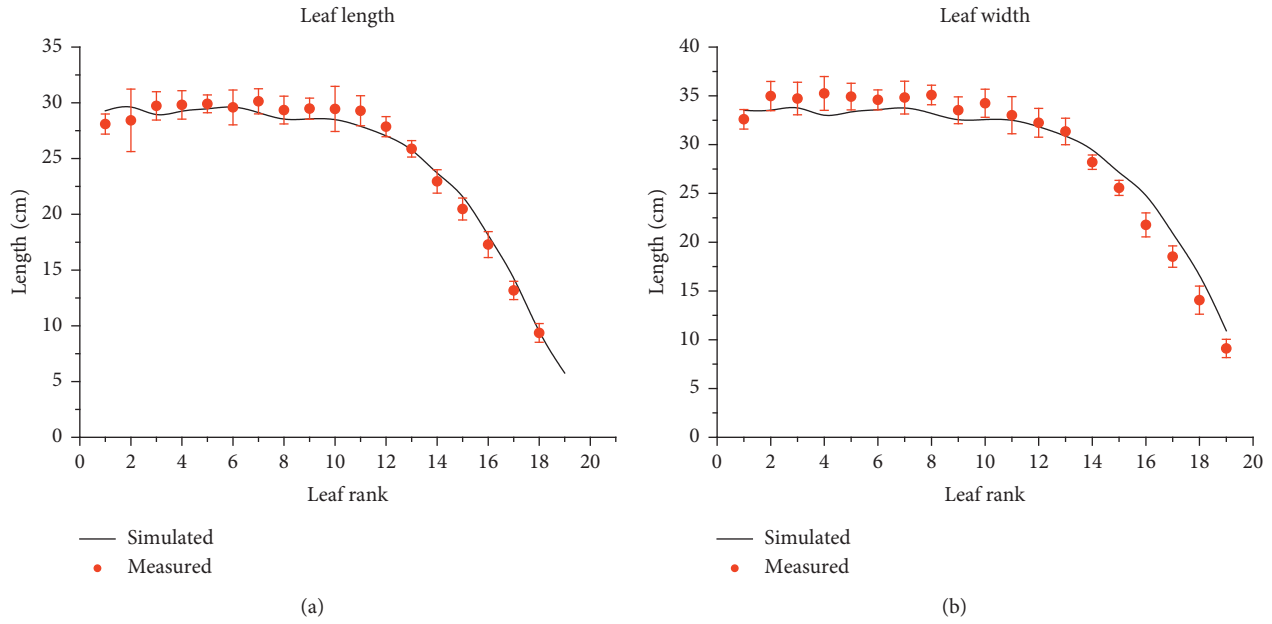


Figure 6: Continued.

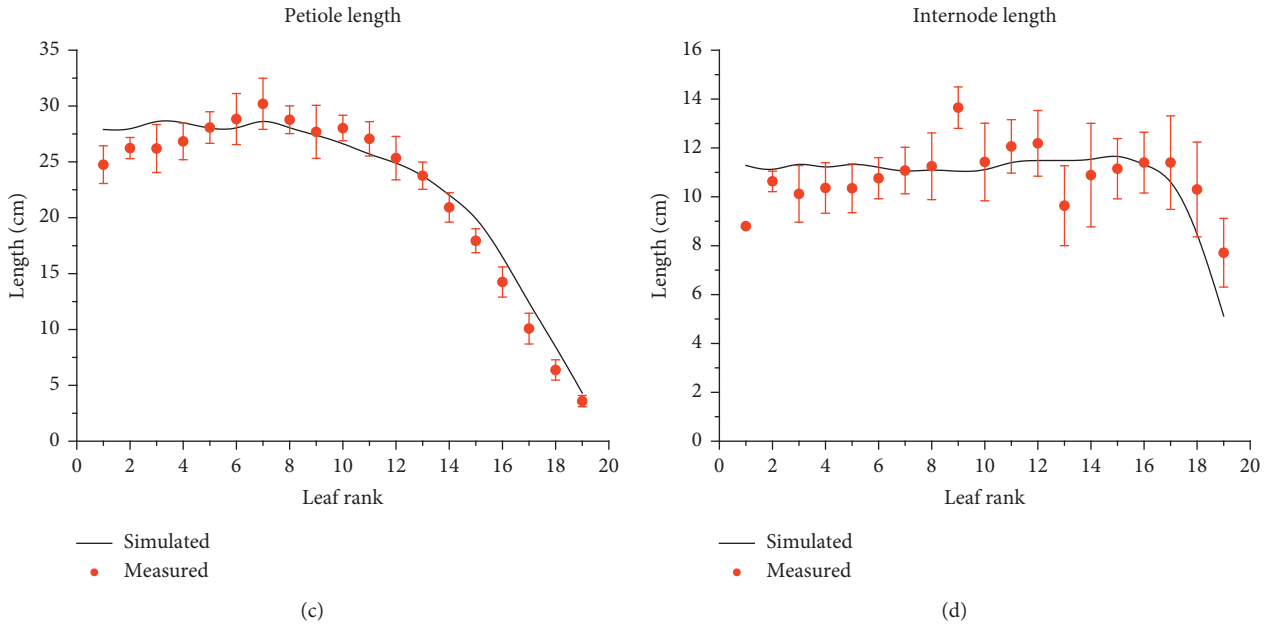


Figure 6: Comparison between the simulated and the measured value of leaf length, leaf width, petiole length, and internode length. Error bars indicate the SD.

Table 4: The bias, RMSD, and accuracy of the model simplifications for leaf length, leaf width, petiole length, and internode length.

Structural parameters	Simulated		
	Bias (cm^2)	RMSD (cm)	Accuracy
Leaf length	−7.14	0.28	0.99
Leaf width	−19.26	0.21	0.99
Petiole length	23.59	0.72	0.99
Internode length	0.97	0.57	0.99

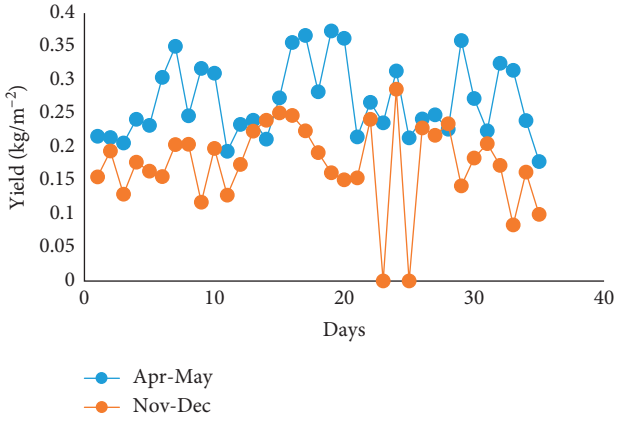


Figure 7: Measured cucumber yield per day in Nov-Dec and Apr-May.

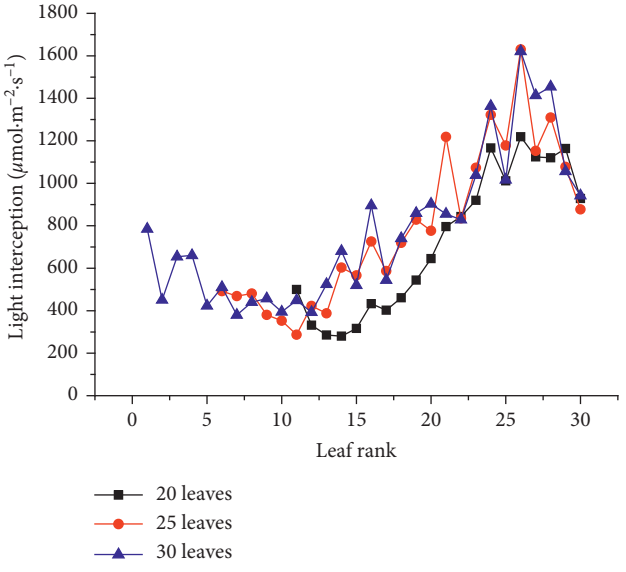


Figure 8: Light interception per leaf in the three canopies.

building a sufficiently accurate 3D model because plant morphologies are diverse for different species, and there is none uniform solution for all kinds of plants [28]. The parametric modeling method applied in this study is also one of the alternative methods for constructing three-dimensional plants. Based on the descriptive model of organ expansion, a plant population with any number stands can be constructed. However, since the organ expansion model is a regression model, its adaptability is limited, and the simulation of the three-dimensional structure model is also constrained by the model. Therefore, we separately constructed the organ expansion curve equations in different scenarios (Apr-May and Nov-Dec) to construct the canopy growth accordingly.

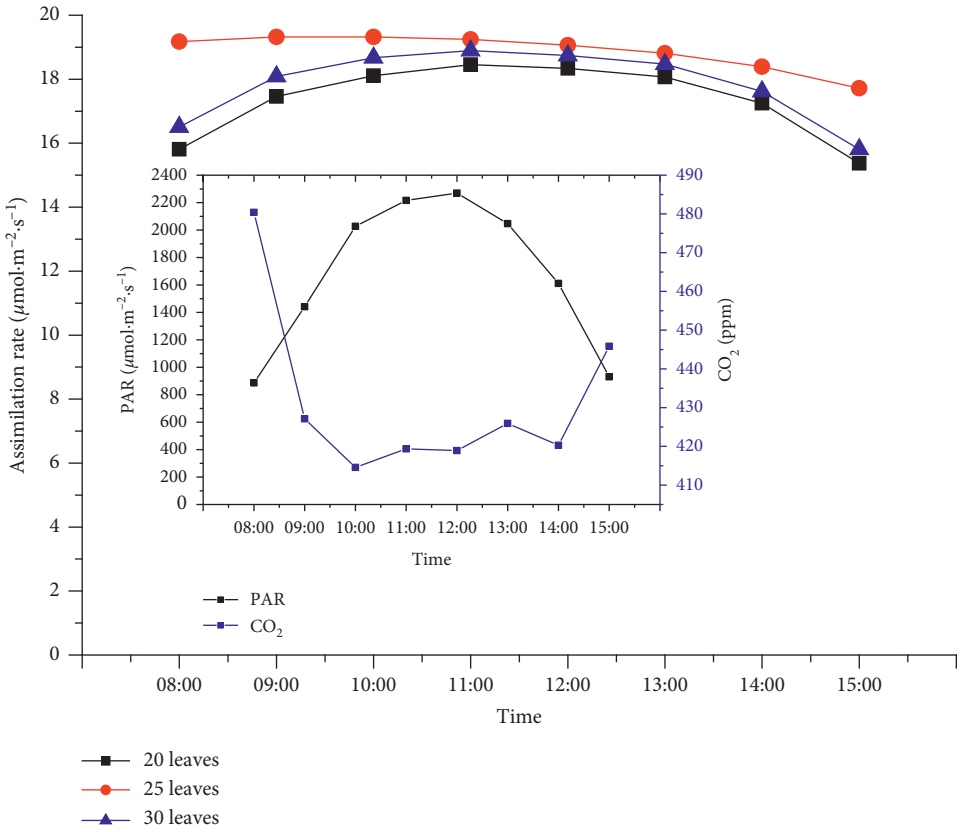


Figure 9: Diurnal variation of carbon dioxide assimilation rate per plant.

4.2. Light Energy on Cucumber Development. The effect of light on plant growth and development not only directly affects the photosynthetic reaction of leaves, but also affects plant morphogenesis. In most of the plant growth models, leaf area is an important parameter directly related to biomass accumulation and structural development [30]. However, the results of this study showed that leaf area is not completely positively related to dry matter. Cucumber grown in winter has larger leaf area but less dry matter, resulting in a smaller specific leaf area (SLA). Therefore, growth is more directly related to specific leaf area than to leaf area [31].

4.3. Photosynthesis Modeling. The photosynthetic capacity of leaves is affected not only by the amount of light intercepted, but also by the age of the leaves and the ambient temperature. In particular, cucumbers belong to C3 type plants, which have stronger photorespiration and growth respiration. They consume more assimilates than C4 plants to maintain growth and protect photosynthetic organs. In many research studies, various forms of equations were used to describe the response of FvCB model parameters to temperature, such as temperature sensitivity (Q10), polynomials, exponential functions, and normal distribution functions [32]. The change in temperature dependence of photosynthesis is caused by several factors such as C_i, $J_{\max}$, and $J_{\max}$ to $V_{c\max}$ ratio, which were also reported to change with growth temperature. The response of each factor to temperature seems to differ among species, and there is no report yet in cucumber. In the research of Zhou et al. [33], it is found that leaf aging significantly affected A_n, A_g, R_{dark}, $V_{c\max}$, $J_{\max}$, R_{dark}/A_n, and g_s, but leaf aging had little effect on the $J_{\max}/V_{c\max}$ ratio, suggesting coordination between $V_{c\max}$ and $J_{\max}$ processes regardless of leaf age. Further studies are needed to reconfirm the leaf aging effect on the temperature sensitivity of R_{dark} and to elucidate the mechanisms involved.

4.4. Optimal Utilization of Light Energy. In a semiclosed greenhouse, the CO_2 concentration in the morning is significantly higher than at noon and afternoon (Figure 9). The main reason is that the respiration of plants at night accumulates a lot of CO_2. From the photosynthesis simulation results, the high concentration of CO_2 in the morning can obviously promote the leaf photosynthesis rate (Figure 9). At this stage, light intensity is the main factor limiting the assimilation rate. By noon, the light intensity increased, but the CO_2 concentration decreased significantly and became the main factor limiting the rate of assimilation. Therefore, increasing light intensity in the morning or supplementing CO_2 at noon can increase photosynthetic output. In the results of leaf light interception and photosynthesis rate analysis (Figures 8 and 10), the light interception and photosynthetic rate of the lower leaves in the 20-leaf canopy were significantly lower than those in the other two canopies. The main reason is that, in the parametric modeling process,

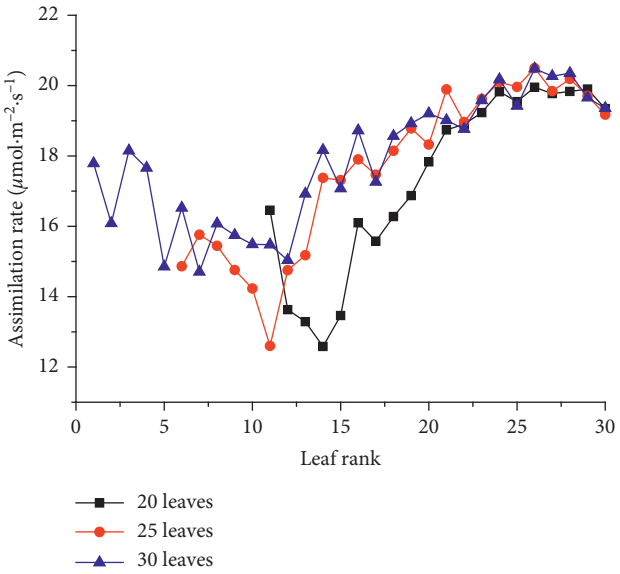


Figure 10: Net assimilation per day of leaves in the three canopies.

the blade azimuth model (φ_L) we choose will make the blade distribution ratio in each azimuth range change with the LAI. As the LAI increases, more leaves will turn orientation towards rows with better lighting conditions to intercept more light. It can be seen that when the light conditions in winter are weak, maintaining a reasonable LAI is more conducive to canopy photosynthesis. From the results of this study, the 25-leaf canopy is the best canopy structure for light interception and photosynthesis. However, considering that the light interception of leaves under 20 leaves is weak and the leaf function is weak due to the effects of leaf senescence, the number of leaves in the canopy during winter should be controlled within 25 leaves.

5. Conclusions

Low temperature and low light energy in the winter of east China are important reasons for the decline in cucumber yield, especially in greenhouses without supplementary light. This study provides an important management basis for the optimal use of light in the greenhouse in winter. Using 3D modeling technology, the virtual structure model of cucumber canopy was established, and combined light interception modeling and photosynthesis modeling, the light environment in the canopy and the photosynthetic performance of leaves were quantitatively studied to obtain cucumber management strategies in winter. By using the FvCB model, Rubisco-limited (A_c) and RuBP regeneration-limited (A_j) assimilation rate of each leaf in three canopies with different leaves per plant were simulated, respectively. Based on the simulated result, we analyzed the effects of light and CO_2 changes on the photosynthesis at different time periods and different leaf positions. Through a comparative analysis of photosynthetic characteristics in the virtual canopy, we have come up with strategies to improve the canopy photosynthetic capacity, and a suitable canopy structure that is conducive to winter canopy photosynthetic yields.

Data Availability

All data used during the study are available from the corresponding author by request (e-mail: qiantingting@saas.sh.cn).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Tingting Qian and Xiuguo Zheng contributed equally to this work.

Acknowledgments

The authors would like to thank Dr. Feng Lu for her support during the preparation of this manuscript. This work was supported by Shanghai Agriculture Applied Technology Development Program of China (no. G2015060402), National Natural Science Foundation of China (no. 61762013), and Shanghai Science and Technology Commission Project of China (no. 19DZ2281300).

References

[1] T. Boulard, J.-C. Roy, J.-B. Pouillard, H. Fatnassi, and A. Grisey, "Modelling of micrometeorology, canopy transpiration and photosynthesis in a closed greenhouse using computational fluid dynamics," *Biosystems Engineering*, vol. 158, pp. 110–133, 2017.

[2] X. Ding, Y. Jiang, D. Hui et al., "Model simulation of cucumber yield and microclimate analysis in a semi-closed greenhouse in China," *HortScience*, vol. 54, no. 3, pp. 547–554, 2019.

[3] L. Marcelis, A. Broekhuijsen, E. Meinen, E. Nijs, and M. Raaphorst, "Quantification of the growth response to light quantity of greenhouse grown crops," in *Proceedings of the International Symposium on Artificial Lighting in Horticulture*, pp. 97–104, Lillehammer, Norway, June 2005.

[4] L. Marcelis, F. Maas, and E. Heuvelink, "The latest developments in the lighting technologies in Dutch horticulture," in *Proceedings of the International ISHS Symposium on Artificial Lighting*, pp. 35–42, Québec City, Canada, November 2000.

[5] T. Hovi, J. Näkkilä, and R. Tahvonen, "Interlighting improves production of year-round cucumber," *Scientia Horticulturae*, vol. 102, no. 3, pp. 283–294, 2004.

[6] X. Hao, X. Guo, X. Chen, and S. Khosla, "Inter-lighting in mini-cucumbers: interactions with overhead lighting and plant density," *Acta Horticulturae*, vol. 1107, no. 1107, pp. 291–296, 2015.

[7] X. Hao, X. Guo, J. Lanoue et al., "A review on smart application of supplemental lighting in greenhouse fruiting vegetable production," *Acta Horticulturae*, no. 1227, pp. 499–506, 2017.

[8] J. H. Kim, J. W. Lee, T. I. Ahn, J. H. Shin, K. S. Park, and J. E. Son, "Sweet pepper (*Capsicum annuum* l.) canopy photosynthesis modeling using 3D plant architecture and light ray-tracing," *Frontiers in Plant Science*, vol. 7, p. 1321, 2016.

[9] T.-T. Qian, S.-L. Lu, C.-J. Zhao, X.-Y. Guo, W.-L. Wen, and J.-J. Du, "Heterogeneity analysis of cucumber canopy in the solar greenhouse," *Journal of Integrative Agriculture*, vol. 13, no. 12, pp. 2645–2655, 2014.

[10] W. Wen, X. Guo, B. Li et al., "Estimating canopy gap fraction and diffuse light interception in 3D maize canopy using hierarchical hemispheres," *Agricultural and Forest Meteorology*, vol. 276-277, Article ID 107594, 2019.

[11] A. J. Burgess, R. Retkute, T. Herman, and E. H. Murchie, "Exploring relationships between canopy architecture, light distribution, and photosynthesis in contrasting rice genotypes using 3D canopy reconstruction," *Frontiers in Plant Science*, vol. 8, p. 734, 2017.

[12] R. P. A. Perez, C. Fournier, L. Cabrera-Bosquet et al., "Changes in the vertical distribution of leaf area enhanced light interception efficiency in maize over generations of selection," *Plant, Cell & Environment*, vol. 42, no. 7, pp. 2105–2119, 2019.

[13] L. Tang, D. Yin, and C. Chen, "Optimal design of plant canopy based on light interception: a case study with loquat," *Frontiers in Plant Science*, vol. 10, p. 364, 2019.

[14] G. Buck-Sorlin, P. H. B. de Visser, M. Henke et al., "Towards a functional-structural plant model of cut-rose: simulation of light environment, light absorption, photosynthesis and interference with the plant structure," *Annals of Botany*, vol. 108, no. 6, pp. 1121–1134, 2011.

[15] T.-W. Chen, M. Henke, P. H. B. de Visser et al., "What is the most prominent factor limiting photosynthesis in different layers of a greenhouse cucumber canopy?" *Annals of Botany*, vol. 114, no. 4, pp. 677–688, 2014.

[16] R. P. A. Perez, J. Dauzat, B. Pallas et al., "Designing oil palm architectural ideotypes for optimal light interception and carbon assimilation through a sensitivity analysis of leaf traits," *Annals of Botany*, vol. 121, no. 5, pp. 909–926, 2017.

[17] Y. Y. Cho, S. Oh, M. M. Oh, and J. E. Son, "Estimation of individual leaf area, fresh weight, and dry weight of hydroponically grown cucumbers *(Cucumis sativus* l.) using leaf length, width, and SPAD value," *Scientia Horticulturae*, vol. 111, no. 4, pp. 330–334, 2007.

[18] T. Qian, X. Zheng, X. Guo, W. Wen, J. Yang, and S. Lu, "Influence of temperature and light gradient on leaf arrangement and geometry in cucumber canopies: structural phenotyping analysis and modelling," *Information Processing in Agriculture*, vol. 6, no. 2, pp. 224–232, 2019.

[19] S. Lu, C. Zhao, X. Guo, W. Wen, and B. Xiao, "A new paradigm for fast interactive design of crops," *Intelligent Automation and Soft Computing*, vol. 16, pp. 1147–1155, 2010.

[20] G. D. Farquhar, S. von Caemmerer, and J. A. Berry, "A biochemical model of photosynthetic CO_2 assimilation in leaves of C3 species," *Planta*, vol. 149, no. 1, pp. 78–90, 1980.

[21] K. Kahlen and H. Stützel, "Modelling photo-modulated internode elongation in growing glasshouse cucumber canopies," *New Phytologist*, vol. 190, no. 3, pp. 697–708, 2011.

[22] L. Lei, C. Qiu, Z. Li et al., "Effect of leaf occlusion on leaf area index inversion of maize using UAV-LiDAR data," *Remote Sensing*, vol. 11, no. 9, p. 1067, 2019.

[23] R. P. A. Perez, E. Costes, F. Thévenу, S. Griffon, J.-P. Caliman, and J. Dauzat, "3D plant model assessed by terrestrial LiDAR and hemispherical photographs: a useful tool for comparing light interception among oil palm progenies," *Agricultural and Forest Meteorology*, vol. 249, pp. 250–263, 2018.

[24] J. A. Jimenez-Berni, D. M. Deery, P. Rozas-Larraondo et al., "High throughput determination of plant height, ground cover, and above-ground biomass in wheat with LiDAR," *Frontiers in Plant Science*, vol. 9, p. 237, 2018.

[25] S. Madec, F. Baret, B. de Solan et al., "High-throughput phenotyping of plant height: comparing unmanned aerial vehicles and ground LiDAR estimates," *Frontiers in Plant Science*, vol. 8, p. 2002, 2017.

[26] B. Elnashef, S. Filin, and R. N. Lati, "Tensor-based classification and segmentation of three-dimensional point clouds for organ-level plant phenotyping and growth analysis," *Computers and Electronics in Agriculture*, vol. 156, pp. 51–61, 2019.

[27] S. Wu, W. Wen, B. Xiao et al., "An accurate skeleton extraction approach from 3D point clouds of maize plants," *Frontiers in Plant Science*, vol. 10, p. 248, 2019.

[28] Y. Wang, W. Wen, S. Wu et al., "Maize plant phenotyping: comparing 3D laser scanning, multi-view stereo reconstruction, and 3D digitizing estimates," *Remote Sensing*, vol. 11, no. 1, p. 63, 2018.

[29] M. Vázquez-Arellano, H. Griepentrog, D. Reiser, and D. Paraforos, "3-D imaging systems for agricultural applications-a review," *Sensors*, vol. 16, no. 5, p. 618, 2016.

[30] L. F. M. Marcelis, E. Heuvelink, and J. Goudriaan, "Modelling biomass production and yield of horticultural crops: a review," *Scientia Horticulturae*, vol. 74, no. 1-2, pp. 83–111, 1998.

[31] S. M. Weraduwage, J. Chen, F. C. Anozie, A. Morales, S. E. Weise, and T. D. Sharkey, "The relationship between leaf area growth and biomass accumulation in *Arabidopsis thaliana*," *Frontiers in Plant Science*, vol. 6, 2015.

[32] K. Hikosaka, K. Ishikawa, A. Borjigidai, O. Muller, and Y. Onoda, "Temperature acclimation of photosynthesis: mechanisms involved in the changes in temperature dependence of photosynthetic rate," *Journal of Experimental Botany*, vol. 57, no. 2, pp. 291–302, 2005.

[33] H. Zhou, M. Xu, H. Pan, and X. Yu, "Leaf-age effects on temperature responses of photosynthesis and respiration of an alpine oak, *Quercus aquifolioides*, in southwestern China," *Tree Physiology*, vol. 35, no. 11, pp. 1236–1248, 2015.

Hindawi
Scientific Programming
Volume 2020, Article ID 8841006, 10 pages
https://doi.org/10.1155/2020/8841006

Research Article

A Bandwidth Statistical Multiplexing and Control Method for Satellite Broadcasting

Yingqiang Wang,[1] **Zhaohua Nian**,[1] **Chang Liu**,[1] **Wei Han**,[1] **and Maowei Lin**[2]

[1]*Space Star Technology Co., Ltd., Beijing, China*
[2]*China Meteorological Administration, Beijing, China*

Correspondence should be addressed to Yingqiang Wang; wangyq@spacestar.com.cn

Received 8 April 2020; Revised 14 June 2020; Accepted 7 July 2020; Published 1 August 2020

Academic Editor: Chenxi Huang

In the field of satellite data broadcasting, the management quality of data broadcasting bandwidth is directly related to the throughput of the broadcasting system and plays an important role in the performance of satellites. In this paper, for the sun-synchronous orbit meteorological satellite broadcasting which has the conventional product files and emergency information, a broadcast bandwidth statistical multiplexing and control method is designed for bandwidth management. It can be used for the management of broadcasting between regular products and emergency information, as well as internal broadcasting among regular products. This paper is the first to apply common multiplexing of PID and channel mode (CMPCM) to satellite broadcasting. The test verified that the broadcast channel of the parameters and the broadcast schedule management channel resources achieved statistical multiplexing of bandwidth, ratio of channel management functions, and data broadcast control. Broadcasting occupation ratio (BOR) and broadcasting file error ratio (BER) improved significantly. This is significant for improving the efficiency of satellite uplink broadcasting.

1. Introduction

The weather is playing an increasingly important role whether in people's daily life or in the operation of the entire national economy. Meteorological disasters cause huge losses to people's lives, property, and national economic development every year.

Now with the development of China's meteorological service, meteorological disaster monitoring, and early warning process, the satellite data broadcasting system as a whole is an important subsystem in the system; it is not only an important domestic and foreign meteorological data transmission distribution system, but also the ground platform for the distribution of observational data, radar data, satellite data, and other data, as well as the broadcasting and transmission of regular meteorological products and emergency information. The satellite data broadcasting system can provide FENGYUN series of meteorological satellite observation data and product broadcasts for different levels of meteorological departments and users in the Asia-Pacific region. The data and products are formatted, packaged, modulated, upconverted, and amplified in real time. The satellite data broadcasting system operates automatically under the bidirectional driving of data and mission schedules and has a highly reliable, all-weather service operation capability.

Broadcast bandwidth is an important scarce resource for satellites especially for geostationary satellite. The total bandwidth of geostationary satellite transponder's link is very limited [1]. Compared with the speed of sun-synchronous orbit sensing satellites, which is often hundreds of megabytes or even several gigabytes, its rates are usually only several megabytes to tens of megabytes. For continuous broadcasting, even $7 * 24$ hours, broadcast method becomes very important. Therefore, whether the data broadcasting system can reasonably utilize the satellite bandwidth becomes particularly important. The level of data broadcasting efficiency is the key to the system's ability to complete the

daily broadcasting task. How to use these tens of megabytes of data bandwidth efficiently is an important content in the development of satellite data broadcasting systems and plays a key role in the application of meteorology.

In order to cope with the signal delay of the satellite link, Donner et al. [2] proposed and analyzed a suitable solution based on the use of a performance enhancing proxy (PEP) and header compression for voice over IP (VoIP) streams to save satellite bandwidth. Park and Kim proposed an adaptive bandwidth allocation scheme according to traffic demands and weather-induced signal attenuation, which can achieve a reasonable solution between the maximization of total capacity and the support of proportional fairness among the beams for satellite downlinks [3]. In [4], Lei and Vazquez-Castro proposed an analytical solution for the optimal carrier allocation and improved it in terms of power gain, spectral efficiency, and traffic matching ratio compared with the conventional system, which is designed based on uniform bandwidth and power allocation. The issue of multibeam power allocation based on traffic demands and channel conditions over satellite downlinks with power and delay constraints is addressed [5]. Li et al. [6] investigated an energy harvesting scheme in a smart grid based on the cognitive relay protocol, developing the analytically optimal bandwidth allocation strategy to maximize the total sum rate of the grid network. Bahnasse et al.'s article contains information on a new intelligent bandwidth allocation model for future network (Smart Allocation) [7]. The included data describe the topology of the network testbed and the obtained results.

Shi et al. proposed two joint power and bandwidth allocation schemes [8] based on the intercluster orthogonal frequency reuse, which considers channel conditions and the traffic demands of user beams. It gives a good compromise between efficiency of frequency bands utilization and cochannel interference mitigation. The joint bandwidth and power allocation iterative algorithm based on duality theory was proposed [9] to obtain the optimal solution of satellite power and bandwidth resources in a multi-spot-beam satellite communication system. A joint power and bandwidth allocation with QoS support algorithm was proposed by Miao [10] using convex optimization method for heterogeneous wireless networks. It satisfied the minimum rate constraint of delay-constraint service traffic and proportional fairness of best-effort service traffic. With the help of the classical multiknapsack problem, heuristic algorithms together with some innovative scaling schemes [11] are presented for the optimization of resource allocation in systems using DVB-S2 links.

These conclusions cannot be directly extended to the meteorological satellite broadcasting, especially for data broadcasting. In satellite broadcasting systems, logical channel is used to manage different types of broadcast data. The types of information in different broadcast channels have different requirements on broadcast timeliness, which can spread minutes to hours. At the same time, the allocation of bandwidth needs to change dynamically with time and be adjusted in real time. The bandwidth of the channel which has no unbroadcast data needs to release resources in time for reuse by other channels.

2. Demand Introduction

Bandwidth reuse statistics are one of the most important functions in a data broadcasting system. Bandwidth reuse statistics function has an important influence on whether the precious satellite channel resources can be fully utilized [12], whether the weather data can reach the destination for the first time, and whether data is correctly transmitted. The system must transmit meteorological regular files and emergency information at the same time and ensure that the bandwidth of the satellite is fully utilized as far as possible. Control of content delivery needs to be prioritized, to ensure that urgent tasks can be issued as soon as possible. The total rate of satellite data broadcasting should reach or approach the maximum bandwidth of the satellite channel as much as possible and make full use of the satellite channel resources, so that all types of meteorological data can reach the receiving station in the shortest time.

Based on the concepts of channel maximum bandwidth, minimum bandwidth, and channel priority, CMACast broadcasting system designed and implemented the bandwidth allocation strategy mode which changes with time [13, 14], which has guiding significance for this paper.

2.1. Functional Requirements. The data that the satellite broadcasting system needs to distribute belongs to multiple different PIDs and channels, which are at different data levels and are subject to different priorities and bandwidth restrictions. To achieve high-performance broadcasting, the system must use an optimized bandwidth statistics algorithm and adjust the broadcast speed of each data packet in real time to make it reach or approach the theoretical broadcast upper limit.

File channels involved in this article include 5 categories: HRIT-1 (high rate information transport 1), HRIT-2, HRIT-3, LRIT (low rate information transport), and EWAIB (emergency weather alert information broadcast). Among them, HRIT-1 and HRIT-3 include regular files and emergency information, HRIT-2 contains regular files only, and EWAIB contains emergency information [15]. Table 1 shows detailed information.

Data broadcasting system software completes the configuration of the parameters of the channel, scheduling management, and broadcasting channel resources and realizes the bandwidth statistical multiplexing function, the channel ratio management function, and the data broadcasting control function. The specific functions are as follows:

(1) *Bandwidth Statistical Multiplexing.* The bandwidth statistical multiplexing function can adjust the broadcast rate of each channel in real time according to the set bandwidth statistical multiplexing strategy and the priority of each channel, as well as the highest and lowest bandwidth parameters during the data broadcast process. It can also allocate

Table 1: Channel and types.

Channel	Content	Type
HRIT-1	LMI level 2 products	Emergency
	AGRI level 1 products	Regular
	HRIT-1 timetable	Regular
	Satellite information	Regular
HRIT-2	GIIRS level 2 products	Regular
	AGRI level 2 products	Regular
	HRIT-2 timetable	Regular
	Satellite information	Regular
HRIT-3	AGRI level 1 products	Regular
	LMI level 2 products	Emergency
	HRIT-3 timetable	Regular
	Satellite information	Regular
LRIT	AGRI LRIT data	Regular
	Satellite information	Regular
	LRIT timetable	Regular
EWAIB	Emergency information	Emergency

conventional broadcast channels and emergency information broadcast. The transmission rate between channels guarantees the rational use of satellite channel bandwidth.

(2) *Channel Ratio Management.* Channel ratio management function can be calculated according to the statistical multiplexing of bandwidth theoretical rate channel for each file and the respective channel number data for the mixing process.

(3) *Data Broadcast Control.* Data broadcast control function may control the broadcast code rate for each broadcast's threads, to ensure that thread data code rate can be broadcasted as expected.

Bandwidth statistical multiplexing includes statistical multiplexing among regular file channels and statistical multiplexing between regular files and emergency information. The statistical multiplexing among the regular file channels mainly calculates the theoretical rate of each channel in real time according to the priority of the file channel, the highest (lowest) bandwidth setting, and the dynamic changes of the channel data. The broadcast of regular files gives the bandwidth needed for emergency information broadcast through its internal statistical reuse by reducing the total code rate of regular file broadcast. When the emergency information broadcast completes, the regular file broadcast rate is automatically restored.

The function of bandwidth multiplexing statistics is calculating the real-time broadcast occupied bandwidth of each channel and reallocating the remaining bandwidth according to the priority of each channel. While achieving dynamic balance, the total bandwidth will occupy the satellite channel bandwidth as much as possible, so that the entire data broadcasting performance is optimal.

Bandwidth statistical multiplexing and speed control are a feedback control process, in which regular data and emergency data adjust real-time broadcasting speed due to available bandwidth, as shown in Figure 1.

Emergency information and regular file channel information are combined for bandwidth multiplexing and

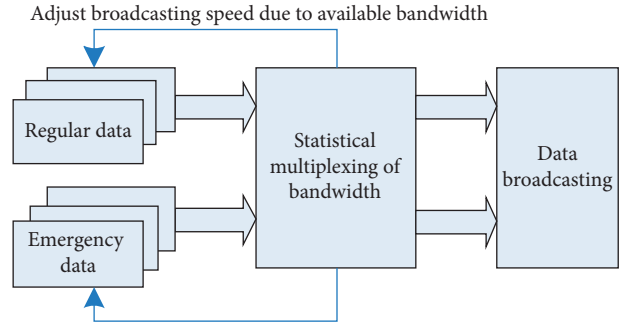


Figure 1: Bandwidth statistical multiplexing process control.

control, which requires priority to ensure that emergency information channels have sufficient bandwidth, and then the remaining bandwidth is allocated among each regular file channel.

2.2. Nonfunctional Requirements. For bandwidth statistics and multiplexing control method, the broadcasting system needs to achieve the following indicators:

(1) Speed of broadcast for emergency messages and general files is 12 Mbps for HRIT-1, 10 Mbps for HRIT-2, 2 Mbps for HRIT-3, 256 Kbps for LRIT, and 2 Kbps for EWAIB.

(2) The utilization rate of bandwidth is more than 96%. In an independent link, the bandwidth sum of different PIDs and channels should not be less than 96% of the total bandwidth.

(3) Meanwhile, file security also needs to be satisfied, and broadcasting success rate should be 100%. Each broadcasting file must be completely processed, and the data received through the small station are exactly the same as the data before broadcasting.

3. System Design

The data broadcasting system provides the principle of statistical multiplexing and priority setting among different broadcasting channels for satellite data broadcasting, which can easily set the channel rate and priority. When there is channel access and exit, the speed of other channels can be automatically adjusted according to the current bandwidth usage and the priority of different channels to achieve the full use of satellite channel bandwidth resources.

For the design of bandwidth statistical multiplexing, factors that need to be considered include the following:

(1) *The Use of Satellite Channel Bandwidth Should Be Maximized.* In the process of statistical multiplexing, the first thing considered is that the total rate of satellite data broadcasting should reach or approach the maximum rate allowed by the satellite channel as much as possible, make full use of the satellite

channel, and be able to broadcast meteorological products at the fastest speed.

(2) *During the Process of Bandwidth Statistical Multiplexing, the Total Out Bandwidth is Stable and Free from Fluctuations.* During statistical multiplexing, it is necessary to ensure that the overall rate of satellite data broadcasting is stable and that the instantaneous total code rate will not increase or decrease due to the rate adjustment of regular files or emergency information, resulting in data loss or waste of bandwidth.

(3) *Bandwidth Statistics Multiplex Response Is Fast and Time-Effective.* The purpose of bandwidth multiplexing statistics is to adjust the code rate of related channels when the channel that broadcasts regular files or emergency information changes, to ensure continuous and stable broadcasting of data on each channel. Therefore, bandwidth multiplexing statistics need to ensure high timeliness and respond quickly to changes in the status of each channel.

(4) *Bandwidth Statistical Multiplexing Does not Affect the Accuracy of the Data.* In the process of bandwidth statistical multiplexing, changing the code rate of each channel will not cause data broadcast errors or data loss due to improper code rate control.

With a full understanding of the requirements and principles of bandwidth statistical multiplexing, this paper has designed a variety of algorithms for bandwidth statistical multiplexing mechanisms, which can make all regular files or emergency information channels use the same bandwidth statistical multiplexing mechanism. Based on a unified priority, each file channel and each emergency information channel are allocated bandwidth according to different mechanisms.

A channel is a logical division of data. When data are finally sent, different channels data are combined into one channel and sent out in the form of network data packets. Bandwidth (code rate) of each channel is expressed as a statistical average rate over a period of time. In implementation, the rate setting of the channel adopts network packets to match the proportion of the code rate of the channel for the entire bandwidth. The specific algorithm is as follows:

(i) Let the rate of N channels be $\mathrm{BR}_1 - \mathrm{BR}_n$ and the total rate be $\mathrm{BR} = \sum \mathrm{BR}_i$, the proportion of each channel in the total bandwidth is

$$b_i = \frac{\mathrm{BR}_i}{\mathrm{BR}}. \quad (1)$$

(ii) It has m network packets as a working period.

(iii) In each period, every channel participates based on the whole package.

(iv) Channel i sends P_i packets in a matching period:

$$P_i = [M \times b_i + 0.5]. \quad (2)$$

The flow of bandwidth statistical multiplexing and channel bandwidth control is shown in Figure 2.

4. Internal Bandwidth Multiplexing among Regular Files

In this paper, according to the characteristics of regular files, without considering PID, the bandwidth statistical multiplexing among channels is mainly designed as follows.

4.1. Absolute Priority. The principle of the absolute priority algorithm is that when bandwidth multiplexing is performed, allocation is completely based on priority. When bandwidth is allocated, the channel with the highest priority is first allocated to the maximum bandwidth, and the remaining bandwidth is allocated to the next priority.

4.2. Priority Weight Allocation Mechanism. The priority weight allocation mechanism is an improvement strategy of the absolute priority mechanism. The absolute priority mechanism is only related to the level of priority, and the specific value of the priority has no practical significance for the mechanism. In the priority weight allocation mechanism, each priority has a weight, which indicates the proportion of bandwidth allocated under the priority. When performing actual multiplexing statistics, the value of the priority is quantified as a weight. Channel weights are allocated for each channel bandwidth.

4.3. Bandwidth Allocation Set by Channel Mechanism. The bandwidth allocation algorithm set according to the channel is similar to the prior weight allocation mechanism. The difference is that the weight of the bandwidth allocation is determined by the set channel bandwidth. The bandwidth allocation is based on the priority from high to low. Channel weights are allocated for each channel bandwidth.

Obviously, the characteristics of the three mechanisms are as follows:

(i) Based on the absolute priority mechanism, a channel with a higher priority can obtain high bandwidth, and a channel with a lower priority can obtain additional bandwidth only when all other channels have obtained the highest bandwidth.

(ii) Based on the priority weight allocation mechanism, the remaining bandwidth can be allocated to each channel in proportion. Even for channels with low priority, additional bandwidth can be obtained by setting the weight.

(iii) Based on the bandwidth ratio allocation set by channels mechanism, the actual rate of each channel approaches its maximum bandwidth when the bandwidth is relatively sufficient.

In summary, when PID is not considered, each mechanism has its own advantages. When the system is running, different priority mechanisms can be selected to meet the

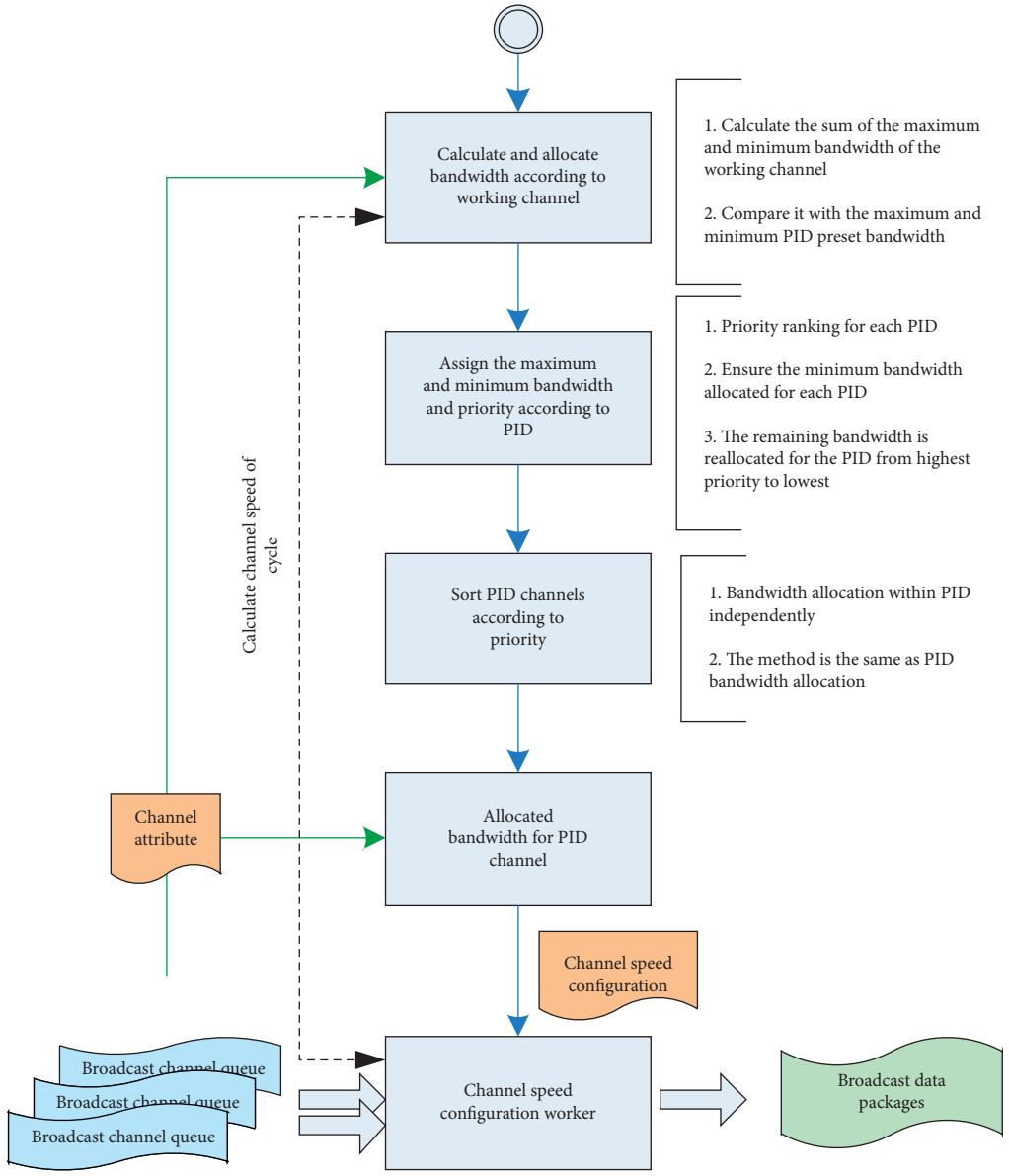


Figure 2: Bandwidth and statistical multiplexing channel control process.

needs of data broadcast. At the same time, it may customize the priority mechanism differently from the above strategy.

After considering the system PID, channel statistical multiplexing has two-level statistical multiplexing, i.e., PID and channel. In order to adapt to different broadcast requirements, this scheme designs two kinds of bandwidth statistical multiplexing modes between PID and channel, which are described as follows.

4.4. PID First and Then Channel Mode (PFM). PID first and then channel refers to first calculating the theoretical rate of each PID according to the bandwidth range, priority, and statistical multiplexing strategy set by each PID and then calculating the theoretical rate of all channels under the PID according to the theoretical bandwidth of each PID. In this mode, the channels between PIDs do not directly participate in multiplexing statistics and can control the rate of each PID better.

4.5. Common Multiplexing of PID and Channel Mode (CMPCM). This mode does not calculate the PID rate separately but instead substitutes the priority of the PID into the channel below it for global multiplexing statistics. This method enables higher priority channels to obtain higher rates, without being constrained by the PID bandwidth range.

The following is experimental tests for the two bandwidth statistical multiplexing methods designed. The configuration information of the participating devices is shown in Table 2.

Table 2: Bandwidth multiplexing test channel configurations.

Name	Model	Configuration
Broadcast server	Inspur NF5280M4	CPU: E5-8860 * 4 Memory: 16 GB OS: SUSE x64
Data modulator	Newtec M6100	
Demodulator	Newtec AZ910	
Receiver	HP Z820	CPU: Xeon E5-2620 * 2 Memory: 4 GB OS: Windows 7

Table 3: The PID bandwidth and priorities.

PID	Minimum bandwidth	Maximum bandwidth	Priority
PID1	26	64	1
PID2	26	64	2

In the experiment, two PIDs were designed, with two channels under each PID, and a total bandwidth of 64 Mbps was set. The specific configurations of the PID and channels are shown in Tables 3 and 4.

The test results of the actual broadcast rate of each PID and channel after statistical bandwidth multiplexing in different ways are shown in Figure 3, and the statistical multiplexing algorithm uses an absolute priority mechanism.

From the occupation of the total bandwidth of the four channels, the following conclusions can be drawn:

(i) First PID after passage way for PID theoretical rate was well controlled at the PID level to ensure PID bandwidth, so that each PID having high priority within the channel gets extra bandwidth.

(ii) PID priority is taken into account in CMPCM. PID priority is higher than channel priority. In this experiment, the order is $A > B > C > D$.

5. Bandwidth Multiplexing between Regular Files an Emergency Information

In order to ensure the timely broadcasting of emergency information, the bandwidth of emergency information must be guaranteed first for the statistical multiplexing between regular files and emergency information. According to this principle, this paper designs the following three statistical multiplexing algorithms between regular files and emergency information.

5.1. No Delay Mechanism (NDM). In order to guarantee that receiving station can receive and process emergency information broadcasted by the master station, master station broadcasts information instantly without delay and buffering.

However, because the code rate of emergency information may not be obtained in advance, in order to ensure that the maximum broadcast bandwidth does not exceed the channel bandwidth capacity, when the master station starts to receive and forward emergency information, it needs to allocate the emergency information bandwidth according to the set maximum bandwidth of the emergency information channel. After the master station software calculates the emergency information, it needs to release the emergency information highest bandwidth and the difference between the emergency information and the actual bandwidth.

In this way, there is no time delay from the master station receiving the emergency information data and it can almost start broadcasting, but this may result in a short period of channel bandwidth waste due to calculating the emergency information data symbol rate. If it is receiving data from the modulator encoder, the data broadcasting system software can obtain the modulator setting rate in advance and adjust the maximum broadcast bandwidth accordingly, which can effectively avoid bandwidth waste.

Table 4: The PID channel bandwidth and priorities.

PID	Channel number	Minimum bandwidth (Mbps)	Maximum bandwidth (Mbps)	Priority
PID1	Channel A	8	32	1
	Channel B	8	32	2
PID2	Channel C	8	32	1
	Channel D	8	32	2

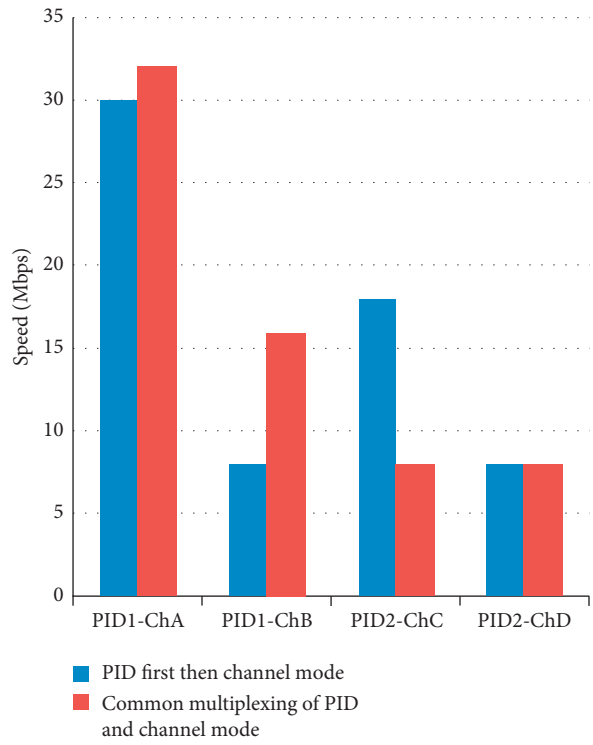


Figure 3: Two bandwidth occupation modes.

5.2. No Bandwidth Wasting Mechanism (NBWM). No channel bandwidth wasting mechanism is different from the processing flow at the beginning. When receiving emergency information, the master station broadcast is not performed,

but rather the data is buffered until the actual code rate is calculated. The bandwidth is allocated according to the actual rate of the emergency information, and broadcasting starts from the beginning of the buffer tail.

This mechanism will not cause the waste of channel bandwidth, but it will have a delay when the receiving end stations obtain emergency information. The delay time equals the calculation time of the emergency information code rate.

5.3. No Delay and No Waste Mechanism (NDNM). No delay and no waste mechanism is similar to NBWM. When the master station starts receiving the emergency information, there is no broadcasting or buffering; only emergency information actual code rate is calculated. According to the actual rate, bandwidth is allocated and then the most recently received data is directly broadcasted. The previously received data are no longer broadcasted and discarded directly.

This mechanism can achieve no delay and no waste of bandwidth, but it will lose some data at the beginning of the emergency information data flow.

In summary, the three types of bandwidth statistical multiplexing mechanisms between regular files and emergency information have their own characteristics, and the corresponding mechanism needs to be selected according to the actual situation, which can make good use of the channel bandwidth and reduce the delay.

In general, the broadcasting without delay mechanism is suitable for emergency information flow. Most of this data needs live broadcasting, and the shorter the delay, the better. No channel bandwidth wasting mechanism is suitable for broadcasting in advance to make good the emergency information file; emergency information file before the broadcast may not get an accurate broadcasted rate, it is not easy to avoid the waste of bandwidth through the bandwidth settings, and such data requires less time delay, so it is reasonable to use no channel bandwidth wasting mechanism. No delay and no bandwidth waste mechanism is suitable when the emergency information beginning has unimportant data, such as redundant data and so on.

Emergency information in this paper contains lightning data and earthquake, flood, mudslide, and other meteorological disaster prediction information. With the requirements of small amount of data and low latency, the broadcasting without delay mechanism is used for the broadcasting.

Since the satellite data cycle is one hour, one hour test time is taken. The specific test results are shown in Figure 4.

Figure 4 shows that NBWM has some time delays, NDNM discards data beginning. All these situations are unacceptable during broadcasting. Therefore, we conclude that NDM is better than the other mechanisms.

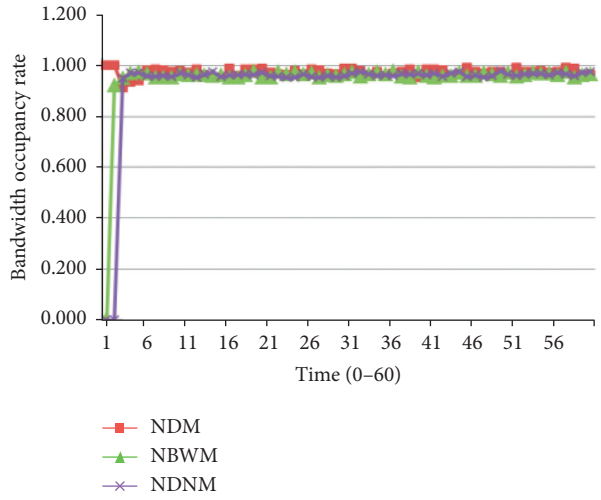


Figure 4: Test results of the three mechanisms.

6. Experiment

6.1. Environment Introduction. In order to verify the broadcasting strategy designed in this paper, we use common multiplexing of PID and channel mode among regular files, and broadcasting without delay mechanism between regular files and emergency information. Data broadcasting software and receiving software were developed, and FENGYUN 4-A satellite was used in this experiment.

We transmitted the broadcast data to the satellite repeater through DVB-S2 modulator, frequency upconverter, power amplifier equipment, and a 13 meter antenna and finally received the data through a frequency downconverter and a DVB-S2 demodulator using a small antenna, as shown in Figure 5.

Since the design principles and mechanisms of HRIT-1, HRIT-2, HRIT-3, LRIT, and EWAIB have no difference, the test part of this paper takes HRIT-1 as a typical object, and its test results can represent the test conditions of other file channels. The performance configuration of the equipment participating in the test is consistent with Table 2.

The broadcast speed of HRIT-1 is 12 Mbps; then one hour data is about 42 Gb. Experiment takes 378 files with 5637144576 total bytes. This paper adopts four combinations of experiment for HRIT-1.

The following four test scenarios are designed: Test 1 contains a single-PID file broadcasted with 12 Mbp. Test 2 contains regular files broadcasted with 4 PIDs. Each rate is 3 Mbps and the total rate is 12 Mbps. Test 3 contains regular files broadcasted with different code rate PIDs: 1 Mbps, 2 Mbps, 4 Mbps, and 5 Mbps. Test 4 contains regular files broadcasted with 3 PIDs, each rate is 3.3 Mbps, 1 PID is used for emergency information broadcasting with the rate 0.1 Mbps, and the total rate is 10 Mbps.

6.2. Metrics for Evaluation. During the test, the following indicators are used to evaluate and compare the bandwidth statistics and multiplexing results.

Broadcasting occupation ratio (BOR) is defined as the ratio between the total bandwidth allocated by the channel and the maximum bandwidth allowed by the network:

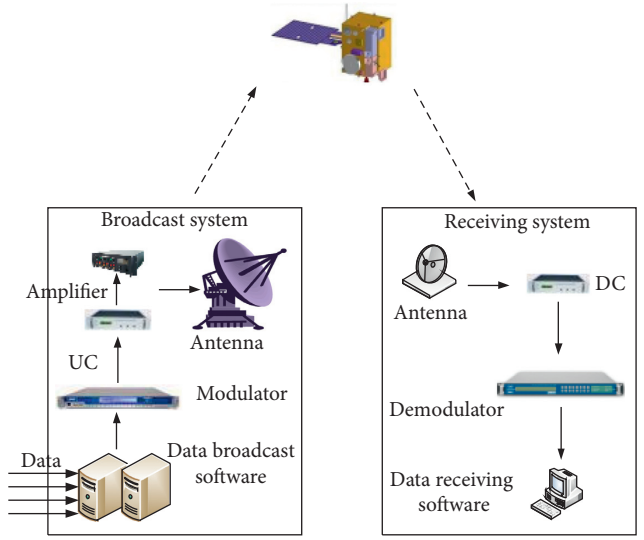


Figure 5: Satellite broadcast test block diagram.

Table 5: The experiment results.

Test no.	PID	BF_i	BS_i (byte)	BR_i (Mbps)	BO_i (Mbps)	BOR_i	BFR_i	BSR_i	BER_i
1	PID1	378	5637144576	12	11.652	0.971	1	1	0
2	PID1	92	1409185367	3	2.923	0.974	1	1	0
	PID2	95	1408351758	3	2.915	0.972	1	1	0
	PID3	93	1412392349	3	2.932	0.977	1	1	0
	PID4	98	1407215102	3	2.92	0.973	1	1	0
3	PID1	30	469785314	1	0.974	0.974	1	1	0
	PID2	63	933895146	2	1.938	0.969	1	1	0
	PID3	119	1823684108	4	3.885	0.971	1	1	0
	PID4	166	2409780008	5	4.865	0.973	1	1	0
4	PID1	115	1859703744	3.9	2.949	0.983	1	1	0
	PID2	119	1862197390	4	2.933	0.978	1	1	0
	PID3	112	1842063726	3.9	2.945	0.982	1	1	0
	PID4	32	73179716	0.2	0.969	0.969	1	1	0

$$\mathrm{BOR} = \sum_i \frac{\mathrm{BO}_i}{B_{\max}}. \tag{3}$$

Broadcasting file ratio (BFR) is defined as the ratio of the number of files received by the small station to the total number of files broadcasted:

$$\mathrm{BFR} = \sum_i \frac{\mathrm{BF}_i}{\mathrm{BF}}. \tag{4}$$

Broadcasting file size ratio (BSR) is defined as the ratio of the total number of file bytes received by the small station to the total number of bytes broadcasted:

$$\mathrm{BSR} = \sum_i \frac{\mathrm{BS}_i}{\mathrm{BS}}. \tag{5}$$

Broadcasting file error ratio (BER) is defined as the ratio of the sum of error bytes in the received file of the small station to the total number of bytes broadcasted:

$$\mathrm{BER} = \sum_i \frac{\mathrm{BER}_i}{\mathrm{BS}}. \tag{6}$$

6.3. Experiment Results. The meteorological products for HRIT-1, HRIT-2, HRIT-3, and LRIT cycle once an hour, so all the tests last for one hour, and the results are shown in Table 5.

According to Table 5, some figures can be drawn from these numbers, such as Figures 6–8.

From the data shown above, it can be concluded that the total broadcast rate is very stable, there is no obvious fluctuation, and BOR is also very accurate, nearly 97%.

6.4. Performance Evaluation. It can be concluded from Table 5 that BFR is 1, BSR is 1, and BER is 0, which proves that the regular files and emergency information can be completely and safely broadcasted to the channel equipment and the satellite.

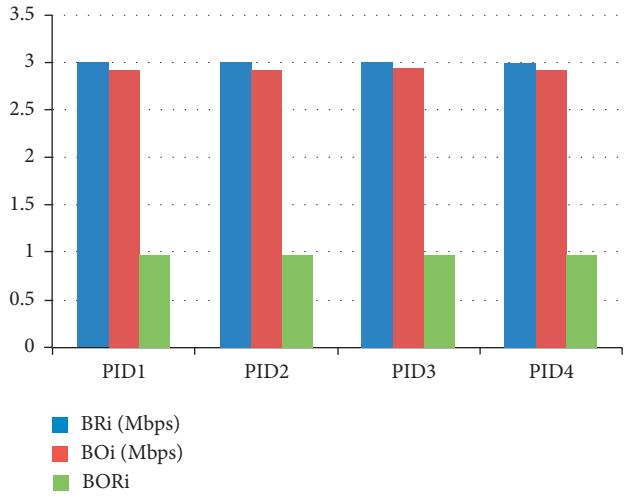


Figure 6: Average bandwidth tests.

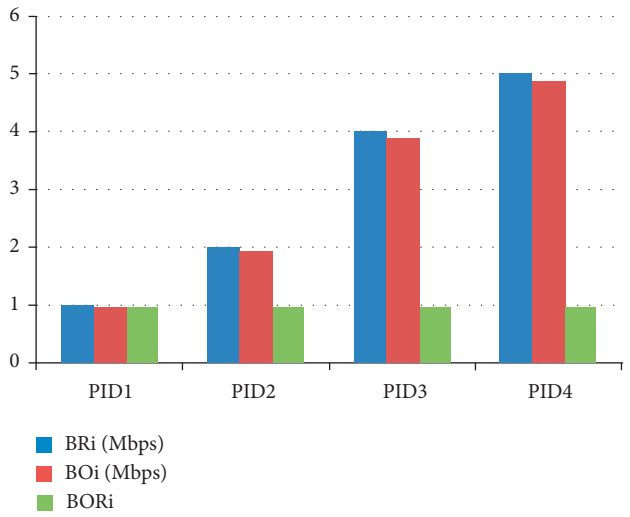


Figure 7: Staircase bandwidth tests.

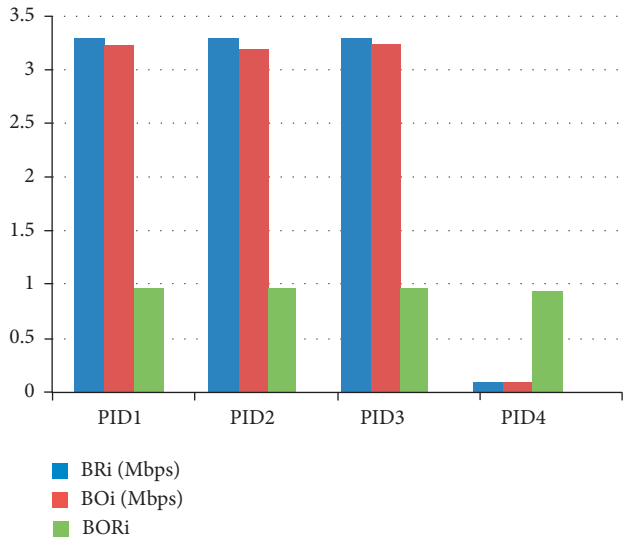


Figure 8: Regular files and emergency information tests.

Figures 6–8 indicate that during the four-combination test, whether single-PID or multi-PID combination, the bandwidth occupation rate of each channel is about 97%, which is better than the requirement utilization rate with 96% described in Section 2.2.

Accordingly, this bandwidth statistics and multiplexing control method meets the broadcasting system nonfunctional requirements well.

7. Conclusion

Satellite bandwidth multiplexing statistics are one of the most important functions of satellite data broadcasting platform. The statistical function of bandwidth multiplexing has an important impact on whether the satellite channel resources can be fully utilized, whether the meteorological data can arrive at the destination for the first time, and whether data is correctly transmitted.

This paper introduced a broadcast bandwidth statistical multiplexing and control method. Common multiplexing of PID and channel mechanism is used for regular file data broadcasting. Broadcasting without delay mechanism is used for emergency information broadcasting. However, broadcast timetable, track data, user authorization, and other information are not taken into account; they do not belong to regular or emergency information.

This broadcast bandwidth management improves satellite broadcasting efficiency. It has become an effective transmission means of meteorological information broadcasting service. It also has meaningful reference for the sun-synchronous orbit satellite communication.

Our future work will include analyzing the characteristics of all broadcast data, considering uplink and downlink quality monitoring [4], and adding adaptive adjustment function for transmission power by the aid of knapsack allocation problem [8], for the purpose of realizing the optimal allocation of broadcast resources.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors wish to acknowledge the cooperation of China Meteorological Administration National Satellite Meteorological Center. This work was supported in part by the National Natural Science Foundation of China Youth Fund under Grants 51608209 and 18BTJ031.

References

[1] D. Roddy, *Satellite Communications*, McGraw-Hill, New York, NY, USA, 2001.

[2] A. Donner, J. A. Saleemi, and J. Mulero Chaves, "TETRA backhauling via satellite: improving call setup times and saving bandwidth," *Journal of Computer Networks and Communications*, vol. 2014, Article ID 562546, 16 pages, 2014.
[3] U. Park, H. W. Kim, D. Sub Oh, and B. J. Ku, "A dynamic bandwidth allocation scheme for a multi-spot-beam satellite system," *ETRI Journal*, vol. 34, no. 4, pp. 613–616, 2012.
[4] J. Lei and M. Vazquez-Castro, "Joint power and carrier allocation for the multibeam satellite downlink with individual SINR constraints," in *Proceedings of the 2010 IEEE international conference on communications (ICC)*, Cape Town, South Africa, May 2010.
[5] J. P. Choi and V. W. S. Chan, "Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2983–2993, 2005.
[6] Y. Li, Z. Wang, L. Luo, Z. Chen, B. Xia, and H. Luo, "Bandwidth allocation of cognitive relay networks with energy harvesting for smart grid," *Journal of Computer Networks and Communications*, vol. 2019, Article ID 5038963, 9 pages, 2019.
[7] A. Bahnasse, F. E. Louhab, H. A. Oulahyane, M. Talea, and A. Bakali, "Smart bandwidth allocation for next generation networks adopting software-defined network approach," *Data in Brief*, vol. 20, pp. 840–845, 2018.
[8] S. Shi, G. Li, Z. Li, H. Zhu, and B. Gao, "Joint power and bandwidth allocation for beam-hopping user downlinks in smart gateway multibeam satellite systems," *International Journal of Distributed Sensor Networks*, vol. 13, no. 5, Article ID 155014771770946, 2017.
[9] H. Wang, A. Liu, and X. Pan, "Optimization of joint power and bandwidth allocation in multi-spot-beam satellite communication systems," *Mathematical Problems in Engineering*, vol. 2014, Article ID 683604, 9 pages, 2014.
[10] J. Miao, Z. Hu, K. Yang, C. Wang, and H. Tian, "Joint power and bandwidth allocation algorithm with Qos support in heterogeneous wireless networks," *IEEE Communications Letters*, vol. 16, no. 4, pp. 479–481, 2012.
[11] S. I. Wayer and A. Reichman, "Resource management in satellite communication systems: heuristic schemes and algorithms," *Journal of Electrical and Computer Engineering*, vol. 2012, Article ID 169026, 10 pages, 2012.
[12] L. Peng and J. Ye, "A bandwidth allocation scheme for the integrated service satellite broadcast network," *Microcomputer Information*, vol. 24, no. 15, pp. P10–P16, 2008.
[13] ran Liu, K. Jiang, and Y. Hu, "Design of data broadcast platform in China meteoroligical administration's satellite data broadcast system(CMACast)," *Digital Communication World*, vol. 7, no. 7, pp. P54–P58, 2013.
[14] C. Wang, Li Xiang, Y. Chen, and K. Jiang, "Design of CMA's broadcast system for meteorological data-CMACast," *Journal of Applied Meteorological Science*, vol. 23, no. 1, pp. P113–P120, 2012.
[15] Z. Xu, Di Xian, Y. Gao, and Y. Qi, "Design and implementation of FENGYUN-4 satellite data service system," *Satellite Application*, vol. 83, no. 11, pp. P55–P60, 2018.

Hindawi
Scientific Programming
Volume 2020, Article ID 8892815, 6 pages
https://doi.org/10.1155/2020/8892815

Research Article

Impact Factors on Posterior Modified Transfacet Debridement for Thoracic Spinal Tuberculosis Basing on Regression and Classification Analysis

Xiaoping Chen,[1] **Jiamin Lin**,[1] **Han Huang**,[2] **and Yunpeng Huang**[3]

[1]*Department of Statistics, College of Mathematics and Informatics & FJKLMAA, Fujian Normal University, Fuzhou 350117, China*
[2]*Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China*
[3]*Department of Spine Surgery, The First Affiliated Hospital of Fujian Medical University, Fuzhou 350005, China*

Correspondence should be addressed to Yunpeng Huang; yawn7770@126.com

Received 15 April 2020; Revised 17 June 2020; Accepted 7 July 2020; Published 1 August 2020

Academic Editor: Chenxi Huang

Posterior transfacet approach has been proved to be a safe and effective access to treat thoracic disc herniation. However, the influencing factors of posterior modified transarticular debridement for thoracic tuberculosis have not been reported in the clinical literature. From 2009 to 2014, 37 patients with TST underwent a posterior modified transfacet debridement, interbody fusion following posterior instrumentation, under the cover of 18 months of antituberculosis chemotherapy. The patients were evaluated preoperatively and postoperatively in terms of Frankel Grade, visual analog scale (VAS) pain score, kyphotic Cobb angle, and bone fusion. Blood loss (positive correlation) and focal debridement (positive correlation) could affect operative time. Operative time (positive correlation) could affect blood loss. While, age (positive correlation), PostE (negative correlation), and T_FocalDebridement (positive correlation) could affect bone fusion. The accuracy of naive bayes classifier model is 86.11%. Our preliminary results show that blood loss and focal debridement could affect operative time; operative time could affect blood loss; age, PostE, and T_FocalDebridement could affect bone fusion; the naive Bayes classifier model can predict the KirkaldyWillis accurately.

1. Introduction

Tuberculosis (TB), an infectious disease caused by mycobacterium tuberculosis (MTB), is the ninth leading cause of global death, ranking above acquired immune deficiency syndrome (AIDS). The spine is the most frequently involved part of extrapulmonary tuberculosis, which is more common in thoracolumbar vertebrae, 30.3% to 55.8% of which are thoracic tuberculosis and about 67% of thoracic tuberculosis patients are accompanied by tuberculosis [1, 2]. Vertebral body destruction will lead to spinal instability, spinal cord damage, and deformity. Therefore, in addition to standardized antituberculosis treatment, surgery is also an important treatment method for thoracic tuberculosis treatment [3].

As an important surgical intervention in the treatment of spinal tuberculosis, surgical treatment aims to effectively remove the lesion, completely relieve the neurospinal compression, restore the height of the affected vertebra as far as possible, and rebuild the stability of the spine. After continuous exploration and development, the current surgical treatment methods are classified according to the approach, mainly including anterior approach, posterior approach, and combined approach. (1) Anterior approach: compared with other methods, anterior approach is the most convenient method for removing lesions because it can directly reach the lesion and has a good operating field of vision [4]. (2) Combined approach: both lesion clearance and deformity correction. However, anterior or anterior combined approaches have a higher rate of primary activity

and mortality [5, 6]. (3) Posterior approach: posterior approaches such as posterolateral, pedicle, or posterior approaches and their variants have been proven to be a safe and effective method for the treatment of thoracic disc herniation (TDH) with a relatively low incidence. In addition, posterior approach can better correct kyphosis, which is the most common complication of TST [7]. Some studies believed that, in terms of operation time, the combined approach was longer than the anterior approach, but there was no statistical difference with the posterior approach. In terms of blood loss, anterior and posterior combined approaches were more common than anterior or posterior approaches. There was no clinical literature to report the influencing factors of posterior modified transarticular debridement for thoracic tuberculosis. This paper considers the influencing factors of operative time, blood loss, bone fusion, and the classification of the KirkaldyWillis.

2. Methods

Between 2009 and 2014, the authors treated 37 consecutive patients with TST via a modified transfacet approach [7]. This study was approved by the Ethics Committee of Affiliated Hospital of Fujian Medical University.

2.1. Preoperative Surgery. Patients were treated with standard chemotherapy regimen of isoniazid (H), rifampicin (R), ethambutol (E), and pyrazinamide (Z) (HREZ) at least 4 weeks before surgery, including isoniazid (300 mg/d), rifampicin (450 mg/d), ethambutol (750 mg/d), and pyrazinamide (750 mg/d). ESR was 44.7 ± 23.3 mm/h. And the mean kyphosis angle was $29.4 \pm 10.9^{\circ}$. Surgery is performed when ESR is significantly reduced (<40 mm/h). Preoperative antituberculosis treatment can reduce mycobacterium tuberculosis in focus and improve surgical safety. Bony spinal fusion was assessed according to the criteria defined by Lee et al. [8], which was graded as shown in Table 1. And in our study, all patients obtained definitive bone fusion.

2.2. Statistical Principle. In order to examine the factors that might affect operative time, the multivariable linear regression model is adopted. Operative time is considered as a continuous and dependent variable, whereas blood loss, PreK, FFUK, and focal debridement are constituted the independent variables. The multivariable linear regression model is also utilized for investigating whether operative time could affect blood loss. Blood loss is considered as a constant and dependent variable, whereas operative time, bone fusion, and PreK constituted the independent variables. The naive Bayes classifier model is also utilized for investigating what affects the KirkaldyWillis grade. All the statistical analyses were performed using the R Project (version 3.5.1, 2018, The R Foundation for Statistical Computing), the R package e1071, corrplot needs to be loaded. The results were considered statistically significant at $p < 0.05$.

Multiple Linear Regression [9] attempts to find a linear equation with known data to describe the relationship between two or more independent variables and dependent variable and use this linear equation to predict the results.

Let the dependent variable be Y, and the k independent variables affecting the dependent variable be $X_1, X_2, \ldots, X_K$, respectively. Let us assume that the influence of each independent variable on the dependent variable Y is linear, that is, under the condition that other independent variables remain unchanged, the mean value of Y changes uniformly with the change of independent variable X_i:

$$Y_i = \beta_i^{(0)} + \beta_i^{(1)} X_i^{(1)} + \beta_i^{(2)} X_i^{(2)} + \cdots + \beta_i^{(K)} X_i^{(K)} + \varepsilon. \quad (1)$$

This is called the population regression model, and $\beta_i^{(0)}, \beta_i^{(1)}, \ldots, \beta_i^{(K)}$ are the regression parameters.

The principle of naive Bayes classifier [10] is a statistical method. Bayes theorem is the theoretical basis of classifier modeling. It uses conditional probability principle, prior information, and sample data information to determine the probability of event occurrence. In addition, naive Bayes model is a kind of generating model, which directly models the joint probability to obtain the target probability. The algorithm is as follows.

$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, among them $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)})^T$, $x_i^{(j)}$ is the jth independent variable of the ith sample, a_{jl} is the jth independent variable of the lth value that it is possible to take $j = 1, 2, \ldots, K$ and $l = 1, 2, \ldots, S_j$; $y_i \in \{c_1, c_2, \ldots, c_P\}$.

(1) Calculate the prior probability and conditional probability:

$$P(Y = c_p) = \frac{\sum_{i=1}^{n} I(y_i = c_p)}{n}, \quad p = 1, 2, \ldots, P,$$
$$P\left(X^{(j)} = a_{jl} \,\middle|\, Y = c_p\right) = \frac{\sum_{i=1}^{n} I(x_i^{(j)} = a_{jl}, y_i = c_p)}{\sum_{i=1}^{n} I(y_i = c_p)}, \quad j = 1, 2, \ldots, K;\ l = 1, 2, \ldots, S_j;\ p = 1, 2, \ldots, P. \quad (2)$$

(2) For a given sample of $x = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^T$, compute

$$P(Y = c_p) \prod_{j=1}^{K} P\left(X^{(j)} = x^{(j)} \,\middle|\, Y = c_p\right), \quad p = 1, 2, \ldots, P. \quad (3)$$

Table 1: The modified criteria of Lee et al for radiological fusion.

Grade	Description
Definitive fusion	Definitive bony trabecular bridging across the graft-host interface, no movement (<3°) on a flexion-extension radiograph and no gap at the interface
Probable fusion	No definitive bony trabecular crossing, but no detectable movement and no identifiable gap at the interface
Possible pseudarthrosis	No bony trabecular crossing, no movement, but identifiable gap at the interface
Definite pseudarthrosis	No traversing trabecular bone, definitive gap, and movement >3°

(3) Determine the classification of sample x:

$$y = \arg\max_{c_p} P\left(Y = c_p\right) \prod_{j=1}^{K} P\left(X^{(j)} = x^{(j)} \,\middle|\, Y = c_p\right). \quad (4)$$

3. Result

We find that blood loss (positive correlation) and focal debridement (positive correlation) could affect operative time, but not bone fusion, gender, PreA, PostA, FFUA, age, hospital stay, segments, follow-up, PreK, PostK, PreE, PostE, FUUK, FFUE, and T_Level. We also find that operative time (positive correlation) could affect blood loss, but not gender, PreA, PostA, FFUA, age, hospital stay, segments, follow-up, PostK, FUUK, PreK, PreE, PostE, FFUE, T_Level, and T_FocalDebridement. We also find age (positive correlation), PostE (negative correlation), and T_FocalDebridement (positive correlation) could affect bone fusion, but not gender, PreA, PostA, FFUA, hospital stay, segments, follow-up, PostK, PreE, FUUK, FFUE, and T_Level.

Final equation 1:

$$\begin{aligned}\text{Operativetime} = {} & 108.27007 + 0.11575 * \text{bloodloss} \\ & + 23.61903 * \text{T\_FocalDebridement}.\end{aligned} \quad (5)$$

As for the influencing factor model of operative time, F test is performed on the whole model (Table 2), its F-statistic is 57.76 on 2 and 34 DF, and p value is $1.161e{-}11$, indicating that the model is significant on the whole, that is, the model is significant highly. The goodness of fit R^2 is 0.7726, indicates that the model has a good fitting effect and can better explain the influencing factors of operative time. In patients with the same T_FocalDebridement, operative time increased by an average of 0.11575 if blood loss increases 1 ml, its p value is 0.000168, which indicates that at the significance level of 0.001, and blood loss has a significant effect on the operative time. Also, operative time increased by an average of 23.61903 as T_FocalDebridement increased by 1 segment when the blood loss is the same, its p value is 0.035237, which indicates that at the significance level of 0.05, and T_FocalDebridement factor has a significant effect on the operative time.

Final equation 2:

$$\text{bloodloss} = -506.0542 + 4.4878 * \text{operativetime}. \quad (6)$$

As for the influencing factor model of blood loss, F test is performed on the whole model (Table 3), and its F-statistic is 99.85 on 1 and 35 DF, and p value is $8.648e{-}12$, indicating that the model is significant highly on the whole, that is, the model is significant. The goodness of fit R^2 is 0.7404, indicates that the model has a good fitting effect and can better explain the influencing factors of blood loss. Blood loss increased by an average of 4.4878 if operative time increased 1 min, its p value is $8.65e{-}12$, which indicates that at the significance level of 0.001, and the operative time factor has a significant effect on the blood loss.

Final equation 3:

$$\begin{aligned}\text{Bonefusion} = {} & 2.75352 + 0.04326 * \text{Age} - 0.08810 * \text{PostE} \\ & + 1.53847 * \text{T\_FocalDebridement}.\end{aligned} \quad (7)$$

As for the influencing factor model of bone fusion, F test is performed on the whole model (Table 4), its F-statistic is 13.12 on 3 and 31 DF, and p value is $8.342e{-}06$, indicating that the model is significant on the whole, that is, the model is significant highly. The goodness of fit R^2 is 0.5026, indicates that the model has a good fitting effect and can better explain the influencing factors of bone fusion. In patients with the same PostE and T_FocalDebridement, bone fusion increased by an average of 0.04326 if age increases 1, its p value is 0.00286, which indicates that at the significance level of 0.001, age has a significant effect on the bone fusion. Also, bone fusion decreased by an average of 0.08810 as PostE increased by 1 when the age and T_FocalDebridement are the same, its p value is 0.03415, which indicates that at the significance level of 0.05, PostE factor had a significant effect on the bone fusion. Also, bone fusion increased by an average of 1.53847 as T_FocalDebridement increased by 1 when the age and PostE are the same, its p value is $3.99e{-}06$ indicates that at the significance level of 0.001, and T_FocalDebridement factor had a significant effect on the bone fusion.

After deleting one sample in which KirkaldyWillis is two, only the samples of KirkaldyWillis = 3 and KirkaldyWillis = 4 are trained. The correlation coefficient matrices of gender, bone fusion, follow-up, FFUK, FFUA, and FFUE are calculated, as shown in Table 5.

As shown in Figure 1, we can find that the interior color of each circle is lighter except the diagonal line, indicating that the correlation coefficient between variables is low and the degree of correlation is not high. The above variables can be used for naive Bayes classification modeling.

The confusion matrix is obtained in Table 6.

Table 2: Model summary for operative time.

	Primary independent variables		
	Coefficient estimate	Standard error	p value
(Intercept)	108.27007	17.71126	<0.001***
Blood loss	0.11575	0.02739	<0.001***
T_FocalDebridement	23.61903	10.76961	0.035*
Statistic	Value	DF	p value
F	57.76	(2, 34)	<0.001***
R^2	0.7726		

Table 3: Model summary for blood loss.

	Primary independent variables		
	Coefficient estimate	Standard error	p value
(Intercept)	−506.0542	103.9502	<0.001
Operative time	4.4878	0.4491	<0.001***
Statistic	Value	DF***	p value
F	99.85	(1, 35)	<0.001***
R^2	0.7404		

Table 4: Model summary for bone fusion.

	Primary independent variables		
	Coefficient estimate	Standard error	p value
(Intercept)	2. 75352	0. 97119	<0.01**
Age	0.04326	0. 01342	<0.01**
PostE	−0.08810	0.03986	<0.05*
T_FocalDebridement	1.53847	0.27879	<0.001***
Statistic	Value	DF	p value
F	13.12	(3, 33)	<0.001***
R^2	0.5026		

Table 5: Sample correlation matrix.

	Gender	Bone fusion	Follow-up	FFUK	FFUA	FFUE
Gender	1.0000000	−0.2805976	−0.1155847	0.0715071	−0.1428571	0.3157139
Bone fusion	−0.2805976	1.0000000	0.2229494	0.2665206	−0.1624512	−0.2556321
Follow-up	−0.1155847	0.2229494	1.0000000	0.3185323	0.0761168	−0.0372737
FFUK	0.0715071	0.2665206	0.3185323	1.0000000	−0.0449019	−0.0057099
FFUA	−0.1428571	−0.1624512	0.0761168	−0.0449019	1.0000000	−0.1733689
FFUE	0.3157139	−0.2556321	−0.0372737	−0.0057099	−0.1733689	1.0000000

Because naive Bayes model is a kind of generating model, it directly models the joint probability to obtain the target probability. Through modeling, we can only get the classification probability of each sample, so as to determine the category, that is, no explicit model can be obtained. The classification accuracy is 86.11%, which indicates the percentage of all correctly predicted samples to all samples; recall is 96.6%, which indicates the correct percentage of all samples predicted with KirkaldyWillis = 4; precision is 87.5%, which indicates the proportion of all KirkaldyWillis samples predicted to be KirkaldyWillis = 4.

4. Discussion

Operative data of 37 patients with TST were recorded with multiple indicators of the operative level. Multiple linear regression is also suitable for analyzing the influence of multiple surgical factors on one operative variable. The number of vertebral segments debridement is the most important factor affecting the operation time. When other surgical factors remain unchanged, the greater the amount of blood loss, the longer the operation time. Bone fusion time is the most important factor affecting blood loss. When other surgical factors remain unchanged, the longer the operation time, the longer the bone fusion time or the greater the preoperative the KirkaldyWillis score, the greater the blood loss. And both models have high R^2 values. The posterior transfacet-modified approach has been reported to be a safe and effective method for the treatment of thoracic spinal tuberculosis (TST) with low morbidity in the precious report [7]. However, the impact factors of the surgical approach for TST have not been intensively analyzed. Multiple linear regressions are indeed applicable for analyzing the impact of multiple factors on one operative variable. The aim

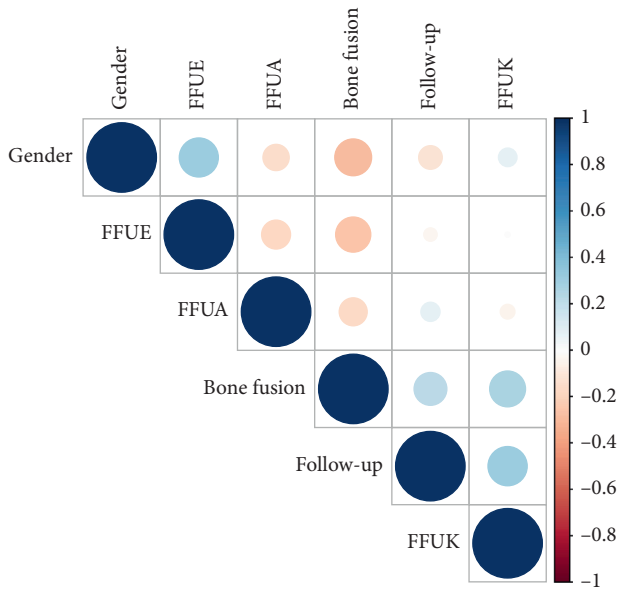


Figure 1: Sample correlation diagram.

Table 6: Confusion matrix.

		Predicted	
		KirkaldyWillis = 4	KirkaldyWillis = 3
Actual	KirkaldyWillis = 4	3	4
	KirkaldyWillis = 3	1	28

of this study was to validate the efficacy and safety of posterior modified transfacet surgical approach for the treatment of TST by using multiple linear regressions. Naive Bayes classifier can classify category data well, with high recall and accuracy. On the whole, the analysis results are consistent with clinical results. Because naive Bayes is sensitive to sample equilibrium, samples with KirkaldyWillis = 2 is deleted before the classification of KirkaldyWillis, so we cannot judge the samples with KirkaldyWillis = 2, which needs to be improved. Machine learning classification methods, such as decision tree, random forest, KNN, boosting, and SVM, can also be used to classify other categories of surgical indicators. The Naive Bayes classification model could be used because the correlation coefficient between variables is low and the degree of correlation is not high. The classification accuracy, recall, and precision for KirkaldyWillis scores are high. Therefore, the Naive Bayes classification model is reasonable to classify the KirkaldyWillis scores.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Ethical Approval

Between 2009 and 2014, the authors treated 37 consecutive patients with TST via a modified transfacet approach [7]. This study was approved by the Ethics Committee of Affiliated Hospital of Fujian Medical University.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors sincerely thank the participants for their help and willingness to participate in this study. This study was supported by the National Natural Science Foundation of China (11601083 and U1805263), Natural Science Foundation of Fujian Province, China (2019J01451), Program for Probability and Statistics: Theory and Application (IRTL1704), and Innovative Research Team in Science and Technology in Fujian Province University (IRTSTFJ).

Supplementary Materials

The supplementary information files contain the table of patients treated in this study. (*Supplementary Materials*)

References

[1] C. S. Restrepo, R. Katre, and A. Mumbower, "Imaging manifestations of thoracic tuberculosis," *Radiologic Clinics of North America*, vol. 54, no. 3, pp. 453–473, 2016.

[2] Y. Yao, H. Zhang, M. Liu et al., "Prognostic factors for recovery of patients after surgery for thoracic spinal tuberculosis," *World Neurosurgery*, vol. 105, pp. 327–331, 2017.

[3] C. Fisahn, F. Alonso, G. A. Hasan et al., "Trends in spinal surgery for Pott's disease (2000–2016): an overview and bibliometric study," *Global Spine Journal*, vol. 7, no. 8, pp. 821–828, 2017.

[4] Y. Gao and Y. Ou, "Comparison between titanium mesh and autogenous iliac bone graft to restore vertebral height through posterior approach for the treatment of thoracic and lumbar spinal tuberculosis," *PLoS One*, vol. 12, no. 4, pp. 1–12, 2017.

[5] R. K. Jain and I. K. Dhammi, "Tuberculosis of the spine: a review," *Clinical Orthopaedics and Related Research*, vol. 460, pp. 39–49, 2007.

[6] A. M. Quamar and S. A. Mir, "The Concept of evolution of thoracolumbar fracture classifications helps in surgical decisions," *Asian Spine Journal*, vol. 9, no. 6, pp. 984–994, 2015.
[7] Y. P. Huang, J. H. Lin, X. P. Chen et al., "Preliminary experience in treating thoracic spinaltuberculosis via a posterior modified transfacet debridement, instrumentation, and interbody fusion," *Journal of Orthopaedic Surgery and Research*, vol. 13, no. 292, pp. 1–9, 2018.
[8] J. S. Lee, K. P. Moon, S. J. Kim, and K. T. Suh, "Posterior lumbar interbody fusion and posterior instrumentation in the surgical management of lumbar tuberculous spondylitis," *The Journal of Bone and Joint Surgery*, vol. 89, no. 2, pp. 210–214, 2007.
[9] X. Q. He, *Applied Regression Analysis (R Language Edition)*, Publishing House of Electronics Industry, pp. 55–84, Beijing, China, 2017.
[10] H. Li, *Statistical Learning Method*, Tsinghua University Publishing House, pp. 47–52, Beijing, China, 2012.

Hindawi
Scientific Programming
Volume 2020, Article ID 8859152, 10 pages
https://doi.org/10.1155/2020/8859152

Research Article

Perceptual Characteristics of Chinese Speech Intelligibility in Noise Environment

Hui Song[1] **and Siyu Zhang**[2]

[1]*Art College, Northeast Electric Power University, Jilin, China*
[2]*Communication Acoustics Laboratory, Communication University of China, Beijing, China*

Correspondence should be addressed to Hui Song; sandysong@neepu.edu.cn

Received 8 April 2020; Revised 3 June 2020; Accepted 11 June 2020; Published 24 July 2020

Academic Editor: Chenxi Huang

Speech intelligibility is affected by various interfering factors in a speech transmission system. Noise is one of the most common affecting factors. Subjective listening experiments were, respectively, carried out in pink noise, speech noise, and white noise-interfering environment. The perceptual characteristics of the initials, finals, tones, and syllable intelligibility were analyzed, and the function relationships between Chinese speech intelligibility and SNR in noise environment were concluded, which could be used to evaluate or predict the Chinese speech intelligibility under noise transmission conditions.

1. Introduction

Speech is a main way of human communication, and speech intelligibility is an important parameter to assess the acoustic quality of a communication system. ISO defines speech intelligibility as a professional term of *the percentage of correctly received speech units in all transmitted speech units* [1]. Speech could be interfered by various factors in the communication system, like noise, reverberation, distortion, and so on. Noise is one of the most common affecting factors. Noise unavoidably exists in our daily life and masks speech signals with different degrees, while speech intelligibility decreases with the noise level increasing. It is very necessary to investigate speech intelligibility in noise environment for the speech transmission system [2].

There were some studies on the influence of noise on speech intelligibility, which include both objective and subjective ones. For objective experiments, Houtgast tested the speech intelligibility in the environment affected both by ambient noise and reverberation [3]. Bradley tested the speech intelligibility in ten classrooms full with students [4]. Ma et al. tested the speech intelligibility in 72 noisy conditions with four different noises including car, babble, train, and street noise [5]. Lavandier and Culling conducted two experiments to investigate the speech reception thresholds in noise condition with speech-shape in different angles [6]. Rennies et al. tested the speech intelligibility in the condition with noise interfering and quiet [7]. For subjective experiments, Rhebergen et al. designed a subjective experiment to investigate the influence of fluctuating noise on speech intelligibility [8]. Ishikawa et al. conducted a subjective experiment to observe the influence of background noise on speech intelligibility by dysphonia [9]. Bradley et al. investigated the combined influences of SNR and room acoustics parameters on speech intelligibility and concluded that the effect of SNR is much more important [10]. Van Wijngaarden et al. compared the speech intelligibility in noise environment for nonnative and native listeners [11]. Rhebergen et al. predicted the speech intelligibility in real-life background noises including animals, machines, and vehicles [12]. Luts et al. tested the French speech intelligibility in noise conditions [13]. Boon used CRM corpus to measure the speech intelligibility masked by speech-spectrum-shaped noise and concluded the functional relationship between speech intelligibility and SNR [14]. Pollack and Pickket measured the speech intelligibility in noisy environment with high levels and summarized the relationship of speech intelligibility, speech SPL, and SNR [15]. Elliott

measured the perceptual characteristics of noise-masked speech of children aged 9~17 years [16]. Payton et al. compared dry signals and speech signal intelligibility in noise, reverberant, and combined conditions and obtained the differences between the results and objective methods results [17]. Van Wijngaarden compared noise-masked speech intelligibility between native and nonnative Dutch speakers [18]. Zeng et al. investigated the influence of noise and reverberation on the Chinese speech intelligibility of elderly and young subjects [19]. Visentin et al. investigated the effects of different types of noise on speech intelligibility in university classrooms [20]. Duquesnoy measured the speech perceptual threshold for sentences in quiet and noisy environment of elderly and young subjects with normal hearing [21]. Kostić performed a subjective experiment with the MOS method and objective experiment with STOI to obtain the influence of music noise and SNR on Serbian speech intelligibility [22]. Most of the objective experiments aimed to get or improve the objective measuring methods of intelligibility, and the subjective experiments investigated the speech perception in noisy environments with different types and SNRs for different purposes. The objective results should be in accordance with subjective perceptual results, and the relationship between speech intelligibility and SNR should be studied further.

Every Chinese word is a monosyllable, which consists of initials, finals, and tones. Take "zhōng" for example, where "zh" is the initials and "ong" is the finals, while "-" is the tone which represents high flat, and "zhōng" is the syllable which is also the Pinyin of Chinese word. Thus, Chinese speech intelligibility includes initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility which is influenced by initials, finals, and tones intelligibility at the same time.

The present study would introduce a series of experiments of Chinese speech intelligibility in noise transmission systems and analyze the perceptual characteristics of speech intelligibility disturbed by different noises. The changing rules of Chinese speech intelligibility with SNR were analyzed, and the perceptual model of Chinese intelligibility was also concluded.

2. Listening Experiments of Intelligibility

The test speech signals were recorded in an anechoic chamber, and the noise would be added in postprocessing stage. Subjective experiments of speech intelligibility were performed in a listening room according to the National Standard of the People's Republic of China Acoustics-Speech articulation testing method [23], which introduces the method of measuring and evaluating the quality of a speech transmission system quantificationally and directly. The standard applies to measuring the speech intelligibility of a speech communication system, like concert hall and human-computer interaction.

KXY lists were used for the speech listening test [23]. Every KXY list has 75 words, and the 75 words were divided into 25 groups including 3 syllables without any actual meaning when reading successively. Each group of 3 syllables were recorded after a prompt number. For example, in speech signal "No.R qiè fán yīng," "R" represents the group number and "qiè fán yīng" are the 3 syllables needed to play back to the subjects. Each list was recorded by two professional announcers.

All the lists were recorded in an anechoic chamber, in which the background noise is less than 20 dB and RT is less than 0.1 s [24]. The speech signals were recorded by two professional announcers (one male and one female) with standard Chinese Mandarin at a standard rate (4 words per second). The speakers' mouth was kept 10 cm away from the microphone. The SPL was 70 dB near the microphone, and thus the SNR of original speech signals was 50 dB.

In the postprocessing stage, the prerecorded speech signals were added with different types of noises and SNRs according the experiments needed in order to simulate the speech signals masked by noise. The subjective experiments were carried out in a quiet room, and the speech was played back to the listener through a headphone. The experiment procedure is shown in Figure 1.

2.1. Noise Types and SNR. Three types of noises were chosen in the experiments including pink noise, speech noise, and white noise. Pink noise commonly exists in natural environment, where the energy decreases from low frequency to high frequency with 3 dB per octave in logarithmic coordinates. Speech noise as a background noise distributes in the public places like concert, hall, and stadium, where the energy mainly focuses on low frequency. White noise is a broadband noise existing in the communication system, and its energy distributes uniformly in linear coordinate. Their time-domain and frequency-domain waveforms are shown in Figure 2. The three noises introduced above are typical noises in daily life.

Before the experimental parameters were selected, a pilot experiment was performed. In the pilot experiment, white noise was chosen as the background noise, and SNR focused on two extreme ranges of [−20,12] and [18, 24]. The results of the pilot experiment showed that the listeners cannot hear the speech when SNR was below -14 dB while the noise did not affect the listeners in distinguishing the speech when SNR exceeded 20 dB. Therefore, the SNR of speech intelligibility experiments in noise environment was chosen from −14 dB to 20 dB with 2 dB or 4 dB increasing and also including the initials speech recorded in studio without extra noise added with SNR of 50 dB. There were 15 transmission conditions. The information of SNR chosen in the experiments is shown in Table 1.

In order to compare and analyze the results, the SNRs were identical in the three types of noise-interfering environments, and each SNR condition included two speech signals with one male speaker and one female speaker.

2.2. Methods. Listening room and experiment procedure were all the same as those in the experiment described in [25]. If a subject writes down the initials completely same as what he heard, his record was regarded as a correct result; otherwise, considered wrong one, and the ratio of correct

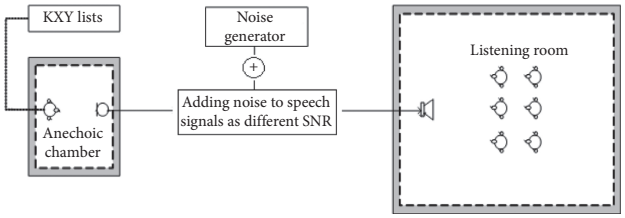


Figure 1: Experiment procedure of the subjective listening experiment.

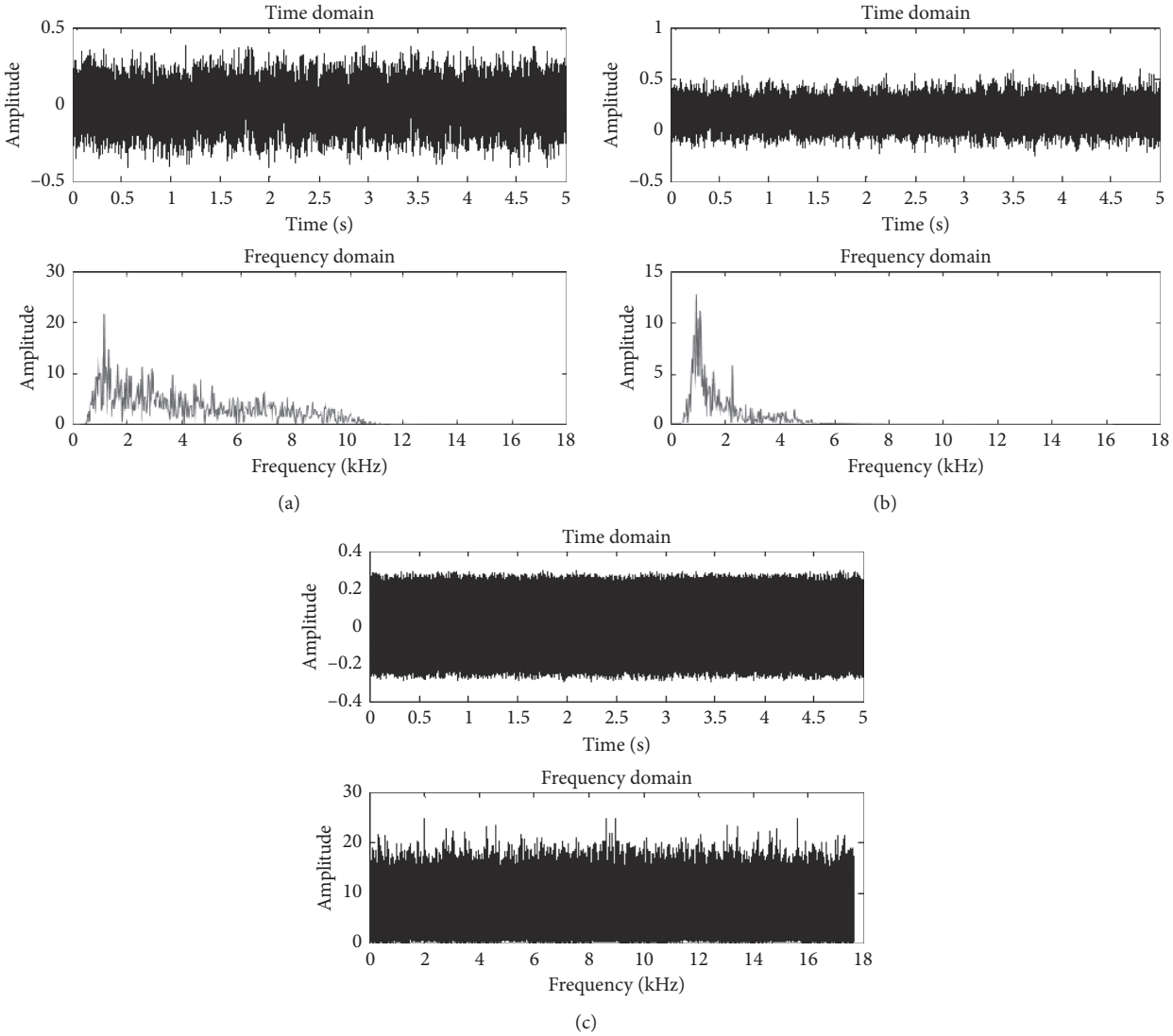


Figure 2: Time-domain and frequency-domain properties of noise. (a) Pink noise. (b) Speech noise. (c) White noise.

Table 1: SNR of speech intelligibility experiments in noise environment.

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SNR (dB)	−14	−12	−10	−8	−6	−4	−2	0	4	8	12	16	18	20	50

answer was called initials intelligibility score. In the same way, finals intelligibility, tones intelligibility, and syllable intelligibility were also calculated as described in following equations [23]:

$$Q = \sum_{i=1}^{n} \frac{(N_i/75) \times 100}{n}, \quad (1)$$

$$Q_{\mathrm{sm}} = \sum_{i=1}^{n} \frac{(N_{i\mathrm{sm}}/75) \times 100}{n}, \quad (2)$$

$$Q_{\mathrm{ym}} = \sum_{i=1}^{n} \frac{(N_{i\mathrm{ym}}/75) \times 100}{n}, \quad (3)$$

$$Q_{\mathrm{sd}} = \sum_{i=1}^{n} \frac{(N_{i\mathrm{sd}}/75) \times 100}{n}, \quad (4)$$

where Q, Q_{sm}, Q_{ym}, and Q_{sd} represent the syllable intelligibility, initials intelligibility, finals intelligibility, and tones intelligibility, respectively, N_i is the quantity of corrected perceptual syllable intelligibility by number i listener, while $N_{i\mathrm{sm}}$ is the quantity of corrected perceptual initials intelligibility by number i listener, while $N_{i\mathrm{ym}}$ is the quantity of corrected perceptual finals intelligibility by number i listener, while $N_{i\mathrm{sd}}$ is the quantity of corrected perceptual tones intelligibility by number i listener, and n is the total number of listeners involved in the experiment.

To guarantee the reliability of the experiments, when one listener's score was three times standard deviation different with the average score of all the listeners, his results were wiped out as invalid data. Then, the intelligibility score was recomputed.

2.3. Subjects. There were three types of noises with SNR 15 in the experiments which means 45 acoustic transmission conditions. A total of 90 KXY lists were involved, while every transmission condition needs two KXY lists (a male speaker and a female speaker).

The series of intelligibility experiments were divided to 3 groups according to the noise type:

(a) There were 14 subjects (7 males and 7 females) who participated in the pink noise-interfering experiment, while the results of 13 subjects (6 males and 7 females) were valid after the standard deviation was tested three times.

(b) There were 13 subjects (6 males and 7 females) who participated in the speech noise-interfering experiment, while the results of 13 subjects (6 males and 6 females) were valid after the standard deviation was tested three times.

(c) There were 14 subjects (7 males and 7 females) who participated in white noise-interfering experiment, while all of the 13 subjects' results (6 males and 7 females) were valid after the standard deviation was tested three times.

All of the subjects were undergraduates, aged from 19 to 23 years, and were all native speakers of Chinese Mandarin without any known hearing problems. The subjects were familiar with Chinese Pinyin spelling rules and are experienced in relevant listening experiments. A simple listening training was taken to the subjects before the formal subjective experiments. A total of 92250 $((14 + 14 + 13) \times 75 \times 30 = 92250)$ stimulus-response events happened.

3. Results

3.1. Intelligibility of Mandarin in Noise Environment. After data processing, syllable intelligibility, initials intelligibility, finals intelligibility, and tones intelligibility of Mandarin in pink noise, speech noise and white noise-interfering environments were calculated, respectively.

3.1.1. Influence of Gender on Perceptual Intelligibility. Because of the differences of speeches between male and female in fundamental frequency, the bandwidth of frequency, spectrum structure, and intelligibility perception of speech may be different for listeners. Comparisons of intelligibility scores tested by the male speaker and female speaker in pink noise, speech noise, and white noise environments are shown in Figure 3.

Figure 3 shows that the variational trend of intelligibility curves between male and female speakers is almost the same. No significant differences ($F(2,90) = 0.616$, $p > 0.05$) can be found in the speech intelligibility perceptual results between male and female announcers after the ANOVA test. Thus, the Chinese speech intelligibility scores were averaged across the results calculated from two genders under each transmission condition.

3.1.2. Intelligibility of Mandarin in Different Noise Environments. Initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility tested in pink noise environment are shown in Figure 4.

It is shown that the speech intelligibility increases as the SNR increases for all the four curves, and the tones intelligibility scores are the highest, then the finals intelligibility, thirdly the initials intelligibility, and the syllable intelligibility the lowest. The perceptual curves of tones and finals intelligibility have the same trend. Tones intelligibility increases significantly as SNR increases and when SNR is less than −8 dB. Tones intelligibility arrives at 0.9, and the curve is relative stability when SNR exceeds −8 dB. Finals intelligibility has a large increase as SNR increases and when SNR is less than −4 dB, while the relative stability of the finals intelligibility is 0.9 when SNR exceeds −4 dB. The perceptual curves of initials and syllable intelligibility have the same trend. Both two curves increase as SNR increases and when SNR is less than 16 dB, while initials intelligibility stabilizes at 0.9 and syllable intelligibility stabilizes at 0.88 and when SNR exceeds 16 dB.

Initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility tested in speech noise environment are shown in Figure 5.

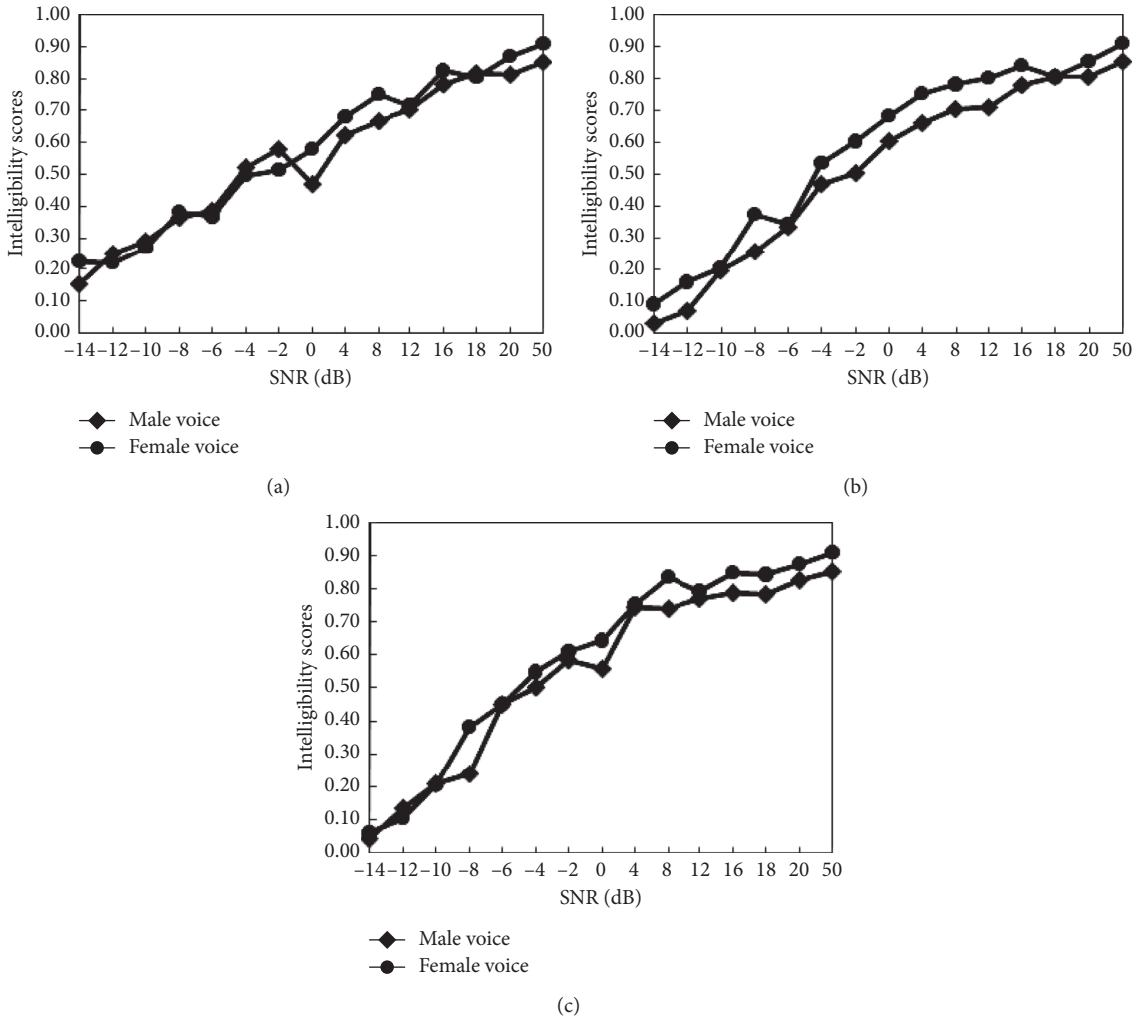


Figure 3: The influences of announcers' gender to Chinese speech intelligibility in noise environments. (a) White noise. (b) Pink noise. (c) Speech noise.

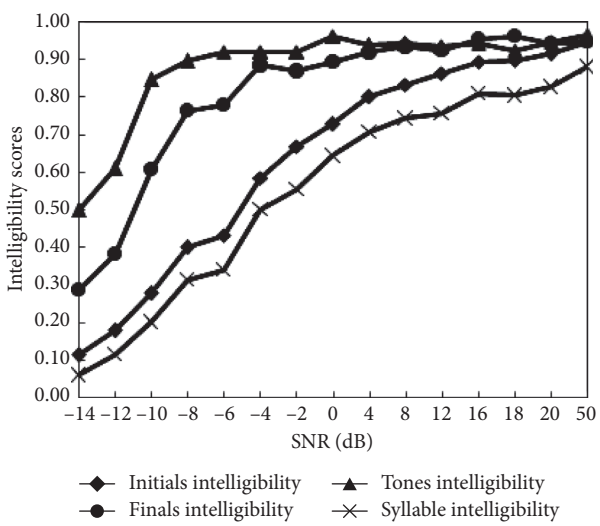


Figure 4: Chinese speech intelligibility in pink noise environment.

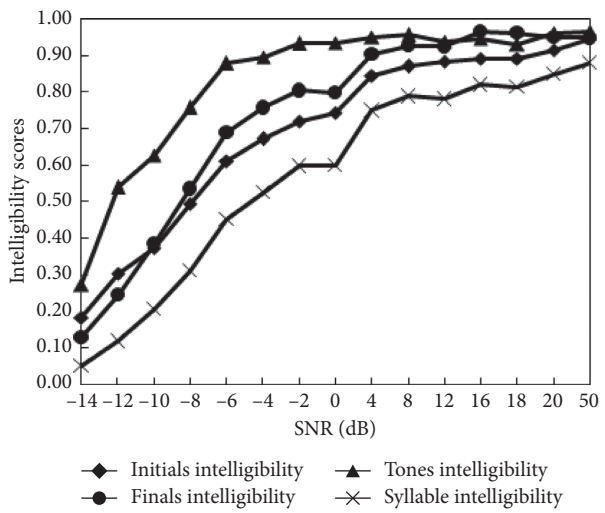


Figure 5: Chinese speech intelligibility in speech noise environment.

The perceptual curves of tones intelligibility, finals intelligibility, initials intelligibility, and syllable intelligibility test in speech noise environment have the same trend with the results tested in pink noise environment. Tones intelligibility increases as SNR increases. When SNR reaches −6 dB, tones intelligibility arrives at 0.9 and the curve is relatively stable. Finals intelligibility, initials intelligibility, and syllable intelligibility have a large increase as SNR increases, when SNR is less than 4 dB. When SNR greater than 4 dB, those curves increase slowly and tend towards stability.

Initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility tested in white noise environment are shown in Figure 6.

It is shown that the speech intelligibility increases as the SNR increases for all the four curves, and the results of tones intelligibility is the highest, then the finals intelligibility, thirdly the initials intelligibility, the syllable intelligibility the lowest. The perceptual curves of tones intelligibility remain in a high score (about 0.9) which means white noise has little influence on tones perception. Finals intelligibility increases significantly as SNR increases, when SNR is less than −8 dB, and increases slowly until the curve tends towards stability when SNR exceeds −8 dB. The measuring curves of initials and syllable intelligibility have the same trend. Both of the two curves increase as SNR increases, when SNR is less than 16 dB, and increase slowly when SNR is more than 16 dB. Initials intelligibility is affected most by noise, and syllable intelligibility is mainly determined by initials.

3.1.3. Influence of Noise Type on the Perceptual Intelligibility. Both temporal and spectrum characteristics of pink noise, speech noise, and white noise are different which lead to different influences on intelligibility perception. The comparison of the influence on initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility by pink noise, speech noise and white noise are shown in Figure 7.

It is shown that the curves of initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility have the same trend when speech signals are interfered by pink noise, speech noise, and white noise. Finals intelligibility and tones intelligibility have a little difference above the three noises especially when SNR is less than −6 dB, because white noise has little influence on finals and tones. Initials intelligibility and syllable intelligibility has no significant difference, and the three curves of syllable almost overlap.

ANOVA on initials intelligibility, finals intelligibility, tones intelligibility, and syllable shows that no significant differences (F (2,45) = 0.211, $p > 0.05$) were observed in initials intelligibility scores above the three noises. The same results also exist in the ANOVA on finals intelligibility ($F(2,45) = 1.071$, $p > 0.05$), tones intelligibility ($F(2,45) = 1.742$, $p > 0.05$), and syllable intelligibility ($F(2,45) = 0.029$, $p > 0.05$). Therefore, Chinese speech intelligibility scores were averaged across all of the data tested in the three noise environments for each SNR.

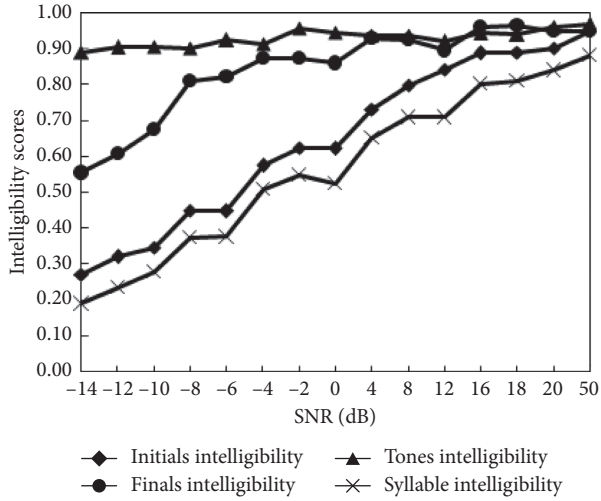


Figure 6: Chinese speech intelligibility in white noise environment.

Figure 7(d) shows that Chinese syllable intelligibility scores seem to be stable at a saturating value when SNR exceeds a certain value and the scores cannot reach 1. Even when no extra noise is interfered, perceptual intelligibility is also influenced by some unavoidable factors like vocal and receiving aspects which lead the systematic error in the experiments. The saturation of perceptual intelligibility was calculated after repeating tests on the original speech signals recorded in the anechoic chamber, and the results are shown in Table 2.

3.2. Perceptual Characteristics of Chinese Intelligibility in Noise Environment. Initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility of Mandrin are calculated, respectively, by averaging the results tested in pink noise, speech noise, and white noise conditions, and their scatter diagram is shown in Figure 8.

It is can be observed that the trends of speech intelligibility increasing with the SNR obey the exponential law. Speech intelligibility scores were fitted according to the least squares fitting method, and the fitted curves are also plotted in Figure 8. Fitting formulas are shown in equations (5) to (8), and fitting precision is shown in Figure 9.

$$Q_N = -0.33 \times e^{-0.07\times \mathrm{SNR}} + 0.92, \quad R^2 = 0.99, \tag{5}$$

$$Q_{N\mathrm{sm}} = -0.29 \times e^{-0.08\times \mathrm{SNR}} + 0.98, \quad R^2 = 0.99, \tag{6}$$

$$Q_{\mathrm{Nym}} = -0.09 \times e^{-0.15\times \mathrm{SNR}} + 0.96, \quad R^2 = 0.99, \tag{7}$$

$$Q_{\mathrm{Nsd}} = -0.01 \times e^{-0.25\times \mathrm{SNR}} + 0.95, \quad R^2 = 0.99, \tag{8}$$

where Q_N represents the syllable intelligibility in noise environment, Q_{Nsm} represents the initials intelligibility, Q_{Nym} represents the finals intelligibility, Q_{Nsd} represents the tones intelligibility, SNR represents the signal to noise ratio, and R^2 represents the fitting precision.

Equations (5) to (8) also provided mathematical models to predict the speech intelligibility of Mandarin according to

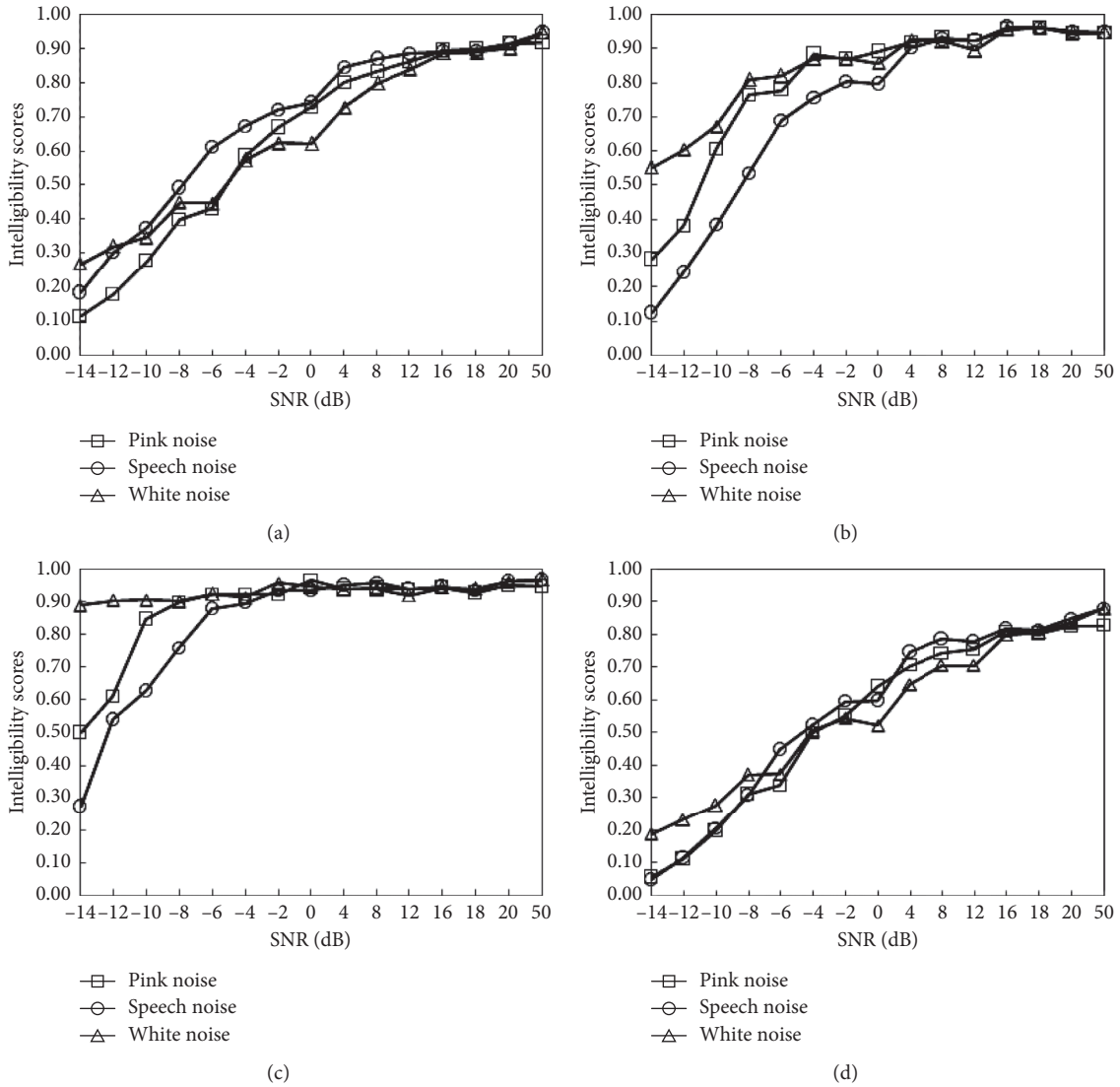


Figure 7: Influence of noise type on Chinese speech intelligibility. (a) Initials intelligibility. (b) Finals intelligibility. (c) Tones intelligibility. (d) Syllable intelligibility.

Table 2: Saturation of the perceptual intelligibility.

	Perceptual intelligibility
Initials intelligibility	0.93
Finals intelligibility	0.96
Tones intelligibility	0.96
Syllable intelligibility	0.86

SNR in the case of noise interference only and it had high precision with R^2 reaching 0.99. To observe the precision of the predicting models intuitively, a scatter diagram of measuring and predicting values of speech intelligibility was plotted in Figure 9.

Figure 9 shows that all of the points distribute around the central line, and it verifies that the predicting models had high precision on predicting initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility of Chinese Mandarin in noise-interfering environment which can be named CIPMNE (Chinese Intelligibility Predicting Model in Noise Environment). In the actual application, syllable intelligibility is only thing used to represent the quality of the transmission system, and the CIPMEN can be only defined as a method of predicting the syllable intelligibility.

4. Discussion

There is already a mature objective method of speech intelligibility called STI which is described in IEC 60268 [26]. As the calculation procedure of STI is too complicated, a simplified method of STI called STIPA is widely used in the actual measurement. STIPA can evaluate the transmission quality of speech with respect to intelligibility and has been verified to be effective for western languages. But, the language system of Chinese is different from western languages.

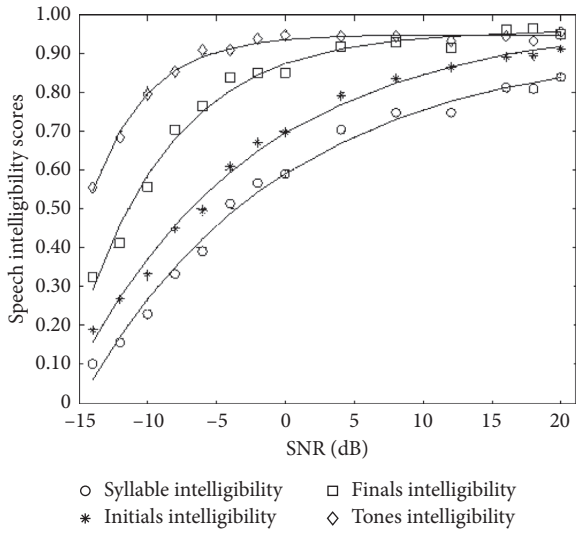


Figure 8: Perceptual Characteristics of Chinese intelligibility in noise environment.

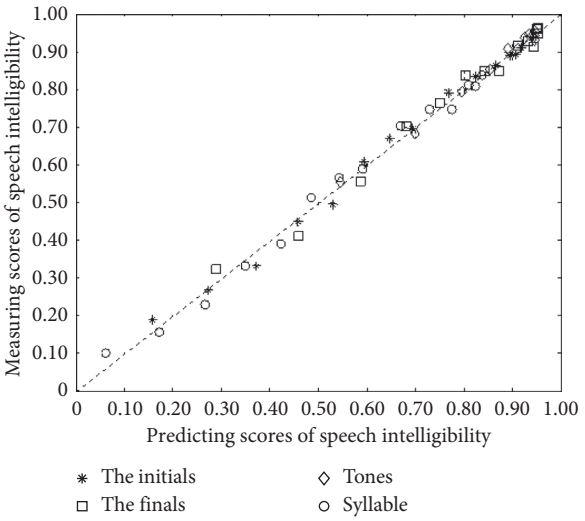


Figure 9: Scatter diagram of measuring and predicting values of speech intelligibility in noise environment.

Which method is fit for Chinese, STIPA or CIPMNE, needed to be tested.

Series data were obtained by STIPA and CIPMNE methods, respectively, under the same SNR, and the comparison of the results obtained by the two methods is shown in Figure 10. The differences between the results obtained from the two methods and subjective syllable intelligibility scores are shown in Figure 11.

Figure 10 shows that the results of CIPMNE and STIPA under the same SNR are obviously different. CIPMNE is greater than STIPA when SNR is less than 16 dB and less than STIPA when SNR is more than 16 dB. Especially when SNR is around 0 dB, the difference almost reached 0.3.

It is should to be noticed in Figure 11 that the curves of CIPMNE and syllable intelligibility are almost have the same trends but different from STIPA. Because CIPMNE is based

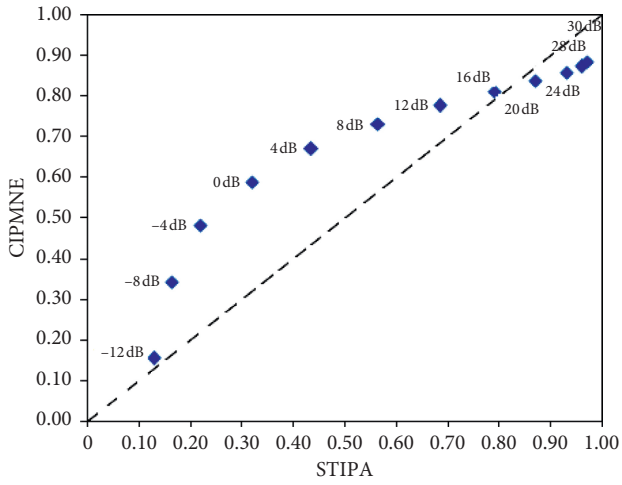


Figure 10: Scatter diagram of CIPMNE and STIPA.

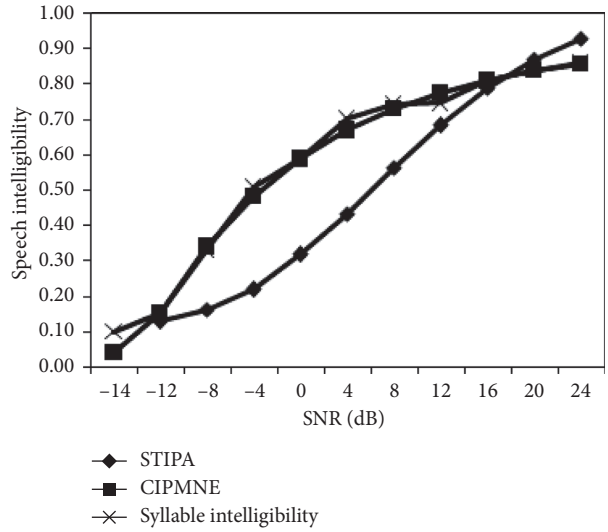


Figure 11: Speech intelligibility measured by STIPA, CIPMNE, and subjective listening experiments.

on the perception of syllable intelligibility. The results of speech intelligibility varying with SNR are different when measured by CIPMNE and STIPA, and the results measured by CIPMNE are closer to subjective results. The aim of the researches on the objective measurement method of the speech intelligibility is to approach the subjective perception results as much as possible, and instead of subjective measurement, maybe CIPMNE is fit for Chinese speech intelligibility measurement in the transmission condition of noise-interfering only, or STIPA needs to be revised according to the Chinese speech intelligibility obtained by the subjective listening experiments.

5. Conclusions

The present work introduced a series of Chinese speech intelligibility subjective experiments in the environment deteriorated by white noise, pink noise, and speech noise. Initials intelligibility, finals intelligibility, tones intelligibility,

and syllable intelligibility were obtained, and the changing rules of Chinese speech intelligibility with SNR were analyzed. The results showed that the influences of the three types of noises on perceptual intelligibility have no significant difference. Tones intelligibility scores are the highest, then the finals intelligibility, thirdly the initials intelligibility, and the syllable intelligibility the lowest, which means that the tones have the strongest antinoise property and the property of syllable is the weakest. Speech intelligibility increases with SNR obeying exponential low values and tends to get saturated gradually. The best-fitting exponential function relationships between initials intelligibility, finals intelligibility, tones intelligibility, and syllable intelligibility and SNR, respectively, were established on the basis of the least square method and a model called CIPMNE was proposed based on these functions to evaluate and predict the Chinese speech intelligibility under noise-interfering transmission conditions. A comparison of CIPMEN and STIPA was also made to analyze the differences between the two methods on the Chinese speech intelligibility measurement.

Data Availability

The data used in this research are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Nature Science Foundation of China(11204278), Education Department of Jilin Province Research Funds (JJKH20170155KJ), and Doctoral Research Funds of Northeast Electric Power University (BXJXM-2017219).

References

[1] H. J. M. Steeneken and T. Houtgast, "Basics of the STI measuring method," in *Proceedings of the Past, Present, and Future of the Speech Transmission Index, International Symposium on STI*, pp. 13–44, Soesterberg, Netherlands, October, 2002.

[2] S. Zhang, H. Song, and Z. Meng, "Relationship between Chinese Mandarin intelligibility and speech transmission index STIPA under simulated tranmission conditions," in *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, IEEE, Chengdu, China, July 2015.

[3] T. Houtgast, "The effect of ambient noise on speech intelligibility in classrooms," *Applied Acoustics*, vol. 14, no. 1, pp. 15–25, 1981.

[4] J. S. Bradley, "Speech intelligibility studies in classrooms," *The Journal of the Acoustical Society of America*, vol. 80, no. 3, pp. 846–854, 1986.

[5] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[6] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 387–399, 2010.

[7] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2999–3012, 2011.

[8] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3988–3997, 2006.

[9] K. Ishikawa, S. Boyce, L. Kelchner et al., "The effect of background noise on intelligibility of dysphonic speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 7, pp. 1919–1929, 2017.

[10] J. S. Bradley, R. D. Reich, and S. G. Norcross, "On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1820–1828, 1999.

[11] S. J. Van Wijngaarden, H. J. M. Steeneken, T. Houtgast, and T. Houtgast, "Quantifying the intelligibility of speech in noise for non-native listeners," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1906–1916, 2002.

[12] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Prediction of the intelligibility for speech in real-life background noises for subjects with normal hearing," *Ear and Hearing*, vol. 29, no. 2, pp. 169–175, 2008.

[13] H. Luts, E. Boon, J. Wable, and J. Wouters, "FIST: a French sentence test for speech intelligibility in noise," *International Journal of Audiology*, vol. 47, no. 6, pp. 373-374, 2008.

[14] D. S. Boon, "Evaluation of speech intelligibility with the coordinate response measure," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2276–2279, 2001.

[15] I. Pollack and J. M. Pickett, "Masking of speech by noise at high sound levels," *The Journal of the Acoustical Society of America*, vol. 30, no. 2, pp. 127–130, 1958.

[16] L. L. Elliott, "Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability," *The Journal of the Acoustical Society of America*, vol. 66, no. 3, pp. 651–653, 1979.

[17] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1581–1592, 1994.

[18] Van Wijngaarden and J. Sander, "Intelligibility of native and non-native Dutch speech," *Speech Communication*, vol. 35, no. 1-2, pp. 103–113, 2001.

[19] J. Zeng, J. Peng, and Y. Zhao, "Comparison of speech intelligibility of elderly aged 60-69 years and young adults in the noisy and reverberant environment," *Applied Acoustics*, vol. 159, Article ID 107096, 2020.

[20] C. Visentin, N. Prodi, F. Cappelletti, S. Torresin, and A. Gasparella, "Speech intelligibility and listening effort in university classrooms for native and non-native Italian listeners," *Building Acoustics*, vol. 26, no. 4, pp. 275–291, 2019.

[21] A. J. Duquesnoy, "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 739–743, 1983.

[22] D. Kostić, "The influence of musical noise, type major and minor chord, to the intelligibility of speech in Serbian language," in *Proceedings of the UNITECH*, Gabrovo, Bulgaria, December 2017.
[23] GB/T. 15508-1995, *Speech Articulation Testing Method (In Chinese)*, Standards Press, Beijing, China, 1995.
[24] F. Zhao and B. Shi, "Acoustic processing of anechoic chamber with short reverberation," *Audio Technique*, vol. 5, pp. 22–24, 2004, in Chinese.
[25] S. Zhang and H. Song, "Chinese Mandarin speech intelligibility under band-limited transmission conditions," in *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, IEEE, Chengdu, China, December 2017.
[26] IEC 60268-16-2011, "Sound system equipment–part 16: objective rating of speech intelligibility by speech transmission index," 2011.

Hindawi
Scientific Programming
Volume 2020, Article ID 8847481, 17 pages
https://doi.org/10.1155/2020/8847481

Research Article

Massively Parallel CFD Simulation Software: CCFD Development and Optimization Based on Sunway TaihuLight

Xiazhen Liu,[1,2] Zhonghua Lu,[1,2] Wu Yuan,[1,2] Wenpeng Ma,[3] and Jian Zhang[1]

[1]*Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*College of Computer and Information Technology, Xinyang Normal University, Xinyang, Henan 464000, China*

Correspondence should be addressed to Zhonghua Lu; zhlu@cnic.cn

Received 16 April 2020; Revised 17 June 2020; Accepted 29 June 2020; Published 22 July 2020

Academic Editor: Chenxi Huang

A parallel framework software, CCFD, based on the structure grid, and suitable for parallel computing of super-large-scale structure blocks, is designed and implemented. An overdecomposition method, in which the load balancing strategy is based on the domain decomposition method, is designed for the graph subdivision algorithm. This method takes computation and communication as the limiting condition and realizes the load balance between blocks by dividing the weighted graph. The fast convergence technique of a high-efficiency parallel geometric multigrid greatly improves the parallel efficiency and convergence speed of CCFD software. This paper introduces the software structure, process invocations, and calculation method of CCFD and introduces a hybrid parallel acceleration technology based on the Sunway TaihuLight heterogeneous architecture. The results calculated by Onera-M6 and DLR-F6 standard model show that the software structure and method in this paper are feasible and can meet the requirements of a large-scale parallel solution.

1. Introduction

Computational fluid dynamics (CFD) is a technique for numerical simulation and analysis of fluid mechanics problems. CFD technology is increasingly used in aerospace, meteorological prediction, and other fields [1–4]. Some parallel computing frameworks have been used with parallel CFD [5, 6], e.g., OpenFOAM [7] and SU^2 [8, 9]. Parallel computing methods are increasingly used to solve large-scale computationally intensive problems. Parallel programming environments such as OpenMP [10, 11] and MPI [12, 13] have appeared on parallel machines, networked workstations, and supercomputers. The development of CFD in numerical methods, turbulent models, and mesh generation technology has led to its making strides in the simulation accuracy and capacity of complex geometric shapes. Meanwhile, with the increasing complexity of engineering problems and the rapid advancement of numerical simulation calculation technology, the requirements for simulation accuracy are getting increasingly stringent, and the amount of calculation required has increased geometrically. Parallel CFD technology has become the primary method for solving complex simulation calculations [14, 15].

Current CFD methods use a computational grid to simulate the flows in complex geometries and form different cells through the discreteness of the grid, which is then followed by the numerical calculation [15–17]. In order to simulate the shape structure more realistically, a multiblock approach is generally used to simulate different spatial regions. A commonly used method for solving large-scale CFD problems is applying domain decomposition technology in order to decompose more subblocks and allocating these decomposed calculation regions to different processes or threads for performing parallel calculations [18–20]. The methods of parallel computing are very different under different parallel programming environments. The MPI method allocates one or more computing areas to a processor. As the different computing areas are completely isolated, the data exchange at the boundary needs to be completed through communication. The OpenMP method

uses the shared memory in an environment in which the computing area is actually on one processor, and the grid data are completely shared, rendering data communication unnecessary. Therefore, designing different parallel frameworks is important when using different programming environments and grid types [21, 22].

Supercomputer architectures are generally either homogeneous [23] or heterogeneous [24]. Each core architecture of a homogeneous machine is the same, and their status is equal. They can share the same code or execute different codes on each core [25]. Homogeneous processors can be interconnected using shared storage or through cache [26–28]. Heterogeneous architectures represent a new hardware platform that allows different types of processors to work together efficiently in shared memory [24]. The heterogeneous architecture includes not only the traditional CPU but also other accelerator units such as GPUs and MICs [24, 29]. These different computing units have different instruction set architectures and memory spaces [30]. The CPU and the accelerators have unified access to the system memory through the Memory Management Unit (MMU) [31, 32]. With the evolution of supercomputer technology, the traditional homogeneous architecture has been unable to meet the increasing requirements for computing power and storage. For this reason, the heterogeneous architecture has become the most important technology for the development of supercomputers. Heterogeneous architecture machines account for more than 30% of the current TOP500 list of supercomputers [33]. Of the top five parallel computers, four have heterogeneous architectures.

The main purpose of the research content in this paper is to develop an open-source large-scale parallel solver based on a multiblock structure grid. First, multicore parallel computing based on the homogeneous architecture was implemented. Second, based on the Sunway TaihuLight heterogeneous system [34], the calculation of data on the CPU and accelerator communication were achieved using the MPI + OpenACC/Athread hybrid programming model [35] and direct memory access (DMA) technology [36]. Third, through the use of stencil calculation, the boundary information is exchanged through register communication, and the single instruction, multiple data (SIMD), and assembly instructions are used to optimize the computationally intensive area and further improve calculation performance. Finally, the parallel simulation of the Onera M6 wing model [37] realizes a parallel calculation of 500,000 cores of 1 billion grids and achieves ideal parallel acceleration efficiency.

2. CCFD Software Design and Implementation

2.1. CCFD Software Architecture. CCFD is designed by object-oriented programming [38], and its architecture is shown in Figure 1. The entire software platform consists of an input/output control module, a computational geometry module, a solver module, and a parallel algorithm module. The system architecture of CCFD needs to take into account the characteristics of various calculation methods and physical models and design a simulation software with flexibility, robustness, accuracy, and data security. In addition, the parallel framework design needs to meet the requirements for high parallel scalability, high data transmission efficiency, and fast response time. To increase maintainability and scalability, the CCFD architecture is organized with blocks as the basic data structure units. Based on this, the use and loading of data by each module in the platform greatly simplifies the program interface and reduces the difficulty of development.

The CCFD software platform system is built on block units. The structure and basic unit calling of CCFD are the blocks and boundaries. The internal points and boundary points of the blocks are divided into two data structures: BLOCKs and BCs. The main information such as geometric variables and calculation variables are stored in the block, while the boundary type and starting information are stored in the boundary. In accordance with the advantages and disadvantages of an array of structures (AoS) and a structure of arrays (SoA), the variables of CCFD on the block are organized in the form of SoA. The advantages of this method can be stated as follows. On the one hand, the centralized use and loading of grid variables improves the hit rate and data utilization of cache. On the other hand, it facilitates the packaging and collection operation of parallel communication, which is beneficial to the parallel optimization of the large-scale computation of CCFD.

CCFD integrates various computing models and methods effectively. In order to facilitate the software module specifications, code organization development, and extended maintenance, we organize the entire software system with the application layered model, which makes the code modules independent between layers and interconnects the calling relationships. The solver can be divided into five layers. The first layer is grid input, which controls the file input and the load balancing of block calculation and communication on the processor. In the second layer, the calculation is divided into a multiblock method and conventional calculation methods according to the calculation demand. The third layer is divided into Euler, N-S, and turbulent model based on the flow equation. The fourth layer is a method of selecting spatial and temporal terms for different equations. The fifth layer is a basic algorithm layer that includes a sparse matrix solving method, modular parallel communications, and error information output.

2.2. CCFD Software Process. Software process design needs to fully take into account the actual calculation methods used. For example, CCFD uses a multiblock structure grid based on a second-order precision finite volume method to solve partial differential equations and uses multigrid methods to accelerate convergence techniques. In order to allow large-scale structure blocks to be calculated on parallel machines, the domain decomposition technique is used to divide a large-scale block into smaller-scale subblocks. The adjacency relationships between the subblocks are stored in the information of the block, and the information is shared and updated after each iteration of calculation to ensure the compatibility between the results of the partial differential

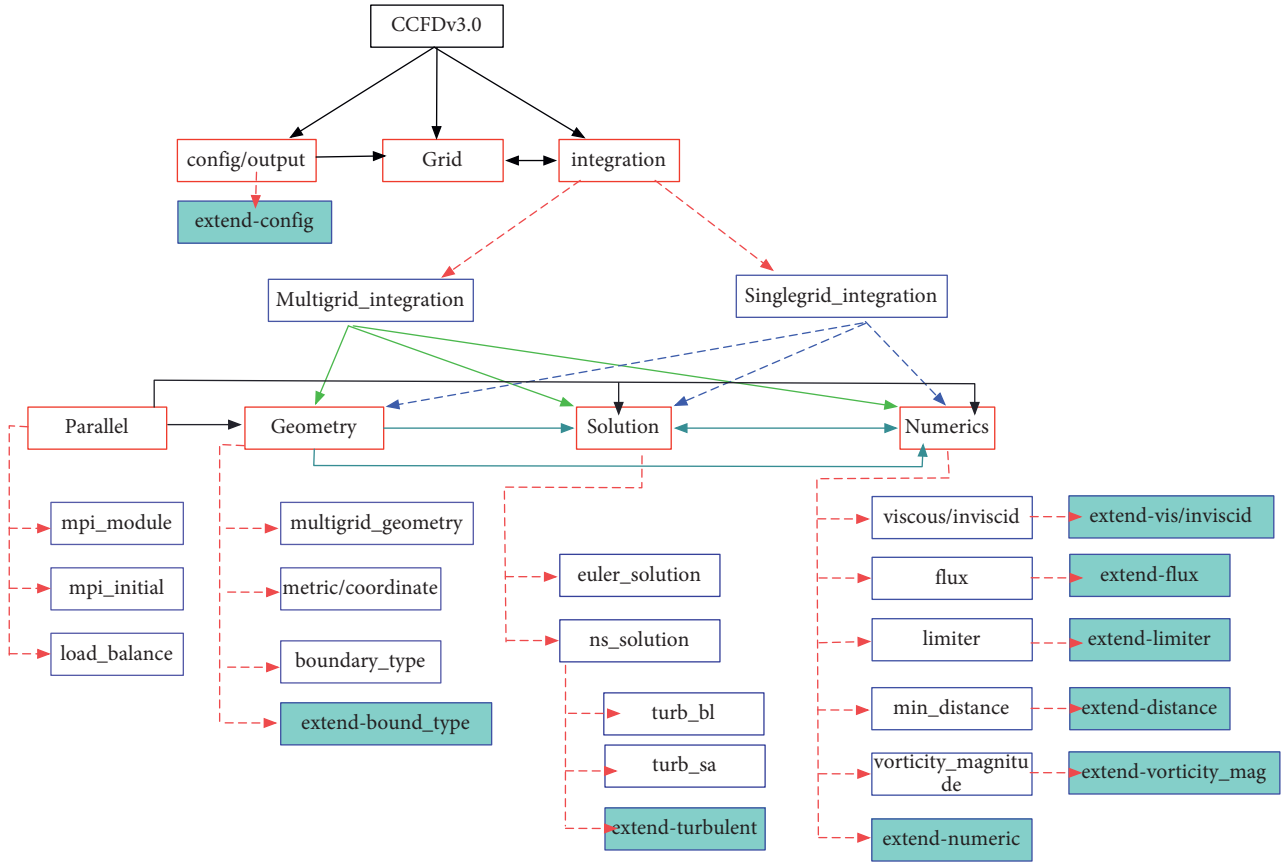


Figure 1: CCFD software framework.

equations and the unpartitioned block. The area of the block is assigned to different processors for calculation. This mapping and allocation process involves the calculation and communication load balancing of the block on the processor. Therefore, the entire software system can be thought of as being roughly divided into five parts: preprocessing, area decomposition, load balancing, flow field calculation, and postprocessing output.

The software process design considers the calculation methods used in practice and meets the communication needs between the multiblocks. Regardless of partition parallel computing or multiblock computing, it is essential to calculate the partial differential equations on each block. The calculation time of subblocks of different sizes differs depending on the process. The larger the scale, the larger the calculation time. If the subblock sizes are not uniform, there will be multiple processes waiting for individual processes. Based on the above characteristics of CCFD software, we designed a software parallel process, as shown in the flowchart in Figure 2.

The unit data based on the block, which is the functional module structure of CCFD, are shown in Figure 3. The structure is divided into eight modules. The calling and connection between each module are based on the block unit. The main body of the calling is the parallel solver. Through the parallel solver, the block geometry module, parallel module, source term solution, right term solution, turbulent viscous, and output control are used to complete the entire solution process. The connection lines in Figure 3 indicate the link relationship between the modules. The entire process largely determines the parallel expansion efficiency of the software system. The allocation strategy of blocks to processors also affects how much the parallel efficiency can be improved. For example, an imbalance in the calculation and the communication load between blocks on the process will cause the process to wait idle for the overall calculation.

3. CCFD Parallel Design

3.1. Load Balancing Strategy. One goal of domain decomposition technology is to subdivide a large size of block into smaller subblocks. The subblocks' computing area is allocated to different processes, or threads, for parallel computing processing [19, 20]. As shown in Figure 4, a three-dimensional structure block is subdivided into multiple smaller subblocks, each having multiple adjacency surfaces, which represent the communication relationship between the subblocks. When the subblocks are calculated on different processors, the adjacent surfaces need to carry out data transmission and communication.

Once the domain is decomposed, a larger number of smaller, more uniform blocks are formed. These subblock computing areas are allocated to different processes, or threads, for concurrent execution. In this particular process, we need to ensure that the calculation and communication

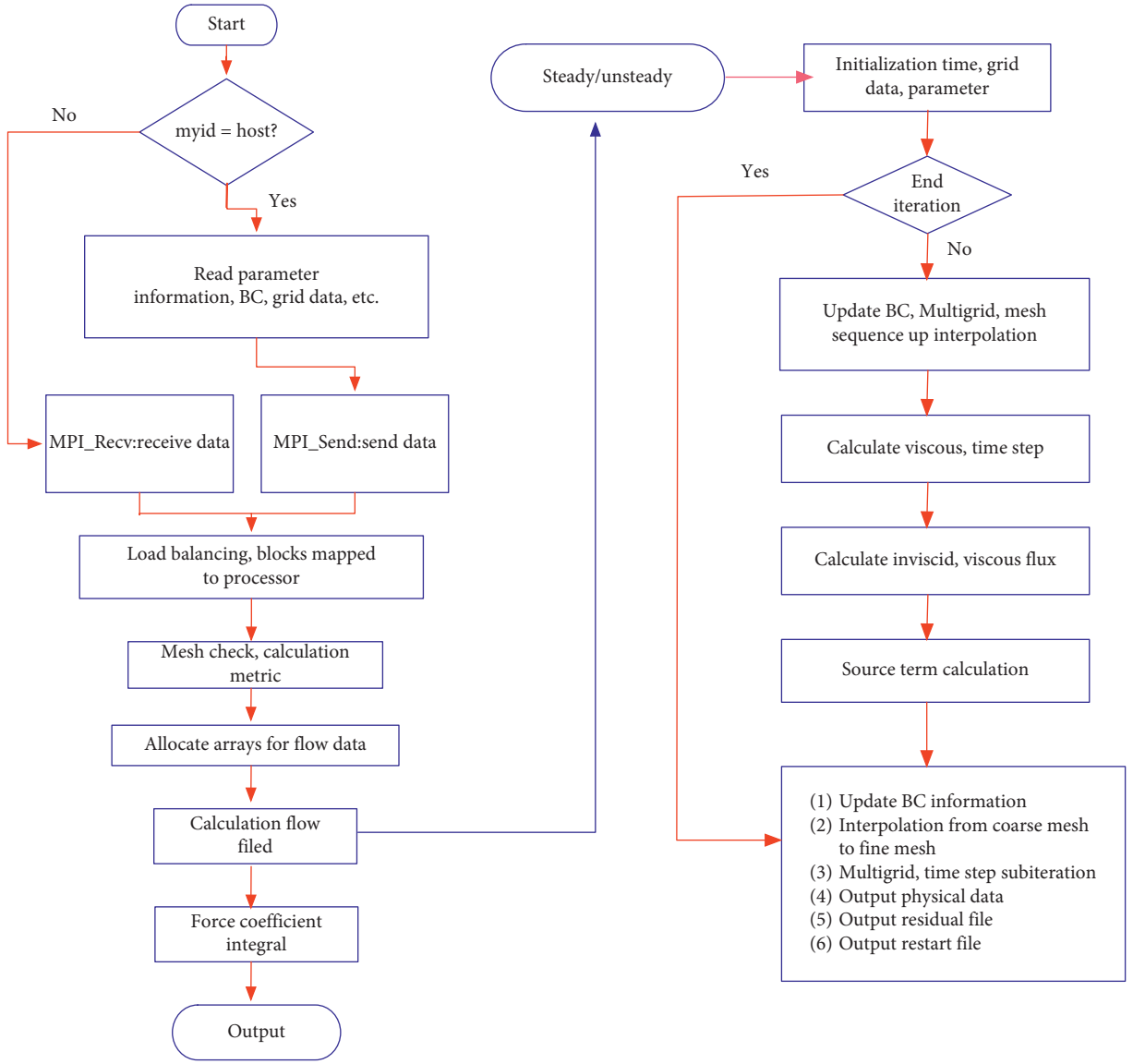


Figure 2: CCFD software process.

allocated to each processor are relatively balanced, and there is no idle waiting of the processor. This process of block and course mapping allocation is called load balancing. Currently, there is much research taking place on load balancing methods, and several algorithms have been proposed, including greedy algorithm, spectral method, genetic algorithm, and ant colony algorithm [39]. These algorithms are applicable to unstructured grids but cannot be directly applied to structured grids. Traditional structure grid load balancing methods generally use a cyclic allocation method or a uniform distribution method, neither of which consider the load balance of the calculation amount and communication [20].

Aiming at the problem of large-scale multiblock structure grid computing and communication load balancing, CCFD scientists proposed an overdecomposition method. CCFD divides the blocks and processor mapping and distribution process into two layers: coarse and fine. At the coarse layer, domain decomposition methods such as recursive boundary are used to subdivide large and uneven blocks into smaller and more uniform subblocks. The algorithm is a method of approximate linear time complexity, and the decomposition time is related to the number of blocks and the maximum dimension of the block. The purpose of the subdivision operation in the coarse layer is to lay a foundation for further balancing the calculation and communication loads. For example, uneven blocks make it difficult to balance the calculation and communication allocation, which will result in a series of processor-independent block sequences. Before using the fine-layer load balancing algorithm, the subblocks of the coarse layer need to be renumbered. The number of grids in each subblock was used as the statistical standard for the calculation amount, and the adjacent faces of each subblock and other subblocks are the statistical standard for communication volume. This information is redrawn into a multidirectional weighted

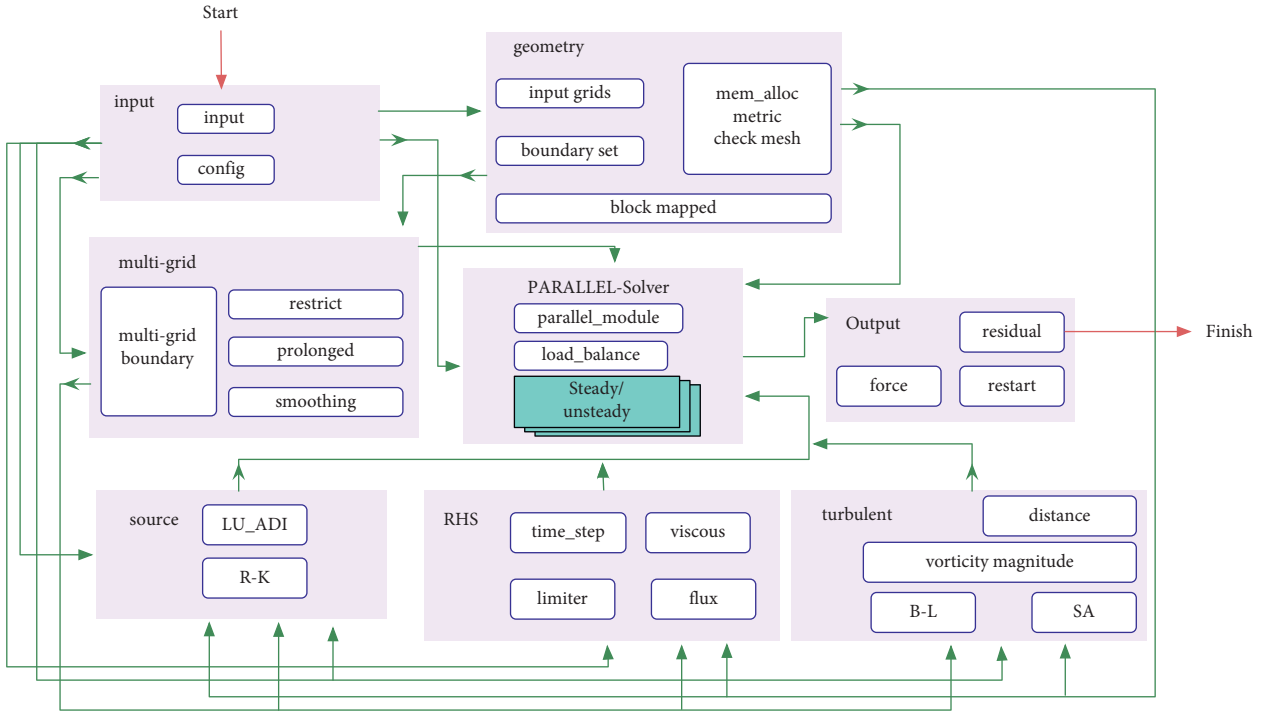


Figure 3: CCFD method and module invocation relationship.

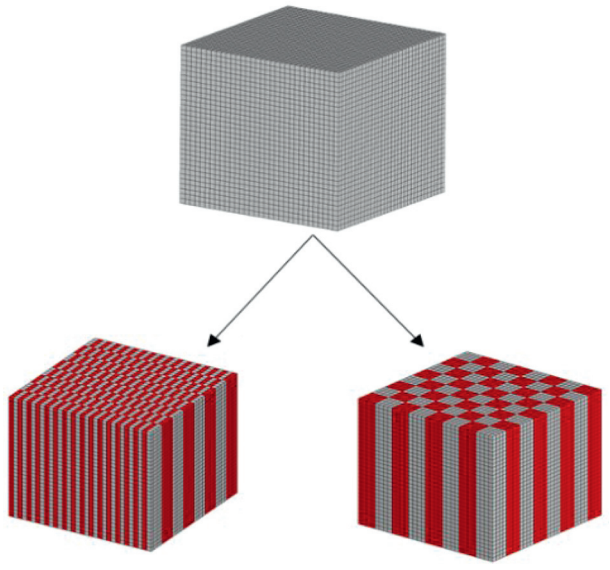


Figure 4: Domain decomposition.

graph based on the subblock topology, and the communication volume and calculation volume are the two weights of the weighted graph [40]. The fine-layer graph partitioning algorithm performs graph segmentation based on the weighted graph formed by the upper layer. The total amount of calculations of the subdomains formed as far as possible is balanced, and the total amount of communication in the subdomains is balanced with the smallest amount of communication between subblocks [41]. Figure 5 shows a multidirectional weighted graph $G = (V, E)$ drawn after preprocessing the CCFD blocks, and finding a partitioning of the vertices of G into n sets in such a way that the sums of the vertices weights in each set are as equal as possible, and the sum of the edges weights crossing between sets is minimized. Therefore, this partitioning of the vertices guarantees the computation in each processor is equal, and the communication between processors is minimized. The method based on k-way graph partitioning has moderate computational complexity, its complexity for computing is $O(|V|\log k)$, and $\log k$ is levels of recursive bisection for the algorithm.

In this paper, we discuss how we found that the calculation scale and the number of subblocks of the calculation example have a direct relationship with the amount of calculation and communication load balancing. For example, the number of subblocks and the size of the subblocks directly affect the distribution of the amount of calculation in the process. The greater the number of subblocks and the balance of the size, the more balanced the distribution of the calculations. However, the greater the number of partitions, the larger the adjacent area of the block. This condition directly leads to an increase in communication and affects the realization of communication load balancing among the blocks. As an example, the block size in the M6 calculation ranges from 11,000 to 13,000, and the area of cut face is between 450 and 625. The block size and the adjacent surface after domain decomposition are relatively balanced, which is conducive to load balancing. Figure 6 shows the actual calculation load distribution of each process in the M6 calculation example, using 256 or 512 threads.

Testing the M6 model, we found that, with the increase in parallel scale, the actual distribution of the computing load will become worse, along with an increase in the

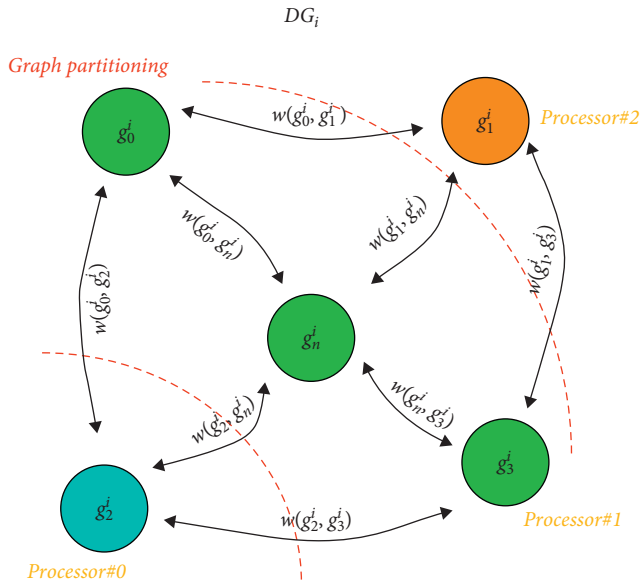


Figure 5: Multiblock structured grids for graph partitioning.

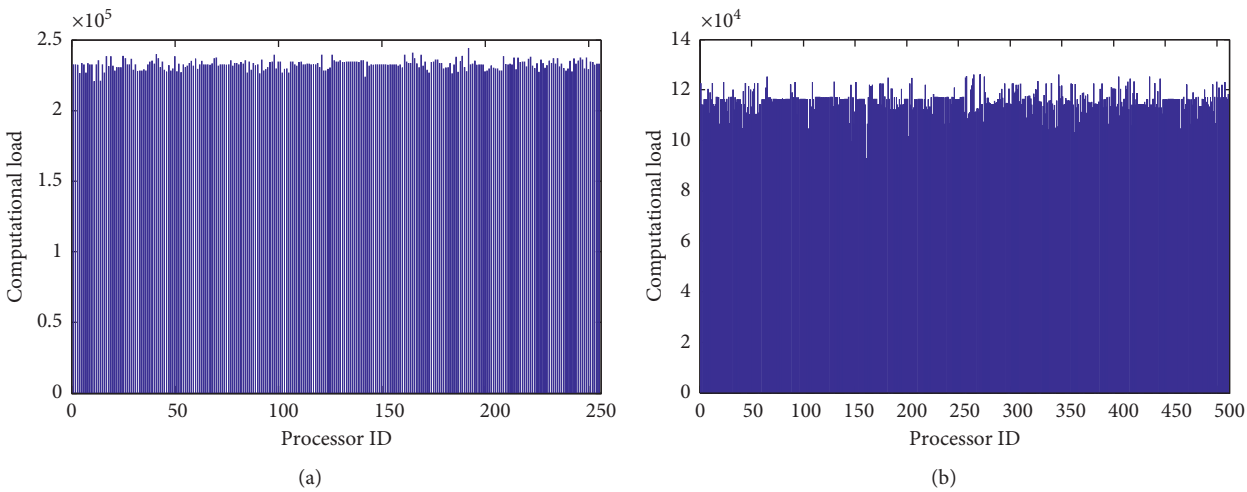


Figure 6: Calculation load distribution in the use of 256 and 512 threads.

number of threads. In this case, 256 threads are the critical point of the calculation load balancing. When the number of threads is less than 256, the overall computational load balance is ideal. However, the actual distribution of computational load will be significantly different when the number of threads is greater than 256. When the subblock size and the total number of blocks in the partition are stable, the number of blocks allocated to each process decreases, with the number of threads increasing. Then, the influence on the distribution of the calculation amount caused by the difference in the number of blocks on each thread and the actual size of the block increases, resulting in a decrease in the degree of interprocess calculation load balancing.

After the domain decomposition, the total calculation amount of the block remains stable, and the absolute amount of communication between the blocks increases with the increase of the number of parallel cores. In terms of communication load balancing between blocks, the degree of communication load balancing decreases with the increase in parallel scale. Figure 7 shows the actual communication load distribution of each process in the M6 calculation example, using 256 and 512 threads.

3.2. Parallel Data Structure. CCFD supports two grid formats: CFD General Symbol Standard (CGNS) grid and PLOT3D. The operation of the software system involves the access to and update of various data, which mainly include control parameters, grid data, flow field data, and temporary variables. The specific block variables are as follows:

```
type, public: blocks_type
type (cell_type), pointer, dimension(:,:,): cell
type (coordinate_type), pointer, dimension(:,:,): coordinate
type (metric_type), pointer, dimension(:,:,): metric
type (variable_types), pointer, dimension(:,:,): variable
```

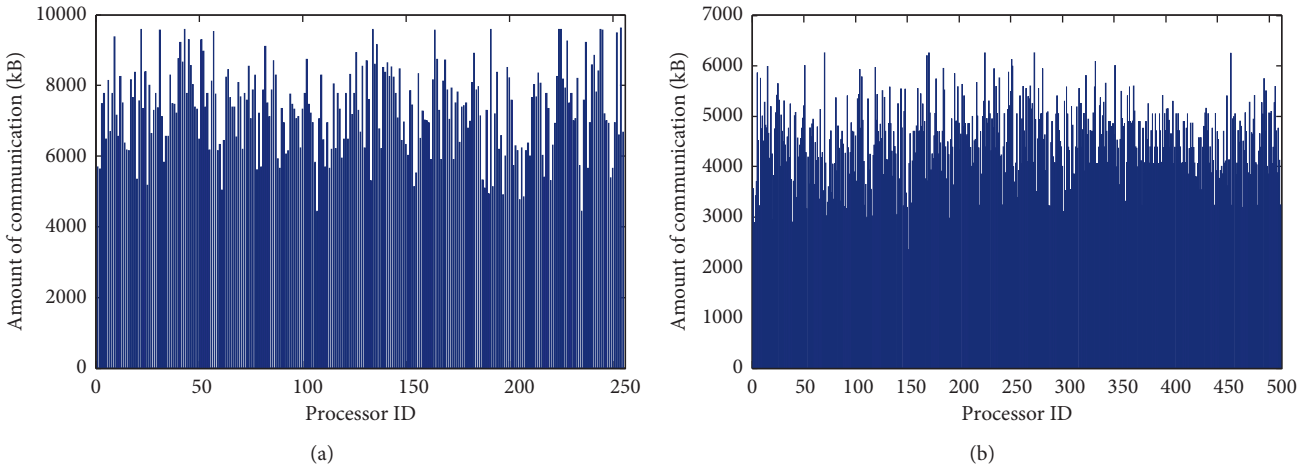


Figure 7: Communication load distribution using 256 and 512 threads.

type (variable_mg_types), pointer, dimension(:,:,): variable_mg

type (turbulent_type), pointer, dimension(:,:,): turbulent

end type blocks_type

All data can be divided into global, local, and exchange data. Global data refer to the data that all processors will save and not modify during calculation, such as control parameters. Local data refer to the data that are only stored locally and are not shared with other processors, such as flow field data in blocks' interior points. Exchange data refer to the data that need to be exchanged with other processes, such as boundary data of adjacent faces.

When large-scale parallel computing is performed in CFD, huge data reads and exchanges offer a great test to the data structure and data flow. The reasonable data structure and data flow design directly affect the reliability and efficiency of the platform operation. In order to ensure the efficiency and security of data, the design of CCFD parallel software is based on the block data structure. The grid data, boundary data, and flow field data are stored in each block unit. As shown in Figure 8, a block stores information such as grid coordinates, boundary information, and primitive variables [40, 41].

3.3. Parallel Communication Optimization. The proportion of blocks computing and communicating is one of the main factors in the parallel expansion performance of software. Ideally, a large amount of local computing is a computational priority problem, and a small amount of communication is good for parallel expansion. If the amount of local computing is small, the large amount of communication is a communication priority problem, which leads to poor parallel scalability. The computing model and method used by parallel software determine the amount of local computing, and the communication mode of parallel software affects the amount of communication. For example, in actual calculations, there are some challenges: many small and frequent data communications, a large amount of data and a fixed number of a block boundary updates, one-to-many broadcast data, and many-to-one master-slave communication. In terms of these challenges, using a different and flexible communication mode facilitates the software's parallel expansion.

CCFD uses the mode of traversing and allocating blocks to communicate. Each process traverses all block numbers. Only the local blocks are calculated, while the nonlocal blocks are used for asynchronous communication according to the actual communication needs. For example, when process 1 communicates with process 2, each process packs the mesh belonging to the adjacency surface relationship with nonblocking transmissions and then directly performs subsequent calculations. It does not detect whether the communication has been completed until the communication data are used locally. If the communication is completed, it can be used directly; otherwise, wait for the communication to complete. In Figure 9, B4 and B6 are the respective senders and receivers of data. Under nonblocking communication, B4 and B6 only need to send data to perform subsequent calculations and do not need to wait for the completion of data reception.

In the actual calculation of CCFD, there is a large amount of small and frequent communication, such as the detection of the local subblock, the updating of the local subblock, the updating of the flow field data, the calculation of the turbulent model, and the multigrid up-down interpolation process, which occurs after the block is read. These small and high-frequency communications are very time-consuming and affect the parallel efficiency of the processor. To solve these problems, we generally use methods such as collective communication, data prefetching, and computational alternative communication. Collective communication packs the number of high-frequency communication times in order to reduce the number of communication requests, which in turn reduces the communication time. The aim of prefetching technology is to obtain data in advance in each calculation iteration and hide the communication time. The alternative method of calculation refers to using calculations to complete the local results, which can

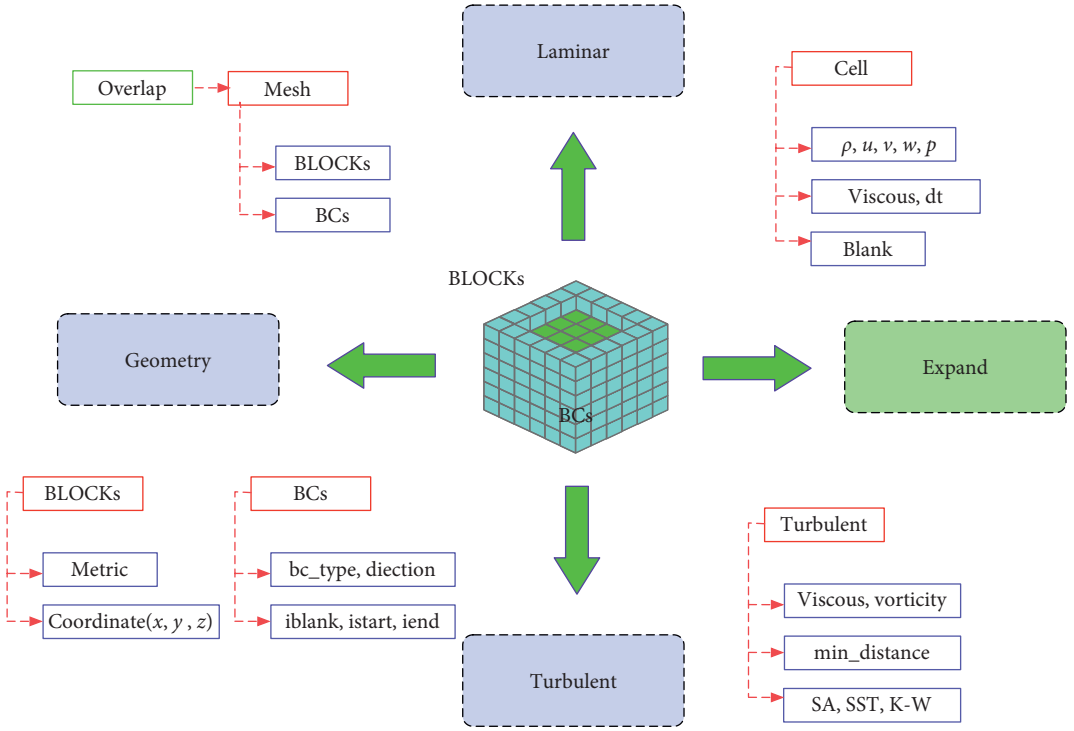


Figure 8: Data structure.

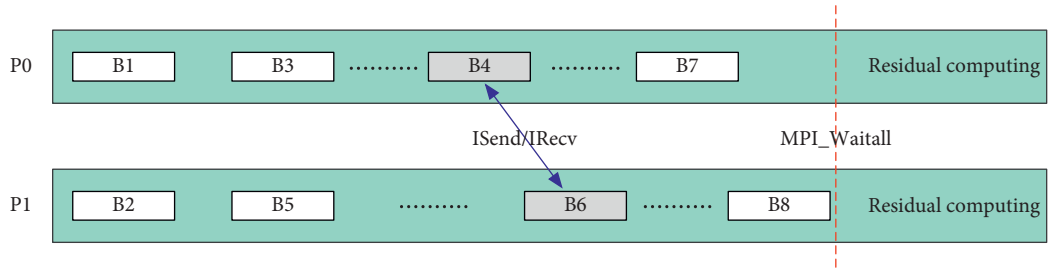


Figure 9: Nonblocking communication mode.

improve data utilization and reduce communication times and data volume.

When there are many uneven blocks in the local process, the program is prone to idle waiting in the blocking or sequential traversal execution mode. For example, the small block quickly ends the calculation and enters the communication. Large blocks have not yet entered the communication, and the data are not ready, which will cause a long wait. The overlap of calculation and communication is shown in Figure 10. Each communication request ends after the data are sent, and the detection of the completed communication is not performed until the data to be received are used in the calculation process. For example, processor P1 and processor P2 send data to each other, but these data are not necessarily used immediately once the communication ends. In this case, we can send the data to the background for continued processing, and the processor can perform subsequent calculations until the data are used and whether they have been received is detected. This method allows data communication and calculation to be performed at the same time. Because the communication time is hidden under the calculation time, it is called the overlapping technology of calculation and communication.

The design and selection of communication modes largely determine the computing efficiency and parallel scalability of parallel software. Different computing modes and methods and data structures require the design of a reasonable and flexible communication mode according to the actual situation. CCFD parallel software adopts flexible communication modes and carries out different parallel optimization designs for different calculation models and methods. Moreover, it also presents the modular design for the general data communication methods, which greatly simplifies the process of program programming and maintenance and achieves ideal parallel efficiency and data utilization.

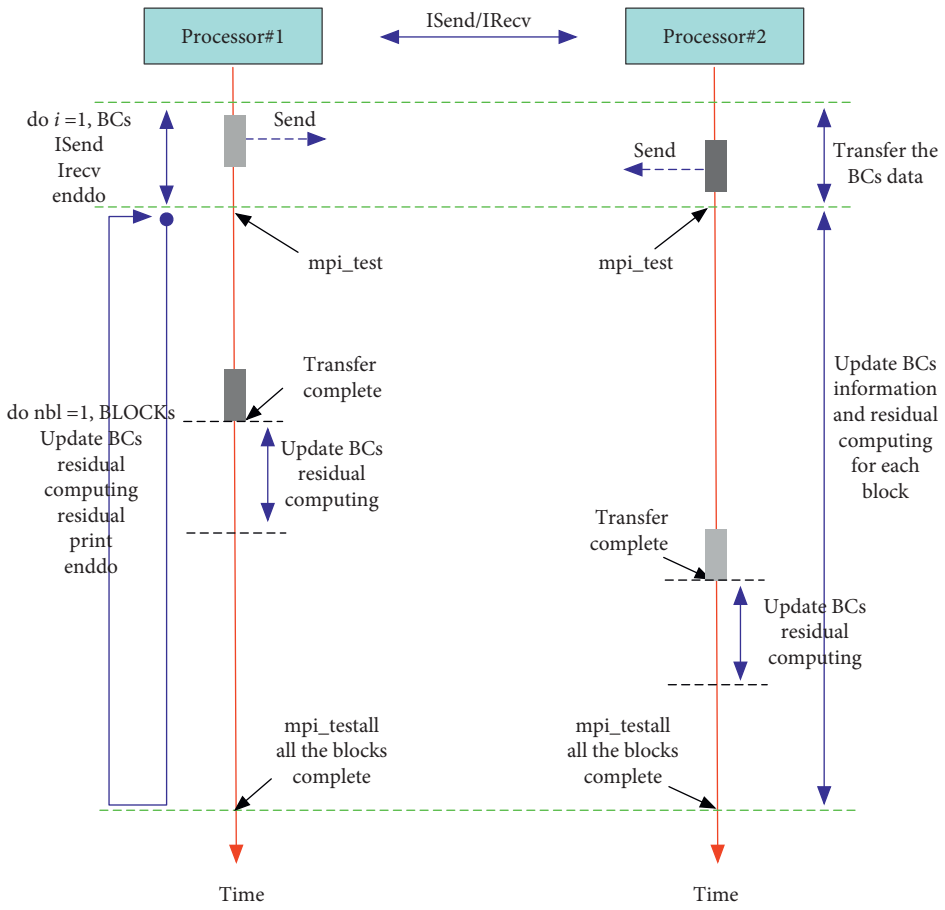


Figure 10: Communication and computation overlap.

4. Optimization on Sunway TaihuLight Heterogeneous Platforms

Sunway TaihuLight is the first 100P heterogeneous system supercomputer in China. It was developed by China's National Research Center of Parallel Computer Engineering & Technology (NRCPC) and is installed at the National Supercomputing Center in Wuxi province. It is powered exclusively by Sunway's SW26010 processors, with an HPL mark of 93.0 petaflops, and kept its number three spot in the TOP500 list of supercomputers [33]. The SW26010 multicore processor is composed of four computing core groups (CGs), and the four CGs are interconnected at high speed through the on-chip network, as shown in Figure 11. Each CG consists of a management processing element (MPE) and a multicore array containing 64 computing processing elements (CPEs). MPE can directly access memory, while CPE can directly access main memory and supports DMA batch data transfer [34].

The SW26010 processor has slow memory access and large memory access instruction delay, a memory access problem that has seriously affected the performance of the CFD solver. The CCFD solver has been transplanted and optimized in Sunway TaihuLight, and the program memory access problem has been solved in three ways: using high-speed storage instead of main memory access; eliminating or reducing memory access operations; and hiding memory access time. The optimization work in this paper is divided into the following aspects. We use MPI to achieve parallel communication between the core groups and use Athread to achieve the master-slave core parallelism in the core group. The data of the slave core are stored in the core group, the subarea is partitioned, and DMA batch transfer eliminates main memory fetch operations. The matrix data continuity access is optimized. The fast Carmack algorithm is used. The vectorized calculation and manual assembly instruction are rearranged. The use of a double cache method to cover fetch time significantly improves the parallel efficiency.

4.1. Data Partitioning and DMA Parameter Tuning. CCFD is optimized based on the memory size of SW26010 and the performance of DMA to achieve domain decomposition and DMA parameter tuning. The block division of the many-core group calculation grid used by CCFD is shown in Figure 12. Using a $32 * 32 * 32$ block as an example, in order to efficiently use the storage space of CPE, 64 slave core blocks divide the grid data along the X-Y cross section. Then, the divided block data are allocated to 64 slave cores, and the 64 slave cores execute the same calculation process synchronously to complete a flow field calculation of the entire block.

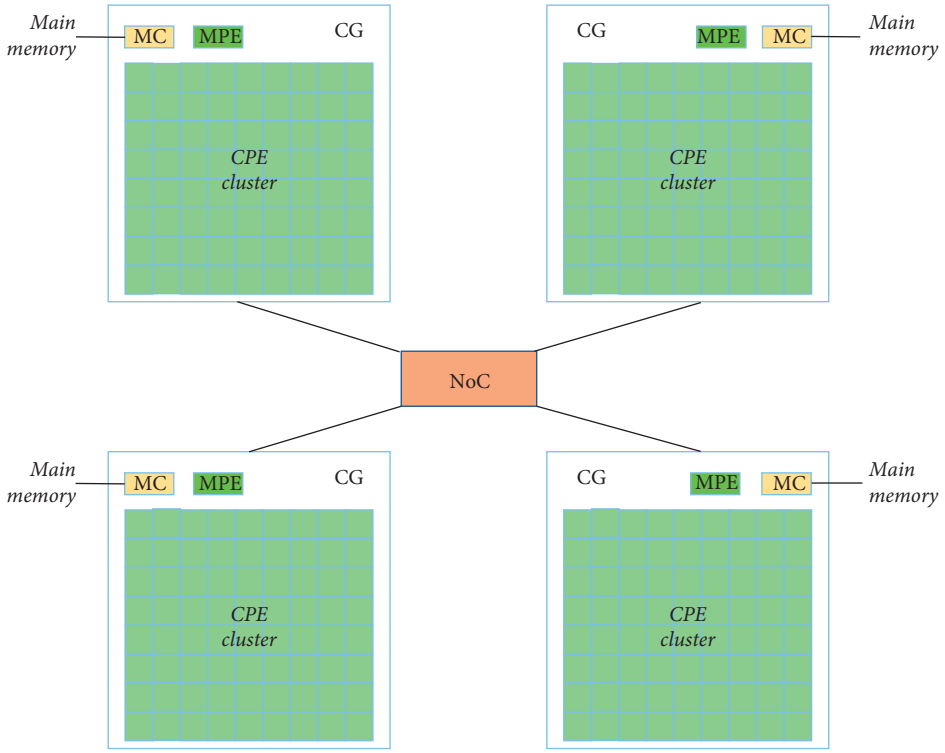


Figure 11: General architecture of the SW26010 processor.

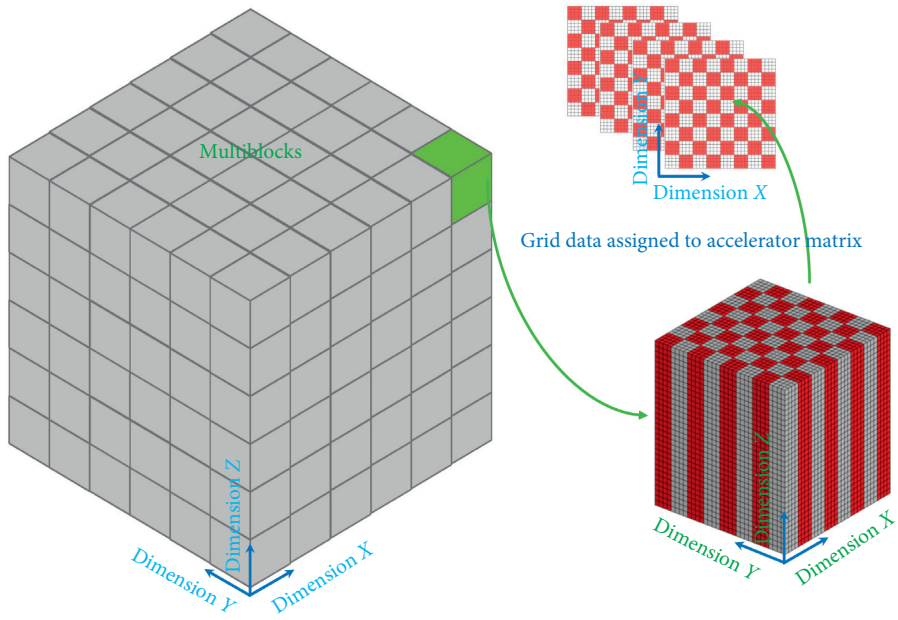


Figure 12: Block's domain decomposition for CPE slave group.

In actual programming, the size partition and direction selection of computational subdomains need to be further tuned, combined with the calculation process and the performance of the DMA. The local data memory (LDM) of the core processing unit is 64 kB, and in the actual flow field calculation, the size of the storage space required for different calculation processes ranges from 12 kB to more than 100 kB. Based on this, the temporarily stored LDM data need to strictly control the calculation scale, flexibly utilize DMA batch transmission, and eliminate the fetch operation to obtain data from the memory, thereby reducing the fetch time. In Figure 11, DMA is called by slave core. X is the lowest dimension of the data block, Y is the secondary dimension, and Z is the highest dimension. The data along the highest dimension are allocated to the same slave processor, and they are calculated along the Z-axis. The data

transmitted to local are calculated cyclically by the slave core processor to complete the iterative solution of the flow field of a single block.

4.2. Register Communication Optimization. After the block subarea data are evenly distributed to the slave core arrays, the slave cores that are physically located adjacent to each other on the on-chip bus have the adjacent global grid data. The SW26010 processor supports the same columns of slave core array registers performing point-to-point discrete communication. The aggregate bandwidth of the register communication theory is much larger than the theoretical aggregate bandwidth of the DMA. This paper uses register point-to-point communication to obtain the boundary data stored on the adjacent slave cores on the *X*-*Y* plane in the slave core array group. Moreover, the register communication and the asynchronous mechanism of arithmetic operation are used to fill the instructions that have nothing to do with the communication data between the RLC_GET/RLC_PUT instructions, to cover the communication time between the register communications.

As shown in Figure 13, using the *XY* plane as an example, the green dots indicate the grid data stored on the slave core LDM. The flow field calculation of the local cell is performed on each slave core, and the update result is returned. However, the local cells are calculated using the data of cells in four directions which contain part of halo cells in Figure 13 and assigned to adjacent slave calculation local cell to complete the calculation of the internal point. For the three-dimensional 13-point model discussed in this paper, after reading one layer of calculation subregion data blocks from the core, it is necessary to read two layers of halo data blocks from the surrounding four directions to complete the calculation.

The SW26010 supports register communication between peers and the same column. On the accelerator array, the registers of accelerators in the same row or the same column realize interconnected communication and can transmit 128 bits of data each time. As shown in Tables 1 and 2, the peer processors place the data in the register and specify the target number of the peer processor and then send the data to the destination. The target processor will collect the data over the Internet and save it to the register of the processor, realizing the communication operation of the register.

4.3. SIMD Calculation and Assembly Instruction Optimization. The SW26010 supports the 256-bit SIMD instruction set, which can perform four double-precision or eight single-precision data operations simultaneously. To ensure the use of SIMD, the calculation process and the address alignment of the variable array need to be alignment-adjusted. At the same time, SIMD_LOAD/SIMD_STORE and the calculation instruction are manually adjusted under the premise of ensuring that the calculation is correct, so that the LOAD/STORE and the calculation time are partially hidden. Data access and storage processing and arithmetic of the SW26010 are operated on two pipelines. To achieve the overlap of the data fetch and arithmetic operations, this paper analyzes the order of instructions generated by the compiler, through manually adjusting the order of instructions, and implements the overlap of data fetch load/store operations and data-independent operations. This method can eliminate data dependencies and improve code efficiency.

To further improve the calculation performance, the calculation code is disassembled. As shown in Table 3, by observing the order of the instruction execution sequence after disassembly, we found the memory access and calculation instructions are completely separated, and nearly six access instructions are issued on the pipeline before one calculation instruction is launched. After ensuring that the data are not relevant, we use inline assembly functions and handwritten assembly codes instead of function interface codes. As shown in Table 4, after manually adjusting the assembly code, the overlap of memory access and calculation instruction time is realized.

5. CCFD Application and Results

To verify the architecture and parallel performance of the CCFD software, this paper selects the DLR-F6 wing-body assembly, the CT-1 standard model, and the ultra-large-scale M6 wing model for testing. The test platforms are the Yuan supercomputer at the Computer Network Information Center in Beijing and the Sunway TaihuLight supercomputer at the National Supercomputing Center in Wuxi.

5.1. Optimization Results on Sunway TaihuLight. Using a block of size $32 * 32 * 32$ as the optimization case, the block data were divided into the $8 * 8$ slave cluster on CGs, the main calculation function and hot spots of CCFD were optimized, the flow field calculation was accelerated, and the entire CFD simulation process was completed. Figure 14 shows the SIMD optimization provided by Sunway's compiler in the Roe scheme calculation part, and the comparison of the average cycles number before and after inline assembly instruction optimization. The optimization effect of SIMD and assembly instructions is significant, in that implementing the overlap of fetching and manual instruction adjustment further optimizes the calculation.

Now that CCFD has completed the acceleration optimization of the main functional modules, including flux calculation, turbulent model, and time advancement method, and the slave core acceleration ratio has finally reached more than 30 to 1. Figure 15 shows the comparison of calculation time before and after the optimization of the main calculation function module of CCFD.

5.2. DLR-F6 Wing Body Assembly. The German Aerospace Center's DLR-F6 wing-body assembly is a typical example of a transonic transport aircraft, which verifies the CFD solver's ability to calculate complex shapes [42]. The calculation conditions are Mach number $\text{Mach} = 0.7510$, Reynolds number $\text{Re} = 0.0212465 \times 106$, angle of attack $\alpha = 1.003^{\circ}$, and side slip angle $\beta = 0^{\circ}$. The calculation method uses Roe scheme, LU_ADI time advancement method, three-layer mesh sequencing strategy, and the Spalart-Allmaras

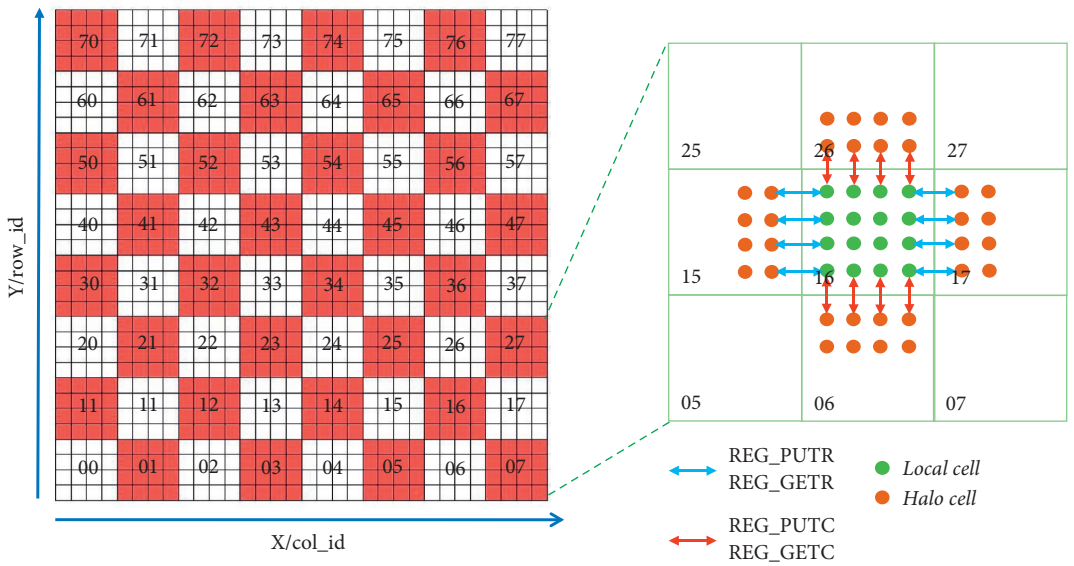


Figure 13: Data dependence and register communication from slave cores.

Table 1: Send data to the same row register.

```
asm volatile(\
"vldd $0, 0(%0)\n\t"
"vldd $1, 0(%1)\n\t"\
"vldd $2, 0(%2)\n\t" "putc $0, %4\n\t"\
"vldd $3, 0(%3)\n\t" "putc $1, %4\n\t"
"putc $2, %4\n\t"\
"putc $3, %4\n\t"\
\)
```

Table 2: Receive data from the same row register.

```
asm volatile(\
"getc $1\n\t"\
"getc $2\n\t"\
"getc $3\nt" "vstd $1, 0(%0)n\t"\
"getc $4\n\t" "vstd $2, 0(%1)\nt"\
"vstd $3, 0(%2)\n\t"\
"vstd $4, 0(%3)\n\t"
\)
```

Table 3: Order of instructions generated by the compiler.

(1). "vldd $0, 0(%2)\n\t"
(2). "vldd $1, 32(%2)\n\t"
(3). "vldd $2, 64(%2)\n\t"
(4). "vldd $3, 0(%3)\n\t"
(5). "vldd $4, 32(%3)\n\t"
(6). "vldd $5, 64(%3)\n\t"
(7). "vmuld $0, $3, $0\n\t"
(8). "vsubd $1, $4, $1\n\t"
(9). "vdivd $2, $5, $2\n\t"
(10). "vmad $0, $1, $2, $2\n\t"

turbulent model. Figure 16 shows the pressure nephogram of the entire assembly.

Figure 17 shows the comparison of the experimental results and calculated pressure coefficients of the different

Table 4: Out-of-order execution of instructions.

(1). "vldd $0, 0(%2)\n\t"
(2). "vldd $3, 0(%3)\n\t"
(3). "vmuld $0, $3,$0\n\t"
(4). "vldd $1, 32(%2)\n\t"
(5). "vldd $4, 32(%3)\n\t"
(6). "vsubd $1, $4,$1\n\t"
(7). "vldd $2, 64(%2)\n\t"
(8). "vldd $5, 64(%3)\n\t"
(9). "vdivd $2, $5, $2\n\t"
(10). "vmad $0, $1, $2, $2\n\t"

positions of the DLR-F6 wing. The CCFD calculation results are consistent with the experimental results, which verifies that CCFD can simulate the flow field calculation of a complex-shaped aircraft.

5.3. CT-1 Standard Model. The CT-1 is a model of high-angle static aerodynamic characteristics released by the China Aerodynamics Research and Development Center in 2005 [43]. We focus on the numerical simulation capabilities of the CFD solution software at high angles of static aerodynamic characteristics. Figure 18 shows the pressure nephogram at different angles of attack.

Figure 19 shows a comparison of the experimental results and the drag coefficients for each angle of attack. By comparing the experimental results with the calculation results of the static aerodynamic characteristics of the CT-1 standard model at high angles of attack, the CCFD calculation results are consistent with the experimental results, which verifies the numerical simulation ability of CCFD to solve the static aerodynamic characteristics at high angle of attack.

5.4. Ultra-Large-Scale Parallel Testing of M6 Wing. The Onera M6 wing standard calculation model [37] uses 650 million grids and 28,160 blocks. It verifies the parallel solver capabilities above

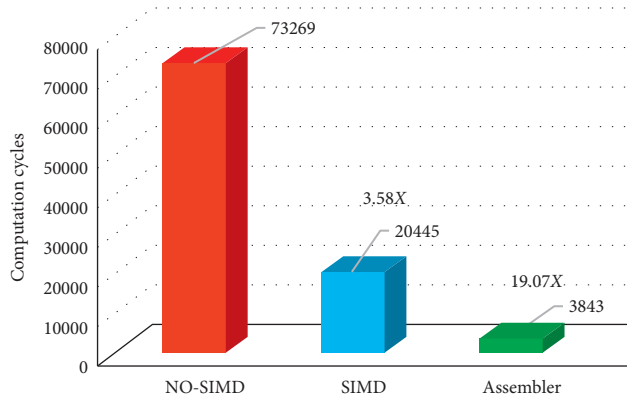


Figure 14: Comparison of cycle before and after SIMD and assembler optimization.

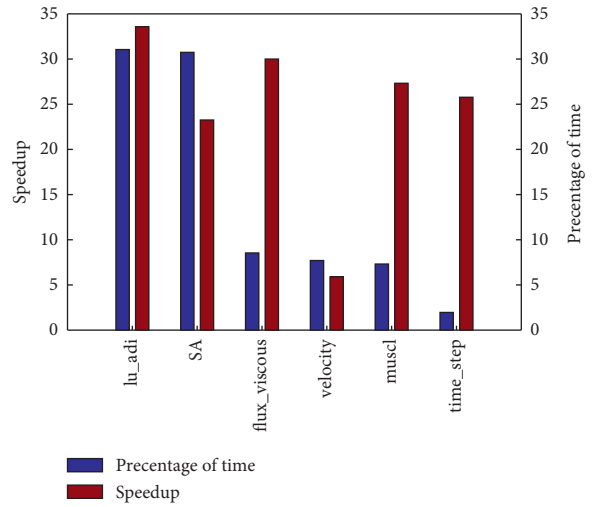


Figure 15: Master-slave core acceleration times and time ratio of the main calculation module after optimization.

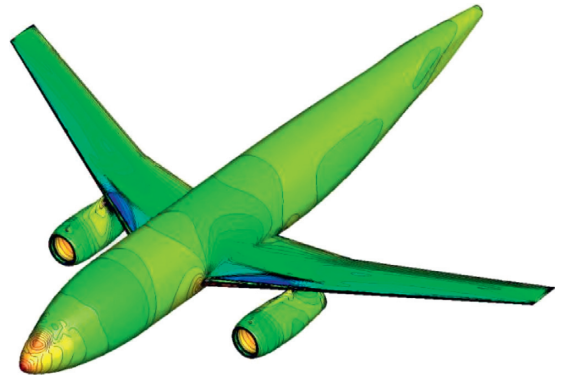


Figure 16: Pressure nephogram for DLR-F6.

the 10,000-core level and the parallel expansion performance of CCFD. The calculation conditions are Mach number Mach = 0.840, Reynolds number Re = 21.70 × 106, angle of attack $\alpha = 3.06^{\circ}$, and side slip angle $\beta = 0^{\circ}$. The calculation method uses Roe scheme, LU_ADI time advancement method, and the Spalart-Allmaras turbulent model.

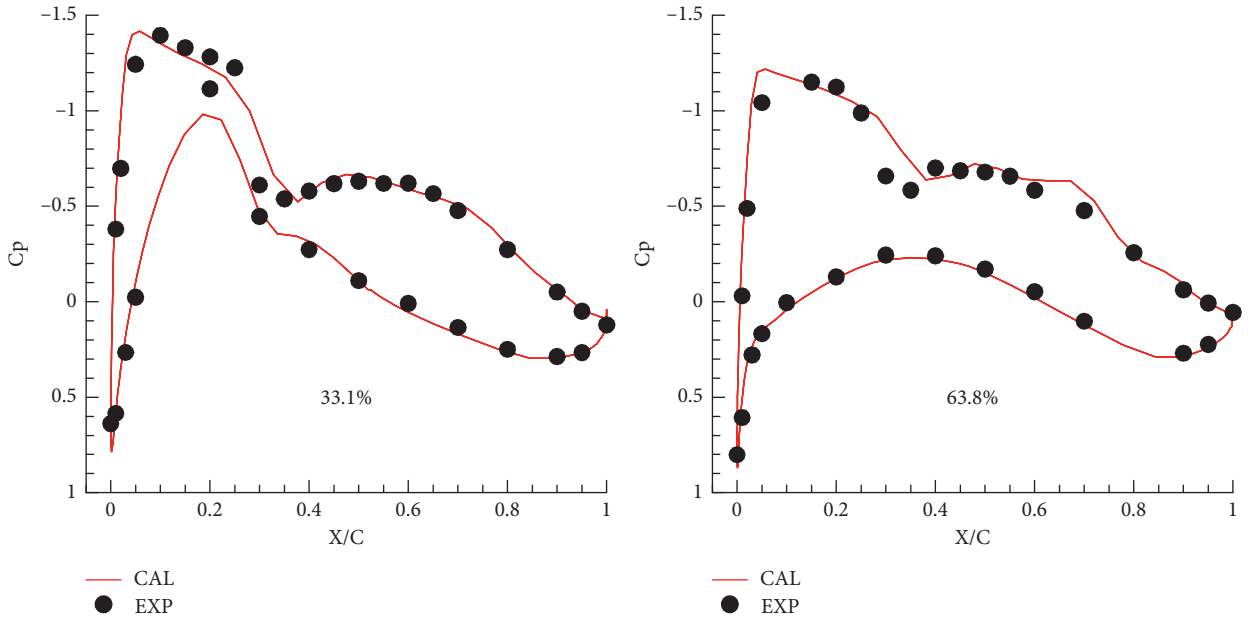


Figure 17: Comparison of the pressure curves of each wing position and the nacelle section with the experimental data.

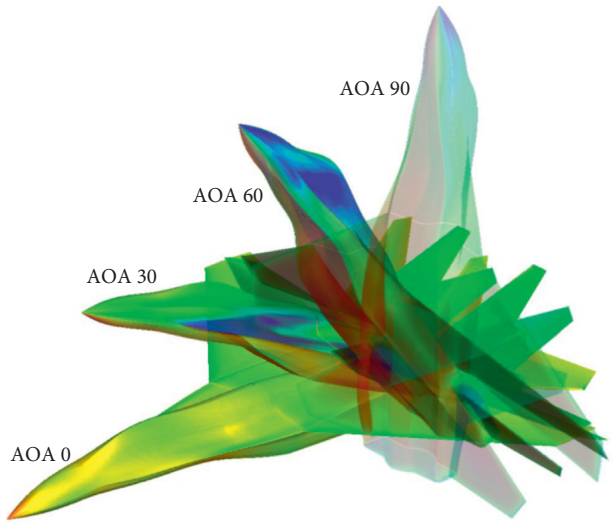


Figure 18: Pressure nephogram at different angles of attack for CT-1.

CCFD is based on the Sunway TaihuLight heterogeneous platform, with its core groups communicating with each other using MPI, and 64 slave cores in the core group accelerated in parallel using Athread. Figure 20 shows that the parallel efficiency of CCFD reached 60% under the Sunway TaihuLight heterogeneous platform with 13,000 cores and 500,500 cores. The parallel result test indicates that the super-large-scale parallel computing scalability of CCFD has met the design requirements.

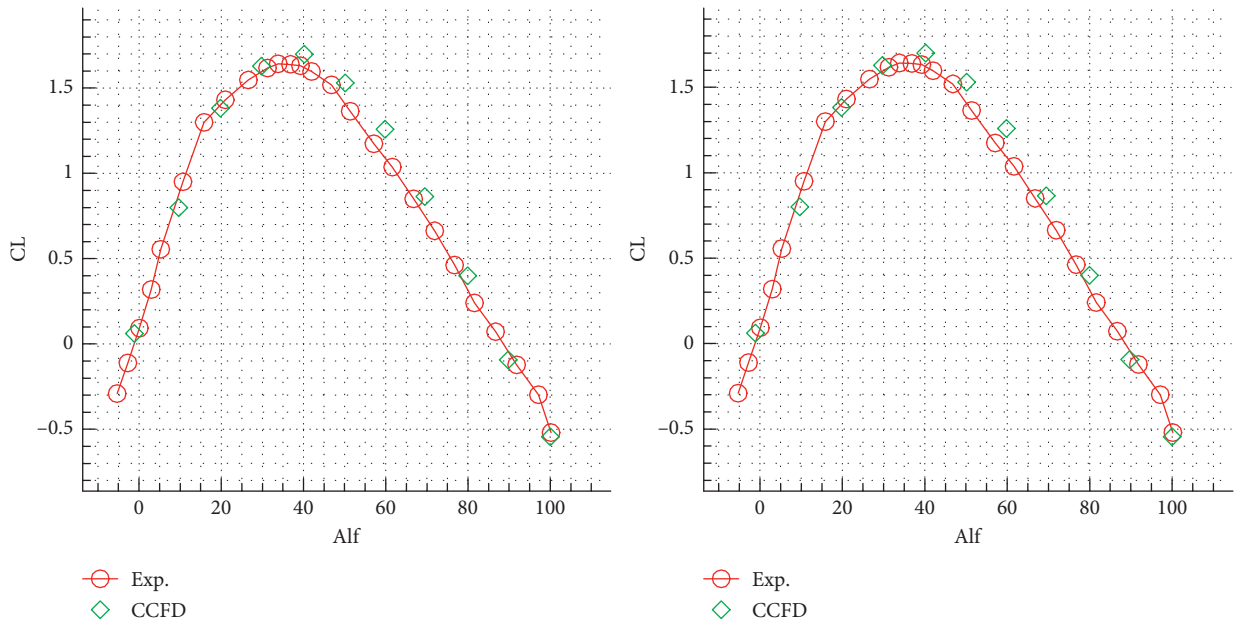


Figure 19: Lift and drag coefficients at different angles of attack for CCFD.

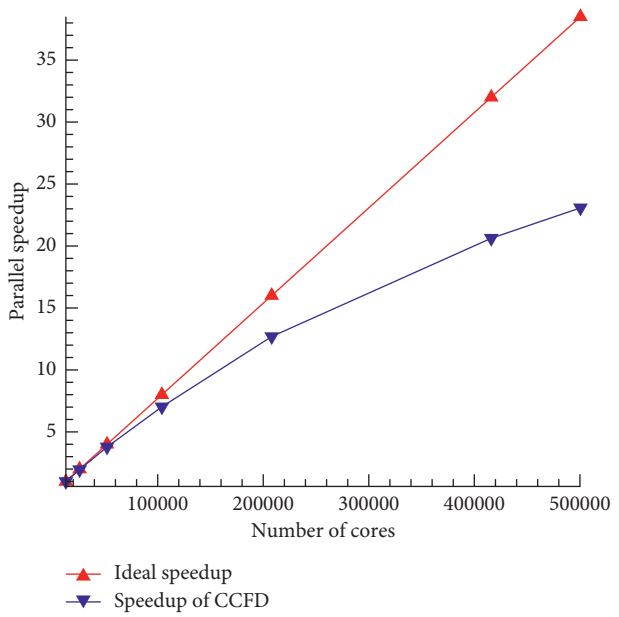


Figure 20: Parallel efficiency of large-scale test for CCFD.

6. Conclusions and Future Work

Using the multiblock structure grid parallel technology of domain decomposition, we design the parallel software framework, software process, parallel data structure, and communication mode of parallel CFD solver software. The designed CCFD can be used for large-scale mega-core parallel computing tasks and has good parallel expansion ability and parallel efficiency. The overdecomposition load balancing strategy for CCFD guarantees the load balancing performance of computing and communication. The use of a mesh sequencing method based on multigrid technology can accelerate the iterative convergence speed. We perform the optimizations on DMA, SIMD, assembly instruction rearrangement, and double buffering in the Sunway TaihuLight heterogeneous architecture. The super-large computational scale test with a maximum of 505,000 cores in parallel across the core group achieved a parallel efficiency of 60% based on 13,000 cores. In the future, we will use direct numerical simulation of turbulence models to achieve more realistic geometric flow simulations, increase the scale of cases and parallelism, and achieve more efficient operation of parallel CFD software in heterogeneous systems.

Data Availability

The grid data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by a grant from the National Key R&D Program of China (No. 2019YFB1704204), the National Natural Science Foundation of China (Nos. 61702438 and 11502267), the Chinese Academy of Sciences strategic pilot technology special (Class B, XDB22020102, and Class C, XDC01040100), and the Informatization Special Project of Chinese Academy of Sciences (No. XXH13506-204).

References

[1] J. Blazek, *Computational Fluid Dynamics: Principles and Applications*, Elsevier Science Publication, Amsterdam, Netherlands, 2005.

[2] H. K. Versteeg and W. Malalasekera, *An Introduction to Computational Fluid Dynamics*, Pearson Education Limited, London, UK, 2nd edition, 2007.

[3] P. Wesseling, "Principles of computational fluid dynamics," in *Computational Mathematics*, Vol. 29, Springer-Verlag, Berlin, Germany, 2001.

[4] C.-L. Lin, M. H. Tawhai, G. Mclennan, and E. A. Hoffman, "Computational fluid dynamics," *IEEE Engineering in Medicine and Biology Magazine*, vol. 28, no. 3, pp. 25–33, 2009.

[5] F. He, X. Dong, N. Zou, W. Wu, and X. Zhang, "Structured mesh-oriented framework design and optimization for a coarse-grained parallel CFD solver based on hybrid MPI/OpenMP programming," *The Journal of Supercomputing*, vol. 76, no. 4, pp. 2815–2841, 2020.

[6] A. Gel, E. J. Hu, E. Ould-Ahmed-Vall, and A. A. Kalinkin, "Modernization and optimization of a legacy open-source CFD code for high-performance computing architectures," *International Journal of Computational Fluid Dynamics*, vol. 31, no. 2, pp. 122–133, 2017.

[7] S. Li, R. Paoli, and M. D'Mello, "Scalability of OpenFOAM density-based solver with Runge–Kutta temporal discretization scheme," *Scientific Programming*, vol. 2020, Article ID 9083620, 11 pages, 2020.

[8] F. Palacios, T. D. Economon, and J. J. Alonso, "Large-scale aircraft design using SU^2," in *Proceedings of the AIAA Aerospace Sciences Meeting*, Kissimmee, FL, USA, January 2015.

[9] F. Palacios, A. Alonso, K. Duraisamy et al., "An open-source integrated computational environment for multi-physics simulation and design," in *Proceedings of the AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, Grapevine, TX, USA, January 2013.

[10] S. Yakubov, B. Cankurt, M. Abdel-Maksoud, and T. Rung, "Hybrid MPI/OpenMP parallelization of an Euler-Lagrange approach to cavitation modelling," *Computers & Fluids*, vol. 80, pp. 365–371, 2013.

[11] Y.-X. Wang, L.-L. Zhang, W. Liu, X.-H. Cheng, Y. Zhuang, and A. T. Chronopoulos, "Performance optimizations for scalable CFD applications on hybrid CPU + MIC heterogeneous computing system with millions of cores," *Computers & Fluids*, vol. 173, pp. 226–236, 2018.

[12] D. Cheng, C. Xu, B. Cheng, M. Xiong, X. Gao, and X. Deng, "Performance modeling and optimization of parallel LU-SGS on many-core processors for 3D high-order CFD simulations," *The Journal of Supercomputing*, vol. 73, no. 6, pp. 2506–2524, 2017.

[13] O. Bessonov, "OpenMP parallelization of a CFD code for multicore computers: analysis and comparison," in *Proceedings of the 11th International Conference on Parallel Computing Technologies, PaCT 2011*, Kazan, Russia, September 2011.

[14] A. Amritkar, S. Deb, and D. Tafti, "Efficient parallel CFD-DEM simulations using OpenMP," *Journal of Computational Physics*, vol. 256, pp. 501–519, 2014.

[15] V. Moureau, P. Domingo, and L. Vervisch, "Design of a massively parallel CFD code for complex geometries," *Comptes Rendus Mécanique*, vol. 339, no. 2-3, pp. 141–148, 2011.

[16] B. Kong, R. O. Fox, H. Feng et al., "Euler-euler anisotropic Gaussian mesoscale simulation of homogeneous cluster-induced gas-particle turbulence," *AICHE Journal*, vol. 63, no. 7, pp. 2630–2643, 2017.

[17] A. Capecelatro, A. Cauble-Chantrenne, A. Jundt et al., "Running large-scale CFD applications on Intel-KNL-based clusters," *International Journal for Numerical Methods in Fluids*, vol. 86, no. 11, pp. 699–716, 2018.

[18] L. Lapichino, A. Quarteroni, and G. Rozza, "Reduced basis method and domain decomposition for elliptic problems in networks and complex parametrized geometries," *Computers & Mathematics with Applications*, vol. 71, no. 1, pp. 408–430, 2016.

[19] S. Badia and H. Nguyen, "Balancing domain decomposition by constraints and perturbation," *SIAM Journal on Numerical Analysis*, vol. 54, no. 6, pp. 3436–3464, 2016.

[20] S. Badia, A. F. Martín, and J. Principe, "Multilevel balancing domain decomposition at extreme scales," *Siam Journal on Scientific Computing*, vol. 38, no. 1, pp. C22–C52, 2016.

[21] O. S. Lawlor, S. Chakravorty, T. L. Wilmarth, and N. Choudhury, "ParFUM: a parallel framework for unstructured meshes for scalable dynamic physics applications," *Engineering with Computers*, vol. 22, no. 3-4, pp. 215–235, 2006.

[22] G. T. Abraham, A. James, and N. Yaacob, "Priority-grouping method for parallel multi-scheduling in grid," *Journal of Computer and System Sciences*, vol. 81, no. 6, pp. 943–957, 2015.

[23] T. Shimokawabe, T. Aoki, and N. Onodera, "High-productivity framework for large-scale GPU/CPU stencil applications," *Procedia Computer Science*, vol. 80, pp. 1646–1657, 2016.

[24] I. Z. Reguly and G. R. Mudalige, "Productivity, performance, and portability for computational fluid dynamics applications," *Computers and Fluids*, vol. 199, Article ID 104425, 2020.

[25] J. Wang, X. Lv, and X. Chen, "Comparative analysis of list scheduling algorithms on homogeneous multi-processors," in *Proceedings of the IEEE International Conference on Communication Software & Networks*, IEEE, Beijing, China, 2016.

[26] L. Chai, A. Hartono, and D. K. Panda, "Designing high performance and scalable MPI intra-node communication support for clusters," in *Proceedings of the IEEE International Conference on Cluster Computing*, IEEE, Tampa, FL, USA, November 2006.

[27] Q. Tang, L.-H. Zhu, L. Zhou, J. Xiong, and J.-B. Wei, "Scheduling directed acyclic graphs with optimal duplication strategy on homogeneous multiprocessor systems," *Journal of Parallel and Distributed Computing*, vol. 138, pp. 115–127, 2020.

[28] A. Kayi, T. El-Ghazawi, and G. B. Newby, "Performance issues in emerging homogeneous multi-core architectures," *Simulation Modelling Practice and Theory*, vol. 17, no. 9, pp. 1485–1499, 2009.

[29] S. G. Ahmad, C. S. Liew, E. U. Munir, T. F. Ang, and S. U. Khan, "A hybrid genetic algorithm for optimization of scheduling workflow applications in heterogeneous computing systems," *Journal of Parallel and Distributed Computing*, vol. 87, pp. 80–90, 2016.

[30] H. A. Tokel, G. Alirezaei, S. Baig, and R. Mathar, "An optimization framework for planning of WAMS with a heterogeneous communication network," in *Proceedings of the IEEE International Conference on Smart Grid Communications*, IEEE, Sydney, Australia, November 2016.

[31] L. Deng, F. H. Bai, F. H. Bai, and Q. Xu, "CPU/GPU computing for an implicit multi-block compressible Navier-Stokes solver on heterogeneous platform," *International Journal of Modern Physics: Conference Series*, vol. 42, Article ID 1660163, 2016.

[32] W. Cao, C.-F. Xu, Z.-H. Wang, L. Yao, and H.-Y. Liu, "CPU/GPU computing for a multi-block structured grid based high-order flow solver on a large heterogeneous system," *Cluster Computing*, vol. 17, no. 2, pp. 255–270, 2013.

[33] TOP500 Supercomputing Sites (EB/OL), 2019, https://www.top500.org.

[34] Z. Xu, J. Lin, and S. Matsuoka, "Benchmarking SW26010 many-core processor," in *Proceedings of the 2017 IEEE international parallel and distributed processing symposium workshops (IPDPSW)*, IEEE, Orlando, FL, USA, pp. 743–752, May 2017.

[35] H. Fu, J. Liao, J. Yang et al., "The Sunway TaihuLight supercomputer: system and applications," *Science China Information Sciences*, vol. 59, no. 7, pp. 113–128, 2016.

[36] C. Yang, W. Xue, H. Fu et al., "10M-core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE Press, Salt Lake City, UT, USA, 2016.

[37] Z. Johan, K. K. Mathur, S. L. Johnsson, and T. J. R. Hughes, "A case study in parallel computation: viscous flow around an ONERA M6 wing," *International Journal for Numerical Methods in Fluids*, vol. 21, no. 10, pp. 877–884, 1995.

[38] W. S. Brainerd, "Object-oriented programming," in *Guide to Fortran 2008 Programming*, Springer, Berlin, Germany, 2015.

[39] W. Deng, J. Xu, Y. Song, and H. Zhao, "An effective improved co-evolution ant colony optimization algorithm with multi-strategies and its application," *International Journal of Bio-Inspired Computation*, vol. 20, no. 5, pp. 1–10, 2020.

[40] D. LaSalle, M. M. A. Patwary, N. Satish, N. Sundaram, P. Dubey, and G. Karypis, "Improving graph partitioning for modern graphs and architectures," in *Proceedings of the 5th Workshop on Irregular Applications: Architectures and Algorithms*, ACM, Austin TX, USA, November 2015.

[41] Z. Shang, "Large-scale CFD parallel computing dealing with massive mesh," *Journal of Engineering*, vol. 2013, Article ID 850148, 6 pages, 2013.

[42] J. C. Vassberg, A. J. Sclafani, and M. A. DeHaan, "A wing-body fairing design for the DLR-F6 model: a DPW-III case study," Report No. AIAA-2005-4730, AIAA, Reston, VA, USA, 2005.

[43] Y. Wang, G. Wang, and Z. Chen, "Numerical simulation of static aerodynamic characteris of CT-1 model at high angles of attack," *Acta Aeronautica Et Astronautica Sinica*, vol. 29, no. 4, pp. 859–865, 2008.

Hindawi
Scientific Programming
Volume 2020, Article ID 8866557, 8 pages
https://doi.org/10.1155/2020/8866557

Research Article

ACT-SVM: Prediction of Protein-Protein Interactions Based on Support Vector Basis Model

Wenzheng Ma,[1] Yi Cao,[1] Wenzheng Bao,[2] Bin Yang,[3] and Yuehui Chen[1]

[1]*School of Information Science, University of Jinan, Jinan 250022, China*
[2]*School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou 221018, China*
[3]*School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China*

Correspondence should be addressed to Wenzheng Bao; baowz55555@126.com

Received 7 April 2020; Revised 13 May 2020; Accepted 22 May 2020; Published 20 July 2020

Academic Editor: Chenxi Huang

The interactions between proteins play important roles in several organisms, and such issue can be involved in almost all activities in the cell. The research of protein-protein interactions (PPIs) can make a huge contribution to the prevention and treatment of diseases. Currently, many prediction methods based on machine learning have been proposed to predict PPIs. In this article, we propose a novel method ACT-SVM that can effectively predict PPIs. The ACT-SVM model maps protein sequences to digital features, performs feature extraction twice on the protein sequence to obtain vector A and descriptor CT, and combines them into a vector. Then, the feature vectors of the protein pair are merged as the input of the support vector machine (SVM) classifier. We utilize nonredundant *H. pylori* and human dataset to verify the prediction performance of our method. Finally, the proposed method has a prediction accuracy of 0.727897 for *H. pylori* data and a prediction accuracy of 0.838799 for human dataset. The results demonstrate that this method can be called a stable and reliable prediction model of PPIs.

1. Introduction

Proteins are the material basis of all life composed of 20 types of amino acids in the level of biology [1]. There are several kinds of proteins with different properties and functions, which play a pivotal role in the cells and tissues of various biological species. Not only is it an important part of the living organism, but also it participates in and carries all important life activities in the life process. However, most proteins often do not perform their functions alone. Instead, it is more common that two or more proteins work together by forming a protein complex, and a large protein-protein interaction network is finally built [2–6]. Obviously, PPIs play a key role in cellular processes and are involved in many important biological processes such as immune response, material transport, and gene expression regulation. Therefore, exploring the interactions between proteins has become one of the most important links in researching the function and mechanism of proteins [7–9]. In addition, PPIs are a major molecular mechanism of virus pathogenic, which makes them one of the important research objects for disease discovery and treatment.

The importance of researching PPIs has advanced the methods for predicting and identifying PPIs [10–13]. In recent years, some high-throughput laboratory biotechnology has been widely utilized in PPIs, such as yeast two-hybrid (Sato et al.; Schwikowski et al.; Coates Hall) [14–16] and coimmunoprecipitation (Free et al.) [17]. However, they all have some defects in common or personality. For example, some methods fail to overcome higher proportion of false negatives and false positives, and some methods require more sample material to extract proteins, which is surprisingly expensive. At the same time, methods such as protein phylogenetic profile (Kim et al.) [18, 19], natural language processing (Daraselia et al.) [20], and protein tertiary structure (Aloy and Russell) [21] have also been favored by researchers. However, if there is no known protein-related biological knowledge, these kinds of methods are difficult to implement, and some of them cannot fully predict PPIs [22, 23].

In addition, with the tireless efforts of researchers, it was found that PPIs can be predicted based on the amino acid sequence of the protein [24–27]. At the same time, machine learning has been utilized by researchers far and wide. Then, a large number of prediction methods based on protein sequences and machine learning algorithms have appeared [13, 28–32]. For example, Cui et al. [33] utilized support vector machine classifier to predict human proteins that interact with viral proteins [34–37]. The L1-logreg classifier proposed by Dhole et al. can effectively predict PPIs and advance related research such as drug design. Xia et al. [38] proposed a sequence-based multiclassifier system called Spinning Forest to infer PPIs [39]. The performance of their method on the *Saccharomyces cerevisiae* and *H. pylori* datasets is better than previously published literature methods. And as an effective machine learning method, deep learning is also utilized in the prediction of PPIs (Du et al.) [40].

In this paper, we propose a novel prediction model which is based on support vector machine to predict PPIs named ACT-SVM. Two different methods were utilized to extract features from protein sequences, and finally we reconstruct them into a feature vector. First, we extract an A vector for each protein sequence in the dataset. Hereafter, we construct composition (C) and transformation (T) descriptors to describe protein sequences. Last, we utilize their combination as the input of the classifier. In general, the area under curve (AUC), accuracy (Acc), specificity (Sp), and Matthew correlation coefficient (Mcc) are utilized to evaluate the performance of our prediction method.

We have additionally constructed 5 different classifiers for comparing the predictive performance, including *k*-nearest neighbor (KNN), artificial neural network (ANN), random forest (RF), naive Bayes (NB), and logistic regression (LR). We utilized H. pylori and human datasets to evaluate our novel predictor. Experimental results demonstrate that the novel model based on support vector machine which is proposed by us is performs best.

2. Methods and Materials

In scientific research, it is extremely important to first define the workflow. Our working flow is demonstrated in Figure 1. First, we obtained nonredundant *H. pylori* and human datasets. Then, we map each protein sequence to digital features by constructing A vector, composition, and transformation (CT) and combine them into one feature vector as the input of the classifier. The following process is to input the extracted digital feature into different classifiers to train different classification models and evaluate them by 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation, respectively. Finally, on the independent test datasets, we sequentially verified the 6 trained models. In addition, we utilize AUC, Acc, Sp, Sn, and MCC indicators to evaluate the performance of our novel predictive silver and five models utilized as a comparison.

2.1. Dataset. As people pay an increasing attention to PPIs, the number of databases utilized to research PPIs is increasing, such as BioGRID, GeneMANIA, and DIP. However, there is inevitable redundancy in the data in these existing databases. To make our prediction tool more effective, we derived nonredundant *H. pylori* and human PPIs dataset utilized by Kong et al. [41]. They downloaded the *H. pylori* and human PPIs dataset from the DIP database and utilized the cd-hit tool to construct nonredundant sequences for these two datasets. After removing redundancy, the *H. pylori* dataset contains 1458 interacting protein pairs and 1457 noninteracting protein pairs, while the human dataset has 3899 interacting protein pairs and 4262 noninteracting protein pairs.

2.2. Sequence Feature Vectors

2.2.1. Construct a Vector. When constructing the A vector, we refer to the physical and chemical properties of the protein. The 20 amino acids that make up the protein sequence are divided into 6 classes, as demonstrated in Table 1.

In this way, according to the category, we replace each amino acid in the sequence with the corresponding C_1, C_2, . . ., C_6. Then, we can obtain a simplified sequence. We utilize f_i to describe the frequency of occurrence of each element in the simplified sequence ($i = 1, 2, ..., 6$) and finally get the A vector. The detailed definitions of f_i and A vector are illustrated by equations (1) and (2).

$$f_i = \frac{m_i}{l}, \tag{1}$$

where l is the length of the protein sequence, m_i is the number of type i amino acids in the protein sequence, $i = 1, 2, ..., 6$. For example, if there is a sequence "MGPDDSKRYE," it can be replaced with C_1, C_6, C_6, C_5, C_5, C_3, C_4, C_4, C_2, and C_5. We can see that there are one C_1, one C_2, one C_3, two C_4, three C_5 and two C_6 in the simplified sequence. Thus, $f_1 = 1 * 100\%/10 = 10\%$, $f_2 = 1 * 100\%/10 = 10\%$, $f_3 = 1 * 100\%/10 = 10\%$, $f_4 = 2 * 100\%/10 = 20\%$, $f_5 = 3 * 100\%/10 = 30\%$, and $f_6 = 2 * 100\%/10 = 20\%$.

A vector can be constructed as

$$A = (f_1, f_2, \ldots, f_i, \ldots, f_6). \tag{2}$$

Then, we got a 6-dimensional A vector to describe the feature of the protein.

2.2.2. Sparse Matrix and Descriptor. First, we construct a $20 \times n$ sparse matrix B, where n is the number of amino acids in the protein sequence. We assume that there is a protein sequence $S = S_1, S_2, \ldots, S_n$. At the same time, we put 20 amino acids in E, $E = \{A, V, L, I, M, C, F, W, Y, H, S, T, N, Q, K, R, D, E, G, P\}$. When the i-th amino acid in E is the same as the j-th amino acid in S, the corresponding element b_{ij} in the sparse matrix takes 1; otherwise, it takes 0. The sparse matrix of this protein sequence is demonstrated in the following:

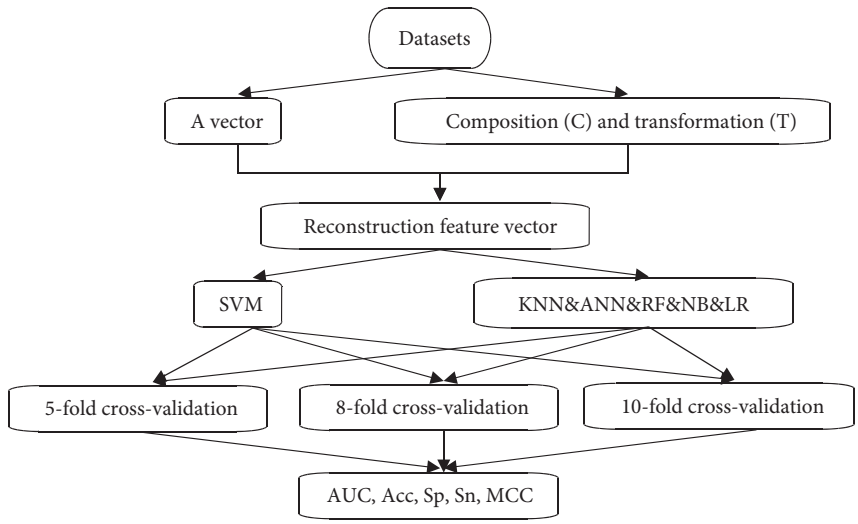


Figure 1: The working flow.

Table 1: Classification of proteins.

Category	Property	Amino acid
C_1	Aliphatic	A, C, I, L, M, V
C_2	Aromatic	F, H, W, Y
C_3	Polar	N, Q, S, T
C_4	Positive	K, R
C_5	Negative	D, E
C_6	Special conformations	G, P

$$B_{20\times n} = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2n} \\ b_{31} & b_{32} & b_{33} & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{20,1} & b_{20,2} & b_{20,3} & \cdots & b_{20,n} \end{pmatrix}, \quad (3)$$

$$b_{ij} = \begin{cases} 0, & E(i) \neq S(j) \\ 1, & E(i) = S(j) \end{cases}$$

Next, we divide each of the 20 row vectors in the sparse matrix into P subvectors. The descriptor consists of composition (C) and transformation (T), and they are extracted from each subvector. Among them, the composition (C) is composed of two parts, including the frequency of 0 and 1 in the subsequence. The transition (T) consists of three parts: the sum of the number of 01 and 10 in the subvector, the number of "11" and the number of "111." Suppose $P = 4$, and the first subsequence of a protein sequence is "MYAHQAAA." Then, the first subvector of the first row vector in the sparse matrix is {0, 0, 1, 0, 0, 1, 1, 1}. Obviously, there are four "0," four "1," two "01," one "10," two "11," and one "111." Therefore, the five parts of the composition and transformation (CT) are $4 * 100\%/8 = 50\%$, $4 * 100\%/8 = 50\%$, 3 $(2 + 1 = 3)$, 2, and 1. Thus, a protein sequence is mapped into a 400-dimensional $(4 * 20 * 5 = 400)$ vector.

2.2.3. Reconstruction of Feature Vectors. For each protein sequence, we extracted two feature vectors, including a 6-dimensional vector A and a 400-dimensional descriptor. Then, we combined them into a 406-dimensional vector as the feature vector of a protein. Finally, the feature vectors of two proteins are connected as a 812-dimensional feature vector, describing the PPIs between them.

2.3. Classifier Construction. Our model is based on SVM. As a linear classifier, SVM is widely utilized in classification problems. Its learning strategy is to maximize the interval. Finally, it can find a geometric hyperplane with the largest distance in the feature space to segment the sample. SVM is extremely stable and sparse. The partitioning hyperplane in the sample space can be described as

$$\boldsymbol{\omega}^T \mathbf{x} + b = 0. \quad (4)$$

Among them, the direction of the hyperplane is determined by $\boldsymbol{\omega}$, and b represents the distance from the origin to the hyperplane. If the hyperplane can correctly classify the samples, one side of the hyperplane is positive samples and the other side is negative samples. Assume that the samples in the sample space are $(\mathbf{x}_i, y_i)$, $y_i \in \{+1, -1\}$, which can be expressed as

$$\begin{cases} \boldsymbol{\omega}^T + b \geq +1, & y_i = +1 \\ \boldsymbol{\omega}^T + b \leq -1, & y_i = -1. \end{cases} \quad (5)$$

The distance from any point in the sample space to the hyperplane can be described by equation (6):

$$d = \frac{\left|\boldsymbol{\omega}^T \mathbf{x} + b\right|}{\|\boldsymbol{\omega}\|}. \quad (6)$$

The closest sample point to the hyperplane is called the support vector. The sum of the distance from the positive sample support vector to the hyperplane and the distance from the negative sample support vector to the hyperplane is called the interval, which can be expressed as

$$\gamma = \frac{2}{\|\boldsymbol{\omega}\|}. \tag{7}$$

The ultimate goal of support vector machine is to find a hyperplane that maximizes the interval, so the support vector machine can be described as

$$\max_{\omega,b} \frac{2}{\|\omega\|}, \tag{8}$$

$$s.t.\ y_i\left(\boldsymbol{\omega}^T \mathbf{x}_i + b\right) \geq 1, \quad i = 1, 2, \ldots, n, \tag{9}$$

where m is the number of samples. Formulas (8) and (9) can also be rewritten as

$$\min_{\omega,b} \frac{1}{2}\|\omega^2\|, \quad s.t.\ y_i\left(\boldsymbol{\omega}^T \mathbf{x}_i + b\right) \geq 1, \quad i = 1, 2, \ldots, n. \tag{10}$$

Through continuous experimentation, we finally set the kernel function of the SVM classifier to a linear kernel function. And combined with our proposed feature extraction method ACT, it showed superior prediction performance on *H. pylori* and human dataset.

2.4. Evaluation of the Predictor. In order to verify the reliability and stability of our proposed predictor, we trained 6 models using *H. pylori* and human dataset and performed 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation [42]. In actual training, the model usually fits the training data better, but it is not particularly ideal for novel data outside the training data. k-fold cross-validation can be utilized to evaluate the generalization ability of models, so as to choose a better model and prevent the model from being too complex and causing overfitting. The basic idea of k-fold cross-validation is to divide the dataset into k parts in equal proportions. Then each part of the data is utilized in turn as the test dataset, and the other $k-1$ parts of the data are utilized as training data. k-fold cross-validation is performed for k trainings to ensure that the k parts of the data have been the test data; the remaining $k-1$ parts have been utilized as training data. The obtained K experimental results are equally divided as the final score of the model ultimately. For k-fold cross-validation, we set k to 5, 8, and 10, respectively, to verify the performance of our model.

In this paper, we employ four evaluation indicators to evaluate the predictive performance of our proposed method, including accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew correlation coefficient (Mcc). Among them, Acc reflects the model's ability to classify positive samples correctly; Sn measures the classifier's ability to recognize positive samples; Sp reflects the model's ability to recognize negative samples; Mcc returns a value between −1 and +1, which is an indicator often utilized to measure the performance of binary classification models. Their definitions are as follows:

$$\begin{aligned}
\mathrm{Acc} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \\
\mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}}, \\
\mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}, \\
\mathrm{Mcc} &= \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}},
\end{aligned} \tag{11}$$

where TP is the number correctly divided into positive samples, FP is the number incorrectly divided into positive samples, FN is the number incorrectly divided into negative samples, and TN is correctly divided into negative samples. In addition, we still utilize the AUC value to evaluate the performance of our proposed model. AUC is defined as the area under the ROC curve. In many cases, the ROC curve does not clearly indicate which classifier works better. As a numerical value, the larger the corresponding AUC value, the better the classifier. Thus, we utilize the AUC value as one of the evaluation criteria of the model.

3. Result and Discussion

3.1. Model Stability Analysis. K-fold cross-validation is widely utilized to compare the performance of different machine learning models on a specific dataset. The principle of k-fold cross-validation is to divide the dataset into equal k shares for k trainings and finally take the average of the K results. However, there may be outliers in the k-time results, which means that this classifier may not have good stability for the prediction of all samples. We utilized *H. pylori* and human dataset to train 6 models and performed 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation to evaluate their performance. We draw boxplots to reflect the stability of 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation of the two datasets in 6 classifiers. Six boxplots were drawn to describe the results of 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation of two datasets in 6 classifiers. Among them, the ordinate of the boxplot is accuracy (Acc), and the abscissa is 6 classifiers. That is to say, each boxplot has 6 boxes, and each box stores the Acc value in the k times of k-fold cross-validation in the classifier. The boxplots of the *H. pylori* dataset on 6 classifiers for 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation are demonstrated in Figure 2(a), and the boxplots for the human dataset are demonstrated in Figure 2(b).

The hollow dots appearing in the boxplots are outliers, the size of the boxes reflects the degree of dispersion of the data, and the height of the boxes represents the accuracy value. From the 5-fold cross-validation box diagram in Figure 2(a), we can see that there are outliers in the 5 Acc values obtained by KNN, NB, and SVM in 5 trainings, while the box of the RF classifier is too large that the data is more discrete. The box size of the ANN and LR classifiers is

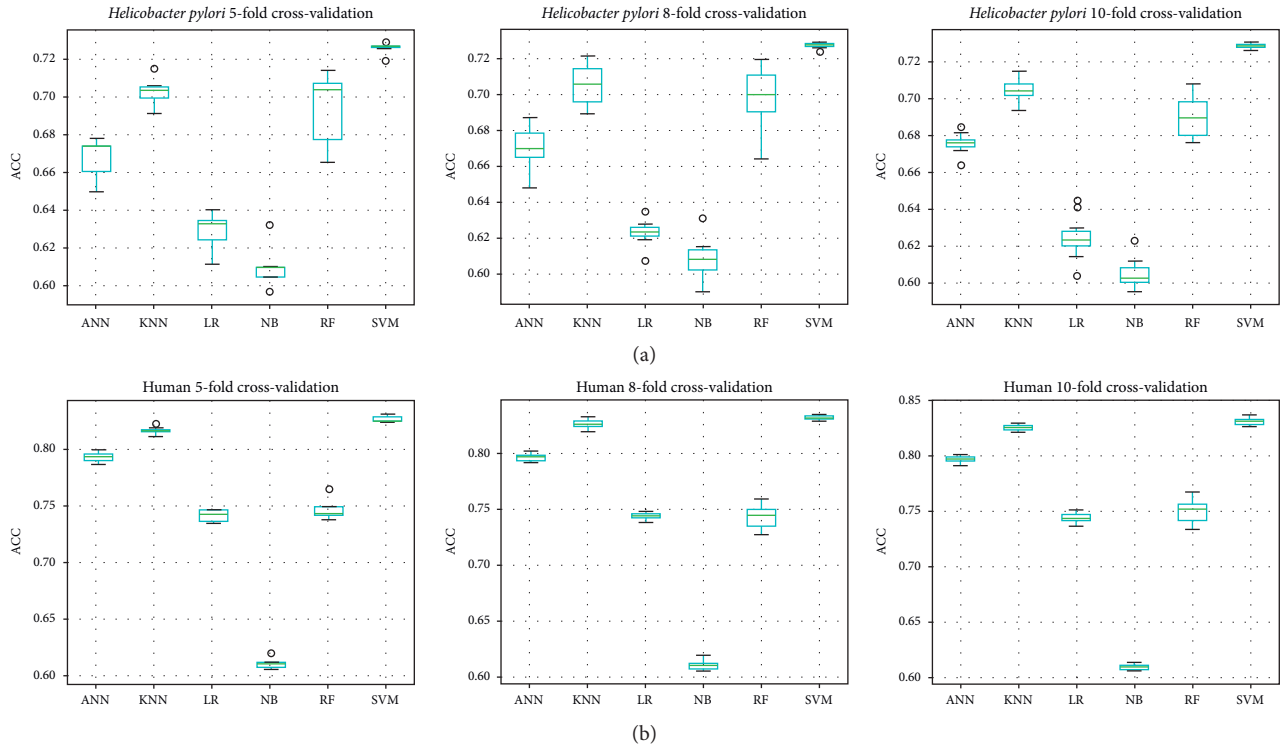


Figure 2: Cross-validated boxplots of *H. pylori* and human dataset.

similar, but from the box height, it can be seen that the accuracy of the ANN is higher. Therefore, on the *H. pylori* dataset, the best performing model using 5-fold cross-validation is ANN. Although the SVM classifier has an outlier in the 8-fold cross-validation, the impact is not significant. Since the outlier has a very small offset and the cabinet is small in size and high in position, SVM still performs best. In this way, we can see from Figure 2 in turn that, on the *H. pylori* dataset, the best performing model using 10-fold cross-validation is SVM. On human dataset, the most stable classifiers with 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation are still SVM. This can prove that the predictor which is based in SVM we proposed performs the most stability in k-fold cross-validation.

Table 2: Performance comparison in classifiers.

Dataset	Classifier	AUC	Acc	MCC	Sn	Sp
H. pylori	ANN	0.7412	0.6738	0.3544	0.5780	0.7698
	SVM	**0.8010**	**0.7279**	**0.4558**	0.7238	0.7320
	KNN	0.7746	0.7055	0.4177	**0.7942**	0.6168
	RF	0.7815	0.7030	0.4104	0.6312	**0.7749**
	LR	0.6969	0.6266	0.2602	0.5129	0.7406
	NB	0.6378	0.6043	0.2089	0.5780	0.6306
Human	ANN	0.8819	0.8008	0.6007	**0.7883**	0.8122
	SVM	0.8938	**0.8388**	**0.6791**	0.7774	0.8950
	KNN	**0.9165**	0.8118	0.6444	0.6575	**0.9531**
	RF	0.8875	0.7849	0.5812	0.6511	0.9073
	LR	0.8177	0.7505	0.5001	0.6915	0.8046
	NB	0.6089	0.6105	0.2642	0.2527	0.9378

3.2. Model Performances. To verify the reliability of our proposed method, we constructed 5 traditional classifiers for comparison, including KNN, RF, ANN, LR, and NB. We utilized *H. pylori* and human datasets and chose 8-fold cross-validation to evaluate the classifiers we constructed. Finally, we utilize 5 evaluation indicators (AUC, Acc, Sn, Sp, and Mcc) to evaluate the predictive performance of each classifier. The experimental results demonstrate that the SVM classifier performs best, as demonstrated in Table 2.

In Table 2, the AUC, Acc, and MCC values of the SVM classifier are the highest of the six classifiers, reaching 0.800963, 0.727897, and 0.455814, respectively, in the *H. pylori* dataset. The KNN classifier has the highest Sn value 0.794168, while the RF classifier has the highest Sp value 0.953052. Although the Sn and Sp values of the SVM classifier are not the highest values, they are not much lower than the highest value, which are 0.723842 and 0.731959, respectively. More importantly, the Sn and Sp values of the SVM classifier are the closest, which means that its ability to correctly predict positive and negative samples is similar. In human dataset, the Acc value of the SVM classifier reached 0.838799, and the MCC value was also the highest among the six classifiers. Although AUC, Sn, and Sp are not the highest values, they are close to the highest values. As in the *H. pylori* dataset, the SVM classifier has the smallest difference in its ability to identify positive and negative samples. From these data, it is clear that the SVM classifier has higher accuracy, pretty good stability, and higher reliability compared to the other five classifiers. Thus, the superior performance of our proposed method has been further verified.

Table 3: Comparison of features.

Dataset	Classifier	Method	AUC	Acc	MCC	Sn	Sp
H. pylori	ANN	FCTP	0.6772	0.6137	0.2337	0.5009	0.7268
		ACT	0.7412	0.6738	0.3544	0.5780	0.7698
	SVM	FCTP	0.7038	0.6549	0.3099	0.6535	0.6564
		ACT	0.8010	0.7279	0.4558	0.7238	0.7320
	KNN	FCTP	0.5747	0.5554	0.1148	0.6913	0.4192
		ACT	0.7746	0.7056	0.4176	0.7942	0.6168
	RF	FCTP	0.7553	0.6601	0.3372	0.5043	0.8162
		ACT	0.7815	0.7030	0.4104	0.6312	0.7749
	LR	FCTP	0.6866	0.62578	0.2547	0.5489	0.7027
		ACT	0.6969	0.6266	0.2602	0.5129	0.7406
	NB	FCTP	0.5186	0.5013	0.0024	0.6072	0.3952
		ACT	0.6378	0.6043	0.2089	0.5780	0.6306
Human	ANN	FCTP	0.8980	0.8275	0.6541	0.8024	0.8504
		ACT	0.8819	0.8008	0.6007	0.7883	0.8122
	SVM	FCTP	0.8320	0.7582	0.5150	0.7274	0.7864
		ACT	0.8938	0.8388	0.6791	0.7774	0.8950
	KNN	FCTP	0.9373	0.8557	0.7181	0.7652	0.9384
		ACT	0.9165	0.8118	0.6444	0.6575	0.9531
	RF	FCTP	0.8950	0.8112	0.6247	0.7357	0.8803
		ACT	0.8875	0.7849	0.5812	0.6511	0.9073
	LR	FCTP	0.8180	0.7444	0.4873	0.7146	0.7717
		ACT	0.8177	0.7505	0.5001	0.6915	0.8046
	NB	FCTP	0.6472	0.6414	0.3525	0.2822	0.9701
		ACT	0.6089	0.6105	0.2642	0.2527	0.9378

3.3. Comparison of Features. For classification issues, the performance of a model is determined by many aspects. One of the very important factors is the choice of feature extraction methods. Feature extraction can transform our original data into features that can better represent the data, improve the prediction accuracy of unknown data, and directly affect the prediction results of the model. Nowadays, researchers have proposed many feature extraction methods, which are dedicated to abstracting the most effective features for classification and recognition from the data. In this paper, we will utilize 6 prebuilt classifiers to compare our feature extraction method ACT with the feature extraction method FCTP proposed by Kong et al. Comparative experimental results are demonstrated in Table 3.

The experimental results demonstrate that, in the *H. pylori* dataset, the five evaluation indexes of the six classifier models using our proposed feature extraction method are better than those using FCTP. In the human dataset, the performance of the model constructed by our method combining SVM and LR is better than that of Kongs' method. In particular, our proposed model ACT-SVM has an Acc value which is 0.08 higher than that of the model using FCTP. Although on the human dataset FCTP performs better on ANN, KNN, RF, and NB, our method also demonstrates good performance with a small gap in indicators in all aspects. Overall, FCTP performed well on the human dataset but performed poorly on the *H. pylori* dataset. Our feature extraction method demonstrates good prediction performance on both datasets and is relatively stable. Therefore, the method we proposed is further proved to be a reliable and stable prediction model for PPIs.

4. Conclusions

In recent years, the problem of identifying PPIs has been valued by researchers and in-depth research. Several efforts to solve this problem have appeared one after another. Although machine learning methods are widely utilized in the prediction of PPIs, there is still a lack of predictors that can accurately and efficiently make predictions. Our proposed model ACT-SVM can effectively predict PPIs. We utilize a combination of A vector, composition, and transition (CT) descriptors as the digital features of the amino acid sequence and utilize them as input to train the SVM model. We evaluate the performance of our proposed method by constructing multiple classifiers using 5-fold cross-validation, 8-fold cross-validation, and 10-fold cross-validation. With these evaluations, we can easily get the conclusion that the model we proposed has the better performance in the majority of situations. The prediction accuracy of our method for *H. pylori* data reaches 0.727897, and the prediction accuracy for human dataset reaches 0.838799. The experimental results demonstrate that our proposed model based on SVM can efficiently predict PPIs. It has good performance on *H. pylori* and human dataset and can be utilized as a research tool to support biomedical and other fields.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China (nos. 61902337 and 61702445), the grant from the Ph.D. Programs Foundation of Ministry of Education of China (no. 20120072110040), the grant of Shandong Provincial Natural Science Foundation, China (no. ZR2018LF005), Natural Science Fund for Colleges and Universities in Jiangsu Province (no. 19KJB520016), and Jiangsu Provincial Natural Science Foundation (no. SBK2019040953).

References

[1] S. Brohee and J. Van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 488, 2006.

[2] N. Sugaya and K. Ikeda, "Assessing the druggability of protein-protein interactions by a supervised machine-learning method," *BMC Bioinformatics*, vol. 10, no. 1, p. 263, 2009.

[3] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[4] Q. C. Zhang, D. Petrey, L. Deng et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[5] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Mäkelä, and S. Hautaniemi, "Integrated network analysis platform for protein-protein interactions," *Nature Methods*, vol. 6, no. 1, pp. 75–77, 2009.

[6] J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks," *PLoS Computational Biology*, vol. 6, no. 6, Article ID e1000807, 2010.

[7] R. K. Barman, S. Saha, and S. Das, "Prediction of interactions between viral and host proteins using supervised machine learning methods," *PLoS one*, vol. 9, no. 11, Article ID e112034, 2014.

[8] Z.-H. You, J.-Z. Yu, L. Zhu, S. Li, and Z.-K. Wen, "A MapReduce based parallel SVM for large-scale predicting protein-protein interactions," *Neurocomputing*, vol. 145, pp. 37–43, 2014.

[9] S. Patel, "DeepInteract: deep neural network based protein-protein interaction prediction tool," *Current Bioinformatics*, vol. 12, pp. 551–557, 2017.

[10] G.-H. Liu, H.-B. Shen, and D.-J. Yu, "Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures," *The Journal of Membrane Biology*, vol. 249, pp. 141–153, 2016.

[11] P. Chatterjee, "PPI_SVM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables," *Cellular and Molecular Biology Letters*, vol. 16, no. 2, pp. 264–278, 2011.

[12] Z.-H. You, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, 2013.

[13] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine*, vol. 83, pp. 67–74, 2017.

[14] T. Sato, M. Hanada, S. Bodrug et al., "Interactions among members of the Bcl-2 protein family analyzed with a yeast two-hybrid system," *Proceedings of the National Academy of Sciences*, vol. 91, no. 20, pp. 9238–9242, 1994.

[15] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[16] P. Coates and P. Hall, "The yeast two-hybrid system for identifying protein-protein interactions," *The Journal of Pathology*, vol. 199, no. 1, pp. 4–7, 2003.

[17] R. B. Free, L. A. Hazelwood, and D. R. Sibley, "Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy," *Current Protocols in Neuroscience*, vol. 46, no. 1, pp. 5–28, 2009.

[18] Y. Kim and S. Subramaniam, "Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships," *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 4, pp. 1115–1124, 2006.

[19] V. S. Rao et al., "Protein-protein interaction detection: methods and analysis," *International Journal of Proteomics*, vol. 2014, Article ID 147648, 12 pages, 2014.

[20] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo, "Extracting human protein interactions from MEDLINE using a full-sentence parser," *Bioinformatics*, vol. 20, no. 5, pp. 604–611, 2004.

[21] P. Aloy and R. B. Russell, "InterPreTS: protein interaction prediction through tertiary structure," *Bioinformatics*, vol. 19, no. 1, pp. 161-162, 2003.

[22] Y.-A Huang et al., "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics17*, vol. 1, p. 184, 2016.

[23] S.-W. Zhang, L.-Y. Hao, and T.-H. Zhang, "Prediction of protein-protein interaction with pairwise kernel support vector machine," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 3220–3233, 2014.

[24] L. Liu, Y. Cai, W. Lu, K. Feng, C. Peng, and B. Niu, "Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection," *Biochemical and Biophysical Research Communications*, vol. 380, no. 2, pp. 318–322, 2009.

[25] X. Li, B. Liao, Y. Shu, Q. Zeng, and J. Luo, "Protein functional class prediction using global encoding of amino acid sequence," *Journal of Theoretical Biology*, vol. 261, no. 2, pp. 290–293, 2009.

[26] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[27] L. Nanni, "Hyperplanes for predicting protein-protein interactions," *Neurocomputing*, vol. 69, no. 1-3, pp. 257–263, 2005.

[28] L. Burger and E. Van Nimwegen, "Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method," *Molecular Systems Biology*, vol. 4, no. 1, p. 165, 2008.

[29] L. Nanni and A. Lumini, "An ensemble of K-local hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 22, no. 10, pp. 1207–1210, 2006.

[30] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-Scale prediction of human Protein–Protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[31] G. Singh et al., "Springs: prediction of protein-protein interaction sites using artificial neural networks," *PeerJ Pre-Prints*, 2014.

[32] S. Dohkan, A. Koike, and T. Takagi, "Prediction of protein-protein interactions using support vector machines," in *Proceedingsof the Fourth IEEE Symposium on Bioinformatics and Bioengineering*, May 2004.

[33] G. Cui, C. Fang, and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model," *BMC bioinformatics*, vol. 13, no. Suppl 7, 2012.

[34] J. R. Bradford and D. R. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, 2005.

[35] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[36] A. Koike and T. Takagi, "Prediction of protein-protein interaction sites using support vector machines," *Protein Engineering Design and Selection*, vol. 17, no. 2, pp. 165–173, 2004.

[37] Z.-H. You, "Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines," *BioMed Research International*, vol. 2015, Article ID 867516, 9 pages, 2015.

[38] J.-F. Xia, K. Han, and D.-S. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor," *Protein & Peptide Letters*, vol. 17, no. 1, pp. 137–145, 2010.

[39] L. Wong et al., "Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor," in *Proceedings of the International Conference on Intelligent Computing*, Springer, Cham, Fuzhou, China, August 2015.

[40] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "DeepPPI: boosting prediction of protein-protein interactions with deep neural networks," *Journal of Chemical Information and Modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

[41] M. Kong, "FCTP-WSRC: protein–protein interactions prediction via weighted sparse representation based classification," *Frontiers in Genetics*, vol. 11, p. 18, 2020.

[42] Z. Lu, S. Lu, G. Liu, Y. Zhang, J. Yang, and P. Phillips, "A pathological brain detection system based on radial basis function neural network," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 5, pp. 1218–1222, 2016.

Hindawi
Scientific Programming
Volume 2020, Article ID 8841368, 12 pages
https://doi.org/10.1155/2020/8841368

Research Article

Dynamic Capabilities and Business Model Innovation of Platform Enterprise: A Case Study of DiDi Taxi

Ping Lin,[1,2] Xiaosan Zhang,[3] Shuming Yan,[1] and Qingquan Jiang[4]

[1]*Newhuadu Business School, Minjiang University, Fuzhou 350108, China*
[2]*Research Center for Energy Economics and Low-Carbon Development, Minjiang University, Fuzhou 350108, China*
[3]*Research Centre for Belt and Road Financial and Economic Development, Xiamen National Accounting Institute, Xiamen 361005, China*
[4]*School of Economics & Management, Xiamen University of Technology, Xiamen 361024, China*

Correspondence should be addressed to Shuming Yan; yanshuming@nbs.edu.cn and Qingquan Jiang; jiangqingquan@xmut.edu.cn

Received 19 March 2020; Revised 24 June 2020; Accepted 30 June 2020; Published 17 July 2020

Academic Editor: Chenxi Huang

Internet has revolutionized business model and given birth to sharing economy. A large number of platform enterprises are growing rapidly but with sustainability problems. Platform enterprises have to continue innovating business models in order to obtain sustainable competitive advantages. In complex and varying environment, dynamic capabilities help enterprises overcome core rigidity and promote business model innovation. This article analyzes the elements of business model innovation of platform enterprises and also the relationship between dynamic capabilities and business model innovation. It concludes that the elements of business model innovation are value proposition, product, partnership, and profit model innovation. Dynamic capabilities promote business model innovation which has different guiding effects on the cultivation of dynamic capabilities. An exploratory case study was conducted, using DiDi taxi as an example, and verified the theory model.

1. Introduction

Internet has transformed trade locations, expanded trade exchange hour, enriched trade item varieties, accelerated trade speed, and reduced intermediaries. Internet has revolutionized previous business model and given birth to sharing economy. Many platform enterprises come into being, and many unicorn enterprises have been developed under the context of sharing economy. Sharing economy influenced at first transportation and housing industry; nowadays, it influences finance, delivery, lifestyle services, and technology industry. From consumption to production, sharing economy has played a role in every field and segment market. In some segment markets, hundreds of platform enterprises appear, grow in a fast pace within a very short amount of time, and eventually become competitive and representational enterprise. Advanced development of Internet technology has provided a solid foundation for platform enterprise's growth but, at the same time, led to brutal competition. Unfortunately, many platform enterprises flourished for only a few years before their decline and even closing-down. Therefore, it can be concluded that the innovation of business model is never once and for all. Under the dynamic and complex environment, enterprises must continue innovating business model to obtain sustainable competitive advantage.

Compared to traditional enterprise, platform enterprises face much more uncertainties, including faster-changing technology environment, lower entering barrier, and intense competition. What is worse, it is much easier to be imitated. On the one hand, the factors that lead to fast development include technology revolution, market globalization, uncertainty of consumer preferences, ever-changing industry, and institutional environment. On the other hand, the interaction between different factors makes the competitive environment more complicated. In order to become flexible

and maintain competitive advantage in such dynamic environment, platform enterprise must foster dynamic capabilities to overcome core rigidity and innovate business model constantly. Therefore, it is necessary to combine dynamic capabilities with business model innovation.

As of now, although there is enormous research focused on dynamic capabilities as well as business model innovation, very few studies explore the intersection and interaction of the two. Meanwhile, there is also a lack of research on relationship between dynamic capabilities and business model innovation for platform enterprise. Therefore, this paper aims to analyze features of platform enterprise under Internet environment, using DiDi taxi as an example, to study how dynamic capabilities can promote business model innovation to obtain sustainable competitive advantage. It concludes that dynamic capabilities promote business model innovation which has different guiding effects on the cultivation of dynamic capabilities. The main contributions of this paper can be summarized in the following two aspects. First, this paper enriches the research of dynamic capabilities in the era of sharing economy. Second, this paper guides enterprises, especially platform enterprises, toward combining business model innovation with dynamic capabilities and taking various measures to ensure the success of business model innovation.

2. Theoretical Analysis

2.1. Business Model Innovation. Different scholars have different interpretations for the concept of business model. J. Magretta [1] believes that business model portrays a story of the enterprise that catches attention of investors. Rappa [2] believes that business model shows how the business operates fundamentally. Osterwalder et al. [3] believe that business model creates value for enterprise to maintain competitive advantages. Zott et al. [4] believe that business model describes the operation of the enterprise from a systematic and holistic perspective.

After analyzing and classifying more than thirty business models, Morris et al. [5] believe that business model can be defined by economy, operation, and strategy. Economic definition classifies business model as economy model. It suggests that business model defines how the enterprise generates profit, and the main factors include operational planning, financial and human resources process, risk management, and internal operation mechanism. Operational definition classifies business model as competitive and development strategy. Its main factors include product differentiation, enterprise vision, and cooperative partners. Strategy definition at first focused only on enterprise's internal profit model, product, and operation management and marketing strategy; nowadays, it focuses mainly on how the company benefits through seizing market opportunities and providing products to consumers. In the case of start-up companies, the business model is the profit model, since the enterprise could only create unique value by identifying consumers' urgent unsolved needs in the market. Although the current research results have not reached a unified understanding, most of them agree on the concept that business model explains how enterprises create value in a systematic and holistic perspective. This paper defines business model as an elaboration on enterprise's production and operation activities.

However, scholars' definitions of business model innovation are mainly based on the universal industries and do not concern the business model innovation of special industries with specific means or approaches. For different industries, the broad dimension of business model innovation is basically the same, but for the research on specific approaches under the broad dimension, there are few theoretical studies. From the industrial period to the current Internet period, decentralization and massive interaction between people, driven by Internet, have led to sharing economy. Sharing economy utilizes Internet and other information technology to integrate and share massive decentralized idle resources and satisfy the sum of economic activities, whose demands are diverse. With Internet as foundation, sharing economy stimulated the birth and growth of many platform enterprises. "Platform" is a vivid depiction: the platform enterprise does not produce any product or create any core value but connects core value creators and consumers. It simply creates a platform for service provider to serve consumers effectively. Throughout the process of creating the platform, platform enterprise develops from "trading platform" to "reliable platform" and then to "trusted platform." When "trading platform" carries out transaction effectively while having constructed qualification examination certification, management, and insurance, it turns into "reliable platform," which reduces transaction cost by a large amount. After establishing the reliable transaction relationships on the platform, participants develop emotional exchange and even spiritual exchange; therefore, "reliable platform" becomes "trusted platform." "Trusted platform" has a "magnetizing effect" on participants, facilitating communication and understanding between two sides of transaction while further promoting trade. For example, Alibaba started as a "trading platform" for consumers, but when it realized that consumers always worry about being deceived, it conducted qualification examination and background check before anyone enters the market, which helps in investigation when fraud occurs. Therefore, Alibaba has established a trust relationship with consumers by providing them with a safe and secure transaction platform. With more transaction, many participants start to communicate and share through the Internet; as a result, participants develop "trusted" relationship with each other. This paper defines platform enterprise as the enterprise which relies on Internet as the foundation, connects communication as well as multiple markets, realizes the integration of multiple industries, satisfies demands from a variety of groups, and profits from such action.

In the Internet period, platform enterprise's business model undergoes some great changes. Platforms link the supply and demand sides of the resource, so people buy products and services on a point to point basis rather than on the corporate basis. Both resource providers and resource demanders could participate in sharing economy at any time and from any place; as a result, larger scale and wider range of idle resources are shared. Enterprises and customers all

become value creators, while products and services are not specifically produced but are instead derived from recycling of idle resources. All in all, the new characteristics of Internet platform enterprises promote the business model innovation. There are unique ways in business model innovation of platform enterprises.

2.1.1. Customer Value Proposition Innovation. Customer value proposition represents an in-depth description of the customer's real needs. It also represents a unique value that enterprises convey to users for them to understand and make a purchase of a product or service, and for the company to enhance brand loyalty. The traditional business model is mainly based on the core capabilities of the enterprise and it provides what it produces to the customers. Usually, the products or services are standardized with no adjustment or specialization to customer need or market change. Under the sharing economy, enterprises are no longer limited to products provided by enterprises. Any resource owner can share his unique idle resources on the platform at a lower price to provide more unique products for the demand side. As a result, a large number of personalized, diversified products can provide diverse experience for participants and satisfy deeper emotional needs of the users. At the same time, the sharing platform increases the variety of the transaction objects and the products, reduces the buyer's information asymmetry, and reduces the transaction cost between seller and buyer, thus improving the satisfaction and economic value of the two parties. There are two main ways for platform enterprises to innovate customer value. The first is to identify customer needs through in-depth research and analysis. The second is to create consumer demand. According to the spiritual needs and other intangible needs of consumers, enterprises find new interest points and seek consumers' recognition. For example, Uber's service is positioned as "Apple in taxi software," and its brand slogan, "your personal driver," has greatly enriched the customer's ride experience.

2.1.2. Product Innovation. Product is the connection between enterprise and consumer. When the consumer demand is found or created, the enterprise would present the customer value proposition with the product accordingly. Product innovation is to provide users with new products or better services through technological innovation, product design update, or service system improvement. One important prerequisite of platform enterprise is to have many users, and another is to have sticky service. Therefore, the products and services provided by the platform should constantly be enriched for the connection between the suppliers and the demanders to be more stable.

As more information can be searched by Baidu in China, users gradually equate search with Baidu, so more and more providers are willing to provide information on Baidu. Similarly, Alibaba relies on making it easier for users to find goods online, and Tencent specializes in connecting users with others to form a strong network effect. After the platform enterprise creates stable connection, there are two other prerequisites for further success. One is whether more and more different types of users can be pulled into the platform system through more products and services, and the other one is whether a transition from bilateral connections to multilateralism can be achieved. With more diverse participants, the platform tends to be stickier and more efficient and gradually develops into a huge ecosystem. Therefore, platform enterprises should pay special attention to product innovation to increase the quantity and stickiness of customers and meet the needs of customers.

2.1.3. Profit Model Innovation. The profit model refers to the income structure, cost structure, and corresponding target profits divided according to different stakeholders. Profit model innovation mainly includes cost control innovation and revenue mode innovation. The cost control innovation of traditional enterprises mainly focuses on large-scale production, outsourcing, acquisition, and other means. The innovation of income model focuses on increasing sales through enhancing publicity, improving service level, and carrying out the promotion activities. There are two main ways for platform enterprise to achieve cost innovation. The first one is to allow both sides to trade through shared platform, and reduce the transaction cost. The second is to "reuse" idle resources, which not only eliminates the production cost, but also improves the value of the product by repeated utilization with high efficiency. The revenue model innovation of platform enterprises mainly includes transaction profit, service charge, and advertisement fee. Transaction profit refers to the transaction commission based on a certain proportion of the supplier or the buyer's revenue. Service charge refers to charging the registered member for provided service. For example, some websites would provide movie or music downloading service for paid members. Advertisement fee establishes a good benefit sharing mechanism by charging of advertising fee from the third advertising parties based on the flow rate, instead of charging consumers directly. Baidu and Google's process of bidding for advertising serves as an example of advertisement fee.

2.1.4. Cooperative Partnership Innovation. Cooperative partnership refers to the long-term cooperative relationship of sharing information, sharing risks, and sharing interests between enterprises because of mutual trust, in order to achieve the common goals of both parties. The innovation of partners in traditional enterprises is mainly aimed at partners with no competitive relationship with the enterprise, such as the upstream suppliers and the downstream distributors, by means of strengthening the quality of the cooperative relationship, removing the useless cooperation relationship, or adjusting the mechanism of the cooperative relationship. And the partnership innovation of the platform enterprise focuses on finding the common interest with the third-party enterprises such as competitors and cross-border enterprises, transforming the competitive relationship into organic win-win cooperation, with each enterprise contributing its own resources, and maximizing the overall interests. Partner innovation can be achieved through

screening potential partners or adjusting relationships with different partners, or through choosing a variety of cooperative ways, which include buying resources from third parties and developing jointly with the third party.

2.2. Dynamic Capabilities. Since the 1990s, the dynamic capabilities theory has been widely accepted by scholars. The dynamic capabilities theory is based on the resource-based view and capability theory. Both theories believe that the competitive advantage of the enterprise comes from the resource bundle and core competence of the enterprise. Teece [6] proposed the dynamic capabilities as the abilities to build, reset, and integrate the internal and external resources in order to cope with the fast-changing environment, to maintain the competitive advantage. Teece [7] integrates the abilities to perceive and grasp opportunities into the connotation of dynamic abilities. When studying the microcosmic foundation of dynamic capabilities, he proposes that such capabilities refer to the abilities to perceive, create, and seize opportunities and to combine, strengthen, protect, and even reconstruct internal and external resources to maintain the competitive advantage. Other scholars regard dynamic capabilities as a convention, process, and pattern. Zollo & Winter [8] define a dynamic capability as a learned and stable pattern of collective activity through which the organization systematically generates and modifies its operating routines in pursuit of improved effectiveness.

At present, scholars have not reached consensus on the dimension of dynamic capabilities. While Teece et al.[6] make it clear that dynamic capabilities are composed of three dimensions of integration, construction, and reconstruction, Wang & Ahmed [9] point out that dynamic capabilities consist of three parts: absorption, adaptation, and innovation. Absorptive capacity refers to the absorbing external knowledge and integrates it into internal knowledge; adaptability emphasizes the matching of the internal elements of enterprises and the external environment; innovation ability refers to the way enterprises rely on internal innovation to obtain market advantages. Teece et al. [10] point out that dynamic capabilities include not only the ability to integrate and reconfigure resources, but also the cognitive dimension, that is, to perceive and identify opportunities and threats. The reason is that when facing a rapidly changing global environment, enterprises could maintain competitive advantage only through responding to opportunities and threats in time. On the other hand, Pavlou & Eisawy [11] and Lin [12] define dynamic capabilities as perception, learning, coordination, and integration capabilities.

Sorting the above literature, this paper finds that the research on dynamic capability is universal. Less attention is paid to the dynamic capability of a specific industry. The environment of the Internet age is full of uncertainties. Accordingly, the management of enterprises is also highly uncertain. The most important thing for the platform enterprise is continuously integrating and restructuring resources and capabilities to maintain current updates. Accordingly, dynamic capability is not only the integration of internal resource, but also the integration of internal and external resources of enterprises. This paper defines dynamic capabilities as the ability to identify and seize market opportunities and integrate the internal and external resources in order to cope with the changing external environment and sustain competitive advantages. The dimensions of dynamic capabilities of platform enterprises are market perception capability, learning and integration capability, coordination capability, and organizational flexibility, and the emphasis of each capability of platform enterprise is different from that of traditional enterprise.

2.2.1. Market Perception Capability. With the industrial economy, manufacturers create value through a series of activities within the value chain. However, in the era of Internet, customers are the source of value creation and part of value creation. The sharing platform creates the conditions for the user to communicate and share in a socialized way. A public and transparent evaluation system is set up between the suppliers and the consumers, and the consumers can also check the product reviews by other consumers at any time before making a purchase decision. Consumers pay more attention to personalized needs and focus more on the product experience, as well as social, leisure, emotional, and other deeper aspects of the experiences. And more consumers take the initiative to participate in the design and research of products. For example, Xiaomi, mobile phone company, will have the users vote in the forum every Friday to determine the direction of research and development; the forum has about more than a million visits per day and there are nearly 300 thousand posts overall. Users of Xiaomi are deeply involved in company's research and development, and vice president Li Wanqiang has said that Xiaomi's purpose is not selling products, but creating a sense of participation. Moraes [13] believes that value is derived from the experience brought to the customer. On the other hand, Gourville [14] believes that value is driven by thoughts of the customer, and they create value and share value with the enterprise.

Therefore, for the platform enterprise, it is very important to have market perception capability, which equips the enterprise with ability to quickly interpret the changes of government or industry policies, grasp the latest trends of competitors, see the demand of consumers for frequent changes, and find the new market opportunities more efficiently and quickly. Moreover, the technology entering barrier of Internet enterprises is low, easy to imitate, and full of uncertainty. Consequently, a good business idea would be produced by dozens of similar enterprises at the same time, which all compete for customer resources. In each vertical segment, there is only one or two leaders, and the third one is easily bought out or goes bankrupt eventually. Therefore, many companies attract consumers quickly by burning money to capture the market share. All in all, it is crucial for platform enterprise to have strong market perception capability.

2.2.2. Learning and Integration Capability. Learning and integration capability refers to the ability of an enterprise to restructure and integrate internal and external resources in

order to respond to environment changes. When the enterprise identifies the new market opportunities in the external environment and defines the innovation of the business model, it should redeploy and reintegrate the resources, by promoting exchanges and cooperation between the marketing, sales, R&D, administration department, etc. It should first determine whether its own resources are enough to ensure the development and implementation of the new products. Next, if the internal resources are insufficient, the enterprise should introduce external resources through the means of learning, cooperation, and purchase. It is crucial for the enterprise to internalize the new resources and integrate them with the internal resources, including very specific knowledge like stakeholders' rights, the customer, the employees, and cross-border, internalizing external knowledge into organizational knowledge. The integration matches the enterprise' resources with its new desired strategic position and strives to maintain enterprise's leading competitiveness.

2.2.3. Coordination Capability. Coordination capability refers to the ability to coordinate with investment institutions, suppliers, advertisers, competitors, cross-border enterprises, and so on. In the Internet age, cooperation capability not only improves the product's adaptability, but also transforms the competitive relationship into a cooperative relationship strategically, making many unrelated and incompatible elements connected to create new values. Therefore, it is important for enterprise to coordinate the relationship between the stakeholders, such as competitors and partners, in order to share qualified resources, cooperate in the competition, and compete in cooperation. The coordination capability includes the ability to select high-quality cooperative partner, to establish and maintain a good lasting relationship with stakeholders, to bargain effectively, and to construct a considerable interest sharing mechanism.

2.2.4. Organizational Flexibility. In this process of change, the corporation has to prevent itself from overshooting and becoming extremely rigid or chaotic. Enterprise that has organizational flexibility reacts to changes effectively by restructuring the business process, organization structure, communication system, and management system. In the Internet economy, learning organization, information sharing cloud platform, and network organization help to improve the organizational flexibility greatly.

2.3. Relationship between Platform Enterprises' Dynamic Capabilities and Business Model Innovation. Now the research on the formation mechanism of dynamic capability is still limited. This paper chooses business model innovation as the starting point of dynamic capability cultivation and discusses the relationship between business model innovation and dynamic capability.

Firstly, in a dynamic environment, the dynamic capabilities of platform enterprises can effectively promote the successful implementation of business model innovation (Najmaei) [15]. For example, the market perception ability can monitor the external environment in real time, grasp the changes in consumer demand, interpret the changes in government or industry policies, grasp the latest trends of the competitors, better satisfy the consumer needs, and find new market opportunities more efficiently and faster than the competitors. This promotes customer value proposition innovation. The integration ability allows enterprise to absorb external knowledge, transform it into internal knowledge, and reorganize and integrate all resources in time to ensure the quality and speed of product development, thus promoting product innovation. The coordination ability is favourable for the platform enterprises to establish good sharing mechanism with the stakeholders, to increase the income source and proportion, and to match the strategic positioning with own resources, thus promoting the income model innovation. Teece [16] mentioned that the strength of a firm's dynamic capabilities helps to shape its proficiency at business model design. Through its effect on organization design, a business model influences the firm's dynamic capabilities and places bounds on the feasibility of strategies. Organizational flexibility adjusts the organizational structure and functional modules according to the dynamic changes of the external environment to respond to the needs of customers fast and flexibly and to promote partnership innovation.

Secondly, platform enterprises choose feasible and new business models according to the change of dynamic capabilities. Warner and Wager [17] proposed a process model with nine micro foundations to reveal the generic contingency factors that can trigger, enable, and hinder the building of dynamic capabilities for digital transformation, revealing that agility is the core mechanism for the strategic renewal of an organization's business model. Business model innovation can be based on the original business model, or also a completely new business model. When the new type of business model is selected, the platform enterprise should determine whether its own capabilities and resources can support the success of the new business model. If the new model can be supported, the platform enterprise will rely on the existing capabilities and resources to carry out the business model innovation. If not, the platform enterprise must first develop its own dynamic capabilities. When the new dynamic capabilities are formed, the enterprise can effectively promote the successful implementation of business model innovation, thus obtaining the leading competitive advantage.

Finally, the different paths of business model innovation have different guiding effects on the cultivation of dynamic capabilities. If the new dynamic capabilities are still unable to provide effective support, it is necessary for the enterprise to revise the cultivation measures of the dynamic capabilities and to carry out a new round of cultivation. Therefore, in order to maintain the leading competitive advantage, platform enterprise has formed an internal dynamic cycle to cultivate dynamic capabilities of leading business model innovation (Figure 1).

In addition, business model innovation cannot be successful with only one try. In order to maintain competitive

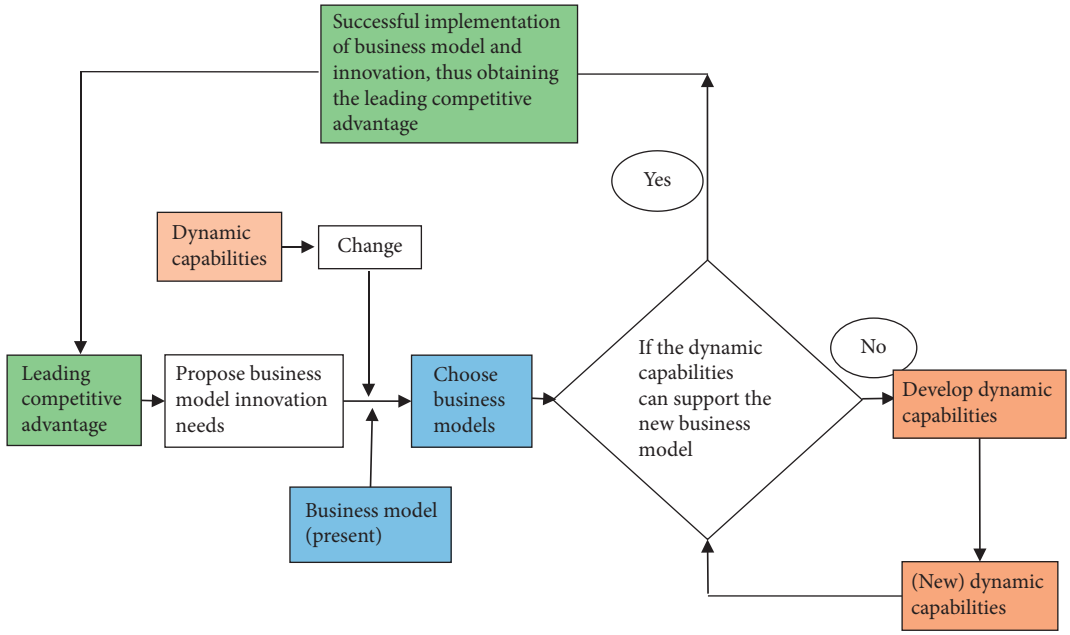


Figure 1: The relationship between platform enterprises' dynamic capabilities and business mode innovation.

advantage in a turbulent environment, enterprises are constantly making use of dynamic capabilities to continue to innovate business model. After the cultivation of the platform enterprise dynamic capabilities (present) is achieved, the dynamic capabilities (New 1) can identify the new changes, reorganize the various resources of the enterprise, and promote a new round of business model (New 1) innovation. Therefore, there is a dynamic cycle between the cultivation of dynamic capabilities of platform enterprises and business model innovation (Figure 2). In this process, business models are continuously innovated, and dynamic capabilities of platform enterprises are continuously improved.

3. Research Design

3.1. Selection of Research Methods. This study takes an exploratory single case study as example. First, it explores the business model innovation of the platform enterprise and the role of dynamic capabilities under the Internet environment. There is few specific answers in the existing literature or publications. A case study allows the investigation of a phenomenon in the real-life setting, so it is suitable for this research to adopt the case study method. Second, in-depth description and analysis of specific phenomena are helpful to understand the dynamic and complex mechanism behind a phenomenon. Eisenhardt and Graebner [18] point out that case study is a very objective and rigorous empirical method that is close to reality. They introduced two different methods of case analysis in detail: phenomenon driven and theory driven case analysis methods. The phenomenon driven type strives to establish the theory from the phenomenon and is applicable to situations where there is no feasible theory. The theory driven type is based on the existing theory to establish the framework and to discover the qualitative data for verifying and developing the theory. This research paper adopts the theory driven case study method.

In order to verify and expand the theory, researchers often choose extreme and enlightening cases that satisfy theoretical requirements. Such cases help to exclude the interference of research factors in the typical situation, so that the researchers can observe and analyze the more valuable scientific phenomena and make the research conclusions more powerful. Therefore, this study chooses DiDi taxi (also referred to as DiDi) as a case study. There are three main selection criteria: Firstly, DiDi has been set up only 6 years ago and has become the largest station mobile travel platform in the world. Secondly, DiDi is a new model of "Internet plus traditional industry." Its business model innovation practices the typical characteristics of the idea of sharing economy under the Internet environment, and it is helpful to analyze the relationship model between the platform enterprise's dynamic capabilities and the business model innovation. Finally, DiDi is facing rapidly changing dynamic environment. Specifically, great changes, regularized and coordinated by government and transportation authorities, have been made in the industry policy environment. The market competition has become much fiercer due to competitors employing other business models or other service providers like shared bicycle platform, while consumers' preferences are always changing.

3.2. Case Introduction. DiDi business includes taxis, special cars, express cars, and buses, getting through the O2O closed loop. Since its inception in 2012, DiDi has served more than 400 domestic cities to meet 300 million of the travel demands of consumers. In 2015, DiDi's online platform completed a total of 1.43 billion orders, making DiDi China's second largest online trading platform after the top one

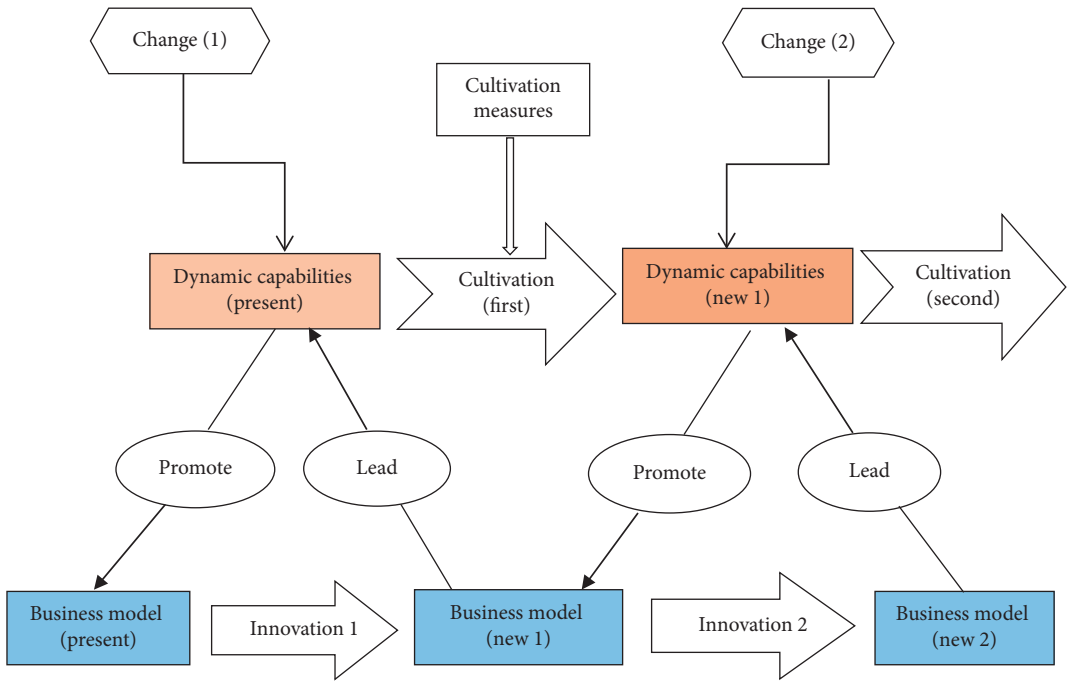


Figure 2: Dynamic cycle between the cultivation of dynamic capabilities and business model innovation.

Taobao. Its valuation in 2017 has exceeded $16.5 billion with more than 200 million users, accounting for more than 87% of the domestic car market and 99% of the online taxi market. The course of its development is shown in Table 1.

3.2.1. Phase 1: Market Cultivation Period. During this period, DiDi's overall growth slowed, but it eventually overcame the biggest challenge by introducing a new-brand platform product which operates with a difficult product logic to cultivate users and products. Since its establishment on July 10, 2012, DiDi developed better product and service to expand its market service for driver side and customer side. On October 28, 2012, version 1.1, following a new version of iOS, supported the new Apple system more efficiently. Moreover, it added instant messaging to the arrival of a taxi, a one button duplication function, and a simplified registration process. On December 2, 2012, version 1.2 introduced three new functions. The first one was the appointment function, which allowed users to make an appointment for the next day or even the day after it. The second was the "increased price" function: during the peak of taxi demand, it offered "increased price" for the users to enhance their success rate in getting a taxi. Last, this version also simplified the registration and landing process, so both drivers and users benefit from the convenience. On December 18, 2012, version 1.3 introduced "call-waiting function," which allowed extending the peak waiting time and was expected to raise the call success rate by 40%. This version also optimized the start speed of the software.

In 2013, from version 1.4 to 2.5, DiDi experienced many new changes. It added "instant calling" and "appointment from different cities" function. It also introduced many new functions including V2.1 sound effect, friends invite, taxi-hailing tips, presaved home and company address, order canceling, custom location, tip adjustment, and flexible reservation time. At this stage, DiDi solved the situation where a significant number of people cannot get the taxi by creating the car-hailing platform with millions of users. It has provided a real value for hundreds of thousands of people enabling them to travel more efficiently. In October 2013, iResearch Global (China) released the only industry report on taxi software: DiDi's market share was 59.4%, exceeding the market share of the other taxi software combined. DiDi's growth was mainly driven by stable market operations (vigorously advancement and multi-party cooperation), as well as constant product improvement.

3.2.2. Phase 2: Accelerated Growth Period. From January to September 2014, the number of DiDi users increased significantly. After completing the transformation from being an available travel tool to a common travel tool, DiDi acquired a larger market share. In 2014, the market competition became fierce and the growth of user base in taxi applications increased rapidly. In the face of powerful competitors like "KuaiDi," the focus of DiDi was on rapid user growth. DiDi succeeded through subsidies and cooperation with WeChat. In January 2014, DiDi reached a strategic cooperation with WeChat to open WeChat payment, making it much convenient for users to pay. DiDi also incorporated red pocket subsidy and even provided free rides. At this stage, DiDi and KuaiDi invested a total of about 2 billion yuan of market subsidies, in order to cultivate a larger market and develop the user's habit of using the software. In March 2014, the number of DiDi users exceeded 100 million and the number of drivers exceeded 1 million,

Table 1: The course of DiDi's development.

Phase	Time	Product
(1) Market cultivation period	2012.09–2012.12 2013.01–2013.12	V1.0: setting online; V1.1, V1.2: online appointment, increased price, simplifying the registration and landing process; V1.3: call-waiting function, extending the peak waiting time V1.4, V2.0: instant calling, appointment from different cities; V2.1: sound effect, friends invite, taxi-hailing tips; V2.2: presaved home and company address; V2.3, V2.4: order canceling, custom location, tip adjustment; V2.5: flexible reservation time
(2) Accelerated growth period	2014.01–2014.08	V2.6: WeChat payment; V2.7: free rides; V2.8: order deletion, optimizing the map
(3) Rapid growth period	2014.09–2015.02	V3.1: special car; V3.2: priority text "call"; V3.3: integral mall; V3.4: reservation car, subtype recommendation, estimate cost; V3.5: new recommended award columns; V3.6: default ticket for payment, new game center
(4) Integration period	2015.02-	V3.7: pay hitching; V3.9: new express, driving agent, possibility of locating all cars at home and company, Alipay, QQ wallet payment, driving service, automatic fare increase system; V4.1: blacklist, making an appointment 7 days in advance, account reminding, chartered car, investment sharing bicycle platform Ofo, car rental, overseas car rental business, artificial intelligence laboratory, driverless car

with average daily orders of 5.2183 million, making DiDi the largest daily order trading platform on the mobile Internet.

3.2.3. Phase 3: Rapid Growth Period. From September 2014 to February 2015, the number of users increased rapidly, with 5-time increase in half a year. In September, DiDi introduced "special car," which helped in its transition from the deep ploughing stock market to the beginning of the development of the incremental market. A series of optimization initiatives were introduced for the "special" user experience, such as subtype recommendation, estimate cost, default payment selection, and priority text "call." Through this series of initiatives, DiDi significantly improved the user's experience and trained a number of users to use the "special car" through the operation level, with aids like subsidy.

3.2.4. Phase 4: Integration Period of Resource Development. Since February 2015, DiDi has begun to diversify its development, aiming at all the subdivision areas of the travel industry and integrating travel resources to create a one-stop mobile travel platform. As the market became mature and saturated, the growth of users was slow and sustained subsidies were no longer a good means. Moreover, the impact of a powerful competitor was more and more significant, like Uber China. In February 2015, DiDi announced its merger with KuaiDi to nurture a bigger market. Since then, a series of services, such as express, pay hitching, bus, and driving agent, have been put online in an all-round way. DiDi became a "one-stop travel platform" for users. The service line of express and pay hitching service targeted the common users, while the special car targeted the middle to high end users. The function of the express car is foundation for the system of "combined carpool." In the future, it can be used by people in the lowest part of society, when the service of the carpool can make the car be users' "bus."

3.3. Data Collection and Analysis. This research strives to use a variety of sources to collect data, mainly including the following channels: (1) enterprise archives, such as enterprise journals, work summaries, internal rules and regulations, and enterprise annual reports; (2) official website of DiDi; (3) media reports; (4) publications of third-party research institutions. Data sources and codes are shown in Table 2.

3.4. Construction Measure and Data Coding. The understanding of key constructs is based on the cross test of the research members and the use of charts for cooperation. Through continuous data analysis and data conceptualization, the logical relationship gradually emerges, and the theory is finally saturated. The measuring dimensions of key constructs are shown in Table 3.

Data were coded in a double-blind way. After encoding the collected material into concepts, the researchers classified the concepts and assigned them to the related existing concepts. Before the formal coding began, the coders involved conducted precoding and compared the results of the coding to test the encoders' understanding of the connotations and dimensions of the study. If the consistency rate of the coding results was more than 90%, the coding process could be formally started. Data coming from the same source with similar meaning was recorded as one item. The results of the coding were presented in a tabular way. After the double-blind coding was completed, the two groups of encoders checked the coded results together. For the incompatible results, the two parties were asked to discuss them and reach a final agreement on the coding results.

This paper analyzes the business model innovation and dynamic capabilities of case enterprises and verifies the suggested theoretical model.

4. Case Findings

4.1. DiDi's Business Model Innovation

4.1.1. Customer Value Proposition Innovation. DiDi is acutely aware of the changes in the market environment of the mobile Internet technology and the development of big data. With the help of the O2O platform, a large number of

Table 2: Data sources and codes.

Source	Code
Enterprise archives	F1
Official website of DiDi	F2
Media reports	F3
Publications of third-party research institutions	F4

redundant resources and services in the society are linked with the needs by lease or purchase, so that the value of idle resources would be recreated, the utilization rate would increase, and the economic mutual benefit would be achieved. On the other hand, enterprises gain profits to support daily operations and win long-term development. The following is performed:

(1) Integrating Redundant Resources. all kinds of idle resources from private individual, car rental companies and taxi companies are integrated through the platform, and then vehicles or related services are provided to the groups in need.

(2) Improving Operation Efficiency. the advantages of the mobile Internet technology and the mining of large data are used to locate and match the supply and demand, effectively reducing the information asymmetry between the supply and demand and greatly improving the travel efficiency.

(3) Paying Attention to Service Experience. the scoring mechanism between driver and passenger is implemented, using the cloud computing and big data technology to divide drivers and passengers based on their scores, providing different service for different grades of drivers and users, and eliminating the phenomenon of "bad money driving out good money" as much as possible.

(4) Advocating Sharing Economy. a flexible, alternative way of work that can be done any time and, more importantly, with low barrier of entry is provided.

4.1.2. Product Innovation. DiDi is keenly aware of the personalized needs of Internet users in a sharing economy. It has carried out a series of product iterations to meet the needs of the market and user growth. With the convenience of data collection provided by its products, DiDi has been collecting feedback from consumers and constantly promoting functional optimization.

In the market incubation period, DiDi found out that, from the user reviews, the problem in V1.1 version is that passengers cannot register or cannot get a car. The fundamental reason of not getting a car was that there were very few DiDi drivers, thus leading to an uneven supply and demand. After introducing the appointment function, DiDi enabled users to find good cars ahead of time while increasing drivers' enthusiasm through increasing price and facilitating transactions. DiDi also solved the registration issue by canceling the registration process. In the period of accelerated growth, when facing strong competitors, DiDi cooperated with WeChat by introducing the WeChat payment function, replacing cash payment and starting a large-scale subsidy of red packet. In this stage, the product became much more well-rounded. In the rapid growth period, DiDi also launched the key product "special car" and had a series of optimization initiatives on the user experience and overall function. Based on different users' attitudes to travel service, type of the vehicle, price, and conveniency, DiDi developed corresponding products for different market segments and launched different travel services to meet personalized needs, like new model recommendation. Consumers could therefore choose vehicle type and grade according to their own consumption level, estimate cost, reservation call, default payment selection, and driver evaluation. During business integration period, based on the understanding of business and users, product version management became more refined with continuous iteration and optimization, as well as the launch of a series of products, such as express car and express bus. A one-stop mobile travel platform was created successfully.

4.1.3. Profit Model Innovation. Traditional manufacturing industries gain profits by selling products, while platform enterprises like DiDi make profit mainly from transaction costs, traffic advertising, service fees, third-party fees, and financial value.

(1) Transaction Costs. DiDi charges drivers 0.5 yuan per transaction fee, plus 1.77% management fees, as well as the revenue sharing with competitors, industry agencies, and third-party enterprises.

(2) Information Service Fee. Information service fee is collected from the passenger and the driver, respectively. In remote areas, bad weather, or busy time, the passengers and taxi drivers can choose to pay the information service fee of 0.5–2 times per single price and obtain the right to receive priority.

(3) Advertising Revenue. The peak of the daily trading volume of the platform reaches about 14000000, and the DiDi charges the third-party partners by charging the traffic according to the flow rate, which will produce considerable advertising revenue.

(4) The Financial Value of a Huge Pool of Capital. DiDi pays its drivers every Tuesday, so that the company has a stable cash pool for 7 days. According to DiDi statistics, the peak of the trading volume reaches 14000000 orders per single day. In 2015, there were 1 billion 430 million order transactions. Thus, the drop trip would have great financial value for the huge pool of capital.

(5) The Value of Information Supported by Big Data. When the user uses DiDi, the software will record the location of the car and the passenger information and can get a point map similar to the commercial map through data analysis. DiDi covers more than 400 cities, with more than 300

Table 3: Constructs and codes.

Constructs	Codes
Customer value proposition innovation	Enterprises clearly understand the target customers and explore customer needs through innovative ways
Product innovation	Product iteration, technical innovation, changing product design, or improving service system
Profit model innovation	Income structure and profit model are novel
Cooperative partnership innovation	Emphasizing win-win cooperation, developing cooperative networks in an innovative way, and constantly introducing diversified partners
Market perception capability	Keeping abreast of customer needs and trends, making accurate predictions of technology trends in the industry, keeping abreast of closely related policies, discovering and taping new market opportunities more efficiently and quickly than competitors
Learning and integration capability	Updating outdated knowledge in real time; quickly learning knowledge from the internal organization, suppliers, and competitors; introducing external resources by means of learning, cooperation, purchasing, and other means to integrate internal and external resources; applying new knowledge to new products
Coordination capability	Coordinating and cooperating with investment institutions, suppliers, advertisers, competitors, cross-border enterprises, etc.
Organizational flexibility	Flexibility in organizational structure, communication systems, management systems, procedures, and methods

million users. The detailed data have great informational value for the commercial distribution of real estate and urban construction of municipal construction departments for the improvement of traffic, and also for the consumer travel planning suppliers.

4.1.4. Partner Innovation. DiDi cooperates not only with enterprises in the industry and third-party enterprises, but also with its competitors through innovation, as follows.

(1) Enterprise Cooperation in the Industry. DiDi cooperated with agencies, travel companies, bus groups, etc. Also, it cooperated with automobile manufacturers and after-sales service providers. This shows that platform enterprises pay special attention to seeking common interests with third-party enterprises. Through such cooperation, each party contributes its own resources and creates greater value in accordance with a certain benefit distribution mechanism.

(2) Cross-Border Cooperation. DiDi cooperated with WeChat, Alipay, and other mobile payment platforms to facilitate user payment; with Baidu and Auto Navy Map to achieve precise location; with Qunar.com to add the function that allowed passengers to make online reservation from a different city; with Tencent, Lenovo, Dell, Jingdong, and other hundreds of enterprises for cross-border cooperation. It also worked with Jiangsu TV channel and WeChat platform to launch the largest red enveloped campaign in 2015, which greatly stimulated consumption. It cooperated with many other high-quality media platforms such as Ali travel and Hunan TV channel to start a harmonious marketing campaign and with service platforms such as Home Inn and Auto Navy Map.

(3) Cooperation with Competitors. DiDi acquired KuaiDi, Uber, and other competitors in the industry to avoid vicious competition and achieved the integration of useful resources and the growth of overall value. The implementation of differential operation and the integration with the traditional taxis, which helped taxis to carry out online management and technological dispatch, avoided the conflicts caused by alternative competition and improved their operation efficiency overall.

All in all, through these measures, DiDi embodies not only the uniqueness of platform enterprise's business model innovation but also the platform enterprise's priority on transforming the competitive relationship into organic win-win cooperation and achieving the maximization of the overall interest.

4.2. Relationship between Dynamic Capabilities and Business Model Innovation. Business model innovation elements and corresponding dynamic capabilities from the case are shown in Table 4.

First, dynamic capabilities promote business model innovation. DiDi constantly cultivates its dynamic capabilities, which promotes and guarantees the success of business model innovation, so that enterprises can obtain sustainable competitive advantages.

Market perception capability promotes customer value proposition innovation. Market perception is the ability of enterprises to discover and seize market opportunities and avoid all kinds of external threats. Customer value proposition is an in-depth description of the real needs of customers. Every DiDi's customer value proposition innovation benefits from its market perception ability. DiDi pays special attention to customer evaluation, listens to consumers' opinions, and realizes Internet user's personalized, diversified, experiential, and deeper emotional needs. Therefore, it quickly sees and seizes market opportunities. In time, it formulates corresponding measures according to the changes in government related policies, competitors' trend, and customer requirements. With the help of the Internet technology, as well as the precise user needs as guidance, it sets the specific user experience as the promotion target and continues to look for new interest demand points, such as

Table 4: The list of business model innovation elements and dynamic capabilities.

Business model innovation	Form	Dynamic capabilities
Customer value proposition innovation	Integrating redundant resources through the platform; using large data mining to improve operational efficiency: providing differential services for different levels of drivers and passengers to improve service experience; and advocating a sharing economy	Market perception capability: listening to customers' opinions, focusing on customer evaluation, and detecting changes in external environment
Product innovation	A series of products, such as taxi, DiDi special car, and express have been iterated, and the service system of the car under the support of big data has been established, which has formed a higher trade barrier	Learning and integration capabilities: improving the poorly evaluated products, obtaining new knowledge from the external environment, such as consumers, competitors, and partners, integrating them into internal knowledge, and ensuring the speed of product iteration Organizational flexibility: quickly responding to customer needs, adjusting organizational structure and only module
Profit model innovation	The proportion of orders fee, information service fees, advertising revenue, the financial value of large pools, and the information value of big data	Coordination capability: finding common interests with competitors and cross-border enterprises and turning the competitive relationship into organic win-win cooperation
Partnership innovation	Partnership with WeChat, Alipay, and other mobile payment platforms; partnership with Micro-Blog, the Front, Ali Travel, Hunan Satellite TV, and other media platforms; partnership with the GAD map, Home Inn, and other living service platforms	Coordination capability: good relationship with Tencent, Alibaba, Sina, Auto Navy Map, and other cross-border enterprises, as well as good negotiation and communication skills
	Merger with competitors in the industry, such as the KuaiDi, integration with the traditional cruising taxi, cooperation with the tourist company, the bus group, the automobile manufacturer, the service provider of the auto after-sale service, and so on	Coordination capability: Good relationship with competitors, taxis, and upstream and downstream companies such as the after-sales market, as well as negotiation and communication skills

WeChat payment and special car. By doing so, DiDi achieves consumer's approval and promotes the innovative customer value proposition rapidly.

Learning and integration capabilities promote product innovation. Learning and integration mean the ability of enterprises to integrate new knowledge with old knowledge and effectively combine internal and external resources, on the basis of receiving and digesting new information. After identifying external environment changes and market opportunities, DiDi focuses on customer evaluation, especially bad reviews, and continues to obtain useful information and identify beneficial resources and opportunities from consumers, competitors, suppliers, partners, and other external markets, effectively translating the new information into the internal knowledge. Moreover, DiDi explores beneficial new knowledge that would be beneficial for product improvement, firm internal operation, and future development and improves firm's ability to adapt to external change. In addition, DiDi internalizes external resources in time while constantly adjusting the business model and updating existing products and services. By doing so, DiDi ensures the speed of product iteration and establishes strategic barriers so that competitors cannot easily obtain the same speed and quality of product development, even though they have the same resources.

Coordination capability promotes profit model and partner innovation. DiDi focuses on finding interests in the third-party enterprises such as competitors and cross-border enterprises, transforming the competitive relationship into organic win-win cooperation. Each subject contributes its own resources and maximizes the overall interests according to a certain benefit distribution mechanism.

Organizational flexibility promotes product innovation. In the dynamic external environment, DiDi's organizational flexibility is manifested by customer evaluation and the fast adjustment done by the sales, marketing, R&D, and organizational administration and functional modules basis. Such fast and elastic changes can respond to customer needs in time and promote product creation.

Then, business model innovation has different guiding effects on the cultivation of dynamic capabilities. In the past 6 years, the business model innovation of DiDi was not a single event. There have been many business model innovations every year. It also requires that DiDi constantly cultivates corresponding dynamic capabilities, which in turn promote and guarantee the success of business model innovation, so that enterprises can obtain sustainable competitive advantages.

5. Conclusion

With the rapid development of Internet technology, the number of platform enterprises increased significantly, and many unicorn enterprises have been developed. Platform enterprises are quite different from traditional manufacturing industries in terms of production cost structure, trading pattern, profit mode, and consumer consumption habits. Under the dynamic environment, the company must make strategic adjustment which means that the company should undertake business model innovation. Most previous research

about dynamic capabilities mainly focused on content and composition and ignored the relationship between dynamic capabilities and business model innovation.

Based on the previous literature, this paper shows that the elements of business model innovation of platform enterprise are value proposition innovation, product innovation, partnership innovation, and profit model innovation. The dimensions of dynamic capabilities of platform enterprises are market perception capability, learning and integration capability, coordination capability, and organizational flexibility. Dynamic capabilities promote business model innovation which has different guiding effects on the cultivation of dynamic capabilities. An exploratory case study was conducted, using DiDi taxi as an example, and verified the theory model.

The main theoretical contributions of this paper can be summarized in the following three points.

Firstly, this paper initially introduces business model innovation theory into the cultivation mechanism of dynamic capability. Previous literature which develops dynamic capability theory often consists of two theoretical aspects: resource-based view and capability theory, which are greatly improving and enriching dynamic capability theory. Unlike these two aspects, this paper explores and develops the relational mechanism effect of business model innovation on dynamic capability and develops the dynamic cycle between the business model innovation and the cultivation of dynamic capabilities. Secondly, this paper centers on the new era of digital platform as the specific research context which is full of dynamic and turbulent environment. This new form of economy is largely different from traditional economy both in business model innovation and in cultivation of dynamic capabilities. Thirdly, this paper employs DiDi taxi as an exploratory case study, which represents a wholly new way to develop business model and dynamic capabilities under the complex and turbulent environment. Consequently, this paper qualitatively elaborates the proposed theoretical model of the relationship between business model innovation and dynamic capabilities, which expands the aspects of investigating the relationship between the two.

In addition, from practical aspect, this paper guides enterprises, especially platform enterprises, toward combining business model innovation with dynamic capabilities and taking various measures to ensure the success of business model innovation.

Data Availability

This research strived to use a variety of sources to collect data, mainly including the following channels: (1) enterprise archives, such as enterprise journals, work summaries, internal rules and regulations, and enterprise annual reports; (2) official website of DiDi; (3) media reports; (4) publications of third-party research institutions. Data sources and codes are shown in Table 2 in the article. The data used to support the findings of this study are available from the corresponding authors upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Social Science Foundation of Fujian (Grant nos. FJ2018B062 and FJ2019B101) and Fujian Provincial Department of Science and Technology, Soft Science Research Plan Project (Grant nos. 2019R0093 and 2019R0094).

References

[1] J. Magretta, "Why business models matter," *Harvard Business Review*, vol. 80, no. 5, pp. 86–92, 2002.

[2] M. A. Rappa, "The utility business model and the future of computing services," *IBM Systems Journal*, vol. 43, no. 1, pp. 32–42, 2004.

[3] A. Osterwalder, Y. Pigneur, and C. L. Tucci, "Clarifying business models: origins, present, and future of the concept," *Communications of the Association for Information Systems*, vol. 16, no. 1, pp. 1–25, 2005.

[4] C. Zott, R. Amit, and L. Massa, "The business model: recent developments and future research," *Journal of Management*, vol. 37, no. 4, pp. 1019–1042, 2011.

[5] M. Morris, M. Schindehutte, and J. Allen, "The entrepreneur's business model: toward a unified perspective," *Journal of Business Research*, vol. 58, no. 6, pp. 726–735, 2005.

[6] D. J. Teece, G. Pisano, and A. Shuen, "Dynamic capabilities and strategic management," *Strategic Management Journal*, vol. 18, no. 7, pp. 509–533, 1997.

[7] D. J. Teece, *Managing Intellectual Capital: Organizational, Strategic, and Policy Dimensions*, Oxford University Press, Oxford, U.K, 2002.

[8] M. Zollo and S. G. Winter, "Deliberate learning and the evolution of dynamic capabilities," *Organization Science*, vol. 13, no. 3, pp. 339–351, 2002.

[9] C. L. Wang and P. K. Ahmed, "Dynamic capabilities: a review and research agenda," *International Journal of Management Reviews*, vol. 9, no. 1, pp. 31–51, 2007.

[10] D. J. Teece, "Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance," *Strategic Management Journal*, vol. 28, no. 13, pp. 1319–1350, 2007.

[11] P. A. Pavlou and O. A. El Sawy, "Understanding the elusive black box of dynamic capabilities," *Decision Sciences*, vol. 42, no. 1, pp. 239–273, 2011.

[12] P. Lin, "Measurement and function of dynamic capabilities: empirical data from Chinese enterprises," *Journal of Central South University: Social Science Edition*, vol. 04, pp. 86–93, 2009.

[13] D. M. D. Moraes, "The internet of things - the value is the customer experience," *Appliance Design*, vol. 64, no. 7, pp. 25-26, 2016.

[14] J. T. Gourville, "Eager sellers and stony buyers: understanding the psychology of new-product adoption," *Harvard Business Review*, vol. 84, no. 6, pp. 98–145, 2006.

[15] A. Najmei, "Dynamic Business Model Innovation: An Analytical Archetype," in *Proceedings of the 2011 3rd International Conference on Information and Financial Engineering IPEDR*, pp. 165–171, Shanghai, China, August 2011.

[16] D. J. Teece, "Business models and dynamic capabilities," *Long Range Planning*, vol. 51, no. 1, pp. 40–49, 2018.

[17] K. S. R. Warner and M. Wäger, "Building dynamic capabilities for digital transformation: an ongoing process of strategic renewal," *Long Range Planning*, vol. 52, no. 3, pp. 326–349, 2018.

[18] K. M. Eisenhardt and M. E. Graebner, "Theory building from cases: opportunities and challenges," *Academy of Management Journal*, vol. 50, no. 1, pp. 25–32, 2007.

Hindawi
Scientific Programming
Volume 2020, Article ID 8862468, 7 pages
https://doi.org/10.1155/2020/8862468

Research Article

SubRF_Seq: Identification of Sub-Golgi Protein Types with Random Forest with Partial Sequence Information

Qingyu Cui,[1] Yi Cao,[1] Wenzheng Bao ,[2] Bin Yang,[3] and Yuehui Chen[1]

[1]*School of Information, University of Jinan, Jinan 250024, China*
[2]*School of Information Engineering, Xuzhou University of Technology, Xuzhou 221018, China*
[3]*School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China*

Correspondence should be addressed to Wenzheng Bao; baowz55555@126.com

Received 7 April 2020; Revised 13 May 2020; Accepted 16 June 2020; Published 16 July 2020

Academic Editor: Chenxi Huang

In the recent years, the subject of Golgi classification has been studied intensively. It has been scientifically proven that Golgi can synthesize many substances, such as polysaccharides, and it can also combine proteins with sugars or lipids with glycoproteins and lipoproteins. In some cells (such as liver cells), the Golgi apparatus is also involved in the synthesis and secretion of lipoproteins. Therefore, the loss of Golgi protein function may have severe effects on the human body. For example, Alzheimer's disease and diabetes are related to the loss of Golgi protein function. Because the classification of Golgi proteins has a specific effect on the treatment of these diseases, many scholars have studied the classification of Golgi proteins, but the data sets they used were complete Golgi sequences. The focus of this article is whether there is redundancy in the Golgi protein classification or, in other words, whether a part of the entire Golgi protein sequence can be used to complete the Golgi protein classification. Besides, we have adopted a new method to deal with the problem of sample imbalance. After experiments, our model has certain observability.

1. Introduction

Golgi is an organelle found in eukaryotic cells [1]. The Golgi was initially defined by Camilo-Golgi in 1897 and was named after Golgi in 1898 [2–4]. Considering its large size and unique structure, the Golgi apparatus can be treated as the first organelles which are discovered and observed in detail [5–7]. As part of the inner membrane system, Golgi proteins are encapsulated in membrane vesicles [8], which are sent to their destination. Golgi is located between the secretory pathway, the lysosome, and the endocytosis pathway [9]. Golgi plays an essential role in protein secretion. Meanwhile, such an issue contains a series of related glycosylases [10]. The subcellular position of the Golgi apparatus is different from that of various eukaryotic cells. In most eukaryotic cells, the Golgi apparatus includes cis-Golgi and trans-Golgi [11, 12]. Cis-Golgi is mainly composed of vesicles and multiple vesicles form the Golgi pile. Trans-Golgi is the final vesicle structure, where proteins are encapsulated in transport vesicles and sent to the lysosome, secretory pathway, or cell surface. The Golgi apparatus is closely related in the areas of structure and function [13, 14]. Each independent Golgi stack can contain several types of enzymes. These abovementioned enzymes can process several biological issues [15].

Disorders of protein metabolism are the core link leading to the development of many neurodegenerative diseases [16]. The Golgi apparatus is an essential organelle in the material metabolic pathway and must be closely related to it. Parkinson's disease [17] and Alzheimer's disease [18] are typical of neurodegenerative diseases. Experiments have shown that β-amyloid protein plays a central role in the pathological changes of Alzheimer's disease [19], and its metabolic disorder is closely related to the loss of a certain function of the Golgi apparatus. However, in order to understand the mechanism of Golgi function, an essential step is to find a Golgi-resident [20] and use the types and functions of the Golgi-resident protein to determine the principles of the disease. For example, the cause of the diseases is likely to be a lack of a Golgi-resident protein

[21, 22], resulting in a loss of Golgi function. Therefore, it is important to correctly judge the type of Golgi apparatus [23, 24].

With several years' effort, the prediction of the Golgi type has become one of the most significant hot subjects [25] in the field of computational biology and bioinformatics. Currently, simply knowing whether a protein is a Golgi-resident protein is not enough to fully explain the function of the Golgi body [26–28]. Further analysis of the specific type of Golgi-resident protein is needed. For now, some methods are applied to this subject. Ding et al. proposed the improved Mahalanobis Discriminant (MD) algorithm to predict Golgi-resident protein types in 2011 [29]. Dijk proposed the prediction of the Golgi-resident protein type of type II membrane proteins using structural information and transmembrane domain information in 2008 [30]. Jiao and Du proposed that the general form of Chou pseudoamino acids to predict the Golgi-resident protein type in 2016 [31]. Ding and Jiao used a relatively small data set with 150 Golgi proteins. Yang et al. created a new data set with 304 sub-Golgi proteins for training and 64 sub-Golgi proteins for testing classification models [21]. Ahmad and Hayat [32] proposed a Golgi protein classification model using multi-voting feature selection. Zhou [33] proposed XGBoost conditional covariance minimization based on multifeature fusion to predict Golgi protein types. Whether it is based on an amino acid feature extraction method or after multiple amino acid feature extractions and voting or multifeature fusion, they all use the complete amino acid sequence to extract features, and because they use the complete amino acid sequence to extract features, their models obtain considerable accuracy. However, we all know that the amino acid sequence of a Golgi is very long, and it will take a lot of effort to extract feature information on the entire amino acid sequence.

In this paper, we propose a new model, dubbed sub-RF_seq. In detail, if we do not use a complete protein sequence in feature extraction, some of them can also get considerable accuracy. Throughout this article, our work is summarized as follows: firstly, we propose 529 types of cutting sequences. The training set and test set are cut according to these 529 cutting types. Then, the 529 training sets are encoded. We use EAAC technology to extract features and put them into the RF classifier to train the model. Finally, we use the split to equal validation to balance the data set and test the classification effect of the Golgi apparatus. We use the random forest classifier to get the top 5 cutting sequence methods, and then put the features of these five cutting methods into other classifiers we have constructed and compare which classifier is the best classifier with the partial Golgi protein sequence.

Our workflow is as follows.

2. Methods and Materials

2.1. Data. This experiment uses a new data set created by Ahmad [12]. There are 87 cis-Golgi protein sequences and 217 trans-Golgi protein sequences in the training set. No protein has more than 40% pairing with any other protein in the data set. The 64 sub-Golgi protein sequences were independently used for testing the effect of the classifier, of which 13 were cis-Golgi protein sequences and 51 were trans-Golgi protein sequences. It should be noted that there is no connection between the training set and the test set.

Our work flow chart is shown in Figure 1. Specifically, we need to process the complete sub-Golgi protein sequence. In this step, the 304 sub-Golgi protein sequences in the training set are cut. The cutting method is to cut three positions in the front and three positions from the back to form a new protein sequence. This forms the first partial Golgi sequence. Then, the front three digits are unchanged, the back cleavage digit is increased by one, and it is added to the back cleavage 25 to form 23 new protein sequences and form 23 partial Golgi training sets. Then, the number of front-end cuts is increased by one, and the number of rear-end cuts is from 3–25, until the last front-end cut is 25 digits and the back-end cut is 25 digits. There are 23×23 different cutting methods. 23×23 incomplete Golgi protein sequences were formed. The test set adapts the same cutting methods. Then, use EAAC to extract protein sequence features, input to the classification model to train the model, and then test the effect on an independent test set.

2.2. Feature Extraction

2.2.1. Amino Acid Composition Encoding. The sequence information of the Golgi apparatus contains the types and arrangement order of 20 amino acids [34, 35]. Therefore, the feature extraction algorithm based on the amino acid composition is the simplest and most intuitive method. The amino acid composition simply represents the probability of 20 kinds of amino acids appearing in the sequence [36, 37]. It is a basic Golgi sequence feature extraction algorithm. The amino acid composition maps the Golgi sequence to a point in the 20-dimensional European space. The vector is expressed as follows:

$$V_{\text{aac}} = (v_1, v_2, v_3, \ldots, v_{20})T,$$

$$V_i = \frac{f\_i}{\sum_{j=1}^{20} f\_i}, \tag{1}$$

$$\sum_{j=1}^{20} v_j = 1.$$

Here, f_i is the number of times the ith amino acid appears in the sequence ($i = 1, 2, 3, \ldots, 20$). The amino acid composition is easy to calculate, and it is the most commonly used sequence feature extraction algorithm in Golgi classification research.

2.2.2. Enhanced Amino Acid Content Encoding (EAAC). Chen et al. [38] proposed a new encoding method based on AAC encoding, dubbed EAAC. EAAC coding directly reflects the distribution frequency of 20 amino acid residues. EAAC coding differs from AAC coding in that EAAC coding defines a sliding window of length 8 and calculates the

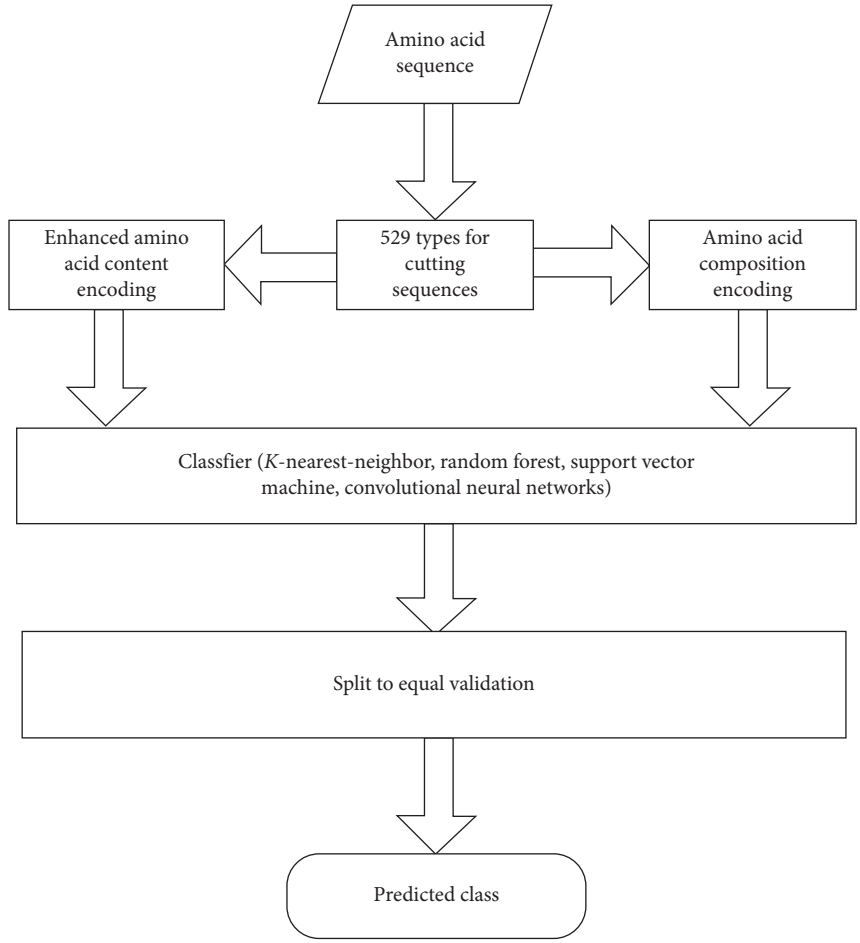


Figure 1: Illustration of the process of the proposed method.

frequency of 20 amino acid residues that appear in each 8-dimensional subsequence segment [39]. The frequency of 20 amino acid residues is continuously sliding in the window from the N-terminus to the C-terminus of each Golgi sequence in the dataset. Therefore, the vector dimension corresponding to a Golgi sequence of x residues is

$$\begin{aligned} L_s &= x - L_v + 1, \\ D_s &= L_s \times 20. \end{aligned} \tag{2}$$

Here, L_v is the size of the sliding window we defined. In EAAC encoding, the value of L_v is 8, x is the length of the Golgi sequence, and the D_s is the dimension of the feature vector.

2.3. Construction of the Classifier. This experiment mainly uses a classifier of random forests. Random forests are called "representative methods for ensemble learning" [40], which is easy to implement and has relatively low overhead. Random forests are an extension of Bagging's idea [41], which is based on decision tree learning, and the algorithm further introduces random attribute selection in the training process of the decision tree [42–44]. The basic idea of random forest is to train the model with data, then get multiple decision trees, and then merge the decision trees to get more stable predictions. In random forests, the performance becomes better as the number of trees increases, and the error becomes smaller. In this experiment, we selected 1000 decision trees to build a random forest model. In addition, we also constructed KNN (K nearest neighbor classification algorithm), SVM (Support Vector Machine Algorithm), CNN (Convolutional Neural Network), and ANN (Artificial Neural Network) classifiers to compare which is in the best classifiers with the part of Golgi protein sequences.

2.4. Evaluation Methods. The positive and negative samples of the training set of this experiment are imbalanced, and the ratio of positive and negative samples is about 1 : 2. In the binary classification problem, the imbalance of positive and negative samples will have a certain impact on the classification effect. It will cause the prediction category towards the category with many samples. Therefore, for the evaluation method, we chose an SE verification method proposed by Sun et al. [45]. The advantage of this verification method is that data processing and cross-validation can be implemented at the same time.

Performance measurement is an evaluation standard for measuring the generalization ability of the model, which

reflects the needs of the task. The use of different performance metrics often leads to different evaluation results. Therefore, it is essential to choose a good set of performance indicators to predict the performance of the model. In this experiment, ACC and AUC were selected for evaluation. ACC and AUC performance indicators have evolved from the confusion matrix [46–51]. In the binary classification problem, when the real situation of data classification in the test set is a positive example, the model prediction result is a positive example, which is called the real example (TP). When the predicted outcome is a counterexample, it is a false counterexample (FN). Similarly, when the true situation of the data classification of the test set is a counterexample, there are false positive examples (FP) and true counterexamples (TN). The accuracy rate formula is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

The recall formula is

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

The formula for the accuracy rate (ACC) is

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{5}$$

The value of AUC is the area of the ROC curve. We often use the value of AUC as the criterion for judging the quality of the model because the ROC curve cannot intuitively see the quality of the model [52, 53]. ROC is a curve drawn with sensitivity as the vertical axis and 1 minus specificity as the horizontal axis.

The formula for sensitivity is

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

3. Results and Discussion

In this section, we mainly describe the effect of the 529 incomplete Golgi sequences we have defined for training the model. Besides, we chose the top 5 cutting methods for classification effects in the sub_RF_seq model for comparison experiments.

3.1. Results. In this experiment, we recorded the AUC values of 529 different cutting methods. In order to intuitively understand the classification effect of these 529 cutting methods, we made a three-dimensional histogram based on the AUC values. The X-axis represents how many bits are cut from the front end of the protein sequence, and the Y-axis represents the number of bits cut from the rear end of the protein. The X-axis represents how many bits are cut from the front end of the protein sequence, and the Y-axis represents the number of bits cut from the back end of the protein. In this way, this three-dimensional histogram shows the classification effect using incomplete Golgi protein sequences. Figure 2 shows that, among these 529 Golgi sequence cutting methods, 202 of the cutting methods have an AUC value greater than or equal to 0.6, and 426 of the cutting methods have a value greater than or equal to 0.5.

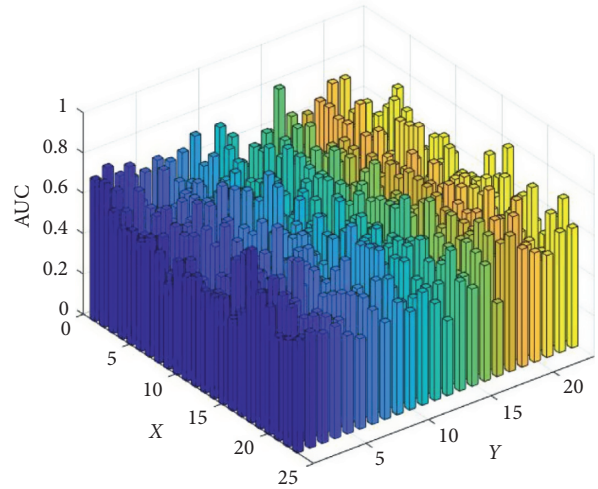


Figure 2: Three-digit histogram of an AUC value of 529 cutting methods.

In addition, we used the random forest classifier to select the top 5 cutting methods for Golgi classification. The values of AUC and ACC for these five cutting methods are shown in Table 1.

3.2. Discussion

3.2.1. Comparison of Model Effects under Different Classifiers. We put the cutting sequence of the top 5 classification effects in the model into the SVM, KNN, CNN, and ANN classifiers and compared which classifiers used the partial Golgi sequence to achieve the best Golgi classification. From Table 2, we found that the RF classifier performs better than several other classifiers. For example, under the premise of a certain cutting sequence method, EAAC coding is selected for the feature coding method. In the 20 + 3 Golgi sequence, the value of ACC in the RF classifier is as high as 82.81%, and the value of AUC is as high as 0.854. The values are better than several other classifiers. However, the classification effect of partial Golgi sequences in other classifications is still considerable. In Table 2, the AUC and Acc values of most classifiers are above 70%, which further to confirm that there is a certain redundancy in the Golgi sequence when it is used to determine the Golgi types.

3.2.2. Classification Effect under Different Encoding Methods. In this experiment, we chose two encoding methods, EAAC and AAC, to see the effect of different amino acid sequence encoding methods of the classification effect. In order to explore the classification effect under different encoding methods, we controlled the variable classifier. Only the RF classifier is selected. From Table 3, we can see the AUC and ACC values of the five cutting methods under the EAAC and AAC. Table 3 shows that, in the EAAC encoding mode, the values of Acc and AUC are higher than those in the AAC

Table 1: Sub-Golgi protein sequence cutting methods ranked by the AUC value.

Classifier	Encoding schemes	Cutting method	AUC	ACC
RF	EAAC	20 + 3	0.854449	0.828125
		4 + 17	0.849170	0.859375
		18 + 25	0.782805	0.828125
		20 + 11	0.782805	0.796875
		11 + 11	0.773002	0.828125

Table 2: Comparison of the effects of different classifiers.

Cutting	RF		SVM		CNN		KNN		ANN	
	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
20 + 3	82.81	0.8544	67.19	0.6448	70.31	0.7179	62.5	0.549	78.13	0.4042
4 + 17	85.94	0.8492	76.56	0.7858	76.56	0.7451	67.19	0.6531	78.13	0.6456
18 + 25	82.81	0.7828	65.63	0.6365	78.13	0.7602	60.94	0.7451	68.75	0.4781
20 + 11	79.69	0.7828	68.75	0.7873	81.25	0.736	62.5	0.7572	60.03	0.7188
11 + 11	82.81	0.773	70.31	0.7451	73.44	0.7315	56.25	0.5038	68.75	0.4962

Table3: Classification effect under different encoding methods.

Cutting	Classifier	EAAC encoding		AAC encoding	
		Acc (%)	AUC	Acc (%)	AUC
20 + 3	RF	82.81	0.8544	65.63	0.4434
4 + 17		85.94	0.8492	70.31	0.4894
18 + 25		82.81	0.7828	71.88	0.7549
20 + 11		79.69	0.7828	68.75	0.586
11 + 11		82.81	0.7730	64.06	0.4563

Table 4: Performance of imbalance of positive and negative samples of the data set on the classification effect.

Cutting	Classifier	Encoding	SEV		10-flod CV	
			Acc (%)	AUC	Acc (%)	AUC
20 + 3	RF	EAAC	82.81	0.8544	78.13	0.7813
4 + 17			85.94	0.8492	67.95	0.6161
18 + 25			82.81	0.7828	79.69	0.4615
20 + 11			79.69	0.7828	77.75	0.7681
11 + 11			82.81	0.7730	80.06	0.5716

encoding method, which directly proves our guess that different encoding methods will affect the classification effect of the model.

3.2.3. Performance of Imbalance of Positive and Negative Samples of the Data Set on the Classification Effect. Due to the imbalance of the positive and negative samples in the data set we used, we used both the SEV verification method and the 10-fold cross-validation method to verify the classification effect of the model. The SEV verification method can deal with the imbalance of the positive and negative samples of the data set, and the 10-fold cross-validation does not have the effect of data preprocessing. Table 4 proves that processing the imbalance of the data set will improve the model's effectiveness. Using SEV is nearly 18% higher than a simple 10-fold cross-validation.

4. Conclusions

In the past, when determining the type of Golgi apparatus, many people used the entire Golgi protein sequence in encoding; a complete Golgi protein sequence has a large number of amino acids, which is very time-consuming when encoding. In this article, we present subRF_seq, which can complete the classification of Golgi using a part of the Golgi protein sequence and has a considerable classification effect. We cut the data set, extract the feature vector from the cut sequence, and finally, train it in a random forest to distinguish trans-Golgi and cis-Golgi. Also, in the binary classification problem, the proportion of positive and negative samples of many training sets cannot reach 1 : 1, which will cause the problem of falsely high AUC values. Our model can effectively overcome this problem. We also used other classifiers and feature extraction techniques to prove our ideas, and the results show that our ideas of using part of the Golgi sequence in feature extraction is feasible because the values of AUC and ACC are considerable in different classifiers and encoding methods. The experimental results prove that Golgi proteins can still be distinguished by using partial Golgi sequences, In other words, there is a certain degree of redundancy in Golgi protein classification on Golgi classification. If we use part of the Golgi sequence in Golgi classification, it will significantly reduce the time.

Data Availability

To data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the grants from the National Science Foundation of China (nos. 61902337 and 61702445)

Jiangsu Provincial Natural Science Foundation, China (no. SBK2019040953), and Natural Science Fund for Colleges and Universities in Jiangsu Province no. 19KJB520016.

References

[1] G. Griffiths and K. Simons, "The trans Golgi network: sorting at the exit site of the Golgi complex," *Science*, vol. 234, no. 4775, pp. 438–443, 1986.

[2] M. Gribskov, I. Mellman, and K. Simons, "The golgi complex: in vitro veritas?," *Cell*, vol. 68, no. 5, pp. 829–840, 1991.

[3] Q. Zou and Q. Liu, "Advanced machine learning techniques for bioinformatics," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 16, no. 4, pp. 1182-1183, 2019.

[4] I. Mishqat, "Camillo Golgi's black reaction for staining neurons," Embryo Project Encyclopedia, Washington, DC, USA, 2017.

[5] S. Bassnett, "The fate of the Golgi apparatus and the endoplasmic reticulum during lens fiber cell differentiation," *Investigative Ophthalmology & Visual Science*, vol. 36, no. 9, pp. 1793–1803, 1995.

[6] J. Rothman, "The Golgi apparatus: two organelles in tandem," *Science*, vol. 213, no. 4513, p. 1212, 1981.

[7] W. He, L. Wei, and Q. Zou, "Research progress in protein posttranslational modification site prediction," *Briefings in Functional Genomics*, vol. 18, no. 4, pp. 220–229, 2019.

[8] M. Rao and C. R. Alving, "Delivery of lipids and liposomal proteins to the cytoplasm and Golgi of antigen-presenting cells," *Advanced Drug Delivery Reviews*, vol. 41, no. 2, pp. 171–188, 2000.

[9] Z. Wang, H. Ding, and Q. Zou, "Identifying cell types to interpret scRNA-seq data: how, why and more possibilities," *Briefings in Functional Genomics*, 2020.

[10] L. Yuan, F. Guo, L. Wang, and Q. Zou, "Prediction of tumor metastasis from sequencing data in the era of genome sequencing," *Briefings in Functional Genomics*, vol. 18, no. 6, pp. 412–418, 2019.

[11] B. H. Hummer, D. Maslar, M. S. Gutierrez, N. F. D. Leeuw, and C. S. Asensio, "Differential sorting behavior for soluble and transmembrane cargoes at the trans-Golgi network in endocrine cells," *Molecular Biology of the Cell*, vol. 31, no. 3, pp. 157–166, 2020.

[12] J. Ahmad, F. Javed, and M. Hayat, "Intelligent computational model for classification of sub-Golgi protein using oversampling and Fisher feature selection methods," *Artificial Intelligence in Medicine*, vol. 78, pp. 14–22, 2017.

[13] S. Deng, H. Liu, K. Qiu, H. You, Q. Lei, and W. Lu, "Role of the Golgi apparatus in the blood-brain barrier: golgi protection may be a targeted therapy for neurological diseases," *Molecular Neurobiology*, vol. 55, no. 6, pp. 4788–4801, 2018.

[14] J. Villeneuve, J. Duran, M. Scarpa, L. Bassaganyas, J. V. Galen, and V. Malhotra, "Golgi enzymes do not cycle through the endoplasmic reticulum during protein secretion or mitosis," *Molecular Biology of the Cell*, vol. 28, no. 1, pp. 141–151, 2017.

[15] Y. Hou, J. Dai, J. He, A. J. Niemi, X. Peng, and N. Ilieva, "Intrinsic protein geometry with application to non-proline cis peptide planes," *Journal of Mathematical Chemistry*, vol. 57, no. 1, pp. 263–279, 2019.

[16] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Transactions on Nanobioscience*, vol. 16, no. 4, pp. 240–247, 2017.

[17] J. M. V. D. Elsen, D. A. Kuntz, and D. R. Rose, "Structure of Golgi α-mannosidase II: a target for inhibition of growth and metastasis of cancer cells," *The EMBO Journal*, vol. 20, no. 12, pp. 3008–3017, 2001.

[18] S. Hoyer, "Is sporadic Alzheimer disease the brain type of non-insulin dependent diabetes mellitus? A challenging hypothesis," *Journal of Neural Transmission*, vol. 105, no. 4, pp. 415–422, 1998.

[19] D. D. Elsberry and M. T. Rise, "Method of treating movement disorders by brain infusion," U.S. Patent No. 6,042,579, U.S. Patent and Trademark Office. Washington, DC, USA, 2000.

[20] B. Radau, A. Otto, E.-C. Müller, and P. Westermann, "Protein kinase Cα-dependent phosphorylation of Golgi proteins," *Electrophoresis*, vol. 21, no. 13, pp. 2684–2687, 2000.

[21] R. Yang, C. Zhang, R. Gao, and L. Zhang, "A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data," *International Journal of Molecular Sciences*, vol. 17, no. 2, p. 218, 2016.

[22] A. E. Cuadra, S.-H. Kuo, Y. Kawasaki, D. S. Bredt, and D. M. Chetkovich, "AMPA receptor synaptic targeting regulated by stargazin interactions with the Golgi-resident PDZ protein nPIST," *Journal of Neuroscience*, vol. 24, no. 34, pp. 7491–7502, 2004.

[23] I. J. Goldstein, C. E. Hollerman, and E. E. Smith, "Protein-carbohydrate interaction. II. Inhibition studies on the interaction of concanavalin A with polysaccharides," *Biochemistry*, vol. 4, no. 5, pp. 876–883, 1965.

[24] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[25] Z. Yuan and R. D. Teasdale, "Prediction of Golgi type II membrane proteins based on their transmembrane domains," *Bioinformatics*, vol. 18, no. 8, pp. 1109–1115, 2002.

[26] P. Cosson, M. Amherdt, J. E. Rothman, and L. Orci, "A resident Golgi protein is excluded from peri-Golgi vesicles in NRK cells," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12831–12834, 2002.

[27] Y. Jiao, P. Du, and X. Su, "Predicting Golgi-resident proteins in plants by incorporating *N*-terminal transmembrane domain information in the general form of Chou's pseudoamino acid compositions," in *Proceedings of the 2014 8th International Conference on Systems Biology (ISB)*, pp. 226–229, Qingdao, China, 2014, October.

[28] C. Y. L. Yuen, P. Wang, B.-H. Kang, K. Matsumoto, and D. A. Christopher, "A non-classical member of the protein disulfide isomerase family, PDI7 of *Arabidopsis thaliana*, localizes to the cis-Golgi and endoplasmic reticulum membranes," *Plant and Cell Physiology*, vol. 58, no. 6, pp. 1103–1117, 2017.

[29] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, "Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition," *Protein and Peptide Letters*, vol. 18, no. 1, pp. 58–63, 2011.

[30] A. D. V. Dijk, D. Bosch, C. J. T. Braak, A. R. V. D. Krol, and R. C. H. J. V. Ham, "Predicting sub-Golgi localization of type II membrane proteins," *Bioinformatics*, vol. 24, no. 16, pp. 1779–1786, 2008.

[31] Y.-S. Jiao and P.-F. Du, "Predicting Golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties," *Journal of Theoretical Biology*, vol. 391, pp. 35–42, 2016.

[32] J. Ahmad and M. Hayat, "MFSC: multi-voting based feature selection for classification of Golgi proteins by adopting the

general form of Chou's PseAAC components," *Journal of Theoretical Biology*, vol. 463, pp. 99–109, 2019.

[33] H. Zhou, C. Chen, M. Wang, Q. Ma, and B. Yu, "Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion," *IEEE Access*, vol. 7, pp. 144154–144164, 2019.

[34] A. Raina and A. Datta, "Molecular cloning of a gene encoding a seed-specific protein with nutritionally balanced amino acid composition from Amaranthus," *Proceedings of the National Academy of Sciences*, vol. 89, no. 24, pp. 11774–11778, 1992.

[35] J. Adachi, P. J. Waddell, W. Martin, and M. Hasegawa, "Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA," *Journal of Molecular Evolution*, vol. 50, no. 4, pp. 348–358, 2000.

[36] V. J. Vivekanand and J. Ramana, "Prediction of lysosomal membrane proteins using machine learning techniques," 2014.

[37] S.-Y. Kung and M.-W. Mak, "Feature selection for self-supervised classification with applications to microarray and sequence data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 297–309, 2008.

[38] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites," *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 6, pp. 451–459, 2018.

[39] H. Neumann, K. Wang, L. Davis, M. Garcia-Alai, and J. W. Chin, "Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome," *Nature*, vol. 464, no. 7287, pp. 441–444, 2010.

[40] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, no. 2012, pp. 1063–1095, 2012.

[42] A. C. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning*, pp. 157–175, Springer, Boston, MA, USA, 2012.

[43] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line random forests," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1393–1400, Kyoto, Japan, 2009, September.

[44] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[45] J. Sun, Y. Cao, D. Wang, W. Bao, and Y. Chen, "K_net: lysine malonylation sites identification with neural network," *IEEE Access*, vol. 8, pp. 47304–47311, 2019.

[46] M. Tahir and M. Hayat, "iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC," *Molecular BioSystems*, vol. 12, no. 8, pp. 2587–2593, 2016.

[47] Y. Jiao and P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, 2016.

[48] I. S. MacKenzie, T. Kauppinen, and M. Silfverberg, "Accuracy measures for evaluating computer pointing devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 9–16, Seattle, WA, USA, 2001, March.

[49] L. J. Siegel, H. J. Siegel, and P. H. Swain, "Performance measures for evaluating algorithms for SIMD machines," *IEEE Transactions on Software Engineering*, vol. 8, no. 4, pp. 319–331, 1982.

[50] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 14, no. 2, pp. 1137–1145, Montreal, Canada, 1995, August.

[51] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[52] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[53] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, Pittsburgh, PA, USA, 2006, June.

Hindawi
Scientific Programming
Volume 2020, Article ID 6385281, 7 pages
https://doi.org/10.1155/2020/6385281

Research Article

Identifying Ethnics of People through Face Recognition: A Deep CNN Approach

Ahmed Jawad A. AlBdairi,[1,2] **Zhu Xiao**,[1] **and Mohammed Alghaili**[1]

[1]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*
[2]*Computer Center, University of Babylon, Hillah, Babil, Iraq*

Correspondence should be addressed to Ahmed Jawad A. AlBdairi; ahmed_albdairi@hnu.edu.cn and Zhu Xiao; zhxiao@hnu.edu.cn

Received 11 February 2020; Revised 29 April 2020; Accepted 15 June 2020; Published 14 July 2020

Academic Editor: Chenxi Huang

The interest in face recognition studies has grown rapidly in the last decade. One of the most important problems in face recognition is the identification of ethnics of people. In this study, a new deep learning convolutional neural network is designed to create a new model that can recognize the ethnics of people through their facial features. The new dataset for ethnics of people consists of 3141 images collected from three different nationalities. To the best of our knowledge, this is the first image dataset collected for the ethnics of people and that dataset will be available for the research community. The new model was compared with two state-of-the-art models, VGG and Inception V3, and the validation accuracy was calculated for each convolutional neural network. The generated models have been tested through several images of people, and the results show that the best performance was achieved by our model with a verification accuracy of 96.9%.

1. Introduction

The scope of face recognition field has been increased recently. Face recognition refers to the ability of identifying any person from an image or a video frame. Many techniques have been used in face recognition. One of the first techniques used is using a 2D pattern recognition problem in which a distance between the important points in an image is used to recognize the face [1], like calculating the distance between eyes and distance between other important points.

Another technique is called holistic matching technique in which complete face region is taken into account as an input data into the catch face system. The most important studies that used this technique are eigenfaces [2], principal component analysis, and linear discriminant analysis [3].

Feature-based structural technique is another technique used in face recognition where the local features of the face are extracted first and their locations and local statistics are fed into a structural classifier.

The holistic and feature extraction techniques are used together to make a new technique called hybrid technique in which 3D images are used. The person's face image is caught in 3D; the system after that will note the important features such as curves or shapes in the face. The system after that detects the image whether it is a photograph or real time, determines the location of the face, and measures the curves and shapes of the important features in that face, converting the face into a numerical representation and matching this numerical representation with a dataset of faces.

The most important technique in face recognition that has been emerged recently is using the convolutional neural network (CNN) [4]. Although a lot of studies used CNN in face recognition, none of these studies has proposed a robust model to identify ethnics of people through their faces with high classification accuracy for people who have some similarities with different ethics.

Motivated by this, we propose two new models for face recognition with regularization and without regularization, in which they have the ability to recognize the ethnics and origins of people through their faces' facial. To specify, the primary contribution of this paper is proposing a face recognition model that can detect the detailed features of the

faces and differentiate between them using RGB images or a real-time face recognition. The ethics of different people can be recognized using this model through extracting the most detailed features of the peoples' faces. A new dataset has been collected for that purpose with high resolution from three different regions in Asia. These images were collected from social media like Facebook and VK (Russian social media website). Finally, we achieved a promising performance on another dataset collected for the test purpose.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 shows the designed network for face recognition. The experiments and results of the new models are given in Section 4. Section 5 concludes the paper.

2. Related Work

A face recognition method has been presented based on dense grid histograms of oriented gradients (HOG) [5]. In that study, the face image has been divided into many dense grids from which the HOG features have been extracted. After that, all these HOG feature grids vectors are composed to realize the feature expression of the whole face, and the k-nearest neighbor classifier is used for recognition. The authors used face dataset in the training stage with complex changes in illumination, time and environment, to test the gamma illumination correction, the spatial gradient direction, the size of the block, the standardization, and the face image resolution to find and analyze the optimal HOG parameters for face recognition. The FERET database is a dataset used for facial recognition system evaluation.

There are many methods in face recognition with high recognition accuracy, which are based on deep learning. One of these methods have a good effect in a restricted environment as well as in the natural environment [6]. The authors improved the method of multipatches by using 4 areas' patches in the face. In order to have a higher performance, they also used a Joint Bayesian (JB) measure in face verification. The model has been trained by the set of CASIA WebFace and tested in the Labeled Faces in the Wild (LFW).

Learning for face recognition has been proposed in another study [7]. The authors argued that the DeepID can be effectively learned through challenging multiclass face identification tasks. Furthermore, the generalization capability of DeepID increases as more face classes are to be predicted at training. They have used about 10,000 face identifications in the training set. The generated model achieved 79.45% verification accuracy on LFW dataset. The deep ConvNet contains 4 convolutional layers with Maxpooling to extract features hierarchically followed by the fully connected DeepID layer and the softmax output layer indicating identity classes.

The developing of an effective feature representations for reducing intrapersonal variations while enlarging interpersonal differences in face recognition has been solved in another study [8] using the deep learning and using both face identification and verification signals as supervision. The Deep IDentification-verification features (DeepID2) are learned by a deep convolutional network. The face identification task increases the interpersonal variations by drawing DeepID2 features extracted from different identities apart, and the face verification task reduces the intrapersonal variations by pulling DeepID2 features extracted from the same identity. The face verification accuracy that has been achieved by testing the method on LFW dataset [9] was 99.15% and this accuracy is different from the validation accuracy. The error rate has been significantly reduced by 67% as compared to the best previously deep learning results [7].

Another approach for face recognition was presented in which the convolutional neural network (CNN) and a logistic regression classifier (LRC) are combined [4]. The CNN used to extract the features in order to detect and recognize the face images and LRC [10, 11] is used to classify the features learned by the convolutional neural network. The structure of the CNN used in this study is composed of four layers: input layer, two convolutional layers, and one subsampling layer. The first layer is considered as 64×64; therefore, the dataset was resized to that size to be compatible with the proposed structure and the output layer is a fully connected layer with 15 feature maps with the size of 1×1.

In ours study, we build two models, with dropout and without dropout layers to find out the effect of this layer in the training. This study is concerned with the recognition of the ethnics of people through their facial features through these two models. We used a new CNN with regularization like dropout layers and without regularization to find out the most accurate performance. During training, we used Adam optimizer [12] with a learning rate of 0.001 and categorical cross-entropy loss function. The generated models can detect the detailed features of the faces from RGB images or through a camera.

3. Convolutional Deep Learning for Face Recognition

3.1. Ethnics Identification Using Deep Learning. Our deep learning layers consist of twelve layers. Four of these layers are Conv layers, each followed by the Maxpooling layer, and some of these Conv layers are also followed by the dropout layer after the Maxpooling layer to extract the facial features. A drop connect layer is placed after the four Conv layers as a separator between them and the two fully connected layers. The output of the drop connect layers is passed to a flatten layer to flatten the output before they pass to the first fully connected layer. Between the two fully connected layers, another dropout layer is used. The softmax output layer is used to identify the classes. The purpose of using dropout layers is to get rid of the overfitting during the training. Figure 1 shows the whole structure of the network layers that predict n classes (e.g., n is 3). The number of predicted classes n can be extended to contain as many nationalities as possible.

The input to this network is an image of $128 \times 128 \times 3$ size (e.g., 3 feature maps). The patch size is 3×3 with the same padding in every Conv layer and stride is 1 which make the output of the Conv layer roughly the same size as the input. The output of each Conv layer is passed to Max-

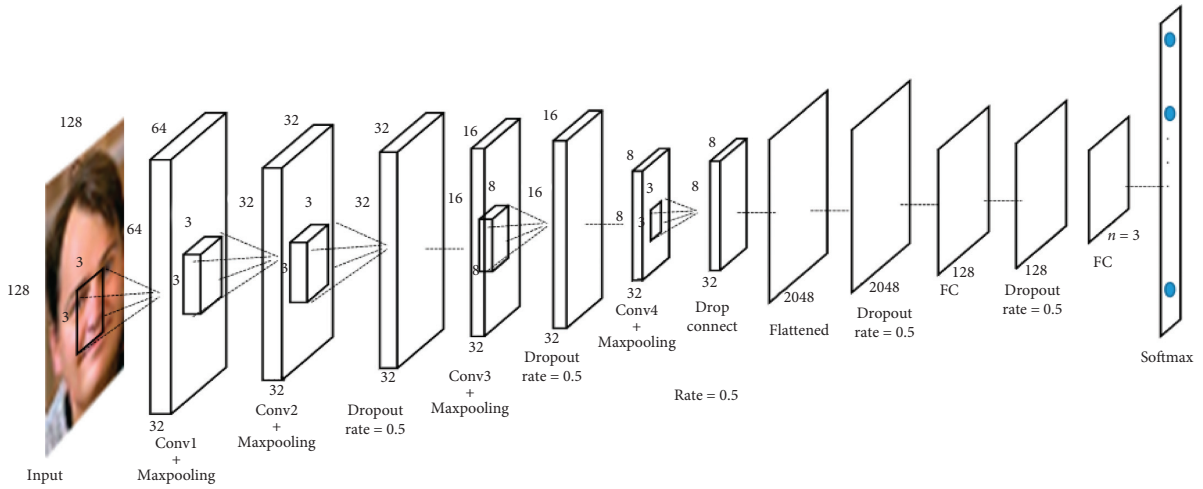


Figure 1: . The ConvNet layers. The small cuboid inside a square denotes the window map size of each Conv layer.

Pooling layer to minimize the input size. After that, the output of each Maxpooling layer is fed to ReLU activation function. The Conv layer with feature map equation is

$$f(x)^{j(r)} = \max\left(0, b^{j(r)} \sum_{n=1}^{\infty} k^{ij(r)} * x^{i(r)}\right), \tag{1}$$

where $f(x)^{j(r)}$ is the j^{th} output patch of the convolutional layer in a particular region r and $x^{i(r)}$ is the i^{th} input patch in a particular region r to the convolutional layer. The input of the first convolutional layer is an image of the size 128×128 divided into regions according to the size of window patch which is 3×3, as it is shown in Figure 1. $b^{j(r)}$ is the bias of the j^{th} output patch in the same particular region r. $k^{ij(r)}$ is the convolution kernel between the i^{th} input patch and the j^{th} output patch, whereas the multiplication of $k^{ij(r)}$ and $x^{i(r)}$ denotes the convolution.

The output of each convolutional layer is passed to the Maxpooling. The formula of the Maxpooling layer is as follows:

$$f(x)^{i}_{jk} = \max_{0 \le m,n < sz} \left(x^{i}_{j \cdot sz+m, k \cdot sz+n}\right). \tag{2}$$

The neurons in i^{th} the output patch $f(x)^{i}$ pool over $sz \times sz$ local region in the i^{th} input patch x^{i}. The output of the Maxpooling layer in each Conv layer is passed to ReLU nonlinearity $f(x) = \max(0, x)$. The ReLU sets all negative values in the input x to zero and all other values are kept constant, and it shows better fitting abilities than the sigmoid function [13].

Some of the Conv output is passed to a dropout to prevent the overfitting in the network. The number of dropout layers used is three where two of them are used after the second and the third Conv layer, and the third one is used between the last two fully connected layers.

The last layers are the two fully connected layers with dropout layer between them. The equation can be represented as follows:

$$fc = \max\left(0, \sum_{i} x^{i-1} \cdot w^{i-1, j-1}\right) + \max\left(0, \sum_{i} \text{DOut}_{\text{rate}}\left(x^{i} \cdot w^{i,j}\right)\right), \tag{3}$$

where x^{i-1} and $w^{i-1,\,j-1}$ denote the neurons and the weights of the previous layer, respectively. The output of the first fully connected layer is passed to $\text{DOut}_{\text{rate}}$ where the rate is 0.5 and the output of $\text{DOut}_{\text{rate}}$ is passed to the last fully connected layer. x^{i} and $w^{i,j}$ denote the neurons and the weights of the first fully connected layer before passing them to the $\text{DOut}_{\text{rate}}$ layer.

The output of the ConvNet is n-way softmax to predict the ethic of the face among n different ethics. The softmax works as follows:

$$y_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp\left(x_j\right)}, \tag{4}$$

where x_i is a vector of the inputs to the output layer and it denotes the most important features used to recognize the face. The output of that vector is calculated in x_j where x is the index of the output in n, e.g., number of classes.

3.2. Dropout Layers in the Network. Sometimes in the testing phase, the results are not accurate due to the training error. Researchers argue that because of overfitting [14], strong regularization like dropout [15] is used to overcome this problem. The idea of dropout is to drop out some neurons in a neural network wherein neurons are chosen randomly with probability $q = 1 - p$. When the neuron is dropped out, that means its input and output connection will be ignored and that will allow each neuron to learn something useful on its own without relying too much on other neurons to correct its shortcomings [16, 17]. Figure 2 illustrates the idea of dropout.

The input and output of each patch are computed as follows before we apply dropout:

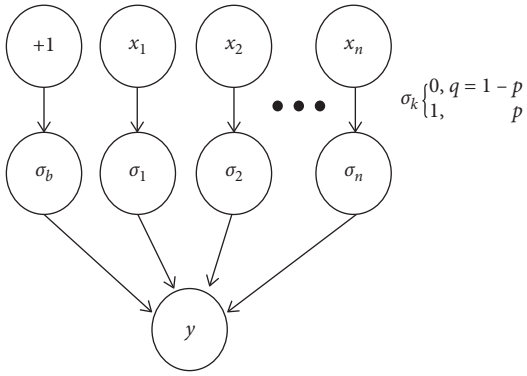


Figure 2: Neurons' training with dropout. The hidden neurons are randomly dropped out with Bernoulli's distribution p.

$$x^{l+1} = w^{l+1} y^l + b^{l+1}, \tag{5}$$

$$y^{l+1} = \mathrm{AF}\left(x^{l+1}\right), \tag{6}$$

where l denotes the index of the network layer. x^{l+1} is the input patch and y^{l+1} is the output patch at a hidden layer $l = 1, \ldots, l-2$, the layer being l. w^{l+1} is the weight and b^{l+1} is the bias. AF denotes the activation function. The following operations occur when the dropout is performed:

$$\sigma_i^l \approx \mathrm{Bernoulli}\,(p), \tag{7}$$

$$y'^l = \sigma^l \oplus y^l, \tag{8}$$

$$x^{l+1} = w^{l+1} y'^l + b^{l+1}, \tag{9}$$

$$y^{l+1} = \mathrm{AF}\left(x^{l+1}\right), \tag{10}$$

where $\oplus$ is the multiplication of an element by element and σ_i^l is a Bernoulli random variable of the i^{th} neuron at layer l with probability being 1.

3.3. Training Two Networks. The first network is the layers consisting of twelve layers including dropout layers. The training accuracy rate of this network is 96.9% and the validation accuracy rate is 96.9% with a validation loss of 0.221 which means the overfitting has been drastically eliminated as it is shown in Figure 3. In the second network, all the dropout layers are omitted, and the training accuracy is checked. The training accuracy rate in that network is 100%, the validation accuracy rate is 96.9%, and the least validation loss is 0.525. That means the overfitting is very high, and accordingly, the error rate of the created model from that network is more than that in the first network. Figures 3 and 4 show the training accuracy and validation accuracy for each network. The training accuracy in Figure 4 in epoch number 18 is 100% and that accuracy rate did not change until the end of the training which means the overfitting is very high and consequently the error rate is more than the error rate in the first network.

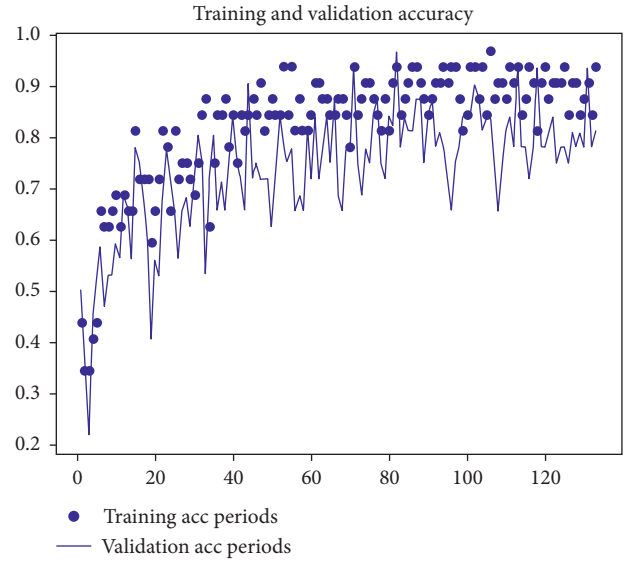


Figure 3: Training and validation accuracy for the first network.

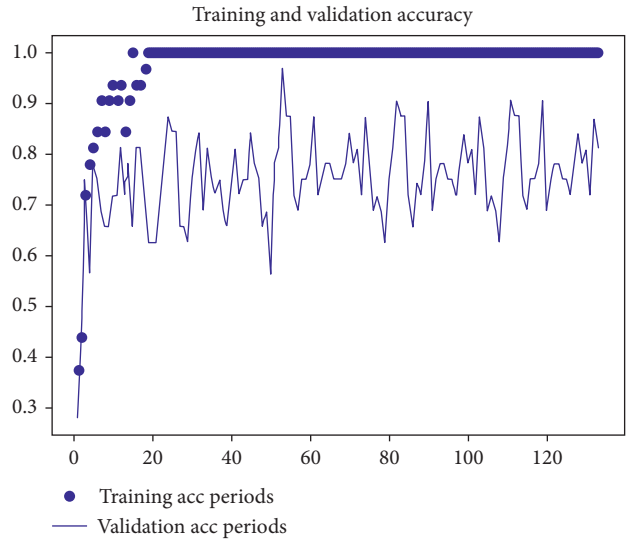


Figure 4: Training and validation accuracy for the second network.

4. Experiment

4.1. Experimental Training Dataset. Although there are many large-scale facial image databases available online, but all these databases are not appropriate to meet the objective of this study. Therefore, we manually collected 3141 photos from different resources. We collected 1081 Chinese facial images, 1021 Pakistani facial images, and 1039 Russian facial images. After collecting the images, they were processed to extract the faces from the whole images. The total images after that were divided into two sets; the first set was used for training stage and we took 70% of the whole images and the other 30% of the images as the second set for validation stage. Figure 5 shows a subset of the new dataset.

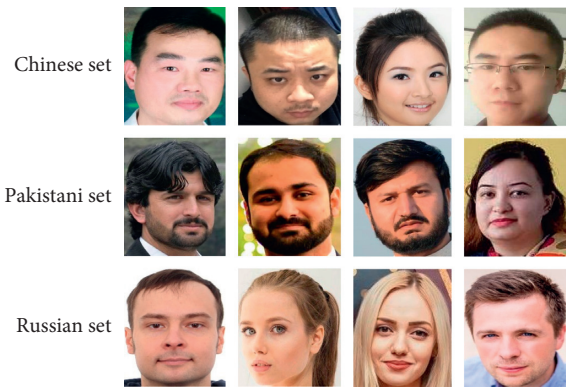


Figure 5: Three different subset images were collected from three different regions.

4.2. Comparison with the State-of-the-Art Approaches. Two state-of-the-art approaches were selected, and the last four layers for each approach have been frozen and used our fully connected layers to determine the number of output according to the number of classes in the dataset. These approaches are VGG [18] and Inception V3 [19]. The training was made in Tesla K80 GPU which is freely provided by Google Colaboratory. The results show that our approach has the highest validation accuracy and the least validation loss. Table 1 shows the results of training of our network and the two state-of-the-art approaches.

The comparison between our approach and the two state-of-the-art approaches VGG and Inception V3 is shown in Table 1 where it was observed that our approach has the highest validation accuracy (96.6%) with the less validation loss (0.22) as shown in Figure 3 with regularization. Figure 4 shows that our approach without regularization has the same validation accuracy (e.g., 96.6%), but the loss function value is different (0.525) which indicates that there is an overfitting problem, whereas the validation accuracy of VGG and Inception V3 are (91.48%) and (61.92%) with validation loss of (0.23) and (0.81), respectively, as shown in Figures 6 and 7.

Tables 2 and 3 summarize the total number of images for each category, the number of images that are predicted correctly and the number of images that are predicted incorrectly for the two models. The confusion matrix for both models is calculated to visualize the performance of each model.

The performance metrics that were widely used to evaluate the predicting results of the models were precision and recall. The results are summarized in Table 4.

Furthermore, a statistical significance test was conducted to compare the results of the two models. From the evaluation, the first model with the dropout layers has the highest accuracy rate with (90.65%), while the second model without the dropout layer has the lowest accuracy rate with (76.70%).

In this study, we need to insert some dropout layers into some specific places on our CNN to overcome the overfitting barrier and get high results. It is difficult to use some CNNs architecture like ResNet or SENet because they are heavy and take long time in training, and it is difficult to control the overfitting problem easily in such architecture due to the

Table 1: A comparison between our approach and the two state-of-the-art approaches over the validation rate and the validation loss values.

Approach	Validation acc. (%)	Loss
VGG	91.48	0.23
Inception V3	61.92	0.81
Our network	96.9	0.22

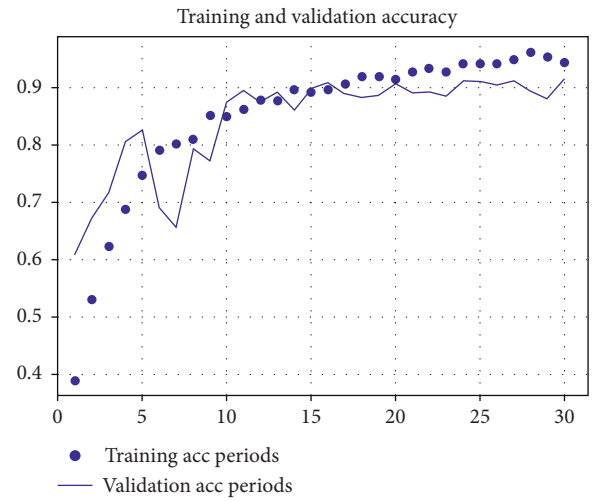


Figure 6: Training and validation accuracy for VGG.

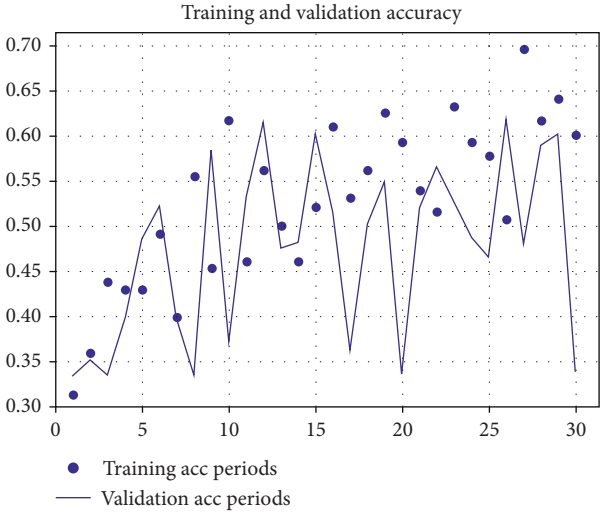


Figure 7: Training and validation accuracy for Inception V3.

Table 2: Number of images predicted correctly and incorrectly for the network with dropout layers.

Nationality	Total images	Correctly predicted	Incorrectly predicted
Chinese	540	511	29
Russian	642	561	81
Pakistani	582	527	55

Table 3: Number of images predicted correctly and incorrectly for the network without dropout layers.

Nationality	Total images	Correctly predicted	Incorrectly predicted
Chinese	540	388	152
Russian	642	467	115
Pakistani	582	498	114

Table 4: Statistical significance test for each model.

	Model with dropout	Model without dropout
TP	511	388
FP	29	152
TN	1088	968
FN	81	144
Recall (FP rate)	0.863176	0.729323
FP rate	0.025962	0.136079
Kappa	0.904659	0.762491
Accuracy rate	90.64626%	76.70068%
Precision	0.946296	0.718519

difficulties in changing their architecture. VGG and Inception V3 are also very heavy networks in training and it is difficult to change their architecture to control the overfitting problem too.

This paper is based on Cohen's methods [20]. Cohen's methods measure the degree of agreements amongst the assigned labels correcting for agreement by chance. In the evaluation, the number of unseen images is 1764, which is not included in the training dataset to evaluate the performance of each model. We found that the number of errors in the image predicting using a second model without dropout layers is larger than the number of errors in the first model with dropout layers.

5. Conclusions

In this paper, we propose a new deep learning convolutional neural network designed to create a new model that can recognize the ethnics of people through their facial features. The new model is compared with two state-of-the-art models, VGG and Inception V3, and the validation accuracy is calculated for each convolutional neural network. Two models from the proposed convolutional neural network are created with dropout layers and without dropout layers to discover the effect of the regularization in the performance of the models.

A new dataset is collected to use in the training phase to identify the ethnics of people through images from three different regions. This dataset is considered as the first dataset collected for ethnics of people and that will be available for the research community. Another unseen dataset is collected to evaluate the performance of our two models, and a statistical significance test is conducted to evaluate the performance of the two models.

Data Availability

The collected dataset has been uploaded to the following ULR: https://drive.google.com/file/d/1brRMSh7XDR7h5awgXudQXBqxAIiYSHy_/view?usp=sharing.

Disclosure

The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61836009 and 61702175), the Fund of State Key Laboratory of Geo-Information Engineering (no. SKLGIE2018-M-4-3), the Open Fund of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education (no. IPIU2019007), the fund of Hubei Key Laboratory of Transportation Internet of Things (no. WHUTIOT-2019004), and the Natural Resources Scientific Research Project of Department of the Natural Resources of Hunan Province (no. 201910).

References

[1] C. A. Hansen, *Face Recognition*, Institute for Computer Science University of Tromso, Tromsø, Norway, 2009.

[2] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 586–591, IEEE, Maui, HI, USA, August 1991.

[3] S. Satonkar Suhas, B. Kurhe Ajay, and B. Prakash Khanale, "Face recognition using principal component analysis and linear discriminant analysis on holistic approach in facial images database," *IOSR Journal of Engineering*, vol. 2, no. 12, pp. 15–23, 2012.

[4] H. Khalajzadeh, M. Mansouri, and M. Teshnehlab, "Face recognition using convolutional neural network and simple logistic classifier," in *Soft Computing in Industrial Applications*, pp. 197–207, Springer, Berlin, Germany, 2014.

[5] Z. Xiang, H. Tan, and W. Ye, "The excellent properties of a dense grid-based HOG feature on face recognition compared to gabor and LBP," *IEEE Access*, vol. 6, 2018.

[6] J. Yan, L. Zhang, Y. Wu et al., "Research on face recognition method based on deep learning in natural environment," in *Proceedings of the IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pp. 501–506, Taichung, Taiwan, November 2017.

[7] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, Columbus, OH, USA, June 2014.

[8] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1988–1996, Montreal, Canada, December 2014.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, MA, USA, 2007.

[10] S. K. Palei and S. K. Das, "Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: an approach," *Safety Science*, vol. 47, no. 1, pp. 88–96, 2009.

[11] S. Permission, "Generative and discriminative classifiers: naive bayes and logistic regression," 2005.

[12] D. Kingma and J. Ba, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097–1105, 2012.

[14] 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*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, https://arxiv.org/abs/1207.0580.

[16] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, pp. 351–359, MIT Press, Cambridge, MA, USA, 2013.

[17] P. Baldi and P. J. Sadowski, "Understanding dropout," in *Advances in Neural Information Processing Systems*, vol. 26, pp. 2814–2822, MIT Press, Cambridge, MA, USA, 2013.

[18] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, Kuala Lumpur, Malaysia, November 2015.

[19] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.

[20] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

Hindawi
Scientific Programming
Volume 2020, Article ID 6348372, 18 pages
https://doi.org/10.1155/2020/6348372

Research Article

Isomap-Based Three-Dimensional Operational Modal Analysis

Cheng Wang,[1,2] **Weihua Fu**,[1] **Haiyang Huang**,[1] **and Jianwei Chen**[3]

[1]*College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[2]*State Key Laboratory for Strength and Vibration of Mechanical Structures, Xi'an Jiaotong University, Xi'an 710049, China*
[3]*Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA*

Correspondence should be addressed to Cheng Wang; wangcheng@hqu.edu.cn

Received 7 February 2020; Revised 2 June 2020; Accepted 9 June 2020; Published 14 July 2020

Academic Editor: Chenxi Huang

In order to identify the modal parameters of time invariant three-dimensional engineering structures with damping and small nonlinearity, a novel isometric feature mapping (Isomap)-based three-dimensional operational modal analysis (OMA) method is proposed to extract nonlinear features in this paper. Using this Isomap-based OMA method, a low-dimensional embedding matrix is multiplied by a transformation matrix to obtain the original matrix. We find correspondence relationships between the low-dimensional embedding matrix and the modal coordinate response and between the transformation matrix and the modal shapes. From the low-dimensional embedding matrix, the natural frequencies can be determined using a Fourier transform and the damping ratios can be identified by the random decrement technique or natural excitation technique. The modal shapes can be estimated from the Moore–Penrose matrix inverse of the low-dimensional embedding matrix. We also discuss the effects of different parameters (i.e., number of neighbors and matrix assembly) on the results of modal parameter identification. The modal identification results from numerical simulations of the vibration response signals of a cylindrical shell under white noise excitation demonstrate that the proposed method can identify the modal shapes, natural frequencies, and ratios of three-dimensional structures in operational conditions only from the vibration response signals.

1. Introduction

Operational modal analysis (OMA) has received widespread attention because it enables the identification of the modal parameters of a structure in its working condition using only the vibration response [1]. Modal parameters (including the modal natural frequencies, modal shapes, and modal damping ratios) are essential for structural vibration control, damage diagnosis, and so on [2, 3]. Recently, blind source separation (BSS) methods such as sparse component analysis (SCA) [4, 5] and independent component analysis (ICA) [6] have been widely used for output-only identification.

Dimensionality reduction techniques are an effective means of overcoming the curse of dimensionality. Current methods can be categorized as either linear or nonlinear dimensionality reduction [7]. Classical linear dimensionality reduction algorithms include principal component analysis (PCA) [8], locality preserving projection (LPP) [9], multidimensional scaling (MDS) [10], and linear discriminant analysis (LDA) [11]. Recently, Wang et al. introduced PCA to OMA [12] and proposed PCA and second-order blind identification- (SOB-I-) based three-dimensional OMA methods [13, 14]. However, the time invariant engineering structures are often three-dimensional, with damping and small nonlinearity [15].

Similarly, there are many nonlinear dimensionality reduction methods, such as locally linear embedding (LLE) [16, 17], Laplacian eigenmaps (LE) [18], kernel PCA [19], and isometric feature mapping (Isomap) [20]. Bai et al. proposed an LLE-based OMA method for three-dimensional structures [21], Zhang et al. optimized the nearest neighbor selection method for LLE-based OMA [22], Dong et al. introduced modal identification and influence factors of LLE algorithm [23], and Guan et al. made comprehensive and systematic comparisons of four statistical learning algorithms (PCA, ICA, SOBI, and LLE) on analyzing their performance for resolving operational modal parameters identification [24]. This paper applies Isomap to OMA.

Isomap is a nonlinear dimensionality reduction algorithm for manifold learning [25]. This method can find meaningful low-dimensional structures hidden in their high-dimensional observations. This idea can be used in OMA. To date, Isomap has been widely used in various fields [26].

Based on Isomap algorithm, this paper proposes a three-dimensional OMA method for complex three-dimensional continuum structures.

The primary contributions of this paper can be summarized as follows:

(1) An Isomap-based OMA method is proposed for the identification of modal shapes, modal natural frequencies, and modal ratios of three-dimensional structures

(2) We identify the correspondence between the low-dimensional embedding matrix and the modal response matrix and between the transformation matrix and the modal shapes

(3) We conduct a theoretical analysis of the characteristics of the Isomap-based OMA method

(4) We analyze the effects of different parameters (e.g., matrix assembly method, number of neighbors, and dimensionality reduction method) on the algorithm

(5) We design numerical simulations of the vibration response signals of a cylindrical shell to verify the effectiveness of our algorithm

The remainder of this paper is organized as follows. In Section 2, the Isomap-based OMA method for three-dimensional modal parameter identification and the characteristics of the proposed method are introduced. The effect of various parameters is investigated in Section 3. Section 4 presents the simulation verification results. Finally, we conclude the paper in Section 5.

2. Isomap-Based Three-Dimensional Operational Modal Analysis

2.1. Stationary Response Signals Decomposition in Modal Coordinate and Operational Modal Identification. According to dynamic structural vibration theory, the dynamic differential equation of a d-degree-of-freedom (DOF) linear system in the physical coordinate system is

$$\mathbf{M}\ddot{\mathbf{X}}(t) + \mathbf{C}\dot{\mathbf{X}}(t) + \mathbf{K}\mathbf{X}(t) = \mathbf{F}(t), \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{D\times D}$ is the mass matrix, $\mathbf{C} \in \mathbb{R}^{D\times D}$ is the damping matrix, $\mathbf{K} \in \mathbb{R}^{D\times D}$ is the stiffness matrix, $F(t) \in \mathbb{R}^{D\times T}$ is the external excitation, and $\ddot{\mathbf{X}}(t), \dot{\mathbf{X}}(t), \mathbf{X}(t) \in \mathbb{R}^{D\times T}$ are the acceleration, speed, and displacement response signals, respectively.

The free and random vibrations of weakly damped systems have a displacement response that can be expressed as follows in modal coordinates:

$$\mathbf{X}(t) \approx \mathbf{\Phi}\mathbf{Q}(t) = \sum_{i=1}^{d} \vec{\phi}_i \vec{q}_i(t), \tag{2}$$

where $\mathbf{\Phi} \in \mathbb{R}^{D\times d}$ is the modal shapes matrix constituted by the modal shape $\vec{\phi}_i$ of each order and $\mathbf{Q}(t) \in \mathbb{R}^{d\times T}$ is the modal response matrix formed by the modal responses $\vec{q}_i(t)$. When the order of each natural frequency is different, the modal shape vector $\vec{\phi}_i \in \mathbb{R}^{d\times 1}$ becomes

$$\begin{aligned} \mathbf{\Phi}^T\mathbf{M}\mathbf{\Phi} &= \mathbf{M} = \operatorname{diag}\left[\vec{m}_1, \vec{m}_2, \ldots, \vec{m}_i \ldots, \vec{m}_d\right], \\ \mathbf{\Phi}^T\mathbf{K}\mathbf{\Phi} &= \mathbf{K} = \operatorname{diag}\left[\vec{k}_1, \vec{k}_2, \ldots, \vec{k}_i \ldots, \vec{k}_d\right]. \end{aligned} \tag{3}$$

The modal responses $\vec{q}_i(t)$ are independent of each other and $\Lambda_{d\times d}$ is a diagonal matrix:

$$E\left[\mathbf{Q}(t)\mathbf{Q}^T(t)\right] = \mathbf{\Lambda}_{d\times d}. \tag{4}$$

Taking advantage of single-DOF (SDOF) identification approach such as FFT, it is easy to recognize the ith natural frequency and modal damping ratio from the modal responses $\vec{q}_i(t)$.

After $\mathbf{Q}(t)$ has been obtained, the modal shape matrix Φ can be estimated using the Moore–Penrose matrix inverse:

$$\mathbf{\Phi} = \mathbf{X}(t)\mathbf{Q}^T(t)\left(\mathbf{Q}(t)\mathbf{Q}^T(t)\right)^{-1}. \tag{5}$$

OMA is to identify modal shape matrix Φ, natural frequencies, and modal damping ratios only from the vibration response signals $\mathbf{X}(t)$.

2.2. Isomap-Based OMA. Isomap is a very popular manifold learning algorithm. Unlike PCA, which is designed only for linear dimensionality reduction, Isomap can solve nonlinear dimensionality reduction problems [27]. The dimension-reducing Isomap algorithm takes the sample set $\mathbf{X}(t) = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\} \in \mathbb{R}^{D\times T}$ as input, where D is the number of dimensions and T is the number of samples. The values of two hyperparameters are then set: the number of neighbors K and the low-dimensional space dimension d. The output of the algorithm is the matrix of samples in low-dimensional space $\mathbf{S}(t) = \{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_T\} \in \mathbb{R}^{d\times T}$. The steps involved in the algorithm can be described as follows:

(1) Build a neighborhood graph G. For each sample point $x_{ij} \in \vec{x}_i \in \mathbb{R}^{D\times 1}, 1 \le i \le T,\ 1 \le j \le D$, we first calculate which points are its neighbors. Based on the Euclidean distance, all points within some fixed radius ε or the K nearest neighbors are selected. The points between x_{ij} and its neighbors are then connected, as shown in Figure 1. These neighborhood relations are represented as a weighted graph G over the data points.

(2) Estimate the geodesic distances. Call the shortest path algorithm to calculate the distance between any two points. This is also called the geodesic distance, that is, the distance to the adjacent point is the Euclidean distance, and that to the nonadjacent point is the shortest path distance. The specific calculation process for the distance matrix $\mathbf{B}(t) \in \mathbb{R}^{T\times T}$ is given in (6)–(10).

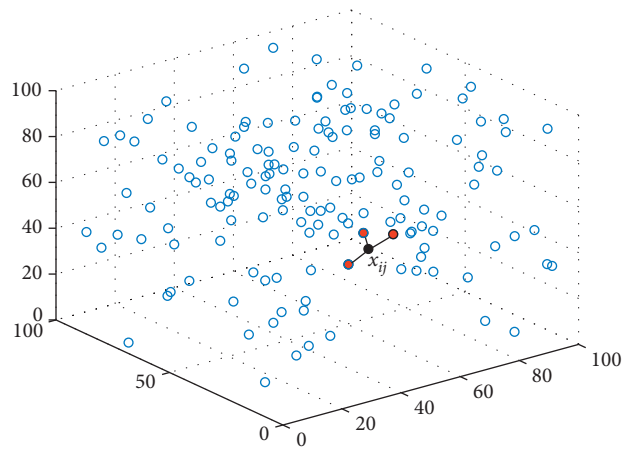


Figure 1: Connection of a sample to its neighbors in the Isomap algorithm.

(3) Call the classical multidimensional scaling (MDS) algorithm to obtain the matrix $\mathbf{S}(t)$ of the sample set in the low-dimensional space.

The following can be obtained from the steps of MDS. The aim of the MDS algorithm is to obtain the representation of the sample in the d-dimensional space, $\mathbf{S}(t) = \{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_T\} \in \mathbb{R}^{D\times T}, d < D$. The geodesic distance of any two samples in the d-dimensional space is equal to the distance in the original space, which means $\|s_i - \vec{s}_j\| = \text{dist}_{ij}$ and dist_{ij} is the distance between any two samples.

Setting $\mathbf{B}(t) = \mathbf{S}(t)^T\mathbf{S}(t) \in \mathbb{R}^{T\times T}$, we have that $b_{ij} = \vec{s}_i^T\vec{s}_j$, so the following equation can be obtained:

$$d_{ij}^2 = \left\|\vec{s}_i - \vec{s}_j\right\|^2 = \left\|\vec{s}_i\right\|^2 + \left\|\vec{s}_j\right\|^2 - 2\vec{s}_i^T\vec{s}_j = b_{ii} + b_{jj} - 2b_{ij}. \tag{6}$$

From the following equations

$$\text{dist}_{i\cdot}^2 = \frac{1}{T}\sum_{j=1}^{T} d_{ij}^2, \tag{7}$$

$$\text{dist}_{\cdot j}^2 = \frac{1}{T}\sum_{i=1}^{T} d_{ij}^2, \tag{8}$$

$$\text{dist}_{\cdot\cdot}^2 = \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T} d_{ij}^2, \tag{9}$$

we obtain that

$$b_{i,j} = \frac{d_{i,\cdot}^2 + d_{\cdot,j}^2 - d_{\cdot,\cdot}^2 - d_{ij}^2}{2}. \tag{10}$$

In summary, the classical MDS algorithm can be split into the following steps:

(1) Using (7)–(9), calculate $\text{dist}_{i\cdot}^2$, $\text{dist}_{\cdot j}^2$, and $\text{dist}_{\cdot\cdot}^2$

(2) Using (10), compute the matrix $\mathbf{B}(t)$

(3) Perform eigenvalue decomposition on $\mathbf{B}(t)$

(4) Construct $\widetilde{\mathbf{\Lambda}}$ as the diagonal matrix of the d largest eigenvalues and $\widetilde{\mathbf{V}}$ as the corresponding eigenvector matrix

(5) $\mathbf{S}(t) = \widetilde{\mathbf{V}}\widetilde{\mathbf{\Lambda}}^{1/2} \in \mathbb{R}^{T\times d}$ contains the low-dimensional coordinates for each sample

We now introduce the Isomap algorithm; an example of using Isomap for nonlinear dimensionality reduction is presented in Figure 2 [25]. (a) of Figure 2 is a dataset of three-dimensional space. After the dimensionality reduction by Isomap algorithm, the dataset of two-dimensional space (Figure 2(b)) still maintains the internal connection of three-dimensional space data.

PCA and classical MDS can be effectively applied to Euclidean structures but fail to extract nonlinear features [21]. The Isomap algorithm is an improved version of the classical MDS method that can solve nonlinear problems. More specifically, the dimensionality reduction principle of the Isomap method makes the Euclidean distance between the low-dimensional data equal to the geodesic distance between the high-dimensional data. When calculating the distance between data points on a high-dimensional manifold, instead of applying the traditional Euclidean distance, Isomap uses the geodesic distance in the differential geometry and estimates this distance using actual input data.

The Isomap algorithm can obtain low-dimensional representations $\mathbf{S}(t) \in \mathbb{R}^{d\times T}$ that are also independent, so there exists

$$\mathbf{H} = \mathbf{X}(t)\mathbf{S}^T(t)\left(\mathbf{S}(t)\mathbf{S}^T(t)\right)^{-1}, \tag{11}$$

where $\mathbf{H}$ is the transformation matrix. Therefore, $\mathbf{X}(t)$ has the following form:

$$\mathbf{X}(t) \approx \mathbf{H}\mathbf{S}(t) = \sum_{j=1}^{d} \vec{h}_j\vec{s}_j(t). \tag{12}$$

Comparing (2) and (12), we can conclude that there is a one-to-one corresponding mapping relationship between the modal shape matrix Φ in modal coordinate and transformation matrix $\mathbf{H}$ in Isomap, the modal responses matrix $\mathbf{Q}(t)$ in modal coordinate, and low-dimensional representations $\mathbf{S}(t)$ in Isomap. Figure 3 shows a physical interpretation of Isomap-based OMA.

2.3. Three-Dimensional Modal Parameter Identification. The OMA of "three-dimensional" structure is a more complex problem, because it requires the assembly of the modal parameter in three directions to calculate the "three-dimensional" modal parameter.

Therefore, we introduce two three-dimensional matrix assembly methods in this section, namely, least-squares matrix substitution (LSMS) and direct matrix assembly (DMA). After the data matrix has been assembled, we use the Isomap algorithm to find the modal parameters.

Continuum-structure mechanical systems can be divided into D discrete observation components at regular intervals. The displacement response of a three-dimensional cylindrical shell in the time domain can be expressed by its

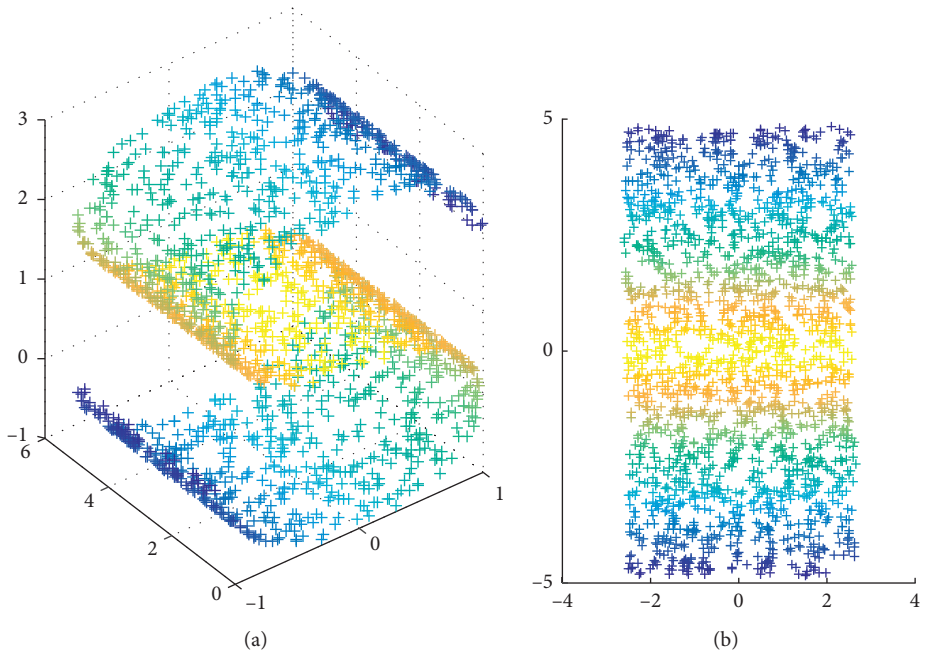


Figure 2: An example of using Isomap for nonlinear dimensionality reduction: (a) original three-dimension data before dimensionality reduction; (b) two-dimension data after dimensionality reduction by Isomap.

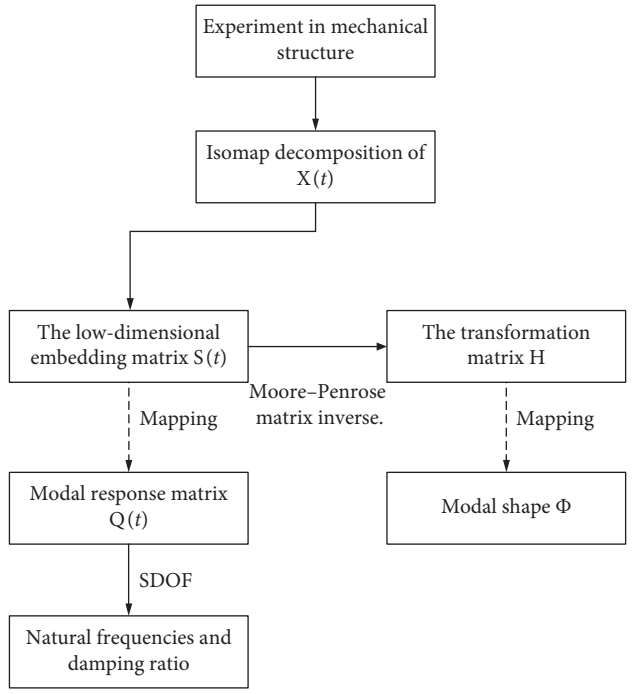


Figure 3: Physical interpretation of Isomap-based OMA.

modal coordinate approximation when D is sufficiently large:

$$\begin{cases} \mathbf{X}(t) = \mathbf{U}\mathbf{Q}(t) = \sum_{i=1}^{d} \vec{u}_i \vec{q}_i(t), \\ \mathbf{Y}(t) = \mathbf{V}\mathbf{Q}(t) = \sum_{i=1}^{d} \vec{v}_i \vec{q}_i(t), \\ \mathbf{Z}(t) = \mathbf{W}\mathbf{Q}(t) = \sum_{i=1}^{d} \vec{w}_i \vec{q}_i(t). \end{cases} \tag{13}$$

In (13), $\mathbf{X}(t), \mathbf{Y}(t), \mathbf{Z}(t) \in \mathbb{R}^{D\times T}$ are the vibration responses in three directions, d is the embedding dimension, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ are the modal shapes in three directions, T is the sampling time, and the modal response matrix $\mathbf{Q}(t)$ of each dimension is the same. Using the minimum square sum of errors as the objective function to be optimized, the Moore–Penrose matrix inverse of (13) is [13]

$$\begin{cases} \mathbf{U} = \mathbf{X}(t)\mathbf{Q}^T(t)\left[\mathbf{Q}(t)\mathbf{Q}^T(t)\right]^{-1}, \\ \mathbf{V} = \mathbf{Y}(t)\mathbf{Q}^T(t)\left[\mathbf{Q}(t)\mathbf{Q}^T(t)\right]^{-1}, \\ \mathbf{W} = \mathbf{Z}(t)\mathbf{Q}^T(t)\left[\mathbf{Q}(t)\mathbf{Q}^T(t)\right]^{-1}. \end{cases} \tag{14}$$

In the LSMS-based Isomap method, a single-dimensional displacement response signal is first decomposed. To reduce the influence of the Gaussian-distributed signal measurement noise, the main and biggest structure's dynamic response dimension is selected in practical engineering cases. Substituting $\mathbf{Q}(t)$ into the other two-dimensional displacement response signals using the Moore–Penrose matrix inverse in (14), the modal shapes of these two dimensions can be identified. Three-dimensional modal shapes $[\vec{u}_i, \vec{v}_i, \vec{w}_i]\, i = 1, 2, \ldots, d$ are then assembled. To identify the operational modal parameters of the three-dimensional structure, the following strategy is adopted and the process of identification is performed for the three-dimensional structure. The process is described in Figure 4, where it is assumed that the vibration response is greatest in the $\mathbf{X}(t)$ direction.

From (13), we know that the modal response matrix $\mathbf{Q}(t)$ of each dimension is the same. Therefore, the DMA method assembles the modal response in three directions as the overall modal response $\mathbf{A}(t) \in \mathbb{R}^{3D\times T}$ of the entire structure [28]:

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{X}(t) \\ \mathbf{Y}(t) \\ \mathbf{Z}(t) \end{bmatrix}. \tag{15}$$

The modal coordinate response and modal shape of the overall structure are solved in one pass:

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{X}(t) \\ \mathbf{Y}(t) \\ \mathbf{Z}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \\ \mathbf{W} \end{bmatrix} \mathbf{Q}(t). \tag{16}$$

Using the Moore–Penrose matrix inverse of (16), we can obtain the modal shapes of all three directions as

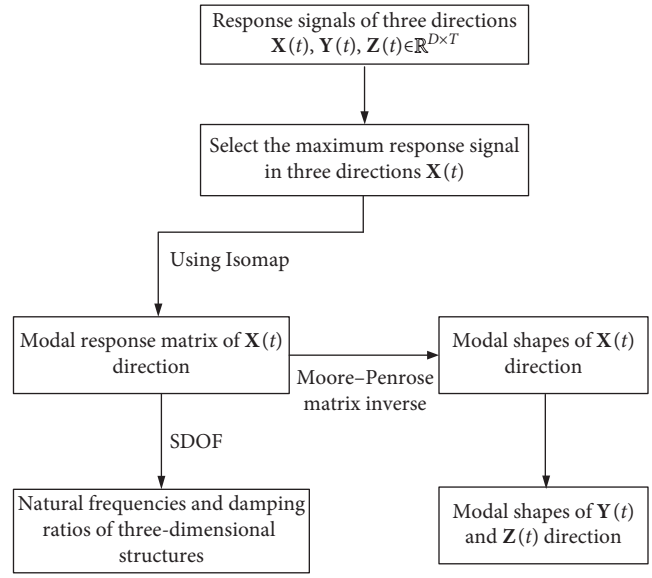


Figure 4: Process of LSMS-based Isomap for three-dimensional structure OMA.

$$\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \\ \mathbf{W} \end{bmatrix} = \mathbf{A}(t)\mathbf{Q}^T(t)\left[\mathbf{Q}(t)\mathbf{Q}^T(t)\right]^{-1}. \tag{17}$$

The process of the DMA-based Isomap method for three-dimensional OMA is described in Figure 5.

2.4. Characteristics of Isomap-Based OMA. Isomap-based OMA has the following characteristics:

(1) The order of the identified modal parameters is different from the theoretical value. The Isomap algorithm is an improved version of classical MDS, and the order of recognition follows the contribution of the principal components, running from small to large.

(2) The amplitude information of the modal shapes is lost. According to (8), the modal shapes of each order are orthogonal to one another and the modal shape vector is unitized.

(3) Some modal information may be missing. When the contribution of an independent component is small, it is difficult to identify its modal parameters.

3. Effect of Parameters on the Algorithm

3.1. Matrix Assembly. The LSMS method selects the maximum vibration response data of the three-dimensional structure to calculate the modal response matrix, which is used to determine the other modal parameters. However, the modal coordinate response of the direction of maximum vibration response is not equal to the global modal coordinate response. Although we obtain this matrix at a small cost, the resulting three-dimensional modal shape must have a large error.

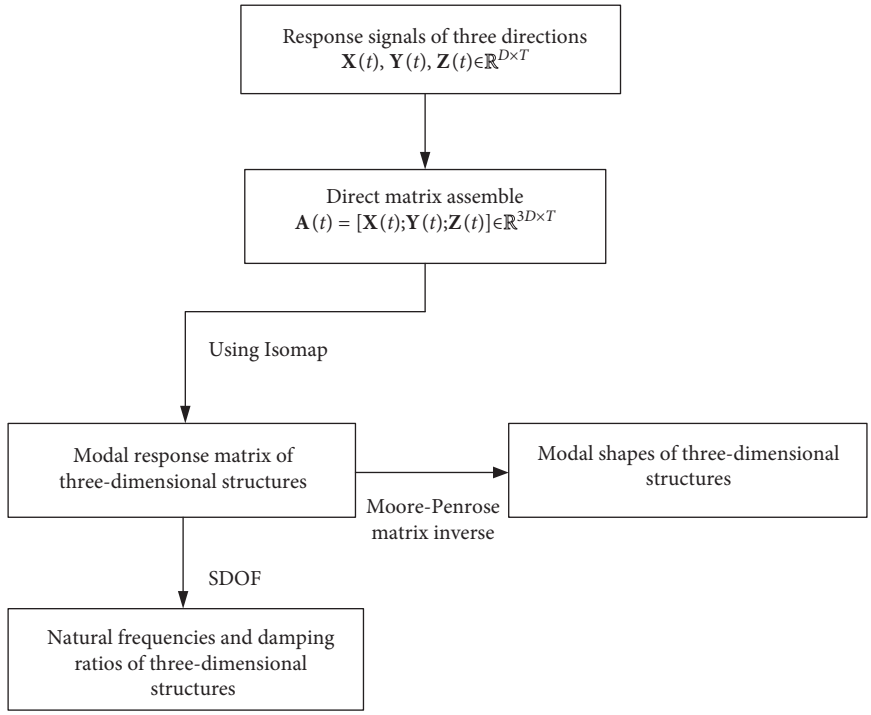


Figure 5: Process of DMA-based Isomap for three-dimensional structure OMA.

The DMA-based method directly obtains the global modal shape and overall modal coordinate response of three-dimensional structures and then uses Isomap to identify the modal parameters. Compared with LSMS, DMA avoids the need for matrix inversion operations and is more robust. Furthermore, this method offers greater accuracy because matrix inversion errors and ill-posed problems [29] are inevitable in the matrix decomposition calculation. Table 1 compares the performance of the two assembly methods [26].

Table 1: Performance comparison of two assembly methods in Isomap.

Matrix assembly methods	LSMS	DMA
Matrix inversion error	✓	×
Ill-posed problems	✓	×
Robustness	Sensitive to measurement noise	Insensitive to measurement noise

3.2. Number of Neighbors K. The neighbor value K refers to the number of data points to which a given point is connected. Each point is connected to the nearest K points to form a graph G, and then some algorithm (e.g., Dijkstra's algorithm and Floyd's algorithm) is used to calculate the shortest path.

If K is too small, the connectivity graph may not be formed, which will affect the calculation of the shortest distance. At the same time, the negative effects of noisy points will be amplified. A sufficiently large value of K can reduce the difference between the path length and the true geodesic distance. If K is too large, the computation time may become unbearable and underfitting may occur.

3.3. Linear and Nonlinear Methods. Both MDS and PCA are linear dimensionality reduction techniques. The PCA method looks for the low-dimensional embedding of the data points which best preserves their variance, as measured in the high-dimensional input space. The classical MDS method finds an embedding that preserves the interpoint distances. We use the Euclidean distance in the proposed method, and the results of PCA are consistent with those from MDS [21]. Our experimental results also confirm this.

Isomap is a nonlinear dimensionality reduction technique. When there are nonlinear features in the three-dimensional structures, the extraction performance of linear methods is poor. The linear and nonlinear relationships between the data also affect the algorithm results.

4. Simulation Verification and Result Analysis of Three-Dimensional Operational Modal Parameter Identification

4.1. Generation of Simulation Data. We conducted simulations to study a cylindrical shell with a complex three-dimensional structure. The cylindrical shell is simply supported at both ends, and a certain number of vibration sensors are positioned on its surface to record the vibration response in three directions, with a vibration exciter used to simulate the working conditions. Figure 6 illustrates this process.

The cylindrical shell has a thickness of 0.005 m, radius of 0.1825 m, length of 0.37 m, and elasticity modulus of

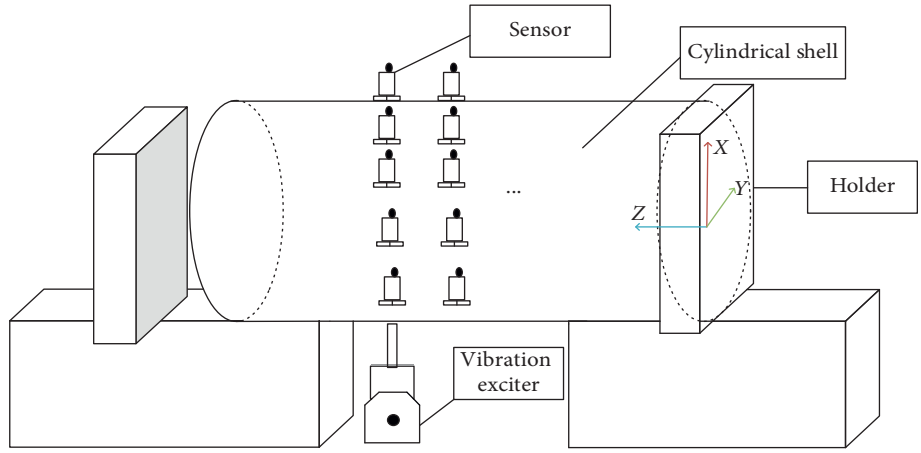


Figure 6: Simulation process.

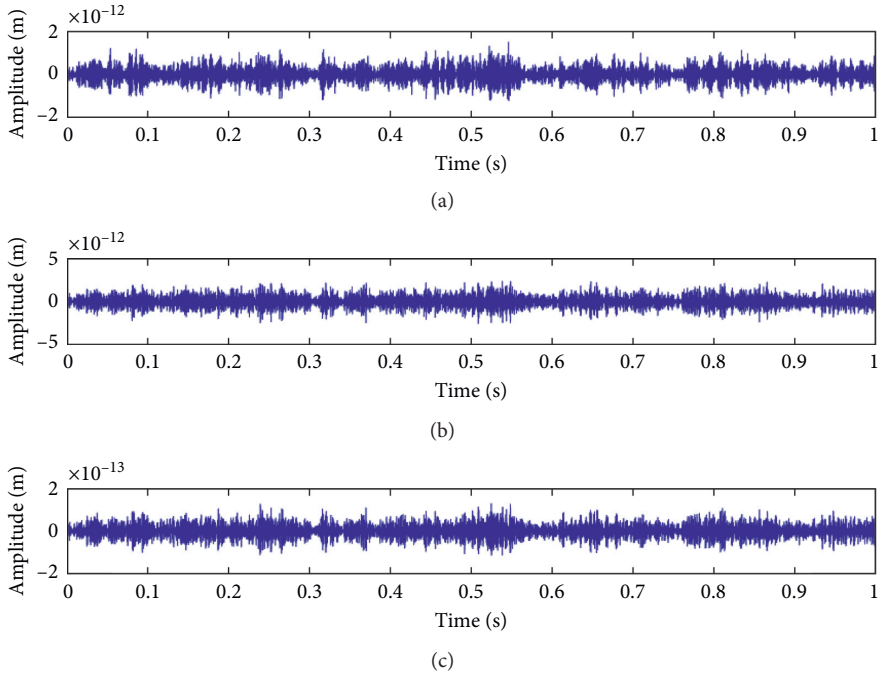


Figure 7: Response signals in three directions. (a) Response in X direction at the 1118th observation point. (b) Response in Y direction at the 1118th observation point. (c) Response in Z direction at the 1118th observation point.

205 GPa. The material has Poisson's ratio of 0.3, density of 7850 kg/m^3, and mode damping ratios of 0.03, 0.05, and 0.1.

The cylindrical shell was considered to be a uniform axial distribution of 38 circles, and each circle had 115 uniformly distributed observation points, giving a total of $D = 4370$ observation points. The sampling frequency was set to 5120 Hz, and the sampling time was set to 1 s. Finally, response signals in three directions were calculated by LMS Virtual Lab using finite element analysis (FEA). Response signals in the three directions at a random observation point are shown in Figure 7.

4.2. Evaluation Index. To evaluate the effect of identification using the proposed method for three-dimensional structures, the mode shapes and natural frequencies were calculated using FEA. These were considered the real modal parameters for comparison with the identified modal parameters. The modal assurance criterion (MAC) reflects the effectiveness of the modal identification given by the proposed method. MAC is defined as

$$\mathrm{MAC}_{\vec{\phi}_i, \vec{\varphi}_i} = \frac{\left(\vec{\phi}_i^{\,T} \vec{\varphi}_i\right)^2}{\left(\vec{\phi}_i^{\,T} \vec{\phi}_i\right)\left(\vec{\varphi}_i^{\,T} \vec{\varphi}_i\right)}, \tag{18}$$

where $\vec{\phi}_i$ is the identified ith-order modal shape and $\vec{\varphi}_i$ is the true ith-order modal shape. The MAC values range from 0 to 1, with higher values indicating a better correlation. The modals are all real, and there are no complexes.

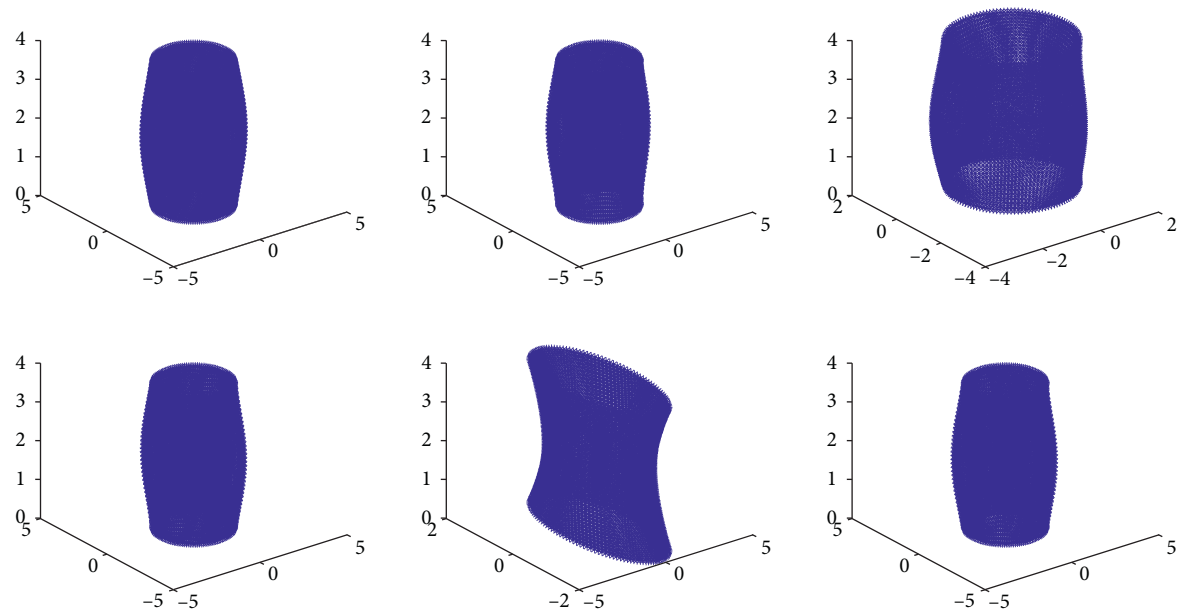


Figure 8: Real modal shapes calculated by FEA.

4.3. Simulation Results of Three-Dimensional Operational Modal Parameter Identification. From the response signals in the three directions shown in Figure 7, it is apparent that the vibration response in the $\mathbf{X}(t)$ direction is greater than those in the $\mathbf{Y}(t)$ and $\mathbf{Z}(t)$ directions. Thus, the LSMS-based method first used Isomap to decompose the $\mathbf{X}(t)$ direction.

The modal shapes and natural frequencies calculated by the FEA method with a damping ratio of 0.03 were considered to be the real values. Figure 8 shows the real modal shapes.

Embedding dimension of 6 and number of neighbors $K = 40$, when the modal response matrix $\mathbf{Q}(t)$ is obtained by Isomap, we apply a fast Fourier transform (FFT) to each of its columns and take the X-coordinate corresponding to the highest amplitude to be the modal natural frequencies of each order.

To enable a better comparison, we rotated the coordinates and obtained the results shown in Figure 9.

The modal shapes and natural frequencies identified by the LSMS-based Isomap algorithm are shown in Figures 10 and 11. Under the same conditions, the modal shapes and frequencies identified by the DMA-based Isomap algorithm are presented in Figures 12 and 13. Tables 2–5 compare the frequency, MAC of modal shapes, modal shapes, and modal ratios, respectively, identified by LSMS-based Isomap, DMA-based Isomap, and FEA.

The modal shape and frequency identified by the LSMS-based Isomap algorithm with an embedding dimension of 5 are shown in Figures 14 and 15.

Tables 6 and 7 compare the MAC and frequencies identified by the LSMS-based Isomap algorithm in different dimensions.

Tables 8–10 compare the MAC of modal shapes, modal frequencies, and modal damping ratios calculated by the LSMS and DMA methods for different numbers of neighbors K.

For the modal ratio, we apply the random decrement technique (RDT) [30] or natural excitation technique (NExT) [31] to each column of the modal response matrix $\mathbf{Q}(t)$ and then obtain the envelope of the curve from the Hilbert transform. Fitting an exponential decay to the envelope, the slope of best fit gives the modal ratio. Figure 16 illustrates this process.

We compare the modal damping ratio identification methods (RDT and NExT) using different dimensionality reduction methods (PCA [13]) in Tables 11–13. The results using LSMS to identify the MAC and frequencies are presented in Tables 14–16, and those using DMA are given in Tables 17–19.

4.4. Analysis of Simulation Results for Three-Dimensional Operational Modal Parameter Identification

(1) From Figures 8–13 and Tables 2–5, we can see that the DMA- and LSMS-based Isomap methods can effectively identify the modal shapes and frequencies. The DMA method does well in latter orders, whereas the LSMS method can better identify the first few orders of modal shapes. For the modal ratios, the DMA assembly method is better able to obtain the 6th-order results.

(2) From Tables 2–5, it can be concluded that the modal shapes and frequencies do not have a one-to-one correspondence. This is because the order of the mode coordinate vector is uncertain.

(3) From Figures 14 and 15 and Tables 6 and 7, the results of the five-dimensional case are the same as those of the first five dimensions of the six-

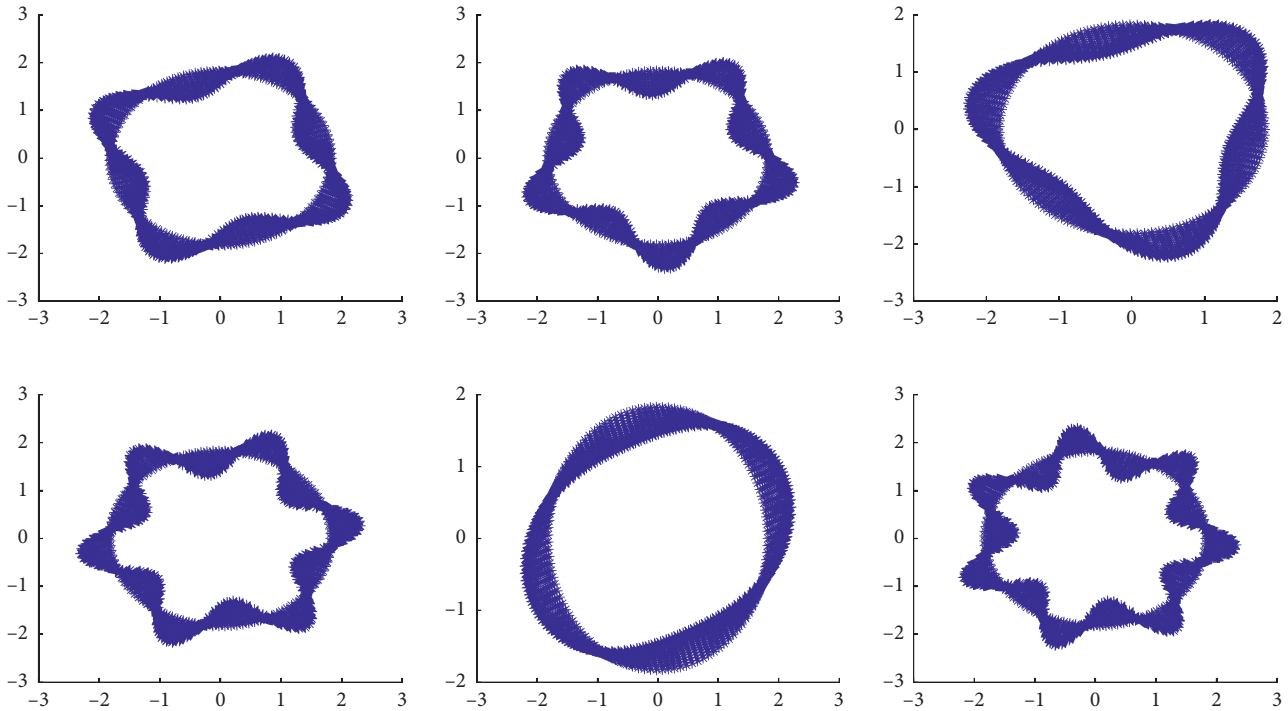


Figure 9: Real modal shapes after rotating the coordinates.

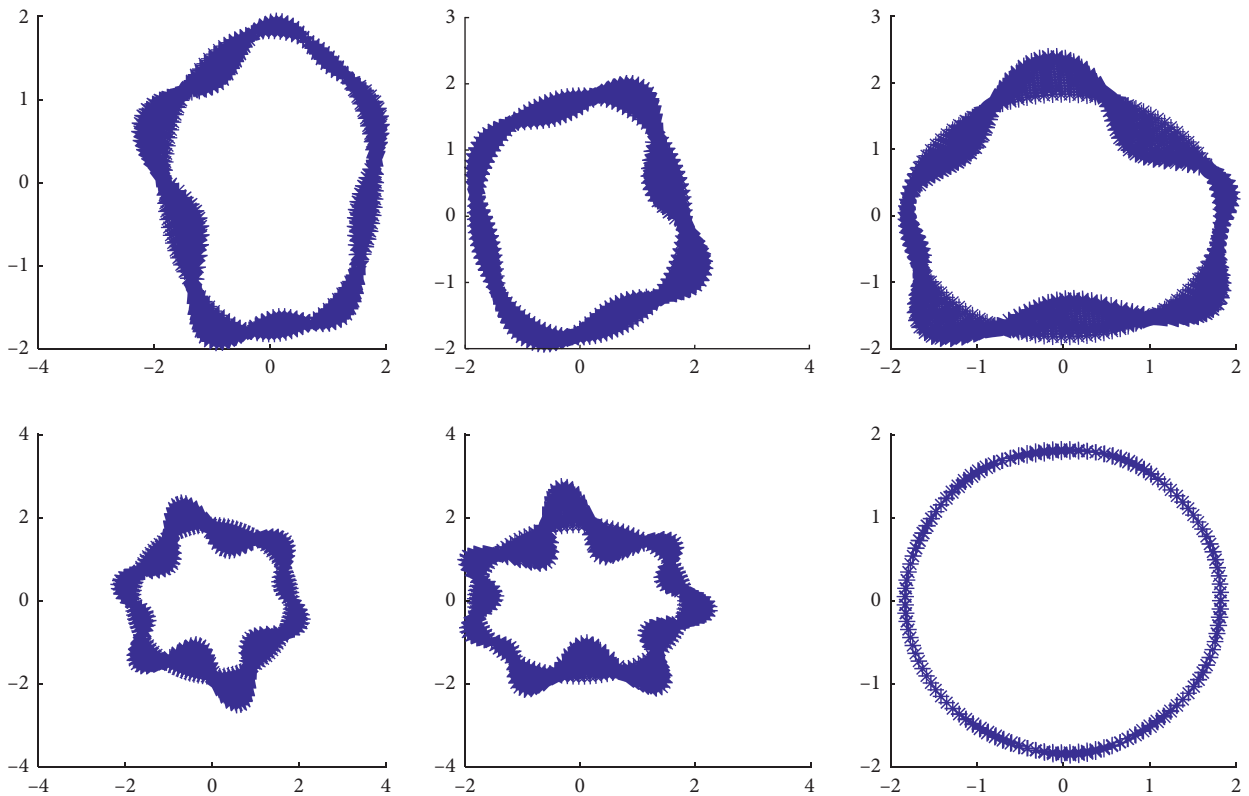


Figure 10: Modal shapes identified by LSMS-based Isomap.

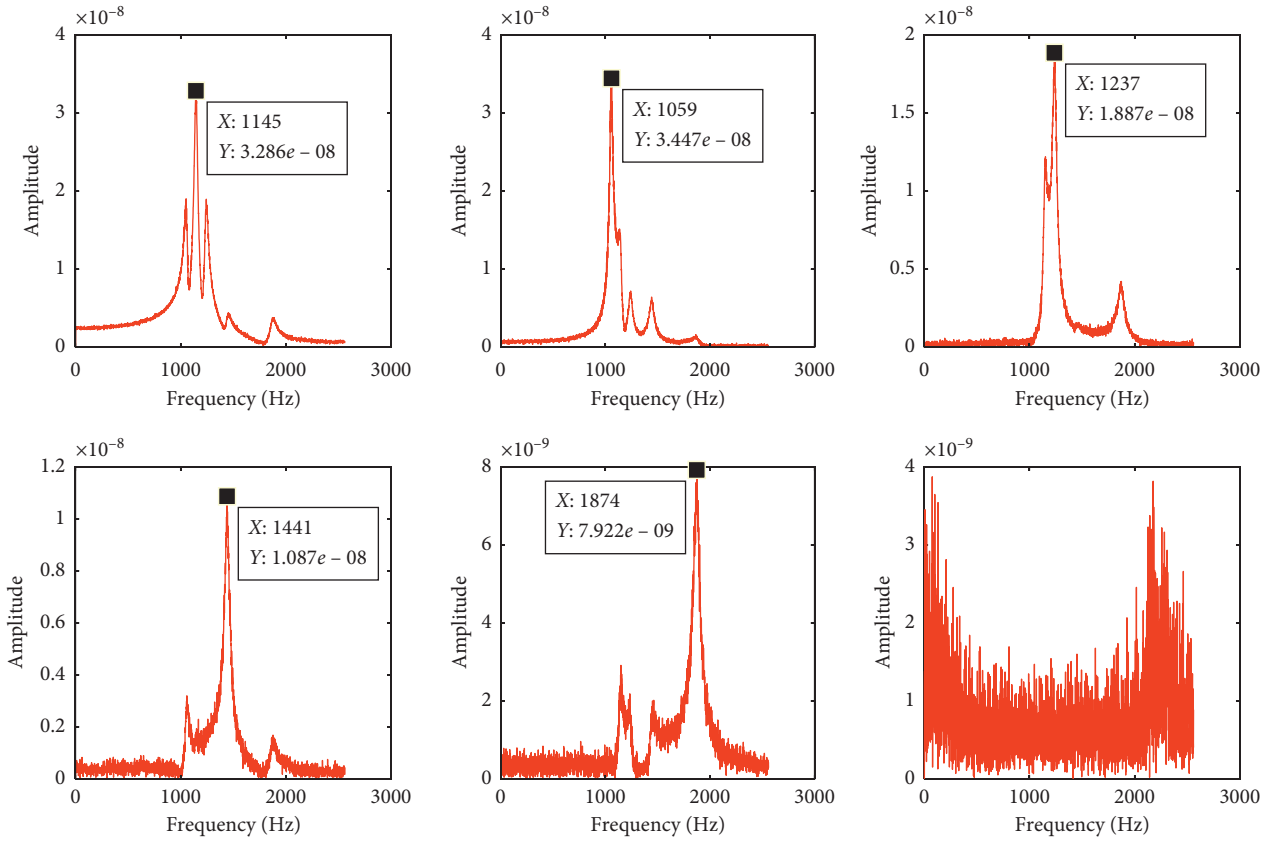


Figure 11: Modal frequencies identified by LSMS-based Isomap.

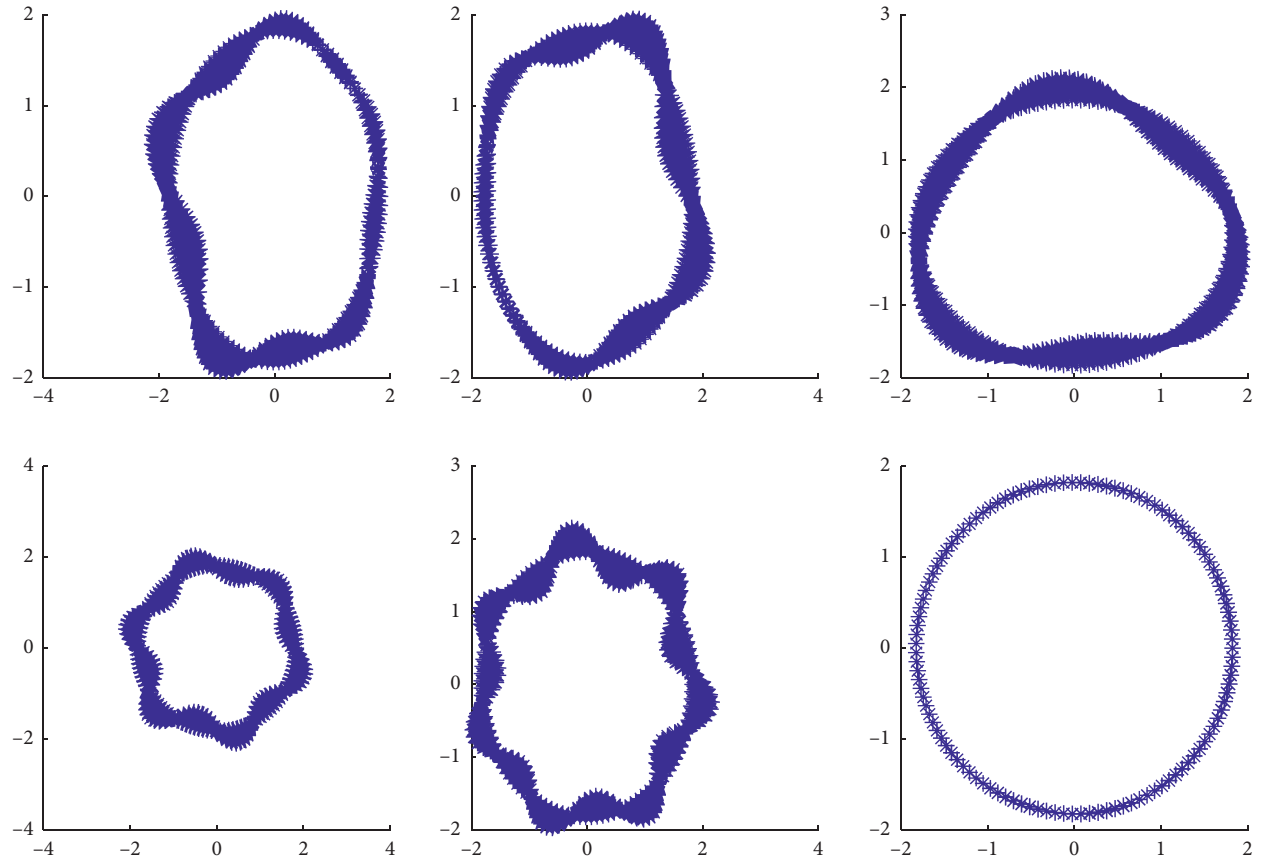


Figure 12: Modal shapes identified by DMA-based Isomap.

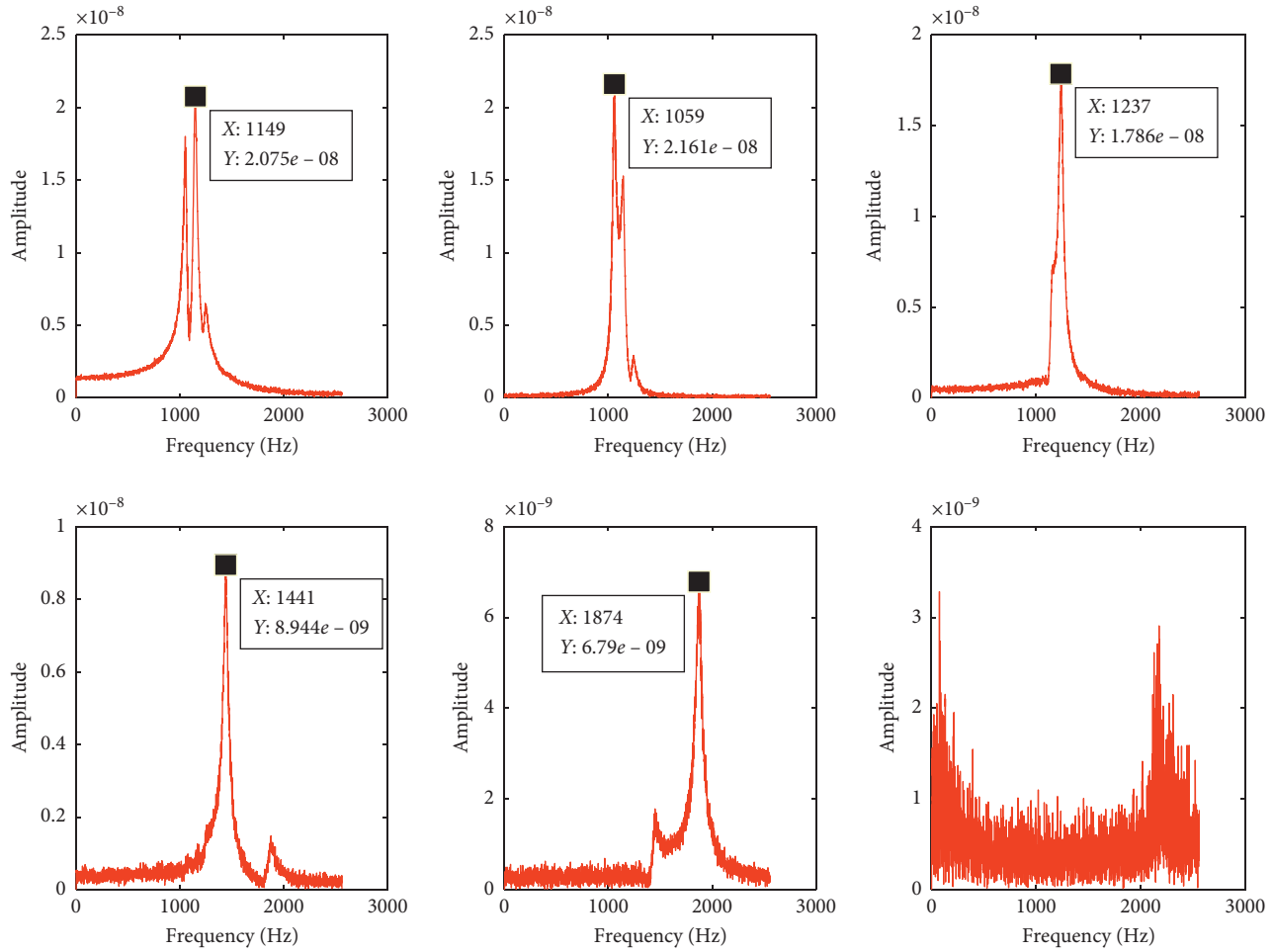


Figure 13: Modal frequencies identified by DMA-based Isomap.

Table 2: Comparison of natural frequencies with different assembly methods.

Orders	Frequency calculated by FEA (frequency/Hz)	Orders of Isomap components	Identified by LSMS-based Isomap (frequency/Hz)	Relative error (%)	Identified by DMA-based Isomap (frequency/Hz)	Relative error (%)
1	1054.9	2	1059	0.387	1059	0.387
2	1145.7	1	1145	−0.061	1149	0.288
3	1239.6	3	1237	−0.210	1237	−0.210
4	1441.9	4	1441	−0.062	1441	−0.062
5	1740.0	6	—	—	—	—
7	1871.7	5	1874	0.123	1874	0.123

The symbol "--" indicates that the result was not recognized or too small.

Table 3: Comparison of MAC of modal shapes with different assembly methods.

Order of real modal shape	Order of identified modal shape	*MAC* identified by the LSMS-based Isomap	*MAC* identified by DMA-based Isomap
1	2	0.8783	0.6564
2	1	0.6837	0.6180
3	3	0.5634	0.7817
4	4	0.4238	0.5241
5	6	0.0014	0.0049
7	5	0.6495	0.8932

Table 4: Comparison of modal shapes with different methods.

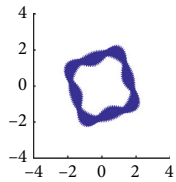
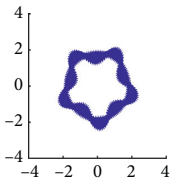
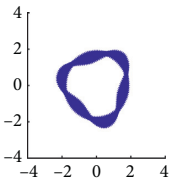
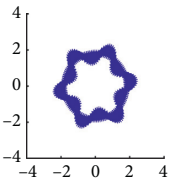
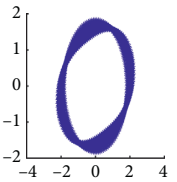
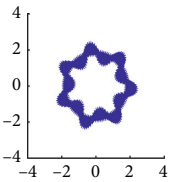
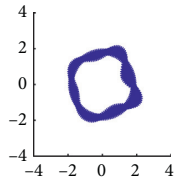
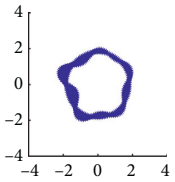
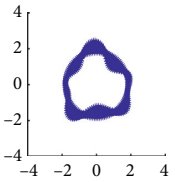
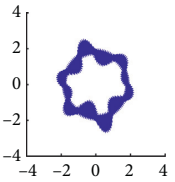
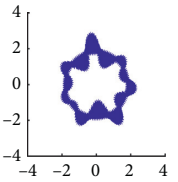
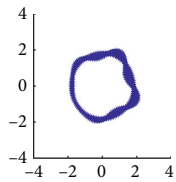
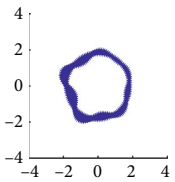
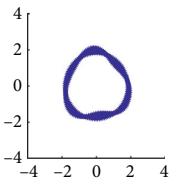
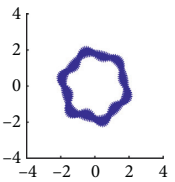
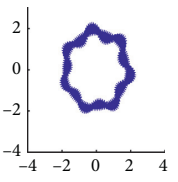
Method	Modal shapes when $K = 40$ and damping ratio is 0.03					
	1th order	2th order	3th order	4th order	5th order	7th order
FEA						
LSMS					—	
DMA					—	

Table 5: Comparison of modal ratios with different assembly methods.

Order of real modal ratios	Order of identified modal ratios	Identified by the LSMS-based Isomap	Identified by DMA-based Isomap
1	2	0.0204	0.0286
2	1	0.0187	0.0196
3	3	0.0270	0.0210
4	4	0.0216	0.0214
5	6	—	0.0678
7	5	0.0282	0.0277

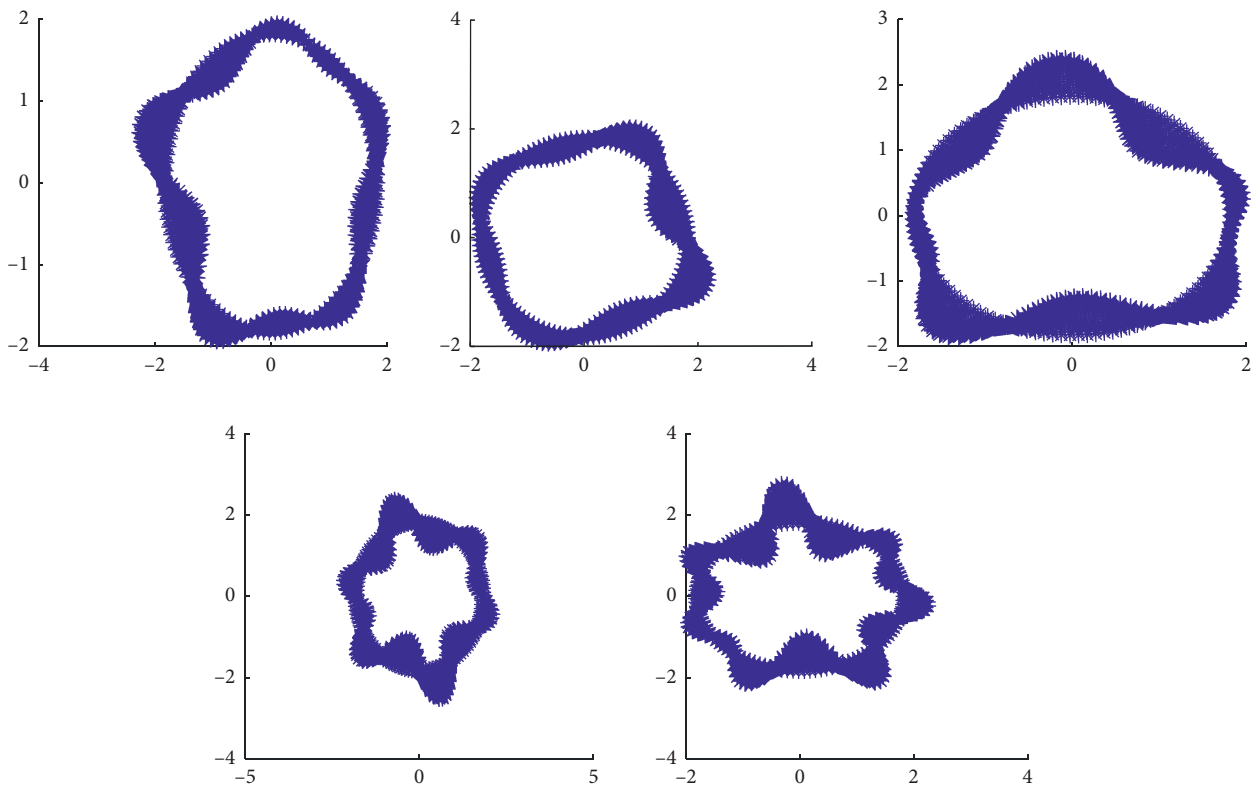


Figure 14: Modal shapes identified when the embedding dimension is 5.

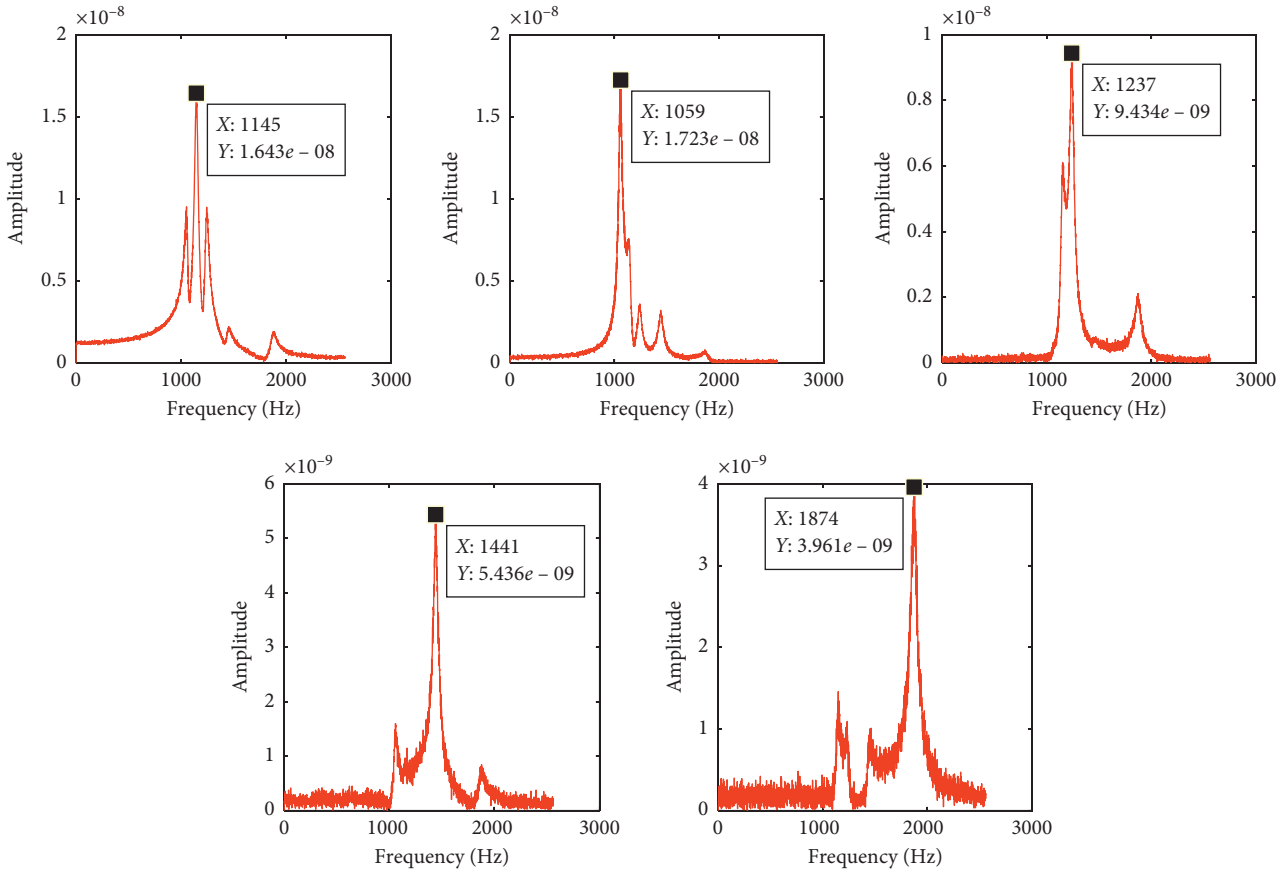


Figure 15: Modal frequencies identified when the embedding dimension is 5.

Table 6: Comparison of MAC at different dimensions.

FEA method	LSMS method			
	$d = 5$		$d = 6$	
Real order	Identify order	*MAC*	Order	*MAC*
1	2	0.8783	2	0.8783
2	1	0.6837	1	0.6837
3	3	0.5634	3	0.5634
4	4	0.4238	4	0.4238
5	6	—	6	0.0014
7	5	0.6495	5	0.6495

Table 7: Comparison of modal frequencies at different dimensions.

FEA method		LSMS method			
		$d = 5$		$d = 6$	
Real order	Real frequencies	Identify order	Frequencies	Order	Frequencies
1	1054.9	2	1059	2	1059
2	1145.7	1	1145	1	1145
3	1239.6	3	1237	3	1237
4	1441.9	4	1441	4	1441
5	1740.0	6	—	6	—
7	1871.7	5	1874	5	1874

Table 8: Comparison of MAC for different neighbor numbers K.

Real order		1	2	3	4	5	7
Identify order		2	1	3	4	6	5
$K = 12$	LSMS	0.8850	0.6863	0.5591	0.4179	0.0003	0.6544
	DMA	0.6609	0.6198	0.7790	0.5221	0.0002	0.8877
$K = 40$	LSMS	0.8783	0.6837	0.5634	0.4238	0.0014	0.6495
	DMA	0.6564	0.6180	0.7817	0.5241	0.0049	0.8932
$K = 200$	LSMS	0.8800	0.6851	0.5632	0.4261	0.0050	0.6534
	DMA	0.6599	0.6208	0.7820	0.5260	0.0042	0.8981
$K = 500$	LSMS	0.8798	0.6851	0.5635	0.4262	0.0519	0.6532
	DMA	0.6594	0.6206	0.7824	0.5263	0.0213	0.8990
$K = 800$	LSMS	0.8796	0.6851	0.5636	0.4265	0.1423	0.6543
	DMA	0.6585	0.6201	0.7827	0.5271	0.0724	0.9007

Table 9: Comparison of modal frequencies for different neighbor numbers K.

Real order		1	2	3	4	5	7
Identify order		2	1	3	4	6	5
$K = 12$	LSMS	1052	1146	1237	1441	—	1874
	DMA	1059	1145	1243	1441	—	1874
$K = 40$	LSMS	1059	1145	1237	1441	—	1874
	DMA	1059	1149	1237	1441	—	1874
$K = 200$	LSMS	1059	1145	1237	1443	—	1868
	DMA	1059	1149	1237	1443	—	1868
$K = 500$	LSMS	1059	1145	1237	1443	—	1868
	DMA	1059	1149	1237	1443	—	1868
$K = 800$	LSMS	1059	1146	1237	1443	—	1868
	DMA	1059	1149	1237	1443	—	1868
	FEA	1054.9	1145.7	1239.6	1441.9	1740.0	1871.7

Table 10: Comparison of modal ratios for different neighbor numbers K.

Real order		1	2	3	4	5	7
Identify order		2	1	3	4	6	5
$K = 12$	LSMS	0.0201	0.0186	0.0272	0.0216	0.0846	0.0294
	DMA	0.0286	0.0195	0.0211	0.0216	0.0668	0.0284
$K = 40$	LSMS	0.0204	0.0187	0.0270	0.0216	0.1140	0.0282
	DMA	0.0286	0.0196	0.0210	0.0214	0.0678	0.0277
$K = 200$	LSMS	0.0204	0.0187	0.0270	0.0213	0.0702	0.0275
	DMA	0.0285	0.0196	0.0210	0.0212	0.0317	0.0275
$K = 500$	LSMS	0.0204	0.0187	0.0271	0.0212	0.0560	0.0274
	DMA	0.0285	0.0196	0.0210	0.0211	0.0364	0.0274
$K = 800$	LSMS	0.0204	0.0188	0.0271	0.0212	0.0592	0.0274
	DMA	0.0286	0.0196	0.0210	0.0211	0.0359	0.0274

dimensional case. Therefore, in Isomap, the selected dimension has no effect on the identification of the modal parameters of the first few orders.

(4) Different K values have little effect on the recognition results (see Tables 8–10). As K increases, the identification accuracy of later orders increases. However, the time complexity of the algorithm also increases.

(5) Tables 11–13 indicate that the modal ratios can be effectively identified using both RDT and NExT. The different dimensionality reduction methods (PCA or MDS and Isomap) and assembly methods (LSMS and DMA) have little impact on the estimated modal ratios.

(6) Isomap-based OMA can effectively identify the three-dimensional modal parameters (see Tables 11–16). The 6th-order modal parameters are not extracted well by Isomap because of their weak nonlinear characteristics.

(7) Tables 11–19 illustrate that, as the real modal ratio increases, the modal shapes, frequencies, and modal ratios become more difficult to identify. This is because a larger damping ratio results in smaller structure responses.

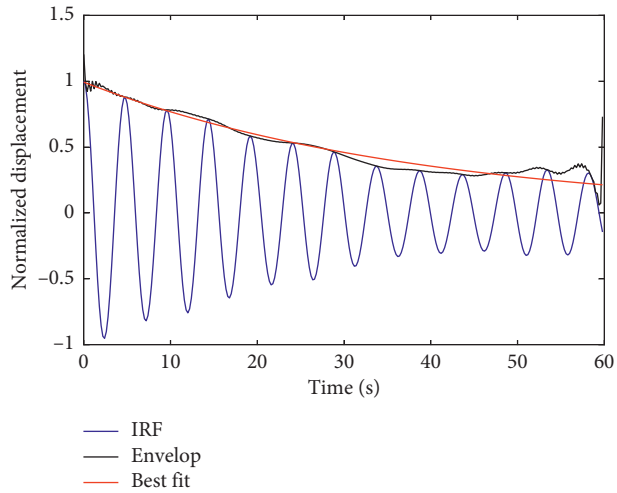


Figure 16: Example of the modal ratio solution process.

Table 11: RDT and NExT results with a real modal damping ratio of 0.1.

Real order	Identified order	Real modal ratio = 0.1							
		PCA (MDS)				Isomap			
		LSMS		DMA		LSMS		DMA	
		RDT	NExT	RDT	NExT	RDT	NExT	RDT	NExT
1	2	0.0532	0.0567	0.0587	0.0624	0.0510	0.0567	0.0605	0.0627
2	1	0.1155	0.1241	0.1045	0.1139	0.1136	0.1245	0.1074	0.1139
3	3	0.0668	0.0639	0.0650	0.0605	0.0682	0.0645	0.0653	0.0612
4	4	0.0716	0.0847	0.0869	0.0967	0.0770	0.0837	0.0956	0.0971
5	6	0.0742	0.0758	0.0748	0.0759	—	—	—	—
7	5	0.1079	0.1192	0.1308	0.1351	0.1369	0.1388	0.1473	0.1420

Table 12: RDT and NExT results with a real modal damping ratio of 0.05.

Real order	Identified order	Real modal ratio = 0.05							
		PCA (MDS)				Isomap			
		LSMS		DMA		LSMS		DMA	
		RDT	NExT	RDT	NExT	RDT	NExT	RDT	NExT
1	2	0.0312	0.0331	0.0312	0.0454	0.0297	0.0331	0.0373	0.0442
2	1	0.0716	0.0854	0.0585	0.0679	0.0825	0.0861	0.0549	0.0673
3	3	0.0388	0.0410	0.0543	0.0611	0.0391	0.0410	0.0524	0.0597
4	4	0.0274	0.0367	0.0625	0.0780	0.0298	0.0368	0.0619	0.0821
5	6	0.0434	0.0450	0.0361	0.0450	—	—	0.0364	0.0471
7	5	0.0492	0.0485	0.0393	0.0453	0.0499	0.0496	0.0416	0.0462

Table 13: RDT and NExT results with a real modal damping ratio of 0.03.

Real order	Identified order	Real modal ratio = 0.03							
		PCA (MDS)				Isomap			
		LSMS		DMA		LSMS		DMA	
		RDT	NExT	RDT	NExT	RDT	NExT	RDT	NExT
1	2	0.0185	0.0204	0.0213	0.0286	0.0181	0.0204	0.0206	0.0286
2	1	0.0183	0.0188	0.0160	0.0197	0.0167	0.0187	0.0163	0.0196
3	3	0.0230	0.0271	0.0212	0.0209	0.0230	0.0270	0.0212	0.0210
4	4	0.0168	0.0212	0.0149	0.0210	0.0160	0.0216	0.0163	0.0214
5	6	0.0270	0.0267	0.0269	0.0267	—	—	0.0550	0.0678
7	5	0.0275	0.0274	0.0290	0.0274	0.0298	0.0282	0.0297	0.0277

Table 14: PCA and Isomap by LSMS with a real modal damping ratio of 0.03.

		LSMS					
Real order	Identified order	PCA (MDS)			Isomap		
		MAC	Frequencies	Relative error (%)	MAC	Frequencies	Relative error (%)
1	2	0.8785	1058	0.294	0.8783	1059	0.389
2	1	0.6843	1145	−0.061	0.6837	1145	−0.061
3	3	0.5639	1237	−0.210	0.5634	1237	−0.210
4	4	0.4274	1442	0.007	0.4238	1441	−0.062
5	6	0.6282	1737	−0.172	0.0014	—	—
7	5	0.6564	1872	0.016	0.6495	1874	0.123

Table 15: PCA and Isomap by LSMS with a real modal damping ratio of 0.05.

		LSMS					
Real order	Identified order	PCA (MDS)			Isomap		
		MAC	Frequencies	Relative error (%)	MAC	Frequencies	Relative error (%)
1	2	0.6311	1059	0.389	0.6316	1065	0.957
2	1	0.3026	1143	−0.236	0.3019	1143	−0.236
3	3	0.1082	1232	−0.613	0.1088	1229	−0.855
4	4	0.0801	1437	−0.340	0.0788	1441	−0.062
5	6	0.0037	1747	0.402	0.0227	—	—
7	5	0.0219	1868	−0.198	0.0211	1879	0.390

Table 16: PCA and Isomap by LSMS with a real modal damping ratio of 0.1.

		LSMS					
Real order	Identified order	PCA (MDS)			Isomap		
		MAC	Frequencies	Relative error (%)	MAC	Frequencies	Relative error (%)
1	2	0.7248	1073	1.716	0.7256	1080	2.379
2	1	0.4890	1140	−0.498	0.4887	1137	−0.759
3	3	0.3430	1199	−3.275	0.3379	1178	−4.969
4	4	0.2435	1414	−1.935	0.2434	1391	−3.530
5	6	0.6293	1747	0.402	0.0002	—	—
7	5	0.2520	1854	−0.946	0.2710	1800	−3.831

Table 17: PCA and Isomap by DMA with a real modal damping ratio of 0.03.

		DMA					
Real order	Identified order	PCA (MDS)			Isomap		
		MAC	Frequencies	Relative error (%)	MAC	Frequencies	Relative error (%)
1	2	0.6554	1059	0.389	0.6564	1059	0.389
2	1	0.6180	1151	0.463	0.6180	1149	0.288
3	3	0.7833	1237	−0.210	0.7817	1237	−0.210
4	4	0.5284	1443	0.076	0.5241	1441	−0.062
5	6	0.7289	1738	−0.115	0.0049	—	—
7	5	0.9038	1873	0.069	0.8932	1874	0.123

Table 18: PCA and Isomap by DMA with a real modal damping ratio of 0.05.

		DMA					
Real order	Identified order	PCA (MDS)			Isomap		
		MAC	Frequencies	Relative error (%)	MAC	Frequencies	Relative error (%)
1	2	0.5395	1052	−0.275	0.5614	1052	−0.275
2	1	0.1402	1143	−0.236	0.1521	1140	−0.498
3	3	0.1928	1225	−1.178	0.1770	1237	−0.210
4	4	0.1213	1434	−0.548	0.1245	1441	−0.062
5	6	0.0005	1877	7.874	0.0003	1889	8.563
7	5	0.0041	1879	0.390	0.0044	1890	0.978

Table 19: PCA and Isomap by DMA with a real modal damping ratio of 0.1.

Real order	Identified order	DMA					
		PCA (MDS)			Isomap		
		MAC	Frequencies	Relative error (%)	MAC	Frequencies	Relative error (%)
1	2	0.5232	1081	1.716	0.5193	1081	2.474
2	1	0.4812	1044	−0.498	0.4804	1138	−0.672
3	3	0.2107	1185	−4.405	0.2033	1176	−5.131
4	4	0.1703	1406	−2.490	0.1788	1392	−3.461
5	6	0.7279	1747	0.402	0.0003	—	—
7	5	0.4739	1854	−0.946	0.5082	1889	0.924

5. Conclusion and Prospects

In this paper, we have described the application of the Isomap algorithm to OMA to identify modal shapes, modal natural frequencies, and modal ratios of three-dimensional structures. Promising results were obtained from simulations of a cylindrical shell emitting nonlinear response patterns. We also compared the influence of various parameters. The simulation results show that the parameters of the Isomap algorithm (number of neighbors K and dimension d) have little effect on the results. However, the structural parameters (damping ratio) and matrix assembly method have a significant influence on the output. Furthermore, we compared the results of PCA (MDS) with Isomap.

In future work, we will attempt to improve the accuracy of our method at higher damping ratios. The method discussed in this paper uses the most basic Isomap algorithm. Improved versions of Isomap may achieve better results. Furthermore, the experiment with the actual structure and the problem of finding the missing mode are worth studying.

Data Availability

All the data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Grant no. 51305142) and the Promotion Program for Young and Middle-Aged Teachers in Science and Technology Research at Huaqiao University under Grant ZQN-PY212.

References

[1] M. Ghalishooyan and M. Abdelghani, "Output-only damping estimation of structures using enhanced frequency domain decomposition method," *Modares Civil Engineering Journal*, vol. 18, no. 2, pp. 183–193, 2018.

[2] H.-F. Lam, J. Hu, and J.-H. Yang, "Bayesian operational modal analysis and Markov chain Monte Carlo-based model updating of a factory building," *Engineering Structures*, vol. 132, pp. 314–336, 2017.

[3] G. P. Martindale, "Structural identification and assessment of the inverted tee girder bridge system," MS thesis, University of Nebraska-Lincoln, Lincoln, NE, USA, 2018.

[4] H. Zhou, K. Yu, Y. Chen, R. Zhao, and Y. Wu, "Output-only modal estimation using sparse component analysis and density-based clustering algorithm," *Measurement*, vol. 126, pp. 120–133, 2018.

[5] K. Yu, K. Yang, and Y. Bai, "Estimation of modal parameters using the sparse component analysis based underdetermined blind source separation," *Mechanical Systems and Signal Processing*, vol. 45, no. 2, pp. 302–316, 2014.

[6] J. Wang, C. Wang, T. Zhang, and B. Zhong, "Comparison of different independent component analysis algorithms for output-only modal analysis," *Shock and Vibration*, vol. 2016, Article ID 6309084, 25 pages, 2016.

[7] W. K. Härdle and Z. Hlávka, "Principal component analysis," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 163–183, 2015.

[8] X. Li, C. Cai, and J. He, "Density-based multi-manifold ISOMAP for data classification," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 897–903, IEEE, Kuala Lumpur, Malaysia, December 2017.

[9] X. He, S. Yan, Y. Hu et al., "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[10] S. Bai, X. Bai, L. J. Latecki et al., "Multidimensional scaling on multiple input distance matrices," *AAAI*, pp. 1281–1287, 2017.

[11] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*, pp. 237–280, Springer, New York, NY, USA, 2013.

[12] C. Wang, J. Gou, J. Bai et al., "Modal parameter identification with principal component analysis," *Journal of Xi'an Jiaotong University*, vol. 47, no. 11, pp. 97–104, 2013.

[13] C. Wang, W. Guan, J. Gou et al., "Principal component analysis based three-dimensional operational modal analysis," *International Journal of Applied Electromagnetics and Mechanics*, vol. 45, no. 1–4, pp. 137–144, 2014.

[14] C. Wang, J. Wang, and Y. Zhang, "Operational modal analysis of three-dimensional structures by second-order blind identification and least square generalized inverse," *Journal of Vibroengineering*, vol. 19, no. 4, pp. 2857–2872, 2017.

[15] J. P. Noël and G. Kerschen, "Nonlinear system identification in structural dynamics: 10 more years of progress," *Mechanical Systems and Signal Processing*, vol. 83, pp. 2–35, 2016.

[16] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[17] C. Sun, P. Wang, R. Yan, R. X. Gao, and X. Chen, "Machine health monitoring based on locally linear embedding with kernel sparse representation for neighborhood optimization,"

Mechanical Systems and Signal Processing, vol. 114, pp. 25–34, 2019.

[18] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[19] S. W. Choi, C. Lee, J.-M. Lee, J. H. Park, and I.-B. Lee, "Fault detection and identification of nonlinear processes based on kernel PCA," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, no. 1, pp. 55–67, 2005.

[20] S. K. Prabhakar and H. Rajaguru, "Comparison of Isomap and matrix factorization with mahalanobis based sparse representation classifier for epilepsy classification from EEG signals," in *Proceedings of the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 580–583, IEEE, Dhaka, Bangladesh, December 2017.

[21] J. Bai, G. Yan, and C. Wang, "Modal identification method following locally linear embedding," *Journal of Xi'an Jiaotong University*, vol. 47, no. 1, 2013.

[22] J. Zhang, L. Dong, Y. Guirong et al., "Improved LLE algorithm and its application," *International Journal of Applied Electromagnetics and Mechanics*, vol. 52, no. 1-2, pp. 685–690, 2016.

[23] L. Dong, C. Hao, and J. Zhang, "Modal identification and influence factors of LLE algorithm," *Structure & Environment Engineering*, vol. 44, no. 05, pp. 41–46, 2017.

[24] W. Guan, J. M. Dong, Y. Han, and J. Zhou, "Data-driven methods for operational modal parameters identification: a comparison and application," *Measurement*, vol. 132, pp. 238–251, 2019.

[25] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[26] K.-K. Xu, H.-X. Li, and Z. Liu, "ISOMAP-based spatiotemporal modeling for lithium-ion battery thermal process," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 569–577, 2018.

[27] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7a–7, 2002.

[28] T. Zhang, C. Wang, and Y. Zhang, "Three-dimensional operational modal analysis based on self-iteration principal component extraction and direct matrix assembly," *Journal of Vibroengineering*, vol. 19, no. 8, 2017.

[29] K. A. Basov, "ANSYS and LMS virtual Lab," *Geometric Modeling*, p. 240, 2006.

[30] S. R. Ibrahim, "Random decrement technique for modal identification of structures," *Journal of Spacecraft and Rockets*, vol. 14, no. 11, pp. 696–700, 1977.

[31] G. H. James, T. G. Carne, and J. P. Lauffer, "The natural excitation technique (NExT) for modal parameter extraction from operating structures," *Modal Analysis-The International Journal of Analytical and Experimental Modal Analysis*, vol. 10, no. 4, p. 260, 1995.

Hindawi
Scientific Programming
Volume 2020, Article ID 8842463, 10 pages
https://doi.org/10.1155/2020/8842463

Research Article

How to Construct a Power Knowledge Graph with Dispatching Data?

Shixiong Fan,[1] Xingwei Liu,[1] Ying Chen,[2] Zhifang Liao,[2] Yiqi Zhao,[2] Huimin Luo,[2] and Haiwei Fan[3]

[1]*Beijing Key Laboratory of Research and System Evaluation of Power Dispatching Automation Technology, China Electric Power Research Institute, Haidian District, Beijing 100192, China*
[2]*School of Computer Science and Engineering, Central South University, Hunan 410000, China*
[3]*State Grid Fujian Electric Power Co., Ltd., Fuzhou 350003, China*

Correspondence should be addressed to Ying Chen; 1074647728@qq.com and Zhifang Liao; zfliao@csu.edu.cn

Received 24 April 2020; Revised 7 May 2020; Accepted 26 May 2020; Published 14 July 2020

Academic Editor: Chenxi Huang

Knowledge graph is a kind of semantic network for information retrieval. How to construct a knowledge graph that can serve the power system based on the behavior data of dispatchers is a hot research topic in the area of electric power artificial intelligence. In this paper, we propose a method to construct the dispatch knowledge graph for the power grid. By leveraging on dispatch data from the power domain, this method first extracts entities and then identifies dispatching behavior relationship patterns. More specifically, the method includes three steps. First, we construct a corpus of power dispatching behaviors by semi-automated labeling. And then, we propose a model, called the BiLSTM-CRF model, to extract entities and identify the dispatching behavior relationship patterns. Finally, we construct a knowledge graph of power dispatching data. The knowledge graph provides an underlying knowledge model for automated power dispatching and related services and helps dispatchers perform better power dispatch knowledge retrieval and other operations during the dispatch process.

1. Introduction

Smart grids have made important progress in the research and integration of dispatch automation systems [1]. According to the relevant dispatch documents such as the power dispatching control rules and the experience of the dispatcher, together with the dispatching system data and the operation state of the power grid, the dispatcher judges whether the current operation state of the power system needs dispatching and what kind of dispatching behavior is to be executed. In the actual power dispatching scenarios, power dispatching tasks are still highly dependent on the dispatcher's business knowledge and dispatching experience. Most dispatchers only understand local business knowledge [2] and cannot effectively respond to other dispatching business or global business.

With the continuous integration of multiple dispatching services, the expertise of experts or dispatchers also needs to be integrated to meet the needs of simultaneously handling complex multiservice power dispatching problems. To provide dispatcher reference for dispatchers, digital power experts have written the texts of power dispatch, which summarize and contain all aspects of the global power dispatch business. Hence, studying knowledge organization methods for global dispatching texts, building knowledge models based on multiple dispatching behaviors, and implementing knowledge expressions that flexibly and clearly express business logic will help to improve the degree of automation of dispatching systems and provide global knowledge support for intelligent grid dispatching.

The text of power dispatching is characterized by knowledge-intensive and abundant knowledge types, and it is a kind of unstructured data. Compared with structured data with strict format and specification, the expression mode of power dispatching text is more flexible and is more difficult to read and understand. So, it is necessary to explore a natural language processing method for the dispatching text and a behavioral knowledge organization method that is suitable for the characteristics of the power dispatching behavior.

To solve these problems, we use the knowledge graph to organize knowledge of power dispatching behavior. Traditional relational databases face the problems of repeated data, weak data relationships, and difficult updates. Comparatively, the knowledge graph organizes knowledge in a graph topology, which is more in line with the structure of the power system. It regards relationships as an important knowledge element, which can better describe knowledge entities and their relationships such as power environment, dispatching roles, and power dispatching behavior. Its knowledge storage and retrieval are more flexible. First, we analyze the text of power dispatching behavior, combine natural language processing technology to build a power domain dictionary, mine power domain phrases, and identify entities based on domain phrases and domain dictionaries to achieve coreference resolution. Then, we analyze the text characteristics of dispatching behavior, define and organize the relationship of power dispatching behavior, and use the graph structure to store entities and relationships, thereby constructing a knowledge graph of power dispatching behavior.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the constructing method of knowledge graph of power dispatching behavior. Section 4 introduces the experiment and evaluation results, data set, experimental design, experimental details, and experimental results. Section 5 introduces an experimental summary and future work outlook.

2. Related Work

Google first proposed the concept of the knowledge graph in 2012 [3], which can formally describe things and their related relationships in the real world. The knowledge graph uses <entity, relationship, entity> triples to store knowledge, and it uses entities as nodes, relationships as edges to build a knowledge network, which conforms to the behavior rules of general subjects, actions, and action objects, and uses graph structures to describe the relationship. At present, many well-known knowledge graph projects organize a large amount of data, extract knowledge from them for organization and management, and provide users with high-quality intelligent services, such as understanding the semantics of search and providing more accurate search answers. In recent years, due to the development of crowdsourcing [4] and open-source ecosystems [5–7], the related research of constructing knowledge graphs by crowdsourcing and knowledge graphs in software has become a new research topic in the field of knowledge graphs, which also shows that knowledge graphs are flexible to organize domain knowledge.

Entities are the basic units in the knowledge graph, including attributes, attribute values, and the correspondence between related entities. Wang [8] designed a named entity recognition system based on text structure features. This method needs to design the characteristics of text in different fields separately, which is not universal. As the scale of data grows, the study on multilayer architecture and deep learning is extraordinarily important and necessary [9]. In order to reduce manual rules and improve the generalization capabilities of the model, many methods based on deep learning have been used for named entity recognition in recent years. For example, Lample et al. [10] proposed a neural network structure based on bidirectional LSTM and CRF for entity recognition. This method does not rely on artificial features and domain-specific knowledge and has excellent versatility. In the same period, Chiu and Nichols [11] proposed a bidirectional LSTM and CNN hybrid model for automatic detection of the word and character-level features, eliminating the need for most feature engineering. To overcome the problem of missing Chinese information when using the English relationship extraction method, Han et al. [12] generated a large-scale Chinese relationship extraction data set based on the Chinese Encyclopedia and proposed an attention model based on the entity character features in the Chinese relationship extraction method. Leng and Jiang [13] proposed an improved SDAE model for entity-relationship extraction. This method eliminates the need to annotate the relationship manually and can extract the relationship between the entities automatically through contextual features.

There are many types of power grid equipment, and the relationship between the types of equipment is complicated. There is little research on knowledge graphs in the field of power grids. Tang [14] merged existing multisource heterogeneous power equipment related data to construct a power equipment knowledge graph to improve the data storage efficiency and extraction process. Given the problem of flattening and efficient utilization of power asset information, Yang [15] proposed the general process of constructing professional knowledge graphs in the power field and proposed a multisource heterogeneous power asset information fusion method based on knowledge fusion. Li et al. [2], based on the underlying data and business logic data of the smart grid dispatching control D5000 system, used a top-down and bottom-up method to construct a dispatch knowledge graph. These power knowledge graph construction methods based on power system data such as dispatch management systems face several challenges: huge data size, complex data types, diverse knowledge content, low-value density, and low data quality. Therefore, the research on text data with higher value density has attracted much attention in recent years. Wang et al. [16] proposed a method of entity recognition, coreference resolution, and relationship extraction based on the records of defects in electrical equipment and automatically constructed the knowledge graph of defects in electrical equipment to improve the retrieval quality of the records of defects in electrical equipment. Existing research usually processes unstructured text according to different text types and modes in their application scenarios, and most of them are not universal. Hence, it is necessary to study and design a knowledge graph construction method for power dispatching behavior based on power dispatching text.

3. Methodology

In this section, this paper will introduce the knowledge graph constructing method of power dispatching behavior, and its technical roadmap is shown in Figure 1. First, a

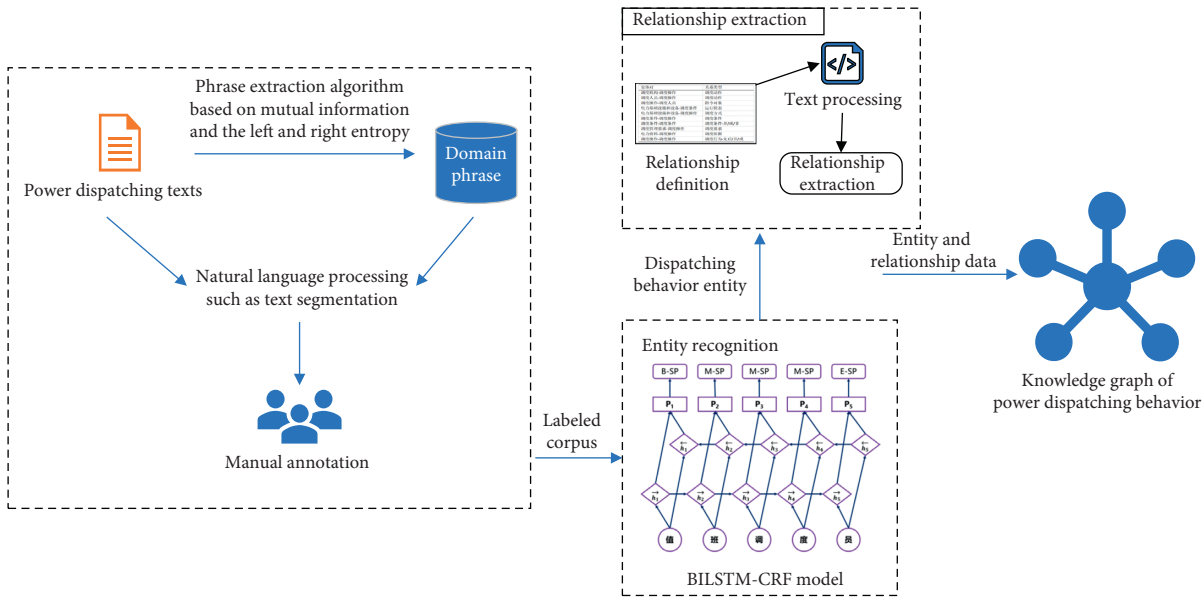


Figure 1: Technical roadmap for constructing a knowledge graph of power dispatching behavior.

phrase extraction algorithm based on mutual information and left and right entropy is used to extract power domain phrases, construct a professional dictionary for power dispatch, and prepare a corpus of power dispatch behavior. Then, the BiLSTM-CRF model is built to train labeled data and identify and extract entities in the power dispatching domain. Finally, by analyzing and summarizing the entity relationships, the power dispatching behavior relationships are extracted, and a graph database is used to store and construct a knowledge graph structure.

3.1. Corpus Construction. Constructing a high-quality domain text corpus is a prerequisite for acquiring knowledge entities of power dispatching behavior. However, there is a lack of labeled data in the field of power dispatching, and manual labeling consumes much energy. It is affected by the complexity of the domain entity category and the professionalism of the labeling personnel. Therefore, this paper uses a phrase extraction algorithm based on mutual information and left and right entropy to get candidate phrases and selects and annotates them manually to get high-quality annotation data.

3.1.1. Phrase Extraction Algorithm Based on Mutual Information and Left and Right Entropy. Most of the grid dispatching entities are nested combinations of multiple words. Therefore, in the traditional corpus labeling process, the original corpus must first be segmented to clarify the boundary of the words, which is convenient for manual labeling later. Existing word segmentation toolkits, such as jieba word segmentation tools, mostly use dictionary-based word segmentation methods. The dictionaries used are cross-domain general dictionaries, most of which commonly used vocabularies and lack professional vocabulary in the power field. For example, the word "operation instruction ticket" will be divided into three words "operation," "instruction," and "ticket" when using the general dictionary. If we use the original corpus in the field of power dispatching to directly segment words, the effect of this method is not satisfactory. Therefore, in the stage of the cold start of corpus labeling, in order to obtain labeled corpus, this paper uses a novel unsupervised word discovery algorithm, which is a phrase extraction algorithm based on mutual information and left and right entropy.

The algorithm first calculates the mutual information between the words in the corpus. The formula is as follows:

$$\mathrm{PMI}x, y = \log_2 \frac{p(x, y)}{p(x)p(y)}. \tag{1}$$

In Formula 1, $p(x, y)$ is the probability of two words appearing, and $p(x)$ is the probability of a single word appearing. We use specific examples to explain this. There are three dispatching behavior words: "Provincial Dispatching," "Dispatcher on duty," and "Provincial Dispatcher on duty." If the word frequency of "Provincial Dispatching" is 10, the word frequency of "Dispatcher on duty" is 20, and the word frequency of "Provincial Dispatcher on duty" is 5, the total number of words is N, and the total number of double words is M, then we have the following formula:

$$\begin{aligned}&\mathrm{PMI\ Provincial\ Dispatching, Dispatcher\ on\ duty}\\ &= \log_2 \frac{5/M}{(10/N) * (20/N)}.\end{aligned} \tag{2}$$

The mutual information can reflect the relationship between two words well. The higher the mutual information value is, the higher the correlation between X and Y is, the more likely X and y are to form phrases. On the contrary, the lower the mutual information value is, the lower the correlation between X and Y is, the more likely there is a phrase boundary between X and y.

Mutual information indicates the relevance of the two words. Also, we need to calculate the degree of freedom of the word. The degree of freedom refers to the degree of diversity of adjacent words that appear on the left and right sides of the word. If the left and right sides of a candidate word are different words in different sentences, the smaller the connection between the word and other words, the greater the internal connection between the candidate words, that is, the greater the possibility that the candidate words have boundaries and are a single word.

We use the left and right entropy to measure the degree of freedom. Entropy can describe information uncertainty. In information theoretic learning, correntropy has been a widely used nonlinear similarity measure method due to its robustness [17]. The larger the left entropy and right entropy of a candidate, the more uncertain the words that may appear on the left and right sides of the candidate and the higher the degree of freedom. The formula for calculating the left and right entropy is as follows:

$$\begin{aligned} E_L(W) &= -\sum_{\forall a \in A} P(aW \mid W) \cdot \log_2 P(aW \mid W), \\ E_R(W) &= -\sum_{\forall b \in B} P(Wb \mid W) \cdot \log_2 P(Wb \mid W). \end{aligned} \tag{3}$$

Taking the left entropy as an example, suppose that the "Dispatcher on duty" has several kinds of collocations: "National Dispatcher on duty," "Province Dispatcher on duty," and "Temporary Dispatcher on duty," then the left entropy of the word "Dispatcher on duty" is as follows:

$$\begin{aligned} & -E_L(\text{Dispatcher on duty}) \\ &= P(ND \mid \text{Dispatcher on duty}) \\ &\quad \cdot \log_2 P(ND \mid \text{Dispatcher on duty}) \\ &\quad + P(PD \mid \text{Dispatcher on duty}) \\ &\quad \cdot \log_2 P(PD \mid \text{Duty Officer}) \\ &\quad + P(TD \mid \text{Dispatcher on duty}) \\ &\quad \cdot \log_2 P(TD \mid \text{Dispatcher on duty}). \end{aligned} \tag{4}$$

The final input result is the score of a series of words. The calculation formula of the score is as follows:

$$\text{score} = \text{PMI} + \min(\text{left entropy, right entropy}). \tag{5}$$

These scores are sorted from high to low. We add the top 100 words to the jieba word segmentation dictionary and then perform word segmentation processing on the original corpus text to facilitate the manual labeling of the role of words in the later period.

3.1.2. Manual Annotation. Due to the fuzzy boundary of Chinese words and a large number of cross-nesting structures in the grid dispatching entity, the complexity of the identification task increases. Furthermore, the data set in this paper contains multiple categories of entities. Consequently, according to the word segmentation results obtained by the unsupervised phrase extraction method, the word segmentation results need to be returned to the original corpus after manual inspection. Then, we use the BMESO labeling mechanism to convert it to the input format required by the model and finally get a labeled Training data set. The definition of the BMESO annotation model is shown in Table 1.

The labeling tool we use is YEDDA. For named entities in the field of power dispatching, we have summarized many dispatch documents and dispatch glossary classification methods. Then, a number of collaborators form a team to collaborate to review and determine the dispatching behavior entities and finally they are classified as follows:

(1) Scheduling mechanism (SM): including China's five major power generation groups, regional power generation groups, State Grid Corporation of China, Regional Power Grid Corporation, management organizations, and departments at all levels

(2) Scheduling personnel (SP): including leaders of various organizations, technical personnel at all levels, and dispatching personnel on duty at all levels

(3) Scheduling operation (SO): including but not limited to dispatching operation related to the protection device

(4) Facilities (Fac): such as transformer, bus, line, circuit breaker, switch, knife gate, protection device, primary equipment, secondary equipment, electrical equipment, boiler equipment, steam (water, gas) turbine equipment, power transmission equipment, transmission equipment, converter equipment, power system, chemical treatment, and fuel transportation

(5) Management requirements (MR): including scheduling management scope (equipment name), scheduling management mode, and scheduling instructions

(6) Electric power data (EPD): such as power-related documents, systems, and operation tickets

(7) Scheduling condition (SC): the objective conditions for certain dispatching under the power performance, such as the conditions for the power outage and power stations or substations on both sides of the line

(8) Equipment state (ES): such as operation, maintenance, standby, charging, power transmission, power failure, and other equipment states

3.2. Entity Extraction. In the past few years, the rapid development of machine learning has attracted the attention of many researchers [18]. In order to identify and extract knowledge entities, this paper uses a Bidirectional Long Short-Term Memory (BiLSTM) model and Conditional Random Fields (CRF) model as a named entity recognition model. We use the annotated data of the annotated corpus above for model training and extraction of knowledge entities in the field of power dispatching behavior. The BiLSTM model is composed of forward LSTM and backward LSTM. The LSTM model can memorize the long-term dependence of sentences from front to back, but it cannot encode information from back to front. Compared with a single LSTM model, the BiLSTM model can obtain bidirectional semantic

Table 1: The definition of the BMESO annotation model.

Annotation	Meaning
B	The first word of the entity
M	The internal words of the
E	The suffix of the entity
S	Single entity words
O	Nonphysical constituent words

dependence and obtain more comprehensive text information. However, the BiLSTM model does not guarantee that the prediction results obtained at each output layer are correct, and some prediction results that do not meet the constraints of the training set may appear. Therefore, the CRF model can be introduced to learn the constraining rules, thereby reducing the output of the model the probability of an illegal sequence. The annotated corpus constructed above prepares for the building of an entity recognition model. At the same time, annotated data is used for model training to identify entities with domain knowledge of power dispatching behavior.

The BiLSTM + CRF model is mainly composed of three layers, and the schematic diagram of the model is shown in Figure 2. The first layer is the embedding layer. The word vector is trained by inputting the pretrained character vector and word vector, and the dictionary obtained in the previous corpus labeling process is added to make the generated word vector more capable of expressing semantics.

The second layer in the middle is the forward and backward LSTM layer. In order to make full use of word meaning and word order information, the input sequence of the character vector and the word vector of the matching dictionary are subjected to feature fusion through network calculation.

The BiLSTM layer automatically extracts sentence features, uses the char embedding sequence $(x_1, x_2, x_3, \ldots, x_n)$ of each word in a sentence as the input of each time step of Bi-LSTM, and then uses the hidden state sequence $(\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}, \ldots, \overrightarrow{h_n})$ output by the forward LSTM and the reverse LSTM $(\overleftarrow{h_1}, \overleftarrow{h_2}, \overleftarrow{h_3}, \ldots, \overleftarrow{h_n})$. The hidden state output at each position is stitched by a position to obtain a complete hidden state sequence: $(\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3}, \ldots, \overrightarrow{h_n}) \in \mathbb{R}^{n*m}$.

The output of this layer is the score of each label of a word by selecting the highest label score as the label of the word.

Finally, the CRF layer is introduced for sentence-level sequence annotation. The parameter of the CRF layer is a $(k+2) \times (k+2)$ matrix A, k is the number of labels in the label set, and Aij represents the transfer score from the i-th label to the j-th label. When labeling a location, you can use the label that has been labeled before. The reason for adding 2 is to add a start state to the beginning of the sentence and an end state to the end of the sentence. Adding the CRF layer can consider the order between the labels of the output words of the Bi-LSTM layer, adding some constraints to the last predicted label to ensure that the predicted label is legal.

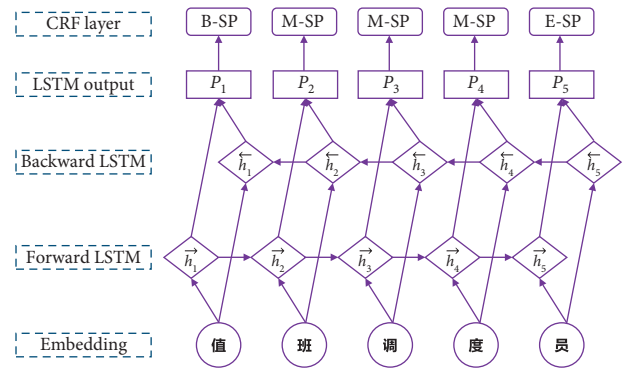


Figure 2: Schematic diagram of the BiLSTM-CRF model.

This paper introduces the Dropout mechanism to prevent overfitting. The Dropout mechanism prevents overfitting by randomly deleting hidden neurons in the network with a certain probability. The neurons in the input layer and the output layer of the network remain the same. In this way, the hidden neurons deleted in each iteration cycle are different, which increases the randomness of the network and improves the generalization ability of the network. The model code is shown in Algorithm 1.

3.3. Relationship Extraction. In order to mine the relationship of power dispatching behavior, this paper needs to analyze the language characteristics of the relationship description of power dispatching text. Since the power dispatching text is an unstructured natural language text written in Chinese, it has the characteristics of the Chinese language grammar and the power field. Its specific characteristics are as follows.

The sentence contains a large number of power domain entities. In a sentence related to scheduling behavior, there may be three or more behavior subjects and objects at the same time. The relationship network formed by the relationship between any two entities in the sentence is complicated. However, the entity-relationship category is relatively straightforward, and the entity relationships between the restricted entity categories mostly belong to one category.

Each sentence in the dispatching text corresponds to a dispatch behavior, and each segment corresponds to a type of dispatch scenario with various types. Understanding the dispatch statement requires professional knowledge of electricity, and it is difficult for nonprofessionals to learn. The Chinese grammatical structure is more flexible and sophisticated than English, with many grammatical phenomena such as condition, sequence, causality, and passive. Different writers have different language habits, and different scheduling behaviors will also use different expressions. At present, there is a lack of available syntactic knowledge rule base in the field of electric power.

The dispatching text which is the basis of dispatching behavior is based on the summary of real-world dispatching behaviors. The content is refined, the data volume is

Input: self.
Output: Trained model.
(1) Initialize the model.
(2) Define the Embedding layer.
(3) Add the Embedding layer to the model.
(4) Add forward LSTM to the model//units = 128, return_sequences = True.
(5) Add Dropout.
(6) Add backward LSTM to the model//units = 64, return_sequences = True.
(7) Add Dropout.
(8) Add TimeDistributed layer to the model.
(9) Define the CRF layer and Add the CRF layer to the model.
(10) Parameter status of each layer of the output model.
(11) Return model.

Algorithm 1: How to build BiLSTM + CRF named entity recognition model.

inadequate, and there is a lack of labeled data. The characteristics of multiple entities in the sentence make it challenging to label entity-relationship data. Machine learning algorithms commonly used in the general field often require large amounts of labeled data and cannot be directly applied to power dispatch texts.

Based on the above characteristics, we define the types of power dispatching behavior relations, as shown in Table 2. In the knowledge graph, the edges representing the relationship have directions, and the relationship edges in different directions may have different relationship types.

According to the above definition, most of the two entities have only one type of relationship. If two entities appear in a general sentence and their entity type meets the predefined relationship, it can be considered that there is a predefined relationship between the two entities. To extract the entity relationship, if there are multiple entities of the same type in a sentence, there may be a special relationship between these entities, such as a union. When analyzing power dispatching behavior sentences, words such as "common," "parallel," "and," and "or" are often used in the sentence to express the order, parallel, and other relationships. If there are related words in the sentence that represent particular sentence patterns such as juxtaposition, negation, and time, it can be determined that the sentence has a special relationship and a particular relationship type. For the dispatcher and dispatch operation entity, the relationship type between the two types of entities is judged according to the position characteristics of the entity in the sentence. If the dispatcher entity is before the dispatch operation entity, the relationship arrow is directed by the dispatcher to the dispatch operation. Otherwise, the relationship arrow is determined by the dispatch. The operation is directed to the dispatcher. Therefore, this paper sorts out and extracts the entity relations of power dispatching behavior.

3.4. Knowledge Graph Construction and Retrieval. After extracting power dispatching behavior entities and relationships, we use a graph database to store entity and attribute information and rely on entity relationships to connect directed edges between entity nodes, thereby constructing a knowledge graph structure. We use a graph database query language to provide a retrieval method based on knowledge graphs. Neo4j database is one of the more popular graph databases, with good performance and a friendly user interface. We use the Neo4j database as a storage database to construct a knowledge graph for power dispatching and use the declarative graph query language Cypher provided by the Neo4j database for knowledge graph retrieval.

4. Experiment

Based on the knowledge graph construction method proposed above, this paper presents the experimental work of labeling corpus construction, knowledge entity extraction, and knowledge graph construction of power dispatching behaviors with the power dispatch text data set. In this section, we will detail the experimental design, experimental details, and experimental results.

4.1. Data Sets and Data Preprocessing. In this paper, we crawled 29 documents related to power dispatching behavior such as power grid dispatching procedures, basic knowledge of dispatching, and disposal plans of dispatching failure. These documents were written by professional power dispatchers, and these documents fully describe the power dispatching business process, dispatching requirements, and dispatching behavior of dispatchers in the dispatching process. In this paper, the above documents are used as the original corpus for entity extraction and knowledge graph construction experiments. In order to facilitate the follow-up work, we unify the document format, remove the spaces and numbers in the document, and leave only character-type data.

4.2. Experiments and Result Analysis

4.2.1. Construction of Power Dispatching Behavior Annotated Corpus and Entity Extraction. There are a large number of unlabeled entity vocabularies in the field of power grid

Table 2: Predefined types of power dispatching behavior relations.

Entity pairs	Relational type
Scheduling mechanism—scheduling operation	Scheduling action
Scheduling personnel—scheduling operation	Scheduling action
Scheduling operation—scheduling	Instruction object
Personnel facilities—scheduling condition	Running state
Facilities—scheduling operation	Scheduling mode
Scheduling condition—scheduling operation	Scheduling condition
Scheduling condition—scheduling condition	Scheduling condition—and/or/not
Management requirements—scheduling operation	Scheduling requirements
Electric power data—scheduling operation	Scheduling basis
Scheduling operation—scheduling operation	Scheduling behavior—order/and/or

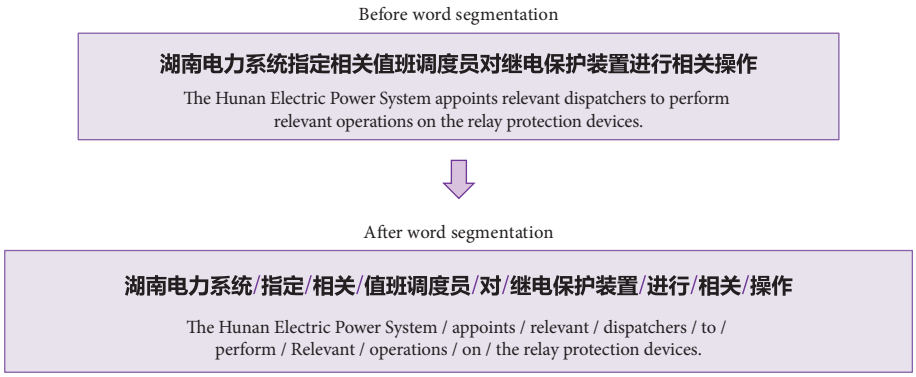


Figure 3: Partial word segmentation results of power dispatch text.

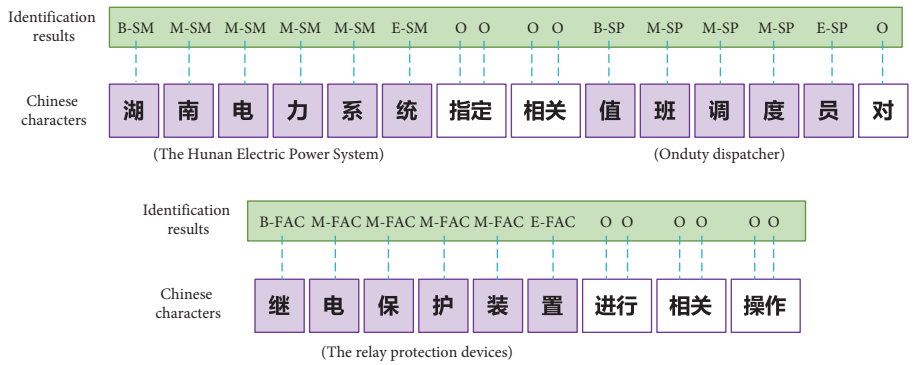


Figure 4: Partial recognition results of power dispatching behavior entities.

dispatching in the obtained power dispatch text data set. Due to professional domain issues, these documents have no distinct word boundaries. Then, we use a phrase extraction algorithm based on mutual information and left and right entropy to extract domain words and use the extracted domain words as a custom dictionary of Chinese words segmentation tool named "jieba" to assist in document segmentation. As can be seen from the word segmentation results in Figure 3, the use of the phrase extraction algorithm can improve the quality of word segmentation and separate the professional vocabulary in the power field such as Hunan Power System and Relay Protection.

According to the entity category of power dispatching behavior defined in this paper, we complete the construction of the labeled corpus of power dispatching behavior by manually labeling the corpus after word segmentation. Shown in Figure 3, we use the code to build the BiLSTM-CRF model, using an annotated corpus as the training set, to realize the entity recognition of text for power dispatching behavior. The recognition effect of the final model is shown in Figure 4. The entity extraction method in this paper can extract the entity vocabulary of power dispatching behavior from the power dispatching sentence and classify the entities. It can be seen that the entity extraction method in this

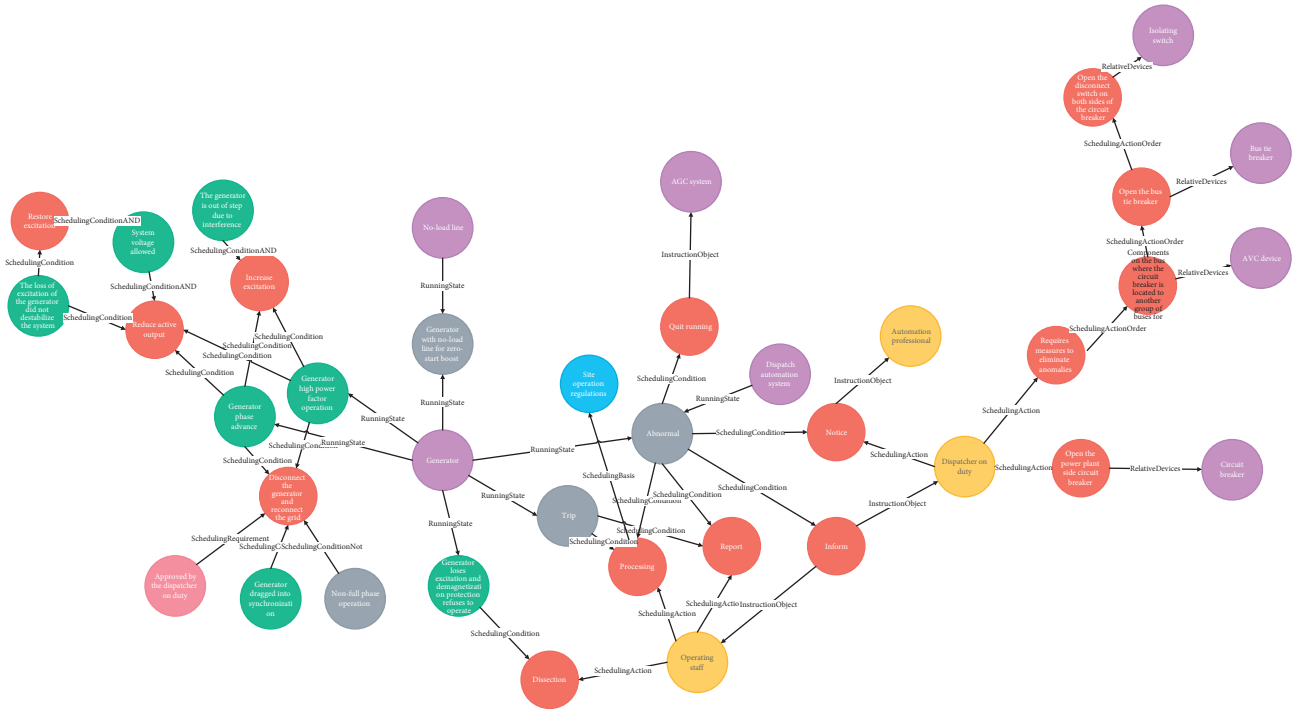


Figure 5: Partial knowledge graph of power dispatching behavior.

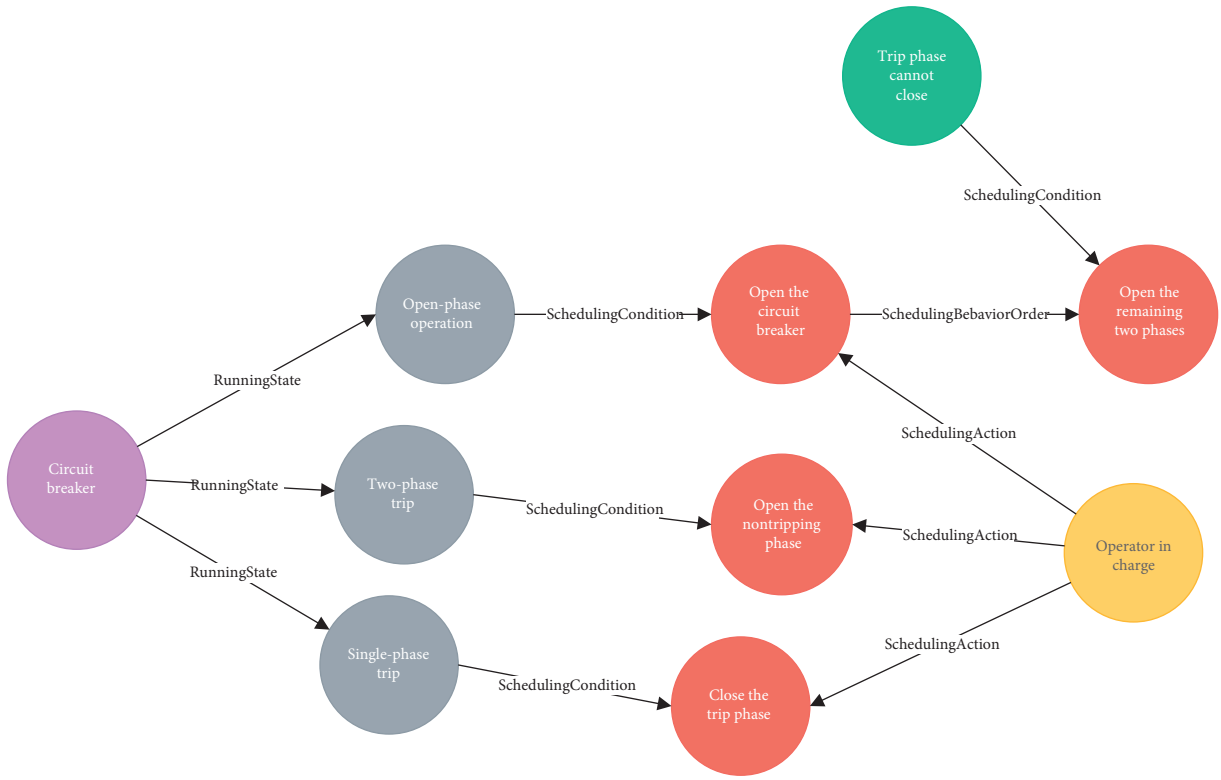


Figure 6: Power dispatching behavior graph under the circuit breaker non-full phase operation scenario.

paper can extract the entity vocabulary of power dispatching behavior from the power dispatching sentence and classify the entities.

4.2.2. Construction of Knowledge Graph of Power Dispatching Behavior. According to the relationship extraction method mentioned above, we extract the entity-relationship of the

power dispatching behavior based on the power dispatching texts and entity recognition results and form triples with the entity pairs. The graph database Neo4j is used to store the data and construct a knowledge graph structure. The result of the knowledge graph construction of power dispatching behavior is shown in Figure 5.

The nodes of different colors represent entities of different entity categories. Entities are connected by directed edges that represent relationships between entities to form the graph structure of the knowledge graph. The knowledge graph can store knowledge information such as knowledge entities and relationships. It is easy to see that, compared with other forms of databases such as original text and tables, knowledge graphs link discrete data, and knowledge representation and knowledge storage are more intuitive and efficient, without the need for intermediate data conversion and processing.

This paper adds a "scheduling scenario" attribute to the relationship of the knowledge graph, to facilitate querying the possible scheduling behavior in a certain scheduling scenario in the knowledge graph. Taking the scheduling scenario of the "non-full phase operation occurs during circuit breaker operation" as an example, we executed the Cypher query language of the neo4j database to conduct the query. The specific query statement is as follows:

$$\text{match } p = ()-[*\{\text{reSta: "non}-\text{full phase operation occurs during circuit breaker operation"}\}]-()\text{return } p. \quad (6)$$

According to the query statement to get the power dispatching behavior knowledge in this scenario, the query result is shown in Figure 6. It can be seen from a simple retrieval example that the power dispatch behavior knowledge graph constructed in this paper has both semantic information and relationship information, which can retrieve richer information and return intuitive visualization results. In addition to the example retrieval method, the knowledge graph query method is very flexible and can be queried based on entity node attributes, relationship attributes, path depth, etc., to obtain richer knowledge information. In the face of complex power dispatching business, the knowledge graph constructed in this paper will provide knowledge about the dispatching behavior of related businesses and effectively help dispatchers to conduct power dispatching.

5. Conclusion

This paper explores the construction method of knowledge graph based on power dispatching behavior. In order to obtain the annotated corpus, a phrase extraction algorithm based on mutual information and left and right entropy is used in this paper to annotate the corpus, by which the corpus is constructed semiautomatically. Based on the bidirectional long and short time memory network and conditional random field model, the entity is trained and identified. The relations of entities are extracted according to the text of power dispatching behavior, to store the data and construct the knowledge graph of power dispatching behavior.

According to the constructed knowledge graph, we can search more efficiently the knowledge related to the power dispatching behavior, provide the underlying knowledge model for the dispatching automation system, and further improve the intelligence of the power dispatching. There are also some problems and threats in this paper. The data set we used is small, and the diversity of knowledge content requires more knowledge data support. In addition, due to the lack of updated data, we cannot study the update process of the knowledge graph, and the relationship extraction method in this article depends on text mode and rules. In the future, we will conduct further research and improvement on the existing problems, continue to explore a more efficient and automated relationship extraction model, and study a more effective construction method of knowledge graph based on power dispatching.

Data Availability

The data set contains some books of Grid Dispatching Regulations published by STATE GRID Corporation of China and its subsidiaries, such as "Dispatching Regulation of Hunan Power Grid" for Hunan province of China.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

For this paper, Shixiong Fan conceived and designed the research study; Shixiong Fan, Zhifang Liao, Xingwei Liu, and Ying Chen collected data; Shixiong Fan, Xingwei Liu, Zhifang Liao, Ying Chen, and Yiqi Zhao designed the methodology and experiment; Shixiong Fan, Xingwei Liu, Ying Chen, Yiqi Zhao, and Huimin Luo completed the experiment; Shixiong Fan and Haiwei Fan conducted application deployment; Ying Chen, Yiqi Zhao, and Huimin Luo wrote and modified the initial paper; Zhifang Liao, Ying Chen, and Huimin Luo revised the paper. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported in part by the Basic Prospective Project of SGCC (no. 5442DZ180017) and in part by the

Science and Technology Research Foundation of SGCC (5442DZ180024-I).

References

[1] Y. Bi, L. Jiang, X. Wang, and L. Cui, "Design and investigation on service-oriented architecture-based smart grid dispatching and control System," *Automation of Electric Power Systems*, vol. 39, no. 2, pp. 92–99, 2015.

[2] X. Li, J. Xu, Z. Guo, W. Ning, and Z. Wang, "Construction and application of knowledge graph of power dispatch automation system," *China Electric Power*, vol. 52, no. 2, pp. 70–77, 2019.

[3] T. Steiner, "Adding realtime coverage to the google knowledge graph,"*Proceedings of the 11th International Semantic Web Conference (ISWC 2012)*, Boston, MA, USA, September 2012.

[4] Z. Liao, Z. Zeng, Y. Zhang, and X. Fan, "A data-driven game theoretic strategy for developers in software crowdsourcing: a case study," *Applied Sciences*, vol. 9, no. 4, p. 721, 2019.

[5] Z. Liao, Z. Wu, Y. Li, Y. Zhang, X. Fan, and J. Wu, "Core-reviewer recommendation based on Pull Request topic model and collaborator social network," *Soft Computing*, vol. 24, no. 8, pp. 5683–5693, 2020.

[6] Z. Liao, B. Zhao, S. Liu et al., "A prediction model of the project life-span in open source software ecosystem," *Mobile Networks and Applications*, vol. 24, no. 4, pp. 1382–1391, 2019.

[7] Z. Liao, L. Deng, X. Fan et al., "Empirical research on the evaluation model and method of sustainability of the open source ecosystem," *Symmetry*, vol. 10, no. 12, p. 747, 2018.

[8] N. Wang, "Company name identification in Chinese financial domain," *Journal of Chinese Information Pro Cessing*, vol. 16, no. 2, pp. 1–6, 2002.

[9] X. Luo, Y. Li, W. Wang, X. Ban, J.-H. Wang, and W. Zhao, "A robust multilayer extreme learning machine using kernel risk-sensitive loss criterion," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 1, pp. 197–216, Jan. 2020.

[10] G. Lample, "Neural architectures for named entity recognition," 2016, https://arxiv.org/abs/1603.01360.

[11] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[12] X. Han, Y. Zhang, W. Zhang, and T. Huang, "An attention-based model using character composition of entities in Chinese relation extraction," *Information*, vol. 11, no. 2, p. 79, 2020.

[13] J. Leng and P. Jiang, "A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm," *Knowledge-Based Systems*, vol. 100, pp. 188–199, 2016.

[14] Y. Tang, "Enhancement of power equipment management using knowledge graph," in *Proceedings of the 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, IEEE, Chengdu, China, May 2019.

[15] Y. Yang, "Multi-source heterogeneous information fusion of power assets based on knowledge graph,"*Proceedings of the 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, IEEE, Zhengzhou, China, October 2019.

[16] H.-F. Wang, "An error recognition method for power equipment defect records based on knowledge graph technology," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 11, pp. 1564–1577, 2019.

[17] X. Luo, J. Sun, L. Wang et al., "Short-term wind speed forecasting via stacked extreme learning machine with generalized correntropy," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4963–4971, 2018.

[18] M. Chen, Y. Li, X. Luo, W. Wang, L. Wang, and W. Zhao, "A novel human activity recognition scheme for smart health using multilayer extreme learning machine," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1410–1418, 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8865413, 15 pages
https://doi.org/10.1155/2020/8865413

Research Article

Flow Chart Generation-Based Source Code Similarity Detection Using Process Mining

Feng Zhang,[1,2] Lulu Li,[1] Cong Liu,[3] and Qingtian Zeng[1,2]

[1]*College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China*
[2]*Shandong Key Laboratory of Wisdom Mine Information Technology, Qingdao 266590, China*
[3]*School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China*

Correspondence should be addressed to Cong Liu; liucongchina@163.com and Qingtian Zeng; qtzeng@163.com

Received 13 April 2020; Revised 2 May 2020; Accepted 13 May 2020; Published 7 July 2020

Academic Editor: Chenxi Huang

Source code similarity detection has extensive applications in computer programming teaching and software intellectual property protection. In the teaching of computer programming courses, students may utilize some complex source code obfuscation techniques, e.g., opaque predicates, loop unrolling, and function inlining and outlining, to reduce the similarity between code fragments and avoid the plagiarism detection. Existing source code similarity detection approaches only consider static features of source code, making it difficult to cope with more complex code obfuscation techniques. In this paper, we propose a novel source code similarity detection approach by considering the dynamic features at runtime of source code using process mining. More specifically, given two pieces of source code, their running logs are obtained by source code instrumentation and execution. Next, process mining is used to obtain the flow charts of the two pieces of source code by analyzing their collected running logs. Finally, similarity of the two pieces of source code is measured by computing the similarity of these two flow charts. Experimental results show that the proposed approach can deal with more complex obfuscation techniques including opaque predicates and loop unrolling as well as function inlining and outlining, which cannot be handled by existing work properly. Therefore, we argue that our approach can defeat commonly used code obfuscation techniques more effectively for source code similarity detection than the existing state-of-the-art approaches.

1. Introduction

Research studies on source code similarity detection can be tracked back to the 1970s, and such techniques have a wide range of applications in the source code plagiarism detection of computer programming teaching and software intellectual property protection. Existing source code similarity detection techniques mainly involve attribute counting [1] and structure metrics [2–13]. Structure metrics are most commonly used approaches that mainly contain string-based, tree-based, and graph-based code similarity measure approaches. In the current teaching of computer programming courses, the online judge (OJ) system [14] that implements online submission and automatic assessment of the programming assignments has been widely used. Meanwhile, OJ uses the code similarity detection tool to find the plagiarism in programming assignments. To support code similarity detection, most existing OJ systems use string-based and tree-based detection approaches. However, the antiobfuscation effectiveness of these approaches is weak because they can only deal with some simple code obfuscation techniques [15–18]. In computer programming teaching, students sometimes use some complex code obfuscation techniques, e.g., opaque predicates, loop unrolling, and function inlining and outlining, to reduce the similarity between code fragments. However, existing approaches cannot handle these complex obfuscation techniques. The main reason for this problem is that existing approaches measure code similarity only by the static features of the source code, such as the text or structure, and they do not consider the runtime dynamic features of the source code. Thus, existing approaches cannot cope with the above complex obfuscation techniques.

To solve this problem, we propose a novel source code similarity detection approach for computer programming teaching using process mining. The dynamic features of source code are obtained through the running of the code, and they are used as the basis for measuring the similarity between two code fragments. Specifically, given two pieces of source code, we first obtain their running logs by source code instrumentation and running. Next, process mining is used to obtain their flow charts that reflect their dynamic features at runtime. Finally, the similarity between two flow charts can be measured by a graph similarity algorithm, and the similarity value is taken as the final similarity of these two pieces of code.

The rest of this paper is structured as follows. Section 2 introduces the classification of code obfuscation techniques and existing code similarity detection approaches. Section 3 proposes the basic idea and overall framework of the source code similarity detection approach based on process mining. Next, Section 4 introduces the approach in detail, and the effectiveness of our approach is verified by experiments in Section 5. Finally, the whole work is summarized in Section 6.

2. Related Work

The performance of antiobfuscation is an important metric to evaluate source code similarity detection [19]. Therefore, we first summarize most commonly used code obfuscation techniques in this section. Then, we introduce the existing source code similarity detection approaches and their ability to fight against code obfuscation techniques, based on which we summarize the problems of existing approaches.

2.1. Code Obfuscation. Jones summarized ten commonly used approaches to plagiarizing programming assignment of students [20]. On this basis, another two code obfuscation techniques are proposed in the literature [15, 21]. To sum up, there are mainly fourteen kinds of obfuscation techniques, which are shown in Figure 1 from easy to difficult: (1) verbatim copying; (2) changing comments; (3) changing white space and formatting; (4) renaming identifiers; (5) reordering code blocks; (6) reordering statements within code blocks; (7) replacing constants; (8) changing the order of operators or operands in expressions; (9) changing data types; (10) adding redundant statements; (11) splitting expressions; (12) replacing control structures with equivalent structures; (13) loop unrolling; (14) function inlining and outlining.

We further classify adding redundant statements and loop unrolling in detail in this paper. First, adding redundant statements can be divided into adding sequential statements, adding reachable branches, and opaque predicates. Specifically, adding sequential statement refers to adding sequentially executed code that can be run without affecting the results; adding reachable branch refers to adding redundant executable branches that are added without changing the result; opaque predicate [21] is a control flow obfuscation technique, and it adds unreachable paths or branches to the source code without changing the final result. Second, loop unrolling is divided into partial unrolling and full unrolling, and it reduces the number of cycles and increases the source code by copying the code in a loop body [15].

2.2. Source Code Similarity Detection. The first approach for source code similarity measure is attribute counting [22], which mainly measures the similarity by counting various metrics of the source code. Because too much structure information of the source code is discarded, the accuracy of these methods is low. Therefore, researchers propose the approaches based on structure metrics that are mainly based on strings, trees, and graphs.

The string-based source code similarity detection approaches compare the text of the code. Now this kind of approaches is relatively mature, and there are some commonly used software systems, such as JPlag [2], MOSS [3], and Sim [4]. Similarly, in terms of cloning detection, a string-based cloning detection system that can be used for large code is proposed [23]. Although string-based source code similarity detection approaches have higher space-time efficiency, they rarely reflect the semantic and syntactic information of the code. As a result, it is difficult for these approaches to fight against some complex code obfuscation techniques.

The tree-based similarity detection approaches transform the source code into a tree structure and measure the similarity between trees. For example, The AST-CC algorithm [5] transforms source code into an abstract syntax tree and improves the efficiency of syntax tree comparison by transforming storage format. This approach has stronger antiobfuscation ability than string-based code similarity detection approaches. In addition, in terms of code clone detection, a tree-based code clone detection approach is also proposed in [24, 25]. However, the tree-based approaches mainly measure the similarity through the subtrees. As a result, they cannot fight against structural obfuscation techniques [20], such as adding redundant statements and loop unrolling. Consequently, this kind of approaches can only solve simpler code obfuscation techniques.

The graph-based source code similarity detection approaches transform the code into a graph structure and compares the similarity between graphs. At present, existing approaches use two kinds of graphs: program dependency graph (PDG) and control flow graph (CFG). First, GPLAG proposed by Liu et al. constructs a PDG [6] according to the source code and calculates the similarity between two PDGs. Although a PDG contains the semantic information of the source code, GPLAG can only detect the similarity between two single functions but not multiple functions [7]. Meanwhile, it is still susceptible to the code obfuscations techniques that retain semantics, such as opaque predicates [8] as well as function inlining and outlining [9]. In addition, in terms of code cloning detection, PDGs and program slicing are used to find isomorphic PDG subgraphs that represent clones [26]. Second, Lim et al. propose to use CFG to indicate the control structure of the source code [10] and

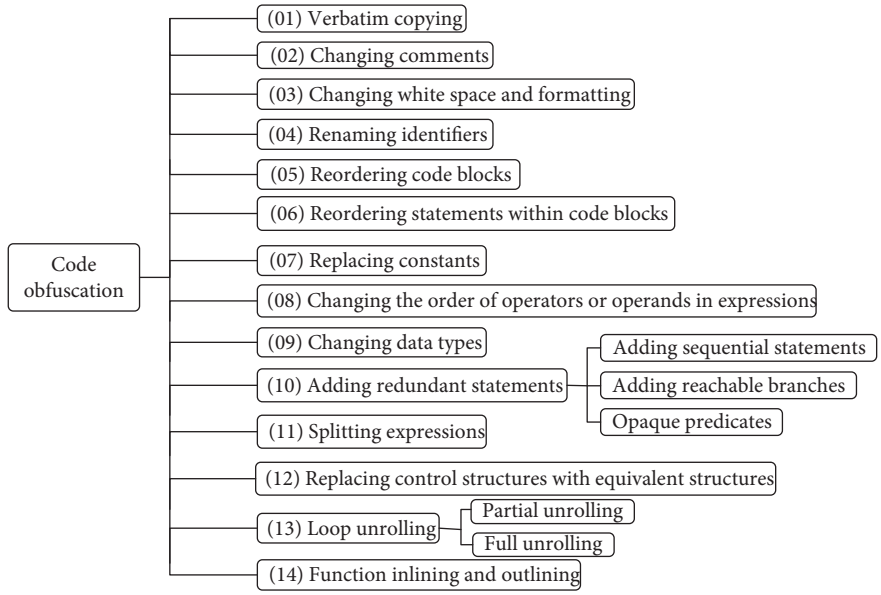


Figure 1: Code obfuscation techniques.

calculate the similarity between code fragments by analyzing the k-length flow paths of CFG. Similarly, Qiu et al. [27] propose a fault-tolerant graph matching approach. The matching subpaths of two CFGs are obtained by matching the basic blocks (a sequence of statements executed sequentially with only one entry and one exit) of two codes, and then the code similarity is calculated by weighted similarity of each path. In these approaches, the control logic of the source code that reflects the possible execution paths of the code is considered. However, because such kind of approaches cannot get the actual running path of the source code, they are still unable to resist some complex code obfuscation techniques, such as opaque predicates and reordering code blocks. In addition, an approach to measuring cross programming languages code similarity is proposes based on the static flow chart of source code [11]. Similarly, this approach only considers the static flow char of the code, making it difficult to fight against opaque predicates and function inlining and outlining. To sum up, graph-based code similarity detection approaches cannot deal with some code obfuscation techniques including opaque predicates, loop unrolling, and some other complex code obfuscation techniques.

3. An Approach Overview

3.1. Basic Idea. Existing work measures source code similarity only by the static features, such as the text or structure of source code, while it does not consider the runtime dynamic features of the source code. Figure 2 shows an example of source code to print the sum of $1 + 2 + 3 + \ldots + 100$, as well as the plagiarized code that uses two obfuscation techniques: loop unrolling and adding redundant statements. Two code fragments are shown in Figure 2(a) and Figure 2(c), respectively. After adding redundant statements, the text and structure of these two code fragments have some differences. Take their flow charts as an example. The flow chart of the source code is shown in Figure 2(b), and the flow chart of the obfuscated code is shown in Figure 2(d). The use of the two obfuscations reduces the similarity between the flow charts of the source code and obfuscated code. Therefore, the traditional source similarity detection approaches based on static features of code cannot deal with these obfuscation techniques. To solve this problem, we propose to obtain dynamic features of source code through the running of code and use the dynamic features to measure the similarity between two code fragments.

The dynamic features of source code are the key to the proposed approach. For two code fragments, we first obtain their running logs by source code instrumentation and running. Then, we use the process mining to obtain their flow charts that indicate their dynamic features at runtime. Process mining extracts process-related information from the event logs to discover, monitor, and improve the actual processes [28]. By process mining, the execution sequences of the activities in the execution logs based on existing process instances can be mined to obtain the dependencies and execution orders between the activities. Thus, the process model that expresses the logical relationship between the activities can be obtained. Borrowing the idea of process mining, if we regard the output logs of a piece of code by running once as the activity sequence in a process instance and obtain a group of output logs after running the code multiple times, these logs can be used as the input of a process mining algorithm to get a flow chart that reflects the actual running process of the code. Based on the above analysis, we propose to use process mining to get the flow chart by mining the output logs that are produced through running the source code and take the flow chart as the dynamic features of the code. Thus, the similarity between flow charts can be used to measure the similarity of the source code. In this way, the more accurate similarity of two pieces of code can be obtained.

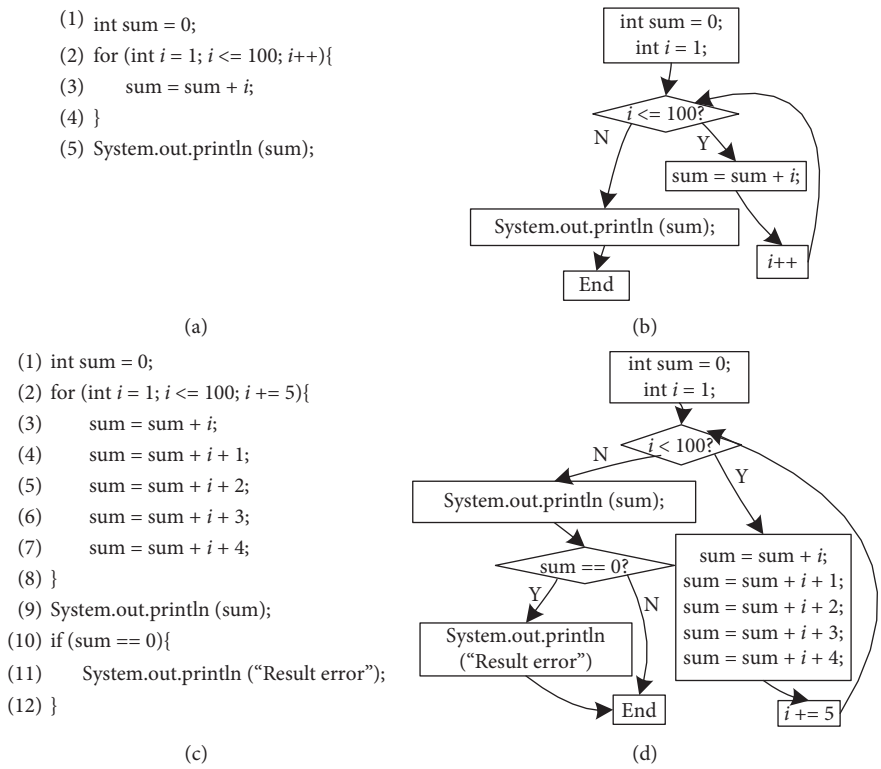


Figure 2: An example of code obfuscation.(a) Source code, (b) source code flow chart, (c) obfuscated code, and (d) obfuscated code flow chart.

3.2. Framework Design. A general framework for source code similarity detection using process mining is shown in Figure 3. First, the two pieces of source code are preprocessed to remove redundant statements. Then, some output statements are automatically inserted into these two preprocessed code fragments (this process is called "code instrumentation"). By running two code fragments, their running logs can be obtained. Next, the flow charts of the two code fragments are obtained by mining the logs of these two code fragments by a process mining algorithm. Finally, a graph similarity measure algorithm is used to calculate the similarity between these two flow charts, and the value is regarded as the final similarity between these two code fragments.

4. Proposed Approach

The source code similarity detection approach based on process mining is introduced in detail in this section according to the above framework. The approach mainly contains the following four steps: (1) preprocessing redundant statements based on PDGs; (2) automatic source code instrumentation; (3) the code flow charts generation based on process mining; (4) similarity calculation of code flow charts. In addition, three examples are given to show the effectiveness of the approach in fighting against the above three complex code obfuscation techniques.

4.1. Preprocessing Redundant Statements Based on PDGs. Adding redundant statements is a common code obfuscation technique. For example, redundant variable declaration statements can reduce the similarity of two code fragments. Therefore, it is necessary to remove redundant statements. For the code assignments submitted by students in OJ system, the last statement is usually a *return* or an *output* statement. Therefore, if the source code is converted into a PDG, the nodes in the PDG can be traversed in depth-first order from the final node of the PDG that corresponds to the final statement, and a new graph can be obtained. The statements that correspond to the nodes not in the graph can be regarded as redundant statements and can be deleted. Therefore, if there is no data dependence and control dependence between the redundant statements and the core code, we can convert the code into a PDG and find redundant statements. Algorithm 1 presents the PDG-based preprocessing algorithm for removing redundant statements, which converts the source code *oc* into a PDG $G(V, E, \mu, \delta)$. In this algorithm, V is the set of statement nodes, E is the set of dependency edges between nodes in V, μ is the type mapping function for statement nodes, and δ is the type mapping function (data dependency or control dependency) for edges [6].

We take the code fragment in Figure 4(a) as an example. Given the code fragment to get the maximum value of two variables: *a* and *b*, the statements on lines 2, 8, and 9 are redundant statements and they have no data dependence with the last *return* statement in line 10. Figure 4(b) is the PDG generated by the code fragment, and Figure 4(c) is the PDG after deleting the redundant statements. Finally, the redundant statements on lines 2, 8, and 9 are removed.

4.2. Automatic Source Code Instrumentation. A certain amount of running logs is needed to obtain the flow chart of

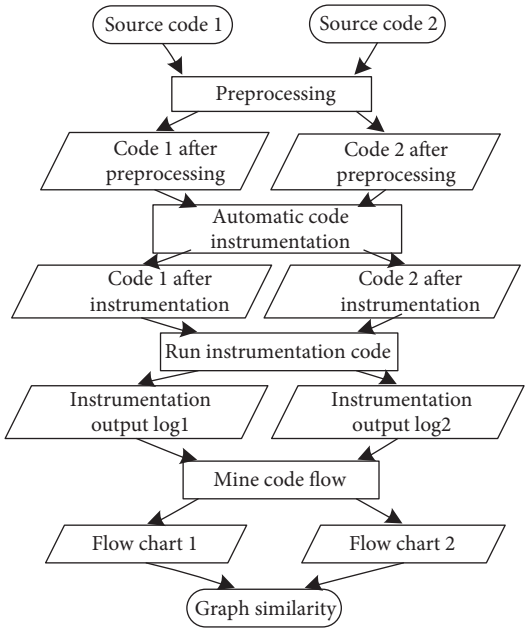


Figure 3: A general framework of source code similarity detection using process mining.

Input: The original code *oc*
Output: The target code *tc*
(1) Generate the PDG of *oc*: $G(V, E, \mu, \delta)$;//convert *oc* to a PDG
(2) Denote the end node as $G.v_e$, the target node set as V'; //V' is the node set that are dependent on the last node
(3) **for** each $v \in V$, $t = \mu(v)$ **do**
(4) **if** $(t == return || t == output)$ **then**//determine whether the current node *t* is a *return* statement or an *output* statement
(5) $v_e = v$;
(6) break;
(7) **end if**
(8) **end for**
(9) $V' = \boldsymbol{DFS}(V)$//get node set that are dependent on the last node
(10) **for** each $v \in G.V$ **do**
(11) **if**$(v \notin V')$ **then**//delete nodes that do not depend on the last node
(12) $tc = oc$.delete(v)
(13) **end if**
(14) **end for**

Algorithm 1: Redundant statements preprocessing based on PDGs.

a code fragment by process mining. However, there are not enough print statements that can indicate the flow of the code assignments. Therefore, it is necessary to insert some print statements into the code fragment to output some running logs of the code. Meanwhile, the output logs need to indicate the execution path of the code fragment, so that they can be used as the input of a process mining algorithm. By existing instrumentation software as well as the definition of nodes in PDG [6], we define two basic types of source code instrumentation statements: *Assign* and *Output*.

(1) **Assign ($\mathbf{Type_1}$,..., $\mathbf{Type_n}$)** represents variable declaration or variable assignment. $\mathbf{Type_1}$,..., $\mathbf{Type_n}$ are the variable types of *Assign*, where $Type_i \in$ {the variable types provided by a program language}.

(2) **Output** represents the print statement in a code fragment.

Based on two instrumentation statements, we give the following rules for source code instrumentation.

(1) Instrumentation statements *Assign* ($Type_i$) are inserted after assignment and variable declaration statements. If there is no data dependency between multiple consecutive assignment statements in a basic block [27], an instrumentation statement is inserted after the last assignment statement, and meanwhile the variable types of all assignment statements are merged; i.e., *Assign* ($Type_1$,..., $Type_n$) is inserted at the end of these statements.

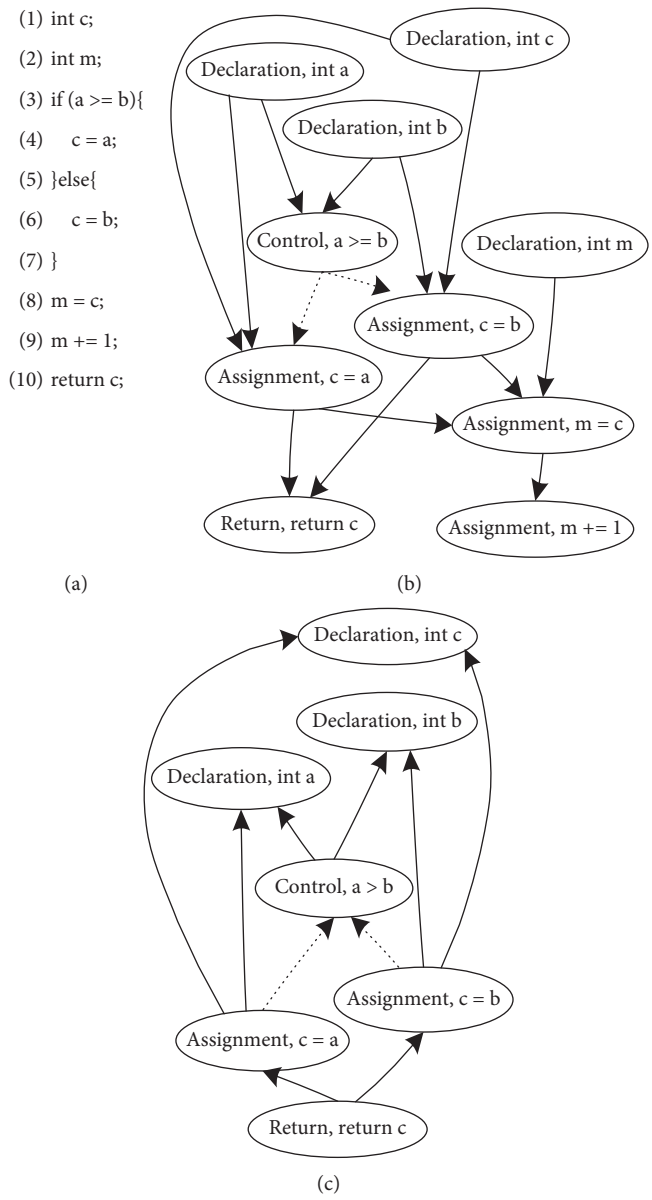


Figure 4: An example of PDG-based redundant statements preprocessing. (a) Source code. (b) PDG. (c) Graph after preprocessing.

(2) For the output statement, *Output* instrumentation is inserted after it.

(3) Instrumentation statements of the same type are numbered sequentially.

Figure 5 shows an example of automatic source code instrumentation of a piece of Java code based on the above instrumentation rules. The first four lines of the code are consecutive assignment statements in a basic block. The assignment statements in lines 1 and 2 have no data dependency, so they are combined, and *Assign1(int, int)* is inserted according to the first rule. The statements in lines 3 and 4 have data dependency and two instrumentation statements, i.e., *Assign2(Scanner)* and *Assign3(int)*, are inserted after them, respectively. Since there are two basic blocks in the *if* branch (line 5–9), two *Assign* instrumentation statements are inserted for the assignment statement in each basic block. For the output statement in the last sentence, the *Output* instrumentation is inserted after it.

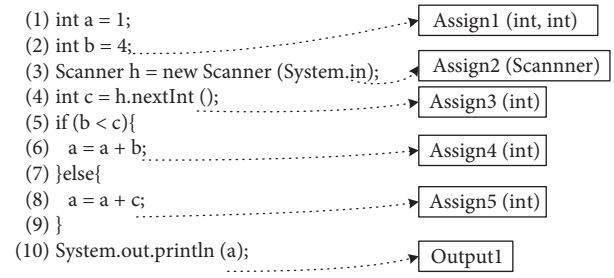


Figure 5: An example of source code instrumentation.

Input: instrumentation output sequences by running the source code, *Seq*
Output: CDFC = (*V*, *E*)
(1) Denote the following directly threshold as T_f, the threshold of dependence as T_d, the number of following directly as *num*[][] = 0, the dependence as *d*[][], the instrumentation output set as *IO*, the node set of CDFC *as V*, and the edge set of CDFC as *E*.
(2) **for** each *trace* in *Seq* **do**//traverse the instrumentation output sequences of each running of the code
(3) **for** *i* = 0; *i* < *trace.size*-1; *i*++ **do**//traverse the adjacent instrumentation outputs in the instrumentation output sequences
(4) *num*[*trace*[*i*]*trace*[*i*+1]]++//record the following directly number of every two instrumentation outputs
(5) **if** *trace*[*i*] not exist in *V* **then**
(6) add *trace*[i] to *V*;
(7) **end if**
(8) **end for**
(9) **end for**
(10) **for** each io_1 in *IO* **do**
(11) **for** each io_2 in *IO* **do**
(12) **if** $io_1 == io_2$ **then**//the output of two instrumentation outputs is the same
(13) $d[io_1][io_2] = num[io_1][io_2]/(num[io_1][io_2] + 1)$//calculate the dependence of every two instrumentation outputs
(14) **if** $num[io_1][io_2] \geq T_f$ and $d[io_1][io_2] \geq T_d$ **then**
(15) *add* (io_1, io_2) to *E*
(16) **end if**
(17) **end if**
(18) **if** $io_1 != io_2$ **then**
(19) $d[io_1][io_2] = (num[io_1][io_2] - num[io_2][io_1])/(num[io_1][io_2] + num[io_2][io_1] + 1)$//calculate the dependence of every two instrumentation outputs
(20) **if** $num[io_1][io_2] \geq T_f$ and $d[io_1][io_2] \geq T_d$ **then**
(21) *add* (io_1, io_2) to *E*
(22) **end if**
(23) **end if**
(24) **end for**
(25) **end for**
(26) return CDFC = (*V*, *E*)

Algorithm 2: CDFC mining algorithm based on the heuristic process mining.

4.3. Source Code Flow Chart Generation Using Process Mining. The output logs of the source code can be obtained by running the code with instrumentation (instrumentation code for short). Then, a process mining algorithm can be used to mine the output logs that are made up of *Assign* and *Output* instrumentation statements to get the actual running process of the code that is expressed by a process model.

Existing process mining algorithms mainly include the α algorithm and the heuristic algorithm [28]. Among them, the α algorithm cannot deal with short loops with length of one or two. As a result, the algorithm cannot deal with the cases where the loops are executed only once or twice in the flow chart of the code. Heuristic mining algorithms [28] consider the frequency of events and sequences when building a process model, by which the actual running processes such as loops and branches in the source code can be mined. Therefore, we use the heuristic mining algorithm to mine the output logs and generate the code flow chart based on a causal network [28]. Since the flow chart is mined from the output logs produced by running the source code, it can indicate the dynamic features of the code. Therefore, we call such kind of flow charts *Code Dynamic Flow Chart*, which is defined as follows.

Definition 1. CDFC (Code Dynamic Flow Chart) = (*V*, *E*), where *V* is the set of *Assign* or *Output* nodes and $E \subseteq V \times V$ is the set of edges that represent the sequential relationship of nodes in *V*.

The output logs obtained by each running of the instrumentation code can be regarded as the running logs of a process instance that is required by a process mining algorithm. Meanwhile, each *Assign* and *Output* in the running logs of the instrumentation code (instrumentation output for short) can be regarded as an activity that is executed in the process instance. The sequence composed of *Assign* and *Output* is called the instrumentation output sequence, and it is regarded as the input the process mining algorithm. Based on this idea, we propose the CDFC mining algorithm that takes the logs obtained by running the instrumentation code as input, and outputs the corresponding CDFC based on a causal network. The instrumentation code needs to be run multiple times, and thus a set of output logs that are required by the process mining algorithm can be obtained. Then, the CDFC is obtained using the heuristic mining algorithm according to the occurrence times and dependence of *Assign* and *Output* in the instrumentation output sequence. Algorithm 2 shows the CDFC mining process based on the heuristic process mining algorithm.

Algorithm 2 first counts the directly following number (lines 2–6) of each two adjacent *Assign* or *Output* in the instrumentation output sequence. Then, the dependency (lines 10 and 16) between two instrumentation outputs is calculated. Finally, the connection between the instrumentation outputs is built. Thus, the code flow chart is obtained according to the directly following number and the dependency values (lines 11–13, lines 17–19).

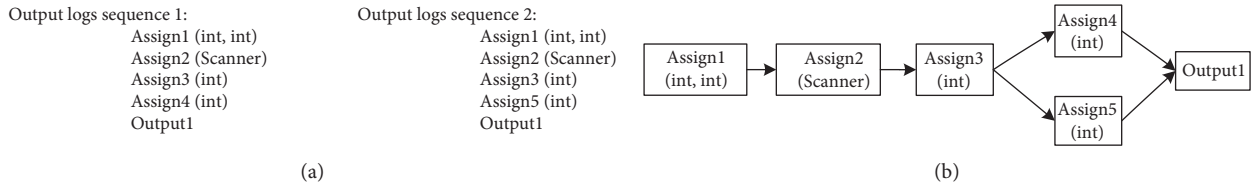


Figure 6: An example of CDFC mining. (a) Code instrumentation output logs. (b) CDFC.

Input: $\mathrm{CDFC}_1 = (V_1, E_1)$, $\mathrm{CDFC}_2 = (V_2, E_2)$, where, $V_1 = \{v_{11},..., v_{1n}\}$ and $V_2 = \{v_{21},..., v_{2m}\}$
Output: Minimum cost path $\lambda_{\min}$ that CDFC_1 is transformed to CDFC_2
(1) $OPEN \longleftarrow \{\emptyset\}$, $\lambda_{\min} \longleftarrow \emptyset$;
(2) **if** v_{1i} is a subset of v_{2j} **then**
(3) $v_{1i} = v_{2j}$;
(4) **end if**
(5) **for** each node $w \in V_2$ **do**
(6) $OPEN \longleftarrow OPEN \cup \{v_{11} \longrightarrow w\}$;
(7) $OPEN \longleftarrow OPEN \cup \{v_{11} \longrightarrow \varepsilon\}$; //the operation and cost value of v_{11} that replaces and deletes each vertex w in CDFC_2 are stored into *OPEN*
(8) **end for**
(9) **while** true **do**
(10) $OPEN \longleftarrow \lambda_{\min} = \mathrm{argmin}_{\lambda \in \mathrm{OPEN}}\{g(\lambda) + h(\lambda)\}$; //g($\lambda$) is the accumulated distance value from the starting point to the current search point, $h(\lambda)$ is the estimated distance value from the current point to the end point, and the point with the smallest function value is the next search point
(11) **if** $\lambda_{\min}$ is a complete edit path **then**
(12) return $\lambda_{\min}$;
(13) **else**
(14) $\lambda_{\min} = \{v_{11} \longrightarrow v_{2i},..., v_{1k} \longrightarrow v_{2k}\}$
(15) **if** $k < |V_1|$ **then**
(16) **for** each $w \in V_2 \backslash \{v_{2i},..., v_{2k}\}$**do**//$V_2 \backslash \{v_{2i},..., v_{2k}\}$ represents unmatched nodes in V_2
(17) $OPEN \longleftarrow \lambda_{\min} \cup \{v_{1k+1} \longrightarrow w\}$;
(18) $OPEN \longleftarrow \lambda_{\min} \cup \{v_{1k+1} \longrightarrow \varepsilon\}$;
(19) **end for**
(20) **else**
(21) **for** each $w \in V_2 \backslash \{v_{21},..., v_{2k}\}$ **do**
(22) $OPEN \longleftarrow \lambda_{\min} \cup \{\varepsilon \longrightarrow w\}$;
(23) **end for**
(24) **end if**
(25) **end if**
(26) **end while**

Algorithm 3: Graph edit distance of two CDFCs.

Figure 6 gives an example, in which Figure 6(a) shows two instrumentation output sequences of the instrumentation code in Figure 5. Through two test cases with the input 5 and 3, respectively, two different instrumentation output sequences are obtained. As is shown in Figure 6(a), either *Assign*4 or *Assign*5 is executed in each instrumentation output sequence. Therefore, a mutually exclusive branch is included in the CDFC obtained based on Algorithm 2, which is shown in Figure 6(b).

4.4. CDFC Similarity Measure Based on Graph Edit Distance. The similarity between two code fragments can be measured by the similarity between two CDFCs. Because a CDFC is a directed graph, the similarity measure approach for directed graphs can be used. In recent years, researchers have proposed many approaches for graph similarity measure, among which graph edit distance (GED) [29] is a commonly used algorithm. GED calculates the transformation intensity that is needed to transform one graph into another, and therefore it measures the dissimilarity between two graphs. We use GED-based algorithm to calculate the similarity of two CDFCs.

Let $\mathrm{CDFC}_1 = (V_1, E_1)$ and $\mathrm{CDFC} = (V_2, E_2)$. The definition of graph edit distance between CDFC_1 and CDFC_2 is shown as follows:

$$\mathrm{d}_{\lambda\min}(\mathrm{CDFC}_1, \mathrm{CDFC}_2) = \min_{\lambda \in \gamma(\mathrm{CDFC}_1, \mathrm{CDFC}_2)} \left\{\sum_{e_j \in \lambda} c(e_j)\right\} \tag{1}$$

Among them, $\lambda_{\min}$ denotes the minimum cost and path among all the complete editing paths; γ (CDFC_1, CDFC_2)

denotes all the editing paths from $CDFC_1$ to $CDFC_2$; and $c(e_j)$ denotes the cost of editing e_j. In the proposed approach, for the *Assign* nodes in a CDFC, if the variable types in an *Assign* node are the subset of the variable types in another one, these two *Assign* nodes represent the same node. Because each edge in a CDFC points from one node to another, the cost of *replacement*$(v \longrightarrow w)$, *deletion*$(v \longrightarrow \varepsilon)$, and *addition*$(\varepsilon \longrightarrow v)$ of nodes is the same as that of *deleting*$(v \longrightarrow \varepsilon)$ and *adding*$(\varepsilon \longrightarrow v)$ of edges. In this paper, the operation cost function of the node and that of the edge are set to one.

The graph edit distance of two CDFCs based on GED is calculated in Algorithm 3. Its basic idea is to make every possible editing operation be considered by processing the nodes in $CDFC_1$ one by one. Specifically, a search tree is built dynamically and the process of traversing the search tree is equivalent to the process of solving an editing path.

In Algorithm 3, the first vertex v_{11} of $CDFC_1$ is selected by traversing the search tree. The operation and cost value of v_{11} that replaces and deletes each vertex w in $CDFC_2$ are stored in an initial empty set (lines 5–8). Then, the nodes with the lowest operation cost are selected to be traversed according to the heuristic search algorithm, and the *replacement* and *deletion* operations of the remaining nodes are inserted into the *OPEN* set (lines 17 and 18). After all the node operations of *replacement* and *deletion* in $CDFC_1$ are completed, the remaining nodes after processing in $CDFC_2$ are inserted (lines 21–23). Finally, the edit distance path $\lambda_{\min}$ of $CDFC_1$ and $CDFC_2$ is obtained.

Given the flow chart of the source code $CDFC_1 = (V_1, E_1)$ and the flow chart of the obfuscated code $CDFC_2 = (V_2, E_2)$, $\lambda_{\min}$ is the edit distance between $CDFC_1$ and $CDFC_2$. $|CDFC| = |CDFC.V| + |CDFC.E|$ denotes the module value of CDFC, where $|CDFC.V|$ denotes the number of nodes in CDFC, and $|CDFC.E|$ denotes the sum of the length of all edges in CDFC. The similarity of the source code and the obfuscated code can be calculated as

$$sim\left(CDFC_1, CDFC_2\right) = 1 - \frac{\lambda_{\min}}{\max\left(\left|CDFC_1\right|, \left|CDFC_2\right|\right)}. \tag{2}$$

4.5. Effectiveness in Fighting Against Code Obfuscation. In this section, we analyze the effectiveness of the proposed approach in fighting against three relatively complex code obfuscation techniques, opaque predicates, loop unrolling, and function inlining and outlining, by three examples.

Firstly, in terms of opaque predicates, unreachable paths can be added to reduce the similarity of two code fragments. Taking the code fragment in Figure 2 as an example, an unreachable *if* branch is added (lines 10–12) in Figure 2(c). In addition, the statements in lines 2–4 of Figure 2(a) represent a loop structure, and the statements in lines 2–8 of Figure 2(c) constitute the code that is obfuscated by copying the loop body while the result from running the code does not change. The source code and obfuscated code in Figure 2 with inserted instrumentation code by the proposed approach are as shown in Figures 7(a) and 7(d). Because the *if* branch of the obfuscated code cannot be executed, the code in the *if* branch will not be executed. In addition, because the loop structure is changed, the code instrumentation output sequences are different after running the two instrumentation code fragments. There are one hundred and two lines in the source code instrumentation output sequences, while there are only 22 lines in the obfuscated code instrumentation output sequences, which are shown in Figures 7(b) and 7(e), respectively. The CDFCs of the two code fragments obtained by the heuristic process mining algorithm are shown in Figures 7(c) and 7(f), respectively. Because the variable types in *Assign2* node in the CDFC of the source code are a subset of that in *Assign2* nodes in the CDFC of the obfuscated code, the similarity between the two CDFCs is 100% according to Algorithm 3. Therefore, the proposed approach can fight against the opaque predicates and loop unrolling used in the example of Figure 2.

Secondly, we analyze the effectiveness of the proposed approach in fighting against function inlining and outlining. Figure 8 shows an example of the source code to get the maximum value between two input integers. The obfuscated code uses the *if* branch structure in the source code as a method invocation, and a new function is added. Therefore, the textual similarity between these two code fragments is reduced. However, the CDFCs obtained by the proposed approach are still the same, as shown in Figure 8(c). Therefore, the proposed approach can fight against function inlining and outlining used in this example.

5. Experiment and Evaluation

We conduct a group of experiments to evaluate the effectiveness and efficiency of the proposed approach for source code similarity detection in computer programming teaching. First, the experimental setup is introduced in 5.1. The data set used in the experiments is introduced in 5.2. Next, in 5.3, the effectiveness in fighting against code obfuscation techniques of the proposed approach is compared with that of existing source code similarity measure approaches. In addition, the efficiency of the proposed approach is verified compared with Sim and GPLAG in 5.4. Finally, the conclusions of the experiments are given.

5.1. Experimental Setup. For a pair of source code and obfuscated code, their similarity is calculated by the proposed approach and existing state-of-the-art approaches. Because Sim and GPLAG are representatives of the string-based and graph-based similarity measure approaches, respectively, we choose these two approaches to be compared with our approach. Thus, we can verify the effectiveness of the proposed approach in source code similarity detection and the ability to fight against code obfuscation techniques.

We conduct three experiments. First, for a data set of source code, we use three obfuscation techniques in the first experiment to modify the source code: opaque predicate, loop unrolling, and function inlining and outlining. In this way, we can construct the source code and obfuscated code pairs. Then, the similarity of each pair of code is calculated

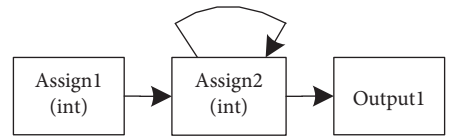
```
int sum = 0;
System.out.println ("Assign1(int)");
for (inti = 1;i < = 100;i++){
    sum = sum + i;

System.out.println ("Assign2(int)");
}
System.out.println (sum);
System.out.println ("Output1");
```

(a)

```
Assign1 (int)
Assign2 (int)
Assign2 (int)
...
...
...
Assign2 (int)
Output1
```

(b)



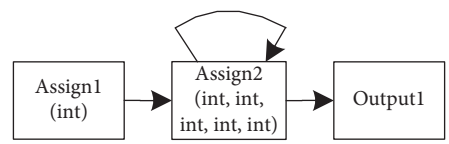
(c)

```
int sum = 0;
System.out.println ("Assign1(int)");
for (inti = 1;i < = 100;i + = 5){
        sum = sum + i;
        sum = sum + i + 1;
        sum = sum + i + 2;
        sum = sum + i + 3;
        sum = sum + i + 4;
        System.out.println ("Assign2(int,int,
        int,int,int)");
}
System.out.println (sum);
System.out.println ("Output1");
if (sum == 0){
        System.out.println ("Resulterror");
        System.out.println ("Output2");
}
```

(d)

```
Assign1 (int)
Assign2 (int, int, int, int, int)
Assign2 (int, int, int, int, int)
...
...
...
Assign2 (int, int, int, int, int)
Output1
```

(e)



(f)

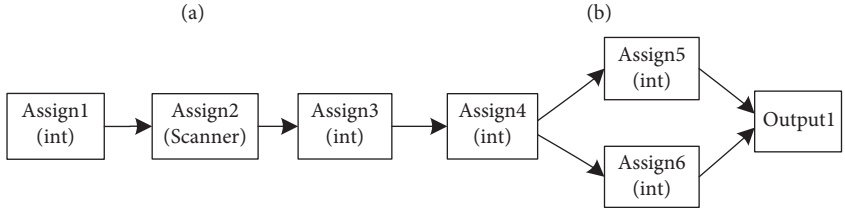
Figure 7: Examples of opaque predicates and loop unrolling. (a) Source code instrumentation. (b) Source code instrumentation output sequence. (c) Source code flow chart. (d) Obfuscated code instrumentation. (e) Obfuscated code instrumentation output sequence. (f) Obfuscated code flow chart.

```
(1)  int a = 2;
(2)  Scanner m = new Scanner (System.in);
(3)  int b = m.nextInt();
(4)  int c = m.nextInt();
(5)  if (b < c){
(6)      a = a + b;
(7)  }else{
(8)      a = a + c;
(9)  }
(10) System.out.println (a);
```

(a)

```
(1)   int a = 2;
(2)   Scanner m = new Scanner(System.in);
(3)   int b = m.nextInt();
(4)   int c = m.nextInt();
(5)   a = Min(b,c);
(6)   System.out.println(a);
(7)  private static int Min(int b, int c){
(8)   if (b < c){
(9)       b = b + 2;
(10)      return b;
(11)   }else{
(12)      c = c + 2;
(13)      return c;
(14)  }
(15)  }
```

(b)



(c)

Figure 8: An example of function inlining and outlining. (a) Source code. (b) Obfuscated code. (c) Code flow chart.

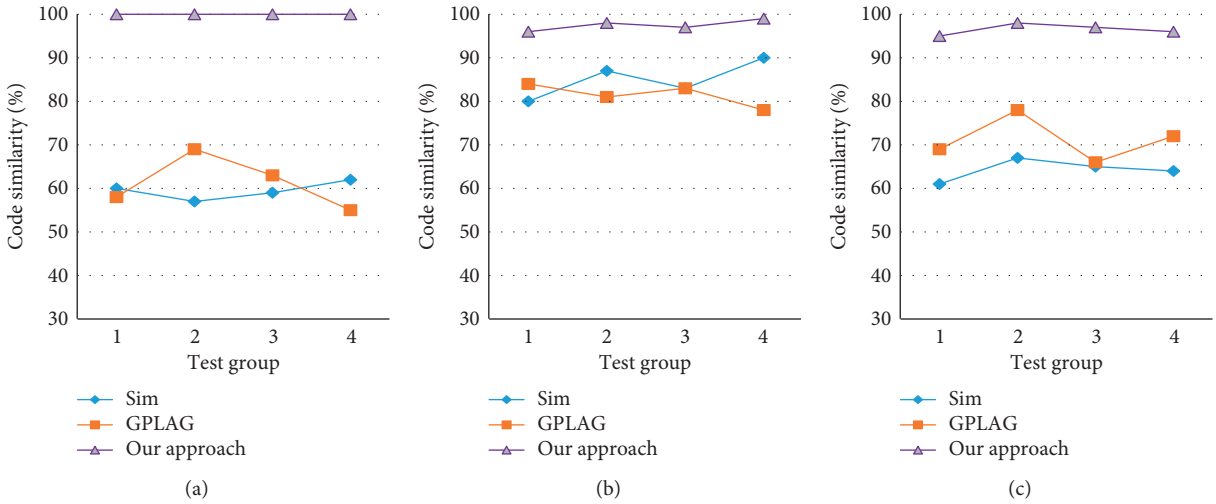


Figure 9: Experimental results of three approaches in fighting against opaque predicates, loop unrolling, and function inlining and outlining. (a) Opaque predicate. (b) Loop unrolling. (c) Function inlining and outlining.

by the proposed approach, Sim, and GPLAG to verify the effectiveness of the proposed approach in fighting against the above three obfuscation techniques.

Second, to verify the practicability of the proposed approach in source code similarity detection, experiment 2 evaluates the antiobfuscation ability of the proposed approach by some commonly used obfuscation techniques that existing approaches can fight against. Specifically, Sim, GPLAG, and the proposed approach are used to compare the average similarity between the source code and the obfuscated code that is processed by seven commonly used code obfuscation techniques.

The first two experiments demonstrate the ability of the proposed approach to fight against a single code obfuscation technique. To further verify the effectiveness of the proposed approach against multiple code obfuscation techniques, we select more complex source code and use a few obfuscation techniques to process the source code in the third experiment. Then, we use Sim, GPLAG, and our approach to calculate code similarity and compare their effectiveness in antiobfuscation.

In addition, the efficiency of Sim, GPLAG, and our approach is compared in the fourth experiment to verify the efficiency of the proposed approach.

5.2. Data Sets. We take the code assignments submitted by students in the OJ system as data sets. Because code obfuscation techniques used by students in the data set are not complete, we obfuscate the code manually and construct three source code sets. It should be noted that each exercise problem in OJ system contains a set of test cases, in which the input data of each test case can be used as the input of the proposed approach when running the source code. Therefore, through running all test cases, the instrumentation output sequences of the code can be obtained.

The first code set is used for the first experiment. We randomly select twenty Java programming problems from our OJ system and divide them into four groups, and therefore there are five problems in each group. Then, the code assignment of a student that is selected randomly is modified using three obfuscation techniques: opaque predicate, loop unrolling, and function inlining and outlining. As a result, for each obfuscation technique, an experimental data set consisting of twenty pairs of source code and obfuscated code in four groups is obtained.

The second code set is used for the second experiment. We randomly select twenty Java problems with medium length from our OJ system. Meanwhile, the code assignment of a student which is randomly selected is obfuscated using the following seven code obfuscation techniques: (1) changing comments; (2) renaming identifiers; (3) reordering statements within code blocks; (4) replacing control structures with equivalent structures; (5) replacing constants; (6) changing data types; (7) splitting expressions. As a result, there are seven different obfuscated code fragments for each problem. For each obfuscation technique, twenty pairs of source code and obfuscated code are obtained in this experiment.

The third code set is used for the third experiment. We choose three problems with high complexity from our OJ system. Meanwhile, we select a code assignment of a student randomly for each problem and use the following five code obfuscation techniques to modify the code: renaming identifiers, reordering statements within code blocks, opaque predicates, replacing control structures with equivalent structures, and function inlining and outlining. Thus, three pairs of source code and obfuscated code are obtained in this experiment.

5.3. Experimental Results. Figure 9 shows the result of the first experiment, in which the ordinate axis stands for the average similarity between source code and obfuscated code for each group of five questions, and the abscissa axis represents four code groups. The source code is obfuscated by opaque predicate, loop unrolling, and function inlining

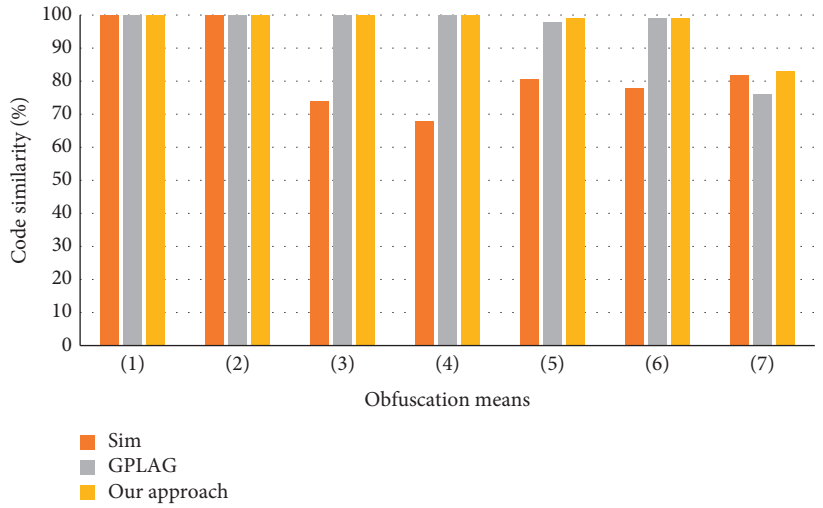


Figure 10: Experimental results of three approaches in fighting against seven obfuscation techniques.

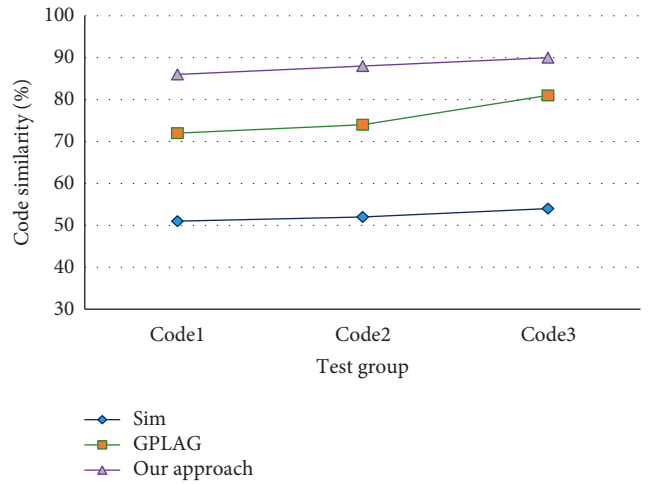


Figure 11: Experimental results of three approaches in fighting against multiple code obfuscation techniques.

Table 1: Efficiency of the three approaches.

Code	Code byte	Matches	Sim (s)	GPLAG (s)	Our approach (s)
Code 1	3147	58	1.2	1.87	2.32
Code 2	3871	72	1.95	2.79	3.14
Code 3	3655	69	1.72	2.71	2.91

and outlining. Opaque predicates and loop unrolling obfuscation techniques add redundant strings into the source code, while the function inlining and outlining disturbs the order of some strings in the source code. At the same time, the structures of PDGs are also changed. As a result, the antiobfuscation ability of Sim and GPLAG in fighting against these three kinds of code obfuscation techniques is poorer than the proposed approach. The approach in this paper can obtain the actual flow of the code by running the code, and therefore it can detect opaque predicates. For the source code obfuscated by loop unrolling as well as function inlining and outlining, some redundant code that has data dependence with the final *return* or *print* statements of the code may be added to the obfuscated code. As a result, this part of the redundant code cannot be deleted by the preprocessing, making the similarity result of the proposed approach slightly lower. However, the detection precision of the proposed approach is still better than that of Sim and GPLAG. To sum up, compared with Sim and GPLAG, the proposed approach has better effectiveness in fighting against opaque predicate and loop unrolling as well as function inlining and outlining.

Figure 10 shows the similarity between the source code and the obfuscated code given by three code similarity detection approaches in the second experiment. Among them, the ordinate axis represents the average similarity of twenty pairs of source code and obfuscated code that is obfuscated by one code obfuscation technique, and the abscissa axis represents the seven code obfuscation techniques. It can be seen that Sim, GPLAG, and the proposed approach can deal with the simple code obfuscation techniques including changing comments and renaming identifiers; Sim cannot fight against adjusting the order of statements in code blocks, replacing control structures with equivalent structures, replacing constants, and changing data types, while GPLAG and the proposed approach can defeat these four obfuscation techniques. Moreover, these three approaches cannot fully fight against splitting expressions because this obfuscation technique splits a complex expression into multiple simple expressions and add multiple lines of code with data dependency. To sum up, the proposed approach can defeat the code obfuscation techniques that can be handled by existing approaches.

Figure 11 shows the result of the third experiment, in which the ordinate axis stands for the similarity value between the source code and the code obfuscated by multiple obfuscation techniques. The code similarity values given by Sim are all less than 60%. GPLAG cannot be affected by reordering statements within code blocks and replacing control structures with equivalent structures, and therefore the code similarity detection effectiveness is better than that of Sim. However, GPLAG measures the code similarity by the static PDG of the source code, and it cannot obtain the actual running characteristics of source code. Therefore, GPLAG cannot fight against opaque predicates and function inlining and outlining, making the source code similarity of each group lower than that of our approach.

5.4. Efficiency Evaluation. We use three groups of source code and obfuscated code pairs in the third experiment to evaluate the efficiency of the proposed approach. Specifically, we count the average byte length of the code in each group and the number of nodes in each generated CDFC and compare the execution time of the proposed approach with that of Sim and GPLAG. The experiment is carried out on a computer with 3.0 GHz Intel(R) Core(TM) i5-7267U CPU, 8 G memory, and Win 10. Each group of experiments is carried out ten times, and the average running time is taken as the result. The experimental results are shown in Table 1. Sim is the fastest because it is a string-based approach to measuring the similarity. In the proposed approach, source code instrumentation, running code, and the process mining are needed, and therefore the execution time is more than that of Sim and GPLAG. However, the proposed approach can provide acceptable performance in detecting the similarity of code assignments in computer programming teaching.

6. Conclusion

In the teaching of computer programming courses, the code assignments submitted by students through online judge system are generally the code to solve a specific problem. This kind of code is usually short, and the overall complexity is not high. However, students may use some complex code obfuscation techniques, such as opaque predicates, loop unrolling, and function inlining and outlining, to reduce the similarity between source codes. Aiming at source code similarity detection in computer programming teaching, we propose an approach to measure source code similarity based on process mining. Through the running of source code and the mining of output logs, the dynamic features that indicate the actual flow of the source code are obtained and used to calculate the code similarity. The results show that the proposed approach can fight against not only the code obfuscation techniques that existing approaches can defeat, but also more complex opaque predicates and loop unrolling as well as function inlining and outlining, which existing approaches cannot defeat. Therefore, the proposed approach has stronger antiobfuscation ability compared with the existing approaches. From the perspective of the efficiency, the complexity of the proposed approach is higher than that of the existing approaches because it involves code instrumentation, code running, process mining, and graphs similarity measure. Therefore, the performance needs to be improved further. However, the approach can meet the actual requirement of code similarity detection in computer programming teaching. Meanwhile, the proposed approach can be combined with the existing approaches in the practical application of OJ and other systems. For example, some existing approach can be used first to measure the code similarity. Then, the proposed approach can be used further when the obtained similarity is lower than a given threshold. In addition, the dynamic flow chart of the source code obtained by process mining represents the dynamic feature of the source code by the graphical process, which can also provide the basis for code plagiarism.

As an improvement of existing static analysis-based approaches, the proposed approach explores source code similarity measure based on dynamic features. However, there are still some problems that need to be further investigated. First of all, the approach cannot defeat some more complex obfuscation techniques, such as adding sequential statements that have data dependence with the core process, adding reachable branches, and splitting expressions. Second, the efficiency of the proposed approach needs to be further improved because of its higher time complexity. Therefore, the following possible future works should be explored. First, combined with the existing static similarity detection approaches, the code similarity detection approaches combining static and dynamic features can be investigated. In addition, there is some work that uses machine learning techniques [30] to measure the similarity of the source code and obtain higher precision. Therefore, we can strengthen our proposed approach by machine learning techniques, especially the deep learning techniques [31]. Second, the source code instrumentation statements and rules of the proposed approach are relatively simple. As a result, for two code fragments, the obtained similarity may be higher than their real similarity in some cases. Therefore, the source code instrumentation statements and rules can be

further optimized to obtain more accurate similarity and deal with more complex code obfuscation techniques. Finally, the efficiency of the approach needs to be further improved to be applied to the similarity detection of larger-scale source code.

Data Availability

We took the code assignments submitted by students in the OJ system as data sets, and an encrypted version of the data sets is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Education Ministry Humanities and Social Science Research Youth Fund Project of China ("User-Steering Multi-Source Education Data Integration Approach Research in Big Data Environment" with grant number 19YJCZH240 and "Research on the Dynamic Evolution Tracking and Evaluation Method of Government's Internet Word-of-Mouth in Dealing With Emergencies Based on Big Data" with grant number 18YJAZH017); Qingdao Social Science Planning Research Project (grant number QDSKL1901123); the NSFC (grant numbers U1931207, 61902222, and 31671588); Sci. & Tech. Development Fund of Shandong Province of China (grant numbers 2016ZDJS02A11 and ZR2017MF027); Taishan Scholars Program of Shandong Province (tsqn201909109 and ts20190936); SDUST Research Fund (grant number 2015TDJH102); the Education and Teaching Research "Constellation" Project of Shandong University of Science and Technology (grant number QX2018M22); and SDUST Excellent Teaching Team Construction Plan (grant number JXTD20180503).

References

[1] S. Engels, V. Lakshmanan, and M. Craig, "Plagiarism detection using feature-based neural networks," *ACM SIGCSE Bulletin*, vol. 39, no. 1, pp. 34–38, 2007.

[2] L. Prechelt, G. Malpohl, and M. Philippsen, "Finding plagiarisms among a set of programs with JPlag," *Journal of Universal Computer Science*, vol. 8, no. 11, pp. 1016–1038, 2002.

[3] K. W. Bowyer and L. O. Hall, "Experience using MOSS to detect cheating on programming assignments," in *Proceedings of the FIE'99 Frontiers in Education. 29th Annual Frontiers in Education Conference. Designing the Future of Science and Engineering Education. Conference Proceedings*, vol. 3, pp. 18–22, San Juan, PR, USA, November 1999.

[4] D. Gitchell and N. Tran, "Sim," *ACM SIGCSE Bulletin*, vol. 31, no. 1, pp. 266–270, 1999.

[5] J. Feng, B. Cui, and K. Xia, "A code comparison algorithm based on AST for plagiarism detection," *Emerging Intelligent Data and Web Technologies*, vol. 1, pp. 393–397, 2013.

[6] C. Liu, C. Chen, J. Han et al., "Detection of software plagiarism by program dependence graph analysis," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 872–881, Philadelphia, PA, USA, August 2006.

[7] Z. Zhang, H. H. Yan, and X. W. Zhang, "Code similarity detection by program dependence graph," in *Proceedings of the International Conference on Computer Engineering and Information Systems*, pp. 255–261, Vienna, Austria, May 2016.

[8] Y. C. Jhi, X. Wang, X. Jia et al., "Value-based program characterization and its application to software plagiarism detection," in *Proceedings of the 33rd International Conference on Software Engineering*, pp. 756–765, Honolulu, HI, USA, May 2011.

[9] X. Wang, Y. C. Jhi, S. Zhu et al., "Detecting software theft via system call based birthmarks," in *Proceedings of the Twenty-Fifth Computer Security Applications Conference*, pp. 149–158, Tokyo, Japan, December 2009.

[10] H.-i. Lim, H. Park, S. Choi, and T. Han, "A method for detecting the theft of Java programs through analysis of the control flow information," *Information and Software Technology*, vol. 51, no. 9, pp. 1338–1350, 2009.

[11] Q. Song, *Research on Cross-Language Code Similarity Detection Method Based on Program Flow Chart*, Shandong University of Science and Technology, Qingdao, China, 2019.

[12] J. Ming, F. Zhang, D. Wu, P. Liu, and S. Zhu, "Deviation-based obfuscation-resilient program equivalence checking with application to software plagiarism detection," *IEEE Transactions on Reliability*, vol. 65, no. 4, pp. 1647–1664, 2016.

[13] D. Fu, Y. Xu, H. Yu, and B. Yang, "Wastk: a weighted abstract syntax tree kernel method for source code plagiarism detection," *Scientific Programming*, vol. 2017, pp. 1–8, 2017.

[14] S. Wasik, M. Antczak, J. Badura, A. Laskowski, and T. Sternal, "A survey on online judge systems and their applications," *Acm Computing Surveys*, vol. 51, no. 1, pp. 1–34, 2018.

[15] C. Ragkhitwetsagul, J. Krinke, and D. Clark, "A comparison of code similarity analysers," *Empirical Software Engineering*, vol. 23, no. 4, pp. 2464–2519, 2018.

[16] V. J. Marin and C. R. Rivero, "Towards a framework for generating program dependence graphs from source code," in *Proceedings of the 4th ACM SIGSOFT International Workshop*, pp. 30–36, Lake Buena Vista, FL, USA, November 2018.

[17] M. Novak, D. Kermek, and M. Joy, "Calibration of source-code similarity detection tools for objective comparisons," in *Proceedings of the 41st International Convention on Information and Communication Technology*, Electronics and Microelectronics, Opatija, Croatia, pp. 0794–0799, May 2018.

[18] M. J. Mišić, D. V. Nikolov, J. Ž. Protić et al., "Parallelization of GST algorithm for source code similarity detection," in *Proceedings of the 24th Telecommunications Forum*, IEEE, Belgrade, Serbia, pp. 1–4, November 2016.

[19] M. Novak, M. Joy, and D. Kermek, "Source-code similarity detection and detection tools used in academia: a systematic review," *Acm Transactions on Computing Education*, vol. 19, no. 3, pp. 27–37, 2019.

[20] E. L. Jones, "Metrics based plagiarism monitoring," *Journal of Computing Sciences in Colleges*, vol. 16, no. 4, pp. 253–261, 2001.

[21] G. Myles and C. Collberg, "Detecting software theft via whole program path birthmarks," *Lecture Notes in Computer Science*, vol. 3225, pp. 404–415, 2004.

[22] G. Whale, "Software metrics and plagiarism detection," *Journal of Systems and Software*, vol. 13, no. 2, pp. 131–138, 1990.

[23] T. Kamiya, S. Kusumoto, and K. Inoue, "CCFinder: a multilinguistic token-based code clone detection system for large scale source code," *IEEE Transactions on Software Engineering*, vol. 28, no. 7, pp. 654–670, 2002.

[24] M. Gabel, L. Jiang, and Z. Su, "Scalable detection of semantic clones," in *Proceedings of the 30th International Conference on Software Engineering (ICSE'08)*, pp. 321–330, New York, NY, USA, May 2008.

[25] L. Jiang, Z. Su, G. Misherghi et al., "DECKARD: Scalable and accurate tree-based detection of code clones," in *Proceedings of the 29th International Conference on Software Engineering*, Minneapolis, MN, USA, May 2007.

[26] R. Komondoor and S. Horwitz, "Using slicing to identify duplication in source code," in *Proceedings of the 8th International Symposium on Static Analysis*, pp. 40–56, Paris, France, July 2001.

[27] D. Qiu, J. Sun, and H. Li, "Improving similarity measure for Java programs based on optimal matching of control flow graphs," *International Journal of Software Engineering and Knowledge Engineering*, vol. 25, no. 7, pp. 1171–1197, 2015.

[28] L. Maruster, A. J. M. M. Weijters, W. M. P. V. D. Aalst et al., "Process mining: discovering direct successors in process logs," in *Proceedings of the Discovery Science International Conference, DS 2002*, Lübeck, Germany, November 2002.

[29] Z. Abu-Aisheh, R. Raveaux, J. Y. Ramel et al., "An exact graph edit distance algorithm for solving pattern recognition problems," in *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, vol. 1, pp. 271–278, Lisbon, Portugal, January 2015.

[30] Z. Huang, J. Tang, G. Shan, J. Ni, Y. Chen, and C. Wang, "An efficient passenger-hunting recommendation framework with multitask deep learning," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7713–7721, 2019.

[31] S. Zhao, D. M. Zhang, and H. W. Huang, "Deep learning–based image instance segmentation for moisture marks of shield tunnel lining," *Tunnelling and Underground Space Technology*, vol. 95, no. 1, pp. 1–11, 2020.

Hindawi
Scientific Programming
Volume 2020, Article ID 8818588, 7 pages
https://doi.org/10.1155/2020/8818588

Research Article

Conquering Gender Stereotype Threat in "Digit Sports": Effects of Gender Swapping on Female Players' Continuous Participation Intention in ESports

Lujie Hao,[1,2] Qinghua Lv,[1] Xiaosan Zhang,[3] Qingquan Jiang,[4] Songxian Liu,[4] and Lin Ping[5]

[1]*School of Business, Huaqiao University, Quanzhou 362021, China*
[2]*Economics and Business Administration Institute, Fuzhou University of Foreign Studies and Trade, Fuzhou 350202, China*
[3]*Research Centre for Belt & Road Financial and Economic Development, Xiamen National Accounting Institute, Xiamen 361005, China*
[4]*School of Economics & Management, Xiamen University of Technology, Xiamen 361024, China*
[5]*Chinese Entrepreneurship Research Center, Minjiang University, Fuzhou 350108, China*

Correspondence should be addressed to Xiaosan Zhang; zxs@xnai.edu.cn and Qingquan Jiang; jiangqingquan@xmut.edu.cn

Received 14 April 2020; Revised 30 May 2020; Accepted 16 June 2020; Published 6 July 2020

Academic Editor: Chenxi Huang

As a sportification form of human-computer interaction, eSports is facing great gender stereotype threat and causing female players' withdraw. This study aims to investigate the relationship between gender-swapping and females' continuous participation intention in eSports, the mediating effect of self-efficacy, and the moderating effect of discrimination. The results demonstrate (1) that the effect of gender-swapping on continuous participation intention in eSports was not significant, while gender-swapping had a significant association with self-efficacy, and self-efficacy had a significant association with continuous participation intention in eSports; (2) that gender-swapping had an indirect effect (via self-efficacy) on continuous participation intention in eSports; and (3) that discrimination moderated the effect of self-efficacy on continuous participation intention. Female players who had experienced discrimination displayed higher continuous participation intention in the context of self-efficacy enhanced by gender-swapping.

1. Introduction

ESports is a typical sportification form [1] of human-computer interaction [2, 3]. Similar to traditional sport, gender stereotype threat exists in eSports and causes a significant gender imbalance in the eSports industry. Taking Chinese eSports as an example, female participants account for only 24.3% [4]. Although emerging mobile eSports attract more and more female participants, eSports are still typically male-dominated and are full of negative stereotypes of females, such as that male players are more competent [5]. Therefore, female players often receive less positive comments [6] and experience discrimination in eSports [7]. Females are considered to have a lower skill level and require more protection [8], and hypersexualized female avatars make male players associate female players with sex, and this has a negative impact [9]. Furthermore, discrimination causes female players to experience greater oppression in the game [5], forcing them to withdraw from eSports.

It would seem that gender stereotypes threat discourages female participation in eSports. Therefore, female players sometimes attempt to dispel gender stereotypes by gender-swapping [10–12]. A few initial studies have found that gender-swapping impacts females' gaming psychology and experience [13], including both online and offline behaviors [10]. However, the relationship between gender-swapping and females' continuous participation intention in eSports is little discussed.

The present study aimed to explore the relationship between gender-swapping, self-efficacy, and females' continuous participation intention in eSports in a sample of Chinese female college students. A moderated mediation model was tested, in which self-efficacy mediated the relationship between gender-swapping and continuous participation intention in eSports, and discrimination moderated the relationship between gender-swapping and self-efficacy and the mediation effect of self-efficacy.

1.1. Relationship between Gender-Swapping and Continuous Participation Intention in ESports. Online gaming permits players to mask their actual gender, and players can create an identity opposite to their actual gender through gender-swapping. Gender-swapping is common among players [14]. Griffiths et al. conducted a survey of 8694 players and found that 10% of women had utilized gender-swapping during the game [11]. In Hussain and Griffiths' survey, the gender-swapping ratio was as high as 68% [12]. According to Fox and Tang , females may try to avoid gender discrimination by gender-swapping [15]. In this way, gender-swapping has become a way of making gaming fair for female players [12]. Royse et al. found that a female's psychological experience changes when they control male avatars [13]. In addition, gender-swapping changes the offline behavior of female players. Huh and Williams found that female players who conduct long-term gender-swapping display more masculine behaviors [10].

Although more and more females are participating in eSports, the title of "gamer" is considered to be exclusive to males [16]. For females, eSports is rife with stereotype threat [17, 18] and discrimination [19, 20]. Shen et al. found that although female players have enough game skills, some still feel that their ability is inferior to that of males of the same level [21], which caused dearth of female players in this eSports community [22]. This lack of confidence comes from stereotypes of female players [23]. Perceiving negative gender stereotypes decreases female players' performance [17] and excludes them from the mainstream eSports circle [24], which weakens their participation intention in eSports; in other words, the higher the females' awareness of the threat of gender stereotypes, the more likely they are to quit eSports. Removing this gender stereotype threat will play an important role in enhancing female participation in eSports; gender-swapping plays a role in "gender masking" and allows female players to escape stereotype threats which can enhance females' continuous participation intention in eSports. Thus, we propose the following hypotheses.

Hypothesis 1. Gender-swapping is positively associated with females' continuous participation intention in eSports.

1.2. The Mediating Role of Self-Efficacy. According to the Proteus effect theory, players' behavior and cognition is consistent with their game avatar's characteristics. For example, players who chose a "good-looking" avatar showed more intimate online social behavior [25]. An avatar's gender also affects players' behavior and cognition: Peña et al. found that when randomly assigned to a female avatar, both male and female participants showed a feminine communication paradigm (e.g., emotional and often apologized) [26]. Lehdonvirta et al. confirmed that male players assigned to male avatars are less help-seeking than those with female avatars [27]. One study [10] reported that females who manipulated male avatars were more likely to display male behaviors. These conclusions indicate that players tend to exhibit a behavior consistent with the avatar's gender stereotype. For eSports games, male avatars generally have aggressive attributes and female avatars often have auxiliary attributes, thus creating the stereotype that male avatars are stronger than female avatars. Therefore, when females control male avatars, they behave in a manner consistent with the stereotypes of male avatars. Through gender-swapping, they are classified as nonstereotype target groups and improve their performance by the stereotype lift effect. Lee et al. demonstrated that regardless of participants' gender, when assigned to female avatars, their mathematics test scores were lower than the scores of those who were assigned to male avatars, confirming that an avatar's gender influences a participant's performance [28]. Similarly, in a study in which all participants were female, those who were assigned to male avatars gained better scores than those with female avatars, further confirming that female players display a better performance when engaged in gender-swapping [29]. Kaye and Pennington found that when associated with a nonstigmatized gaming identity, females' performance deficits may be mitigated [17]. Royse et al. argued that female players have a superior feeling when gender-swapping [13], indicating that female players construct a game identity through gender-swapping which enhances their game self-efficacy.

Social cognitive theory emphasizes the influence of self-efficacy on an individual's behavior and tendencies, suggesting that people like to engage in things that they are more capable at, which means that if an individual is more self-effective about accomplishing the task, he/she will have a greater intention to participate. In the field of information systems, Tsai et al. found that self-efficacy affected a user's intention of using technology [30]. In the gaming field, players' judgment of self-ability has been shown to affect their decision to continue playing games. If players consider that they can gain game skills easily, they tend to have a strong continuous intention; otherwise, they quit playing [31]. On the other hand, when encountering a strong adversary or difficult situations, players with a high self-efficacy who keep fighting and gain skill or level promotions also experience a sense of accomplishment [32]. These positive experiences result in a higher gaming frequency and a continuous intention to play [33]. Therefore, female players' self-efficacy is positively correlated with their continuous participation intention in eSports. Hence, we propose the following two hypotheses.

Hypothesis 2. Gender-swapping is positively associated with self-efficacy and self-efficacy is positively associated with the participating intention of female players in

e-sports, and gender-swapping has indirect effects on the participating intention of female players in eSports through self-efficacy.

1.3. The Role of Discrimination. Hussain and Griffiths reported that female players have a higher gender-swapping ratio, and removing the gender stereotype threat is the main motivation for gender-swapping [12]. One study on First Person Shooting games showed that 75.9% of the females surveyed said they had been discriminated against in the game [16]. According to stereotype threat theory, the experience of being discriminated against initiates female players' awareness of negative gender stereotypes, resulting in a negative gaming experience and worse performance [18]. Behm-Morawitz and Mastro confirmed that females who have been discriminated against had lower game self-efficacy [34]; female players wanted to be treated fairly and tried to prove that they could play as good as males [16]. Because of the stereotype lift effect, female players perform better when utilizing gender-swapping. In addition, the more the female players are afflicted with negative gender stereotypes, the stronger the stereotype lift effect they gain through gender-swapping. Peck et al. proved that, in the context of gender stereotype threat priming, females who were assigned to male avatars gained significantly higher memory test scores than those with female avatars [35], indicating that stereotype threat priming moderated the stereotype lift effect. For this study, the experience of discrimination was the factor priming female players' gender stereotype threats. As a result, female players who have experienced discrimination gain stereotype lift and achieve higher self-efficacy would induce higher continuous participation intention than those who have not been discriminated against, which means that discrimination can moderate the effect of g self-efficacy on continuous participation intention and the mediating effect of self-efficacy. Based on this, we propose the following hypotheses.

Hypothesis 3. Discrimination moderates the relationship of self-efficacy with continuous participation intention of female players in eSports.

Hypothesis 4. Discrimination moderates the mediating effect of self-efficacy of gender-swapping on the continuous participation intention of female players in eSports.

2. Methods

2.1. Participants. A total of 637 female undergraduates, enrolled in six large Southeastern public universities in China, mainly aged between 18 and 22 years, participated in the study. These students listed MOBA (Multiplayer Online Battle Arena) and TPS (Third Person Shooting game) as the eSports genre they mainly played. After eliminating the invalid questionnaires, a total of 475 valid questionnaires were collected.

2.2. Measures

2.2.1. Gender-Swapping. ESports players have several avatars, and gender-swapping happens in certain rounds, but not in all rounds. In order to measure the extent of female players' gender-swapping behavior, this study adapted the gender-swapping scale developed by Kuwik and Baluch (2010). The three items measure the time females spent controlling a male avatar (e.g., I would like to spend more time controlling a male avatar), the degree of female avatar ownership (e.g., I would like to have more male avatars than female avatars), and the degree of enjoyment of playing as a male avatar (e.g., I would prefer and enjoy controlling male avatars). Participants rated their answers on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), with a higher score on the inventory indicating a higher extent of gender-swapping [36]. Cronbach's α was 0.908 ($M = 2.48$, SD = 1.00).

2.2.2. Self-Efficacy. Self-efficacy is the players' self-judgment of their ability to achieve a specific task in eSports. This study used an adapted version of the self-efficacy scale compiled by Yao (2006). The five-item scale measured the female players' self-efficacy in eSports (e.g., I have already gained game skills.). Participants rated their answers on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), with a higher score on the inventory indicating better self-efficacy [37]. Cronbach's α was 0.821 ($M = 3.28$, SD = 0.79).

2.2.3. Continuous Participation Intention in ESports. The continuous participation intention in eSports was measured by a scale compiled by Yao (2006). The two-item scale measured the female players' continuous participation intention in eSports (e.g., I will continue to play this game in future). Participants rated their answers on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), with a higher score on the inventory indicating a higher continuous participation intention [37]. Cronbach's α was 0.847 ($M = 3.32$, SD = 0.84).

2.2.4. Discrimination. The experience of discrimination is the priming factor of the stereotype threat for female players. In this study, this variable was a dichotomous yes-no measure that indicated whether the female players had been discriminated against in eSports.

2.3. Statistical Analysis. Pearson's correlations and descriptive statistics were used to test all the variables. Three sets of hierarchical regression analyses were performed to examine the relationships between variables in this study model. The PROCESS macro for SPSS (model 4) was employed to test mediation models [38], and through adding discrimination as a moderator, moderated mediation models were further examined (PROCESS model 14, see Figure 1). In both the mediation analyses and moderated mediation analyses, effects were calculated using

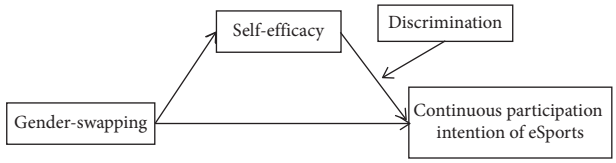


Figure 1: The research model.

5000 bootstrapping samples, generating 95% confidence intervals of the bias-corrected bootstrap type, and were considered significant when 0 fell out of the confidence interval [39].

3. Results

Means, standard deviations, and the intercorrelations for the variables utilized in each analysis can be found in Table 1. Gender-swapping was not significantly associated with a continuous participation intention in eSports ($r = 0.081$, $p > 0.05$). Gender-swapping was associated with self-efficacy ($r = 0.144$, $p < 0.01$), which was associated with a continuous participation intention in eSports ($r = 0.483$, $p < 0.01$).

Then, the proposed moderated mediation model was tested (see Table 2). Gender-swapping was a positive predictor of self-efficacy ($\beta = 0.11$, $p < 0.01$), and self-efficacy was a positive predictor of continuous participation intention (CPI) ($\beta = 0.45$, $p < 0.001$). Moreover, gender-swapping was not significantly associated with CPI ($\beta = 0.016$, $p > 0.05$), and H_1 was not supported. The PROCESS Model 4 was employed to test the indirect effect of gender-swapping on CPI. The indirect effect indicated that self-efficacy mediated the effect of gender-swapping on CPI ($\beta = .052$, SE = 0.021, $p < 0.05$, CI [0.014, 0.096]), and H_2 was supported.

To address H3 and H4, PROCESS Model 14 was conducted. Results showed a significant moderating effect of discrimination between self-efficacy and continuous participation intention ($\beta = -0.22$, SE = 0.09, 95% CI = [−0.40, −0.043]), and H_3 was supported. The interaction was plotted at the no discrimination experience group and discrimination experience group (see Figure 2). As expected, when self-efficacy was high, the discrimination experience group had a higher continuous participation intention compared to the discrimination experience group. Moderated mediation analyses showed that self-efficacy remained a significant mediation effect for both the discrimination group ($\beta = 0.062$, SE = 0.030, 95% CI = [0.01, 0.12]) and the no discrimination group ($\beta = 0.038$, SE = 0.017, 95% CI = [0.0075, 0.076]), but pairwise contrasts between conditional indirect effects were insignificant ($\beta_{\text{contrast}} = -0.025$, SE = 0.018, 95% CI = [−0.067, 0.0007]). Therefore, discrimination did not moderate the indirect effect of self-efficacy (see Table 3), and H_4 was not supported.

Table 1: Descriptive statistics and Pearson's correlations between the variables.

Variable	1	2	3	4
1. Discrimination	1			
2. Gender-swapping	0.03	1		
3. Self-efficacy	−0.049	0.144**	1	
4. Continuous participation intention	−0.09	0.081	0.483**	1
M	1.438	2.48	3.28	3.32
SD	0.497	1.00	0.79	0.84

Note. ** $p < 0.01$, * $p > 0.05$, M = mean, SD = standard deviation.

4. Discussion

ESports, like traditional sports, create an unfair environment for females to experience discrimination [40–44], which can makes them more likely to withdraw from eSports altogether. To escape gender stereotypes, female players may choose male avatars when participating in eSports. Gender-swapping affects a female player's gaming psychology [13] and performance [28, 29]. However, questions regarding the relationship between gender-swapping and a female player's participation in eSports and the influence path remain largely unknown. We formulated a moderated mediation model to answer these questions. This study aimed to extend the knowledge on gender-swapping in the eSports literature.

First, regarding our first aim focusing on the association between gender-swapping and a female's continuous participation intention in eSports, the relationship between the two factors' was not significant ($\beta = 0.068$, $p > 0.05$), suggesting that gender-swapping does not directly affect females' continuous participation intention. The analysis showed that gender-swapping positively predicted self-efficacy, and self-efficacy positively predicted females' continuous participation intention, and the mediation test highlighted the mediation role of self-efficacy. According to the theory of stereotype lift, when female players control male avatars, they classify themselves as part of non-stereotype target groups, gain a stereotype lift effect, and enhance their self-efficacy and game performance; a high level of self-efficacy improves females' continuous participation intention in eSports. The conclusions are consistent with the previous finding [13, 28, 29], revealing that gender-swapping affects continuous participation intention in eSports for female players indirectly.

This study further found that discrimination moderated the relationship between g self-efficacy and continuous participation intention. This finding can be explained by stereotype threat theory [45]. Compared with female players with no discrimination experience, females who have been discriminated against are more aware of the negative stereotypes of female players. Once the female players gain stereotypes lift through gender-swapping, mitigating performance deficits and enhancing self-efficacy [17], they would have a higher continuous participation intention. The result is supported by a previous study [35].

The abovementioned findings have interesting practical implications. Avatars are the self-presentation or extension of players in the virtual world [46], and players can choose an avatar with a gender opposite to their own. In the eSports context, females can improve their self-efficacy through gender-swapping, and felt that they had the ability to compete with their opponents, thus enhancing their

Table 2: Results of simple mediation analysis (process model 4).

(a) Result of simple mediation analysis								
Outcome	SEF				CPI			
Predictor	β(SE)	p	95% CI lower	95% CI upper	β(SE)	p	95% CI lower	95% CI upper
GS	0.11 (0.036)	<0.005	0.043	0.18	0.016(0.035)	0.066	−0.053	0.084
SEF	—	—	—	—	0.46 (0.045)	<0.001	0.38	0.55
(b) Result of the indirect effect								
GS⟶SEF⟶CPI						Effect (SE)	95% CI lower	95% CI upper
						0.052 (0.021)	0.014	0.096

Note. GS = gender-swapping, SEF = self-efficacy, CPI = continuous participation intention, SE = standard error.

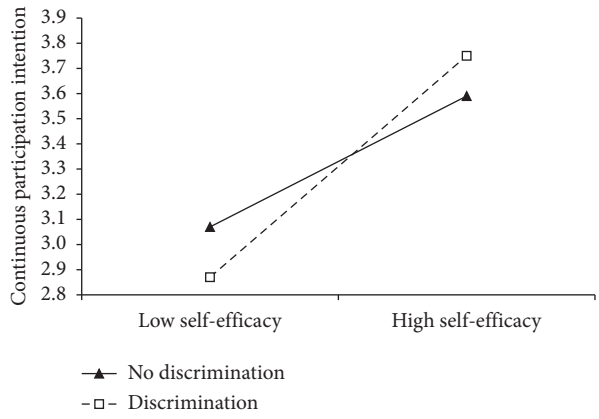


Figure 2: Moderating effect of discrimination for self-efficacy on continuous participation intention.

Table 3: Results of the moderated mediated model (process model 14).

(a) Results of the moderated mediated model analysis								
Outcome	SEF				CPI			
Predictor	β(SE)	p	95% CI lower	95% CI upper	β(SE)	p	95% CI lower	95% CI upper
GS	0.011 (0.036)	0.0017	0.043	0.18	0.018 (0.035)	0.57	−0.049	0.089
DIS	—	—	—	—	0.74 (0.30)	<0.05	0.14	1.33
SEF	—	—	—	—	0.77 (0.13)	<0.001	0.51	1.03
SEF∗DIS	—	—	—	—	−0.22 (0.090)	<0.05	−0.40	−0.043
(b) Results of the conditional indirect effect of self-efficacy at different status of discrimination								
					Discrimination experience	Effect (SE)	95% CI lower	95% CI upper
GS⟶SEF⟶CPI					No (effect 1)	0.038 (0.017)	0.0075	0.072
					Yes (effect 2)	0.062 (0.026)	0.016	0.12
					Effect 1 minus effect 2	−0.24(0.018)	−0.067	0.0007

Note. GS = gender-swapping, SEF = self-efficacy, CPI = continuous participation intention, SE = standard error.

continuous participation intention in eSports. This was especially the case for females who had been discriminated against, for whom the continuous participation intention was higher. Therefore, eSports operators could add certain "masculine" elements to female avatars, such as a more muscular appearance, to moderately increase the ability of female avatars and weaken the impact of gender stereotypes on female players. On the other hand, the gender stereotype threat results the female's underrepresentation and performance in math [47] and STEM [48] that discourage them pursuing science and technology [49]. While online learning becomes popular, educators could create a learning application with avatar attributes; female students would benefit from this human-computer interaction way by embodying male avatars and have better performance [35].

This study has some limitations. First, the current sample was cross-sectional and used a Chinese female college sample, which lacked external validity. Although this group represents the main female participants in the current eSports, future research could expand the participants' ages and occupations. Second, the effect of gender-swapping on the female players' intention to continue participating in eSports may also be influenced by the female player's individual traits. Future research should consider female

players with different personality traits. Third, the scale measuring continuous participation intention only contained two items, and whilst the scale's Cronbach's α was 0.847, showing an acceptable internal consistency, using more scale items in replications may be more effective. Fourth, this study neglected the antecedents of females' gender-swapping. Fox and Tang argued that discrimination is an antecedent of female players' gender-swapping [15], but the association between discrimination and gender-swapping was not significant in this study, which showed that discrimination does not lead to gender-swapping by female players in eSports. It remains a task of future research to explore the antecedents of females' gender-swapping behavior.

5. Conclusions

In summary, this study makes some theoretical contributions. In eSports, as a sportification form of human-computer interaction, females still experience great gender inequity and gender stereotypes. Different from traditional sports participants, part of female players in eSports use gender-swapping to dispel the negative impact of gender inequity. This study provides an empirical framework for testing the indirect relationship between gender-swapping and females' continuous participation intention in eSports and the moderating effect of discrimination. These results can shed light on the underlying mechanism between gender-swapping and females' continuous participation intention in eSports. Our findings may help the eSports industry to create a sustainable business environment by promoting female participation.

Data Availability

The data used to support the findings of this study are available from the first author or corresponding author upon request anytime.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Social Science Foundation of Fujian (grant nos. FJ2018B062 and FJ2019B101), Xiamen Science and Technology Plan Project (grant no. 2018S2247), Outstanding Young Academic Cultivation Plan Project of Fujian (grant no. JYTJQ201804), and Fujian Provincial Department of Science and Technology, Soft Science Research Plan Project (grant nos. 2019R0093 and 2019R0094).

References

[1] G. B. Cunningham, S. Fairley, L. Ferkins et al., "ESport: construct specifications and implications for sport management," *Sport Management Review*, vol. 21, no. 1, pp. 1–6, 2018.

[2] M. Pluss, K. J. M. Bennett, A. R. Novak, D. Panchuk, A. Coutts, and J. Fransen, "Esports: the chess of the 21st century," *Frontiers in Psychology*, vol. 10, p. 156, 2019.

[3] J. Hamari and M. Sjöblom, "What is eSports and why do people watch it?" *Internet Research*, vol. 27, no. 2, 2017.

[4] Tencent, *Report on the Development of China's eSports Industry in 2018*, Tencent, Shenzhen, China, 2019.

[5] L. Vermeulen, E. Núñez Castellar, and J. Van Looy, "Challenging the other: exploring the role of opponent gender in digital game competition for female players," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 5, pp. 303–309, 2014.

[6] O. Ruvalcaba, J. Shulze, A. Kim, S. R. Berzenski, and M. P. Otten, "Women's experiences in eSports: gendered differences in peer and spectator feedback during competitive video game play," *Journal of Sport and Social Issues*, vol. 42, no. 4, pp. 295–311, 2018.

[7] L. K. Kaye, C. E. Gresty, and N. Stubbs-Ennis, "Exploring stereotypical perceptions of female players in digital gaming contexts," *Cyberpsychology, Behavior, and Social Networking*, vol. 20, no. 12, pp. 740–745, 2017.

[8] A. Brehm, "Navigating the feminine in massively multiplayer online games: gender in World of warcraft," *Frontiers in Psychology*, vol. 4, p. 903, 2013.

[9] M. Z. Yao, C. Mahood, and D. Linz, "Sexual priming, gender stereotyping, and likelihood to sexually harass: examining the cognitive effects of playing a sexually-explicit video game," *Sex Roles*, vol. 62, no. 1-2, pp. 77–88, 2010.

[10] S. Huh and D. Williams, *Dude Looks like a Lady: Gender Swapping in an Online Game*, pp. 161–174, Springer, Berlin, Germany, 2010.

[11] M. D. Griffiths, M. N. O. Davies, and D. Chappell, "Breaking the stereotype: the case of online gaming," *CyberPsychology & Behavior*, vol. 6, no. 1, pp. 81–91, 2003.

[12] Z. Hussain and M. D. Griffiths, "Gender swapping and socializing in cyberspace: an exploratory study," *CyberPsychology & Behavior*, vol. 11, no. 1, pp. 47–53, 2008.

[13] P. Royse, J. Lee, B. Undrahbuyan, M. Hopson, and M. Consalvo, "Women and games: technologies of the gendered self," *New Media & Society*, vol. 9, no. 4, pp. 555–576, 2007.

[14] M. Boler, "Hypes, hopes and actualities: new digital Cartesianism and bodies in cyberspace," *New Media & Society*, vol. 9, no. 1, pp. 139–168, 2007.

[15] J. Fox and W. Y. Tang, "Women's experiences with general and sexual harassment in online video games: rumination, organizational responsiveness, withdrawal, and coping strategies," *New Media & Society*, vol. 19, no. 8, pp. 1290–1307, 2017.

[16] M. A. McDaniel, *Women in Gaming: A Study of Female Players' Experiences in Online FPS Games*, pp. 1–34, The University of Southern Mississippi, Hattiesburg, MS, USA, 2016.

[17] L. K. Kaye and C. R. Pennington, ""Girls can't play": the effects of stereotype threat on females' gaming performance," *Computers in Human Behavior*, vol. 59, pp. 202–209, 2016.

[18] L. Vermeulen, E. N. Castellar, D. Janssen, L. Calvi, and J. Van Looy, "Playing under threat. Examining stereotype threat in female game players," *Computers in Human Behavior*, vol. 57, pp. 377–387, 2016.

[19] K. L. Gray, "Deviant bodies, stigmatized identities, and racist acts: examining the experiences of African-American gamers in Xbox live," *New Review of Hypermedia and Multimedia*, vol. 18, no. 4, pp. 261–276, 2012.

[20] J. H. Kuznekoff and L. M. Rose, "Communication in multiplayer gaming: examining player responses to gender cues," *New Media & Society*, vol. 15, no. 4, pp. 541–556, 2013.

[21] C. Shen, R. Ratan, Y. D. Cai, and A. Leavitt, "Do men advance faster than women? debunking the gender performance gap in two massively multiplayer online games," *Journal of Computer-Mediated Communication*, vol. 21, no. 4, pp. 312–329, 2016.

[22] R. A. Ratan, N. Taylor, J. Hogan, T. Kennedy, and D. Williams, "Stand by your man: an examination of gender disparity in league of legends," *Games and Culture*, vol. 10, no. 5, pp. 438–462, 2015.

[23] J. L. Smith, "Understanding the process of stereotype threat: a review of mediational variables and new performance goal directions," *Educational Psychology Review*, vol. 16, no. 3, pp. 177–206, 2004.

[24] L. Vermeulen, S. Van Bauwel, and J. Van Looy, "Tracing female gamer identity. An empirical study into gender and stereotype threat perceptions," *Computers in Human Behavior*, vol. 71, pp. 90–98, 2017.

[25] N. Yee and J. Bailenson, "The Proteus effect: the effect of transformed self-representation on behavior," *Human Communication Research*, vol. 33, no. 3, pp. 271–290, 2007.

[26] J. Peña, J. T. Hancock, and N. A. Merola, "The priming effects of avatars in virtual settings," *Communication Research*, vol. 36, no. 6, pp. 838–856, 2009.

[27] M. Lehdonvirta, Y. Nagashima, V. Lehdonvirta, and A. Baba, "The stoic male," *Games and Culture*, vol. 7, no. 1, pp. 29–47, 2012.

[28] J.-E. R. Lee, C. I. Nass, and J. N. Bailenson, "Does the mask govern the mind?: effects of arbitrary gender representation on quantitative task performance in avatar-represented virtual groups," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 4, pp. 248–254, 2014.

[29] R. Ratan and Y. J. Sah, "Leveling up on stereotype threat: the role of avatar customization and avatar embodiment," *Computers in Human Behavior*, vol. 50, pp. 367–374, 2015.

[30] P. Tsai, C. Tsai, and G. Hwang, "Elementary school students' attitudes and self-efficacy of using PDAs in a ubiquitous learning context," *Australasian Journal of Educational Technology*, vol. 26, no. 3, pp. 297–308, 2010.

[31] D. Choi and J. Kim, "Study on the effect of the cognitive performance, self-efficacy and self-esteem on the players' flow experience during playing online games," *Journal of Korea Game Society*, vol. 13, no. 6, pp. 5–14, 2013.

[32] S. Trepte and L. Reinecke, "The pleasures of success: game-related efficacy experiences as a mediator between player performance and game enjoyment," *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 9, pp. 555–557, 2011.

[33] J. E. Bigné, L. Andreu, and J. Gnoth, "The theme park experience: an analysis of pleasure, arousal and satisfaction," *Tourism Management*, vol. 26, no. 6, pp. 833–844, 2005.

[34] E. Behm-Morawitz and D. Mastro, "The effects of the sexualization of female video game characters on gender stereotyping and female Self-Concept," *Sex Roles*, vol. 61, no. 11-12, pp. 808–823, 2009.

[35] T. C. Peck, M. Doan, K. A. Bourne, and J. J. Good, "The effect of gender body-swap illusions on working memory and stereotype threat," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1604–1612, 2018.

[36] A. Kuwik and B. Baluch, "Gender differences in polish citizen's attitudes towards immigration to the UK," *Europe's Journal of Psychology*, vol. 6, no. 4, pp. 192–208, 2010.

[37] T. Yao, *Research on the Continuous Use of Online Games Based on the Extended Theory of Planned Behavior*, Zhejiang University, Zhejiang, China, 2006, in Chinese.

[38] A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, Guilford Publications, New York, NY, USA, 2017.

[39] A. F. Hayes, "An index and test of linear moderated mediation," *Multivariate Behavioral Research*, vol. 50, no. 1, pp. 1–22, 2015.

[40] J. Fox and W. Y. Tang, "Sexism in online video games: the role of conformity to masculine norms and social dominance orientation," *Computers in Human Behavior*, vol. 33, pp. 314–320, 2014.

[41] M. E. Ballard and K. M. Welch, "Virtual warfare," *Games and Culture*, vol. 12, no. 5, pp. 466–491, 2017.

[42] S. Chess and A. Shaw, "A conspiracy of fishes, or, how we learned to stop worrying about #GamerGate and embrace hegemonic masculinity," *Journal of Broadcasting & Electronic Media*, vol. 59, no. 1, pp. 208–220, 2015.

[43] M. Consalvo, "Confronting toxic gamer culture: a challenge for feminist game studies scholars," *Ada: A Journal of Gender, New Media, and Technology*, vol. 1, no. 1, pp. 1–6, 2012.

[44] A. Salter and B. Blodgett, "Hypermasculinity & dickwolves: the contentious role of women in the new gaming public," *Journal of Broadcasting & Electronic Media*, vol. 56, no. 3, pp. 401–416, 2012.

[45] C. M. Steele and J. Aronson, "Stereotype threat and the intellectual test performance of African Americans," *Journal of Personality and Social Psychology*, vol. 69, no. 5, pp. 797–811, 1995.

[46] S. Bengtsson, "Avatar as second suit: power and participation in virtual work," *Games and Culture*, vol. 14, no. 5, pp. 560–578, 2019.

[47] C. Good, J. Aronson, and J. A. Harder, "Problems in the pipeline: stereotype threat and women's achievement in high-level math courses," *Journal of Applied Developmental Psychology*, vol. 29, no. 1, pp. 17–28, 2008.

[48] N. C. Krämer, B. Karacora, G. Lucas, M. Dehghani, G. Rüther, and J. Gratch, "Closing the gender gap in STEM with friendly male instructors? on the effects of rapport behavior and gender of a virtual agent in an instructional interaction," *Computers & Education*, vol. 99, pp. 1–13, 2016.

[49] Y. Yücel and K. Rızvanoğlu, "Battling gender stereotypes: a user study of a code-learning game, "Code Combat," with middle school children," *Computers in Human Behavior*, vol. 99, pp. 352–365, 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8894633, 7 pages
https://doi.org/10.1155/2020/8894633

Research Article

Use Chou's 5-Step Rule to Classify Protein Modification Sites with Neural Network

Chuandong Song and Bin Yang

School of Information Science and Engineering, Zaozhuang University, Zaozhuang, Shandong 277160, China

Correspondence should be addressed to Bin Yang; batsi@126.com

Received 15 April 2020; Revised 15 June 2020; Accepted 17 June 2020; Published 3 July 2020

Academic Editor: Chenxi Huang

Lysine malonylation is a novel-type protein post-translational modification and plays essential roles in many biological activities. Having a good knowledge of malonylation sites can provide guidance in many issues, including disease prevention and drug discovery and other related fields. There are several experimental approaches to identify modification sites in the field of biology. However, these methods seem to be expensive. In this study, we proposed malNet, which employed neural network and utilized several novel and effective feature description methods. It was pointed that ANN's performance is better than other models. Furthermore, we trained the classifiers according to an original crossvalidation method named Split to Equal validation (SEV). The results achieved AUC value of 0.6684, accuracy of 54.93%, and MCC of 0.1045, which showed great improvement than before.

1. Introduction

Protein post-translational modification (PTM) is a key mechanism to regulate protein functions by the covalent and generally enzymatic modification. Hundreds of types of PTMs have been discovered and reported in this field [1–6]. They played vital roles in influencing almost all aspects of cell biology and pathogenesis, e.g., gene expression, cell division, and cell signaling [7–10]. As one of a newly identified PTM type in both eukaryotic and prokaryotic, Lysine malonylation (Kmal) has wide connections with various biological processes, where some Kmal sites are potentially associated with cancer. Therefore, it is very critical to identify and understand Kmal sites in the studies of biology and diseases [11–16].

Different from traditional experimental methods, computational approach of PTMs provides a fast and low-cost strategy for experimental designing, as the PTM site prediction can be abstracted as a typical classification problem. Meanwhile, there are a list of machine learning approaches which can be successful utilized in this field. For instance, Logistic Regression (LR) was used in ModPred for 23 different modifications using sequence-based features, physicochemical properties, and evolutionary features as features [17]. Musite, which is a general and kinase-specific protein phosphorylation site prediction, applied Support Vector Machine (SVM) models utilizing three types of features: K-nearest neighbor score, disorder scores, and amino acid frequencies. In the previous work, Wei et al. presented PhosPred-RF for predicting phosphorylation sites, which utilized the evolutionary information features from position specific scoring matrices [18, 19]. Deep learning method was also applied in this area, such as the recently published tool MusiteDeep, which was utilized for general and kinase-specific phosphorylation site prediction [20].

In this article, we employed artificial neural network (ANN) classifier based on Stochastic Gradient Descent (SGD) algorithm for protein Kmal site prediction. It was pointed that we investigated a wide range of types of feature extraction schemes and finally choose EBAG + Profile and EAAC methods to train our predictors. Furthermore, we employed another two classifiers, including SVM and kNN for comparative experiments, with the same feature extraction schemes. Besides, in view of the fact that the Kmal prediction problem can be regarded as a binary classification problem, we adopted the original SEV method to solve the inherent imbalance problem of positive and negative

samples in the training set. The result of our experiment shows ANN performed better than SVM and kNN predictors. Overall, ANN can be a useful tool for identifying Kmal sites.

2. Methods and Materials

There are 4 steps in our research, which is depicted in Figure 1. The first step is dataset construction and procession, where the training set and testing set were generated. And then we encode the dataset according to two feature extraction methods. The next step is to construct three models, which were trained by the training set. Finally, all the classifiers would be tested by crossvalidation and independent testing set. Five assessment metrics would be used to evaluate the performance of our classifiers.

2.1. Dataset Construction. In this work, we derived lots of Kmal peptides from mice and human species according to a proteomic assay. Referring to the procedure established by Chen et al. [21], we built a benchmark dataset. There are 67322 Kmal sites in the training set, where the sites with high confidence were regarded as positive sites and other lysine residues were collected as negative sites. For each sample, we extracted 31-residue peptides (−15 to +15) with the lysine site in the center from the representatives. As a result, 5023 positive peptides and 62299 negative peptides were retained for further analyses. We can easily find that the ratio of positive and negative samples in training set approaches 1 to 12. Therefore, the trained models would be tested by Split to Equal validation and independent test, where 35955 peptides (including 2798 positive peptides and 33157 negative peptides) were employed as the independent testing dataset.

2.2. Feature Encodings

2.2.1. EBAG + Profile Encoding. EBAG + Profile encoding is an integration scheme consisted of two different feature encoding method utilized by Han et al. [22]. One is Encoding Based on Attribute Grouping (EBAG) [23], which divides 20 types of amino acids into 5 groups depending on various physical and chemical properties. Table 1 shows the grouping result based on EBAG.

The other encoding method is Profile, which counts the frequency of each amino acid residue occurred in the protein peptides. And then the frequency was used as the representation of this residue in the sequence so that each peptide with 31 residues can be transformed into a 31-dimension vector. The way to combine EBAG and Profile is replacing source amino acid peptides into EGBA sequence and then encoding the sequence according to the Profile method. As a result, a peptide with 31 residues was converted to a vector of 31 dimensions as the EBAG + Profile encodings.

2.2.2. EAAC Encoding. A typical encoding scheme of prediction for PTMs was AAC encoding [24], which reflects the frequency of 20 amino acid residues surrounding the modification site. In this work, we coded each amino acid by

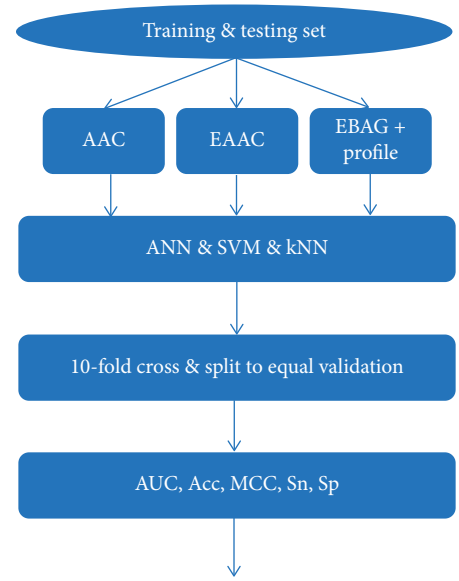


Figure 1: The working flow of our works.

Table 1: Groups of amino acid residues according to EBAG + Profile encoding.

Group	Amino acid residue	Label
C1	*A, F, G, I, L, M, P, V, W*	Hydrophobic
C2	*C, N, Q, S, T, Y*	Polar
C3	*D, E*	Acidic
C4	*H, K, R*	Basic
C5	*X*	Gaps

the EAAC method proposed by Zhen et al. [25], which is based on the AAC encoding. As 8-size window continuously slides from the N-terminus to C-terminus of each peptide in the dataset, the EAAC method counted the frequency of the 20 amino acid residues. Accordingly, the dimension of features can be calculated as follows:

$$N_s = L_p - L_s + 1,$$
$$D_{\text{eaac}} = N_s \times 20, \quad (1)$$

where L_p refers to the length of each peptide, L_s is the length of sliding windows, and D_{eaac} is the dimension of feature vector. As we set L_s to 8, a peptide with 31 residues would be corresponded to 24 $(31 - 8 + 1)$ sliding windows and converted to a matrix of 24×20 dimensions.

2.3. Construction of Classifiers

2.3.1. Artificial Neural Network. ANN is a traditional machine learning algorithm that was widely utilized in lysine PTM prediction applications. In this article, we construct an ANN model with four layers, i.e., input layer, output layer, and two hidden layers. The input layer received the feature sequence generated from different encoding method. The two hidden layers owe both 100 neurons and adopt "reLu" as their activation function. The output layer owes a single unit, outputting the probability score of each site.

2.3.2. Support Vector Machine. SVM is a well-established and commonly employed algorithm based on structural risk minimization from statistical learning theory [20]. SVM can transform the samples into a high-dimensional feature space and then construct an Optimal Separating Hyperplane (OSH) to maximize its distance from the closest training samples. Here, based on Tensorflow [26] and Scikit-learn [27], we employed SCV as our SVM model, where the applied kernel function was linear kernel.

2.3.3. K-Nearest Neighbor Algorithm. kNN algorithm is another widely employed algorithm that calculates the distances of samples to cluster them [28]. If we obtain the training dataset $D = \{v_1, v_2, ..., v_n\}$and a testing sample x, we can utilize KNN to calculate the distances between x and all the instances in D. Therefore, as the nearest neighbor (shortest distance) in the training dataset, the query sample will be assigned to the same class. In this work, we also construct a kNN model implemented by Tensorflow and Scikit-learn. The parameters of our kNN were set to their default values.

3. Crossvalidation Methods

In general, when the classification model is built, researchers will divide the dataset into two parts as training set and testing set. Process of dataset usage partition is depicted in Figure 2. To make full use of the training set samples, we usually train the model through 10-fold crossvalidation. The samples in training set are divided into training set and validating set by the crossvalidation method. And then, the training set is used to train the model, while the validating set is used to verify the effect of the model and obtain the validation scores. After the crossvalidation is completed, the trained model will pass the testing set to evaluate its performance and get the testing scores.

In view of the fact that the classifier is always more sensitive to the category containing more samples and less sensitive to the category containing fewer samples in binary classification problems, it is necessary to preprocess the training set before inputting the data with unbalanced positive and negative samples into the classifier. In the previous work, we proposed a new feature extraction method named SEV (Split to Equal Validation), which can well solve the problem of imbalanced training samples in PTM sites prediction research studies. In this experiment, we also adopt the SEV method and at the same time used 10-fold crossvalidation to do comparative experiments. The working flow of SEV is as follows.

(Note: the pos means the positive and the neg means the negative in Figure 3.)

The experiment shown in Figure 3 is a part of the whole experiment, which corresponds to the process of using training set to obtain classifier and its validation scores after the dataset is divided into training set and testing set in Figure 3. In details, SE validation consists of five steps. Assuming that the ratio of negative samples to positive samples in the training set is close to n:1, (1) the first step is to divide the negative samples into n groups; (2) in the second step, each positive sample is combined with the positive sample to generate n balanced subsets; (3) subsequently, model 1 will be trained by subset 1 and verified by subset 2; model 2 will be trained by subset 2 and verified by subset 3, and so on; (4) according to the n balanced subsets, n models were trained and validated; (5) finally, each model will be tested by independent testing sets, and the average of their scores will be utilized to evaluate their performance.

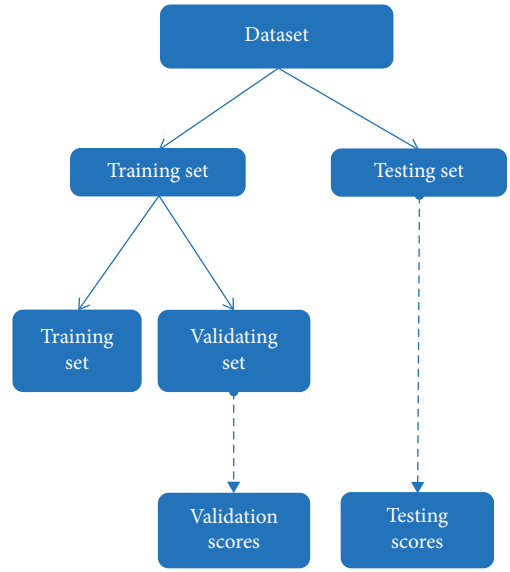


Figure 2: Process of dataset usage partition.

3.1. Performance Assessment of Predictors. There are a set of four metrics [29] directly that are often utilizing to quantitatively evaluate the performance of predictors: Sn (sensitivity), also known as TPR (True Positive Rate), reflects the proportion of true positive samples (TP) determined by the model to all the positive samples in the dataset; Sp (specificity), also known as TNR, reflects the proportion of true negative samples judged by the model in all negative samples; Acc (accuracy) is the proportion of correct samples determined by the model to the total samples; and MCC (Mathew's Correlation Coefficient) reflects the correlation coefficient between the actual predicted samples and the expected predicted samples:

$$\begin{aligned}
\mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\
\mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \\
\mathrm{Acc} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \\
\mathrm{MCC} &= \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}}, \\
F1 &= \frac{2 * \mathrm{pre} * \mathrm{rec}}{\mathrm{pre} + \mathrm{rec}},
\end{aligned} \tag{2}$$

where TP, FP, TN, and FN represent the true positives, false positives, false negatives, and true negatives, respectively.

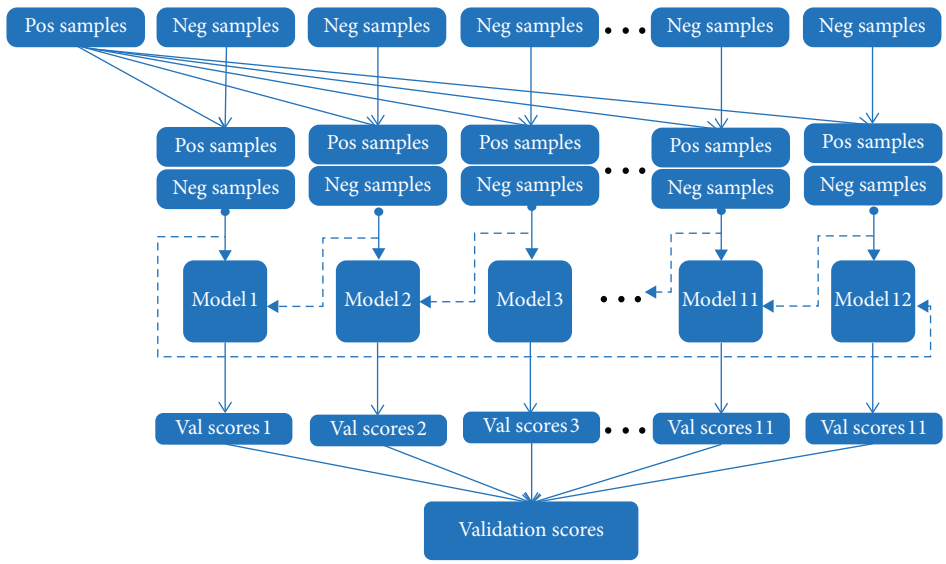


Figure 3: The working flow of Split to Equal validation.

The pre means precision and the rec means recall in the classification model. On the contrary, ROC curves and AUC value were also adopted to evaluate the performance of the predictors.

4. Results and Discussion

4.1. Performance of the Three Classification Models Based on Different Encoding Schemes. In this study, we firstly constructed three machine learning models, i.e., ANN, SVM, and kNN algorithm, and then trained them according to Amino Acid Composition (AAC) encoding scheme that considered the hydrophobicity and charged character of the amino acid. Split to Equal Validation (SEV) and independent training sets were utilized to assess the performance of models above, where AUC, Acc, MCC, Sn, and Sp were adopted as assessment metrics. The results of the independent testing results were depicted in Table 2.

Based on the results of this experiment, we speculate that feature extraction schemes are very important factors affecting the final classification accuracy. Therefore, we adopt the EBAG + Profile encoding method, which utilized the physical and chemical properties of amino acids. EAAC encoding method, which is based on AAC encoding and accords to the probability of occurrence of specific amino acids in the peptide sequence, was also adopted in this experiment. The testing scores were depicted in Tables 3 and 4, respectively.

As we know, a larger AUC value means that the current classification algorithm is more likely to rank positive samples in front of negative samples, so as to get better classification results. Therefore, it is obvious that classifiers under EAAC encoding scheme plays better performance than the other two schemes for getting higher AUC values. And other experimental results such as MCC and Acc value and EAAC encoding also obtained higher scores and showed similar advantages than others.

On the contrary, there is one thing certain that ANN's classification effect is better than SVM and kNN under EAAC encoding scheme. As for independent test, when taking EAAC, the AUC value of ANN is **0.7471**, while SVM and kNN algorithms obtain the AUC value of 0.6322 and 0.6317. All of these results in Tables 2–4 show that different types of classifiers have great impact on the prediction performance. In this work, ANN is the best classifier.

4.2. Comparing the Results of SEV with 10-Fold Cross-validation and Scaling the Number of Training Samples by SEV. In this research, we utilize the SEV verification method to preprocess the training samples, and thus train the classifier model. In addition, on the premise that other conditions remain unchanged, we used 10-fold cross-validation instead of SEV to do the same experiment. The experimental results are shown in Table 5:

Among them, the experiment is based on the neural network model, using 10-fold crossvalidation and SEV methods, respectively. It can be seen that although the Acc equivalent of the 10-fold cross is too high, its AUC value does not perform well. This is because, in the case of extremely unbalanced positive and negative samples, the classifier will guess the kind of samples with higher probability in the training set, but its classification ability is not outstanding. The SEV verification method can overcome this problem well. Although Acc equivalence is not as good as 10-fold cross verification, SEV has more advantages in the AUC value, which can best represent the classification ability of the model in the real sense.

More importantly, in order to further explore the influence of imbalance between positive and negative samples in the training set, we used the SEV method to scale the training samples. We verified the fact that unbalanced training data would finally lead to very low Sn and very high Sp of the classifier, further causing these evaluation metrics to lose its significance. In order to further explore how far

Table 2: The testing results of three models based on AAC encoding.

AAC	Classifier	AUC	Acc (%)	MCC	Sn (%)	Sp (%)	*F*1
Testing scores	ANN	0.5833	54.94	0.0598	56.35	54.83	55.58
	SVM	0.6149	53.92	0.0856	62.86	53.17	57.61
	kNN	**0.6224**	47.64	0.0906	71.20	45.67	55.65

Table 3: The testing results of three models based on EBPR encoding.

EAPR	Classifier	AUC	Acc (%)	MCC	Sn (%)	Sp (%)	*F*1
Testing scores	ANN	**0.6552**	56.57	0.1226	67.23	55.68	60.91
	SVM	0.5041	82.71	0.0056	12.06	88.61	21.23
	kNN	0.5874	64.69	0.0705	46.38	66.21	54.55

Table 4: The testing results of three models based on EAAC encoding.

EAAC	Classifier	AUC	Acc (%)	MCC	Sn (%)	Sp (%)	*F*1
Testing scores	ANN	**0.7471**	63.54	0.2002	74.16	62.65	67.92
	SVM	0.6322	56.21	0.1028	63.61	55.59	59.33
	kNN	0.6317	43.60	0.0931	76.19	40.88	53.21

Table 5: The testing scores of 10-fold crossvalidation and Split to Equal Validation.

Scores	Validation methods	AUC	ACC (%)	MCC	SN (%)	SP (%)	F1
Validation scores	10F CV	0.5751	21.11	0.0776	95.23	15.14	26.13
	SEV	**0.8465**	64.35	0.3260	100.00	61.48	76.15
Testing scores	10F CV	0.6965	90.12	0.1060	10.91	96.73	19.61
	SEV	**0.7471**	63.54	0.2002	74.16	62.65	67.92

reaching the positive and negative sample ratios affect the classifier's performance, we calculate the five metrics, i.e., AUC, Acc, MCC, Sn, and Sp by adjusting different positive and negative sample ratios from 1 : 1 to 1 : 12 under SEV.

As can be seen from Table 6, as the proportion of positive and negative samples in the training set increases, the AUC value of the model gradually decreases from 0.7471 to about 0.7000, the MCC value gradually decreases to about 0.1300, while the ACC value continuously increases to 89.43%. This also verifies our previous conclusion: when the positive and negative samples of the training set are extremely unbalanced, the classifier will tend to guess the kind with more samples, but its classification effect is not good.

In addition, Table 6 also shows another information, that is, when the ratio of positive and negative samples in the training set reaches 1 : 9, the AUC value of the classifier will also tend to be stable, only fluctuating around 0.7000 without further decline. This means that, in this experiment, although the ratio of positive and negative samples has been changing towards a more unbalanced direction, the performance of the classifier will not decrease indefinitely, but will tend to be stable after reaching a certain threshold.

In a word, through this experiment, we can further verify that the positive and negative sample ratios have far-reaching influence on the results in the binary classification problem, and the SEV method can solve this problem well.

Table 6: The testing scores based on different proportion of positive and negative samples.

Scale-up ratio	AUC	Acc (%)	MCC	Sn (%)	Sp (%)	*F*1
1:1	0.7471	63.54	0.2002	74.16	62.65	67.92
1:2	0.7399	76.67	0.1982	52.99	78.65	63.32
1:3	0.7324	82.60	0.1828	38.61	86.28	53.35
1:4	0.7290	84.33	0.1689	32.65	88.64	47.72
1:5	0.7228	85.54	0.1623	28.93	90.26	43.82
1:6	0.7205	86.71	0.1476	24.13	91.94	38.23
1:7	0.7196	87.27	0.1499	23.12	92.62	37.00
1:8	0.7193	86.64	0.1539	25.18	91.77	39.52
1:9	0.7012	87.32	0.1340	20.91	92.86	34.13
1:10	0.6984	88.08	0.1382	19.61	93.80	32.44
1:11	0.6972	88.96	0.1337	16.87	94.97	28.65
1:12	0.7033	89.43	0.1292	15.10	95.64	26.08

5. Conclusions

The currently available PTM prediction approaches are mainly based on ML that requires preprocessing amino acid data into digital features. Here, we adopted two feature extraction schemes according to different ideas of physical and chemical characteristics and occurrence frequency and then constructed three ML classifiers, while the application of the SEV method solves the problem of sample imbalance in binary classification. The results not only showed that

feature extraction methods and classifier types play important roles on prediction results but also indicated the direction of our next work. In addition to proposing new feature encoding schemes, more classifiers can be utilized in this field, including depth learning (DL) classifiers such as CNN or RNN. In addition, the SEV method will be improved and applied to the new machine learning model. In total, the outstanding performance of ML in prediction of Kmal sites suggests that computational methods can be applied widely to this field.

Data Availability

The data utilized to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This work was supported by the talent project of "Qingtan scholar" of Zaozhuang University, Shandong Provincial Natural Science Foundation, China (no. ZR2015PF007), the PhD research startup foundation of Zaozhuang University, and Zaozhuang University Foundation (nos. 2014BS13 and 2015YY02).

References

[1] C. Peng, Z. Lu, Z. Xie et al., "The first identification of lysine malonylation substrates and its regulatory enzyme," *Molecular & Cellular Proteomics*, vol. 10, no. 12, p. M111.012658, 2011.

[2] H.D. Matthew and D. Yingming, "Metabolic regulation by lysine malonylation, succinylation, and glutarylation," *Molecular & Cellular Proteomics Mcp*, vol. 14, 2015.

[3] H. Mujahid, X. Meng, S. Xing, X. Peng, C. Wang, and Z. Peng, "Malonylome analysis in developing rice (Oryza sativa) seeds suggesting that protein lysine malonylation is well-conserved and overlaps with acetylation and succinylation substantially," *Journal of Proteomics*, vol. 170, pp. 88–98, 2018.

[4] T. Arendt, H. G. Zveuintshva, and T. A. Lkontovich, "Dendritic changes in the basal nucleus of meynert and in the diagonal band nucleus in alzheimer's disease-a quantitative golgi investigation," *Neuroscience*, vol. 19, no. 4, pp. 1265–1278, 1986.

[5] X. Bao, Q. Zhao, T. Yang, Y. M. E. Fung, and X. D. Li, "A chemical probe for lysine malonylation," *Angewandte Chemie*, vol. 125, no. 18, pp. 4983–4986, 2013.

[6] P. Boevink, K. Oparka, C. S. Santa, B. Martin, A. Betteridge, and C. Hawes, "Stacks on tracks: the plant Golgi apparatus traffics on an actin/ER network," *Plant Journal for Cell & Molecular Biology*, vol. 15, pp. 441–447, 2010.

[7] M. Bretscher and S. Munro, "Cholesterol and the golgi apparatus," *Science*, vol. 261, no. 5126, pp. 1280-1281, 1993.

[8] M. Canuel, S. Lefrancois, J. Zeng, and C. R. Morales, "AP-1 and retromer play opposite roles in the trafficking of sortilin between the golgi apparatus and the lysosomes," *Biochemical and Biophysical Research Communications*, vol. 366, no. 3, pp. 724–730, 2008.

[9] K. Caroline, M. Katy, A. Shireen et al., "Foot-and-mouth disease virus replication sites form next to the nucleus and close to the golgi apparatus, but exclude marker proteins associated with host membrane compartments," *Journal of General Virology*, vol. 86, 2005.

[10] L. Citores, L. Bai, V. Sørensen, and S. Olsnes, "Fibroblast growth factor receptor-induced phosphorylation of STAT1 at the golgi apparatus without translocation to the nucleus," *Journal of Cellular Physiology*, vol. 212, no. 1, pp. 148–156, 2007.

[11] G. Werner and K. Werner, "Changes in the nucleus, endoplasmic reticulum, golgi apparatus, and acrosome during spermiogenesis in the waterstrider, Gerris najas deg. (Heteroptera: gerridae)," *International Journal of Insect Morphology and Embryology*, vol. 22, no. 5, pp. 521–534, 1993.

[12] W. G. Whaley and M. Dauwalder, "The golgi apparatus, the plasma membrane, and functional integration," *International Review of Cytology*, vol. 58, pp. 199–245, 1979.

[13] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques," *Acm Sigmod Record*, vol. 31, pp. 76-77, 2011.

[14] J.-Y. Xu, Z. Xu, Y. Zhou, and B.-C. Ye, "Lysine malonylome may affect the central metabolism and erythromycin biosynthesis pathway in saccharopolyspora erythraea," *Journal of Proteome Research*, vol. 15, no. 5, pp. 1685–1701, 2016.

[15] R. Yang, C. Zhang, R. Gao, and L. Zhang, "A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data," *International Journal of Molecular Sciences*, vol. 17, no. 2, p. 218, 2016.

[16] S. Ya and P.-F. Jiao, "Predicting golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties," *Journal of Theoretical Biology*, vol. 391, 2016.

[17] M. Bujnicki, Dunin-Horkawicz, de Stanislaw et al., "tRNA-modpred: a computational method for predicting posttranscriptional modifications in tRNAs," *Methods A Companion to Methods in Enzymology*, vol. 107, 2016.

[18] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Transactions on Nanobioscience*, vol. 16, no. 4, pp. 240–247, 2017.

[19] S. Banerjee, S. Basu, D. Ghosh, and M. Nasipuri, "PhospredRF: prediction of protein phosphorylation sites using a consensus of random forest classifiers," in *Proceedings of the International Conference & Workshop on Computing & Communication*, Kassel, Germany, March 2015.

[20] D. Wang, S. Zeng, C. Xu et al., "MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction," *Bioinformatics*, vol. 33, 2017.

[21] C. Zhen, Z. Yuan, Z. Zhang, and J. Song, "Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features," *Briefings in Bioinformatics*, vol. 4, 2014.

[22] R. Z. Han, D. Wang, Y. H. Chen, L. K. Dong, and Y. L. Fan, "Prediction of phosphorylation sites based on the integration of multiple classifiers," *Genetics & Molecular Research*, vol. 16, 2017.

[23] S. C. Bagley and R. B. Altman, "Characterizing the microenvironment surrounding protein sites," *Protein Science*, vol. 4, no. 4, pp. 622–635, 2008.

[24] L.-N. Wang, S.-P. Shi, H.-D. Xu, P.-P. Wen, and J.-D. Qiu, "Computational prediction of species-specific malonylation sites via enhanced characteristic strategy," *Bioinformatics*, vol. 33, p. btw755, 2016.

[25] C. Zhen, H. Ningning, H. Yu et al., "Integration of a deep learning classifier with A random forest approach for predicting malonylation sites," *Genomics Proteomics & Bioinformatics*, vol. 16, 2018.
[26] L. Rampasek and A. Goldenberg, "TensorFlow: biology's gateway to deep learning?" *Cell Systems*, vol. 2, no. 1, pp. 12–14, 2016.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[28] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, 2012.
[29] J. Chen, H. Liu, J. Yang, and K.-C. Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," *Amino Acids*, vol. 33, no. 3, pp. 423–428, 2007.

Hindawi
Scientific Programming
Volume 2020, Article ID 7845384, 8 pages
https://doi.org/10.1155/2020/7845384

Research Article

CenterFace: Joint Face Detection and Alignment Using Face as Point

Yuanyuan Xu,[1,2] **Wan Yan**,[3] **Genke Yang**,[2] **Jiliang Luo**,[1] **Tao Li**,[4] **and Jianan He**[4]

[1]*College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China*
[2]*Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China*
[3]*Xiamen Star Clouds Network Technology Co., Ltd., Xiamen 361005, China*
[4]*Central Laboratory of Health Quarantine, Shenzhen International Travel Health Care Center and Shenzhen Academy of Inspection and Quarantine, Shenzhen Customs District, Shenzhen 518033, China*

Correspondence should be addressed to Yuanyuan Xu; yyxu@hqu.edu.cn

Received 5 February 2020; Revised 3 June 2020; Accepted 17 June 2020; Published 2 July 2020

Academic Editor: Chenxi Huang

Face detection and alignment in unconstrained environment is always deployed on edge devices which have limited memory storage and low computing power. This paper proposes a one-stage method named CenterFace to simultaneously predict facial box and landmark location with real-time speed and high accuracy. The proposed method also belongs to the anchor-free category. This is achieved by (a) learning face existing possibility by the semantic maps, (b) learning bounding box, offsets, and five landmarks for each position that potentially contains a face. Specifically, the method can run in real time on a single CPU core and 200 FPS using NVIDIA 2080TI for VGA-resolution images and can simultaneously achieve superior accuracy (WIDER FACE Val/Test-Easy: 0.935/0.932, Medium: 0.924/0.921, Hard: 0.875/0.873, and FDDB discontinuous: 0.980 and continuous: 0.732).

1. Introduction

Face detection and alignment is one of the fundamental issues in computer vision and pattern recognition and is often deployed in mobile and embedded devices. These devices typically have limited memory storage and low computing power. Therefore, it is necessary to predict the position of the face box and the landmark at the same time, and it is excellent in speed and precision.

With the great breakthrough of convolutional neural networks (CNN), face detection has achieved remarkable progress in recent years. Previous face detection methods have inherited the paradigm of anchor-based generic object detection frameworks, which can be divided into two categories: two-stage method (Faster-RCNN [1]) and one-stage method (SSD [2]). Compared with the two-stage method, the one-stage method is more efficient and has higher recall rate, but it tends to achieve a higher false positive rate and to compromise the localization accuracy. Then, Hu and Ramanan [3] used a two-stage approach to the Region Proposal Networks (RPN) [1] to detect faces directly, while SSH [4] and S3FD [5] developed a scale-invariant network in a single network to detect faces with mutiscale from different layers.

The previous anchor-based methods have some drawbacks. On the one hand, in order to improve the overlap between anchor boxes and ground truth, a face detector usually requires a large number of dense anchors to achieve a good recall rate. For example, more than 100 k anchor boxes is designed in RetinaFace [6] for a 640×640 input image. On the other hand, the anchor is a hyperparameter design that is statistically calculated from a particular dataset, so it is not always feasible to other applications, which goes against the generality.

In addition, the current state-of-the-art face detectors has achieved considerable accuracy on the benchmark WIDER FACE [7] by using heavy pretrained backbones such as VGG16 [8] and resnet50/152 [9]. First, these detectors are difficult to use in practice because the network consumes too

much time and the model size is also too large. Secondly, it is not convenient for face recognition application without facial landmark prediction. Therefore, joint detection and alignment, as well as better balance of accuracy and latency, are essential for practical applications.

Inspired by the anchor-free universal object detection framework [1, 10–15], this paper proposes a simpler and more effective face detection and alignment method named CenterFace, which is not only lightweight but also powerful. The network structure about the CenterFace is shown in Figure 1, which can be trained end-to-end. We use the center point of the face's bounding box to represent the face, then facial box size and landmark are regressed directly to image features at the center location. So, face detection and alignment are transformed to the standard key point estimation problem [16–18]. The peak in the heat map corresponds to the center of the face. The image features at each peak predict the size of the face and the face key points. This approach was fully evaluated and the latest detection performance were shown on a number of benchmark datasets for face detection, including FDDB [19] and WIDER FACE.

In summary, the main contributions of this work can be summarized as four-fold:

(i) By introducing the anchor-free design, face detection is transformed into a standard key point estimation problem, using only a larger output resolution (output stride is 4) compared to previous detectors

(ii) Based on the multitask learning strategy, the face as point design is proposed to predict the faceBoxes and five key points at the same time

(iii) This paper proposes a feature pyramid network using common layer for accurate and fast face detection

(iv) Comprehensive experimental results based on popular benchmarks FDDB and WIDER FACE, as well as CPU and GPU hardware platforms, have demonstrated the superiority of the proposed method in terms of speed and accuracy

2. Related Works

2.1. Cascaded CNN Methods. The method of cascade convolutional neural network (CNN) [20–22] uses cascaded CNN framework to learn features in order to improve the performance and maintain efficiency. However, there are some problems about cascaded CNN-based detector. (1) The runtime of these detector is negatively correlated with the number of faces on the input image. The speed will dramatically degrade when the number of faces increases. (2) Because these methods optimize each module separately, the training process becomes extremely complicated.

2.2. Anchor Methods. Inspired by generic object detection methods [2, 14, 15, 23–27], which embraced all the recent advancement in deep learning, face detection has recently achieved remarkable progress [3–5, 28]. Different from generic object detection, the ratio of the face scale is usually from 1 : 1 to 1 : 1.5. The latest methods [6, 28] focus on single-stage design, which densely samples' face locations and scales on feature pyramids, demonstrating promising performance and yielding faster speed compared to two-stage methods [29, 30]. However, dense samples result in long time consuming.

2.3. Anchor-Free Methods. In our view, Cascaded CNN methods are also a kind of anchor-free methods. However, these method uses sliding window to detect human faces and relies on image pyramids. It has some shortcomings such as slow speed and complex training process. LFFD [31] regards the RFs as natural anchors which can cover continuous face scales, which is just another way to define anchor, but the training time is about 5 days with two NVIDIA GTX1080TI. Our CenterFace simply represents faces by a single point at their bounding box center; then, facial box size and landmark are regressed directly from image features at the center location. Thus, face detection is transformed into a standard key point estimation problem. And the training time of a NVIDIA GTX2080TI is only one day.

2.4. Multitask Learning. Multitask learning uses multiple supervisory labels to improve the accuracy of each task by utilizing the correlation between tasks. Joint face detection and alignment [17, 20] is widely used because alignment task, paralleling with the backbone, provides better features for face classification task with face point information. Similarly, Mask R-CNN [32] significantly improves the detection performance by adding a branch for predicting an object mask.

3. CenterFace

3.1. Mobile Feature Pyramid Network. We adopted Mobilenetv2 [33] as the backbone and Feature Pyramid Network (FPN) [14] as the neck for the subsequent detection. In general, FPN uses a top-down architecture with lateral connections to build a feature pyramid from a single scale input. CenterFace represents the face through the center point of the face box, and face size and facial landmark are then regressed directly from image features of the center location. Therefore, only one layer in the pyramid is used for face detection and alignment. We construct a pyramid with levels {P-L}, $L = 3, 4, 5$, where L indicates pyramid level. Pl has $1/2^{L}$ resolution of the input. All pyramid levels have $C = 24$ channels, and we define classification loss, box regression loss, and landmark regression loss only on P2.

3.2. Face as Point. Let $[x_1, y_1, x_2, y_2]$ be the bounding box of face. Facial center point lies at $c = [(x_1 + x_2)/2$ and $(y_1 + y_2)/2]$. Let $I \in R^{W \times H \times 3}$ be an input image of width W and height H. Our aim is to produce the heat map $Y \in [0, 1]^{W/R \times H/R}$, where R is the output stride, and we use the default output stride of $R = 4$. During training, the prediction $\widehat{Y}_{x,\,y} = 1$ corresponds to a face center, while $\widehat{Y}_{x,\,y} = 0$ is background. For each ground

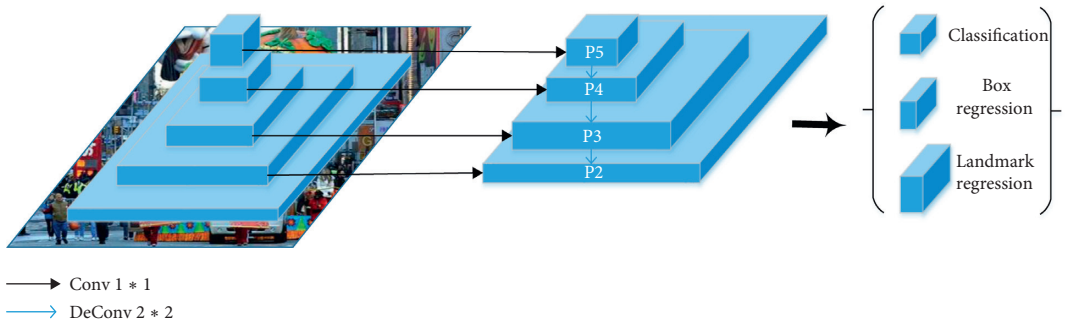


Figure 1: Architecture of the CenterFace.

truth $Y_{x,y}$, we calculate the equivalent heat map by using y an unnormalized 2D Gaussian to represent the ground truth. The training loss is a variant of focal loss [15]:

$$L_c = \begin{cases} -\left(1-\hat{Y}_{xy}\right)^{\alpha}\log\left(\hat{Y}_{xy}\right) & \text{if } Y_{xy}=1, \\ -\left(1-Y_{xy}\right)^{\beta}\left(\hat{Y}_{xy}\right)^{\alpha}\log\left(1-\hat{Y}_{xy}\right) & \text{otherwise,} \end{cases} \tag{1}$$

where α and β are hyperparameters of the focal loss, which are designated as $\alpha=2$ and $\beta=4$ in all our experiments following Law and Deng [34].

To gather global information and to reduce memory usage, downsampling is applied to an image convolutionally, and the size of the output is usually smaller than the image. Hence, a location (x, y) in the image is mapped to the location $(x/n, y/n)$ in the heatmaps, where n is the downsampling factor. When we remap the locations from the heatmaps to the input image, some pixel may be not alignment, which can greatly affect the accuracy of facial boxes. To address this issue, we predict position offsets to adjust the center position slightly before remapping the center position to the input resolution:

$$o_k = \left(\frac{x_k}{n} - \left\lfloor\frac{x_k}{n}\right\rfloor, \frac{y_k}{n} - \left\lfloor\frac{y_k}{n}\right\rfloor\right), \tag{2}$$

where o_k is the offset and x_k and y_k are the x and y coordinate for face center k. We apply the L1 Loss at ground-truth center position.

3.3. Box and Landmark Prediction. To reduce the computational burden, we use a single size prediction $S \in R^{W/4\times H/4}$ for facial box and landmarks. Each ground-truth bounding box is specified as $G=(x1, y1, x2, y2)$. During training, our goal is to learn a transformation that maps the networks position outputs $(\hat{h}, \hat{w})$ to center position in the feature maps:

$$\begin{aligned} \hat{h} &= \log\left(\frac{x_2}{R} - \frac{x_1}{R}\right), \\ \hat{w} &= \log\left(\frac{y_2}{R} - \frac{y_1}{R}\right), \end{aligned} \tag{3}$$

where R is the stride of networks, which are designated as $R=4$.

Different from box regression, the regression of the five facial landmarks adopts the target normalization method based on the center position:

$$\begin{aligned} \widehat{lm_x} &= \frac{lm_x}{\text{box}_w} - \frac{c_x}{\text{box}_w}, \\ \widehat{lm_y} &= \frac{lm_y}{\text{box}_h} - \frac{c_y}{\text{box}_h}, \end{aligned} \tag{4}$$

where lm_x and lm_y are the x and y coordinates for face landmark, c_k and c_k are the x and y coordinates for face center, and box_w and box_h are width and height of the face. We also use smooth L1 loss to facial box and landmark prediction at the center location.

For any training face center, we minimise the following multitask loss:

$$L = L_c + \lambda_{\text{off}}L_{\text{off}} + \lambda_{\text{box}}L_{\text{box}} + \lambda_{lm}L_{lm}, \tag{5}$$

where λ_{off}, λ_{box}, and λ_{lm} is used to scale the loss, and we use 1, 0.1, and 0.1, respectively, in all our experiments.

3.4. Training Details

3.4.1. Dataset. The proposed method is trained on the training set of WIDER FACE benchmark, including 12,880 images with more than 150,000 valid faces in scale, pose, expression, occlusion, and illumination. RetinaFace [6] introduces five levels of face image quality and annotates five landmarks on faces.

3.4.2. Data Augmentation. Data augmentation is important to improve the generalization. We use random flip, random scaling [35], color jittering, and randomly crop square patches from the original images and resize these patches into 800×800 to generate larger training faces. Faces that are less than 8 pixels are discarded directly.

3.4.3. Training Parameters. We train the CenterFace using Adam optimiser with a batch-size 8 and learning rate $5e-4$ for 140 epochs, with the learning rate dropped $10x$ at 90 and 120 epochs, respectively. The downsampling layers of MobilenetV2 are initialized with ImageNet pretrain and the

up-sampling layers are randomly initialized. The training time is about one day with one NVIDIA GTX2080TI.

4. Experiments

In this section, we firstly introduce the runtime efficiency of CenterFace and then evaluate it on the common face detection benchmarks.

4.1. Running Efficiency. The existing CNN face detectors can be accelerated by GPUs, but they are not fast enough in most practical applications, especially CPU-based applications. As described below, our CenterFace is efficient enough to meet practical requirements and its model size is only 7.2 MB. In Table 1, comparing with other detectors, our method can exceed the real-time running speed (>100 FPS) at different resolutions by using a single NVIDIA GTX2080TI. Owing to the DSFD, PyramidBox, S3FD, and SSH are too slow when running on CPU platforms, and we only evaluate the proposed CenterFace, FaceBoxes, MTCNN, and CasCNN at VGA-resolution images on CPU and the mAP means the true positive rate at 1000 false positives on FDDB. As listed in Table 2, our CenterFace can run at 30 FPS on the CPU with state-of-the-art accuracy.

Table 1: Running efficiency on GTX2080TI.

Approach	640×480 (ms)	1280×720 (ms)	1920×1080 (ms)
DSFD [29]	78.08	187.78	393.82
PyramidBox [28]	50.51	142.34	331.93
S3FD [5]	21.75	55.73	119.53
LFFD [31]	7.60	16.37	31.41
CenterFace	5.51	6.47	8.79

Table 2: Running efficiency on CPU.

Approach	CPU-model	mAP (%)	FPS
CasCNN [21]	E5-2620@2.00	85.7	14
MTCNN [20]	N/A@2.60	94.4	16
FaceBoxes [36]	E5-2660v3@2.60	96.0	20
CenterFace	I7-6700@2.6	98.0	30

Table 3: Evaluation results on FDDB.

Method	Disc ROC curves score	Cont ROC curves score
DFSD [29]	0.984	0.754
PyramidBox [28]	0.982	0.757
S3FD [5]	0.981	0.754
MTCNN [20]	0.944	0.708
Faceboxes3.2 [36]	0.960	0.729
LFFD [31]	0.973	0.724
CenterFace	0.980	0.732

4.2. Evaluation on Benchmarks

4.2.1. FDDB Dataset. FDDB contains 2845 images with 5171 unconstrained faces collected from the Yahoo news website. We evaluate our face detector on FDDB against the other state-of-the-art methods, and the results are shown in Table 3 and Figure 2, respectively. We also add DFSD, PyramidBox, and S3FD detectors, whereas these detectors are much slower due to the larger backbone and denser anchors. Our CenterFace can also achieve good performance on both discontinuous and continuous ROC curves, i.e., 98.0% and 72.9% when the number of false positives equals to 1,000 and it outperforms LFFD, FaceBoxes, and MTCNN evidently.

4.2.2. WIDER FACE Dataset. Until now, WIDER FACE is the most widely used benchmark for face detection. The WIDER FACE dataset is split into training (40%), validation (10%), and testing (50%) subsets by randomly sampling from 61 scene categories. All the compared methods are trained on the training set. For testing on WIDER FACE, we follow the standard practices of [6] and employ flip as well as multiscale strategies. Box voting [36] is applied on the union set of predicted faceBoxes using an IoU threshold at 0.4. We report the results on the testing sets in Table 4, respectively. The proposed method CenterFace achieves 0.932 (Easy), 0.921 (Medium), and 0.873 (Hard) for testing set. Although it has gaps with state-of-the-art methods, but consistently outperforms SSH (using VGG16 as the backbone), LFFD, FaceBoxes, and MTCNN. Additionally, CenterFace is better than S3FD that uses VGG16 as the backbone and dense anchors on hard parts.

Furthermore, we also test on WIDER FACE not only with the original image but also with a single inference, and our CenterFace also produces the good average precision (AP) in all the subsets of both validation sets, i.e., 92.2% (Easy), 91.1% (Medium), and 78.2% (Hard) for the validation set. Figure 3 shows some qualitative results on the WIDER FACE dataset.

4.2.3. AFLW Dataset. To evaluate the accuracy of face alignment, we compare CenterFace with MTCNN on the AFLW dataset. The mean error is measured by the distances between the estimated landmarks and the ground truths and normalized with respect to the interocular distance. As shown in Figure 4, we give the mean error of each facial landmark on the AFLW dataset [37]. CenterFace significantly decreases the normalized mean errors (NME) from 6.2% to 6.9% when compared to MTCNN.

4.3. Parameter, FLOPs, and Model Size. In this section, the comparison method is studied from the perspective of parameters, computation, and model size. Edge devices always have limited storage. We use FLOPs to measure the computation at resolution 640×480. The number of parameters is closely related to the size of the model. However, the model size may vary slightly with different libraries, and less parameters do not mean less computation. All the information is presented in Table 5.

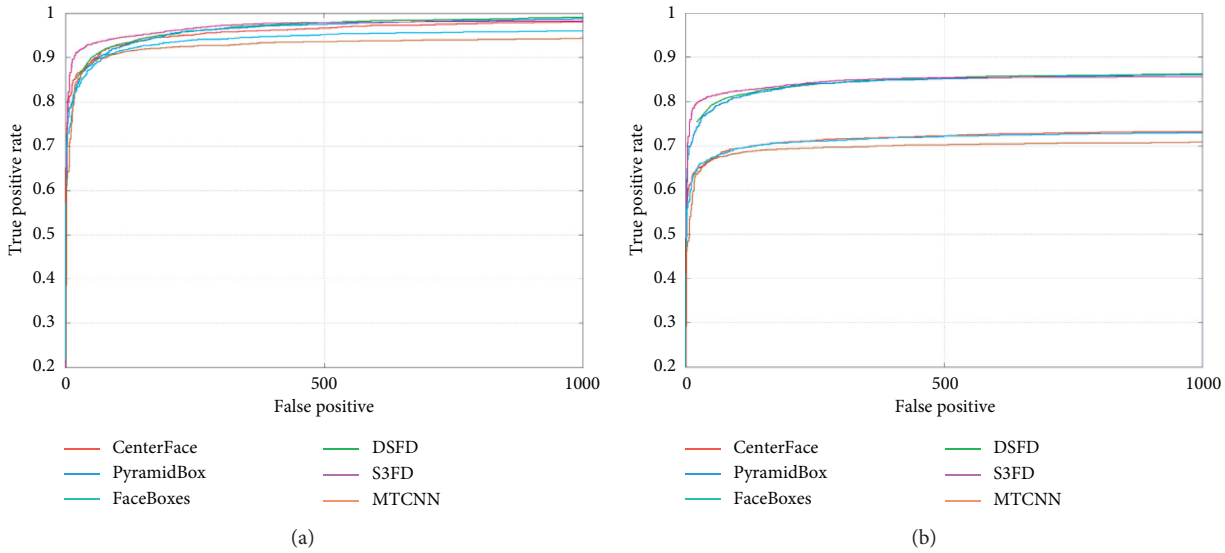


Figure 2: Evaluation on the FDDB dataset. (a) Discontinuous ROC curves. (b) Continuous ROC curves.

Table 4: Performance results on the testing set of WIDER FACE.

Method	Easy	Medium	Hard
RetinaFace [6]	0.963	0.956	0.914
DSFD [29]	0.960	0.953	0.900
PramidBox [28]	0.956	0.946	0.887
S3FD [5]	0.928	0.913	0.840
SSH [4]	0.927	0.915	0.844
MTCNN [20]	0.851	0.820	0.607
FaceBoxes [36]	0.839	0.763	0.396
LFFD [31]	0.896	0.865	0.770
CenterFace	0.932	0.921	0.873

(a)

Figure 3: Continued.

(b)

Figure 3: Face detection results on WIDER face.

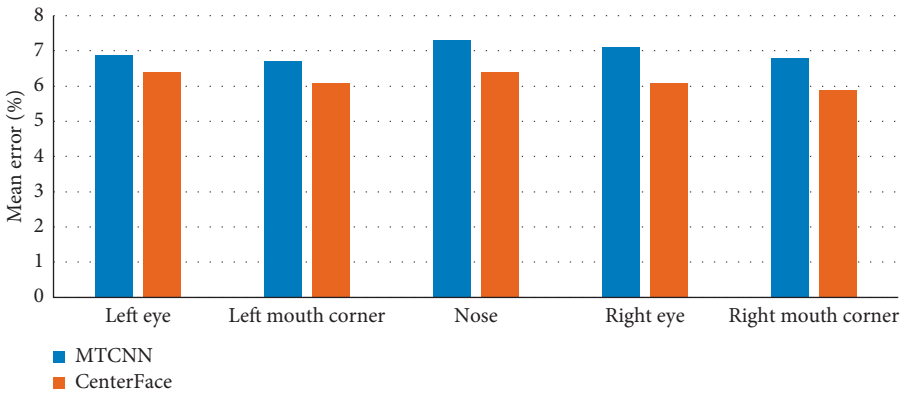


Figure 4: Evaluation on AFLW for face alignment.

Table 5: Number of parameters, FLOPs, and model sizes.

Method	Parameters (M)	FLOPs (G)	Model sizes (MB)
DSFD [29]	141.38	140.19	458
PyramidBox [28]	57.18	236.58	218
S3FD [5]	22.46	96.6	86
SSH [4]	19.75	99.8	79
LFFD [31]	2.15	9.25	9
CenterFace	1.83	2.06	7.2

For the most advanced methods DSFD and PyramidBox, they have a large number of parameters, FLOPs, and model sizes. Evidently, the proposed method has much more efficient computation and light network, which demonstrates the superiority of the concise network design.

5. Conclusion

This paper introduces the CenterFace that has the superiority of the proposed method, performs well on both speed and accuracy, and simultaneously predicts facial box and landmark location. Our proposed method overcomes the drawbacks of the previous anchor-based method by translating face detection and alignment into a standard key point estimation problem. CenterFace represents the face through the center point of the face box, and face size and facial landmark are then regressed directly from image features of the center location. Comprehensive and extensive

experiments are made to fully analyze the proposed method. The final results demonstrate that our method can achieve real-time speed and high accuracy with a smaller model size, making it an ideal alternative for most face detection and alignment applications.

Data Availability

The data used to support the findings of this study have been deposited in the http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/WiderFace_Results.html repository.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2018YFC0809200) and Natural Science Foundation of Shanghai (16ZR1416500).

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] W. Liu, D. Anguelov, D. Erhan et al., "SSD: Single shot multibox detector," in *Computer Vision–ECCV 2016*, Springer, Berlin, Germany, 2016.

[3] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[4] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector,", October 2017.

[5] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[6] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: single-stage dense face localisation in the wild," 2019, https://arxiv.org/abs/1905.00641.

[7] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: a face detection benchmark," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[10] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Computer Vision—ECCV, 2018*, Springer, Berlin, Germany, 2018.

[11] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," 2019, https://arxiv.org/abs/1903.00621.

[12] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[13] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, December 2015.

[14] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affifinity fields," 2018, https://arxiv.org/abs/1611.08050.

[17] A. Newell, Z. Huang, and J. Deng, "Associative embedding: end-to-end learning for joint detection and grouping," 2017, https://arxiv.org/abs/1611.05424.

[18] X. Zhou, A. Karpur, L. Luo, and Q. Huang, "Starmap for category-agnostic keypoint and viewpoint estimation," in *Computer Vision–ECCV 2018*, Springer, Berlin, Germany, 2018.

[19] V. Jain, "FDDB: A Benchmark for Face Detection in Unconstrained Settings," UMASS Amherst Technical Report, University of Massachusetts, Amherst, MA, USA, 2010.

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[21] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Computer Vision–ECCV 2014*, Springer, Berlin, Germany, 2014.

[22] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, June 2016.

[24] J. Huang, V. Rathod, C. Sun et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[25] L. Liu, W. Ouyang, X. Wang et al., "Deep learning for generic object detection: a survey," 2018, https://arxiv.org/abs/1809.02165.

[26] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Computer Vision–ECCV 2018*, Springer, Berlin, Germany, 2018.

[27] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Consistent optimization for single-shot object detection," 2019, https://arxiv.org/abs/1901.06563.

[28] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: a context-assisted single shot face detector," in *Computer Vision–ECCV 2018*, Springer, Berlin, Germany, 2018.

[29] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8231–8238, 2019.

[30] C. Zhang, X. Xu, and D. Tu, "Face detection using improved faster RCNN," 2018, https://arxiv.org/ftp/arxiv/papers/1802/1802.02142.pdf.

[31] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "LFFD: a light and fast face detector for edge devices," 2019, https://arxiv.org/abs/1904.10633.
[32] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," 2017, https://arxiv.org/abs/1703.06870.
[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.
[34] X. Zhou, D. Wang, and P. Krahenbuhl, "Objects as points," https://arxiv.org/abs/1904.07850.
[35] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: efficient multi-scale training," 2018, https://arxiv.org/abs/1805.09300.
[36] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "FaceBoxes: a cpu real-time face detector with high accuracy," in *Proceedings of IEEE International Joint Conference on Biometrics*, pp. 1–9, IEEE, Denver, CO, USA, October 2017.
[37] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization," in *Proceedings of the In IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2144–2151, Barcelona, Spain, November 2011.

Hindawi
Scientific Programming
Volume 2020, Article ID 1394617, 7 pages
https://doi.org/10.1155/2020/1394617

Research Article

Quantile Regression Analysis of Depression and Clinical Symptom Degree in Chinese Patients with Spinocerebellar Ataxia Type 3

Xiaoping Chen, Lihui Zheng, and Jianqi Yao

Department of Statistics, College of Mathematics and Informatics & FJKLMAA, Fujian Normal University, Fuzhou 350000, China

Correspondence should be addressed to Lihui Zheng; fbx20180024@yjs.fjnu.edu.cn

Received 26 February 2020; Accepted 24 April 2020; Published 10 June 2020

Academic Editor: Chenxi Huang

Spinal cerebellar ataxia type 3 is a common SCA subtype in the world. It is a neurodegenerative disease characterized by ataxia. Patients exhibit common neuropsychological symptoms such as depression and anxiety. Some patients have suicidal tendencies when they are severely depressed. So, it is very important to study the severity of depression and clinical symptoms (SARA), to find out the patient's psychological state in time and to help patients actively respond to treatment. A total of 97 Chinese SCA3 patients were enrolled in the study. The Beck Depression Scale was used to investigate the prevalence of depression in the confirmed patients. The distribution of depression data in these patients was investigated. Then, the quantifier was used to model the depression status of Chinese SCA3 patients. An analysis was conducted to identify the key factors affecting depression under different quantiles. Studies have shown that SARA and gender are important factors affecting depression; the effect of initial SARA is small, then the degree of influence increases, and the degree of influence decreases in the later period, but it is always positively correlated with depression; the development of women's SARA is gentler than that of men, and the degree of depression is lower than that of men.

1. Introduction

Machado–Joseph disease or spinocerebellar ataxia 3 (MJD/SCA3) is a clinically heterogeneous, neurodegenerative disorder characterized by varying degrees of ataxia, ophthalmoplegia, peripheral neuropathy, pyramidal dysfunction, and movement disorder. MJD/SCA3 is caused by a CAG repeat expansion mutation in the protein coding region of the ATXN3 gene located at chromosome 14q32.1 [1]. So far, the pathogenesis and pathological effects of such diseases have not been clarified, and there was no effective treatment. Conventional treatment can only improve the clinical symptoms.

Previous studies on SCA3 focused the pathogenesis of SCA3, CAG mutation amplification, and ethnic differences. In the study of the role of the proteasome in the pathogenesis of SCA3/MJD, it was found that the proteasome plays a direct role in suppressing polyglutamine aggregation in disease. Also, the result suggested that modulating proteasome activity is a potential approach to altering the progression of this and other polyglutamine diseases [2]. The CAG repeat amplified in the SCA3 gene caused the patient's restless legs syndrome (RLS) and sleep impairment, and impaired sleep in SCA3 was associated with older age, longstanding disease, and brainstem involvement [3]. There were significant differences in predominant spinopontine atrophy, lack of dystonic features, and larger CAG repeat expansion between families with spinal cord ataxia 3 in two different ethnic origins in the United States [4]. The frequency of trinucleotide repeats in different ataxic categories in Chinese patients was studied to find that the frequency of SCA3/MJD is substantially higher than that of SCA1 and SCA2 in patients with autosomal dominant SCA from Chinese kindreds, who are non-Portuguese. Dementia and hyporeflexia were more frequent in patients with SCA2, while spasticity, hyperreflexia, and Babinski signs were more frequent in patients with SCA3/MJD, and those might be helpful in clinical work to primarily distinguish patients with

SCA3/MJD and SCA2 from others with different types of SCA [5].

SCA3 not only affected patients' ataxia and cognitive ability but also led to depression or anxiety with the development of the disease, seriously affecting patients' physical and mental health. Foreign studies, namely, the study of stuneuropsychological and neurobehavioral characteristics of SCA3 patients and the degree of emotional dysfunction of patients, found that, in addition to anxiety and depression, SCA3 patients also showed the development of increased apathy [6]. The research on the influence between SCAs depression and ataxia found that depression was very common in SCA and would not develop in 2 years. The suicidal ideation of SCA3 patients was significantly higher than that of other types, and there were differences in the suicidal ideation of SCA patients in different regions (North America and Europe). The effect of depression on the progression of ataxia varies by the SCA type [7]. Depression in patients with SCA3 has also been studied in China. Through the establishment of multiple linear equations, it was found that depression is caused by the movement disorder caused by ataxia. In addition, the two had interaction, and gender and ICARS were important factors affecting depression [8].

In summary, domestic and foreign scholars have conducted a large number of in-depth studies on the factors affecting the incidence of SCA3 and other related content. However, there are a few studies on depression in Chinese SCA3 patients, and the main method is linear regression. There are a few studies combining the quantile regression method with the influencing factors of depression in Chinese SCA3 patients. Considering the quantile, regression can effectively avoid the heteroscedasticity and nonnormal distribution of the data, and with the change of the quantile used, it can more accurately describe the influence of independent variables on dependent variables and characterize the conditional distribution. Therefore, it is necessary to study the factors affecting the quantile regression in SCA3 patients with depression.

2. Objects and Methods

2.1. Study Subjects. Participants in this study included 97 patients with confirmed SCA3, all of whom over 18 years of age. This study was approved by the ethics committee of the first affiliated hospital of Fujian Medical University. Written informed consent forms were signed by all subjects.

2.2. Observation Method. Beck depression inventory (BDI) was used to evaluate the severity of depression in SCA3 patients. There were 21 questions in the whole scale, and each question was divided into different grades. According to the total score of the scale, the severity of depressive symptoms can be evaluated. In this study, a BDI score greater than or equal to 19 was defined as clinically related depression [8].

The severity of the patient was assessed using the ataxia rating scale (SARA). SARA is a semiquantitative neurologic functional assessment scale that describes and quantitatively evaluates symptoms of cerebellar ataxia.

2.3. Statistical Analyses. Firstly, the data set was descriptively analyzed. The K-S discriminant method was used to analyze the normal distribution of data variables, and the mean and median values of the variables in the nondepressive group and the depression group were calculated.

Secondly, using univariate and multivariate quantile regression, different vacancies were taken, and the significance and correlation of each variable were judged according to the test P value. The density function maps of depression and SARA were plotted, as well as the density function maps by gender. The density function maps of depression BDI scores at 10% and 90% of the SARA were plotted. The dependent variable was the depression BDI score (Beck), and the independent variables were the gender, SARA score, disease duration, age of onset, small CAG, and large CAG.

Quantile regression [9] was a modeling method to estimate the relationship between regression variable x and dependent variable y at any probability level. Suppose the distribution function of the random variable was $F(y) = \text{Prob}(Y \le y)$, and define the τ quantile of y to be

$$Q(\tau) = \inf\{y: F(y) \ge \tau\}, \quad 0 < \tau < 1. \tag{1}$$

When τ was equal to 1/2, that was the median. In the quantile regression model, the loss function was defined as a piecewise linear function:

$$\rho_\tau(u) = (\tau - I(u < 0))u, \quad u = y_i - \widehat{\xi}, \tag{2}$$

where $\widehat{\xi}$ was the expectation $E(y)$ of the dependent variable and I was the indicator function. The basic idea of regression analysis was to minimize the distance between the sample value and the fitting value, so the optimization problem from the loss function expectation can be converted to another form:

$$\min \sum_{i=1}^{n} \rho_\tau\left(y_i - \widehat{\xi}\right). \tag{3}$$

Given the information set x and condition $Q_\tau(\varepsilon \mid x) = 0$, the conditional quantile function of y can be expressed as $Q_y(\tau \mid x) = x'\beta_\tau$.

The software tool was the R 3.5.1 version. Model results were significant at $P < 0.05$ and were statistically significant.

3. Result

3.1. Variable Descriptive Statistics. Of the 97 patients with SCA3, 41 were nondepressed and 56 were depressed. The depression rate was 57.73%, and SCA3 patients had a higher depression rate. The mean course of the patients with depression (9.259 ± 8.0) was significantly larger than that of the nondepressed group (6.329 ± 6); SARA (13.22 ± 13) was higher than that of nondepressed patients (8.085 ± 7). The depression degree (33.21 ± 32) in the depression group was 4 times higher than that in the nondepression group

Table 1: Descriptive statistics of variables.

	Distribution	Nondepressed group	Depression group
Number	——	41	56
Gender (M/F)	——	19/22	35/21
CAG small	Skewed	20.9 ± 20	21.57 ± 14
CAG large	Normal	74.34 ± 74.34	74.84 ± 75
Age at onset (years)	Normal	35.24 ± 34	33.59 ± 34.5
Disease duration	Skewed	6.329 ± 6	9.259 ± 8.0
SARA	Skewed	8.085 ± 7	13.22 ± 13
BDI score	Normal	8.878 ± 9	33.21 ± 32

Table 2: Univariate regression results.

	OLS	$Q_\tau(0.1 \mid X)$	$Q_\tau(0.25 \mid X)$	$Q_\tau(0.5 \mid X)$	$Q_\tau(0.75 \mid X)$	$Q_\tau(0.9 \mid X)$
CAG small	−0.005	$1.202e-17$	0.233	−0.116	0.071	−0.02
CAG large	0.032	0.333	0.75	0.25	−0.833	−1.33 ∗∗ (0.00092)
Age at onset (years)	0.059	0	−0.154	−0.2	0.36	0.429 ∗∗ (0.0005)
Disease duration	0.924 ∗∗ (0.00101)	0.538	1.190 ∗ (0.012)	1.333 ∗∗ (0.00036)	0.9375 ∗ (0.012)	0.571 ∗∗(0.005)
SARA	0.952 ∗∗ ($3.25e-06$)	0.727 ∗∗ (0.0043)	0.868 ∗∗ (0.0024)	1.273 ∗∗ ($3.044e-05$)	0.765 ∗∗ (0.004)	0.647 ∗∗ (0.00057)

Note. (1) If the P value is less than 0.01, it is considered that the effect of the independent variable is highly significant, which is represented by "∗∗." (2) If the P value is less than 0.05, it is considered that the effect of the independent variable is significant, which is indicated by "∗."

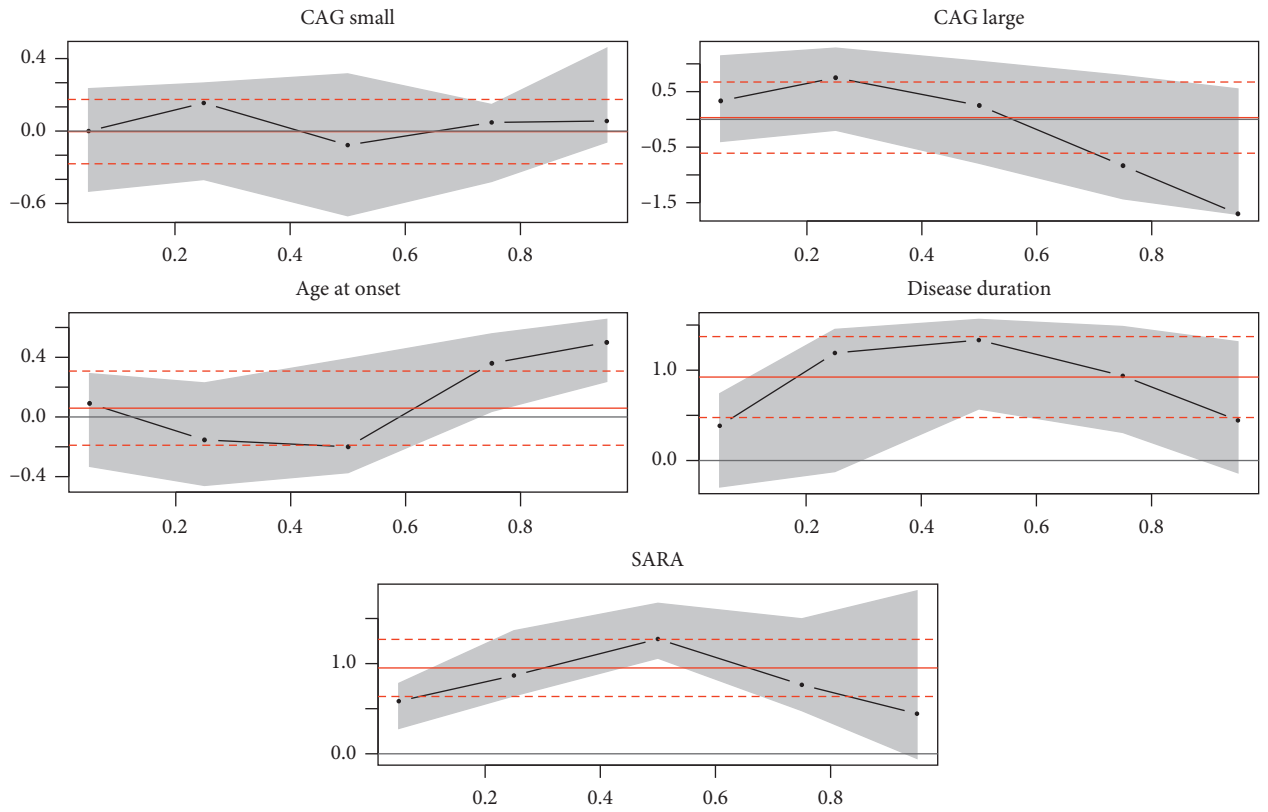


Figure 1: Variation of the quantile coefficient.

(8.878 ± 9), and the depression degree was relatively deep. In the data, the small CAG, the disease duration, and SARA were skewed. The large CAG, the age of onset, and Beck were normally distributed. Therefore, it is more appropriate to use the quantile regression relative least squares regression (the data are shown in Table 1).

Table 3: Estimated values of each quantile parameter.

	OLS		$Q_\tau(0.05 \mid X)$		$Q_\tau(0.1 \mid X)$		$Q_\tau(0.25 \mid X)$	
	Estimated value	P	Estimated value	P	Estimated value	P	Estimated value	P
(Intercept)	14.223	2.15e−07	−3.26	0.189	0.2	0.920	4.143	0.170
Gender (F)	− 6.196∗	0.0181	− 6.282∗	0.038	−3.4	0.296	− 4.143	0.184
SARA	1.036∗∗	4.48e − 07	1.026∗∗	1.098e − 07	0.8∗∗	4.08e − 07	0.905∗∗	7.7e − 4
	$Q_\tau(0.5 \mid X)$		$Q_\tau(0.6 \mid X)$		$Q_\tau(0.75 \mid X)$		$Q_\tau(0.9 \mid X)$	
	Estimated value	P	Estimated value	P	Estimated value	P	Estimated value	P
(Intercept)	11.571	3.3e−03	31.795	0.58844	24	5.0e − 09	34	1.25e −10
Gender (F)	− 6.0	0.102	−10.506∗∗	0.00049	−9∗	0.019	−5	0.18
SARA	1.238∗∗	3.1e − 06	1.010∗∗	0.00000	1∗	1.48e−05	0.67∗	0.02

Note. (1) If the P value is less than 0.01, it is considered that the effect of the independent variable is highly significant, which is represented by "∗∗." (2) If the P value is less than 0.05, it is considered that the effect of the independent variable is significant, which is indicated by "∗."

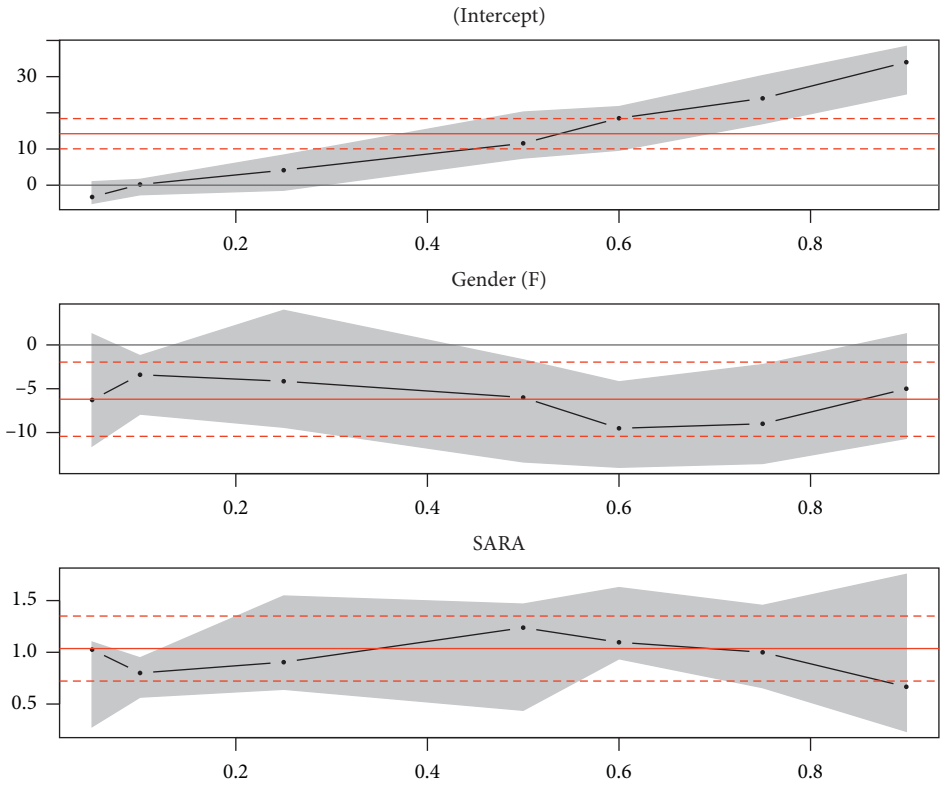


Figure 2: Variation of the quantile coefficient.

3.2. Univariate Regression Analysis. The Beck depression score was used as the dependent variable, and the CAG small, CAG large, age of onset, course of disease, and SARA were used as independent variables. Quantile regression and OLS regression were used to model the variables. The results were as shown in Table 2. From the regression results of the OLS model, it can be seen that in the linear regression model, the duration of disease, and SARA have a significant positive effect on depression. The longer the duration of disease is, the greater will be the SARA score and the more severe the depression will be. However, the CAG small, CAG large, and age of onset had no effect on depression.

The OLS model required that all variables obey normal distribution, but some of the data were biased. The validity of the OLS model fitting was limited, while the quantile regression did not require normal distribution. Columns 3–7 showed the quantile regression coefficients, and P values in parentheses. Compared with the regression coefficients of the OLS model, the quantile regression coefficients changed. The variation of the coefficients is shown in Figure 1, where the red horizontal line represents the OLS regression coefficient and the curve represents variation of the quantile regression coefficient.

In the quantile regression, the CAG was not significant at each quantile. The CAG and the onset age were not significant at other quantile, and only had an effect on depression at 90%. At this time, CAG was negatively correlated (−1.33). The age

of onset was positively correlated (0.429). From the low quantile to the high quantile, the CAG large coefficient symbol changed from positive to negative, and the onset age coefficient symbol changed from negative to positive.

SARA was significant at each quantile. The course of the disease was 10%, and the others were significant. The two variables were positively correlated with depression and the same as the univariate linear regression coefficient. When considering the univariate influence on depression, SARA and disease duration were important variables.

3.3. Multivariate Quantile Regression. In the establishment of multiple quantile regression, variable modeling with higher significance should be selected as far as possible. So, the course of disease and SARA were selected as independent variables. In the study of Lin et al. [8], gender was found to be an important factor affecting depression. So, gender was also selected as an independent variable. Through preliminary modeling, it was found that the disease course coefficient is not significant. Therefore, the disease course was excluded, and only the quantile model of SARA and gender was established.

The *R* software was used to conduct quantile regression on SCA3 patient data. Table 3 lists the parameter estimation, *P* value, and significance results of different quantile points $\tau = (0.05, 0.1, 0.25, 0.5, 0.6, 0.75, 0.9)$ in quantile regression.

Comparing the results of parameter estimation under 7 quantiles, it can be seen that SARA has always played a significant role, and the gender was significant when the quantile was 0.05, 0.6, and 0.75. Figure 2 is the parameter estimation chart of each variable under each quantile. In the figure, it can be seen that although the coefficients of the SARA points that have a significant effect on depression are changing, they were always positive, indicating a positive correlation with depression. When the quantile was small, it can be seen that the SARA coefficient has a decreasing trend. It showed that the effect of SARA on depression was small in the initial period of time. With the increase of SARA, the degree of depression will be aggravated, and the decrease in the coefficient value indicated that the development of SARA to the later stage has a reduced impact on depression.

Gender (female) was negatively correlated with depression, and only became significant when the coefficient value was less than about −6. When the depression score was used as the dependent variable, gender and SARA were used as independent variables to establish linear multiple regression. The coefficient of gender (female) was −6.196, the quantile at this time was between about 0.5 and 0.8, and the degree of depression in women was lower than that in men.

3.4. Density Function Diagram. Furthermore, the density function diagram of BDI scores and SARA, as well as the density function diagram of male and female, was drawn. From Figure 3, it can be more intuitively seen that the BDI score distribution was close to the normal distribution, and SARA presented a skewed distribution. The peak BDI scores of women and men were different, about 10 for women and 35 for men. The distribution of the SARA density function was relatively flat for women and steep for men.

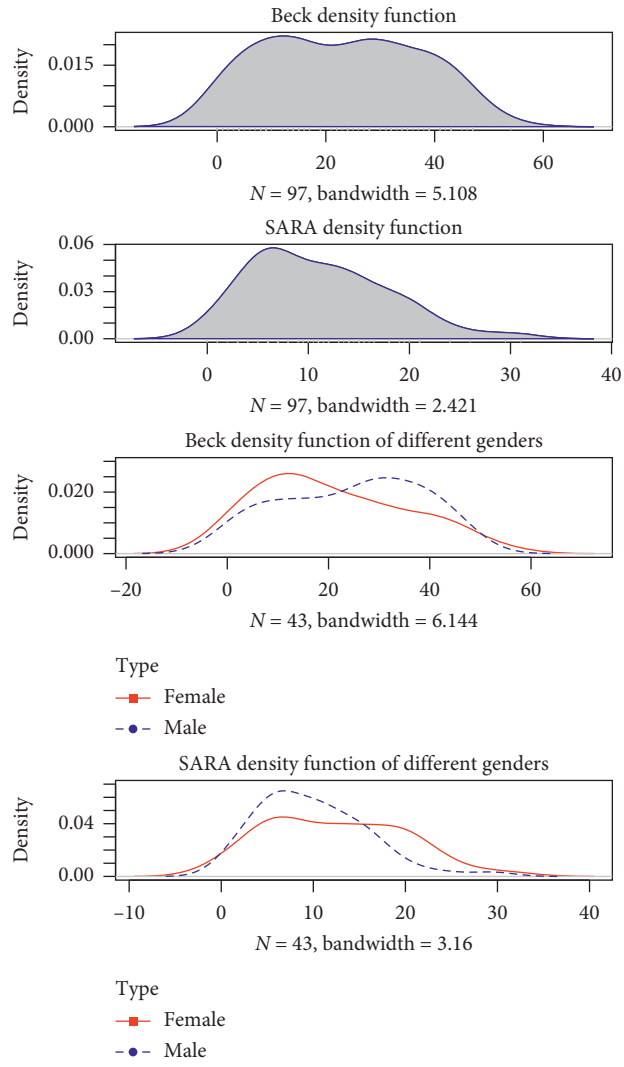


Figure 3: BDI score and the SARA density function.

In order to study the effect of SARA on depression, the effects of clinical symptom severity on low and high scores were compared, and the comparison of BDI scores at 10% and 90% of SARA was drawn (90% is the blue dotted line). From Figure 4, it can be found that the BDI scores of SARA at different points were significantly different. The higher the value of SARA, the higher the BDI score. The right figure reflected the density image of the BDI score in the case of a large SARA value and small SARA value. The larger SARA value was, the larger the BDI score was. The BDI score with a small SARA value was concentrated around 5, while the BDI score with a large SARA value was concentrated around 40.

4. Discussion

In this study, SCA3 patients had a higher rate of depression, and SARA had a significant effect on depression. In the early stage of development, SARA had less effect on depression. At this time, the patient's depressive symptoms were not

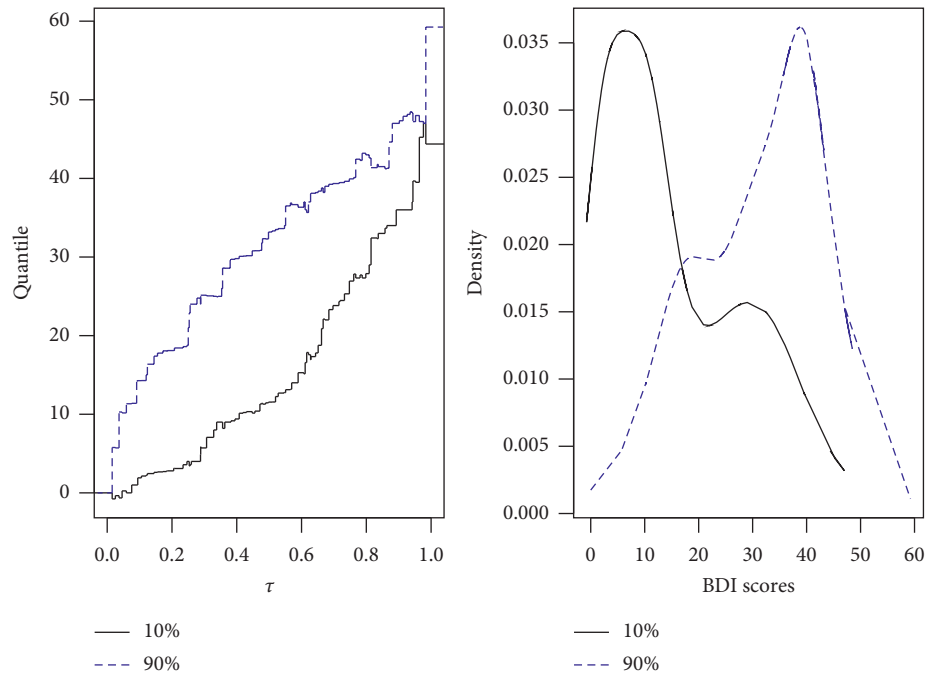


Figure 4: BDI score density map for SARA at 10% and 90%.

obvious. With the aggravation of clinical symptoms, depression of patients will be more affected and may be accompanied by anxiety, apathy, and other manifestations. By the end of the disease, SARA still had an effect on depression, but the degree of influence was reduced. Lo et al. [7] mentioned in their study that depression was very common in SCA and will not develop in 2 years. Therefore, it was speculated that the initial period when SARA has a small impact on depression may be 2 years. After 2 years, we need to pay attention to the impact of SARA on patients' depression. Some SCA3 patients in this article were also accompanied by suicidal tendencies, lethargy, fatigue, etc, which seriously affected the quality of life and physical health of patients. From the early stage of the disease, it was necessary to pay attention to the depressive symptoms and psychological state of SCA3 patients. The treatment of depressive symptoms may be beneficial to improve the clinical symptoms of patients and reduce the risk of suicide.

Gender was also an important factor affecting patients' depression. In this study, the depression degree of women was lower than that of men. Considering that the graph of SARA density function of women presents a relatively flat state and that of men was rather steep, the difference in SARA development between men and women may be the reason for the difference in depression. Therefore, SARA development of SCA3 patients of different genders can be considered in subsequent studies and targeted treatment programs for depression.

Data Availability

The data used to support the findings of this study are included within the supplementary information files.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors sincerely thank the participants for their help and willingness to participate in this study. This study was supported by the National Natural Science Foundation of China (11601083 and U1805263), the Program for Probability and Statistics: Theory and Application (IRTL1704), and Innovative Research Team in Science and Technology in Fujian Province University (IRTSTFJ).

Supplementary Materials

The authors provide the data in the supplementary information files that they submit alongside their manuscript. (*Supplementary Materials*)

References

[1] A. D'Abreu, M. C. França, H. L. Paulson et al., "Caring for Machado–Joseph disease:Current under-standing and how to help patients," *Parkinsonism and Related Disorders*, vol. 16, no. 1, pp. 2–7, 2010.

[2] Y. Chai, S. L. Koppenhafer, S. J. Shoesmith, M. K. Perez, and H. L. Paulson, "Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions in SCA3/MJD and suppression of polyglutamine aggregation in vitro," *Human Molecular Genetics*, vol. 8, no. 4, pp. 673–682, 1999.

[3] L. Schols, J. Haan, O. Riess, G. Amoiridis, and H. Przuntek, "Sleep disturbance in spinocerebellar ataxias: is the SCA3

mutation a cause of restless legs syndrome?" *Neurology*, vol. 51, no. 6, pp. 1603–1607, 1998.

[4] J. J. Higgins, L. E. Nee, O. Vasconcelos et al., "Mutations in American families with spinocerebellar ataxia (SCA) type 3: SCA3 is allelic to Machado-Joseph disease," *Neurology*, vol. 46, no. 1, pp. 208–213, 1996.

[5] B. Tang, C. Liu, L. Shen et al., "Frequency of SCA1, SCA2, SCA3/MJD, SCA6, SCA7, and DRPLA CAG trinucleotide repeat expansion in patients with hereditary spinocerebellar ataxia from Chinese kindreds," *Archives of Neurology*, vol. 57, no. 4, pp. 540–544, 2000.

[6] T. M. Zawacki, J. Grace, J. H. Friedman, and L. Sudarsky, "Executive and emotional dysfunction in Machado-Joseph disease," *Movement Disorders*, vol. 17, no. 5, pp. 1004–1010, 2002.

[7] R. Y. Lo, K. P. Figueroa, S. M. Pulst et al., "Depression and clinical progression in spinocerebellar ataxias," *Parkinsonism & Related Disorders*, vol. 22, pp. 87–92, 2016.

[8] M.-T. Lin, J.-S. Yang, P.-P. Chen et al., "Bidirectional connections between depression and ataxia severity in spinocerebellar ataxia type 3 patients," *European Neurology*, vol. 79, no. 5-6, pp. 266–271, 2018.

[9] L. Hao and D. Q. Naiman, *Quantile Regression*, pp. 40–52, SAGE Publications Inc, Thousand Oaks, CA, USA, 2007.

Hindawi
Scientific Programming
Volume 2020, Article ID 8658040, 9 pages
https://doi.org/10.1155/2020/8658040

Research Article

Relation Extraction Based on Fusion Dependency Parsing from Chinese EMRs

Pengjun Zhai,[1] Xin Huang,[1,2] Beibei Zhang,[1] and Yu Fang[1]

[1]*Department of Computer Science and Technology, Tongji University, Shanghai 200082, China*
[2]*School of Software, Jiangxi Normal University, Nanchang 330047, China*

Correspondence should be addressed to Yu Fang; fangyu@tongji.edu.cn

Received 8 February 2020; Accepted 20 April 2020; Published 8 June 2020

Academic Editor: Chenxi Huang

The Electronic Medical Record (EMR) contains a great deal of medical knowledge related to patients, which has been widely used in the construction of medical knowledge graphs. Previous studies mainly focus on the features based on surface semantics of EMRs for relation extraction, such as contextual feature, but the features of sentence structure in Chinese EMRs have been neglected. In this paper, a fusion dependency parsing-based relation extraction method is proposed. Specifically, this paper extends basic features with medical record feature and indicator feature that are applicable to Chinese EMRs. Furthermore, dependency syntactic features are introduced to analyse the dependency structure of sentences. Finally, the F1 value of relation extraction based on extended features is 4.87% higher than that of relation extraction based on basic features. And compared with the former, the F1 value of relation extraction based on fusion dependency parsing is increased by 4.39%. The results of experiments performed on a Chinese EMR data set show that the extended features and dependency parsing all contribute to the relation extraction.

1. Introduction

Electronic Medical Record (EMR) contains a vast of medical entities that provide rich medical knowledge. It is worth noting that there are certain interdependent relations between entities rather than isolated ones, which truly reflects the medical knowledge and the judgment and application of medical knowledge by doctors. The relations between entities in EMPs represent the health of patients from different perspectives. Relation extraction plays a fundamental role in medical knowledge graph (MKG) construction and completion and supports many other tasks, such as question answering, semantic understanding of texts, and recommender systems.

Entity relation in EMRs mainly includes the relation between treatment and disease, treatment and symptom, test and disease, test and symptom, and disease and symptom. At present, the machine learning method is widely used in the field of medical texts [1–4], including the task of relation extraction of English EMRs [5], and most of the feature selections rely on English medical dictionaries and data sets [6] as well as syntactic analysis [7]. However, the relation extraction of Chinese EMRs is still scarce, which is reflected in two aspects: one is the relation between two specific entities and the other is neglecting the unique features of Chinese EMR texts and sentences.

To cope with the above shortcomings, we proposed a fusion dependency analysis method for relation extract of Chinese EMRs. The underlying idea is to extend features according to the unique features of Chinese EMRs, such as medical records feature, indicators feature, and extended context feature. Considering that the entity relations in two sentences with similar structure and context are often the same and the structural similarity of sentences in Chinese EMR is high, the sentence structure information is fused based on the feature extension. Among the methods of machine learning, some research studies [8, 9] have verified that SVM is a better method for entity relation extract; thus, this paper directly adopts SVM to train the model and predict.

2. Related Work

The concept of relation extraction was first put forward at the Message Understanding Conference (MUC) and supported by the Defense Advanced Research Projects Agency (DARPA) at the end of the 1980s. After that, the Automatic Content Extraction Conference (ACE) promoted the development of relation extraction technologies. Recently, the development of knowledge graph (KG) once again emphasizes the importance of relation extraction.

2.1. Relation Extraction of English EMRs. The relation extraction methods of EMRs are evolved from the early methods based on rules and dictionaries to the current classification based on machine learning, where entity relation refers to the relation between entity pairs appearing in a sentence. For the relation extraction of English EMRs, an SVM model [10] was utilized to identify the relationships among disease, symptom, test, and treatment. In this research, semantic lexical features, the order of entity pairs appearing in sentences, and syntactic features have been added as classifier and present an SR classifier, which can recognize 84% of the relations in the BIDMC corpus and achieve microaveraged F-measures of 0.89. A model was described in a study [11] to identify the semantic relations among medical concepts, including problems, tests, and treatments, from the medical texts and to analyse three types of relations which are the relation between treatment and problem, test and problem, and problem and problem. To extract the above relations, a hybrid method was proposed based on machine learning, dictionary, and rules [12]. In the view of the I2B2 (Informatics for Integrating Biology and the Bedside) 2010 (https://www.i2b2.org/NLP/Relations/), Rink [6] used GENIA15 to pre-processed the medical record texts, and then selected the context similarity as the new feature based on the lexical feature and context feature. The task of feature extraction used knowledge bases such as Wikipedia, WordNet and general inquirer [13]. This model also uses the SVM model to achieve the F-measures of 0.74. The relations between concepts in UMLS were used as a substitute feature to solve the problem that some entities in EMRs do not have rich context features [14], and the experimental results obtained an F-measures of 0.67.

2.2. Relation Extraction of Chinese Text. At present, the research studies of relation extraction in Chinese mainly focus on the open domain and the methods of relation extraction in Chinese EMRs are still in the preliminary stage. A pipeline of NLP techniques was employed [15], a.k.a., word segmentation, POS-tagging, and syntactic parsing, to extract entity relations for an open domain. This system was considered as the first attempt to handle Chinese open relation extraction. In the medical field, the dependency graph was used to automatically learn the syntactic pattern of relation extraction and extracted the relation between disease and symptom by this model [16]. Also, a rule-based method was used to extract medical information for unstructured text data in EMRs [17]. The bootstrapping framework based on semisupervision was proposed in a study [18], combined TCM bibliographic literature database in China and MEDLINE (https://jgc128.github.io/mednli/), to discover the knowledge of gene functional including extracting the relation between symptom and gene, symptom and disease, and disease and gene. According to the characteristics of the relation between entities in the EMRs, a semisupervision learning method was used [19], SVM was adopted as the classifier to predict the labeled samples combined with auxiliary classification information, and then the classification after adding the samples with low confidence to the training set was repeated, which shows that the entity relation can be extracted effectively by the method of classifying and calculating entity co-occurrence.

3. Methods

In this section, we first introduce the preprocessing method of Chinese EMR data. Second, we briefly describe the basic features for relation extraction of Chinese EMRs. And we extend the features based on the basic features, according to the characteristics of Chinese EMR texts. Finally, by fusing sentence structure information, a method of relation extraction based on dependency parsing is proposed. The relation extraction process is shown in Figure 1.

3.1. Data Preprocessing. The data set used in this paper for the research of entity relation extraction comes from XML EMR texts that were preprocessed initially and the files of entity and entity relation that have been tagged from EMR by a semiautomatic annotating method, which is described in Section 4.1. Among them, discharge summaries and progress notes [20] are selected as the Chinese EMR texts. The discharge summary includes the basic information of the patient at the time of admission, the diagnosis of the doctor, the tests and treatments received in the process of hospitalization, the basic information and the doctor's advices at the time of discharge, and the final treatment results. The details of discharge summaries are shown in Figure 2. The process note mainly records the clinical manifestations of the patients during hospitalization and the medical behaviours such as test and treatment received.

The process of data preprocessing is roughly divided into three parts. First, the EMR texts should be segmented by using "。," "；," and "\n" as the boundary of sentences. Then, entity pairs need to be identified from the EMR texts. Finally, the EMR texts that have completed sentence segmentation are tagged with word segmentation and part-of-speech, with the help of NLPIR (http://ictclas.nlpir.org/) that is a word segmentation tool.

3.2. Relation Types. Relation extraction is used to find the relation between entities from the text, while the relation extraction of the EMRs entity mainly studies the relation between entities such as disease, symptom, test, and treatment recognized from the EMRs. These entity relations reflect the health information of patients and medical treatment measures for patients, as well as the professional

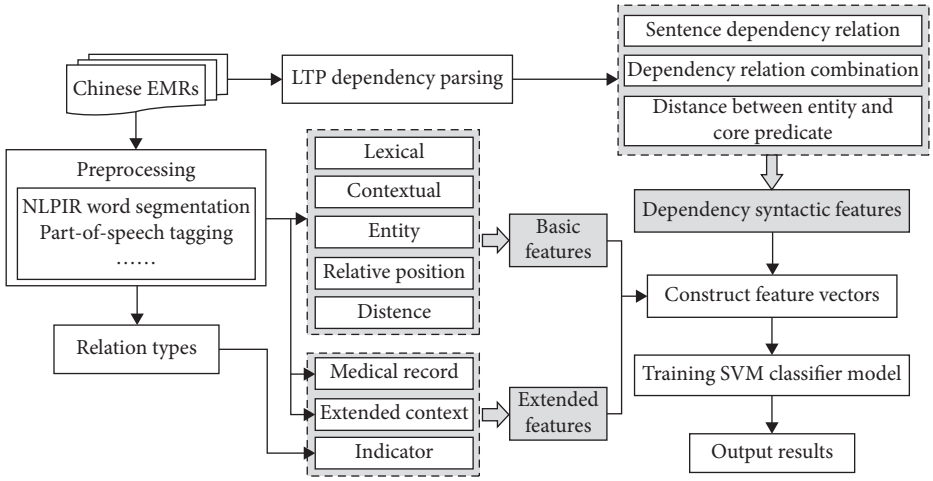


Figure 1: Relation extraction process.

Name	Section office : hepatobiliary department No.1 of general surgery (inpatients)	Inpatient ward :	Bed number :	Admission number :

Name	Gender : male	Age : 68	Radiation number :
Admission time	28/02/2015	Discharge dates : 01/03/2015	
Condition of hospitalization	The patient came to the hospital because of "The mass of right neck has existed for 5 years and increased significantly in more than 2 months" after physical tests, it was found that there was a 0.6*0.6 cm bag on the right neck. No redness or swelling,tender, no skin rash, no ulceration		
Admitting diagnosis	Right cervical mass		
Major laboratory tests and special examination	Null		
Diagnosis and treatment process	After admission, relevant tests were carried out, and the right neck mass was removed under local anesthesis immediately. The operation was smooth and the patient returned safely. Now the patient is in a stable condition and can be discharged from the hospital and become an outpatient		
Pathological results	Come to the hospital ten days later to see the pathology reports		
Discharge diagnosis	Right neck sebaceous cyst		
Hospital discharge	The patient has no discomfort		
Discharge orders	1 Follow-up, Monday, Chief physician shi baomin ward round and records. 2 Health education, five days after operation have stitches out 3 Rehabilitation guidance: avoid touching water		
Treatment results	Cured		

Attending surgeon:	*	Residents:	*

Figure 2: Discharge summary.

knowledge of doctors. For the first time, the assessment task of I2B2 2010 systematically classifies the entity relation of EMRs, including the relation between medical problem and medical problem, medical problem and test, and medical problem and treatment. According to the characteristics of Chinese EMR texts, this paper divides the medical problem in I2B2 2010 into two categories as disease and symptom and then redefines the relation between medical entities as the relation between treatment and symptom, treatment and symptom, test and disease, test and symptom, and disease and symptom. The specific definitions are shown in Table 1.

3.3. Basic Features. Features play an irreplaceable role in the task of relation extraction, especially for Chinese EMR texts. This paper first introduces the basic features of entity relation extraction for Chinese EMRs and references the features of open-domain text for relation extraction,

Table 1: Entity relation type and description.

Relation type	Relation representation	Representation description
Relation between treatment and disease	TrID	Treatment improves disease
	TrWD	Treatment worsens disease
	TrCD	Treatment causes disease
	TrAD	Treatment applied to disease
	TrNAD	Due to disease, not adopting treatment
Relation between treatment and symptom	TrIS	Treatment improves symptom
	TrWS	Treatment worsens symptom
	TrCS	Treatment causes symptom
	TrAS	Treatment applied to symptom
	TrNAS	Due to symptom, not adopting treatment
Relation between test and disease	TeRD	Test confirms disease
	TeCD	For confirming disease, adopt test
Relation between test and symptom	TeRS	Test discovers symptom
	TeAS	Due to symptoms adopt test
Relation between disease and symptom	DCS	Disease causes symptom
	SID	Symptom indicates disease

which are mainly divided into lexical feature, contextual feature, entity feature, and location feature.

(1) Lexical: this involves the two entities themselves, which play a certain role in the relation extraction between them, because even if two specific entities appear in different places, the relation between them may be the same. For instance, the relation between "感冒 (cold)" and "发烧 (fever)" in "患者因感冒而发烧 (patients have fever due to cold)" is usually "DCS (disease causes symptoms)", so this paper also takes the two entities themselves as a feature.

(2) Contextual: in Chinese EMR texts, the bag-of-words and part-of-speech in a certain range before and after two entities play a key role in the extraction of the relation between the two entities. The entity relation is judged by the context information, which refers to three bag-of-words and part-of-speech before and after two entities in this paper.

(3) Entity: the entity feature refers to the type of entity, which is an extremely important feature because the entity relation in this paper is classified by the two types of entity. Among them, the entities of test and treatment type only have relations with two types of entities that are disease and symptom, and there is a relation between disease and symptom instead of the relation between test and treatment. This feature has important guiding significance for the boundary judgment and specific type of judgment of entity relation.

(4) Relative position: the relative position of two entities, E1 and E2, has a certain indicative function for entity relation extraction in a sentence of Chinese EMR texts. For most sentences in the Chinese EMR data set of this paper, the disease entities and symptom entities appear in front of test entities and treatment entities, while the disease entities generally appear in front of symptom entities. For example, the disease entity "胆结石 (gallstone)" is in front of the treatment entities "全胆囊切除术 (total cholecystectomy)" in the Chinese EMR text "1974年因胆结石于瑞金医院行全胆囊切除术 (in 1974, total cholecystectomy was performed in Ruijin hospital due to gallstones)," and the relation between two entities is "TrAD (treatment applied to disease)". There are four categories of relative positions of two entities in this paper: E1 is on the left of E2, E1 is on the right of E2, E1 is in E2, and E2 is in E1.

(5) Distance: the distance between two entities refers to the number of words between them. In general, the more words there are between two entities, the farther apart they are, and the less likely there is a relation between them. The distance between two entities is expressed by measuring the numbers of words between two entities after word segmentation, in which words contain punctuation marks.

3.4. Extended Features. In order to achieve the task of extracting entity relation of Chinese EMRs more accurately, after analysing the texts of Chinese EMR, this paper extends the features of EMRs based on the basic features that are named extended features, which are mainly divided into medical record features, indicator features, and extended context features.

(1) Medical record: the chapter in which the entity located has a certain effect on entity relation extraction of Chinese EMRs. For example, in the "出院情况 (discharge situation)" chapter of discharge summary in Chinese EMRs, the probability of relation related to improvement is higher than that related to worsening. In addition, the modification information of an entity is also unique information in EMRs, which is a description of the entity. To sum up, the medical record features refer to chapters and modifications of entities.

(2) Indicator: the mapping of entity context words and the indicator word base for entity relation are

regarded as indicator features in Chinese MERs. According to the characteristics of Chinese MERs, the judgment of entity relation is related to the context words of two entities. There are some indicators that can directly classify the relation between two entities. If there are indicators such as "好转(improved)," "有所缓解(relieved)," "明显好转(obviously improved)," and "控制稳定(stable control)," the entity relation is generally "TrID" or "TrIS". If there are indicators such as "控制不佳 (poor control)," "效果一般 (general effect)," "未见明显变化 (no obvious change)," the entity relation is generally "TrWD" or "TrWS." After analysis and statistics, the indicator word base of all entity relation is established, and the mapping of the two entity's words in the indicator word base is regarded as an extended feature.

(3) Extended context: in a sentence, there are many entities juxtaposition, which makes it impossible to find the words with indicative meaning to the relation in the entity context. For instance, in the sentence "患者7天前无明显诱因出现腹胀 (the patient had no obvious inducement to develop abdominal distention), 伴右上腹钝痛 (accompanied by dull pain in the right upper abdomen), 伴乏力,食欲减退 (fatigue and anorexia), 伴皮肤,巩膜黄染 (yellow staining of skin and sclera), 无腹泻,黑便 (no diarrhea or black stool), 无寒战 (no shivering), 高热 (fever), 恶心 (nausea), and 呕吐 (vomiting)," there are many words which are useless for relation extraction in a certain range near the entity. Therefore, this paper extends the general context feature and selects verbs near the entity as the extended context feature.

3.5. Dependency Parsing. Most of the Chinese EMRs are long sentences, and the content and form of the sentences are relatively patterned, especially the structure of sentences that are mostly similar. Therefore, it is worth adding the structure information of sentences to the task of entity relation extraction from Chinese EMRs. Dependency parsing reveals the syntactic structure of a sentence by analysing the dependency among its components. In a word, it is to recognize the grammatical components such as "subject predicate object" and "attributive adverbial complement" and analyse the relationships between them. It claims that the dominator of a sentence is the core verb [21] and that all the dominators depend on the core verb in one way or another.

The language technology platform (http://ltp.ai/) (LTP) of the Harbin University of Technology is a complete set of Chinese language processing system developed by the social computing and information retrieval research center of the Harbin University of Technology. It provides rich, efficient, and accurate natural language processing technologies, including Chinese word segmentation, part-of-speech tagging, dependency parsing, and semantic role tagging. Using the LTP to analyse the dependency of the sentence "the patient having symptoms of wheezing and fever was given anti-infection treatment and relieved after the treatment of antiasthmatic (患者出现喘息, 伴发热, 予抗感染, 平喘治疗后缓解)." The results are shown in Figure 3.

Dependency parsing is to analyse the structural information of a sentence, recognize the "subject predicate object" and "attributive adverbial complement," and analyse the relationships between the components. According to the dependency parsing of example sentences in Figure 3, the core predicate of the sentence is "出现 (has)," the dependency of entity "喘息(wheezing)" and "出现(has)" is VOB, and the dependency of the entity "抗感染(anti-infection)" is VOB as well. Table 2 shows the annotation relation obtained from dependency parsing by LTP.

3.6. Dependency Syntactic Features. In this section, sentence structure and features will be integrated to get dependency syntactic features for better mining syntactic construction and semantic features, where the sentence structure is reflected in dependency parsing and sentence similarity calculation by using the algorithm of edit distance. The specific dependency syntactic features are defined as follows:

(1) Sentence dependency relation of binary entities: this is referred to the syntactic relations between two entities in the syntactic structure of a sentence after dependency parsing. For instance, the dependency relation of entity "喘息 (wheezing)" after parsing is VOB in the above example (Figure 3) and the dependency relation of the entity "发热 (fever)" is COO. Therefore, this paper takes the dependency parsing value of each entity in the entity pairs as a feature.

(2) Dependency relation combination of entity pair: the last feature is to take the dependency relation of entity pair as a feature input, while this feature refers to the dependency relation combination of entity pair, which is sequential. Because of this sequential, the syntactic structure of entity pairs in sentences can be shown more clearly by analysing the combinatorial feature than by analysing the independent dependency relation feature. For example, the dependency relation of entity pair <喘息(wheezing), 抗感染(anti-infection)> in the above example (Figure 3) is VOB-VOB, indicating that both entities act as an object in VOB. Different types of relationships have different dependency relation combinations, so this dependency syntactic feature can better reflect the differences of relation types between different entities.

(3) The distance between a binary entity and core predicate: after a lot of research studies and experiments on dependency parsing, it is found that the core predicate plays an important role in the extraction of entity boundary and entity relations. In a sentence, the distance between the entity and the core predicate is obviously different from that between the entity and the common predicate, so this paper takes the former as a feature. After the core

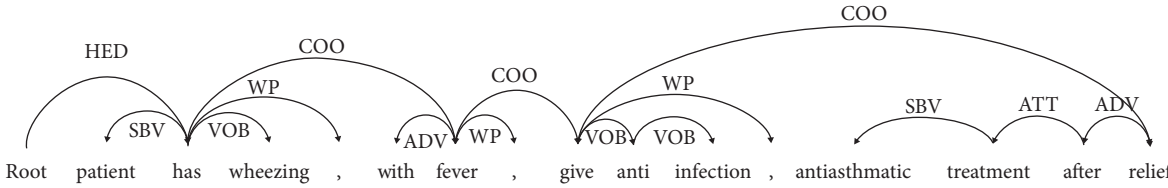


Figure 3: Dependency parsing of example sentences.

Table 2: Tag set of dependency parsing.

Labels	Relation types	Description
ADV	Adverbial-centered relation	Adverbial
ATT	Attribute-head relation	Attribute
COO	Coordinate relation	Coordinate
HED	Head relation	Head
SBV	Subject-verb relation	Subject-verb
VOB	Verb-object relation	Verb-object
WP	Punctuation	Punctuation

predicate of a sentence is obtained by dependency parsing, the distance between the entity and the core predicate is calculated by calculating the number of words between them based on the location of the core predicate.

3.7. SVM Model. The objective of the support vector machine model [22] is to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. In order to find a plane that has the maximum margin, i.e., the maximum distance between data points of both classes, we turn it into a convex quadratic programming problem.

Given a training sample set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in \mathbb{R}^n$ is the ith feature vector and y_i is the label of classes, denoted as $y_i \in \{+1, -1\}$, $i = 1, 2, \ldots, n$, the hyperplane is defined as follows:

$$w^T x + b = 0, \#, \tag{1}$$

where $w = (w_1, w_2, \ldots, w_d)$ is the normal vector of the hyperplane and defines the direction of the hyperplane and b is the intercept that determines the distance between the hyperplane and the origin. Due to the correctness of classification being judged by observing whether $w^T x + b$ and y are both positive or negative numbers, a function of margin γ' should be defined as follows:

$$\gamma' = y\left(w^T x + b\right), \#. \tag{2}$$

In order to unify the measurement, constraints are added to the normal vector w:

$$\gamma = \frac{y\left(w^T x + b\right)}{||w||_2} = \frac{\gamma'}{||w||_2}, \#. \tag{3}$$

The idea of the SVM is to maximize the margin, so that the distance from all points to the hyperplane is greater than or equal to a certain distance; then, all classification points are classified on both sides of the support vector, i.e.,

$$\begin{aligned} &\max \gamma = \frac{y\left(w^T x + b\right)}{||w||_2}, \\ &s.t.\ y_i\left(w^T x_i + b\right) = \gamma'^{(i)} \geq \gamma', \quad i = 1, 2, \ldots, n, \#. \end{aligned} \tag{4}$$

If the function of margin $\gamma' = 1$, then equation (4) is reduced to

$$\begin{aligned} &\max \frac{1}{||w||_2}, \\ &s.t.\ y_i\left(w^T x_i + b\right) \geq 1, \quad i = 1, 2, \ldots, n, \#. \end{aligned} \tag{5}$$

Considering that maximizing the $1/||w||_2$ is equal to minimizing the $1/2||w||_2$, the SVM model for solving the maximum partition hyperplane problem can be expressed as the following constrained optimization problem:

$$\begin{aligned} &\min \frac{1}{2}||w||_2, \\ &s.t.\ y_i\left(w^T x_i + b\right) \geq 1, \quad i = 1, 2, \ldots, n.\#. \end{aligned} \tag{6}$$

4. Results

In this section, we carry out three comparative experiments based on basic features, extended features, and dependency syntactic features. The experimental results show that structural information is very important for entity relation extract of Chinese EMRs, an irreplaceable role in the task of relation extraction, especially for Chinese EMR texts.

4.1. Data Set. We evaluate our approach of entity relation extraction on the medical dataset from the existing research [23]; this dataset is semiautomatic and annotated from Chinese EMRs of a grade-three general hospital in Shanghai for a whole year, and the entity set is obtained through the method of feature-enhanced entity recognition. The detailed information of the data set is shown in Table 3. We use 70% of the dataset as training data and 30% for testing. For readers interested in this data set, it is recommended to read academic study [23].

4.2. Baseline. In this paper, the task of entity relation extraction can be transformed into a multiclassification problem. The machine learning tool of LibSVM [24] is used to automatically build multiple binary classifiers according to the number of categories, which can be directly used for

Table 3: Entity instance distribution.

Entity types	Summary of discharge	First disease process	Total
Disease	905	1519	2424
Symptom	1407	2225	3632
Test	599	986	1585
Treatment	1045	1264	2309
Total	3956	5994	9950

multivalue classification. Therefore, this paper uses a LibSVM tool to train and test the SVM model, which has certain requirements for the data format of training and test data set, and the data format of input files is shown in Figure 4. Each row of data in Figure 4 represents a training vector, and the 'label' represents the identification of each classification label in this multiclassification, the 'index' is the number of features, and the 'value' is the value of features. In this paper, all data sets trained and tested by LibSVM are transformed into data files of this format for experiments after feature extraction and feature vector construction.

In order to compare the effects of extended features and sentence structure information on the experimental results of entity relation extraction in Chinese EMRs, three contrast experiments are set up in this paper. The first experiment is the baseline experiment, which selects the basic features including lexical feature, contextual feature, entity feature, and location feature. And the second experiment adds the extended features based on the basic features, while the last experiment adds the dependency parsing to the features to form the dependency syntactic features. The results of experiments are evaluated by 3 types of indicators [25]: Precision (P), Recall (R), and F1.

4.3. Results and Analysis. The experimental results of relation extraction based on different features for the data set are shown in Table 4. As we expected, the method of fusing dependency parsing outperforms the relation extraction method based on basic features or extended features. For the baseline, the extraction effect of entity relations of TrCD, TrNAD, and TrNAS is poor. This is because these three types of relationships appear less frequently (less than 5 times) in the tagging corpus. While the precision of TeRS is high, not only because this relation type appears more frequently in the training corpus but also because the characteristics of this relation are obvious, in which sentence pattern is basically "胸片示 (chest X-ray shows): 双肺纹理增多 (bilateral lung marking are increased), and 模糊 (blurred)". In addition, the extraction effect of SID and TeRD is better, which is also due to the obvious surface features and more training data. However, the relation extraction precision of TrID, TrIS, TrWd, and TrWS is low because of the existence of long sentences in Chinese EMRs, and only the contextual features of before and after words are not obvious.

While after adding the extended features proposed in this paper, in addition to the three unextracted relation types of TrCD, TrNAD, and TrNAS, the precision and recall rate of all other relation types have been improved, among which the improvement effect of the four types is stronger: TrID, TrWd, TrIS, and TrWS. This is because the medical record features in the extended features (including chapter information and entity modification information) have some influence on the location of entity relation. For example, in the chapter of "出院情况 (discharge situation)," the incidence of entity relation related to improvement is higher than that related to worsening. The indicator features in extended features are more effective for the relation types of improving and worsening because there are related demonstratives (好转 (improvement), 稳定 (stability), 一般 (general), 不佳 (poor), etc.) before and after the entities of improving and worsening. The verb features of entity to the front and back in the extended features are also of great significance to the entity relation extraction of Chinese EMRs. Due to the long sentence in the texts of Chinese EMR, the words before and after many entity pairs are meaningless for entity relation extraction, while the verbs before and after entity pairs generally have certain indicative meanings.

<lable1> <index1>:<value1> … <index n>:<value n> …
… … … … … … …
<lable2> <index1>:<value1> … <index n>:<value n> …
… … … … … … …

Figure 4: Data format of training and test data set.

As shown in Table 4, the precision and recall rate of all entity relations have been significantly improved after adding the dependency syntactic features. Dependency parsing is mainly to mine deeper structure information of sentences based on the surface semantic features. Obviously, the three dependency syntactic features added in this paper still greatly improve the precision rate of TrID, TrWd, TrIS, and TrWS, as well as the improvement effect on the TeRD and TeRS. It is because the sentence patterns of "treatment discover symptoms" and "treatment confirmed diseases" are very similar and unified. Many sentences are the patterns of "a certain test: symptom description or disease description" or "test shows: symptom description or disease description," so this characteristic can be mined by the dependency parsing.

The values of F1 for the relation extraction based on different features show the trend of the effects of entity relation extraction. In the case of limited training corpus, the performance of each entity relation is improved after fusing extended features and dependency syntactic features. Particularly, it is more effective for the several relation types (TrID, TrWD, etc.) that are relatively few in the corpus. However, our method is not very effective in the extraction of three types of relation TrCD, TrNAD, and TrNAS, because the number of these three relation types in the corpus is too small. The future research direction can be focused on how to generate relevant corpus or mine deeper features when the number of the corpus is small.

Table 4: Performance on relation extraction for different features.

Relation types	BFs			BFs + EFs			BFs + EFs + DSFs[1]		
	P	R	F1	*P*	R	F1	*P*	R	F1
TrID	52.76	61.67	56.87	**59.75**	68.32	63.75	**64.34**	72.09	67.99
TrWD	49.31	51.31	50.29	**56.79**	60.57	58.62	**61.09**	65.83	63.37
TrCD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TrAD	68.57	64.97	66.72	71.19	72.76	71.97	76.93	77.93	77.43
TrNAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TrIS	59.13	68.12	63.31	**64.41**	71.43	67.74	**68.01**	76.83	72.15
TrWS	51.97	59.32	55.40	**55.46**	62.82	58.91	**62.24**	67.42	64.73
TrCS	69.13	73.97	71.47	72.73	86.72	79.11	77.03	90.43	83.19
TrAS	58.34	61.45	59.85	62.61	64.86	63.72	66.78	68.92	67.83
TrNAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TeRD	73.21	77.61	75.35	77.89	80.75	79.29	**82.89**	84.79	83.83
TeCD	59.32	61.75	60.51	63.09	64.09	63.59	68.98	69.76	69.37
TeRS	81.78	89.31	85.38	84.98	93.73	89.14	**86.92**	92.67	89.70
TeAS	61.79	62.45	62.12	65.89	67.63	66.75	71.89	72.83	72.36
DCS	58.71	62.75	60.66	61.72	68.53	64.95	65.27	72.46	68.68
SID	75.82	78.97	77.36	80.62	81.41	81.01	84.81	85.42	85.11
Total	63.07	67.20	65.07	67.47	72.59	69.94	72.09	76.72	**74.33**

BFs represent basic features, extended features are represented as EFs, and DSFs are used to represent dependency syntactic features.

5. Conclusions

This paper implements the extraction of entity relations in Chinese EMRs. The relation types of extraction include the relations between treatment and disease, treatment and symptom, test and disease, test and symptom, and disease and symptom. And the machine learning method is used to transform the task of relation extraction into the classification of entity pairs, which mainly uses the SVM model for training and testing. The similarity of sentences brings a lot of hints to entity relation, i.e., generally, the relation between two entities in sentences with similar sentence structures and semantics is the same. First, this paper proposes four basic features of general text, such as lexical feature and location feature. Second, due to the juxtaposition of many entities or words in Chinese EMR texts, the simple context information is redundant and noisy, so the extended feature is proposed, which is composed of chapter information and indicator feature. In addition, because the basic features and extended features are the only superficial semantic features, but ignoring the information of sentence structure, LTP tool is used to analyse the dependency parsing of Chinese EMR texts and introduce the dependency syntactic features. In this paper, an SVM model is adopted to train and test entity relation extraction. Three comparative experiments are designed for the above three types of features. The results show that the extended features and dependency syntactic features proposed in this paper improve the accuracy and recall rate of entity relation extraction of Chinese EMRs to a certain extent. However, the training set and test set used in this paper are limited in scale. In the future, it is necessary to study the deep learning method for a largescale corpus to extract entity relations more efficiently.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank the hospital for its contribution, which provides electronic medical records that are used as the data set for the experiments in this paper. This research was funded by the National Key Research and Development Program of China (No. 2019YFB2101600).

Supplementary Materials

Due to the privacy of the data set of our medical EMRs, we selected some experimental data as samples. The details of the supplementary materials file are as follows: (1) the discharge folder is the data of discharge summary, which includes the data of training, test, discharge summary relation, and discharge summary entity. The files in the folders of train and test are the discharge summaries of patients, which are used to train and test models including condition of hospitalization, admitting diagnosis, diagnosis and treatment process, discharge diagnosis, hospital discharge, and discharge orders. The files in the folders named dischargeEntity are the medical entities of discharge summaries. Every line in the files corresponds with entity information in discharge summaries, which are tagged with "C = entity P = start: end T = entity type A = entity assertion," where C represents the concepts of entities in discharge summaries, P means the start and end position of entities in medical EMR texts, and T and A stands the type of entities and the modification of entities, respectively. The files in the folders named dischargeRelation are the entity relations of discharge summaries. Every line in the files corresponds with the relation between entities in medical discharge summaries, which are tagged with E = {entity[strat-end]entity type; ...;}‖R = ‖E = {entity[strat-end]entity type;...;}, where the first E represents the first entity, including the entity concept, the start-end position, and type of entity. Similarly, the second E represents the second entity. And the middle R represents the relation type between the two entities. (2) The progress

folder is the data of progress record, which includes the data of training, test, progress record relation, and progress record entity. The files in the folders of train and test are the progress records of patients, which are used to train and test models including characteristics of case, preliminary diagnosis, and plan of diagnosis. The files in the folders named progress Entity are the medical entities of progress records. Every line in the files corresponds with entity information in progress records, which are tagged with "C = entity P = start: end T = entity type A = entity assertion," where C represents the concepts of entities in progress records, P means the start and end position of entities in medical progress records, T and A stands the type of entities and the modification of entities respectively. The files in the folders named progressRelation are the entity relations of progress records. Every line in the files corresponds with the relation between entities in progress records, which are tagged with E = {entity [strat-end]entity type;...;}$\|R = \|E$ = {entity[strat-end]entity type;...;}, where the first E represents the first entity, including the entity concept, the start-end position, and type of entity. Similarly, the second E represents the second entity. And the middle R represents the relation type between the two entities. (*Supplementary Materials*)

References

[1] S. Gupta and A. K. Manjhvar, "Relation classification from unstructured medical text using feature based machine learning approach," in *Proceedings of the International Conference on Trends in Electronics and Informatics*, ICEI), Tirunelveli, Tamilnadu, pp. 1135–1138, May 2017.

[2] P.-H. Chen, H. Zafar, M. Galperin-Aizenberg, and T. Cook, "Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports," *Journal of Digital Imaging*, vol. 31, no. 2, pp. 178–184, 2018.

[3] M. Lu, Y. Fang, F. Yan, and M. Li, "Incorporating domain knowledge into natural language inference on clinical texts," *IEEE Access*, vol. 7, pp. 57623–57632, 2019.

[4] C. X. Huang, X. Huang, Y. Fang et al., "Sample imbalance disease classification model based on association rule feature selection," *Pattern Recognition Letters*, vol. 133, pp. 280–286, 2020.

[5] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, "UIMA Ruta: rapid development of rule-based information extraction applications," *Natural Language Engineering*, vol. 22, no. 1, pp. 1–40, 2016.

[6] B. Rink, S. Harabagiu, K. Roberts et al., "Automatic extraction of relations between medical concepts in clinical texts," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 594–600, 2011.

[7] M. Jiang, Y. Huang, J.-w. Fan, B. Tang, J. Denny, and H. Xu, "Parsing clinical text: how good are the state-of-the-art parsers?" *BMC Medical Informatics and Decision Making*, vol. 15, no. S1, 2015.

[8] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 16, 2010.

[9] M. Jiang, Y. Chen, M. Liu et al., "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.

[10] O. Uzuner, J. Mailoa, R. Ryan, and T. Sibanda, "Semantic relations for problem-oriented medical records," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 63–73, 2010.

[11] X. Zhu, C. Cherry, S. Kiritchenko, J. Martin, and B. de Bruijn, "Detecting concept relations in clinical text: insights from a state-of-the-art model," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 275–285, 2013.

[12] T. Mikolov, K. Chen, G. Corrado et al., "Efficient estimation of word representations in vector space," 2013, http://arxiv.org/abs/1301.3781.

[13] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis*, Cambridge, London, UK, 1966.

[14] D. Demner-Fushman, E. Apostolova, R. Islamaj Dogan et al., ""NLM's system description for the fourth i2b2/VA challenge,"" in *Proceedings of the 2010 I2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, i2b2, Boston, MA, USA, 2010.

[15] Y. H. Tseng, L. H. Lee, S. Y. Lin et al., "Chinese open relation extraction for knowledge acquisition," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, pp. 12–16, Gothenburg, Sweden, April 2014.

[16] M. Hassan, O. Makkaoui, A. Coulet et al., *Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs*, 2015.

[17] X. Y. Bao, W. J. Huang, K. Zhang et al., "A customized method for information extraction from unstructured text data in the electronic medical records," *Journal of Peking University. Health sciences*, vol. 50, no. 2, pp. 256–263, 2018.

[18] X. Zhou, B. Liu, Z. Wu, and Y. Feng, "Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks," *Artificial Intelligence in Medicine*, vol. 41, no. 2, pp. 87–104, 2007.

[19] R. J. Ryan, "Groundtruth budgeting: a novel approach to semi-supervised relation extraction in medical language,"Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2011.

[20] http://www.gov.cn/zwgk/2010-03/04/content_1547432.htm.

[21] B. Hu and X. Liao, *Modern Chinese*, Higher Education Press, Beijing, China, 2002.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] B. Zhang, M. Lu, and Y. Fang, "A feature-enhanced entity recognition method for Chinese electronic medical records," in *Proceedings of the 9th International Conference on Information Technology in Medicine and Education (ITME)*, IEEE, Hangzhou, China, pp. 9–14, October 2018.

[24] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training support vector machines," *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.

[25] Ö Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514–518, 2010.

Hindawi
Scientific Programming
Volume 2020, Article ID 6929750, 13 pages
https://doi.org/10.1155/2020/6929750

Research Article

A Microservice-Based Big Data Analysis Platform for Online Educational Applications

Kehua Miao, Jie Li, Wenxing Hong, and Mingtao Chen

Department of Automation Xiamen University, Xiamen 361005, Fujian, China

Correspondence should be addressed to Wenxing Hong; hwx@xmu.edu.cn

Received 9 February 2020; Accepted 19 May 2020; Published 3 June 2020

Academic Editor: Chenxi Huang

The booming development of data science and big data technology stacks has inspired continuous iterative updates of data science research or working methods. At present, the granularity of the labor division between data science and big data is more refined. Traditional work methods, from work infrastructure environment construction to data modelling and analysis of working methods, will greatly delay work and research efficiency. In this paper, we focus on the purpose of the current friendly collaboration of the data science team to build data science and big data analysis application platform based on microservices architecture for education or nonprofessional research field. In the environment based on microservices that facilitates updating the components of each component, the platform has a personal code experiment environment that integrates JupyterHub based on Spark and HDFS for multiuser use and a visualized modelling tools which follow the modular design of data science engineering based on Greenplum in-database analysis. The entire web service system is developed based on spring boot.

1. Introduction

In recent years, data science and big data technology stacks have achieved explosive growth. In data science, especially machine learning, it mainly benefits from the improvement of computing power, especially the rapid development of GPU (Graphics Processing Unit), and the popularity of deep learning [1]. In terms of big data technology stacks, new types of big data tools such as Spark have gradually replaced some components in the traditional Hadoop ecosystem [2], and the update iteration speed is persistent. Based on the rapid development of technology, the modularization and subdivision of work methods become more and more obvious. Traditional working methods, from setting up an experimental environment to data acquisition, data processing, modelling training, and data prediction, are often performed in a unified manner [3]. And this way of working is obviously not suitable for new data science research methods. At present, people pay more attention to teamwork, so infrastructure environment sharing, data sharing, and model sharing have become mainstream workflows. Of course, a personal workspace based on sharing conditions is also essential.

However, the development of big data [4] and the deployment of operating environments often require development, operation, and maintenance personnel with professional knowledge to build and maintain [5], which is more difficult for ordinary teams, especially novice or student team. People often waste a lot of time on infrastructure and environmental construction, and these troubles are often accompanied by huge operation and maintenance problems. People need a multiuser based infrastructure environment platform [6].

For data analysis of actual business scenarios, people's thinking is focused on not only code but also text, pictures, and even mathematical formulas. People need a working code area for a markdown-like environment [7]. The Jupyter Notebook provides such a working environment, but the Juypter Notebook is aimed at single-user members. For web-based platforms [8], JupyterHub is required [9]. For workers with big data needs, they also need to integrate the experimental environment based on Hadoop [10], Spark, and JupyterHub.

In data sharing and data stream processing, the security protection of data is often the most troublesome problem in the analysis of actual business scenarios. People often do not want to perform compatibility processing and desensitization of various data migrations. In team collaboration, the platform needs to be integrated with tools that can analyze data within the database. For team collaboration, data flow processing and model sharing are often accompanied. These abstract contents are often not suitable for collaborative analysis in a short time. People need a visualized modelling environment.

In this paper, we present Qunxian Platform. Qunxian Platform is a big data analysis platform based on microservices. It helps to realize the sharing of data, computing power, and infrastructure resource. Moreover, it enables users to use a more friendly environment to record and share the experimental process and the visualized model. It uses JupyterHub and visualized modelling as its two main applications. The platform's infrastructure environment components are the big data component and the data science environment. In order to easily adapt to the characteristics of rapid environmental changes, the platform uses microservices [11] as a technical dependency; in particular, it uses Docker [5, 12]. The environment's distributed file system uses the HDFS (Hadoop Distributed File System) [13]. The parallel computing architecture uses Apache Spark. Based on Spark and HDFS, Jupyter Notebook is used as the user's code engineering environment. And JupyterHub manages a multiuser notebook environment in a unified manner. In terms of visualized modelling, the Greenplum-based in-database analysis method is used to implement the modelling module of the visualized modelling application with the built-in MADlib algorithm and custom algorithm interface [14]. Because the entire platform system will be oriented to different users and different user-environments, users will share computing and data resources of the server. This system locates B/S (Browser/Server) type systems [15]. Based on the B/S type system, the platform is built using a framework based on spring boot and uses element-admin as a back-end front-end solution [16]. Moreover, it uses Greenplum and MADlib to implement in-database data calculation to protect the data, quickly use the data, and store the model. Based on Element-UI [17], the drag-and-drop method of the back-end module is realized [18], and the visualized data modelling is further realized.

The remaining part of this paper is arranged in the following format. Section 2 presents related works in this field. In Section 3, we visit the platform's architecture without the app layer and support layer. Then, we describe the two web-based applications in Section 4. After that, we give our experiment configuration on Google cloud from Section 5. And finally, we conclude with Section 6.

2. Related Works

As described in the introduction, there is a certain demand to get advanced tools for storing and sharing some kinds of environment for data science and big data research, which is configured by specific engineering staff on cloud computing infrastructure for individuals and research teams. A lot of systems have been developed in recent years and solved some of the problems mentioned above, including big data environment infrastructure constructing [4, 10, 19], microservice-based platform [5, 11, 20], and online educational programming platform [21–23]. However, several issues remain, which are not addressed well by existing approaches and systems.

The first issue is supporting a platform which is based on data science computing and big data ecosystem environment as online services. The use of the service-oriented approach can enable wide-scale sharing and deployment of the new data model as well as automation of scientific tasks and composition of applications into new services [24]. However, the existing web-based toolkits or platforms either do not construct a stable big data ecosystem environment or provide the raw and cumbersome command shell, which is unfriendly for students to getting started with.

The second issue is teamwork corporation based on big data environments platform. The existing online programming platforms always are designed for private use [21, 24, 25]. However, the existing platform usually lacks personal volume storage for a long time as a personal application. Moreover, the personal application does not agree with teamwork nowadays. In this way, to share personal ideas in the teamwork, especially for students in an educational way [21], we need to share the visualized model without the raw code only.

Qunxian addresses the described issues by relying on microservice. On the infrastructure building, we implement the docker-compose service. On the web service building, we implement the spring boot.

Spring framework offers flexibility to all the configure beans in many ways such as annotations and XML [26]. With the number of features being increased, the complexity also gets increased and configuring spring applications becomes error-prone and tedious [27]. Spring boot helps to address the complexity of configuration. Also, spring boot is much more helpful to integrate the microservice applications [28].

To help to build the concept of bringing the science to data, on the web front-end, users will try to manipulate the data in a visualized way. To make it happen, we implement the vue.js to help the front-end communicate with the back-end. Technically, vue.js is a progressive framework that helps developers to build a user interface for website development. It is designed to be applied layers by layers from the bottom-up [17]. In particular, it focuses more on the view layers. These features help to integrate many modern toolchains and various supporting class libraries. Also, it helps to reduce the difficulty of web development. To build a production-ready front-end solution [29], we integrate the element-admin, which is based on Vue and uses the UI Toolkit Element-UI. Vue-Element-Admin is a back-end front-end solution based on Vue and Element-UI. It implements the latest front-end technology stack, built-in i18 international solutions [30], dynamic routing, permission verification, and refined typical business models and provides rich functional components, which can help web

developers quickly build enterprise-level product prototypes. All the tools and frameworks help to build a user-friendly web-based platform.

3. The Platform's Architecture

Qunxian follows the Platform as a Service (PaaS) and Figure 1 gives a high-level overview of the Qunxian platform architecture, which is described in the sequel.

3.1. Hardware Layer

3.1.1. Cloud-Based Big Data Platform. Cloud computing is the provision of computing services providing rapid innovation, elastic resources, and economies of scale through the cloud. For cloud services, web developers usually pay only how much they use, which helps reduce operating costs, enables the infrastructure to operate more efficiently, and adjusts the use of services based on changes in business needs [31]. Its simplicity is the type of computing that depends on shared computing resources instead of local servers. Qunxian Platform decided to take the cloud computing service as the hardware layer and it would get the below-mentioned benefits:

(i) It greatly reduces IT and labor costs

(ii) It is more scalable and offers better and secure storage

(iii) Collaboration and effective communication platforms are provided

(iv) Best work practices and flexibility are received

(v) Access to automatic updates for your IT requirements is included

Nowadays, for cloud computing services, Google Cloud Platform (GCP), Amazon Web Services (AWS), and Microsoft Azure hold a ruling position among the cloud companies. Qunxian chose the Google Cloud Engine, the module of GCP, as the hardware service.

3.1.2. Docker. Based on cloud service, we decided to use Docker as a deployment methodology to manage the hardware layer efficiently. All individual modules (i.e., big data components, such as Hadoop and Spark, web architecture, and JupyterHub service concerned) rely on different operating environments, such as different Java versions. To coordinate them smoothly, without any compatibility problem, we apply the Docker container to separate them into individuals. Also, as the foundation of microservice nowadays, Docker helps to simplify the operation and maintenance work, which is the main work after the first time to deployment because of the frequent updating. By the benefit of the Docker container, it is an easy case to update the individuals that needed to be updated instead of all the system architecture. Also, the Dockerfile and YAML files help to record the deployment process for the reference of next time deployment.

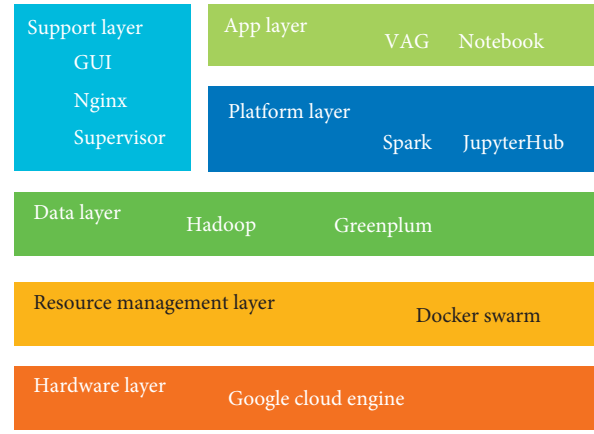


Figure 1: High-level Qunxian platform architecture.

3.2. Resource Management Layer. With the concern of the development stage and the number of users, we do not apply Kubernetes, which is the most popular Docker orchestration tool, to deploy Docker containers on a cluster. We apply Docker Swarm, which is the default orchestration tool of Docker, to construct our Resource Management Layer. On top of Docker Swarm, we implement Docker Compose as a collection of communicating containers that can be scaled and scheduled dynamically [32].

3.3. Data Layer. In the data layer, we deploy the HDFS, the Hadoop distributed file system, and Greenplum, a massively parallel Postgres for analytics relational database [19]. HDFS and Greenplum work with Spark, set at the platform layer, independently.

3.4. Platform Layer. In the platform layer, we deploy the JupyterHub, which is a multiuser server for Jupyter notebooks, as the back-end with containerized big data environment based on Docker.

Table 1 gives Docker components of the platform, which build up the basis big data ecosystem environment of Qunxian.

4. Two Web-Based Applications

Based on the platform layer, we introduce the app layer, which consists of two web-based applications. We describe them as follows.

4.1. Online Programming Application. For students to get the most of Qunxian's high-performance computing (HPC) resource, Qunxian applied JupyterHub, the platform layer's component, the data science, and big data ecosystem environment as a web-based application with the help of Docker.

Based on the platform layer, a default way of working with Apache Spark is to launch a cumbersome command shell from the terminal, which makes it very hard to present information. To go beyond that and make data analysis more shareable and reusable, we choose Jupyter Notebook.

Table 1: Big data Docker component.

Docker name	Base images	Features
Hadoop-Base	Debian	Hadoop basis with OPENJDK
Datanode	Hadoop-Base	HDFS slave server
Namenode	Hadoop-Base	HDFS master server
Spark-Base	python	Monitor the containers and Spark Base
Spark-Master	Spark-Base	Put Spark under supervision
Spark-Worker	Spark-Base	Expose ports for each application

Jupyter Notebook, which spawns by JupyterHub, provides an easy way to use interactive programmatic interface which is accessible from the browser that alleviates the burden for students to implement their Scala or python job via cumbersome command-line methods.

Figure 2 gives a high-level overview of online programming applications on the app layer based on JupyterHub and Docker. It allows the Authenticator to be a GitHub user, which helps the team to cooperate with their work based on the project. Docker Spawner spawns single-user Jupyter Notebook servers in separate Docker containers based on their separate Docker volume, which helps to persist every user's notebook directories. Also, for the persistence of JupyterHub data, we deploy a single Docker volume on the host.

4.2. Visualized Modelling Application. Visualized modelling application is a data science modelling tool, which helps students to understand the high-level machine learning pipeline designed by teachers or other researchers. We apply the classic machine learning module, which includes data processing, feature extraction, and model selection. Students can explore their data on this tool without building their data research way from scratch. Also, for research way, we developed a new module for students and researchers to explore their new idea on model building.

Figure 3 gives a pipeline case on the web-based visualized modelling application. It is a simple case for students to understand and repeat the experiment on each part. From model constructing to model deployment, every single part in data science can be selected from the given choice or build new components based on the pipeline.

5. Experiments

We performed an experimental evaluation of Qunxian with all server-side platform components being deployed on Google Cloud Engine. The boot disk is based on CentOS 7 OS image with the standard persistent disk of 2TB. And the machine type is n1-standard-2 with 64vCPU, 240 GB memory. We add tags and firewall rules to allow specific network traffic from the Internet, including MySQL service [33], Greenplum service, web service, and Hadoop service on the VPC (Virtual Private Cloud) network/Firewall rules in Google cloud platform. A minimal tuning of the system and the Nginx web proxy server [10] was done to support a high number of concurrent connections. In the following part, we will first describe the service deployment and then give two examples of the main applications on the platform.

5.1. JupyterHub Service. The JupyterHub module, on the app layer, is based on the data layer. To integrate with the HDFS and Spark, we do this part of the experiment by docker-compose. And JupyterHub-docker-compose-example.yml file shows part of the configuration. In this paper, we do not show the details. To make the data sustainable of JupyterHub, we choose the PostgreSQL as a database. In the yml file, we leave out some detailed information, such as the environment variable, volumes, and networks' configuration. (Algorithm 1).

5.2. Database Service

5.2.1. MySQL Service. For the web service, we use MySQL service to help build safety information storage, especially the data of users. We pull the MySQL:5.6 image from the DockerHub to build the service. After giving birth to a new container from the image-registry, we make a SQL script to build the ZDSW database, which includes system, applications, users, and other entities' information. In Figure 4, we show part of the physical data model from the database, which is one of SQL script execution results.

5.2.2. Greenplum Service. To put the in-database data processing in force, we ask for the help of Greenplum, which is able to integrate with the MADlib extension without any compatible problem. And MADlib is an open-source library for in-database data processing. Moreover, Greenplum is a relational data warehouse, which follows massively parallel processor architecture. It helps us to make full use of a big data environment. Due to its PostgreSQL kernel, we take advantage of the development platform tool, pgAdmin4, to integrate the database in the front-end on a specific web port. In this experiment, we try to make it easy to repeat. So, we apply the pseudo distribution in a single node with the help of Docker and docker-compose. We prepare a gpdb-docker-compose-example.yml file, which is shown below, for configuration reference. To make it successful, a startGPDB.sh file, which helps Docker to start the database, the base image, and other configuration should be prepared. (Algorithm 2).

The following part shows an example of using pgadmin4 to access Greenplum:

(1) Prerequisites

(2) Starting Docker-compose

(3) Configure Greenplum

(4) Configure pgadmin4

At the prerequisites part, we install the docker-compose and then pull the images of pgadmin4 and GPDB 5.x OSS. Prepare a script, such as docker-compose-pgadmin4.yml,

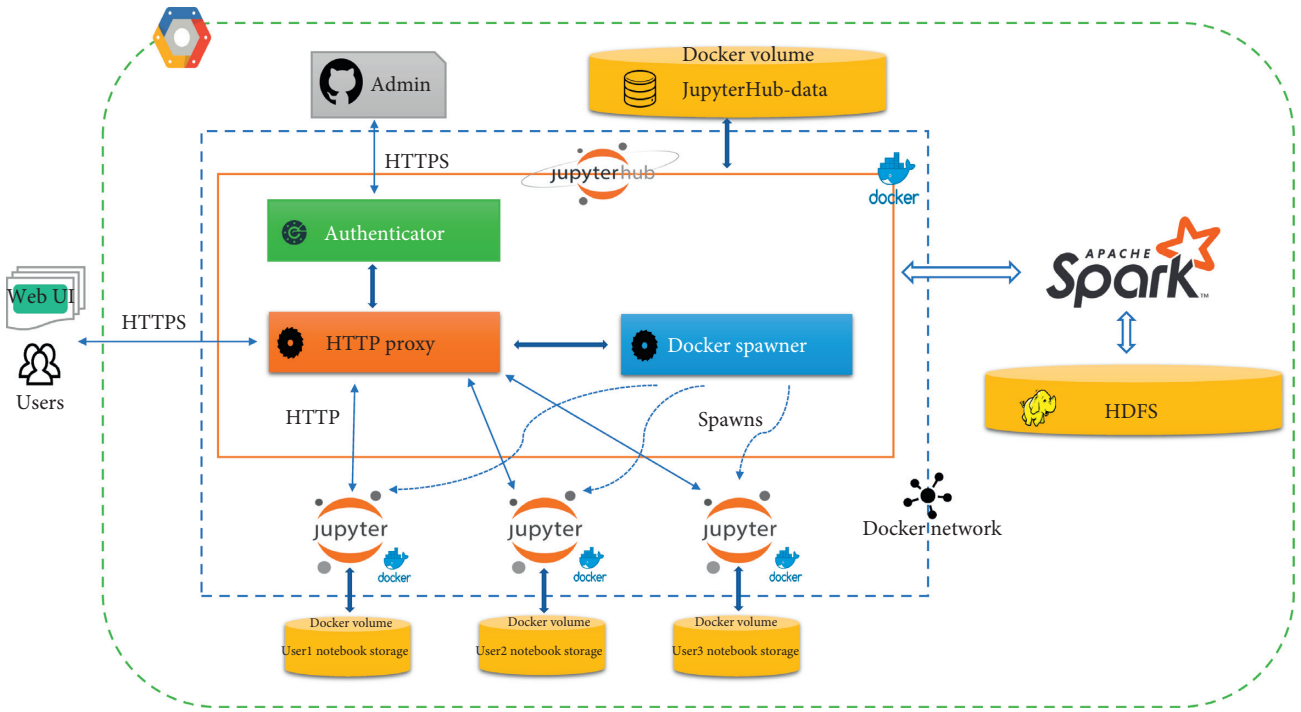


Figure 2: Overview of platform layer architecture.

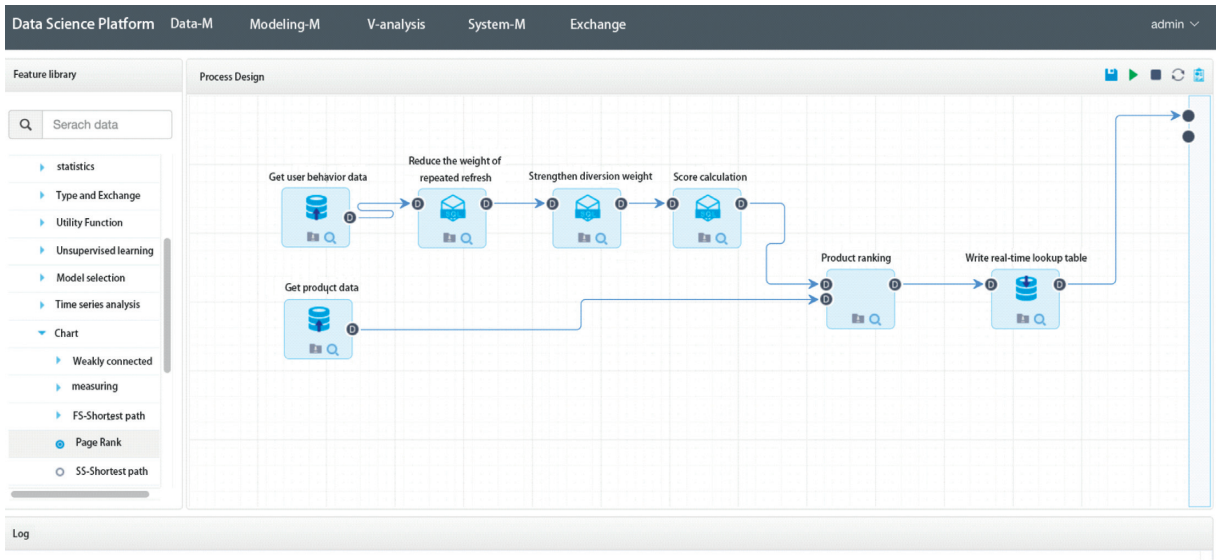


Figure 3: A pipeline case on the web-based visualized modelling application.

which configures the Greenplum and pgadmin4 well, to integrate them successfully.

After the database preparation, as the MySQL service, we need to execute a SQL script. The script is to prepare a data set to test the in-database machine learning.

Figure 5 shows the result of script result on the web-based pgadmin4 client dashboard.

5.3. Web Service. In the web-based project part, we apply spring boot [34] and webpack frameworks [35] to build the platform. The flowing part describes the details of the front and back-end service applications.

5.3.1. Spring Boot Back-End. On the back-end of web service, we build a Java project based on spring boot. The project includes three parts: data-control, resource, and web-app. The data-control helps the visualized Analysis-tools workflow going. The resource connects the back-end and front-end. Web-app is an essential part of the webpack project.

```
version: '3'
services:
  hub-db:
    image: postgres
    container_name: Jupyterhub-db
    restart: always
    env_file:
      -Jupyterhub/secrets/postgres.env
    volumes:
      –"db : ${DB_VOLUME_CONTAINER}"
  hub:
    depends_on:
      –hub-db
    build:
      context:./Jupyterhub
      dockerfile: Dockerfile
      args:
      Jupyterhub_VERSION: ${Jupyterhub_VERSION}
    restart: always
    image: Jupyterhub
    container_name: Jupyterhub
    ports:
      –"443:443"
      –"7070:7070"
    links:
      -hub-db
    env_file:
      -Jupyterhub/secrets/postgres.env
      -Jupyterhub/secrets/oauth.env
  command: >
    Jupyterhub-f/srv/Jupyterhub/Jupyterhub_config.py
  spark-master:
  build:
      context: /spark-master
      dockerfile: Dockerfile
  image: spark-master
  container_name: spark-master
  hostname: spark-master
  ports:
      –"8585 : 8080"
      –"7077 : 7077"
  volumes:
      -/mnt/spark-apps : /opt/spark-apps
      -/mnt/spark-data : /opt/spark-data
  spark-worker-1:
  spark-worker-2:
  spark-worker-3:
  namenode:
    image: hadoop-namenode : 2.0.0-hadoop3.1.1-java8
    container_name: namenode
  volumes:
      -"namenode:/hadoop/dfs/name"
  env_file:
      -/hadoop.env
  datanode:
    image: hadoop-datanode : 2.0.0-hadoop3.1.1-java8
```

Algorithm 1: JupyterHub-docker-compose-example.yml.

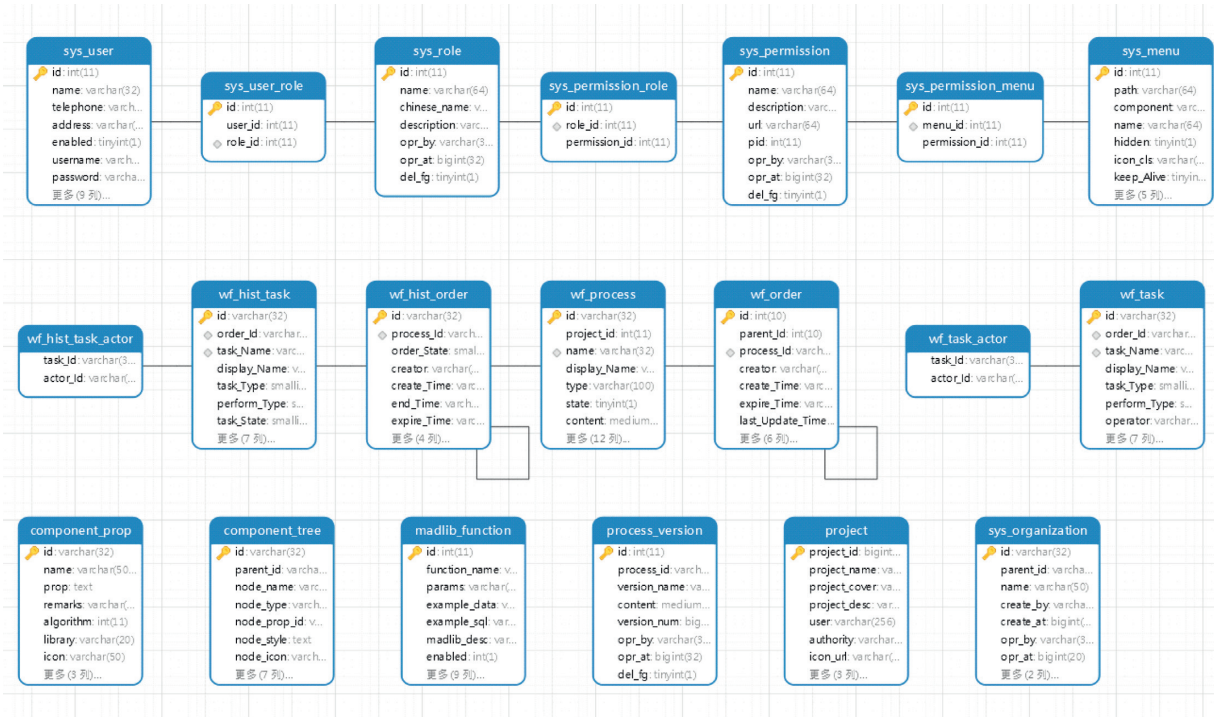


Figure 4: Physical data model (partial).

```
version: '3'
services:
  pgadmin4:
    image: pgadmin4
hostname: "pgadmin4"
    environment:
      -PGADMIN_DEFAULT_e-Mail = xxx@domain.com
      -PGADMIN_DEFAULT_PASSWORD = SecretPassword
    ports:
      -"80 : 80"
    volumes:
      -.. : /code
    tty: true
    privileged: true
  networks:
    mynetwork:
    aliases:
      -pgadmin4
  gpdb5:
    image: "gpdb5oss"
    command: bin/bash-c "startGPDB.sh && bin/bash"
    hostname: "gpdbsne"
    container_name: gpdbsne
    ports:
      -"5432 : 5432"
      -"5005 : 5005"
      -"5010 : 5010"
    expose:
      -"5432"
    volumes:
```

Algorithm 2: Continued.

```
      -../code
    privileged: true
    tty: true
    networks:
    mynetwork:
    aliases:
      -gpdbsne
      -gpdbsne.localdomain.com
networks:
    mynetwork:
    driver: bridge
```

Algorithm 2: gpdb-docker-compose-example.yml.

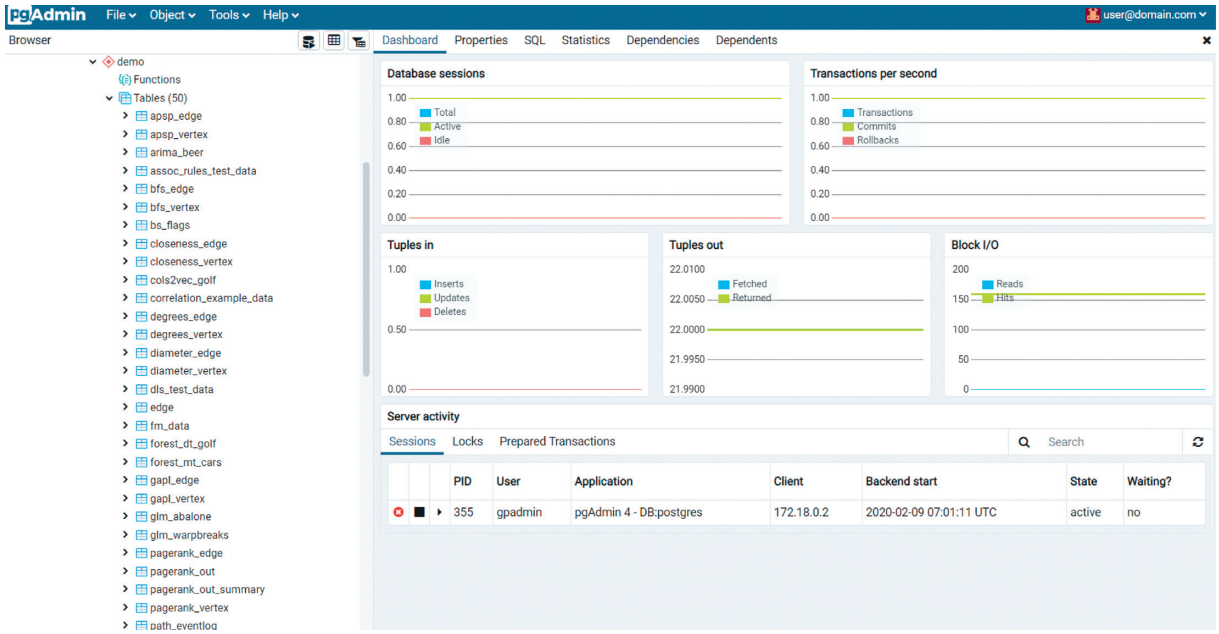


Figure 5: The web-based pgadmin4 client dashboard.

5.3.2. Element-UI-Admin Front-End. To reduce the time of web page development, we apply Element-UI to build an HTML, CSS, and JavaScript project for developing responsive [36], projects on the web. On the foundational of the UI-tool kit, we implement the open-source, Element-UI-Admin, to manage the front-web way of data management.

5.3.3. Deployment. After building the project and configuring the application properties, which is the configuration of database service and the features of low coupling and high cohesion, we deploy it with a special port on the Google Cloud Server Engine.

5.4. Online Programming Experiment on Spark and Hadoop. Since our educational application is based on Spark, Hadoop, and JupyterHub, we will do our experiment on Jupyter Notebook, spawned by JupyterHub on the website interface produced by Qunxian.

Figure 6 shows the SparkContext job and Hadoop job on the Jupyter Notebook, which is exposed by web port. And we can find the Spark job more directly on the port by HTTP way. Figure 7 shows the jobs works.

5.5. Visualized Modelling Application Experiment Example. Since the visualized modelling application is based on Greenplum database and the environment-friendly experiment experience with pure python.

In the experiment, we need to transform the algorithms to the format of pl/python [37], which is shown in Figure 8. To make it successful, the data running environment needs to be initialized to ensure that the environment dependencies required for the experiment are already available on the web platform. If there are missing experimental related environments, the server needs to install the relevant dependencies first. And the web back-end management system should grants experimental users environment

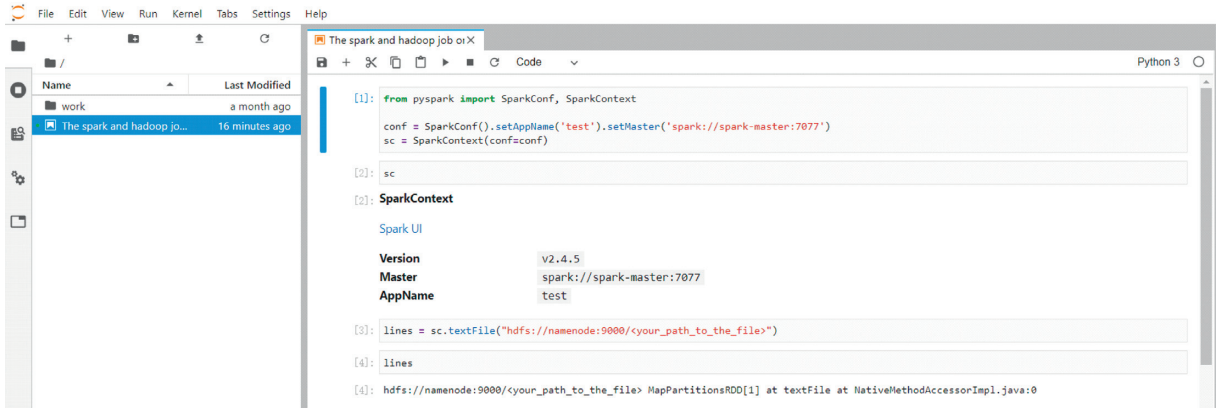


Figure 6: The Spark and Hadoop job on Jupyternotebook.

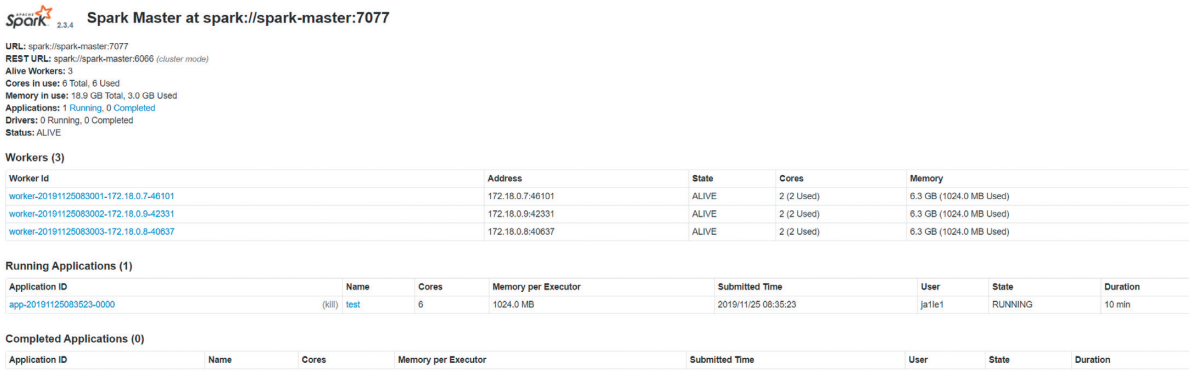


Figure 7: The Spark job on website back-end.

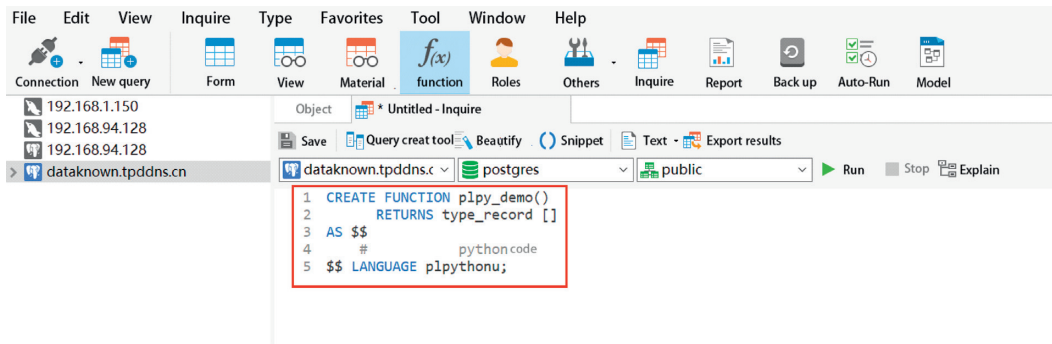


Figure 8: The transformation of algorithms.

administrator rights, because only super administrator users can use it to register with the database using PL/python.

The visualized modelling application supports integrating the configured pl/python algorithms with Qunxian. In the following part, we introduce a deployment way of new functions, shown by the HDP algorithm in the visualized modelling application.

First, we should add the HDP algorithm in the management of algorithms, which is part of data management. Second, we should apply it into the tree of algorithm to become a component of the existing algorithms. Last, we can make full use of it when building the visualized model. The processing part is shown by sequence in Figures 9–11.

5.6. Web Service Stress Test. In order to adapt the microservice characteristics of the adaptive system and the convenience of testing, we still use Docker-based methods to deploy JMeter-based [38] test cases. Select Alpine as the base of the Docker image and configure related environment variables, such as JMETER VERSION and JMETER HOME, which is convenient for decoupling test units and operation and maintenance units. Because JMeter is developed based

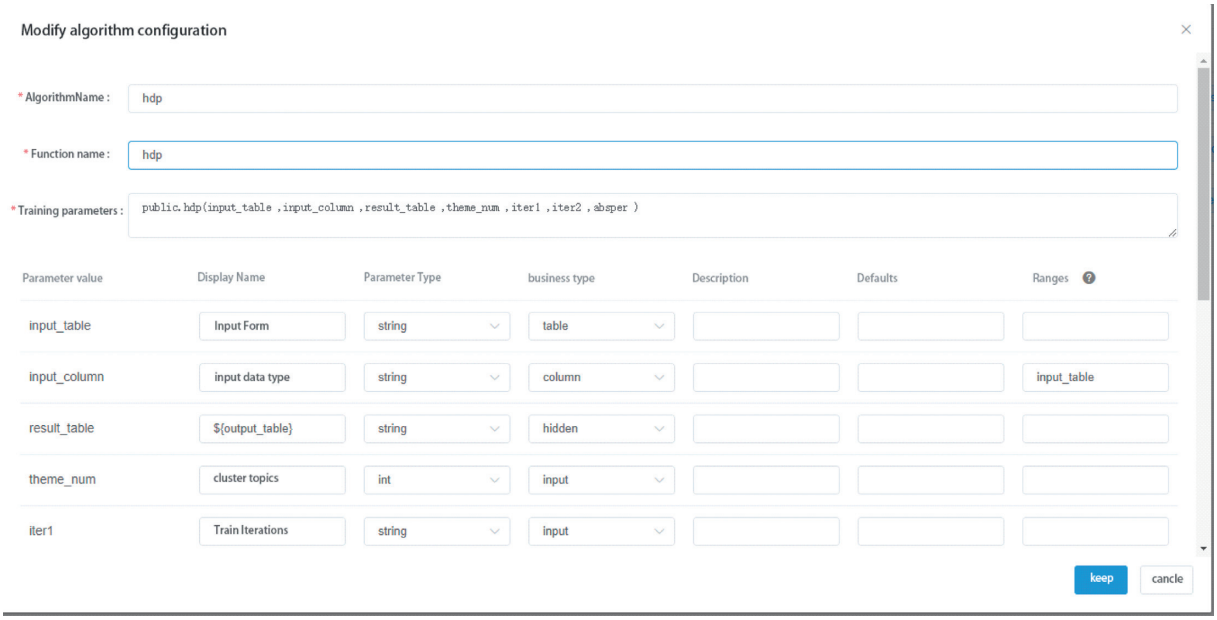


Figure 9: Step1 of processing of the HDP algorithm.

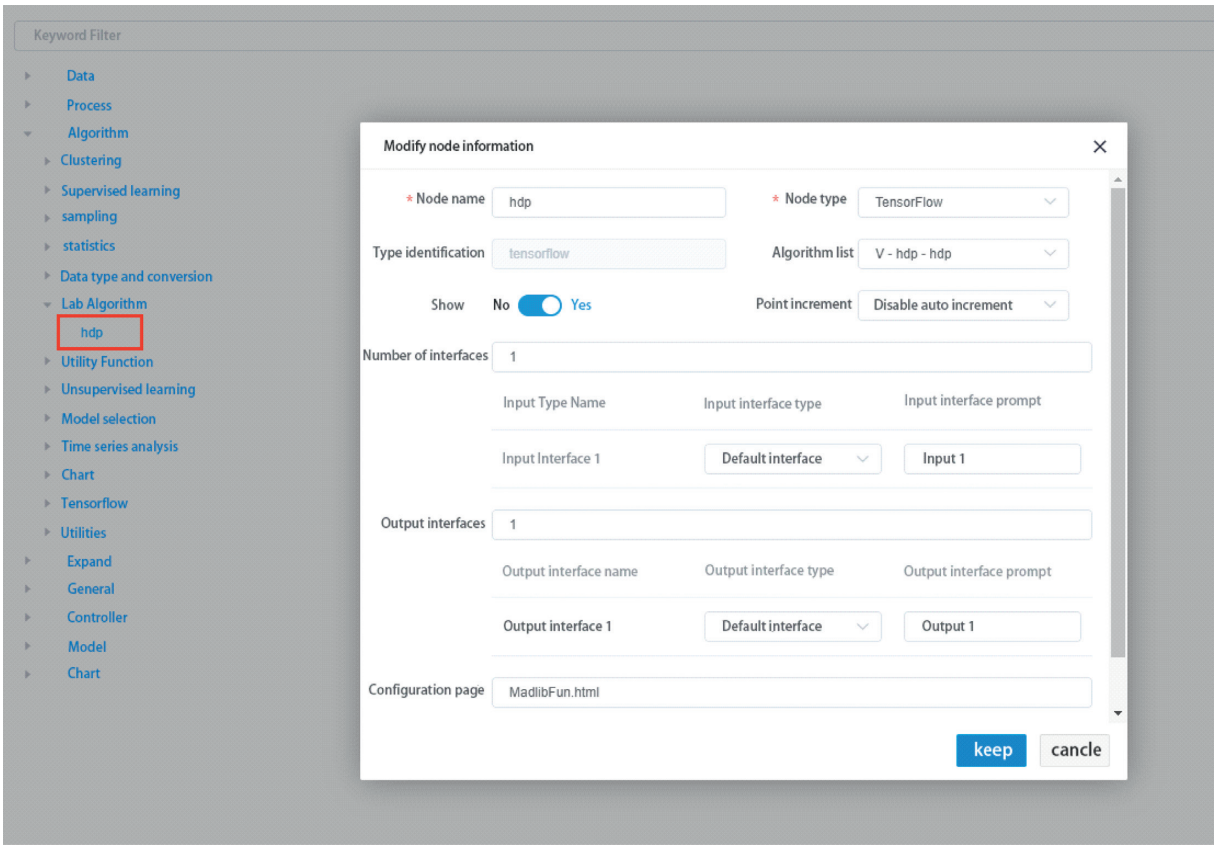


Figure 10: Step2 of processing of the HDP algorithm.

on the Java language, the service environment also needs to be configured with the Java SE Runtime Environment (JRE) [39]. In specific test cases, two scripts can be prepared: one is a test script, and the other is a run script. Test scripts are used to test the writing of rules for specific use cases, such as the hard coding of the system web port. The run script mainly

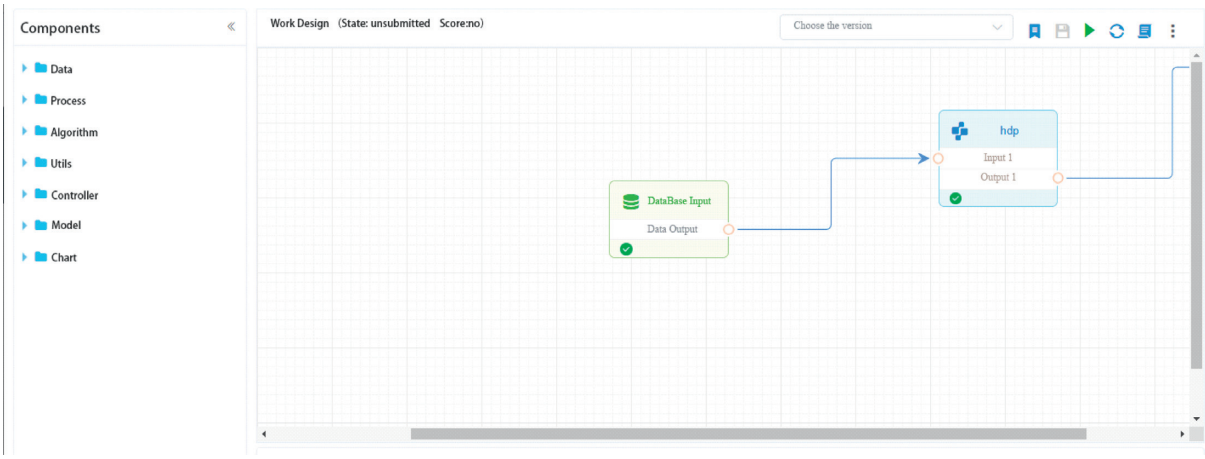
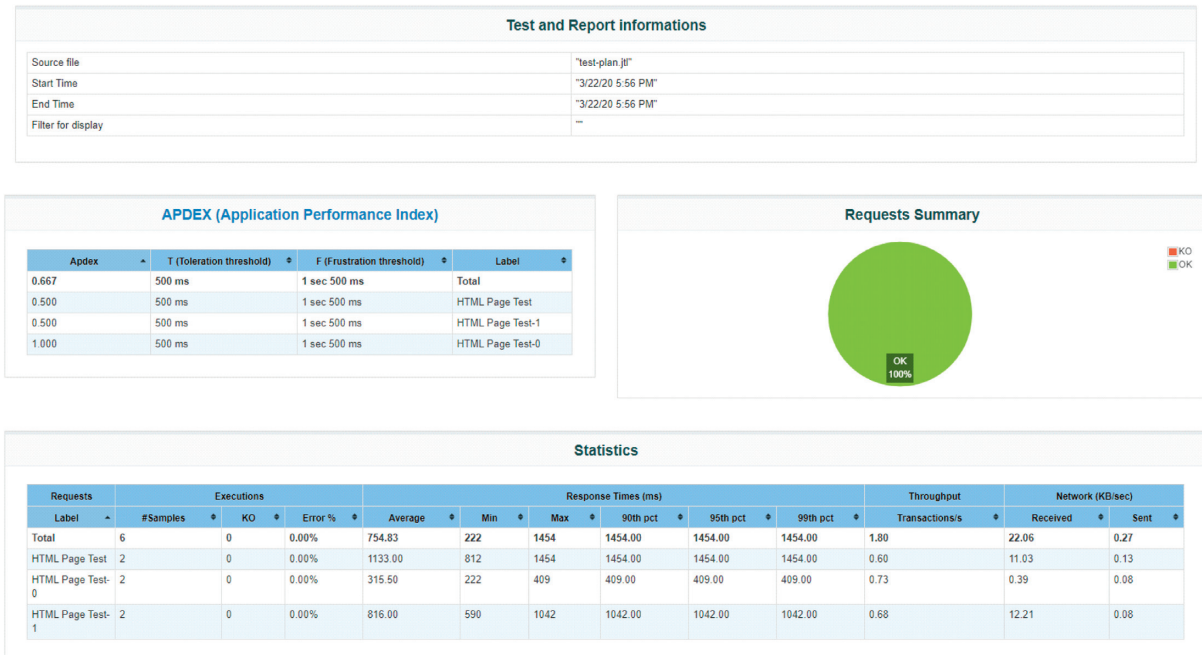


Figure 11: Step3 of processing of the HDP algorithm.



Test and Report informations

Source file	"test-plan.jtl"
Start Time	"3/22/20 5:56 PM"
End Time	"3/22/20 5:56 PM"
Filter for display	""

APDEX (Application Performance Index)

Apdex	T (Toleration threshold)	F (Frustration threshold)	Label
0.667	500 ms	1 sec 500 ms	Total
0.500	500 ms	1 sec 500 ms	HTML Page Test
0.500	500 ms	1 sec 500 ms	HTML Page Test-1
1.000	500 ms	1 sec 500 ms	HTML Page Test-0

Requests Summary

Statistics

Requests	Executions			Response Times (ms)						Throughput	Network (KB/sec)	
Label	#Samples	KO	Error %	Average	Min	Max	90th pct	95th pct	99th pct	Transactions/s	Received	Sent
Total	6	0	0.00%	754.83	222	1454	1454.00	1454.00	1454.00	1.80	22.06	0.27
HTML Page Test	2	0	0.00%	1133.00	812	1454	1454.00	1454.00	1454.00	0.60	11.03	0.13
HTML Page Test-0	2	0	0.00%	315.50	222	409	409.00	409.00	409.00	0.73	0.39	0.08
HTML Page Test-1	2	0	0.00%	816.00	590	1042	1042.00	1042.00	1042.00	0.68	12.21	0.08

Figure 12: The result of web service stress test.

writes common run routines such as the startup and destruction of microservices.

Figure 12 shows the result of the web service stress test with the help of JMeter. In requests summary, all the test requests passed as required. In the statistics form, all the response times are within the controllable range.

6. Conclusion

We have presented Qunxian, a new microservice-based big data analysis platform, which is deployed on Google Cloud Engine running across distributed computing resources. And we construct the big data ecosystem environment based on Docker. On the foundation of platform infrastructure, we introduce two web-based applications. For students who want to get started with a big data ecosystem environment without any barriers on educational purposes, we introduced JupyterHub, with which every user can program online on their environment and keep their data persistence on private data volume. For researchers who want to corporate efficiently while sharing their models based on their specific format of data, we introduce the other web-based application, Visualized Modelling tool, in which data science research people can share their ideas without the cold data and code only. We do our experiment on Google Cloud Platform and check its feasibility.

For future work, we will deploy the platform on Kubernetes clusters, which is a portable, extensible, open-source platform for managing containerized workloads and services, if the platform serves more students. In this way, the

automating deployment and scaling of the two containerized applications may help infrastructure get well utilized.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partly supported by the National Key Research and Development Program of China (Grant no. 2018YFC0830300), Science and Technology Program of Fujian, China (Grant no. 2018H0035), Science and Technology Program of Xiamen, China (Grant no.3502Z20183011), and Fund of XMU-ZhangShu Fintech Joint Lab.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016.

[2] S. Chen, C. Wu, and Y. Yu, "Analysis of plant breeding on hadoop and spark," *Advances in Agriculture*, vol. 2016, Article ID 7081491, 6 pages, 2016.

[3] T. M. Mitchell, *Machine Learning*, 1997.

[4] J. M. Cavanillas, E. Curry, and W. Wahlster, *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, Springer, Berlin, Germany, 2016.

[5] J. Kobusiński, J. Brzeziński, and A. Kobusińska, "On increasing dependability of web services-an approach to design a failure detection service 2018," in *Proceedings of the 16th IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC 2018)*, IEEE, Athens, Greece, pp. 504–511, August 2018.

[6] M. D. Scott, M. A. Foltz, J. Affaki et al., "System for universal remote media control in a multi-user, multi-platform, multi-device environment," U.S. Patent 10-031-647, 2018.

[7] J. M. Perkel, "Why Jupyter is data scientists' computational notebook of choice," *Nature*, vol. 563, no. 7732, pp. 145–147, 2018.

[8] J. Lu, G. Liu, K. Wu et al., "Location-aware web service composition based on the mixture rank of web services and web service requests," *Complexity*, vol. 2019, Article ID 9871971, 16 pages, 2019.

[9] M. Milligan, "Interactive hpc gateways with jupyter and jupyterhub," in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, pp. 1–4, New Orleans, LA, USA, July 2017.

[10] Y. Zhang, M. Wei, C. Cheng et al., "Exploiting delay-aware load balance for scalable 802.11 PSM in crowd event environments," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 6128437, 11 pages, 2017.

[11] C. Pahl and P. Jamshidi, "Microservices: a systematic mapping study," in *Proceedings of the 2016 IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA)*, pp. 137–146, Macau, China, November 2016.

[12] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux Journal*, vol. 2014, no. 239, p. 2, 2014.

[13] K. Shvachko, H. Kuang, S. Radia et al., "The hadoop distributed file system," in *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, IEEE, Incline Village, NV, USA, pp. 1–10, May 2010.

[14] J. Hellerstein, C. Ré, F. Schoppmann et al., "The MADlib analytics library or MAD skills, the SQL," 2012, http://arxiv.org/abs/1208.4165.

[15] S Butlin, "Method of browser-server communication," U.S. Patent 10-151-11, 2002.

[16] K. Artto, T. Ahola, and V. Vartiainen, "From the front end of projects to the back end of operations: managing projects for value creation throughout the system lifecycle," *International Journal of Project Management*, vol. 34, no. 2, pp. 258–270, 2016.

[17] E. You, *Vue. js*, Diakses Dari Httpsvuejs Org Pada Tanggal, Washington, DC, USA, 2018.

[18] H. Zhang, T. H. Lee, K. Chow et al., "Drag and drop interaction between components of a web application," U.S. Patent 10-048-854, 2018.

[19] M. Ferguson, *Architecting a Big Data Platform for Analytics*, p. 30, A Whitepaper prepared for IBM, Armonk, NY, USA, 2012.

[20] D. Guo, W. Wang, G. Zeng et al., "Microservices architecture based cloudware deployment platform for service computing," in *Proceedings of the 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, IEEE, Oxford, UK, pp. 358–363, March 2016.

[21] S. Kim, J. W. Kim, J. Park et al., "Elice: an online CS education platform to understand how students learn programming," in *Proceedings of the Third (2016) ACM Conference on Learning@Scale*, pp. 225–228, Edinburgh, UK, April 2016.

[22] T. Staubitz, H. Klement, R. Teusner et al., "CodeOcean-A versatile platform for practical programming excercises in online environments," in *Proceedings of the 2016 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, Abu Dhabi, UAE, pp. 314–323, April 2016.

[23] Z. Zou, Y. Zhang, J. Li et al., "EasyHPC: an online programming platform for learning high performance computing," in *Proceedings of the 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, IEEE, Hong Kong, China, pp. 432–435, December 2017.

[24] I. Foster, "Service-oriented science," *Science*, vol. 308, no. 5723, pp. 814–817, 2005.

[25] J. Schwieren, G. Vossen, and P. Westerkamp, "Using software testing techniques for efficient handling of programming exercises in an e-learning platform," *Electronic Journal of E-Learning*, vol. 4, no. 1, pp. 87–94, 2006.

[26] M. Keith, M. Schincariol, and M. Nardone, *XML Mapping Files*, pp. 593–654, Apress, New York, NY, USA, 2018.

[27] H. Suryotrisongko, D. P. Jayanto, and A. Tjahyanto, "Design and development of backend application for public complaint systems using microservice spring boot," *Procedia Computer Science*, vol. 124, pp. 736–743, 2017.

[28] F. Gutierrez, *Pro Spring Boot*, Apress, New York, NY, USA, 2016.

[29] M. T. Nygard, "Release it!: design and deploy production-ready software," Pragmatic Bookshelf, Raleigh, NC, USA, 2018.

[30] S. Mahajan, A. Alameer, P. McMinn et al., "Automated repair of internationalization presentation failures in web pages

using style similarity clustering and search-based techniques," in *Proceedings of the 2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*, IEEE, Västerås, Sweden, pp. 215–226, April 2018.

[31] A. Fox, R. Griffith, A. Joseph et al., *Above the clouds: A Berkeley View of Cloud Computing*, University of California, p. 2009.

[32] N. Naik, "Building a virtual system of systems using docker Swarm in multiple clouds,," in *Proceedings of the 2016 IEEE international Symposium on Systems Engineering (ISSE)*, IEEE, Edinburgh, UK, pp. 1–3, October 2016.

[33] L. Stanescu, M. Brezovan, and D. D. Burdescu, "Automatic mapping of MySQL databases to NoSQL MongoDB," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, Gdansk, Poland, pp. 837–840, September 2016.

[34] D. Omilusik, "Spring boot," U.S. Patent 4-660-299, 1987.

[35] F. Feng and L. Zou, "Design of scratch interactive online learning platform based on webpack and react technology," *Computer Knowledge and Technology*, vol. 2018, no. 20, p. 25, 2018.

[36] D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: improving the ultrafast bootstrap approximation," *Molecular Biology and Evolution*, vol. 35, no. 2, pp. 518–522, 2018.

[37] K. K. Das, R. Raghu, and C. J. Rawles, "Imaging subsurface properties using a parallel processing database system," U.S. Patent 9 720-117, 2017.

[38] E. H. Halili, *Apache JMeter: A Practical Beginner's Guide to Automated Testing and Performance Measurement for Your Websites*, Packt Publishing Ltd, Birmingham, UK, 2008.

[39] A. Kasko, S. Kobylyanskiy, and A. Mironchenko, *OpenJDK Cookbook*, Packt Publishing Ltd, Birmingham, UK, 2015.

Hindawi
Scientific Programming
Volume 2020, Article ID 8894757, 7 pages
https://doi.org/10.1155/2020/8894757

Research Article

Incorporating Research Reports and Market Sentiment for Stock Excess Return Prediction: A Case of Mainland China

Huilin Song,[1] Diyun Peng,[2] and Xin Huang [3]

[1]*School of Management, Nanchang University, Nanchang 330031, China*
[2]*School of Economics Management, Nanchang University, Nanchang 330031, China*
[3]*School of Software, Jiangxi Normal University, Nanchang 330031, China*

Correspondence should be addressed to Xin Huang; 1610466@tongji.edu.cn

Received 18 March 2020; Revised 13 May 2020; Accepted 19 May 2020; Published 31 May 2020

Academic Editor: Chenxi Huang

The prediction of stock excess returns is an important research topic for quantitative trading, and stock price prediction based on machine learning is receiving more and more attention. This article takes the data of Chinese A-shares from July 2014 to September 2017 as the research object, and proposes a method of stock excess return forecasting that combines research reports and investor sentiment. The proposed method measures individual stocks released by analysts, separates the two indicators of research report attention and rating sentiment, calculates investor sentiment based on external market factors, and uses the LSTM model to represent the time series characteristics of stocks. The results show that (1) the accuracy and F1 evaluation indicators are used, and the proposed algorithm is better than the benchmark algorithm. (2) The performance of deep learning LSTM algorithm is better than traditional machine learning algorithm SVM. (3) Investor sentiment as the initial hidden state of the model can improve the accuracy of the algorithm. (4) The attention of the split research report takes the two indicators of investor sentiment and price as the input of the model, which can effectively improve the performance of the model.

1. Introduction

Stock price prediction is a method of predicting stock prices in the future based on stock price information at the past or current moment. Traditional quantitative investment methods are mostly based on experience in forecasting future stock prices. Such methods often have weak antirisk capabilities, poor long-term forecasting capabilities, and slow analysis speed and are not convenient for dissemination and promotion. Randomly appearing statistical and financial-based stock analysis methods belong to the traditional machine learning category. Most of them use autoregressive models, random fluctuation models, and Markov models to make predictions. Compared with empirical methods, this method is faster. It is fast and accurate, but the disadvantage is that it can process less information and cannot fully deal with many factors of external market data that cause stock price fluctuations. Thanks to the massive financial data provided by the continuous development of big data technology, it is possible for artificial intelligence methods to enter the financial analysis field. Therefore, more and more researchers have begun to use machine learning or deep learning methods to analyze stock prices and make predictions. Related methods in the field of artificial intelligence have demonstrated incomparable excellent performance on large-scale datasets. This has been verified in areas such as images [1, 2] and text [3–5]. It is foreseeable that artificial intelligence-related methods can solve many problems in current stock price prediction models. Because of its policies, internal environment, and investor attributes, stocks have different rules in different markets. The stock market in mainland China belongs to an emerging capital market. The imperfect regulatory policies and the characteristics of most investors are retail investors, and making media reports to a great extent can affect the trend of stock prices. Ding and Sun's [6]research shows that in the Chinese A-share market, the behavior of ordinary investors in buying and selling stocks will be largely affected by research reports issued by

financial institutions. Compared with the news media reports, which focus on the occurrence of events and describe the original events, the research reports are more focused on financial and market-related attributes related to stock prices, with the purpose of predicting stock prices. At the same time, as an information publisher, securities analysts have a more professional industry background and richer information channels than ordinary financial news reporters, so for ordinary investors, direct and professional research reports are important reference objects for investment decisions.

In order to better predict the stock price of the Chinese mainland stock market, we propose a method that combines research reports and market sentiment to predict abnormal stock returns. The proposed method measures individual stocks released by analysts and splits the research report. Attention and rating sentiment, calculate investor sentiment based on external market factors, and use the LSTM model to represent the time series characteristics of the stock. We selected A-share data from July 1, 2014, to September 30, 2017, for experiments and compared different algorithms and different inputs. Based on the experimental results, we found that

(1) The accuracy of the proposed algorithm and the evaluation of F1 are better than the benchmark algorithm

(2) The performance of deep learning LSTM algorithm is better than traditional machine learning algorithm SVM

(3) Investor sentiment as the initial hidden state of the model can improve the accuracy of the algorithm

(4) The attention of the split research report and the two indicators of investor sentiment and price are used as input for the model, which can effectively improve the performance of the model

The rest of this article is organized as follows. Section 2 reviews the literature that separately introduced the impact of machine learning-based stock price predictions and research reports on stock prices. Section 3 introduces our proposed method. Section 4 presents the experimental design and details. Section 5 presents the experimental results and discussion. Section 6 gives our conclusions and directions for future work.

2. Related Work

2.1. Machine Learning-Based Stock Price Prediction. In the traditional machine learning field, Xiang [7]used an improved gradient boosting decision tree (GBDT) to predict stock prices. This model can mine the relevant features of the current stock series, but the GBDT model structure itself is not suitable for solving serial data problems like stocks.

Du et al. [8] used a Bayesian learning (BL) model to predict stock prices in the research. This model is actually similar to the autoregressive integrated moving average (ARIMA) model. Based on statistical knowledge, it learns the characteristics of the stock sequence. However, the BL model itself is actually not suitable for sequence data. In the field of deep learning, Tsantekidis et al. [9] in their research proposed a stock price prediction model based on a CNN encoder. CNN is a very effective model for image input. In order to adapt it to sequence data, first use the encoder to encode the sequence data, and then use CNN for training. This method is very similar to the signal and system. In filtering theory, the sequence data can be regarded as a time signal, and the CNN can be regarded as a filter for convolution. Bao et al. [10] in combination with the long short-term memory (LSTM) of the autoencoder (AE) constructed a special algorithm based on the recurrent neural network (RNN). The neural unit structure makes it very suitable for processing sequence data such as stocks. This method even adds an autoencoder to encode the stock sequence through training and then uses the LSTM network for training. Based on the basic deep learning models, more studies have considered the basic characteristics of the stock market and incorporated them into the method. Zhang and Tan [11] used historical price data to predict the future return ranking of stocks through a new stock selection model based on deep neural networks. Li et al. [12]have established a system that uses deep learning architecture to improve feature representation and uses extreme learning machines to predict market impact. They concluded that the feature representation of deep learning together with extreme learning machines can provide better accuracy of market impact predictions. Li et al. [13] emotion vectors are obtained through sentiment analysis of news articles, and sentiment vectors are added to the LSTM model to predict stock prices. The experiments on the Hong Kong stock market have shown good performance.

2.2. Stock Price Prediction Based on Research Reports. Lee et al. [14] believe that after being affected by media sentiment, investors will form a subjective and objective comprehensive judgment on future capital flows and investment risks, which is called "investor sentiment." When investor sentiment is extremely optimistic or pessimistic, stock price volatility increases. At the same time, for the purpose of promotion, commission income, contracting customers or business, etc., the research report written by the securities company is not always neutral, and they will convey information with serious selective deviations to the market to meet the needs of investors, and such deviations are often optimistic. Using data from the "Abreast of the Market" column on the Wall Street Journal's website as a sample, Tetlock [15] constructed a media pessimism index and found that abnormally high or abnormally low media pessimism can cause temporary activity in market trading behavior. Hribar and McInnis [16]in their research found that when investor sentiment is high, analysts' optimism tends to be more obvious. The existence of optimism tends to distort stock prices and seriously affects investor decisions. Zhao et al. [17] proposed that in companies with high stock price synchronization, analysts' optimism tends to have a weaker impact on the accuracy of their subsequent earnings forecasts. Xu et al. [18] believe that optimism tends to lead to

high transaction volumes, but it also easily leads to negative news of listed companies not being disclosed in a timely manner, and the risk of future stock price crashes. Lu and Chen [19] found that the impact of extreme optimism and extreme pessimism on the stock price index is asymmetric, and short-term extreme pessimism has a negative relationship with the stock price index.

2.3. Task. The ups and downs of stocks determine the rate of return on stocks, and most stock prices fluctuate with changes in the stock market environment. When the stock market is in a "bull market," most stocks will rise, and when the stock market is in a "bear market," most stocks will follow the trend and fall. Simply predicting the rise or fall of a stock on the next trading day cannot objectively reflect the stock income. Therefore, this article uses stock excess returns as a research object to explore whether the excess returns obtained by individual stocks in a certain time interval in the future are positive, negative, or par. The calculation of excess returns for individual stocks is as follows:

$$\mathrm{AR}_{k,t} = R_{k,t} - \mathrm{ER}_{k,t}, \tag{1}$$

where $\mathrm{AR}_{k,t}$ is the abnormal rate of return on day t of stock k, $R_{k,t}$ is the actual rate of return on day t of stock k, and $\mathrm{ER}_{k,t}$ is its expectation yield (or expected normal return). There are many methods to calculate the expected normal rate of return. In order to exclude the part of the return that is related to market returns, this article uses the Malkiel and Fama [20] market model to measure. Therefore, the actual yield of individual stocks can be expressed as

$$R_{k,t} = \alpha_k + \beta_k R_{m,t} + \mu_{k,t}, \tag{2}$$

where $R_{m,t}$ is the market return rate on day t, $\mu_{k,t}$ is a random error term, and estimates of α_k and β_k obtain the values $\widehat{\alpha}_k$ and $\widehat{\beta}_k$. The models that measure expected normal returns are

$$\mathrm{ER}_{k,t} = \widehat{\alpha}_k + \widehat{\beta}_k R_{m,t}. \tag{3}$$

Finally, calculate the cumulative abnormal return of stock k during the event window:

$$\mathrm{CAR} = \sum \mathrm{AR}_{k,t}. \tag{4}$$

Due to the high turnover rate of ordinary investors in China's A-share market and the average holding time is about one month, this article selects the excess returns from one trading day to the 5, 15, and 30 trading days as the forecast target.

3. Methods

3.1. Overview. In order to better understand the time series relationship between stock prices, our method uses LSTM networks [21] as the basic unit of the model, and based on this, a research report and market sentiment are quantified into the process of stock prices. Figure 1 shows the structure of our method model.

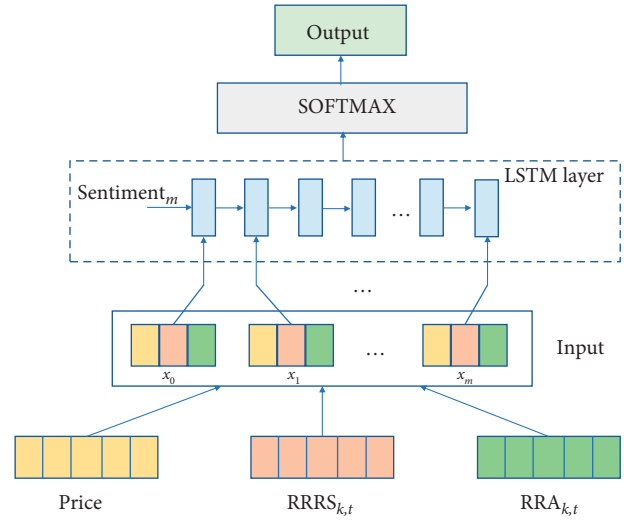


Figure 1: A stock price prediction model incorporating research reports and market sentiment.

First of all, in order to better indicate the status of a stock in the current stock market, we sort the stock price (Price), research report rating sentiment ($\mathrm{RRA}_{k,t}$), and research report attention ($\mathrm{RRRS}_{k,t}$) which is concatenated to get x_t, and the calculation method of x_t is shown in the following formula:

$$x_t = \left[\mathrm{Price}; \mathrm{RRA}_{k,t}; \mathrm{RRRS}_{k,t}\right]. \tag{5}$$

Then, x_t is used as the input of LSTM, and the initial hidden state of LSTM is investor sentiment $\mathrm{Sentiment}_m$. The reason for this is that relative to the positive volatility of the stock price, $\mathrm{Sentiment}_m$ is stable for a period of time and can be regarded as an indicator of market sentiment in the short term. Finally, the output of the LSTM is calculated by the SOFTMAX function to obtain the final output of the model. The calculation methods of LSTM, $\mathrm{Sentiment}_m$, $\mathrm{RRA}_{k,t}$, and $\mathrm{RRRS}_{k,t}$ will be described in detail in the following sections.

3.2. Long Short-Term Memory Network. The LSTM proposed by Hochreiter and Schmidhuber [21] in 1997 can effectively deal with the long-term dependencies in the sequence, and its structure is shown in Figure 2.

The core of LSTM lies in its memory unit, and related information is transmitted backward through the memory unit. Theoretically, the memory unit can transfer information during the entire sequence propagation process so that the information at the previous time can be used to predict the output at the later time, so it can solve the short-term memory problem of the traditional recurrent neural network. In addition, during the backward transfer of information in the memory unit, the LSTM adds or deletes information in the memory unit through three gates. These gates can be seen as different neural networks, which can be trained to automatically learn what information to keep or forget. The process of LSTM processing information is as follows. First, the LSTM will use the "forget gate" to determine which information should be removed. The input is

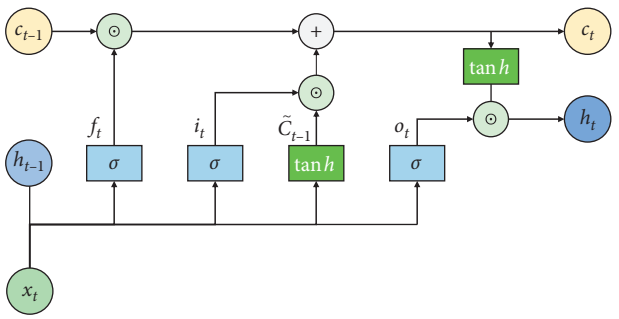


Figure 2: Long short-term memory network structure.

mapped between 0 and 1 by the Sigmoid function. The trend to "1" means to retain the information; otherwise, it means to forget information. The "input gate" is used to determine which information needs to be updated. The Sigmoid function is used to determine whether it needs to be retained. Then, the tan h function maps the input value to [−1, 1], thereby generating a new memory unit state and adding it to the original memory unit. Next, to update the value of the memory unit, first multiply the memory unit by the forget gate, discard the information that needs to be forgotten, and then add the input information obtained from the input gate to obtain the new memory unit value. Finally, the "output gate" decides which memory unit information to output, that is, the hidden state. The calculation formula of LSTM is as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \tag{6}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \tag{7}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right), \tag{8}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{9}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \tag{10}$$

$$h_t = o_t \odot \tanh\left(C_t\right), \tag{11}$$

where f_t, i_t, o_t, and C_t, respectively represent the forget gate, input gate, output gate, and memory unit. For the t time step, h_{t-1} represents the hidden state at the previous time step, W represents the weight matrix, σ represents the Sigmoid function, and $\odot$ represents the point multiplication operation.

3.3. Measure of Research Report. We measure the research report from the two dimensions of attention and rating sentiment. The attention of the research report can measure the popularity of the stock in the entire market by analysts. The rating sentiment indicates an analyst's judgment on the future trend of the target stock.

3.3.1. Attention of Research Report. Different stocks in the market receive different degrees of attention. We calculate the ratio of the absolute number of research reports on a stock day to the total number of all stock research reports in the A-share market for the current month as a measure of the stock's attention, $\mathrm{RRA}_{k,t}$ and $\mathrm{RRRS}_{k,t}$. The larger the value, the higher the stock analyst's attention. $\mathrm{RRA}_{k,t}$ is calculated as shown in the following formula:

$$\mathrm{RRA}_{k,t} = \frac{N_{k,t}}{N_{A,m}}, \tag{12}$$

where $N_{k,t}$ is the total number of research reports of stock k on the t-th day and $N_{A,m}$ is the total number of research reports of all stocks in the A-share market in that month (m).

3.3.2. Rating Sentiment Measures in Research Reports. The text of the research report released by the analyst contains the optimistic or pessimistic attitude of the individual company's operating status, future prospects, earnings expectations, investment recommendations, and risk warnings. The report contains two important key pieces of information: first, the current rating, and second, the rating change. Among them, the current rating provides investment advice on buying, selling, or holding of individual stocks; the rating change indicates the current rating and the previous rating change are reported. In the previous literature, when discussing the impact of stock investment ratings on their abnormal returns, the basic ratings and rating changes were always analyzed separately, making it difficult for investors to choose when the two ratings were inconsistent. For example, when an analyst gives a "Hold" rating to individual stocks, and the rating changes to "Down," it is more difficult for investors to determine whether to buy or sell, so this article innovatively proposes a "research report rating sentiment" index, $\mathrm{RRRS}_{k,t}$, taking into account the two major factors of basic rating and rating change; the calculation method of $\mathrm{RRRS}_{k,t}$ is shown in the following formula:

$$\mathrm{RRRS}_{k,t} = \frac{\sum_1^{N_{k,t}} R_{k,t} \cdot C_{k,t}}{N_{k,t}}, \tag{13}$$

where $R_{k,t}$ is the base rating and $C_{k,t}$ is the rating change. When there are multiple rating results for a stock within a day, this article will average the ratings of these research reports to find a comprehensive rating sentiment.

3.4. Measure of Investor Sentiment. In the actual trading process, investor sentiment will affect investors subjective judgments on future returns. When investor sentiment rises or becomes more pessimistic, it will trigger its "irrational" behavior of information and cause market anomalies. For the calculation method of investor sentiment quantification, this article draws on the ideas of Hai-Yuan [22], selects Shanghai and Shenzhen cities A transaction volume (VOLUME), the number of new investor accounts (NEWIN), and consumer confidence index (CCI). The five indicators of closed-end fund discount rate (FUND) and broad market turnover rate (HS_TVR) were used for principal component factor analysis, and the initial investor sentiment index for each month was calculated using the

respective variance contribution rate as the weight. Then, macroeconomic control was introduced, regression analysis was performed on the variables, and the calculated residual value was used as the investor sentiment index. Finally, the simple investor average of the period i lagging behind the study report date (where $i = 3$) was used to obtain the final investor sentiment index.

First of all, the abovementioned five indicators are standardized, and principal component factor analysis (PCA) is performed on these standardized variables. The three principal components with the highest variance explanations are selected, and the respective feature values are used as weights to obtain the factor load after weighted averaging, and as the principal component coefficients of the preliminary sentiment index, the preliminary sentiment index is shown in the following formula:

$$\begin{aligned}\text{Sentiment} = {} & -0.173\text{FUND} + 0.242\text{NEWIN} \\ & + 0.737\text{VOLUME} + 1.047\text{HS\_TVR} + 0.512\text{CCI}.\end{aligned} \tag{14}$$

Then, control the impact of macroeconomic variables. Take the abovementioned preliminary sentiment indicators as the explanatory variables, and the consumer consumption index (CPI), the amount of new credit (IC), the rate of economic growth (GDP), and the money supply (M_2) as the explanatory variables (standardize the data in advance to eliminate dimensional impact), regression analysis of Sentiment_m, the residual sequence can be used as an indicator of investor sentiment: CSI (China Sentiment Index).

$$\text{Sentiment}_m = \alpha_0 + \alpha_1\text{CPI} + \alpha_2\text{IC} + \alpha_3\text{GDP} + \alpha_4 M_2 + \varepsilon_m, \tag{15}$$

where Sentiment_m is the preliminary investor sentiment index for m months, α_0 is a constant, and α_{1-4} is the regression coefficient to be estimated.

Finally, considering the lag of investor sentiment, the sentiment index of the three months before the month on which the research report was published was selected to calculate the average value as the final investor sentiment index.

4. Experiments

4.1. Data Collection. This article selects the research report on 2,225 Chinese A-share companies issued by 66 securities institutions between July 1, 2014, and September 30, 2017, as the research object. The data on the number of research reports published, the date of publication, the title, the basic rating, and the rating changes are from the Oriental Fortune website. Economic data such as stock returns, market value of stocks in circulation, and stock turnover rate are taken from the wind database. For the selected time period, the trend of the Chinese A-share market can be roughly divided into two phases: July 2014 to June 2015 is the rising period of the stock market, which belongs to the "bull market," and July 2015 to September 2017 is the decline of the stock market period, belonging to the "bear market." The sample time spans a bull-bear cycle, which can more effectively verify the robustness of the algorithm.

4.2. Data Preprocessing

4.2.1. Data Culling. This article deletes some anomalous data: first, unrated or ambiguous research reports; second, new stock data, because during the continuous daily limit of new stock sales, stock price fluctuations do not truly reflect market fluctuations, and during this period, few investors can successfully buy new stocks, so the new stocks issued from July to September 2017 are uniformly eliminated; third, the individual stock research report at the time of long-term suspension, because the long-term suspension of stocks cannot be traded, and the price of the stock cannot be compared with the average market price without fluctuation (temporarily suspended stocks are not excluded).

4.2.2. Rating Consolidation. In all the reports collected, a total of 27 different basic ratings were included, and we obtained a total of 14 different ratings after synonym merger. At the same time, we sorted out four different rating changes. Referring to the research by You et al. [23], we use discrete values to assign ratings and rating changes. This article assigns the "Neutral" rating to 1.0 and increases or decreases by 0.1 according to the intensity change to obtain the basic rating G, $G \in [0.6, 1.9]$. The specific values of the rating are shown in Table 1:

For the four rating changes of "Up," "First," "Maintenance," and "Down," we assign "maintenance" to 1.0 and increase or decrease by 0.1 according to the rating change to get the rating change C, $C \in [0.9, 1.2]$; the specific value of the rating change assignment is shown in Table 2:

4.3. Detail. All data are divided into training set, validation set, and test set according to the ratio of 80%, 10%, and 10%. We use categorical cross-entropy as a loss function to optimize the target parameters during the model's back-propagation, which is defined as

$$\text{Loss} = -\frac{1}{n}\sum_{t=1}^{n} y_t \log p\left(\widehat{y}_t\right), \tag{16}$$

where y_t is ground-truth in the form of one-hot, and $\widehat{y}_t$ is the model's predicted probability that the excess return is a "positive," "negative," and "par" vector. During the model training process, the Adam [24] function was selected for optimization, where the initial learning rate was set to $1e{-}4$ and the minimum batch size was set to 32.

5. Results and Discussion

In order to measure the performance of the model from different angles, this paper chooses the classic classification algorithm SVM and the vanilla LSTM model as the benchmark method to compare with our work. Table 3 shows the results of the experiment. The results show that in the excess return forecast on the 5th, 15th, and 30th, our proposed method achieves the best performance regardless of the accuracy rate or F1 measurement. In the comparison of the 5th, 15th and 30th, the accuracy of the excess return prediction on the 15th by all methods is the highest. This is

Table 1: Assignment of research report ratings.

Rating	Value	Rating	Value
Strong buy	1.9	Cautious overweight	1.2
Buy	1.8	Hold	1.1
Long-term buy	1.7	Neutral	1.0
Overweight	1.6	Recommended with caution	0.9
Buy with caution	1.5	Wait and see	0.8
Highly recommended	1.4	Underweight	0.7
Recommend	1.3	Sell	0.6

Table 2: Assignment of research report rating changes.

Change	Up	First	Maintenance	Down
Value	1.2	1.1	1.0	0.9

Table 3: Comparison of excess return forecast results by different methods.

Methods	5 days		15 days		30 days	
	Accuracy	$F1$	Accuracy	$F1$	Accuracy	$F1$
SVM	0.631	0.642	0.7081	0.6965	0.6660	0.6705
LSTM	0.6598	0.6671	0.7217	0.7201	0.6876	0.6911
OURS	**0.6833**	**0.6792**	**0.7381**	**0.7399**	**0.7087**	**0.6986**

Table 4: Comparison of excess return forecast results in different market environments.

Methods	5 days		15 days		30 days	
	Accuracy	$F1$	Accuracy	$F1$	Accuracy	$F1$
			Bull market			
SVM	0.6810	0.6931	0.7335	0.7398	0.7161	0.7176
LSTM	0.6938	0.6978	0.7534	0.7583	0.7216	0.7271
OURS	**0.7155**	**0.7203**	**0.7801**	**0.7868**	**0.7337**	**0.7312**
			Bear market			
SVM	0.6688	0.6724	0.7186	0.7113	0.6922	0.6950
LSTM	0.6814	0.6957	0.7476	0.7561	0.7284	0.7189
OURS	**0.7076**	**0.7055**	**0.7465**	**0.7493**	**0.7195**	**0.7112**

Table 5: Comparison of excess return forecast results with different inputs.

Methods	5 days		15 days		30 days	
	Accuracy	$F1$	Accuracy	$F1$	Accuracy	$F1$
LSTM	0.6598	0.6671	0.7217	0.7201	0.6876	0.6911
LSTM + Sentiment$_m$	0.6601	0.6766	0.7316	0.7287	0.6891	0.6934
LSTM + Sentiment$_m$ + RRA	0.6768	0.6784	0.7301	0.7287	0.6965	0.7011
OURS_full	**0.6833**	**0.6792**	**0.7381**	**0.7399**	**0.7087**	**0.6986**

related to the release cycle of the research report. According to statistics, the average cycle of all stock research reports in the dataset is 18.9 days. When the period exceeds 20 days, there will be multiple reports overlapping, and the latest report will affect stock price fluctuations and thus affect excess returns.

We included a "bull market" and a "bear market" in our selected trading cycle. Different market states present different trading sentiments, so we further trained the "bull market" and "bear market" data separately and tested them on the test set, Table 4 shows the comparison results. Compared with the training of "bull market" and "bear market" data aggregation, the accuracy of all methods after training separately according to different market conditions has improved. Similarly, our proposed method achieved the best performance in the 5th, 15th, and 30th excess return forecasts, with the highest accuracy rate as the "bull" 15th excess return forecast. The table shows that the overall accuracy of the "bull market" is higher than that of the "bear market," and this conclusion is consistent with the research results of Hai-Yuan [22].

In order to verify the effectiveness of the increased research report metrics and investment sentiment on the model, we delete the corresponding inputs and compare them. Table 5 shows the results. Among them, LSTM + Sentiment$_m$ indicates that the original hidden

LSTM uses Sentiment$_m$ as the initial hidden state of the model and the input of the model is only the stock price. Based on Sentiment$_m$, the input of the model is the concatenation of the stock price and RRA, and OURS_full represents our complete model. The results show that the addition of Sentiment$_m$, RRA, and RRRS can gradually improve the accuracy of the model.

6. Conclusion

Regarding the prediction of excess returns in the mainland Chinese stock market, first of all, this article measures the research report released by the analyst and splits the research report into two indicators: the attention degree of the research report and the rating sentiment; secondly, we quantify the external environment that may affect stock price changes as investor sentiment; then, we split the research report indicators and investor sentiment as the input and initial hidden state of LSTM; finally, in the comparison of experiments, our proposed method achieved the best performance.

Data Availability

All data in this article are from public websites (https://www.eastmoney.com/).

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1097–1105, Lake Tahoe, NV, USA, December 2012.

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[3] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition," in *Proceedings of the Twelfth Annual Conference of the International Speech Communication Association*, Florence, Italy, August 2011.

[4] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th international conference on Machine learning—ICML '07*, pp. 641–648, Corvallis, OR, USA, June 2007.

[5] S. Huilin, P. Diyun, H. Xin, and F. Jun, "Research on weibo hotspot finding based on self-adaptive incremental clustering," *Journal of Shanghai Jiaotong University (Science)*, vol. 24, no. 3, pp. 364–371, 2019.

[6] L. Ding and H. Sun, "A study of the effect of recommending stocks to China' s stock Market%," *Manage. World*, vol. 000, no. 5, pp. 111–116, 2001.

[7] L. Xiang, *Multi-factor Quantitative Stock Selection Plan Planning Based on XGBoost Algorithm*, Shanghai Normal University, Shanghai, China, 2017.

[8] B. Du, H. Zhu, and J. Zhao, "Optimal execution in high-frequency trading with Bayesian learning," *Physica A: Statistical Mechanics and Its Applications*, vol. 461, pp. 767–777, 2016.

[9] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Forecasting stock prices from the limit order book using convolutional neural networks," in *Proceedings of the 2017 IEEE 19th Conference on Business Informatics (CBI)*, pp. 7–12, Thessaloniki, Greece, July 2017.

[10] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLoS One*, vol. 12, no. 7, 2017.

[11] X. Zhang and Y. Tan, "Deep stock ranker: a LSTM neural network model for stock selection," in *Data Mining and Big Data*, pp. 614–623, Springer, Berlin, Germany, 2018.

[12] X. Li, J. Cao, and Z. Pan, "Market impact analysis via deep learned architectures," *Neural Computing and Applications*, vol. 31, no. 10, pp. 5989–6000, 2019.

[13] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong," *Information Processing & Management*, Article ID 102212, 2020.

[14] C. M. C. Lee, A. Shleifer, and R. H. Thaler, "Investor sentiment and the closed-end fund puzzle," *The Journal of Finance*, vol. 46, no. 1, pp. 75–109, 1991.

[15] P. C. Tetlock, "Giving content to investor sentiment: the role of media in the stock market," *The Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[16] P. Hribar and J. McInnis, "Investor sentiment and analysts earnings forecast errors," *Management Science*, vol. 58, no. 2, pp. 293–307, 2012.

[17] L. Zhao, Z. Li, and J. Liu, "The managers preferences, the optimization in the evalution of the investment level and the obtainment of the private information," *Manage. World*, vol. 4, pp. 33–47, 2013.

[18] N. Xu, X. Jiang, Z. Yi, and X. Xu, "Conflicts of interest, analyst optimism and stock price crash risk," *Economics Research Journal*, vol. 7, no. 127, p. r140, 2012.

[19] J. Lu and J. Chen, "Asymmetric relationship between extreme investor sent iment and stock index," *Systems Engineering*, vol. 2, pp. 13–22, 2013.

[20] B. G. Malkiel and E. F. Fama, "Efficient capital markets: a review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] Y. I. N. Hai-Yuan, "A study on effect of media reports on investor sentiment:evidence from China's stock market," *Journal of Xiamen University (Arts & Social Sciences)*, vol. 2, p. 11, 2016.

[23] J. You, Y. Qiu, and C. Liu, "'Changed face phenomena of security analysts' forecasting behaviors: a reputation game model and evidences," *Journal of Management Science. China*, vol. 16, no. 6, 2013.

[24] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv Prepr. arXiv1412, 6980, 2014.

Hindawi
Scientific Programming
Volume 2020, Article ID 8895875, 12 pages
https://doi.org/10.1155/2020/8895875

Research Article

Detection and Classification of Early Decay on Blueberry Based on Improved Deep Residual 3D Convolutional Neural Network in Hyperspectral Images

Shicheng Qiao,[1,2] Qinghu Wang,[1] Jun Zhang,[1] and Zhili Pei [1]

[1]*Inner Mongolia University for Nationalities, College of Computer Science and Technology, Tongliao 028043, China*
[2]*College of Information and Electric Engineering, Shenyang Agricultural University, Shenyang 110866, China*

Correspondence should be addressed to Zhili Pei; zhilipei@imun.edu.cn

Received 21 March 2020; Accepted 8 April 2020; Published 20 May 2020

Academic Editor: Chenxi Huang

Recently, the automatic detection of decayed blueberries is still a challenge in food industry. Early decay of blueberries happens on surface peel, which may adopt the feasibility of hyperspectral imaging mode to detect decayed region of blueberries. An improved deep residual 3D convolutional neural network (3D-CNN) framework is proposed for hyperspectral images classification so as to realize fast training, classification, and parameter optimization. Rich spectral and spatial features can be rapidly extracted from samples of complete hyperspectral images using our proposed network. This combines the tree structured Parzen estimator (TPE) adaptively and selects the super parameters to optimize the network performance. In addition, aiming at the problem of few samples, this paper proposes a novel strategy to enhance the hyperspectral image sample data, which can improve the training effect. Experimental results on the standard hyperspectral blueberry datasets show that the proposed framework improves the classification accuracy compared with AlexNet and GoogleNet. In addition, our proposed network reduces the number of parameters by half and the training time by about 10%.

1. Introduction

Blueberries are popular worldwide for their excellent flavor and high nutritional value [1]. Most of blueberries used for fresh consumption are hand-picked and transported over long distances. Damage during transportation will accelerate fruit decay and reduce overall quality [2]. Therefore, it is important to identify rotten blueberries from healthy blueberries to remove low-quality blueberries from the fresh blueberry supply chain [3].

Under the current industrial standard, the internal decay of blueberry is usually judged by the human touch or by observing the dark rotten tissue of blueberry [4, 5]. The decayed tissue of blueberry becomes darker and more obvious, similar to black, and easier to observe with the naked eye. However, it takes a lot of manpower and time to identify the degree of decay, and it will become inaccurate after several hours of continuous inspection [6]. In addition, the inspection efficiency is very low, and the inspection of early decay is not accurate. The development of hardness measurement method accelerates the detection process of fruit quality evaluation and makes fruit classification more accurate, including blueberry hardness and texture analyzer [7], tomato acoustic pulse response measurement [8], and peach inspection method based on frequency resonance [9]. These methods can provide more accurate hardness measurement, but many hardness measurement technologies need direct touch with fruits, which may cause blueberries to be damaged.

Some researchers at home and abroad used nondestructive testing technology such as machine vision and Hyperspectral Imaging to detect fruit disease or maturity [10] and achieved some excellent results. Georgina et al. [11] used machine vision technology to extract 14 types of features such as color, shape, and texture of citrus and then used Classification And Regression Trees(CART), naive Bayes

(NB), and multilayer perceptron (MLP) to detect citrus canker, black spot, and sclerosis. Lewers et al. [12] used machine vision technology to detect pomegranate disease, where K-means and threshold segmentation methods were used in the experiment to extract the lesion area of pomegranate, and the discrete wavelet transform is adopted to get a set of visual features of the lesion as the input vector of the support vector machine (SVM) model so as to identify pomegranate disease. Lorente [13] uses the hyperspectral imaging system to obtain the hyperspectral images of sound, slight-decayed, moderate-decayed, and severe-decayed peaches, the threshold segmentation method to detect the disease area of peaches, and then the successive projections algorithm (SPA) to extract six characteristic wavelengths, establishes the partial least squares discriminant analysis (PLS-DA) model to identify the disease, and further improves the identification rate of the rotten peaches. Wang et al. [14] also uses hyperspectral imaging technology to obtain the spectral data of the region of interest, where five characteristic wavelengths are extracted by the permutation test method, and the multiple partial least squares regression discriminant analysis model is used to detect the citrus-rot disease caused by fungal infection. Wang et al. [15] used hyperspectral imaging technology to obtain apple spectrum data. Firstly, the threshold segmentation method is used to segment Apple lesion area and extract hyperspectral data; then, the successive projection algorithm is adopted to extract three characteristic wavelengths from the full wavelength; finally, an improved linear discriminant analysis combined with the support vector machine and BP artificial neural network model to detect apple disease. Liu et al. [16] used hyperspectral imaging technology to detect and distinguish the crack, peel spots, malformation, hidden damage, and normal fruit of nectarine. Ten characteristic wavelengths were extracted and the top ten principal component values were obtained by principal component analysis. The disease areas of Nectarine were extracted by threshold segmentation. Finally, the principal component value and six texture indexes (mean, contrast, correlation, energy, homogeneity, and entropy) are fused to establish the ELM model to detect and distinguish the external defect samples and intact samples.

Hyperspectral imaging technology covered the range of 420–1000 nm was employed to detect the nectarine fruit in the literature [17]. 400 RGB images were acquired through a total of 400 samples, which included four types of defective features and sound features. After acquiring hyperspectral images of nectarine fruits the spectral data were extracted from region of interest (ROI). Using Kennard Stone algorithm, all kinds of samples were randomly divided into training set (280) and testing set (120). First of all, according to the calculation of partial least squares regression (PLSR), 10 wavelengths at 497 nm, 534 nm, 657 nm, 677 nm, 696 nm, 709 nm, 745 nm, 823 nm, 868 nm, and 943 nm were selected as the optimal sensitive wavelengths (SWs), respectively. Subsequently, the image of the 876 nm wavelength was selected as the feature image; then, principal component analysis (PCA), Sobel edge detector, and region growing algorithm were carried out among defective and normal nectarines to extract the defective region. Moreover, ten principal components (PCs) were selected based on PCA and seven textural feature variables (mean, contrast, correlation, energy, homogeneity, and entropy) were extracted by using gray level co-occurrence matrix (GLCM), respectively. Finally, the ability of hyperspectral imaging technique was tested by using the extreme learning machine (ELM) models. The ELM classification model was built on the basis of the combination of PCs and textural features. The results show the correct discrimination accuracy of defective samples was 91.67 %, and the correct discrimination accuracy of normal samples was 100%. The research revealed that the hyperspectral imaging technique is a promising tool for detecting defective features in nectarine which could provide a theoretical reference and basis for design in the classification system of fruits in further work [18–20].

In the abovementioned research studies, whether machine vision technology or hyperspectral imaging technology is used, the disease areas of citrus, pomegranate, and other medium-sized fruits need to be separated from the normal areas. Because the color characteristics of the disease areas are obviously different with that of normal areas, the disease areas can be easily separated by threshold segmentation. However, the skin color of blueberry is darker, and the color characteristics of its normal area and disease area are similar, so it is difficult to segment blueberry disease effectively by using the conventional threshold segmentation method [15]. With the development of intelligent signal processing technology, using the convolutional neural network (CNN), we can overcome the abovementioned shortcomings [21]. CNN has a very prominent performance in machine vision tasks by using the local receptive field model to simulate human brain image processing. For example, the two-dimensional convolutional neural network (2D-CNN) is used to mine the spatial features of the principal component band, and the spectral features are fused by the feature fusion technology to classify the images [21]. In the two-channel CNN method, one-dimensional convolutional neural network is used for spectral feature information, while the spatial feature information is extracted, fused, and classified by 2D-CNN, which also achieves good results. However, there is a disadvantage in these methods: before extracting features, principal component analysis (PCA) and other methods should be used to select the principal component band to reduce the dimension, otherwise too many parameters will be introduced, which is difficult to train and optimize the deep network.

The advantage of 3D-CNN convolution kernel in extracting hyperspectral image features is that the spectral information and spatial information are extracted synchronously, which gives full play to the advantages of 3D hyperspectral image [22]. The 3D-CNN feature model directly extracts the spectral spatial features of hyperspectral image end-to-end, which has better classification effect than 2D-CNN features. Spectral spatial-based residual network introduces the residual structure into the 3D-CNN network and uses two 3D convolution kernels of spectral and spatial features to extract deep features, which can improve the recognition accuracy for mildew blueberries [23].

There are some problems in the existing 3D-CNN model, for example, the number of network layers is generally shallow, hyperparameter optimization is time-consuming and laborious, and the accuracy needs to be further improved. To solve the abovementioned problems, the traditional hyperspectral 3D convolution method is improved to obtain the deep features with stronger representation, and it combines the tree structured Parzen estimator (TPE) adaptively and selects the super parameters to optimize the network performance [23]. In addition, aiming at the problem of few samples, this paper proposes a novel strategy to enhance the hyperspectral image sample data, which can improve the training effect.

The contributions of this article are summarized as follows:

(1) An improved Deep Residual 3D Convolutional Neural Network is proposed. The input image of the model is the original hyperspectral image, no dimensionality reduction method is needed, and the image space and spectral characteristics are retained. The extracted features are more representative of hyperspectral images. It makes full use of spectral and spatial 3D correlation information instead of just their separate and independent feature information.

(2) It can avoid introducing excessive parameters, prevent overfitting, and improve computing efficiency; compared with 2D-CNN, 3D-CNN is more suitable for hyperspectral image processing tasks.

(3) Rich spectral and spatial features can be rapidly extracted from samples of complete hyperspectral images using our proposed network. This combines the tree structured Parzen estimator(TPE) adaptively andselects the super parameters to optimize the network performance. In addition, aiming at the problem of few samples, this paper proposes a novel strategy to enhance the hyperspectral image sample data, which can improve the training effect.

2. Blueberry and Its Hyperspectral Imaging Features

Since this study uses hyperspectral imaging mode to detect rotten areas of blueberries, this section needs to introduce blueberries and their hyperspectral imaging functions. Blueberry is a typical climacteric fruit. In the process of maturity, the physical and chemical properties of the inside of the fruit are constantly changing, the color is gradually changing from green to blue or dark purple, and the picking period of blueberry is relatively concentrated. Figure 1(a) shows fresh blueberries on fruit trees. Because the temperature in picking season is high in summer, the fruits are easy to soften or even brown after picking. In the process of transportation, storage, and sales, they are also prone to rot and disease. Because the picking time of fruit is one of the key factors that lead to the taste of fruit, picking the fruit in advance will lead to too stiff and sour taste, affect the flavor and value of the fruit, and it is difficult to meet the eating requirements; picking the fruit too late will lead to overripeness, be easy to deteriorate, and be inconvenient for storage, so it is not easy to carry out subsequent processing. Figure 1(b) shows the mildewed blueberry. Therefore, sorting blueberries after picking is of great significance to increase the added value of blueberries.

Hyperspectral imaging integrates image processing and spectroscopic techniques to obtain the hyperspectral 3D cube data (hypercube). Hyperspectral data cube is not really images that represent spatial 3D. Strictly speaking, the hyperspectral image should be a 2.5D image data. In terms of images, most of the digital images usually are RGB (red, green, and blue) images, which are made up of three basic colors. That is to say, an RGB image can be divided into red, green, and blue components, and each component can generate a gray image. In digital images, this grayscale image is composed of a 2D data matrix, and each data in the matrix is commonly referred to as a pixel. For example, a 256×256 RGB image, its actual data storage size is $256 \times 256 \times 3$, where 3 represents its three RGB components. If these 3 components are extended to hundreds or thousands of continuous bands, such as 100 continuous bands, the data of the image will be expanded to $256 \times 256 \times 100$, and this 100 is the expansion of the spectrum, which makes the image add rich spectral information. The x and y of a hyperspectral image represent its image in the pixel dimension. If you take a point from the image dimension, this point can be connected in the spectral dimension to get the spectrum at this point.

3. Deep Residual 3D Convolutional Neural Network

3.1. 3D Convolutional Neural Network. 2D-CNN, as a classic deep learning in image processing, has outstanding performance in a variety of machine vision tasks, such as image classification, object detection, and dense captioning tasks [24–26]. The advantage of 2D-CNN is that the features can be directly extracted from ordinary images to complete end-to-end processing. The structure of 2D-CNN is shown in Figure 2, where N is the size of the convolution kernel in the convolution layer and L is the number of output channels of the convolution layer. The convolution process can be written by the following equation:

$$V_{1j}^{xy} = f\left(\sum_{m}\sum_{h=0}^{H_{1-1}}\sum_{w=0}^{W_{1-1}} K_{1jm}^{hw} V_{(1-1)m}^{(x+h)(y+w)} + b_{1j}\right), \quad (1)$$

where m is s the number of channels; H_{1-1} and W_{1-1} are the size of the convolution kernel, respectively; k and b are the linear coefficients.

Each channel needs to train a convolution kernel when performing 2D convolution processing. If 2D-CNN is used directly in the hyperspectral image classification task, a large number of parameters will be introduced into the calculation because of the many channels in the hyperspectral image. Too many parameters not only make the network more prone to overfitting and affect the accuracy but also greatly reduce the training speed and calculation efficiency of the network.

(a) (b)

Figure 1: Comparison of blueberries with different levels. (a) Fresh blueberries on fruit tree; (b) mildewed blueberries.

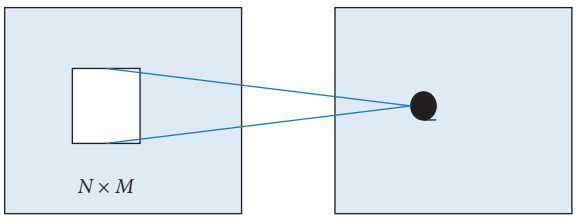


Figure 2: The structure of 2D-CNN.

Usually, in order to solve this kind of problem, scholars take dimension reduction as preprocess before inputting hyperspectral image. For example, they use the PCA method to extract 3 principal component channels in the hyperspectral image, use random PCA (randomized PCA, R-PCA) to keep 10 or 30 principal component channels, and then use 2D-CNN for classification. Since 2D-CNN only performs convolution operations in space and simple linear operations in the spectral dimension, the obvious disadvantages of this type of method is that it will cause the loss of spectral data, which will affect the recognition results [27].

Differing from 2D-CNN, the 3D-CNN convolution structure is shown in Figure 3, where N and M are the plane of the convolution kernel in the convolution layer and spectral dimensions, and L is the number of output channels of the convolution layer. The model can be described as follows:

$$V_{1j}^{xyz} = f\left(\sum_{m}\sum_{h=0}^{H_{1-1}}\sum_{w=0}^{W_{1-1}}\sum_{r=0}^{R_{1-1}} K_{1jm}^{hwr} V_{(1-1)m}^{(x+h)(y+w)(z+r)} + b_{1j}\right), \tag{2}$$

where R_{1-1} is the spectral dimension of convolution kernel.

The 3D-CNN algorithm, which has one more convolution kernel dimension R_{1-1} than the 2D-CNN, can solve the above problem because it has the following advantages:

(1) The input image is the raw hyperspectral image, without the need to use the dimension reduction method, and the image space and spectral features are preserved.

(2) The extracted features are more representatives of hyperspectral images. 3D-CNN is different from 2D-CNN. Instead of plane convolution, it performs convolution operations in both spatial and spectral dimensions to extract the features of the "spectral" combination of hyperspectral images. It makes full use of spectral and spatial 3D correlation information instead of just their separate and independent feature information.

(3) It can avoid introducing excessive parameters, prevent overfitting, and improve computing efficiency. Assuming that the size of the convolution kernel is 3, the number of hyperspectral channels is 200, and the number of output channels is 32, the first 2D-CNN operation requires $3 \times 3 \times 200 \times 32 = 57600$ parameters, and 3D-CNN operation requires $3 \times 3 \times 3 \times 1 \times 32 = 864$ parameters.

Therefore, compared with 2D-CNN, 3D-CNN is more suitable for hyperspectral image processing tasks. However, as the network structure deepens, the vanishing gradient problem will appear, which can affect the training effect of deep neural networks, so introducing the residual error structure is particularly critical.

3.2. Residual 3D-CNN Structure. In deep learning, the deeper the network structure, the more accurate the extracted features and the better the classification results. However, as the network structure continues to deepen, gradients will diffuse or explode during the backpropagation process, resulting in bad effect on network training. After the residual structure is proposed, due to the existence of shortcuts, the gradient is more easily and effectively propagated, which is good to solve the problem. In order to build a deeper network structure, this paper also introduces residuals into 3D-CNN and designs the residual 3D convolution structure block.

According to the design rules for the size of the convolution kernel in 2D-CNN, several consecutive 3×3 convolution kernels have the same field of view as the large convolution kernel and contain fewer parameters and fewer more complex nonlinear features. Research results show that

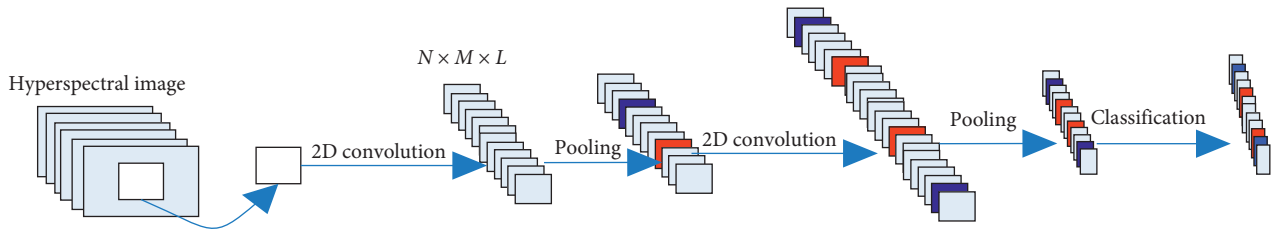


Figure 3: The structure of 3D-CNN.

the $3 \times 3 \times 3$ small convolution kernel is the optimal choice for the spatiotemporal feature learning of video input. In addition, many algorithms for CT 3D image detection also use $3 \times 3 \times 3$ convolution kernels and have achieved good results. Because hyperspectral images and video and the CT images have plane image information and similar 3D data structures, as a reference, this paper designs all the convolution kernel structures used for spectral feature extraction in the network to the size of $3 \times 3 \times 3$.

The residual convolutional structure block is shown in Figure 4. In the residual structure of this paper, there are two forms of shortcut, one is the identity residual block, whose input and output dimensions remain the same, as shown in Figure 4(a). The other is the convolutional residual block, which has different input and output dimensions. The purpose of the design is to change the number of channels. The shortcut of the convolutional form uses $l \times l \times l$ convolution kernel, which will not introduce a large number of parameters, as shown in Figure 4(b). The deepening or complication of the network structure will necessarily introduce some additional hyperparameters, such as the size of the convolution kernel of each convolutional layer and the number of channels, so these hyperparameters need to be selected more reasonably.

In order to improve the calculation efficiency, the network does not directly perform a convolution operation with the size of 3 on each convolutional layer input but uses a bottleneck structure, which will effectively reduce the number of parameters and computational complexity. Assume that there are 256 features as inputs, and if only $3 \times 3 \times 3$ convolution operations are performed, $256 \times 3 \times 3 \times 3 \times 256 = 1769472$ convolution operations must be performed; if the bottleneck structure is adopted, then only $(256 \times l \times l \times l \times 64) + (64 \times 3 \times 3 \times 3 \times 64) + (64 \times 1 \times 1 \times 1 \times 256) = 143360$ convolution operations are performed. The bottleneck structure is used in NIN, GoogleNet, and ResNet [12, 28]. This structure can effectively reduce the computational complexity and enhance the nonlinear expression ability of the network to a certain extent.

In addition, a batch normalization layer (BN) is introduced after each convolutional layer. BN can effectively prevent vanishing gradient and gradient explosion. Although it introduces additional calculations, it can make the overall convergence rate of the model faster. It is worth noting that the network uses ELU (exponential linear units) instead of ReLU (rectified linear unit) as the nonlinear activation function. Although the ReLU function has very good characteristics and is widely used, when its existence input is negative, the derivative will become 0 and no longer change, which will lead to the problem that neurons die and will never be activated. To solve this problem, the ELU function presents a "Soft saturation" state at the part of less than 0, making the derivative not become 0, thus keeping the neuron alive [29].

4. Detection and Classification Based on 3D Deep Residual Model

The input of the network is a 3D data matrix in 3D-CNN, which is obtained by taking a pixel in the original image as the center and its size as $S \times S \times L$, where L is the number of hyperspectral image channels and S is the size of plane dimension. However, the amount of calculation and recognition accuracy introduced by different sizes of the visual field range are also different. According the tradeoff among the accuracy rate, operation efficiency, and other factors, this paper finally fixed the dimension to 7×7 in the multispectral image.

As we all know, the deep learning model has the two optimization tasks. One is the optimization of internal parameters, such as the allocation of weights in neural networks; the other is the optimization of hyperparameters, such as the structural parameters and learning rate of neural networks. The optimization of hyperparameters has always been a difficult point in deep learning, such as the number of channels and the size of the convolution kernel in equation (2); in addition, there are also choices for weight initialization methods, regularization methods, and different training methods. Setting these parameters requires rich training experience, professional knowledge, and a large number of experiments. Therefore, the TPE algorithm is introduced for adaptive hyperparameter optimization, which is used to quickly select the suitable hyperparameters, and it is more time saving and labor saving compared to manually adjusting the hyperparameters. In addition, the training effect is also better.

It is assumed that $\lambda_1, \lambda_2, \cdots, \lambda_n$ represent the hyperparameters selected in the model; $\Lambda_1, \Lambda_2, \cdots, \Lambda_n$ represent the selection domain of each hyperparameter; then, the hyperparameter selection domain space of the model is defined as $\Lambda = \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n$. When k-fold cross-validation method is used for hyperparameter $\lambda \in \Lambda$, the optimization problem of hyperparameters can be expressed as the follows:

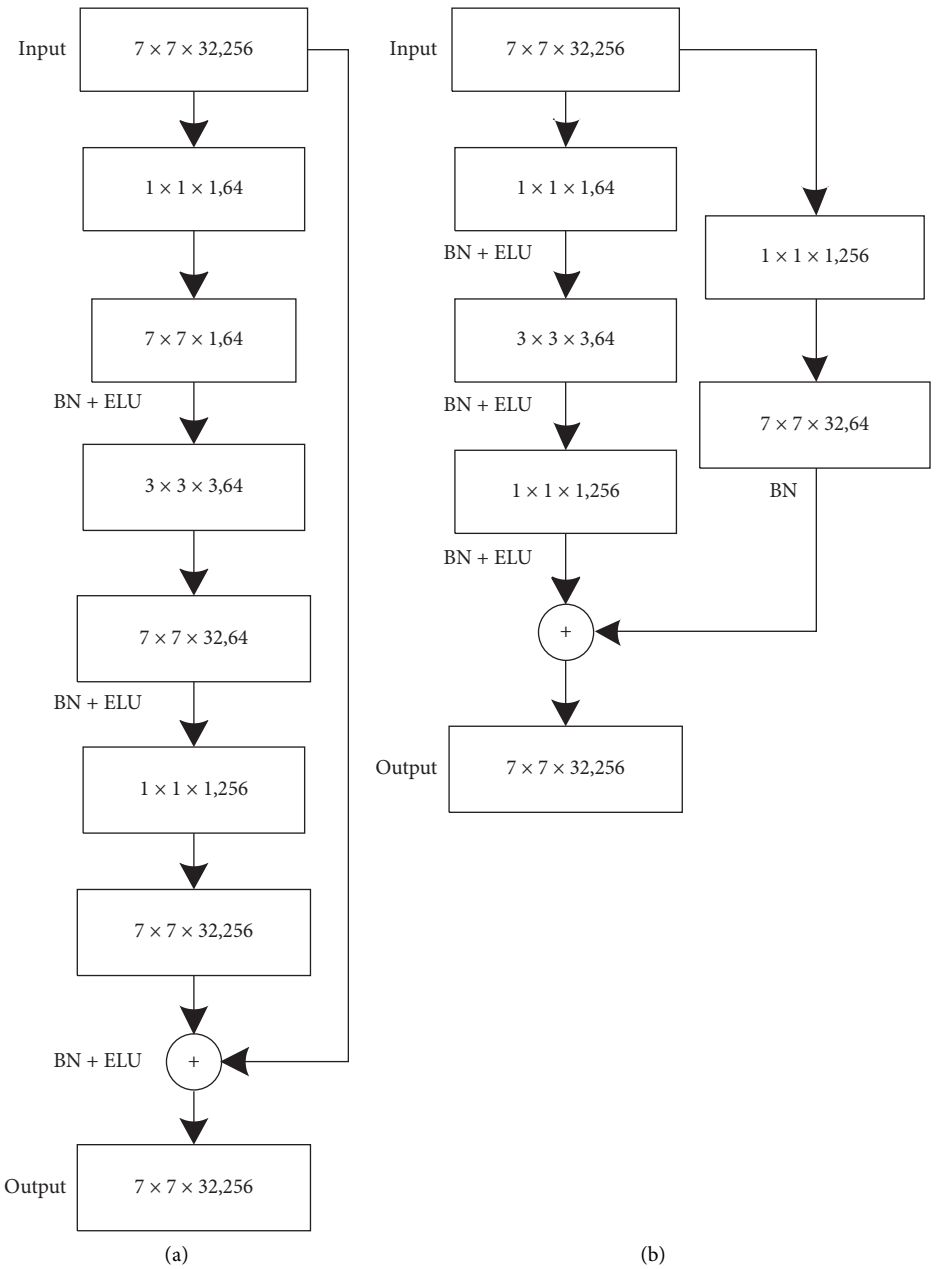


Figure 4: Residual convolutional structure. (a) Identity residual block; (b) convolutional residual block.

$$f(\lambda) = \frac{1}{k}\sum_{i=1}^{k} L\left(\lambda, D^{i}_{train}, D^{i}_{validation}\right), \tag{3}$$

where $L(\cdot)$ is the loss function in training and D_{train} and $D_{validation}$ are denoted as samples in the training set and validation set, respectively.

Recently, the most commonly used hyperparameter optimization methods are still manual search and grid search (violent search), but their efficiency is extremely low, so hyperparameter optimization has always been a very tedious process.

The TPE algorithm is a sequential model-based global optimization algorithm (Smoa). The Smoa algorithm uses the previous hyperparameters to recommend the next hyperparameters through optimization criteria. Different Smoa algorithms use different optimization criteria. TPE algorithm takes expected improvement (EI) as optimization criterion. After each iteration, the algorithm returns the hyperparameter selection of the best EI. In this way, by continuously recommending hyperparameters with the best EI standard, the algorithm can find the optimal hyperparameter faster than grid search. Compared to the random forest algorithm, TPE adopts 2 probability distributions to simulate the posterior probability, which has better modeling strategies and advantages in hyperparameter optimization. The types of hyperparameters can be integers and

continuous real numbers, for example, the number of neurons uses integers and the dropout ratio uses continuous real numbers, and the optimization method of the classifier can use SGD, RMSProp, Adam, etc.

4.1. Structure of Our Model. The network input first is proposed by a convolution layer with the convolution kernel of $l \times l \times 7$ and the step size of $l \times l \times 2$ and a maximum pooling layer with the kernel of $l \times l \times 3$ and the step size of $l \times l \times 2$. The purpose is to reduce the number of channels and improve the operation efficiency. Then, two groups of residual structural units are designed, where each unit is composed of two convolutional residual structural blocks. The first group of residual structural unit is set to $l \times l \times 3$, whose purpose is to extract and fuse spectral features; the second group of residual structural unit is set to $3 \times 3 \times 3$, which is used to extract the spectral features of the hyperspectral image. Finally, a $7 \times 7 \times 1$ global pooling layer and a fully connected layer (FC) are used for classification; each hidden layer uses the strategy in the literature [28] to initialize the convolution kernel parameters and regularize the specification term L_2 of 1×10^{-4}. The activation function is expressed as the exponential linear unit, and the Adam optimizer is selected to train our model in experiment.

4.2. Training Process and Algorithm Framework. According to the structure of hyperparameters which is manually initialized, a search space of hyperparameters is defined for automatic adjustment. There are nearly 10,000 possibilities in the search space. The algorithm and TPE algorithm use the same dataset to search for 50 iterations. 100 epochs are used in training operation. Finally, their recognition accuracy rate is obtained. The hyperparameter with the highest accuracy rate is selected as the hyperparameter of the network.

In this paper, the Softmax layer is used as a classifier. Because it is superior to other classifiers such as support vector machine (SVM) when dealing with multiclassification problems, it has a wide application in deep learning. Its function is defined as follows:

$$S_i = \frac{e^{V_1}}{\sum_i^c e^{V_1}}, \tag{4}$$

where V_i is the output value of classifier in class i; c is the number of class; and S_i is the relative probability.

The algorithm calculates the relative probability for the output value of each class, and the class with the highest relative probability is the classification results.

For pixel-level classification in hyperspectral images, the overall steps can be divided into 3 steps:

Step 1: a patch region with a size of $7 \times 7 \times L$ from the hyperspectral image is extracted as the network input, and the class label of the central pixel is extracted as the object class, where L is the number of channels of the original hyperspectral image.

Step 2: the basic structure of feature extraction is our improved 3D residual convolution structure, and its schematic diagram is shown in Figure 4. The TPE algorithm is adopted to optimize hyperparameters, which can realize end-to-end hyperspectral "spectrum" feature extraction.

Step 3: the network is trained using crossentropy loss and backpropagation; finally, the detection and classification results are obtained. The Softmax layer turns the output of the deep network into a probability distribution, where the distance between the predicted probability distribution and the real probability distribution can be calculated by crossentropy.

5. Experiment Results and Analysis

5.1. Hyperspectral Curve Analysis. As we all know, there is noise interference between the mildew region and the sound region of blueberry in the wavelength range of 400–450 nm. In order to not affect the accuracy of subsequent detection, the spectral data of this waveband range is removed. In addition, the spectral reflectance of the blueberry mildew area in the visible band (450–760 nm) is slightly higher than that of the sound area. In the near infrared band (760–1000 nm), the spectral reflectance of the sound region is higher than that of the mildew region [30]. The reason for the difference of spectral reflectance between the blueberry mildew area and sound area is that the color of the blueberry mildew area is slightly different from that of sound area, and the main components and physical and chemical properties of the blueberry mildew area are changed due to the decay of blueberry disease so that the spectral reflectance is changed. Therefore, the spectral data of 450–1000 nm range were used to establish a training and testing dataset so as to detect the mildewed blueberry. The Hyperspectral Imaging System is used to collect spectral images and is shown in Figure 5.

5.2. Dataset. Training a deep learning network requires a large number of image samples, but the collected blueberry data is often insufficient in practical application. In order to obtain more data so that the deep learning model has strong generalization ability, the obtained blueberry hyperspectral images are expanded. The MATLAB software was used to perform angle rotation, scale transformation, mirror transformation, and adding noise to expand the number of obtained images. Finally, the image is reshaped to the same size 256×256. These images are divided into the training set and the testing set, whose number is shown in Table 1.

5.3. Parameter Setting. The network parameter settings proposed in this paper are as follows: depth = 40, growth_rate = 12, bottleneck = True, reduction = 0.5, batch size is set to 16, learning rate is set to 0.001, and maximum number of iterations is set to 10,000 times; in order to improve optimization efficiency, the ADAMDAM optimization algorithm is adopted. This optimization method is performed using an improved stochastic gradient descent

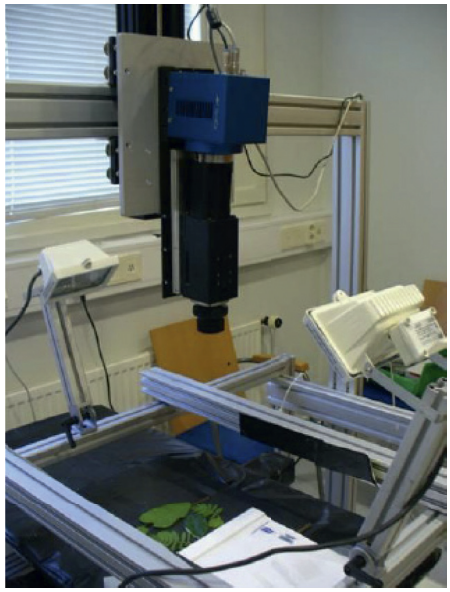


Figure 5: Hyperspectral Imaging System for blueberry classification.

Table 1: Dataset of blueberry on different situations.

	Sound	Slight-decayed	Moderate-decayed	Severe-decayed
Training	15820	7821	3575	6951
Testing	526	425	358	398
Total	16346	8246	3933	7349

algorithm, which can iteratively update the neural network weights based on the training data.

The input of the network is a 3D data matrix with the size of $S \times S \times L$, where L is the number of hyperspectral image channels and S is the field of view. The computation complexity has very close relation with the size of fields of view, so its size needs to be further experimentally determined.

In order to verify the generalization ability of our algorithm, all datasets are divided into three parts: dataset 1, dataset 2, and dataset 3. Figure 6 shows the accuracy of running 10 epochs on different blueberry hyperspectral datasets with different input sizes S. It can be found that the larger the input size, the faster the accuracy rate of the algorithm rises before 3 epochs and the faster the model can converge. The time taken to train 10 epochs and the time spent on testing with different input sizes are shown in Figure 6. It can be found that the larger the size of input, the longer the training time spends. Since the input of larger size converges faster than the input of smaller size, it also requires more training and testing time. Therefore, according the tradeoff between the recognition performance and calculation efficiency, the input size of the hyperspectral image is fixed as to $7 \times 7 \times L$.

5.4. Quantitative Evaluation Indexes. In order to evaluate the performance of our proposed model, the FPPI is adopted as evaluation standard, which focuses on the frequency of occurrence of FP (False Positive). For the mildew detection rectangle obtained for each image, the evaluation criteria used in this paper are Detection Rate (DR) and False Positive Per Image (FPPI), and the relationship is as follows:

$$\mathrm{DR} = \frac{\mathrm{TP}}{(TP + FN)}, \tag{5}$$

$$\mathrm{FPPI} = \frac{FP}{(FP + TN)}, \tag{6}$$

where TP represents the number of positive samples detected correctly; TP + FN represents the number of all positive samples in the picture; and FP + TN represents the number of false positives. In addition, overall accuracy (OA), average accuracy (AA), and kappa coefficient(K) are also selected as quantitative evaluation indexes [31].

5.5. Qualitative and Quantitative Comparison Analysis. In order to better verify the performance of our proposed algorithm in this paper, AlexNet [32], GoogleNet [33], 3D-CNN [34], and ResNet [35] are selected as comparison models; the accuracy of the four algorithms is given from the corresponding paper and open source, and the accuracy is provided by 5 independent tests. The overall classification accuracy is the ratio between the prediction accuracy and the total number on all test sets. The average classification accuracy is the ratio between the correct prediction of each class and the total number of each class, and finally the average value of all class accuracy is taken; kappa coefficient represents the proportion of error reduction, and its calculation is based on the confusion matrix.

In this paper, the neural network AlexNet has 8 layers; the first 5 layers of the convolution layer extract the image features and use the pooling layer to reduce the dimension of the image features; multiple convolutions make the image features become more abstract from the concrete, which can better characterize hyperspectral images. As shown in Table 2, with the increase of the number of iterations, the accuracy of the network has been increasing to 100%. In fact, due to the lack of hyperspectral blueberry image, there is an overfitting situation in the process of training. The overfitting will cause all moderate-decayed blueberries to be classified into severe-decayed blueberries when the trained network is adopted to classify sound, slight-decayed, moderate-decayed, and severe-decayed blueberries.

When the number of iterations of the network reaches 200, the fitness of the training model is not very high. When classifying the blueberry hyperspectral images, the network cannot classify the blueberry correctly. When the sound blueberry hyperspectral images are input into the network for recognition after the network training is completed, more than 50% blueberries are classified as sound and more than 40% are classified as decayed blueberries, but the sound probability is greater than the decayed probability, so it can be judged as sound conditions, and the purpose of accurate classification can be achieved.

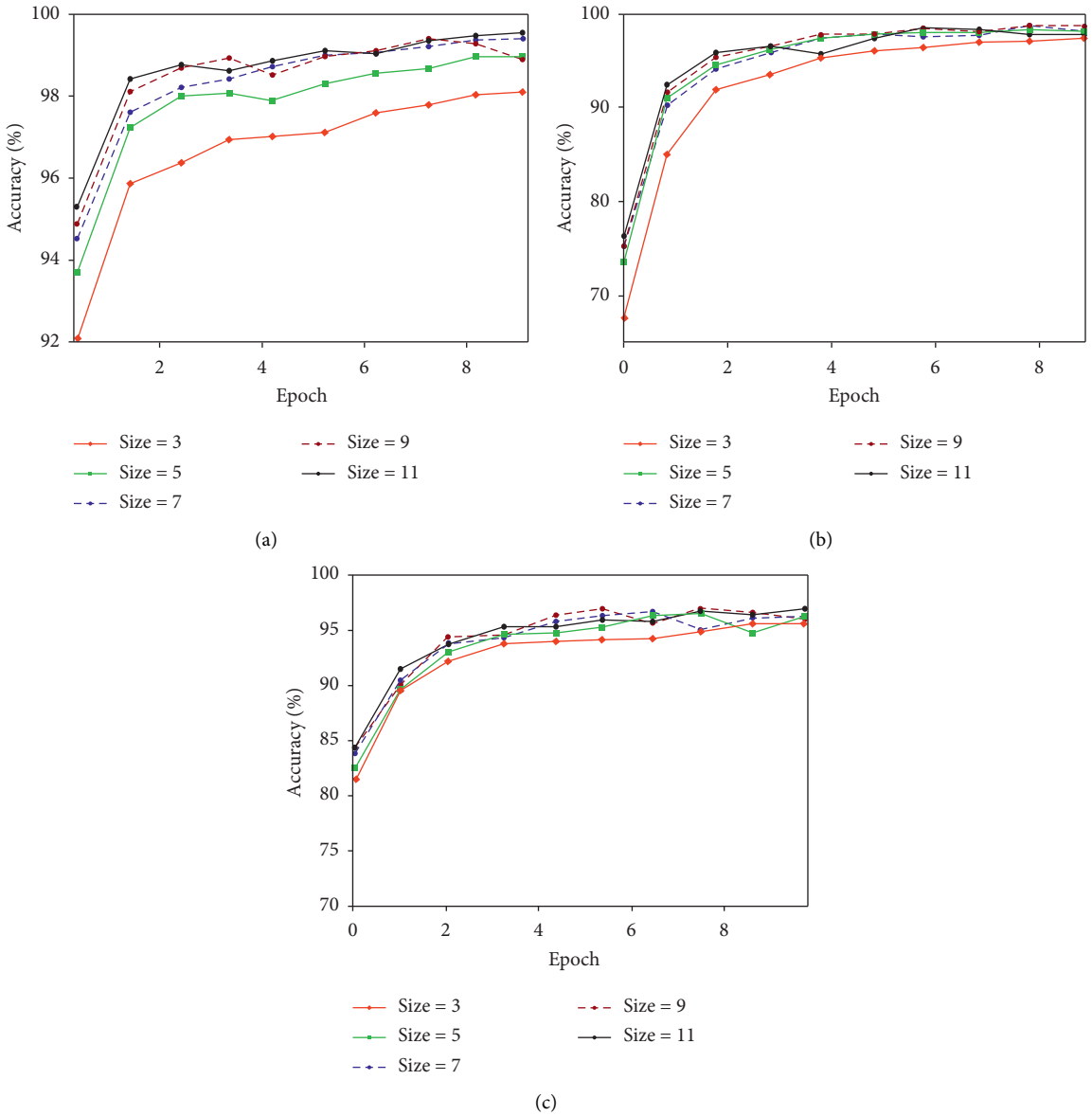


Figure 6: Accuracy of different epochs on different datasets. (a) Dataset 1; (b) dataset 2; (c) dataset 3.

CaffeNet also has 8 layers. The output of each layer is the input of the next layer. The data format has four dimensions in each spectral layer; the first dimension is the number of images, the second dimension is the number of channels, and the third and fourth dimensions are the width and height of images. In deep learning, loss function is often nonconvex, and there is no analytical solution, which needs to be solved by the optimization method. In this paper, the forward algorithm and backward algorithm are called alternately to update the parameters so as to reduce the loss value as much as possible and finally get the local optimal solution. In the process of network iteration, 10-fold crossvalidation is used to verify the performance. It can be seen from this that the accuracy of the network increases rapidly, and the network tends to converge in the process of training and finally reaches 100%. However, due to the lack of data, the increasing number of iterations will lead to overfitting. Because of the huge parameters of the network in the process of overfitting, the data fitting results of the training set are good, but the prediction results of the samples outside the dataset are very poor, where there is a great probability of classification errors. ResNet uses the residual neural network to perform nondestructive detection of blueberries. The detection accuracy rate is up to 90%, and the effect is better. The texture features of the sound blueberry image are obviously different with moderate-decayed and severe-decayed blueberries. It is easy to identify the mildew blueberries using ResNet technology, and the detection effect on the slight-decayed blueberry is poor. The proposed model in this paper is an improved 3D-CNN method for nondestructive detection of blueberries, and its four types of blueberries have better classification performance. Table 3 shows the accuracy under different comparison models. Our proposed algorithm obtains the

Table 2: Classification accuracy under different number of iterations.

Number of iteration	Accuracy	Prediction probability
100	0.81	0.84
200	0.87	0.90
300	0.92	0.95
400	0.93	0.96
500	0.95	0.98
1000	0.95	1
1500	0.99	1
2000	0.97	1
3000	0.95	1

Table 3: Accuracy under different comparison models.

Class	AlexNet	GoogleNet	3D-CNN	ResNet	Proposed
Sound	85.3	83.4	82.7	97.2	98.25
Slight-decayed	96.3	94.2	89.7	83.1	95.1
Moderate-decayed	69.6	70.4	62.0	84.4	93.6
Severe-decayed	58.0	66.9	65.5	84.4	89.4
OA (%)	77.8	81.9	80.6	85.6	95.2
AA (%)	61.3	64.2	68.3	79.8	91.5
K (%)	74.5	79.3	77.9	84.6	94.6

best classification results, which is 17.2%, 20.2%, and 19.8% higher than GoogleNet in OA, AA, and kappa coefficients, respectively. Compared with GoogleNet, our proposed algorithm greatly improves the classification accuracy. Compared with ResNet, our indicators increased by 13%, 14.4%, and 9.6%, respectively. In other words, our proposed algorithm has the best OA, AA, and kappa coefficients.

In order to analyze the blueberry mold recognition performance of the algorithm proposed in this paper, Figure 7 shows the relationship between the detection rate and FPPI (False Positives per Image). Table 4 is the prediction probability of the blueberries in testing set, which is verified by different models. It can be seen from the experimental results that there is an overfitting situation in GoogleNet and AlexNet. Because the GoogleNet network reaches 22 layers, it can learn a lot of features at the same time, but the amount of training samples in this experiment is relatively small, which also leads to overfitting during network training. Both the ResNet network and the proposed network can accurately identify the decay of blueberry hyperspectral images, but the accuracy of ResNet is not as good as the proposed algorithm in this paper. When FPPI = 1, the detection rate of the proposed detection algorithm is 96.69%, and the best result of the comparison algorithm is the ResNet algorithm, the result is 95.42%, while the detection rates of the GoogleNet, AlexNet, and 3D-CNN are 89.12%, 91.88%, and 92.15%.

5.6. Generalization Performance. This paper tests the classification effect of the trained network on different datasets to verify the generalization ability of the model.

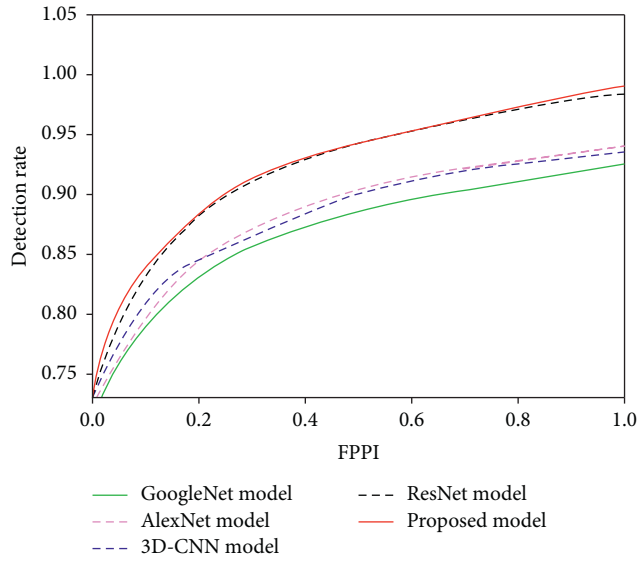


Figure 7: Performance comparison in the database.

This paper uses the model trained on blueberry dataset 1 to classify dataset 2, and dataset 3, respectively. The classification layer is different, so the transfer training method is used to replace the classification part of the network model and fine-tune. The parameters of other parts of the network are not updated. The dataset is still divided into 20% training, 10% verification, and 70% test samples. Experiment results are shown in Table 5. It can be found that the hyperspectral classification model has a high accuracy rate for blueberries, which proves that its "spatial spectrum" feature extraction part has a certain generalization ability.

Table 4: Detection rate of different models when FPPI = 1.

Models	AlexNet (%)	GoogleNet (%)	3D-CNN (%)	ResNet (%)	Proposed (%)
Detection rate	89.12	91.88	92.15	95.42	96.69

Table 5: Comparison for generalization ability of the model.

Indexes	Dataset		
	1	2	3
OA (%)	98.37	96.35	95.31
AA (%)	97.85	92.61	92.09
K (%)	98.29	96.25	95.22

6. Conclusions

An improved deep residual 3D convolutional neural network (3D-CNN) framework is proposed for hyperspectral images classification so as to realize fast training, classification, and parameter optimization. Rich spectral and spatial features can be rapidly extracted from samples of complete hyperspectral images using our proposed network. This combines the tree structured Parzen estimator (TPE) adaptively and selects the super parameters to optimize the network performance. In addition, aiming at the problem of few samples, this paper proposes a novel strategy to enhance the hyperspectral image sample data, which can improve the training effect. Experimental results on the standard hyperspectral blueberries datasets show that the proposed framework improves the classification accuracy compared with AlexNet and GoogleNet. In addition, our proposed network reduces the number of parameters by half and the training time by about 10%.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Scientific Research Foundation of Inner Mongolia University for Nationalities" (no. NMDYB18023); Scientific Research Foundation of Inner Mongolia University for Nationalities" (no. NMDYB19037); Higher Education Science Research Project of Inner Mongolia Autonomous Region of China (no. NJZY19155); Higher Education Science Research Project of Inner Mongolia Autonomous Region of China (no. NJZY18160); CERNET Innovation Project (no. NGIINGII20170612); and Science Research Project of Inner Mongolia University for the Nationalities (no. NMDGP1706).

References

[1] Z. Wang, M. Hu, and G. Zhai, "Application of deep learning architectures for accurate and rapid detection of internal mechanical damage of blueberry using hyperspectral transmittance data," *Sensors*, vol. 18, no. 4, pp. 1126–1135, 2018.

[2] R. Lab, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2009.

[3] D. Lorente, "Selection of optimal wavelength features for decay detection in citrus fruit using the ROC curve and neural networks," *Food and Bioprocess Technology*, vol. 6, no. 2, pp. 530–541, 2013.

[4] S. Bargoti and J. P. Underwood, "Image segmentation for fruit detection and yield estimation in Apple orchards," *Journal of Field Robotics*, vol. 34, no. 6, pp. 1039–1060, 2017.

[5] C. Nandi, B. Tudu, and C. Koley, "A machine vision technique for grading of harvested mangoes based on maturity and quality," *IEEE Sensors Journal*, vol. 34, p. 1, 2016.

[6] J. D. Tygar, "Adversarial machine learning. " internet computing," *IEEE*, vol. 15, no. 5, pp. 4–6, 2011.

[7] N. El-Bendary, "Using machine learning techniques for evaluating tomato ripeness," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1892–1905, 2015.

[8] O. Gupta, "Machine learning approaches for large scale classification of produce," *Scientific Reports*, vol. 8, no. 1, p. 5226, 2018.

[9] M. Rahnemoonfar and C. Sheppard, "Deep count: fruit counting based on deep simulated learning," *Sensors*, vol. 17, no. 4, p. 905, 2017.

[10] M. S. Hossain, M. H. Al-Hammadi, and G. Muhammad, "Automatic fruits classification using deep learning for industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 45, p. 1, 2018.

[11] K. S. Lewers, Y. Luo, and B. T. Vinyard, "Evaluating strawberry breeding selections for postharvest fruit decay," *Euphytica*, vol. 186, no. 2, pp. 539–555, 2012.

[12] S. Cao, "Effect of ultrasound treatment on fruit decay and quality maintenance in strawberry after harvest," *Food Control*, vol. 21, no. 4, pp. 0–532, 2010.

[13] D. Lorente, "Comparison of ROC feature selection method for the detection of decay in citrus fruit using hyperspectral images," *Food & Bioprocess Technology*, vol. 6, no. 12, pp. 3613–3619, 2013.

[14] B. Wang, J. Xue, and S. Zhang, "Detection of decay and disease pear jujube based on hyperspectral imaging technology," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 7, no. 6, pp. 2094–2107, 2013.

[15] P. Liu, H. Zhang, and K. B Eom, "Active deep learning for classification of hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 45, pp. 1–13, 2013.

[16] X. Zhou, "Deep learning with grouped features for spatial spectral classification of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 1, pp. 97–101, 2017.

[17] X. Yang, "Hyperspectral image classification with deep learning models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 99, pp. 1–16, 2018.

[18] H. Huang, "Spatial-spectral feature extraction of hyperspectral image based on deep learning," *Laser & Optoelectronics Progress*, vol. 54, no. 10, p. 101001, 2017.

[19] S. Liu and Q. Shi, "Multitask deep learning with spectral knowledge for hyperspectral image classification," 2019.

[20] D. Zhang, "Hyperspectral image classification using spatial and edge features based on deep learning," *International Journal of Pattern Recognition & Artificial Intelligence*, vol. 45, 2018.
[21] Y. Chen, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.
[22] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 54, no. 8, pp. 4544–4554, 2016.
[23] J. Yue, S. Mao, and M. Li, "A deep learning framework for hyperspectral image classification using spatial pyramid pooling," *Remote Sensing Letters*, vol. 7, pp. 7–9, 2016.
[24] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 99, p. 1, 2017.
[25] W. Zhao, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *International Journal of Remote Sensing*, vol. 36, no. 13, pp. 3368–3379, 2015.
[26] R. Venkatesan and S. Prabu, "Hyperspectral image features classification using deep learning recurrent neural networks," *Journal of Medical Systems*, vol. 43, p. 7, 2019.
[27] F. Bei, "Semi-supervised deep learning classification for hyperspectral image based on dual-strategy sample selection," *Remote Sensing*, vol. 10, no. 4, pp. 574–585, 2018.
[28] C. Shi and S. Pun, "Multi-scale hierarchical recurrent neural networks for hyperspectral image classification," *Neurocomputing*, vol. 10, 2018.
[29] D. Adam, "Bull. "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, vol. 12, no. 10, pp. 3-4, 2011.
[30] S. Yang, "Fuzzy signature-based discriminative subspace projection for hyperspectral data classification," *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, vol. 9, pp. 4196–4202, 2011.
[31] E. Xie, "Unsupervised hyperspectral feature selection based on fuzzy c-means and grey wolf optimizer," *International Journal of Remote Sensing*, vol. 12, 2019.
[32] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, p. 2, 2012.
[33] C. Szegedy, "Going deeper with convolutions," 2014.
[34] Y. Li, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognition*, vol. 80, 2016.
[35] M. E. Paoletti, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 740–754, 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8861987, 10 pages
https://doi.org/10.1155/2020/8861987

Research Article

Face Detection and Recognition Based on Visual Attention Mechanism Guidance Model in Unrestricted Posture

Zhenguo Yuan

School of Mechanical and Electrical Engineering, Guangdong Industry Polytechnic, Guangzhou, China

Correspondence should be addressed to Zhenguo Yuan; 2016001053@gdip.edu.cn

Received 14 April 2020; Revised 28 April 2020; Accepted 6 May 2020; Published 20 May 2020

Academic Editor: Chenxi Huang

Performance of face detection and recognition is affected and damaged because occlusion often leads to missed detection. To reduce the recognition accuracy caused by facial occlusion and enhance the accuracy of face detection, a visual attention mechanism guidance model is proposed in this paper, which uses the visual attention mechanism to guide the model highlight the visible area of the occluded face; the face detection problem is simplified into the high-level semantic feature detection problem through the improved analytical network, and the location and scale of the face are predicted by the activation map to avoid additional parameter settings. A large number of simulation experiment results show that our proposed method is superior to other comparison algorithms for the accuracy of occlusion face detection and recognition on the face database. In addition, our proposed method achieves a better balance between detection accuracy and speed, which can be used in the field of security surveillance.

1. Introduction

There are still some challenging problems in face detection and recognition technology mainly due to the nonrigid features and the influence of complex background [1–3]. Traditional face detection algorithms mostly use a semi-supervised learning method. Since the traditional method needs to design different artificial features for different tasks, such as grayscale features, contour features, and HOG features, these features are easily affected by the imaging angle, and the generalization ability is poor. At the same time, the object occlusion will also lead to the missed detection, thereby reducing the accuracy of the detector. Therefore, it is of great practical significance to study the occlusion problem for face detection and recognition task [4].

The face detection and recognition model based on machine learning is a popular research direction in the field of computer vision [5]. By directly extracting features from the detection area and then using machine learning algorithms to classify and recognize, the accuracy of the model classification can be improved to a certain extent, but the characterization ability of features directly affects the recognition accuracy of the system [6]. Compared with detection and recognition algorithms for shallow learning models such as boosting, decision trees, and neural networks, deep learning represented by convolutional neural networks implements the deep nonlinear network structures through operations such as local receptive fields and weigh sharing. The hierarchical strategy can learn the most essential feature representation in the data set [7]. At present, mainstream deep learning-based face detectors usually adopt a two-stage network structure, which is divided into face detection and face recognition.

Most convolutional neural networks use a classification loss function to measure the difference between the predicted value and the actual value and then complete the classification of the image through the training process to expand the distance between different types of images. Wang et al. [8] used 3-dimensional face information as a feature where the robustness and accuracy of the algorithm are improved through a large amount of data training. Corrow et al. [9] used DeepID for face recognition by partitioning different parts of the face, extracting features separately, and

then using the Bayes algorithm to perform complex operations on the features, and finally obtaining face feature information, effectively improving the accuracy of recognition. However, none of the above algorithms solves the recognition problem under nonlimiting conditions. Therefore, how to increase the distance between classes while reducing the distance of intraclasses in the recognition process is the important topic of the face recognition task. Abbad et al. [10] realized the feedback of the loss function during the training process by adding a loss verification method and used the positive samples to reduce the distance between the classes, but this method is more dependent on the samples. Madhavan and Kumar [11] proposed a ternary loss algorithm that unifies the training data into triple elements; each triple contains positive value, negative value, and sample anchor point, which can effectively reduce the intraclass distance.

Although the above method can solve some nonlimiting problems, it has poor performance in convergence speed, especially when the number of network layers is too large, and vanishing gradient phenomenon will occur. In order to solve this kind of problem, Su et al. [12] proposed a multi-inception structure-based convolutional network neural algorithm for face recognition. By transforming the traditional Softmax loss method and combining Softmax and TripletLoss, a larger interclass distance and a smaller intraclass distance can be obtained. Experiment result proves that the algorithm increases the depth and width of the network, and intraclass spacing can be effectively reduced during the training process. Based on the above description, it can be seen that the convolutional neural network faces different problems when processing different data and application scenarios. Some scenes pay more attention to calculation speed, and some pay more attention to detection accuracy. More scholars strive to find a universal model with high performance in all aspects. This is also the ultimate goal of this study.

Aiming at the problem that occlusion affects the accuracy of face detection and recognition, this paper proposes a deep network with multilevel feature fusion. This network uses a visual attention mechanism to guide the model to highlight the visible area of the occlusion face; the detection recognition problem can be simplified to a high-level semantic feature detection problem, and the position and scale of the face are predicted by means of activation maps, avoiding additional parameter settings. A large number of simulation experiment results show that the proposed method is better than the existing mainstream method in the detection and recognition of the occlusion face on the public data set and has achieved a faster detection speed, which can be used in the field of security surveillance.

The innovations of this study are summarized as follows:

(1) In view of the detection omission caused by occlusion in face detection, a solution is proposed, that is, a visual attention mechanism guidance model is proposed, which uses the visual attention mechanism guidance model to highlight the visible area of the blocked face, thereby improving face detection and recognition accuracy.

(2) The new model parameters have been simplified. Through the improved analysis network, the face detection problem is simplified to the advanced semantic feature detection problem, and the position and scale of the face are predicted by the activation map to avoid additional parameter settings.

2. Face Detection Network

The YOLO-V3 network is a better deep learning model in the field of object recognition, the network evolved from the YOLO and YOLO-V2 networks [13]. Compared with the deep learning network based on region proposal, the YOLO network transforms the detection problem into a regression problem. The network does not need to adopt exhaustive candidate regions but directly generates the confidence and bounding box coordinates of the object through regression. Compared with the Faster-RCNN network, the detection speed is greatly improved [14].

The YOLO detection model is shown in Figure 1. The network divides each image in the training set into an $S \times S$ ($S = 13$) grid. If the center of the real object falls into the grid, the grid is responsible for detecting the category of the object. Multiple bounding boxes are predicted in each grid, and each predicted bounding box is scored to demonstrate that the bounding box completely contains the confidence of the object, which is defined as follows:

$$C = P_r(\text{object}) \times \text{IoU}_{\text{pred}}^{\text{truth}}, \quad P_r(\text{object} \in \{0, 1\}), \tag{1}$$

where $P_r(\text{object})$ indicates the probability of the object contained in the bounding box. If there is an object in the bounding box, we have $P_r(\text{object}) = 1$; otherwise $P_r(\text{object}) = 0$; $\text{IoU}_{\text{pred}}^{\text{truth}}$ indicates the Intersection over Union (IoU) between the prediction result and the benchmark frame. The confidence reflects whether the grid contains objects and the accuracy of prediction boundary box. When multiple bounding boxes detect the same object, YOLO uses a nonmaximum suppression method to select the best bounding box.

Although YOLO has obtained a faster detection speed, its detection accuracy is not as good as Faster R-CNN. In order to solve this problem, YOLO-V2 introduces the idea of the anchor mechanism in the Faster R-CNN network and uses the k-means clustering method or fuzzy c-means method [15] to generate a suitable prior bounding box. Therefore, the number of anchor boxes required by the YOLO-V2 algorithm to achieve the same IoU is reduced. YOLO-V2 improves the network structure and replaces the fully connected layer in the YOLO output layer with a convolutional layer [16]. In addition, YOLO-V2 also introduces batch normalization, dimensional clustering, fine-grained features, multiscale training, and other strategies; compared with YOLO, YOLO-V2 greatly improves the detection accuracy. YOLO-V3 is an improved model based on YOLO-V2. By using multiscale prediction to detect the

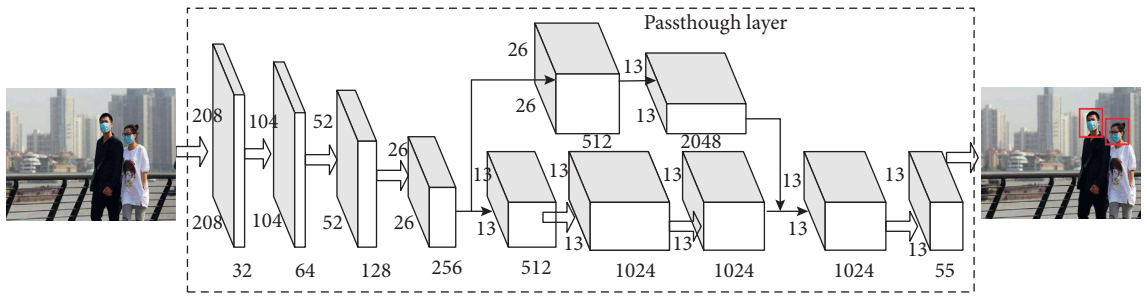


Figure 1: Face detection and recognition process.

final object, its network structure is more complicated than YOLO-V2. YOLO-V3 can predict the bounding boxes of different scales, which can detect small objects more effectively than YOLO-V2, but there are still missing detections for partially occluded face objects [17].

3. Occlusion Face Detection and Recognition Algorithm Combined with the Visual Attention Mechanism

Figure 2 is our proposed face detection and recognition model in this paper, which consists of two parts: feature extraction network and face analysis network. The input image is extracted by the feature extraction network to extract high-level semantic features, and the feature-guided attention module is used for feature fusion; the face analysis network predicts the face position, height, and offset heat map on the basis of the obtained high-level semantic feature and obtains the face boundary box.

3.1. Feature Extraction Network. The feature extraction network is designed on the basis of feature pyramid networks (FPN) [18–21], including basic networks and visual attention networks. ResNet50 has excellent performance in visual tasks such as image classification, so it is used as the basic network. ResNet50 can be divided into 5 levels, and the downsampling rate of each level relative to the input image is i; {1, 2, 3, 4, 5} represents the number of levels. In order to make full use of the location information of the shallow feature map and the semantic information of the deep feature, the shallow and deep feature maps are used to guide the attention network for feature fusion. The process can be described as follows: firstly, the number of C3 and C4 feature channels can be reduced to 256 by the convolutional layer with the size of the 1×1 convolution kernel, which can reduce the amount of calculation; then, the backbone network feature maps (namely, P3 and P4) after bilinear interpolation and upsampling 2 times are, respectively, input into the guided attention module to feature fusion [22–24].

3.2. Visual Attention Network. Different feature channels of the convolutional network will have different responses to a specific area of the face, which is to say that the occlusion form of the object can be described through different feature channels, and the occlusion form $O(n)$ is defined as the following equation:

$$o(n) = [v_0, p_0, v_1, p_1, \ldots, v_k, p_k], \tag{2}$$

where p_i denote different areas of face objects and $v_i \in \{0, 1\}, i \in [0, k]$ is used to indicate whether the partial area i of the face is visible.

The weights of traditional CNN channels are usually fixed and the same, which limits the network's ability to express different occlusion forms [19, 21, 23–26]. Patil et al. [27] recalibrated the weight of each channel so that the feature channel expressing the visible area of the occlusion object has a greater contribution to the final convolution feature, which can highlight the occlusion object in the background. The channel weighting process can be expressed as the following equation:

$$F_o(n) = \mathbf{\Omega}_n \mathbf{F}_c, \tag{3}$$

where $\mathbf{F}_c$ is the channel feature and $\mathbf{\Omega}_n$ is the channel weighting vector corresponding to the occlusion form n. The visual attention module is to get the attention vector $\mathbf{\Omega}_n$ through learning and finally achieve the reweighting of the feature channel for $\mathbf{\Omega}_n$ so that the network can adaptively express different occlusion forms. However, the existing models only consider the relationship between channels and ignore the importance of spatial information for the feature map. Because the spatial information of the feature map is helpful for the network to locate the region of interest, the feature channel attention mechanism and the spatial attention mechanism are used in the feature description task in literature [28]. Similarly, spatial attention mechanism is applied to object detection tasks in literature [29], so as to guide the network to highlight the useful features for current tasks. On the basis of the above description and analysis, the feature space information is used in the face detection and recognition task to highlight the occlusion of the face object area, and the spatial attention module is constructed to achieve the face detection and recognition task. The spatial attention module obtains the spatial attention map from the spatial information of the statistical feature map, which is used to reactivate the input features, so as to guide the network to focus on the occluded face and suppress the background interference.

As shown in Figure 3, the visual attention network consists of two submodules: channel and spatial attention. The input of the visual attention network is two feature maps

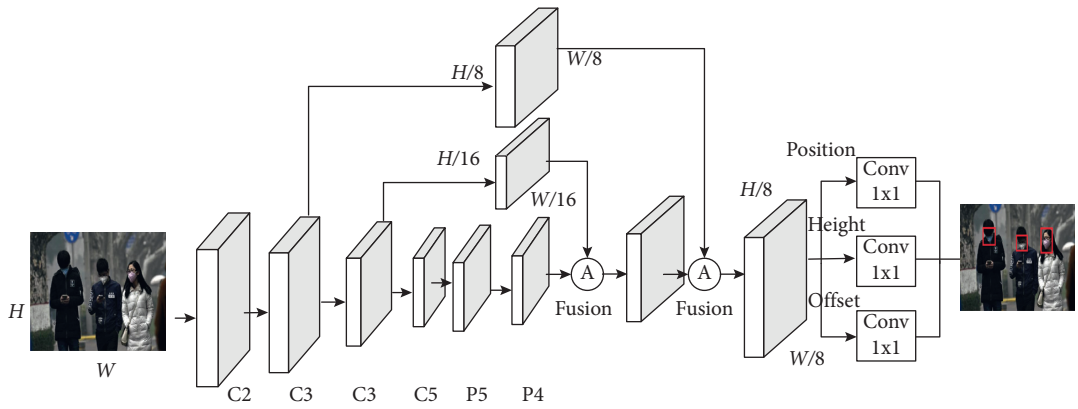


Figure 2: Model framework.

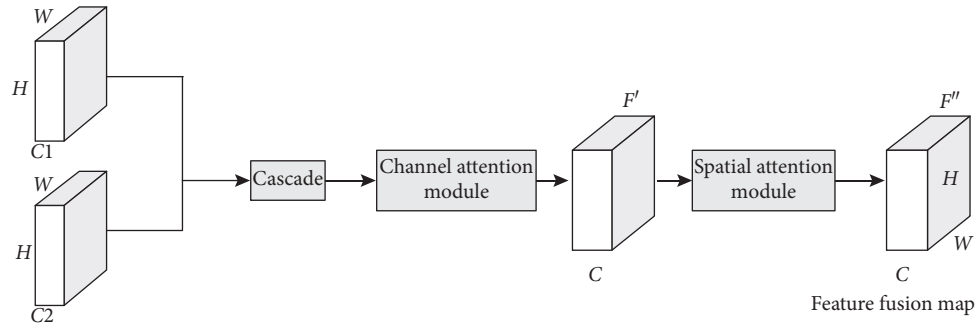


Figure 3: The overall structure of the visual attention module.

(such as C4 and P4) from the shallow convolution layer and the deep convolution layer, respectively. Firstly, the input features are connected in the channel dimension to get $F \in R^{H\times W\times C}$, and then F is input to channel attention module. After a series of operations, the spatial attention module is to achieve the feature fusion fusion. Therefore, by using the attention module to model the correlation between feature channels and the spatial information of the feature map, the network can not only enhance the feature representation of the relevant areas but also obtain the location information of the area of interest [29]. While making full use of the useful features to deal with the problem of face occlusion, it also suppresses the useless clutter information, which is conducive to improving the accuracy of face and recognition detection.

Figure 4 is the feature channel attention module proposed in this paper. For the input feature map F, the global information of each feature channel is obtained by global average pooling and maximum pooling operations to form the channel descriptor z_c^{avg} and $z_c^{\max}$, and then the feature channel attention vector $\Omega_c \in R^{1\times 1\times C}$ is obtained through the two fully connected layers FC1 and FC2. Finally, the deep learning method makes the network to automatically characterize the occlusion form of different samples. The specific steps are shown in equation (3).

$$\Omega_c = \sigma\left(W_2\left(\delta\left(W_1 z_c^{\text{avg}}\right)\right) + W_2\left(\delta\left(W_1 z_c^{\max}\right)\right)\right), \tag{4}$$

where σ and δ are sigmoid function and ReLU function, respectively. $W_1 \in R^{\text{Clr}\times C}$ and $W_2 \in R^{C\times \text{Clr}}$ represent two fully connected layer parameters, where r is the ratio of downsampling dimensionality reduction. $\Omega_c \in R^{1\times 1\times C}$ is used to weight the input feature F channel by channel to obtain F'. The process can be written as the following equation:

$$F' = \Omega_s \otimes F', \tag{5}$$

where $\otimes$ represents the dot-product channel by channel.

Since the useful information for partial-occluding face objects is usually obscured by the background, the network also needs to determine the spatial location of the useful information while enhancing the feature expression of the occlusion object through the channel attention module. Unlike the channel attention mechanism, the spatial attention mechanism is mainly used to highlight the areas in the feature map that are related to the current task, which is to guide the network to focus on the visible area of the occlusion object [30].

In the spatial attention module, the maximum pooling operation is firstly performed on the input feature map F' in the channel dimension to obtain the feature map $F_{\max}{}' \in R^{H\times W\times 1}$, which is used to count the spatial information of the feature map; then, the feature map is input into a 3×3 convolution layer f_c and output by the sigmoid function to obtain the spatial attention map $M_s \in R^{H\times W\times 1}$:

$$M_s = \sigma\left(f_c\left(F'_{\max}\right)\right), \tag{6}$$

where σ is the sigmoid function. Finally, the spatial attention map M_s is used to reactivate the input F' to obtain the final feature map F'':

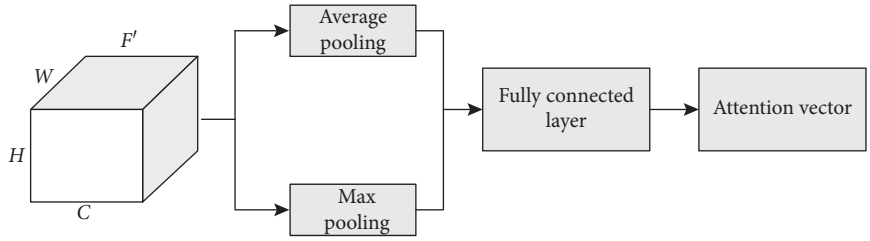


Figure 4: Feature channel attention module structure.

$$F'' = M_s \otimes F', \tag{7}$$

where $\otimes$ represents the dot product between feature maps.

Face detection and recognition task is regarded as a high-level semantic feature detection problem. On the basis of obtaining semantic features, the final prediction bounding box is obtained through the face analysis network. In this paper, the position, height, and position offset of the face are firstly predicted, and the size of the bounding box is obtained by simple geometric transformation, and then, the simple recognition network can get the high-precision recognition effect [31]. Specifically, after the predicted height h of the face is obtained, the width $w = h \cdot \alpha$ of the bounding box can be calculated by the length-width ratio of the bounding box. If the output feature map of the feature extraction network is $F_{\text{final}} \in R^{H/s \times W/s}$, three heat maps are predicted by three parallel 1×1 convolutional layers and correspond to center position $H_c \in R^{H/s \times W/s}$, height $H_h \in R^{H/s \times W/s}$ and position offset $H_{\text{offset}} \in R^{H/s \times W/s}$, respectively. s is the sampling rate of the output activation map relative to the input image. By predicting the heat map, the limitation of the prior frame adopted by the traditional method is avoided, and a more flexible face detection and recognition is realized in the same network.

3.2.1. Position Prediction. Face location prediction is achieved through the location heat map H_c. In this paper, the position prediction problem is simplified as a binary classification problem.

The position of the object center on the feature map F_{final} is (x_c, y_c). And the object center pixel is selected as a positive sample and the other positions as negative samples. The cross-entropy loss function is used to optimize the training position-prediction branch. The training true value H_c^{gt} is generated by a 2D Gaussian function, and the truth value at any position can be obtained by calculating equation (8):

$$H_c^{gt}(i, j) = \max\left(G\left(i, j; x_c, y_c, \sigma_w, \sigma_h\right)\right), \tag{8}$$

$$G\left(i, j, x_c, y_c, \sigma_w, \sigma_h\right) = e^{-\left(\left(\left(i - x_c\right)^2/\sigma_w^2\right) + \left(\left(j - y_c\right)^2/\sigma_h^2\right)\right)}, \tag{9}$$

where (x_c, y_c) is the central location of the object and σ_w and σ_h are variances of the width and height of the object, respectively. In order to alleviate the imbalance of positive and negative samples in the training process, focal loss was defined as the predicted loss function of the center position:

$$\begin{cases} L_c = \dfrac{-1}{N} \displaystyle\sum_{i=1}^{W/s} \sum_{j=1}^{H/s} BW, & \\ \left(1 - p_{i,j}\right)^{\alpha} \log\left(p_{i,j}\right), & H_c^{gt}(i, j) = 1, \\ \left(1 - H_c^{gt}(i, j)\right)^{\beta} p_{i,j}^{\alpha} \log\left(1 - p_{i,j}\right), & \text{others}, \end{cases} \tag{10}$$

where $p(i, j)$ indicates the prediction score of the object center at (i, j) in the prediction heat map, N is the number of objects in the picture, and α and β are the balance factors, generally set to 2 and 4.

3.2.2. Height Prediction. Given the position of the face k in the height heat map which is (x_k, y_k), its corresponding true value is $H_h^{gt}(x_k, y_k) = \log(h_k)$, where h_k denotes the height of the object k. In this paper, the true value within the radius r of (x_k, y_k) is set to $\log(h_k)$, and the radius r is set according to the width of the object, which is generally set as $r = 0.5w_k$. Our proposed model in this paper uses the L1 loss function for training. The loss function is denoted as follows:

$$L_h = \frac{1}{N} \sum_{k=1}^{N} L1\left(h_k, H_h^{gt}\left(x_k, y_k\right)\right), \tag{11}$$

where h_k is the predicted height of the object k in the heat map and N is the number of objects in the image.

3.2.3. Deviation Prediction. Since the convolutional network is usually a downsampling process, the position (x, y) on the input image is mapped into the heat map, whose position can be expressed as $(x/s, t\,y/s)$, where s is the downsampling rate of the network. When the position on the activation map is remapped back to the input image, an error will be generated, especially affecting the detection and recognition result of the dim-small face. To alleviate this problem, the position prediction of the object is corrected by predicting the deviation/offset of the center position [32], and the corresponding true value can be rewritten as the following equation:

$$H_{\text{offset}}^{gt}\left(x_k, y_k\right) = \left(\frac{x_k}{s} - \left\lfloor \frac{x_k}{s} \right\rfloor, \frac{y_k}{s} - \left\lfloor \frac{y_k}{s} \right\rfloor\right). \tag{12}$$

Finally, the multitask loss function weighted optimization can be adopted to train our proposed network, where the weighted loss function can be denoted as follows:

$$L = \lambda_c L_c + \lambda_h L_h + \lambda_o L_{\text{offset}}, \tag{13}$$

where λ_c, λ_h, and λ_o are weighting factors, which are set to 0.01, 1, and 0.12, respectively。

4. Experimental Results and Analysis

4.1. Experimental Data Set and Parameter Settings. In order to evaluate the performance of the face detection and recognition algorithm based on the visual attention-guided mechanism proposed in this paper, LFW (labeled faces in the wild database), CMUFD database (CMU face detection database) [23], and UCFI database (UCD color face image) [25] are used as face detection and recognition data sets. It consists of 500 images with a resolution of 2048×1024. Since CMUFD contains a large number of partial occlusion face images, it is selected as the verification and comparison test of the proposed method; UCFI contains about 350,000 face samples, where the standard testing set consists of 4,024 images with a resolution of 640×680 in a simple scenario. In order to verify the generalization of our proposed method, some testing experiments are performed on the UCFI data set. Face objects of all training data have been accurately marked. Except for the object area, the rest is marked as background, which means that the labeled data set can be used for training and testing of face detection and recognition models.

The network proposed in this paper selects pretrained ResNet50 as the backbone network. Its parameters are set as follows: depth = 40, growth_rate = 12, bottleneck = True, reduction = 0.5, minibatch is set to 16; learning rate is set to 0.001, dropout parameter is set to 0.8, and the maximum number of iterations is set to 10,000; in order to improve the optimization efficiency, this paper uses the Adam optimization algorithm. The Adam optimization algorithm is an extension of the stochastic gradient descent algorithm, which can iteratively update the neural network weights based on the training data; the initialization of learning rate is set to 0.25; then, when training to the 30th epoch, the learning rate is changed to 0.025. The nonmaximum suppression algorithm (NMS) is used to filter out the redundant face results. The threshold of Intersection over Union (IoU) is set to 0.5, and only the face results with the object confidence score greater than 0.1 are retained [33].

Our proposed network module is based on the PyTorch deep learning framework. The experimental environment is Xeon (Xeon) E7-8890 v2 @ 2.80 GHz (X4), 128 GB (DDR3 1600 MHz), Nvidia GeForce GTX 1080 Ti, Ubuntu 6.04, 64-bit operating system.

4.2. Evaluation Index. At present, the video surveillance intelligent analysis system has been able to detect and recognize the face with different scales. However, existing algorithms have a large number of false detection for face objects under partial occlusion mainly due to incomplete occlusion of face objects and the similarity of the face and background gray. Therefore, the standard evaluation indicators are selected as performance evaluation, which is the false positive per image (FPPI) of each image, focusing on the frequency of occurrence of false positive, as shown in Table 1.

In the detection stage, the evaluation criteria are the detection rate (DR) and the false detection rate (false positive per image, FPPI):

$$\begin{aligned} \text{DR} &= \frac{TP}{(\text{TP} + \text{FN})}, \\ \text{FPPI} &= \frac{FP}{(\text{FP} + \text{TN})}, \end{aligned} \tag{14}$$

where TP represents the number of positive samples detected correctly, $TP + FN$ represents the number of positive samples included the image, and FP represents the number of false positive samples. In addition, we also use the log-average miss rate (MR) to characterize the performance of the detector for the face. This paper mainly focuses on the occlusion situation for face detection and recognition. Therefore, we define the visible range to characterize the occlusion situation of the face. Given that the proportion of the object visible area to the total area is λ, if $\lambda > 0.7$, it means that the object is in a normal state and is denoted as N; if $0.2 < \lambda < 0.7$, it means that the object is in a serious occlusion state and is denoted as H. For ease of analysis, we will also divide the data set into four types of subsets, which are, respectively, recorded as mixed face data set (mixed), bare occlusion face data set (bare), partially occlusion data set (partial), and severely occlusion data set (heavy) [34].

4.3. Performance Analysis for Face Detection. In order to verify the effectiveness of the feature-guided attention network, the detector with the attention network removed is used as a test baseline (baseline), and Face++ proposed in [17] is used as a comparison method. Baseline adopts the feature fusion method consistent with FPN to build the model, where CA means the channel attention module and SA means the space attention module, and the performance of the detector after adding each module is compared in the experiment, and the test results are shown in Table 2.

Compared with the baseline model, the face detector's missed detection (MR) of the occlusion image has a significant decrease after adding the attention module, indicating that our proposed attention mechanism can effectively guide the detector to focus on the occlusion object. Compared with Face++, the proposed method has a significant reduction in MR under all evaluation criteria, especially under the severe occlusion evaluation criteria, and its missed detection (MR) decreased by 18.2%, indicating that the proposed method has good effectiveness for face detection in the complex scenario.

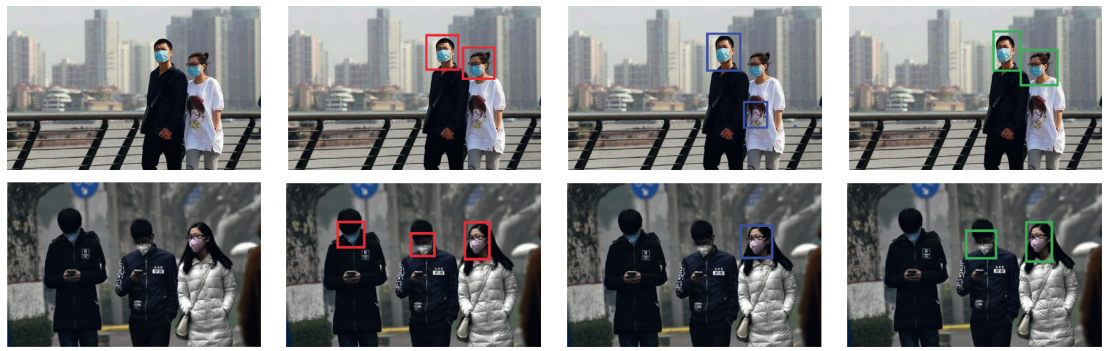
In order to more intuitively understand the attention module on the performance of the face detector, Figure 5 gives visualized face prediction results. Figure 5(a) is a face image, and input it to Baseline, Baseline + CA, and Baseline + CA + SA, respectively, to obtain the position prediction. Through observation, it can be found that the feature response in Figure 5(d) is closer to the face visible

Table 1: Statistics of face detection results.

Results	Benchmark	
	Face (positive)	Background (negative)
Face (positive)	(True positive) TP	(False positive) FP
Background (negative)	(False negative) FN	(True negative) TN

Table 2: Comparative analysis of visual attention network verification results in the face detection module.

Methods	N (%)	H (%)	$N+H$ (%)
Face++	16.11	56.74	38.12
Baseline	12.16	41.05	37.92
Baseline + CA	11.79	39.31	37.83
Baseline + CA + SA	11.53	38.73	37.26



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

Figure 5: Performance comparison for face detection. (a) Face image with occlusion. (b) Baseline + CA + SA. (c) Baseline + SA. (d) Baseline + CA.

area, while there is still background interference in Figures 5(b) and 5(c), but the interference of Figure 5(c) is significantly less than that of Figure 5(b). This also proves that the attention module can guide the network to highlight to the visible part of the occlusion face, while also reducing the impact of background noise on detection performance.

4.4. Performance Analysis for Face Recognition. Since the recognition algorithm based on deep learning has achieved great results in the field of natural image, some classical deep learning algorithms will be used to make a comparison with our proposed algorithm. In order to qualitatively and quantitatively analyze the accuracy of the proposed algorithm for occluded face detection, this paper selects the comparison algorithm as FACEILD [35], Faster-FCC [36], KSDD [37], DNET [14], ResNet [7], and ConvNet [38] to further verify the performance of the proposed method.

From the experiment results in Table 3, it can be seen that the face detection results proposed in this paper are better than the DNET mainly due to the improvement of the face detection accuracy of the attention perception fusion module and the use of the multiscale pyramid pooling layer to capture high-level semantic features. The complementary features can effectively preserve the clear boundary of the face, while the combination of the multiple side output and pyramid pooling layer output can extract rich global context information and adapt to the two classification problems of face recognition. The heavy data set contains the most complex face image in the whole testing data. The face is seriously occluded, especially the image contrast is small and the face is fuzzy, which directly affects the detection and recognition effect of the network. From the precision comparison results in Table 3, it can be seen that the detection rate on the heavy data is the lowest mainly because the occlusion greatly reduces the perception ability of the deep network. However, the face detection based on the visual attention-guided mechanism proposed in this paper is also better than other deep networks. It can be seen that our proposed algorithm achieves 59.78% detection accuracy under the heavy evaluation standard, which is better than the comparison methods. In terms of detection efficiency, the detection and recognition speed of the input image with a resolution of 1024×2048 is 0.22 s, achieving a good balance between speed and accuracy. If the input image with a smaller resolution is detected, the detection speed of this method will be further improved.

In the selected detection image, the shape and scale of the face are quite different, especially the gray level of the face and the adjacent background is similar. According to the maximum analysis of the response map, these appearance changes cannot get the accurate boundary, which leads to ConvNet, KSDD, and FACEILD cannot get the accurate face. However, it is proved that the face with serious

Table 3: Recognition rate under different data sets.

Data set	Models						
	KSDD (%)	DNET (%)	ResNet (%)	ConvNet (%)	Faster-FCC (%)	FACEILD (%)	Proposed (%)
Bare	74.11	85.47	87.22	86.28	83.23	90.15	**90.06**
Mixed	69.12	78.22	81.25	87.24	79.39	87.09	**87.28**
Partial	78.28	85.28	88.20	85.69	78.29	88.36	**89.21**
Heavy	51.91	52.90	59.75	59.11	52.44	58.69	**59.51**

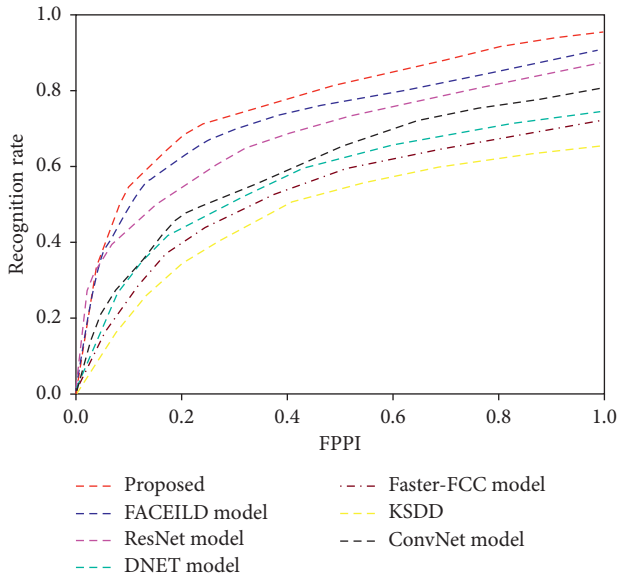


Figure 6: Relationship curve between FPPI and recognition rate on the LFW database.

occlusion such as deformation and low contrast can be accurately detected and recognized. KSDD is a lightweight network structure based on the VGG network. Although it can balance the contradiction between robustness and speed, it is still easy to be disturbed by occlusion, resulting in deviation of the detection center. From the detection results, it can be seen that the ConvNet detection and recognition has deviated from the face center. Our proposed model in this paper uses the attention-guided mechanism to highlight the visual area of occluded faces so that our proposed algorithm in this paper can better adapt to the influence of occluded interference in face detection.

4.5. Generalization Analysis. In order to verify the generalization performance of the proposed method, the proposed method was trained on the LFW training set, and cross-data set experiments were performed on the CMUFD database. The heavy subset consists of face objects with a height greater than 50 pixels and a visible range of [0.20, 0.65]. As shown in Figure 6, FPPI represents the statistical results of face detection and recognition algorithms at different detection rates. In order to facilitate comparison in different deep networks, the experiment mainly discusses the detection results of each algorithm when FPPI = 1 for analysis. The recognition rate of the ResNet algorithm is 91.88%, the recognition rate of the KSSD algorithm is 51.91%, and the detection rate of DenseNet algorithm is only 58.69%. The reason is that most deep detection methods only use the side-output feature and ignore the importance of global structural features. Our proposed paper uses a visual attention mechanism to guide the model to highlight the occlusion object visible area and simplify the face detection and recognition problem to a high-level semantic feature detection problem through an improved analytical network and uses the activation map to predict the location and scale of the face, which can avoid additional parameter settings and further reduce the false detection rate of each image. It can be clearly observed from Figure 6 that the performance of the proposed algorithm is obviously another algorithm.

5. Conclusions

Performance of face detection and recognition is affected and damaged because occlusion often leads to missed detection. In order to improve the accuracy of face detection and recognition, a visual attention mechanism guidance model is proposed in this paper, which uses the visual attention mechanism to guide the model highlight the visible area of the occluded face. The face detection problem is simplified into the high-level semantic feature detection problem through the improved analytical network, and the location and scale of the face are predicted by the activation map to avoid additional parameter settings. A large number of simulation experiment results show that our proposed method is superior to other comparison algorithms for the accuracy of occlusion face detection and recognition on the face database. In addition, our proposed method achieves a better balance between accuracy and speed, which can be used in the field of security surveillance. However, the performance of the proposed algorithm is sensitive to parameters, and its generalization is not high. How to improve this problem will be more conducive to the model applied to other scenarios or data.

Data Availability

All the data used to support the findings of this study are available within the article.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was financially supported by the key project of Education Bureau of Guangdong Province (Exploring the

Training and Practice of BeiDou + Intelligent Logistics Innovative Talents (2018CSLKT3-107) and Application of BeiDou + Blockchain in road transportation (2018GkQNCX072)).

References

[1] A. J. Colmenarez and T. S. Huang, "Face detection and recognition," *NATO ASI Series F Computer and Systems Sciences*, vol. 11, no. 2, pp. 208–218, 1998.

[2] T. Kondo and H. Yan, "Automatic human face detection and recognition under non-uniform illumination," *Pattern Recognition*, vol. 32, no. 10, pp. 1707–1718, 1999.

[3] L. H. Koh, S. Ranganath, and Y. V. Venkatesh, "An integrated automatic face detection and recognition system," *Pattern Recognition the Journal of the Pattern Recognition Society*, vol. 35, no. 6, pp. 1259–1273, 2002.

[4] S. Chaudhry and R. Chandra, "Face detection and recognition in an unconstrained environment for mobile visual assistive system," *Applied Soft Computing*, vol. 53, pp. 168–180, 2017.

[5] M. H. Siddiqi, R. Ali, A. M. Khan, E. S. Kim, G. J. Kim, and S. Lee, "Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection," *Multimedia Systems*, vol. 21, no. 6, pp. 541–555, 2015.

[6] S. Zhang, X. Zhu, Z. Lei, X. Wang, H. Shi, and S. Z. Li, "Detecting face with densely connected face proposal network," *Neurocomputing*, vol. 284, pp. 119–127, 2018.

[7] H. Ling, J. Wu, J. Huang, J. Chen, and P. Li, "Attention-based convolutional neural network for deep face recognition," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 5595–5616, 2020.

[8] H. Wang, D. S. Zhang, and Z. H. Miao, "Face recognition with single sample per person using HOG–LDB and SVDL," *Signal Image & Video Processing*, vol. 13, no. 19, 2019.

[9] S. L. Corrow, A. Albonico, and J. J. S. Barton, "Diagnosing prosopagnosia: the utility of visual noise in the cambridge face recognition test," *Perception*, vol. 47, no. 3, pp. 330–343, 2018.

[10] A. Abbad, O. Elharrouss, K. Abbad, and H. Tairi, "Application of meemd in post-processing of dimensionality reduction methods for face recognition," *Iet Biometrics*, vol. 8, no. 1, pp. 59–68, 2019.

[11] S. Madhavan and N. Kumar, "Incremental methods in face recognition: a survey," *Artificial Intelligence Review*, vol. 284, no. 5, pp. 119–127, 2019.

[12] Y. Su, Z. Liu, and M. Wang, "Sparse representation-based face recognition against expression and illumination," *IET Image Processing*, vol. 12, no. 5, pp. 826–832, 2018.

[13] L. I. Yan, S. Shan, R. Wang, Z. Cui, and X. Chen, "Fusing magnitude and phase features with multiple face models for robust face recognition," *Frontiers of Computer Science*, vol. 12, no. 6, 2018.

[14] Y. Zhang, K. Shang, J. Wang, N. Li, and M. M. Y. Zhang, "Patch strategy for deep face recognition," *IET Image Processing*, vol. 12, no. 5, pp. 819–825, 2018.

[15] Y. Jiang, F.-L. Chung, S. Wang, Z. Deng, J. Wang, and P. Qian, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 688–701, 2015.

[16] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Transactions on Image Processing*, vol. 99, pp. 1001–1015, 2019.

[17] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, and D. Hu, "Extended local binary patterns for face recognition," *Information Sciences*, vol. 24, no. 5, pp. 25–37, 2016.

[18] H. Shao, S. Chen, J. Zhao, W. Cui, and Y. U. Tianshu, "Face recognition based on subset selection via metric learning on manifold," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 12, pp. 102–118, 2015.

[19] A. K. Bobak, A. J. Dowsett, and S. Bate, "Solving the border control problem: evidence of enhanced face matching in individuals with extraordinary face recognition skills," *PLoS One*, vol. 11, no. 2, Article ID e0148148, 2016.

[20] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, "Viplfacenet: an open source deep face recognition sdk," *Frontiers of Computer Science*, vol. 11, no. 2, pp. 208–218, 2017.

[21] A. Rikhtegar, M. Pooyan, and M. T. Manzuri-Shalmani, "Genetic algorithm-optimised structure of convolutional neural network for face recognition applications," *IET Computer Vision*, vol. 10, no. 6, pp. 559–566, 2016.

[22] J. Zhao, Y. Lv, Z. Zhou, and F. Cao, "A novel deep learning algorithm for incomplete face recognition: low-rank-recovery network," *Neural Networks: The Official Journal of the International Neural Network Society*, vol. 94, pp. 115–124, 2017.

[23] Y. Cheng, Z. Li, L. Jiao, H. Lu, and X. Cao, "Enhanced retinal modeling for face recognition and facial feature point detection under complex illumination conditions," *Journal of Electronic Imaging*, vol. 25, no. 4, Article ID 043028, 2016.

[24] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for cross-modal face recognition," *International Journal of Computer Vision*, vol. 96, no. 8, p. 125058, 2016.

[25] X. Dong and H. Zhang, "Weighted neighbor sparse subspace based collaborative representation for face recognition," *Journal of Computational and Theoretical Nanoscience*, vol. 14, no. 4, pp. 1906–1913, 2017.

[26] S. Lou, X. Zhao, Y. Chuang, H. Yu, and S. Zhang, "Graph regularized sparsity discriminant analysis for face recognition," *Neurocomputing*, vol. 173, no. P2, pp. 290–297, 2015.

[27] H. Patil, A. Kothari, and K. Bhurchandi, "Expression invariant face recognition using semidecimated dwt, patch-ldsmt, feature and score level fusion," *Applied Intelligence*, vol. 44, no. 4, pp. 913–930, 2015.

[28] H. Shi, X. Wang, D. Yi, Z. Lei, X. Zhu, and S. Z. Li, "Cross-modality face recognition via heterogeneous joint bayesian," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 81–85, 2017.

[29] G. F. Lu, Y. Wang, and J. Zou, "Graph maximum margin criterion for face recognition," *Neural Processing Letters*, vol. 44, no. 2, pp. 1258–1268, 2015.

[30] H. Ran, W. Xiang, S. Zhenan, and T. Tieniu, "Wasserstein CNN: learning invariant features for NIR-VIS face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 7, pp. 1761–1773, 2018.

[31] S. Karahan, M. K. Yildirim, K. Kirtac, F. S. Rende, and H. K. Ekenel, "How image degradations affect deep CNN-based face recognition?" in *Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, vol. 12, no. 14, pp. 6–23, Darmstadt, Germany, September 2016.

[32] J. Wang and Z. Li, "Research on face recognition based on CNN," in *Proceedings of the IOP Conference*, pp. 170–177, Melbourne, Australia, September 2018.

[33] M. Matsugu, K. Mori, and T. Suzuki, "Face recognition using svm combined with CNN for face detection," *Lecture Notes in Computer Science*, vol. 253, no. 251, p. 1, 2004.

[34] L. Jing, Q. Tao, W. Chang, X. Kai, and W. Fang-Qing, "Robust face recognition using the deep C2D-CNN model based on decision-level fusion," *Sensors*, vol. 18, no. 7, pp. 2080–2093, 2018.

[35] Y. X. Yang, C. Wen, K. Xie, F. Q. Wen, G. Q. Sheng, and X. G. Tang, "Face recognition using the SR-CNN model," *Sensors*, vol. 18, no. 12, p. 1, 2018.

[36] A. Rikhtegar, M. Pooyan, and M. T. Manzuri-Shalmani, "GA-optimized structure of cnn for face recognition applications," *IET Computer Vision*, vol. 10, no. 6, pp. 559–566, 2016.

[37] B. Samik and D. Sukhendu, "Mutual variation of information on transfer-CNN for face recognition with degraded probe samples," *Neurocomputing*, vol. 310, pp. 299–315, 2018.

[38] Z. Lu, X. Jiang, and A. Kot, "Feature fusion with covariance matrix regularization in face recognition," *Signal Processing*, vol. 144, pp. 296–305, 2018.

Hindawi
Scientific Programming
Volume 2020, Article ID 3291426, 13 pages
https://doi.org/10.1155/2020/3291426

Research Article

Fingerspelling Identification for Chinese Sign Language via AlexNet-Based Transfer Learning and Adam Optimizer

Xianwei Jiang,[1,2] Bo Hu,[1,2] Suresh Chandra Satapathy,[3] Shui-Hua Wang,[4,5,6] and Yu-Dong Zhang[7,8,9]

[1]*Nanjing Normal University of Special Education, Nanjing 210038, China*
[2]*Joint Accessibility Key Laboratory, China Disabled Persons' Federation, Nanjing 210038, China*
[3]*School of Computer Engineering, KIIT Deemed to University, Bhubaneswar, India*
[4]*School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, China*
[5]*School of Mathematics and Actuarial Science, University of Leicester, Leicester LE1 7RH, UK*
[6]*School of Architecture Building and Civil Engineering, Loughborough University, Loughborough LE11 3TU, UK*
[7]*School of Informatics, University of Leicester, Leicester LE1 7RH, UK*
[8]*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China*
[9]*Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence should be addressed to Shui-Hua Wang; shuihuawang@ieee.org and Yu-Dong Zhang; yudongzhang@ieee.org

Received 12 February 2020; Accepted 28 March 2020; Published 19 May 2020

Academic Editor: Chenxi Huang

As an important component of universal sign language and the basis of other sign language learning, finger sign language is of great significance. This paper proposed a novel fingerspelling identification method for Chinese Sign Language via AlexNet-based transfer learning and Adam optimizer, which tested four different configurations of transfer learning. Besides, in the experiment, Adam algorithm was compared with stochastic gradient descent with momentum (SGDM) and root mean square propagation (RMSProp) algorithms, and comparison of using data augmentation (DA) against not using DA was executed to pursue higher performance. Finally, the best accuracy of 91.48% and average accuracy of $89.48 \pm 1.16\%$ were yielded by configuration M1 (replacing the last FCL8) with Adam algorithm and using 181x DA, which indicates that our method can identify Chinese finger sign language effectively and stably. Meanwhile, the proposed method is superior to other five state-of-the-art approaches.

1. Introduction

Nowadays, hearing-impaired people account for a large number, especially in China, where there are about 27.9 million deaf people [1]. Their lives, learning, and communication have all encountered unprecedented challenges. Sign language (SL) is their way of expression. But sign language contains a series of content and elements such as hand shape, movement, posture, and emotion, which is a relatively complicated system and not easy to be learned and mastered. Sign language translation and sign language recognition are two important solutions for above challenges. However, the former needs to arrange staff in advance and is expensive, and the latter is receiving more and more attention due to the rapid development of intelligent technology.

Chinese Sign Language (CSL) is designed specifically for Chinese deaf and hearing-impaired groups [2]. It has distinct characteristics, that is, rich in semantics, wide in area, and various in expressions. As an important part of it, Chinese finger sign language is of great significance. First of all, it is simple and easy to learn, which consists of 26 monosyllabic letters (A to Z) and 4 double syllable letters (ZH, CH, SH, NG). The content is clear, unique, and easy to remember. Secondly, the pinyin function is mapped, which can be used as the basis for gesture sign language learning.

The fingerspelling in Chinese Sign Language is different from the letter expression of American Sign Language (ASL). The fingerspelling of American Sign Language is conducive to the direct expression of words or meanings, while the Chinese finger sign language represents the composition of a pinyin element in most of the time. Several finger sign languages constitute a complete meaning, which is not easy to produce ambiguity. In addition, Chinese finger sign language has advantages in representing abstract concepts and terminology.

Sign language recognition (SLR) refers to the utilization of computer technology to translate sign language into texts, images, audio, video, and natural language that can be understood and accepted. Generally, it can be divided into two types: one is based on sensors and another is based on computer vision. Especially, the recognition technology based on computer vision [3–6] is a popular trend, which is flexible to operate, easy to implement, low in cost, and reliable in technology. Many researchers have contributed to these areas by providing superior classification and recognition algorithms and their variants. Using four different support vector machine kernels, a fingerspelling recognition system focusing on Thai fingerspelling sign language achieved the average accuracies of 91.20%, 86.40%, 80.00%, and 54.67%, respectively [7]. But it only had 375 character pictures totally. Based on HMM, K-means, and ant colony algorithm, Li et al. [8] recognized Taiwan Sign Language with accuracy of 91.3%. However, their database just included 11 special Taiwan Sign Language words. The dynamic time warping (DTW) was presented by Lichtenauer et al. [9] and gained a recognition rate of 92.3%. They evaluated sign classification on a set of Dutch Sign Language with 120 diverse signs. In addition, the hierarchical conditional random field (HCRF) method was proposed by Yang and Lee [10]. Rao et al. [11] trained the ANN classifier to match words. Kumar et al. [12] used the hidden Markov model (HMM) method. Lee et al. [13] combined SVM with HMM.

Some other research studies of sign language recognition utilizing deep learning were introduced in [14–17]. Jiang [18] employed the Gray-Level Co-occurrence Matrix (GLCM) and Parameter-Optimized Medium Gaussian Support Vector Machine (MGSVM) method to identify isolated Chinese Sign Language. A 6-layer convolutional neural network with the leaky rectified linear unit (LReLU) technique for Chinese Sign Language fingerspelling recognition was proposed by Jiang [19].

Although these techniques and methods have achieved favorable results, they still have their own problems and shortcomings. For instance, HMM and DTW need to establish an effective template first and are not suitable for real-time system. Feature extraction is quite important for SVM, which will consume a huge amount of resources. The accuracy of HCRF can get great improvement potential. In contrast, with strong self-learning ability and organizational capability, neural networks bring a new dawn. In particular, transfer learning is fueling those mature and efficient neural networks.

In this paper, we proposed a AlexNet-based transfer learning method equipped with data augmentation and Adam optimizer for fingerspelling identification of Chinese Sign Language and gained a stable average accuracy of $89.48 \pm 1.16\%$, which can be considered effective and significant.

Our contributions are as follows. (i) AlexNet-based transfer learning was introduced, which included some advanced techniques such as rectified linear unit function (ReLU), local response normalization (LRN), and dropout. (ii) Optimization algorithm Adam was utilized to accelerate the learning speed and enhance the performance. (iii) Data augmentation (DA) was used to provide sufficient training dataset, which strongly alleviated overfitting and raised the accuracy. (iv) Our study offered a new approach to smooth the barriers of communication between hearing-impaired people and healthy people.

2. Dataset

A self-built dataset of Chinese finger sign language based on computer vision was established, which was consisted of 1,320 images. The dataset is available upon e-mail request. According to the universal sign language standard issued by the state, Chinese finger sign language contains totally 30 sign language letters, 26 of which are single letters (A to Z) and 4 are double syllable letters (ZH, CH, SH, and NG). Figure 1 demonstrates thirty categories of Chinese finger sign language intercepted from sample images. These samples were preprocessed and normalized to 256×256 background-optimized images. Our experiment was executed with this private dataset including 1,320 images. Among them, 1050 images were used for training, and the rest were used for testing. The fingerspelling data used to support the findings of this study are available from the corresponding author upon request.

3. Methodology

3.1. Transfer Learning. An important reason why transfer learning catches the focus is that traditional machine learning applications require a large amount of labeled data, but these datasets may have problems such as distribution differences and training data expiration [20]. Fortunately, transfer learning can solve these issues. It can make full use of the previously labeled data and can guarantee the accuracy of the model on new tasks, which makes it get more and more attention.

Transfer learning is defined as the ability of a system to recognize and apply knowledge and skills learned in previous domains/tasks to novel domains/tasks. Transfer learning can be divided into instance-based transfer, feature-based transfer, and shared parameter-based transfer. The key to transfer learning is to find out which shared knowledge can be moved between different domains, design appropriate algorithms to extract and transfer common knowledge, and avoid negative transfer [21]. There are two general strategies for the application of transfer learning in deep learning. One is fine-tuning, which involves using a pre-trained network on the base dataset and training all layers in the target dataset; the other is freezing and replacing, for

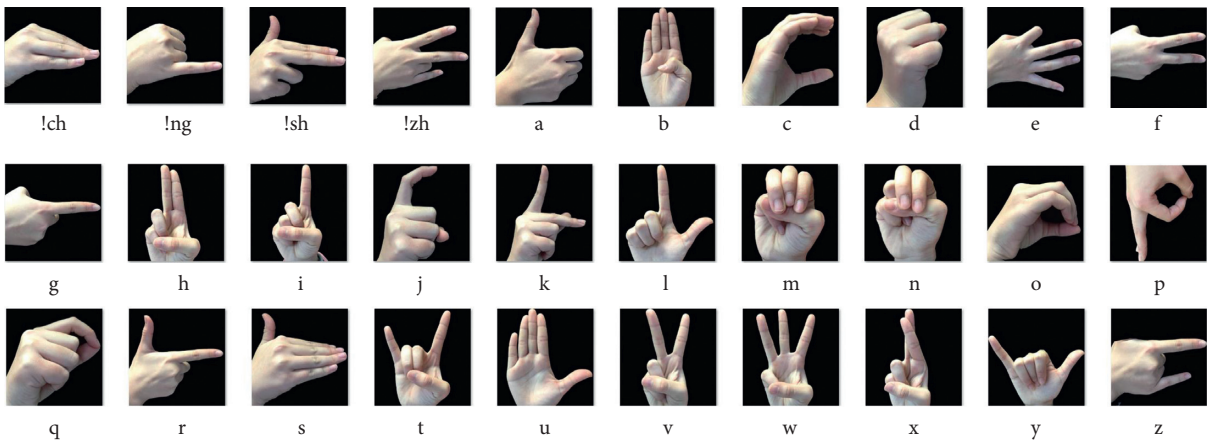


Figure 1: Thirty categories of Chinese finger sign language.

instance, freezing all layers (the weights are not updated) except the last layer and training the last layer or freezing N layers before (that is, the freezing layer was selected by custom) and training the last few layers remaining. Figure 2 shows the application of fine-tuning idea of the transfer learning. It can be seen that mapping from a big dataset to a target small data can be quickly achieved through reasonable fine-tuning, and a new suitable model is established from a pretrained model [22].

In the image processing field, pretraining ImageNet is frequently chosen and its model is initialized. In this study, we initially trained a pretrained model on a subset of ImageNet having 1,000 categories and then transferred the previously learned knowledge to 30 categories of Chinese Sign Language recognition based on a small amount of private data, which is a comparatively simplified task. The entire structure of the neural network was retrained. Thus, the AlexNet-based transfer learning brought advantages for fingerspelling identification of Chinese Sign Language.

3.2. AlexNet: Structure and Layers. AlexNet has attracted much attention due to the fact that it was far ahead of the second winner in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012), which shocked the academic and industrial researchers at the time. It can be thought as a deeper and wider version of LeNet. LeNet contains the basic modules of convolutional neural networks, providing the basis for other deep learning models [23]. From a holistic perspective, AlexNet retains the original idea but introduces more tricks such as ReLU, LRN, and dropout. The rapid development of technologies such as the use of GPUs and resolution of data sample limits has provided an opportunity for AlexNet.

In terms of layers that associate with learnable weights, AlexNet consists of a total of eight layers, including five convolutional layers and three fully connected layers [24]. The structure of AlexNet is illustrated in Figure 3.

Although AlexNet is similar to traditional neural networks, it has its own characteristics. The main aspects on advanced techniques can be summarized as follows:

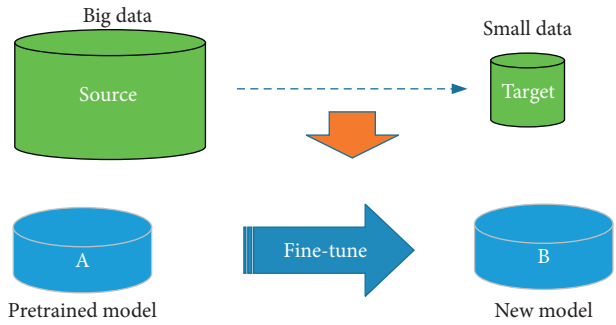


Figure 2: Fine-tuning idea of the transfer learning.

(a) N learnable filters are contained in convolutional layers, and each filter can generate one feature map. Thus, finally a size of $N \times O_W \times O_H$ activation map will be obtained, which implemented the function of feature extraction. Suppose that we have a 3D input with size of $I_W \times I_H \times L$ and a filter with $F_W \times F_H \times L$; $O_W \times O_H$ indicates the size of output activation map. The value of O_W and O_H can be calculated in following formulas:

$$O_W = 1 + \frac{(I_W - F_W + 2P)}{S},$$
$$O_H = 1 + \frac{(I_H - F_H + 2P)}{S}, \quad (1)$$

where the number of filters is defined as N, S indicates the stride size, and P denotes padding size. The entire process of convolution is shown in Figure 4, which started from the input with a series of filters and finally output an activation map.

(b) Replacing the traditional sigmoid function with rectified linear unit function (ReLU) is one of AlexNet's successful practical strategies. Because it is essentially a half-wave rectifier, it significantly speeds up the training process and prevents overfitting. In

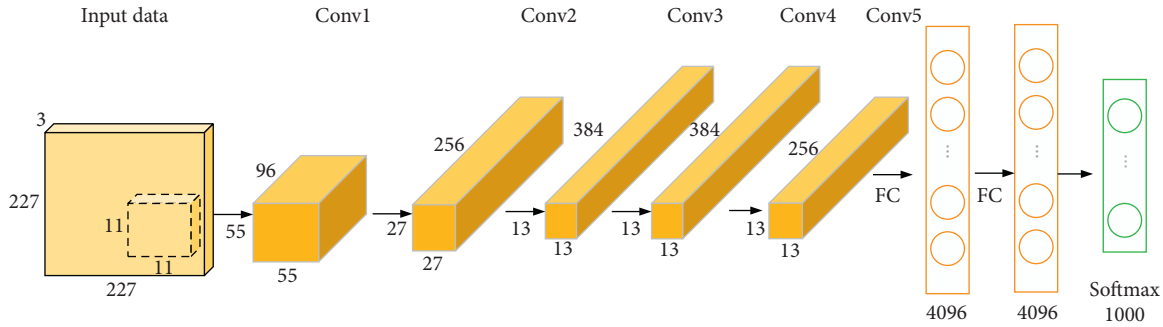


Figure 3: Structure of AlexNet.

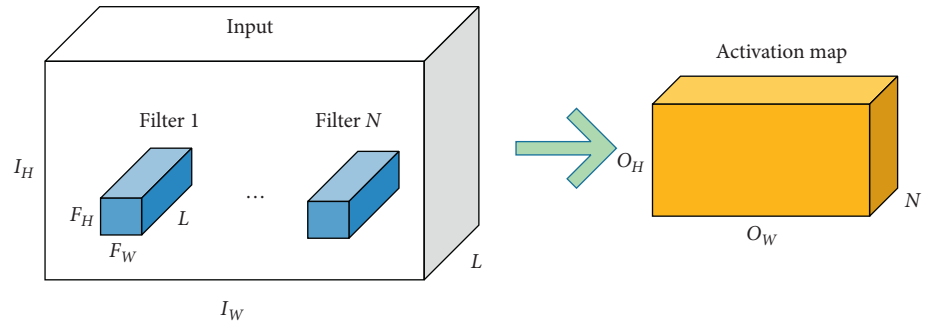


Figure 4: Illustration of convolution operation.

contrast, the sigmoid function is susceptible to gradient disappearance. It has been proven that the convergence speed of ReLU as an activation function in deep neural networks is about 7 times that of the traditional activation function. The formulas for both are expressed as follows:

$$\begin{aligned} \text{Sigmoid}(x) &= \frac{1}{1+e^{-x}}, \\ \text{ReLU}(x) &= \max(0, x). \end{aligned} \tag{2}$$

(c) AlexNet's another superior practical strategy is local response normalization (LRN) which was proposed by Krizhevsky et al. [25] to promote the convergence. Based on the computable neuron α_i with application of kernel i and nonlinear ReLU, the response-normalized neuron β_i can be computed as

$$\beta_i = \frac{\alpha_i}{\left(k + m\sum_{j=B}^{E} {\alpha_j}^2\right)^z}. \tag{3}$$

Among them,

$$\begin{aligned} B &= \max\left(0, i - \frac{n}{2}\right), \\ E &= \min\left(N-1, i + \frac{n}{2}\right), \end{aligned} \tag{4}$$

where n denotes the number of contiguous kernel maps, which is the size of window channel, and N represents the total number of kernels in that layer. According to the usual practice, the parameters are set as follows:

$$\begin{aligned} k &= 1, \\ m &= 10^{-4}, \\ n &= 5, \\ z &= 0.75. \end{aligned} \tag{5}$$

(d) Dropout should be considered a big innovation in AlexNet, which has become one of the must-have structures of neural networks. Dropout was introduced primarily to prevent overfitting. The implementation process is as follows: for those neurons of a certain layer, they are set to 0 with a probability P. Generally,

$$P = 0.5. \tag{6}$$

In this case, the dropout randomly generates the most network structure. These zero-set neurons do not participate in forward and backward propagation, as if they were frozen or discarded.

But at the same time, the number of neurons in the input and output layers is kept unchanged, and the parameters are updated according to the neural network learning method. The process is iterated repeatedly until the end of training. In this way, the entire neural network is reduced in scale and streamlined. It can also be seen as a combination of models that combine multiple models each time by generating different network structures. Thus, it can effectively reduce overfitting. Figure 5 represents a plain neural network; here, we use gray dotted circles to denote dropout neurons and use blue solid circles to indicate retained neurons.

(e) The fully connected layer (FC) can linearly transform one feature space into another, that is, map the learned "distributed feature representation" to the sample tag space. Essentially, it implements the "classifier" function. FC tends to appear at the end to weight the previously designed features. After multiplying the input by the corresponding weight and adding an offset, they can be obtained.

The softmax layer (SL), also called the softmax function, is usually followed after the final fully connected layer. The SL works on a multiple-input multiple-output mode. The softmax function converts the value into a probability, and the node with the highest probability will be selected as the prediction target. The softmax function is defined as follows:

$$P(S_i) = \frac{\exp(V_i)}{\sum_{j=1}^{N} \exp(V_j)}. \quad (7)$$

3.3. Transfer Setting. In this study, only one GPU was used to implement AlexNet because of the acceptable dataset. Table 1 provides the parameters of learnable weights and biases of AlexNet. Totally, weights and biases of AlexNet are 60,965,224 (60,954,656 + 10,568). In Matlab, single-float type is chosen to store all variables. As every variable occupies four bytes, the total 233 MB space was allocated.

The neural network structure needs to be modified, especially all the last fully connected layers. The original fully connected output was designed for 1000 classifications and is not suitable for our classification topic. Thus, we redesigned the relevant layers by fine-tuning. As shown in Table 2, a fully connected layer initialized randomly containing 30 neurons was introduced, and a softmax layer and a classifier layer that can implement 30 categories were utilized. Here, the layers 23, 24, and 25 are layer indexes in the Matlab deep neural network model, which also counts the ReLU layer and pooling layer, so the indexes here are larger than indexes from a Python model.

The options of training were set by practice. Based on the principle that the training epoch of the transfer learning is relatively small, the training epoch was selected as 10. The global learning rate was set to 10^{-4}. Considering that the new layers were randomly initialized with weights and bias while the transferred layers were pretrained, the new layers learning rate were defined 10 times that of the transferred layer.

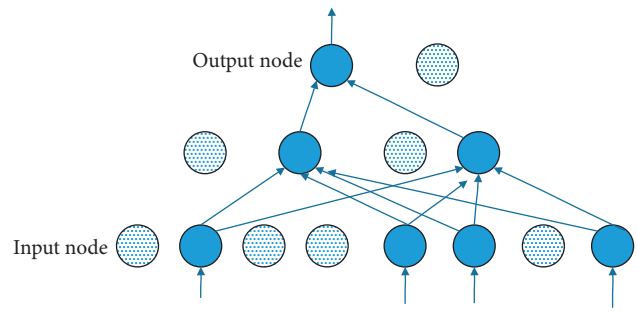


Figure 5: A plain neural network with dropout (gray dotted circles denote dropout neurons and blue solid circles indicate retained neurons).

Table 1: Learnable layers of AlexNet.

Name	Weights	Biases
CL1	$11 * 11 * 3 * 96 = 34{,}848$	$1 * 1 * 96 = 96$
CL2	$5 * 5 * 48 * 256 = 307{,}200$	$1 * 1 * 256 = 256$
CL3	$3 * 3 * 256 * 384 = 884{,}736$	$1 * 1 * 384 = 384$
CL4	$3 * 3 * 192 * 384 = 663{,}552$	$1 * 1 * 384 = 384$
CL5	$3 * 3 * 192 * 256 = 442{,}368$	$1 * 1 * 256 = 256$
FCL6	$4096 * 9216 = 37{,}748{,}736$	$4096 * 1 = 4{,}096$
FCL7	$4096 * 4096 = 16{,}777{,}216$	$4096 * 1 = 4{,}096$
FCL8	$1000 * 4096 = 4{,}096{,}000$	$1000 * 1 = 1{,}000$
Total	60,954,656	10,568

Different transfer learning settings were tried. As shown in Figure 6, setting M indicates that the layers which followed M are replaced and other layers reserved are transferred layers. Transferred layers keep learning rate as 1×10^{-4} while replaced layers initialize with learning rate of 10×10^{-4}. We will test four different transfer learning configurations: M1 to Mn—M1 replaces the last FCL8; M2 replaces FCL7 and 8; M3 replaces FCL6–8; M4 replaces CL5 and FCL6–8.

3.4. Data Augmentation on Training Set. Since the deep neural network model has many parameters, the proposed model needs to contain a considerable number of sample images to achieve optimal performance. Data augmentation (DA) can extend the dataset, which helps to enhance the performance of deep learning and improve the accuracy of classification recognition. For each of the original images selected, we applied the following six DA techniques: PCA color augmentation, affine transform, noise injection, scaling, random shift, and gamma correction.

Take the "*j*" image as an example, which was represented in 4th column and middle row of Figure 1. As each method produced 30 new images, we gained totally 180 new augmented "*j*" images, which can be seen in Figure 7.

The parameters of each DA techniques were set as follows: PCA color augmentation shifted the most present color values in original images. Affine transform exerted deformation to the images and preserved straight lines. The zero-mean 0.01 variance Gaussian noise was employed to every sign language images to generate 30 new noised images. Images were scaled with scaling factor from 0.7 to 1.3

Table 2: Revised layers in AlexNet.

Layer	Original	Revised
23	FCL (1000) with pretrained weights and biases	FCL (30) with random initialization
24	Softmax layer	Softmax layer
25	Classification layer (1,000 classes)	Classification layer (30 classes)

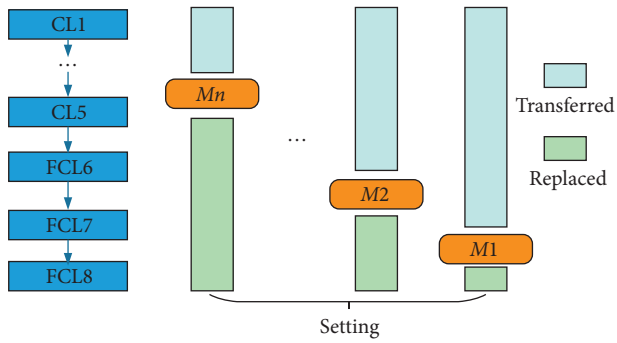
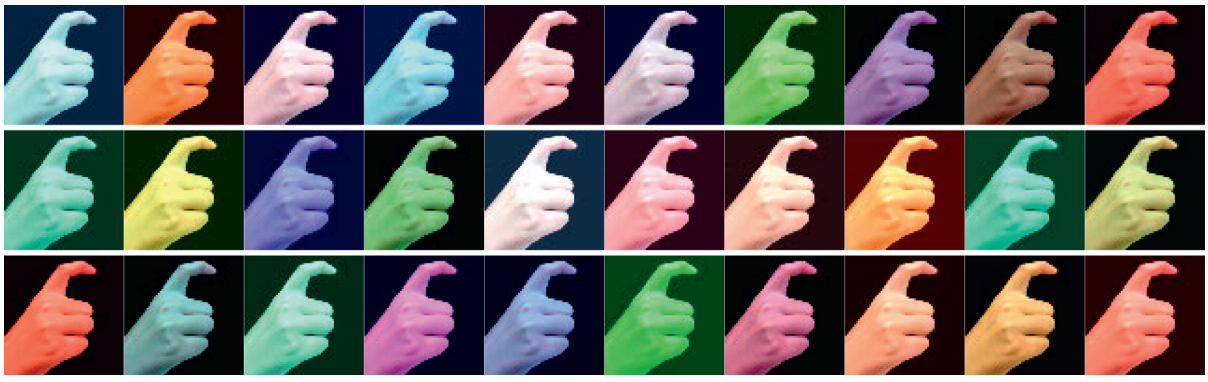
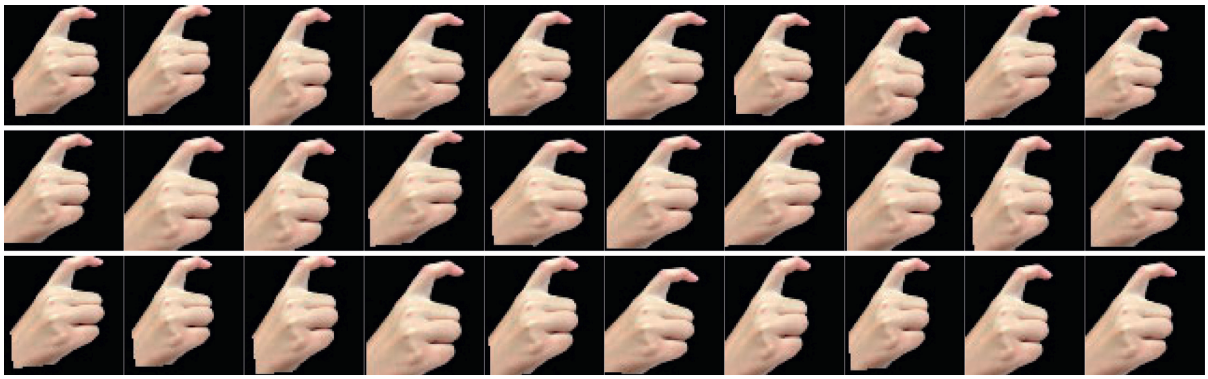


Figure 6: Different transfer learning settings (M1 to Mn—M1 replaces the last FCL8; M2 replaces FCL7 and 8; M3 replaces FCL6–8; M4 replaces CL5 and FCL6–8.).

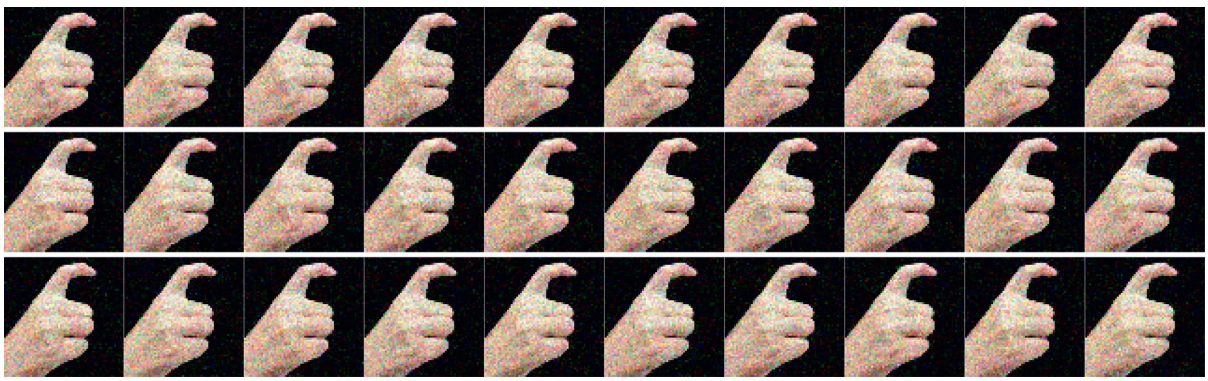


(a)

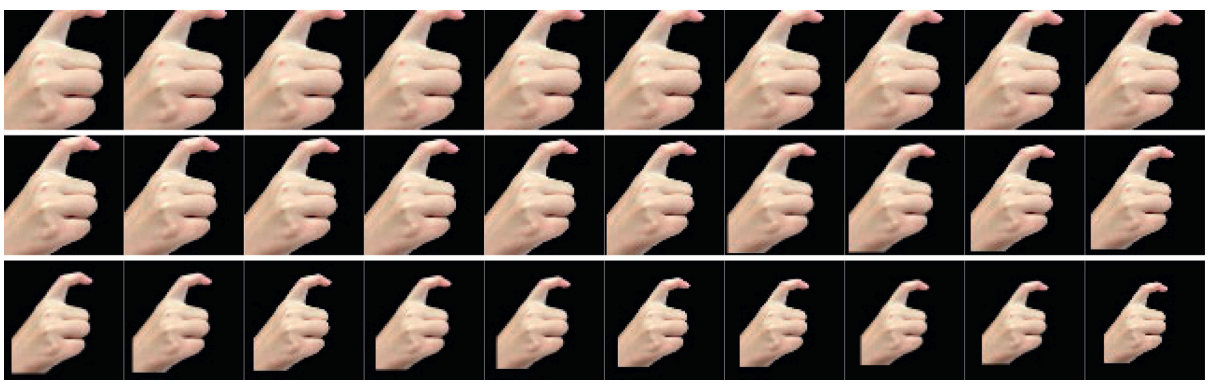


(b)

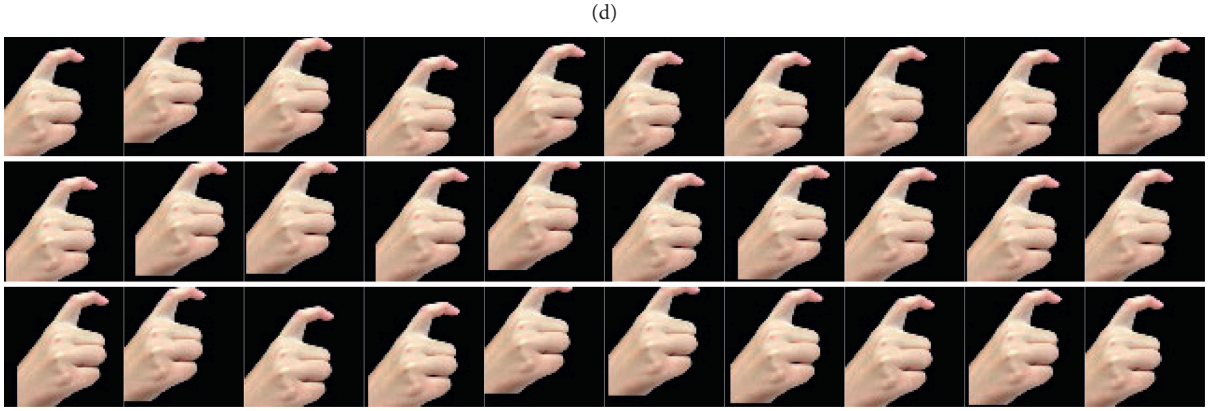
Figure 7: Continued.



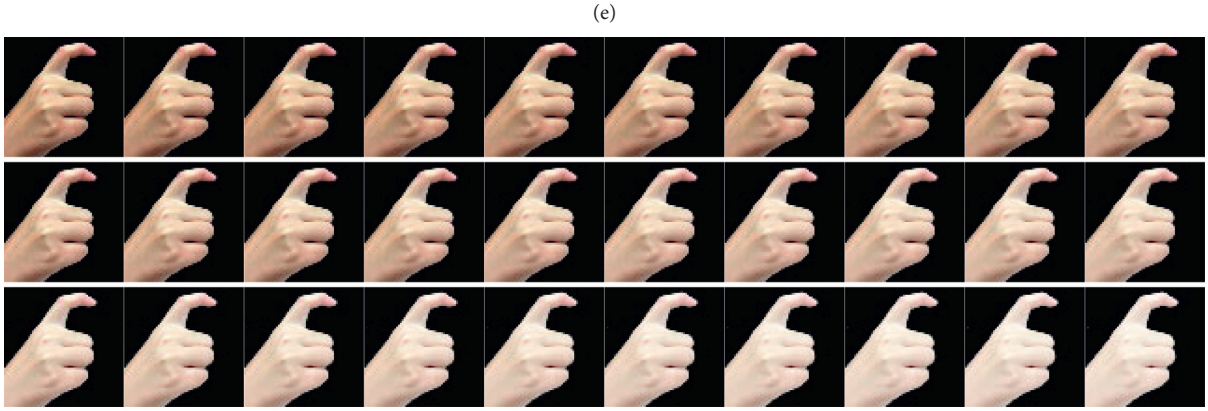
(c)



(d)



(e)



(f)

Figure 7: Data augmentation of "*j*" sample image. (a) PCA color augmentation. (b) Affine transform. (c) Noise injection. (d) Scaling. (e) Random shift. (f) Gamma correction.

with step of 0.02. Then, the hand gesture image was randomly shifted 30 times with the random shift within [−15, 15] pixels. Gamma correction with factor changed in the range of [0.4, 1.6] with increase of 0.04.

As shown in Table 3, which denoted number of images in three partitions: training, augmented training, and test, every original image generated new training sets 181 times. The experiment was repeated ten times with data split resetting randomly.

3.5. Training Algorithms. Deep learning often requires a lot of time and computer resources for training, so optimization algorithm is widely concerned. Adam (adaptive momentum) algorithm occupies less resources and makes the model converge faster [26], which can accelerate the learning speed and improve the effect. Adam is a first-order optimization algorithm that can replace the traditional stochastic gradient descent process. It joins the second moment estimation on the basis of momentum's first-order moment estimation and adds a moment to Adadelta. The learning rate of each parameter is dynamically adjusted by using the first and second moment estimation of the gradient. Bias correction is also added, which makes the parameters relatively stable.

The iterative formulas are as follows:

$$\begin{aligned} g &= \left(h_\theta\left(x^i - y^i\right)\right)x^i, \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1) * g, \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) * g^2, \\ \tilde{m_t} &= \frac{m_t}{1-\beta_1^t}, \\ \tilde{v_t} &= \frac{v_t}{1-\beta_2^t}, \\ \theta_j &= \theta_{j-1} - \tilde{m_t} * \frac{\alpha}{\sqrt{\tilde{v_t}} + \epsilon}, \end{aligned} \tag{8}$$

where g is the calculated gradient, m_t indicates the first moment of gradient g, which is also the expectation of gradient g, v_t denotes the second moment of gradient g, β_1 represents the first-order moment attenuation coefficient, β_2 represents the second moment attenuation coefficient, θ stands for the parameter that needs to be solved (or updated), and $\tilde{m_t}$ and $\tilde{v_t}$ indicate offset correction of m_t and v_t , respectively.

Two other popular training algorithms are stochastic gradient descent with momentum (SGDM) and root mean square propagation (RMSProp). In this study, we used those two algorithms as comparison basis.

SGDM, whose full name is stochastic gradient descent with momentum, introduces first-order momentum on the basis of SGD. The first-order momentum denotes the exponential moving average of the gradient direction at each moment, approximately equal to the average of the sum of the gradient vectors at the most recent T_j moment.

Table 3: Partition of dataset.

Partition	Number of images
Training	1,050
Augmented training	190,050
Test	270

Calculation of m_t is shown in formula above; T_j can be represented as follows:

$$T_j = \frac{1}{1-\beta_1}. \tag{9}$$

In other words, the descending direction at time t is determined not only by the gradient direction of the current point but also by the descending direction accumulated before it. The empirical value of β_1 is 0.9, which means that the direction of decline is mainly the previously accumulated direction of decline.

On the other hand, RMSProp (root mean square propagation) is an optimization algorithm proposed by Geoffrey E. Hinton in Coursera. In order to further optimize the loss function in the update of the problem of excessive swing and further speed up the convergence of the function, RMSProp algorithm used the differential squared weighted average for the gradient of weight W and bias b. As a result, it makes greater progress in the direction where the parameter space is gentler. The sum of squares of historical gradients is smaller because of gentler direction, which leads to a smaller learning drop. Assuming that in the process of iteration t, each formula is derived as follows:

$$\begin{aligned} s_{dw} &= \beta s_{dw} + (1-\beta) dW^2, \\ s_{db} &= \beta s_{db} + (1-\beta) db^2, \\ W &= W - \alpha \frac{dW}{\sqrt{s_{dw}} + \varepsilon}, \\ b &= b - \alpha \frac{db}{\sqrt{s_{db}} + \varepsilon}, \end{aligned} \tag{10}$$

where s_{dw} and s_{db} are the gradient and gradient momentum accumulated by the loss function during the previous iteration $t-1$ and the vector β is an exponential of gradient accumulation. To avoid the denominator becoming 0, ε is going to be a very small number.

RMSProp helps to eliminate the direction of the large swing and is used to correct the swing so that the swing in each dimension is smaller. On the other hand, it also makes the network functions converge faster. RMSProp is very similar to momentum in that it eliminates the wobble in gradient descent, including minibatch gradient descent, and allows to use a higher learning rate a to speed up learning of algorithm.

4. Experiment Results and Discussion

We ran this experiment on the platform of a personal computer whose main components include 2.5GHZ Core i7

CPU, 16 GB memory, and 2 GB AMD Radeon Graphics Processor, under the operating system of Win10. Different settings of transfer learning were tried to gain optimal hyperparameters, and then we carried out 10 runs on the test set using the final model.

Average accuracy (AA) is employed to evaluate the experiment results. AA is the average of the correctly classified categories among the research objectives, which reflects the average of the accuracy of each category. Table 4 provides the average accuracy of 10 runs over the test set. In each run, 270 images were randomly selected from the entire dataset to construct the test set.

4.1. Statistical Results. As can be seen in Table 4, the results of 10 runs show that the highest average accuracy is 91.48% and the minimum value of AA is 87.78%. We marked the highest AA in bold. Additionally, it can be seen that values of AA exceed 90% in 4 times. Finally, the value of mean and standard deviation achieves 89.48 ± 1.16%, which can be regarded as effective.

4.2. Training Algorithm Comparison. In this experiment, we compared Adam algorithm with SGDM and RMSProp algorithms. The comparison is shown in Figure 8.

As can be seen from Table 5, the mean and standard deviation of SGDM, RMSProp, and Adam are 88.33 ± 2.03%, 89.04 ± 0.93%, and 89.48 ± 1.16%, respectively. We can find that Adam algorithm is superior to SGDM and RMSProp. On the on hand, the stability of Adam is more excellent than SGDM. We find that the highest average accuracy of SGDM achieves 92.22% while the lowest value of AA drops to 84.81%. There is a huge amplitude between the two, which means gradient smoothness of SGDM is not enough to achieve stable transition, that is, SGDM's inertia is not maintained enough to accommodate an unstable objective function. One the other hand, the accuracy of Adam is higher than RMSProp. Obviously, in total 10 runs, result of RMSProp has surpassed Adam only three times.

4.3. Setting of Transfer Learning Comparison. Four different transfer learning settings (configurations M1, M2, M3, and M4) were run on the test set. Table 6 and Figure 9 show the results of different settings; it can be seen that the mean and standard deviation values of configurations M1, M2, M3, and M4 are 89.48 ± 1.16%, 88.93 ± 0.75%, 88.59 ± 0.90%, and 86.85 ± 0.94%, respectively. Obviously, the greatest performance was obtained by configuration M1 (replacing the last FCL8) among all four measures. Overall, performance declines as the number of replacement layers increases. Configuration M4 (replacing CL5 and FCL6–8) achieved mean ± SD value of 86.85 ± 0.94% which is the lowest in four configurations. This situation indicates that using most transferred layers from a pretrained model is more efficacious and practical. Moreover, data augmentation expanded our dataset to a sufficient training set, which can strongly avoid overfitting and improve the accuracy.

Table 4: Average accuracy of 10 runs.

Run	AA
1	**91.48**
2	89.26
3	89.26
4	87.78
5	89.26
6	90.00
7	89.26
8	87.78
9	90.00
10	90.74
Mean + SD	89.48 ± 1.16

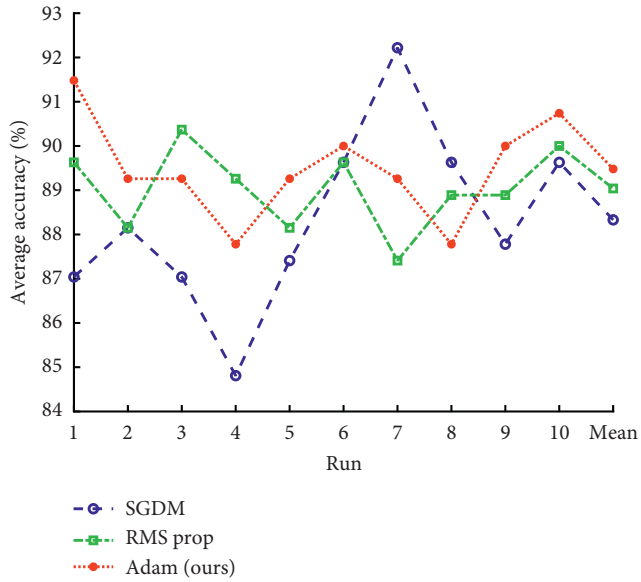


Figure 8: Comparison plot of SGDM, RMSProp, and Adam.

Table 5: Comparison of different training algorithms.

Run	SGDM	RMSProp	Adam (ours)
1	87.04	89.63	91.48
2	88.15	88.15	89.26
3	87.04	90.37	89.26
4	84.81	89.26	87.78
5	87.41	88.15	89.26
6	89.63	89.63	90.00
7	92.22	87.41	89.26
8	89.63	88.89	87.78
9	87.78	88.89	90.00
10	89.63	90.00	90.74
Mean ± SD	88.33 ± 2.03	89.04 ± 0.93	89.48 ± 1.16

4.4. Effect of Data Augmentation. In this experiment, we compared the performances of using data augmentation (DA) against not using DA. The experiment settings were all the same as in Section 4.1, only removing the DA on the training set. The results are shown in Table 7. We tested four different augmentation factors. First, we generate 10 new images for each method of six DAs. Therefore, we obtained

Table 6: Comparison of different transfer learning configurations.

Run	Configuration			
	M1 (ours)	M2	M3	M4
1	91.48	88.52	87.78	86.67
2	89.26	88.89	88.89	86.30
3	89.26	88.52	89.63	86.67
4	87.78	88.89	89.26	85.56
5	89.26	88.15	89.63	86.67
6	90.00	90.37	88.15	88.89
7	89.26	89.26	87.41	87.41
8	87.78	87.78	87.78	86.30
9	90.00	89.26	89.63	86.30
10	90.74	89.63	87.78	87.78
Mean ± SD	89.48 ± 1.16	88.93 ± 0.75	88.59 ± 0.90	86.85 ± 0.94

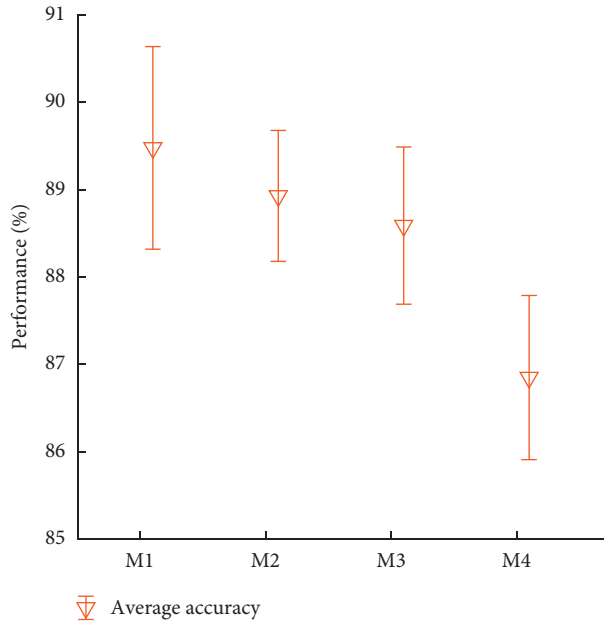


Figure 9: Error bar of average accuracies of different transfer learning configurations.

Table 7: Effect of data augmentation.

Run	Not using DA	Using DA 61x	Using DA 121x	Using DA 181x	Using DA 241x
1	85.19	87.78	88.15	91.48	90.37
2	85.19	85.19	89.63	89.26	89.26
3	85.93	86.30	87.78	89.26	88.89
4	82.22	86.30	85.93	87.78	89.63
5	82.59	87.41	87.41	89.26	89.63
6	84.44	86.67	88.89	90.00	89.63
7	85.56	87.04	88.89	89.26	88.52
8	83.70	85.56	90.00	87.78	88.52
9	84.07	88.15	88.52	90.00	87.78
10	83.70	87.04	88.52	90.74	90.00
Mean + SD	84.26 ± 1.24	86.74 ± 0.94	88.37 ± 1.16	89.48 ± 1.16	89.22 ± 0.79

60 new images; together with original image, we have in total 61x augmentation on the training set. Second, we set the augmentation factor as 121x. Afterwards, we also check the performance of 181x and 241x augmentation.

As can be seen from Figure 10, effect of not using DA is obviously poorer than using DA. When using DA, the performance increases with the improvement of augmentation factor. Nevertheless, 241x augmentation brings the

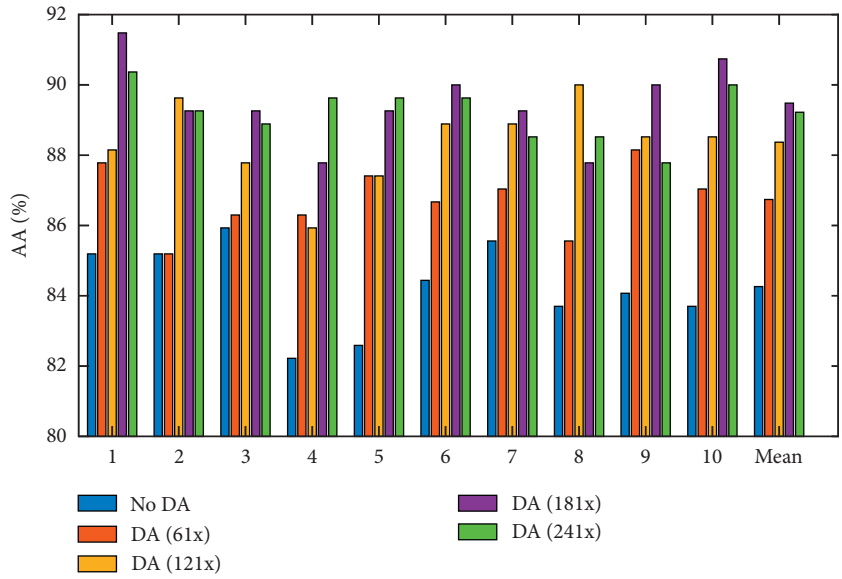


Figure 10: Comparison of DA with different augmentation factors.

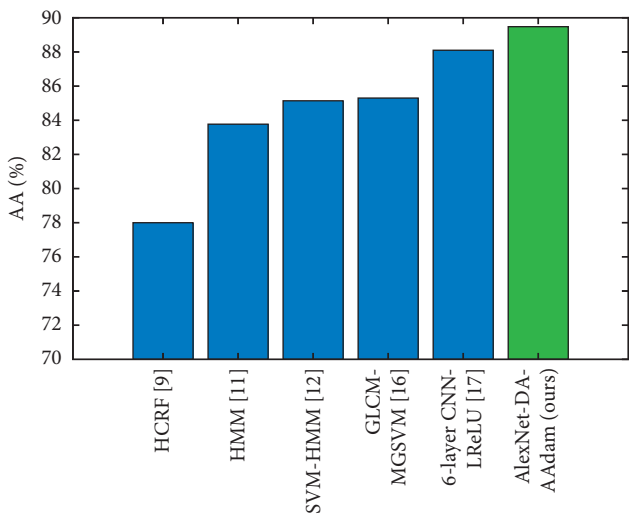


Figure 11: Comparison plot of recent sign language classification methods.

drop of average accuracy, and it also requires the largest computation resources. Thus, setting the augmentation factor as 181x, that is, the purple bar in the diagram, gains the best performance.

4.5. Comparison to State-of-the-Art Approaches. Five other state-of-the-art approaches: HCRF [10], HMM [12], SVM-HMM [13], GLCM-MGSVM [18], and 6-layer CNN-LReLU [19], were sought to compare with proposed method in this experiment. The average accuracy of HCRF, HMM, SVM-HMM, GLCM-MGSVM, and 6-layer CNN-LReLU is 78.00%, 83.77%, 85.14%, 85.3%, and $88.10 \pm 1.48\%$, respectively, which is demonstrated in Figure 11. It can be observed that our method is superior to all the comparisons. Three attributes are helpful to enhance performance of our method. (i) With fine-tuning, the mature pretrained model can quickly transfer previous task to our task, which is comparatively simplified. Therefore, the accuracy of our method is close to the level of the AlexNet model. (ii) The Adam algorithm can boost retraining AlexNet model, which accelerated the learning speed and improved the effect. (iii) Data augmentation extended the dataset and improved the accuracy of classification recognition. Thus, our method achieved a satisfying result.

The shortcoming of our method is AlexNet needs a massive amount of computing resources. In the future, we may try to move our algorithm to cloud computing [27,28] and mobile edge computing [29,30] areas.

5. Conclusion

In this study, an AlexNet-based transfer learning method was proposed, equipped with data augmentation and Adam optimizer. Our method was applied to fingerspelling identification for Chinese Sign Language. The experiment results demonstrated that our method achieved the average accuracy of $89.48 \pm 1.16\%$, which was excellent among the six state-of-the-art approaches. We compared three training algorithms: Adam, RMSProp, and SGDM algorithm, and found that Adam algorithm is more remarkable and stable. We tested four different transfer learning settings and discovered that configuration M1 (replacing the last FCL8) acquired greatest performance. In addition, this strategy is very practical, and it can reduce the large amount of training of the network and make full use of existing models. We also observed that using DA is more effective than not using DA, and setting different augmentation factors leads to distinct performance, in which 181x augmentation achieves the best average accuracy.

In the future, we shall try to verify different transfer learning models, such as ResNet, GoogleNet, and SqueezeNet. We need to solve the issue of setting appropriate

learning rate factor in individual layer. The dataset also needs to be further expanded for getting higher accuracy. We shall try to shift this method to other applicable areas and test other feasible methods.

Data Availability

The fingerspelling data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Jiangsu Overseas Visiting Scholar Program for University Prominent Young & Middle-Aged Teachers and Presidents of China, Natural Science Foundation of Jiangsu Higher Education Institutions of China (19KJA310002), Henan Key Research and Development Project (182102310629), Natural Science Foundation of China (61602250), Jiangsu Natural Youth Science Foundation (BK20180727), Guangxi Key Laboratory of Trusted Software (kx201901), Royal Society International Exchanges Cost Share Award, UK (RP202G0230), Medical Research Council Confidence in Concept (MRC CIC) Award, UK (MC_PC_17171), Hope Foundation for Cancer Research, UK (RM60G0680), and British Heart Foundation Accelerator Award, UK.

References

[1] X. Li, *Research on Chinese Sign Language Recognition for Middle and Small Vocabulary Based on Neural Network*, pp. 1–10, University of Science and Technology of China, Hefei, China, 2017.

[2] M. Huber and C. Havas, "Restricted speech recognition in noise and quality of life of hearing-impaired children and adolescents with cochlear implants - need for studies addressing this topic with valid pediatric quality of life instruments," *Frontiers in Psychology*, vol. 10, no. 2085, p. 9, 2019.

[3] L. N. Wu, "Pattern recognition via PCNN and tsallis entropy," *Sensors*, vol. 8, pp. 7518–7529, 2008.

[4] L. Wu, "Classification of fruits using computer vision and a multiclass support vector machine," *Sensors*, vol. 12, pp. 12489–12505, 2012.

[5] G. Wei, "Color image enhancement based on HVS and PCNN," *Science China Information Sciences*, vol. 53, pp. 1963–1976, 2010.

[6] L. N. Wu, "Improved image filter based on SPCNN," *Science In China Series F-Information Sciences*, vol. 51, pp. 2115–2125, 2008.

[7] P. S. Thongpan Pariwat, "Thai finger-spelling sign language recognition using global and local features with SVM," in *Proceedings of the 9th International Conference on Knowledge and Smart Technology (KST)*, pp. 116–120, Chonburi, Thailand, February 2017.

[8] T.-H. S. Li, M.-C. Kao, and P.-H. Kuo, "Recognition system for home-service-related sign language using entropy-based," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 1, pp. 150–162, 2016.

[9] J. F. Lichtenauer, E. A. Hendriks, and M. J. T. Reinders, "sign language recognition by combining statistical DTW and independent classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 2040–2046, 2008.

[10] H.-D. Yang and S.-W. Lee, "Robust sign language recognition with hierarchical conditional random fields," in *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 2202–2205, Istanbul, Turkey, August 2010.

[11] G. A. Rao, P. Kishore, D. A. Kumar, and A. Sastry, "Neural network classifier for continuous sign language recognition with selfie video," *Far East Journal of Electronics and Communications*, vol. 17, p. 49, 2017.

[12] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, "A position and rotation invariant framework for sign language recognition (SLR) using Kinect," *Multimedia Tools and Applications*, vol. 77, pp. 8823–8846, 2018.

[13] G. C. Lee, F. Yeh, and Y. Hsiao, "Kinect-based Taiwanese sign-language recognition system," *Multimed Tools Applications*, vol. 75, pp. 261–279, 2016.

[14] L. P. Di Wu and P.-J. Kindermans, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, pp. 1583–1597, 2016.

[15] X. Jiang, M. Lu, and S.-H. Wang, "An eight-layer convolutional neural network with stochastic pooling, batch normalization and dropout for fingerspelling recognition of Chinese sign language," *Multimedia Tools and Applications*, vol. 78, 2019.

[16] H. Liu, R. Cui, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1610–1618, Honolulu, HI, USA, July 2017.

[17] W. Z. J Huang, H. Li, and W. Li, "Attention based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 1–10, 2018.

[18] X. Jiang, "Isolated Chinese sign language recognition using gray-level Co-occurrence Matrix and parameter-optimized Medium Gaussian support vector machine," in *Frontiers in Intelligent Computing: Theory and Applications*, pp. 182–193, Springer, Berlin, Germany, 2020.

[19] X. Jiang, "Chinese sign language fingerspelling recognition via six-layer convolutional neural network with leaky rectified linear units for therapy and rehabilitation," *Journal of Medical Imaging and Health Informatics*, vol. 9, pp. 2031–2038, 2019.

[20] M. Byra, M. Galperin, H. Ojeda-Fournier et al., "Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion," *Medical Physics*, vol. 46, pp. 746–755, 2019.

[21] C. Huang, "Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling," *Frontiers in Neuroscience*, vol. 12, p. 818, 2018.

[22] 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.

[23] T. Shanthi and R. S. Sabeenian, "Modified Alexnet architecture for classification of diabetic retinopathy images," *Computers & Electrical Engineering*, vol. 76, pp. 56–64, 2019.

[24] K. Y. Son, M. E. Yildirim, J. S. Park, and J. K. Song, "Flood detection by using FCN-AlexNet," in *Proceedings of the 11th*

International Conference on Machine Vision, pp. 248–255, Bellingham, Germany, November 2019.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2017.

[26] A. Yaguchi, T. Suzuki, W. Asano, S. Nitta, Y. Sakata, and A. Tanizawa, "Adam induces implicit weight sparsity in rectifier neural networks," in *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications*, pp. 318–325, New York, NY, USA, 2018.

[27] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," *IEEE Transactions on Industrial Informatics*, vol. 16, pp. 1321–1329, 2019.

[28] T. Wang, D. Zhao, S. Cai, W. Jia, and A. Liu, "Bidirectional prediction based underwater data collection protocol for end-edge-cloud orchestrated system," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4791–4799, 2020.

[29] Y. Wu, H. Huang, Q. Wu, A. Liu, and T. Wang, "A risk defense method based on microscopic state prediction with partial information observations in social networks," *Journal of Parallel and Distributed Computing*, vol. 131, pp. 189–199, 2019.

[30] T. Wang, H. Luo, X. Zheng, and M. Xie, "Crowdsourcing mechanism for trust evaluation in cpcs based on intelligent mobile edge computing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, pp. 1–19, 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 1895341, 12 pages
https://doi.org/10.1155/2020/1895341

Research Article

Violation Detection of Live Video Based on Deep Learning

Chao Yuan[1,2] **and Jie Zhang**[1,3]

[1]*College of Economics and Management, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Avenue, Nanjing 211100, China*
[2]*School of Design, Jiangnan University, 1800 Lihu Avenue, Wuxi, Jiangsu 214122, China*
[3]*Nanjing University of the Arts, No. 74 Beijing West Road, Nanjing, Jiangsu 210013, China*

Correspondence should be addressed to Jie Zhang; jiezhang@nuaa.edu.cn

Received 2 February 2020; Accepted 16 April 2020; Published 11 May 2020

Academic Editor: Chenxi Huang

With the rapid development of Internet technology, live broadcast industry has also flourished. However, in the public network live broadcast platform, live broadcast security issues have become increasingly prominent. The detection of suspected pornographic videos in live broadcast platforms is still in the manual detection stage, that is, through the supervision of administrators and user reports. At present, there are many online live broadcast platforms in China. In mainstream live streaming platforms, the number of live broadcasters at the same time can reach more than 100,000 people/times. Only through manual detection, there are a series of problems such as low efficiency, poor pertinence, and slow progress. This approach is obviously not up to the task requirements of real-time network supervision. For the identification of whether live broadcasts on the Internet contain pornographic content, a deep neural network model based on residual networks (ResNet-50) is proposed to detect pictures and videos in live broadcast platforms. The core idea of detection is to classify each image in the video into two categories: (1) pass and (2) violation. The experiments verify that the network proposed can heighten the efficiency of pornographic detection in webcasts. The detection method proposed in this article can improve the accuracy of detection on the one hand and can standardize the detection indicators in the detection process on the other. These detection indicators have a certain promotion effect on the classification of pornographic videos.

1. Introduction

At present, Internet technology and Internet of Things technology have achieved rapid development, and various products and consumption models based on these technologies have been deeply applied in various industries [1–12]. People's online entertainment has undergone tremendous changes. From simple video watching and single text reviews to video and commentary barrage videos, it has finally developed into today's barrage live broadcast. The interactive form is constantly approaching offline interaction, and the whole country has entered the "live +" era. According to network statistics, in 2015, China had nearly 200 online live broadcast platforms. The number of live broadcasts has reached 200 million, and the number of users in the online live broadcast industry has reached 300 million. Large-scale live broadcast platforms have nearly 4 million simultaneous online users during peak hours every day, and the number of live broadcasts at the same time exceeds 3000. In 2016, the number of online live broadcast platforms in China increased to 300 and the number of users in the online live broadcast industry reached 300 million. By 2017, the number of live broadcast industry users has reached 500 million. With the vigorous development of live broadcast industry, the illegality of live broadcast content has become increasingly serious. In terms of live webcast video detection, it is still in the manual detection stage, mainly through the supervision of administrators and user reports. The main problems with manual supervision are (1) increased labor costs; (2) prone to missed inspections due to limited human energy; and (3) the huge number of live broadcast platforms and live video, resulting in a huge workload of manual

supervision. Relying solely on manual supervision cannot achieve the goal. In addition, there are a large number of online live broadcast platforms; the content is complex, and the workload of pure human supervision is too large to meet the regulatory requirements of live broadcast platforms. At present, there is an urgent need for pornographic detection methods for online live broadcast.

At present, various methods are used for pornographic image detection, and there are two classification methods in total. One is to divide the detection methods into the following three categories. The first is a rule-based approach [13]. This method first establishes a skin model that can filter nonskin areas and then obtains the skin area from the image to be identified. If it is larger than the threshold, it is considered to be an erotic picture. Although this method is intuitive and easy to implement, the skin area threshold is difficult to set accurately. Many nonerotic images also contain a large amount of skin area, so the accuracy of identifying pornographic images is low. The second method is based on image retrieval [14]. This method first establishes a database with pornographic images, then selects the appropriate image features, and then compares the image features to be identified with the image features in the database. If the similarity exceeds a certain threshold, it is determined as a pornographic image. Due to the diversity of pornographic images, in order to improve the accuracy, a huge database needs to be established. This will lead to excessive memory usage and long recognition time. The third is a learning-based approach [15, 16]. This method needs to design the visual features of pornographic images, and based on these visual features, use the method of machine learning [17–27] to obtain the learned model. In the end, the learned model is used for pornographic image recognition. It can be found through comparison that, although the learning-based method has higher accuracy and detection speed among the three methods, the image features obtained are generally artificially generalized. However, the generalization performance of artificially selected visual features is low and it is difficult to meet the actual requirements.

The second general classification divides pornographic recognition methods into the following three categories: pornographic recognition based on skin color, pornographic recognition based on manually extracted features, and pornographic recognition based on deep learning. The research based on skin color mainly focuses on the classification of pornographic pictures with a large number of skin pixels. Fprsyth and Fleck [28] identify skin tone region of image based on color and texture features to analyze the structure of the human body and the correlation between various parts of the human body, thereby implementing a naked recognition system. Yang and Ahuja [29] proposed using a Gaussian mixture model to learn the distribution of skin pixels of the training image under the CIELuv color space. The learned model is applied to skin pixel detection in the test image. Jones and Rehg [30] introduced the concept of histogram, extracting the skin color regions of erotic pictures and normal pictures and counting the color distribution histograms of their corresponding RGB color spaces. According to the histogram, the distribution of human skin color in the RGB color space is obtained. According to this method, features such as skin color area, maximum skin color connected area, and average skin color probability are calculated. The neural network classifies the features to achieve pornographic detection of the image. Srisaan [31] and Basilio et al. [32] extract skin color regions in the YCbCr color space and HSV color space, respectively. In both studies, skin color detection is achieved by counting the distribution of skin color pixels. To sum up, the detection of pornographic images based on skin color detection relies on the judgment criteria of skin color pixels. Skin color is easily affected by light, and it is difficult to use a simple skin color detection for high-accuracy pornographic images with complex textures. Nonerotic images containing a large number of skin-colored pixels are prone to misjudgment during the detection process. Therefore, it is not reliable to use skin color as the criterion for identifying pornographic images. Pure skin color detection has certain limitations in pornographic image recognition. To enhance recognition rate, more features in image need to be extracted for multidimensional analysis. Therefore, pornographic recognition based on manual feature extraction has appeared. The framework mainly includes two links: feature extraction and classifier, as shown in Figure 1.

Pornographic recognition based on manually extracted features mainly uses different feature extraction algorithms to extract prominent features such as color, texture, and shape and uses different classifiers for the extracted features to realize pornographic image recognition. Karavarsamis et al. [33] obtained the skin color convex hull region of the image by ROI localization, calculated 15 features such as the expectation and variance of the convex hull region and the ratio of nonskinned pixels in the convex hull region in the RGB color space. Finally, the random tree forest was used for classification. Wang et al. [34] proposed a model for identifying nude images based on navel and torso features. Deselaers et al. [35] proposed to extract image information based on visual bag of words (BoW) model, feature vectors in image blocks were used as bag of words, and SVM was used to classify visual word histograms.

Pornographic image recognition based on deep learning mainly classifies pornographic images or videos through convolutional neural network (CNN) and LSTM network models. Moustaf [36] uses the NPDI [37] dataset for fine tuning on the AlexNet [38] and GoogleNet [39] models. The two models are mixed according to different thresholds, and finally the classification is achieved. The model is shown in Figure 2. Wehrmann et al. [40] proposed the Adult Content Recognition with deep neural network model, which is used to extract the features of key frames in the NPDI dataset. Long- and short-term memory networks are used to classify the extracted features. Perez et al. [41] divided video information into static information and dynamic information. Video frames are fed into a CNN model to extract static information, and optical flow and MPEG motion vectors are used to describe video motion information. The motion information is sent to a CNN to extract video motion features. Finally, support vector machines [42] are used to

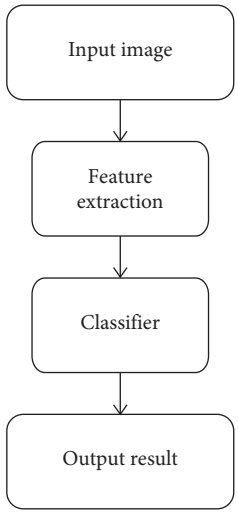


Figure 1: Flowchart of pornographic recognition based on manually extracted features.

classify the extracted features. Pornographic image detection based on deep learning has higher accuracy than skin color and feature-based detection methods.

ResNet-50 learning model based on residual networks is introduced in this study and uses this learning model to detect and classify pictures and videos. The advantages of the model are as follows: (1) it has higher recognition accuracy and efficiency, and the accuracy reaches more than 95%; (2) due to the addition of a residual module in this model, there is relatively little information loss in the process of information transfer between layers, which can avoid the problem that it is difficult for deep networks to accept the information of the previous layers; (3) the model is no longer limited by the dimensions of the image; and (4) by improving the ordinary CNN, the generalization ability of the model after adding the residual network is improved. In this paper, the proposed network model is applied to pornographic content detection on live broadcast platforms, and experiments verify the effectiveness of the proposed scheme. The specific contributions of this article are as follows:

(1) The detection algorithm based on deep residual network can process high-dimensional pictures, and the accuracy and efficiency of recognition are better than other similar algorithms.

(2) Generally, as the number of layers of the learning network increases, the learning efficiency will become lower and lower. The advantage of the residual network is that it is very simple to perform identity transformation in the network structure, reducing the defined network structure by the increase of the number of network layers so that the information in the network structure can reduce the transmission loss as much as possible.

(3) By improving the network model, the recognition accuracy and efficiency of the two types of pictures are further improved.

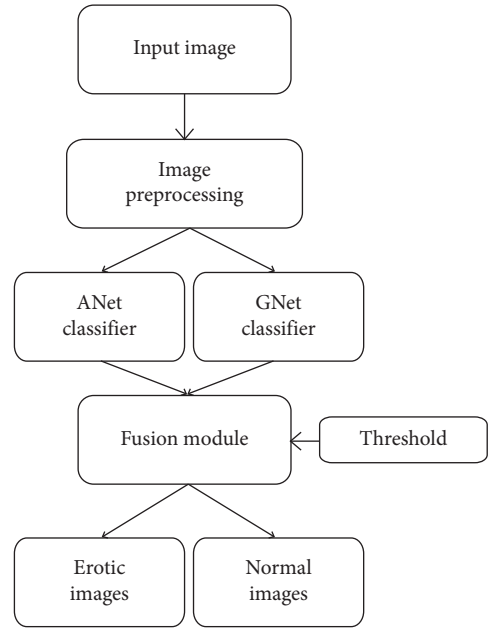


Figure 2: AGNet erotic image classification model.

2. Related Work

2.1. Picture Detection. Pictures are one of the main information carriers that people come into contact with daily, and it is necessary to detect pornographic pictures on the Internet. There are many types of erotic pictures, and the most common of which are pictures of naked bodies. Therefore, some scholars have proposed that the area of exposed skin and the area of all bodies in the picture can be calculated to find the proportion of the area of exposed skin. Given a certain threshold range to determine whether a picture violates the rules, this is a very simple method of judgment, but the disadvantages are also obvious. The detection of pictures is not targeted and the accuracy is difficult to guarantee.

Given a certain number of samples, mark whether these samples are illegal pictures. Before classification, the image must be preprocessed, and the features of the image are extracted and compared using a convolution layer to determine whether the image is illegal. Such an algorithm is more accurate than threshold classification. The specific process is shown in Figures 3 and 4.

2.2. Video Detection. Video is another representation of multimedia and one of the information carriers often seen in people's lives, but the detection of video is relatively tedious. The basic unit of a video is a frame, that is, a picture, which means that a video is composed of many pictures. The video detection process can be broken down into the following steps: first divide the video into frame-by-frame pictures, then preprocess each frame of pictures, and finally classify each frame of pictures. If an offending picture is detected, the administrator can be notified and the offending video can be processed in a timely manner. The flow chart of video detection is shown in Figure 5.

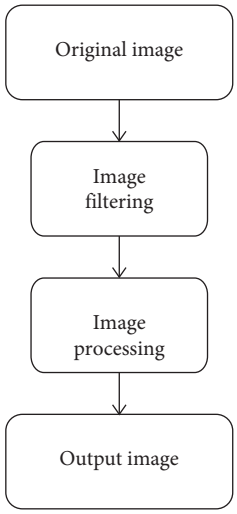


Figure 3: Image preprocessing flowchart.

3. Way of Solving the Problem

3.1. Network Structure. For the detection of illegal videos, the use of deep learning networks is feasible. However, as the depth of the network gradually increases, the efficiency of feature extraction will become lower and lower and the efficiency of learning will gradually decrease. In summary, a deep-learning network alone cannot complete the task well. To address these challenges, this article proposes a new solution; using ResNet-50 [43] based on residual networks, both deep-layer networks can be used to ensure the characteristics of the extracted pictures and as far as possible in the deep-layer network structure so that the defined network structure is not affected by the increase in the number of network layers, thereby improving learning efficiency and accuracy.

3.2. Image Preprocessing and Feature Extraction. Feature extraction the core factor for effective classification. Before feature extraction, a high-quality picture needs to be selected. Pick some pictures with sharp textures and sharp edges. Cut the picture to a suitable size and denoise the picture through a filter. In order to make the effect of feature extraction better, this article performs image enhancement on all pictures. Common methods are contrast raising, histogram equalization, etc. The pictures are divided into two types; the first type is the training set and the second type is the test set. ResNet-50 is used for feature extraction on the training set in deep-learning network. This process is to tell the detection system whether some marked pictures are illegal pictures or qualified pictures, thereby facilitating the detection of the test set and the accuracy of the detection. It directly affects the quality of the detection system. The image processing flow is shown in Figure 6.

3.3. Image Dimensionality Reduction. High-dimensional data is a challenge for detection. Assuming that the dimensions of the image are $512 * 512$, the corresponding one-dimensional vector calculation is 262144 dimensions.

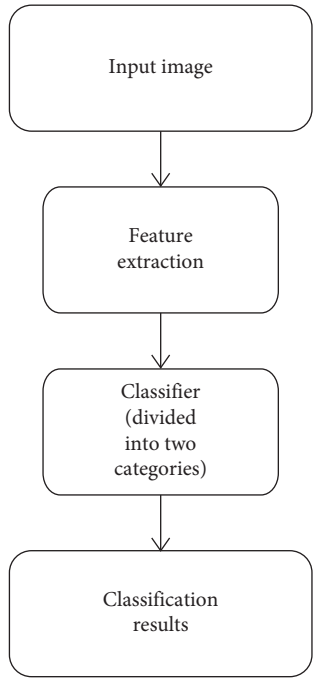


Figure 4: Picture classification flowchart.

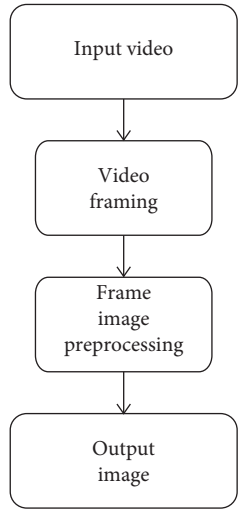


Figure 5: Video preprocessing flowchart.

Such dimensions make it difficult to extract features. Dimension reduction is an essential link. The reduction of dimensions is mainly done by the pooling layer (sampling layer). There are many commonly used feature extraction methods, and the classical ones are the CNN [44], the principal component analysis (PCA) method [45] and kernel principal component analysis [46] (KPCA) and so on. The pooling layer can be directly used to reduce the dimension, and no additional image-processing steps are required. The pooling layer (sampling layer) mainly depends on the size of the convolution kernel and the stride of the convolution kernel to reduce the dimensionality. The reduced dimensional image is easier to process.

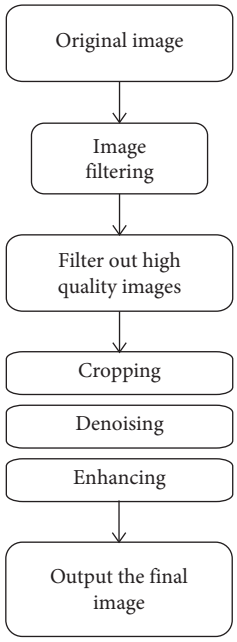


Figure 6: Image processing flowchart.

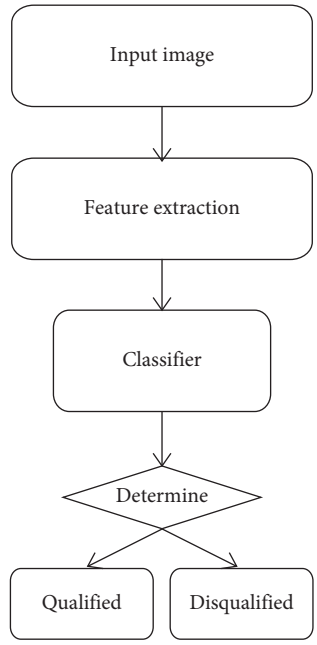


Figure 7: Identification flowchart.

3.4. Residual Network. For a common network structure, when the number of layers is appropriately increased within a certain range, the detection performance can be improved. However, it is not allowed to arbitrarily increase the number of layers of CNN. Continuously increasing the number of layers of the neural network will inevitably cause the accuracy on the training set to saturate or even decrease. When the number of neural network layers continues to increase, it will cause problems such as disappearance or degradation of gradients, gradient explosions, and network training costs, so the detection performance cannot be improved as the number of network layers increases. Based on this background, this paper introduces residual networks. It reduces the number of calculations and computational difficulty. For the residual network, the identity transformation of deep network is used to make the deep network degenerate into a relatively shallow network. The residual network mainly can perfectly solve the side effect caused by the deep network, namely, the degradation problem; then, we can improve the network performance by increasing the network depth.

3.5. Picture Classification. Classify the frames in the video, and the classification results are qualified and illegal. In the residual network structure, a fully connected layer is used to complete the classification. Fully connected layer processes and stores the data and uses it as input to the layer. Then, output a multidimensional vector, which is the number of classifications. Since this experiment is a binary classification problem, the multidimensional vector in the experiment is a two-dimensional vector. In the experiment, softmax [47] was used for discrimination. Since the result of the softmax output is a probability, the probability is used to represent the detection result in a two-dimensional vector. Figure 7 shows specific identification process.

4. Introduction to Detection Algorithms

In order to verify the performance of the proposed detection method, this paper proposes five comparison models, which are (1) pornographic image recognition algorithms based on skin color features; (2) direction gradient histogram (HOG) + support vector machine (SVM); (3) CNN; (4) CNN + residual network algorithm; and (5) detection algorithm based on VGG-16 network. This paper proposes a network structure based on ResNet-50 and introduces five comparison algorithms first.

4.1. Contrast Algorithm 1: Pornographic Image Detection Algorithm Based on Skin Color Features. According to the author's research on pornographic images, erotic images are pictures with more naked body parts. Based on the survey, the author first set a certain threshold and then calculated the proportion of naked leakage in each picture. If the proportion of naked leakage is greater than the set threshold, the picture is considered to be a violation. This method will have certain results when identifying the offending pictures. However, this method has some drawbacks. If the clothes worn by the person in the picture are close to the skin, the recognition result will have a greater impact, and it is likely that the picture is illegal.

4.2. Contrast Algorithm 2: Image Detection Algorithm Based on Directional Gradient Histogram (HOG) and Support Vector Machine (SVM). HOG [48] is the histogram of direction gradient. This method is currently a method in the field of computer vision and pattern recognition. This method can be used to describe the local texture characteristics of the image. The specific description steps are as follows:

Step 1: segment the image.

According to the strategy of segmenting images, there are two kinds of segmentation strategies; the first is nonoverlap and the second is overlap, which means that features can be extracted as many times as possible after image segmentation. This experiment uses the overlap strategy.

Step 2: calculate the direction gradient histogram of the segmented block.

The specific calculation formula is as follows:

$$M(x, y) = \sqrt{I_x^2 + I_y^2},$$

$$\theta(x, y) = \tan^{-1}\frac{I_y}{I_x} \in [0, 360^\circ), \quad \text{or} \in [0, 180^\circ). \tag{1}$$

Among them, I_x and I_y represent the gradient values in the horizontal and vertical directions, $M(x, y)$ represents the magnitude of the gradient, and $\theta(x, y)$ represents the direction of the gradient.

Step 3: composition characteristics.

According to the block segmentation of the image, the results of the block segmentation are connected end-to-end to form the characteristics that describe the image; then, these characteristics are the basis for our judgment of image classification, which is the role of the image example mentioned earlier.

SVM is support vector machine. Take binary classification as an example. The data participating in binary classification are called positive samples and negative samples, respectively. If positive and negative samples are given in the experiment, the role of the support vector machine is to find a "hyperplane." Classify the positive and negative samples as well as possible so that the positive and negative samples are distributed as uniformly as possible on the hyperplane.

4.3. Contrast Algorithm 3: Image Detection Algorithm Based on Convolutional Neural Network. CNN is an excellent network structure among many neural networks. For CNN, the experiments in this paper mainly use convolutional layers, pooling layers (sampling layers), and fully connected layers. The CNN is specifically introduced below.

For the convolutional layer, the features of the original image are extracted by setting the size of the convolution kernel and the step size during the convolution. After the convolution, the dimension of the image will be slightly reduced according to the size of the convolution kernel.

The primary role of the pooling layer or sampling layer is to reduce the dimensions and reduce the image; secondly, simplify the complexity of the network, reduce the amount of calculation, and reduce memory consumption; at the same time, expand the receptive field to achieve translation invariance, rotation invariance, and scale invariance; implement nonlinearity and more.

The fully connected layer maps the learned "distributed feature representations" into the sample space and acts as a "classifier" in the entire CNN to find out whether the picture is specifically qualified or unqualified.

4.4. Contrast Algorithm 4: Image Detection Algorithm Based on Convolutional Neural Network and Residual Network. Through the improvement of the CNN, the residual network part is added to form a new algorithm. The network structure is introduced below. Since the CNN was introduced above, only the residual part is introduced in this section.

When we use a standard optimization algorithm to train a common network, such as using gradient descent (or other popular optimization algorithms), without using residuals, or without these shortcuts or jump connections, you will find that when the network depth deepens, mistakes first decrease and then increase. However, in theory, as the network depth deepens, the better the training effect is. However, in fact, if there is no residual network, for a normal network, the deeper the depth, the harder it is to train with an optimization algorithm. Moreover, as the depth of the network deepens, training errors will increase.

However, the situation is different by using ResNets. Even if the network is deeper, the training performance remains stable, and the training error is reduced, even if the network is deep to 100 layers. Some people have even performed experiments in 1000-layer neural networks, which allow us to guarantee good performance while training deeper networks. As the network gets deeper, the scale of the network connection will become huge, but ResNet is indeed very effective in training deep networks.

4.5. Contrast Algorithm 5: Image Detection Algorithm Based on VGG-16. This model has a total of 16 layers of network structure, so it is called VGG-16 [49]. It was proposed by Oxford University in 2014, and its network is simple and practical. Taking a $224 \times 224 \times 3$ picture as an example, the specific calculation process is as follows:

Step 1: input a $224 \times 224 \times 3$ picture, and after two convolutions with 64 convolution kernels, use one pooling. After the first convolution, c1 has $(3 \times 3 \times 3)$ trainable parameters.

Step 2: after two convolution kernel convolutions of 128, one pooling is used.

Step 3: after three convolutions of 256 convolution kernels, use pooling.

Step 4: repeat the three 512 convolution kernels twice and perform pooling again.

Step 5: the last entry is three full connections.

4.6. Algorithm in This Paper: Image Detection Algorithm Based on ResNet-50. The ResNet algorithm proposed in 2015 is relatively simple and very practical. Based on this, ResNet-50 or ResNet-101 is derived, and many methods are constructed based on this. Detection, segmentation, recognition,

Table 1: Key parameters of the experiment.

Experimental parameters	Number of initial convolution kernels	Training times	Learning efficiency	Image size	Classification number	Classification
Value	64	10,000	$1e-4$	208×208	2	Softmax

Table 2: Detection results of 6 algorithms.

Experiment name	Accuracy $p1$ (%)	Recall rate $p2$ (%)	Accuracy $p3$ (%)	Misclassification $p4$ (%)	$F1$ score (%)
CNN	85.39	88.24	86.26	14.61	86.79
CNN + ResNet	92.51	91.91	92.37	7.87	92.21
HOG + SVM	81.65	80.15	83.21	18.35	80.89
Skin color features	72.50	74.56	70.15	27.50	73.52
VGG-16	—	—	—	—	—
ResNet-50	97.08	97.79	96.18	2.92	97.43

and other application scenarios are beginning to use ResNet. All this shows that ResNet is really easy to use and has great potential for development. Theoretically, with the deepening of the network layers, the accuracy of image classification should gradually increase, but in practice, with the deepening of the network, the accuracy of the training set decreases. However, it is definitely not the case of the network due to overfitting because overfitting will cause the training set to have a high-accuracy rate. Traditional convolutional networks or fully connected networks will have more or less problems of information loss and loss when transmitting information and also cause "gradient disappearance" or "gradient explosion," resulting in deep network training. ResNet solves the abovementioned problems to a certain extent. By using the idea of bypassing the information at the input to the output, the integrity of the information can be protected. The entire network only learns the difference between input and output, thereby streamlining learning goals and reducing complexity.

In view of the abovementioned situation, the residual network will have a relatively better result. The residual unit establishes a direct correlation channel at the data input and data output ends. This makes a powerful network structure focus on learning the residuals between input and output. For example, generally we use $F(X, W_i)$ to represent the residual mapping, then the output is $Y = F(X, W_i) + X$. When the number of input and output channels is the same, we can naturally use X to add directly. When the number of channels between them is different, we need to consider establishing an effective mapping function so that the number of processed input X and output Y channels is the same, that is, $Y = F(X, W_i) + W_s \times X$.

According to the abovementioned introduction of ResNet, it can solve the problem of the deepening of the recognition accuracy of the network. One of the reasons is that the residual network can directly transfer shallow data information to relatively deep levels through associated channels, unlike traditional networks. Shallow layers of information need to be transmitted backward layer by layer. When the information is transmitted to the deeper layer, the information in the ordinary network is bound to cause losses. This explains why the recognition accuracy of ordinary networks decreases as the network deepens.

ResNet-50 is a 50-layer network structure based on residual networks, including 50-layer network structures such as convolutional layers, pooling layers, and residual networks. It is a relatively mature network model. In this paper, the network model is used to detect illegal videos. According to the experimental results, ResNet-50 will have a relatively good effect.

5. Experiment and Analysis

5.1. Dataset. The dataset is used as the basis for data identification. A data-rich dataset has a decisive effect on the recognition result. The dataset in this article contains a training dataset and testing one. In the training dataset, two types of pictures are classified and labeled and used as training. Then, we use the test dataset to check the trained network. Since there is no official library for this kind of dataset, all the datasets are used to find images by themselves, filter the images, and perform a series of preprocessing processes on the images to make ordinary images a usable dataset.

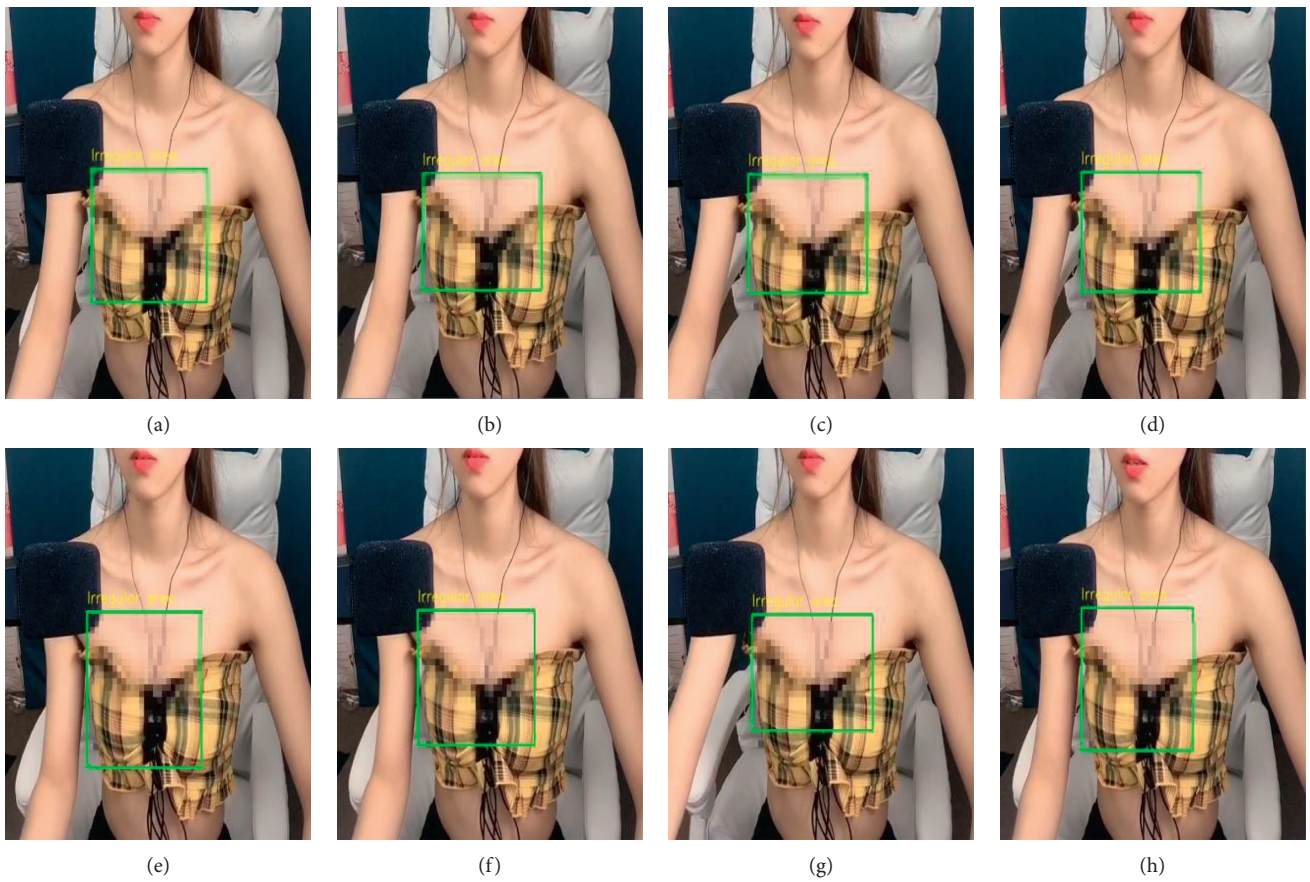
The dataset contains a training set: 280 images, of which 140 are qualified pictures and 140 are unqualified pictures. The test set contains 267 images, of which 136 are qualified images and 131 are unqualified images. For data of the training set, image enhancement is required. All the images of the training set are enhanced trying to avoid the phenomenon of network "overfitting." By transforming the training image to obtain a network structure with stronger generalization ability, it is also a means to enhance the network when the number of datasets is not large. There are several methods for data enhancement:

(1) Image enhancement using standardization

(2) Use geometric transformation (translation, flip, and rotation) to enhance the data of the image

(3) Use random adjustment of brightness to enhance the image

(4) Use random adjustment contrast to enhance the image



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

(e) (f) (g) (h)

Figure 8: Example 1 of violation results detected.

5.2. Evaluation Criteria of Experimental Results. For the test results, the following evaluation criteria have been formulated.

Criterion 1. Accuracy P (1) = number of correctly identified pictures/number of participating pictures × 100%.

However, the recognition accuracy P (1) cannot be simply used as the evaluation criterion. First of all, a single evaluation standard will bias the evaluation results, and if for a wrong classifier, the detection results of all pictures are qualified, then when the test set of qualified pictures is detected, it will show excellent performance. Then, we cannot consider such a classifier to be a good classifier, and such a classifier is often not a qualified classifier. As a result, the classification accuracy rate of qualified pictures and violation pictures and the misclassification rate of the overall classification of the classifier appear, respectively. The specific definitions are as follows.

Criterion 2. Recall rate P (2) = number of correctly identified qualified pictures/number of participating qualified pictures × 100%.

Criterion 3. Recognition accuracy P (3) = number of correctly identified offending pictures/number of participating in identifying offending pictures × 100%.

Criterion 4. Misclassification rate P (4) = number of mispredicted pictures/total number × 100%.

Criterion 5. F_1 is the harmonic mean of precision and recall, and the specific formula is as follows:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{2}$$

where precision is the accuracy rate, that is, the proportion of the number of correctly identified violation pictures in the identification of violation pictures. Recall is the recall rate, that is, the proportion of correctly identifying qualified pictures among the proportion of participating in identifying qualified pictures.

5.3. Experimental Parameters. The key parameters used in the experiment are shown in Table 1. The number of initial convolution kernels determines the number of features extracted by the convolution layer, so as many features as possible can be extracted to facilitate the subsequent detection and classification steps, so this study uses 64 initial convolution kernels. Learning efficiency is the speed of network learning. Due to the number of trainings, the learning efficiency can be set smaller. If you set a large learning efficiency, as the number of learning times increases, it will be difficult to reduce the loss function, and it will always oscillate in an interval.

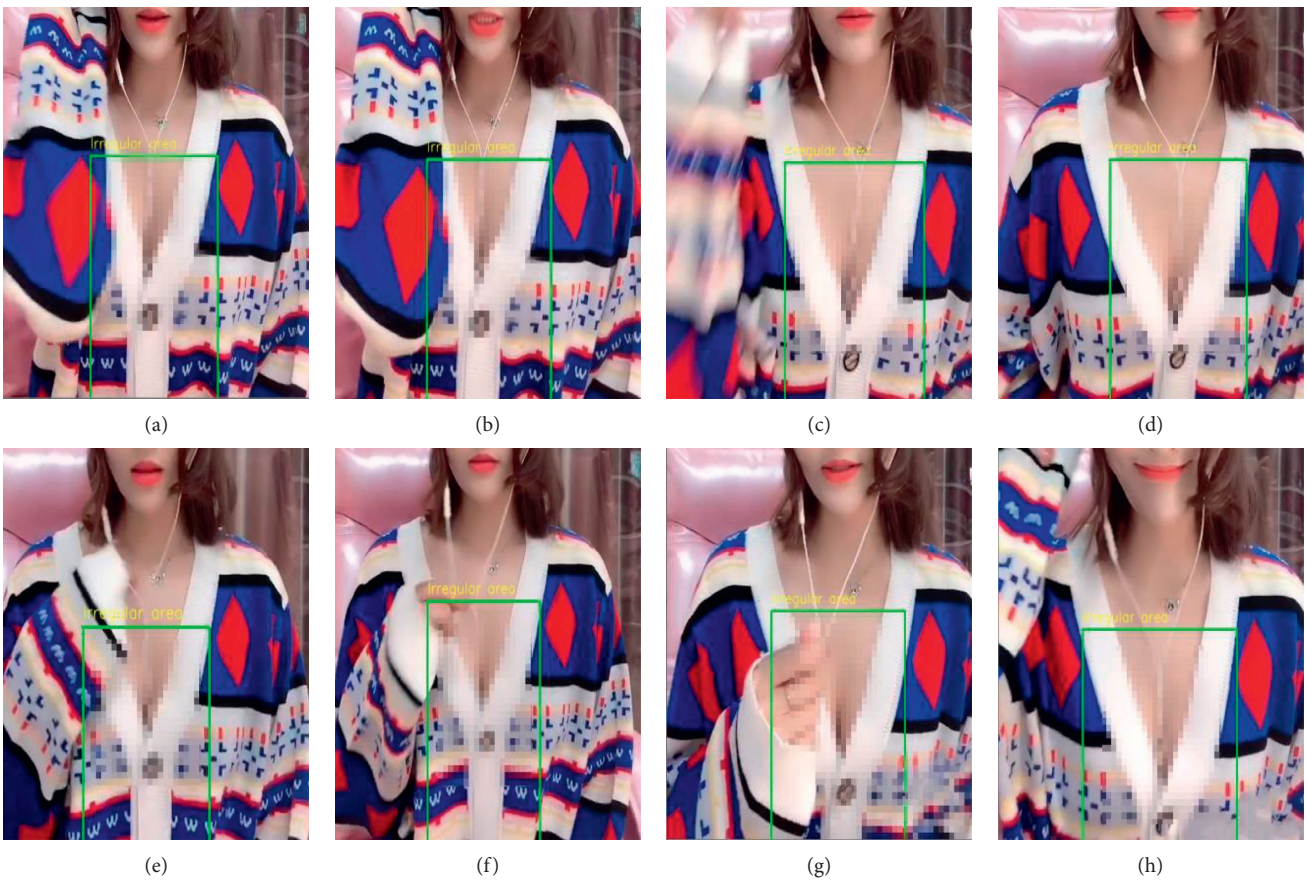


Figure 9: Example 2 of violation results detected.

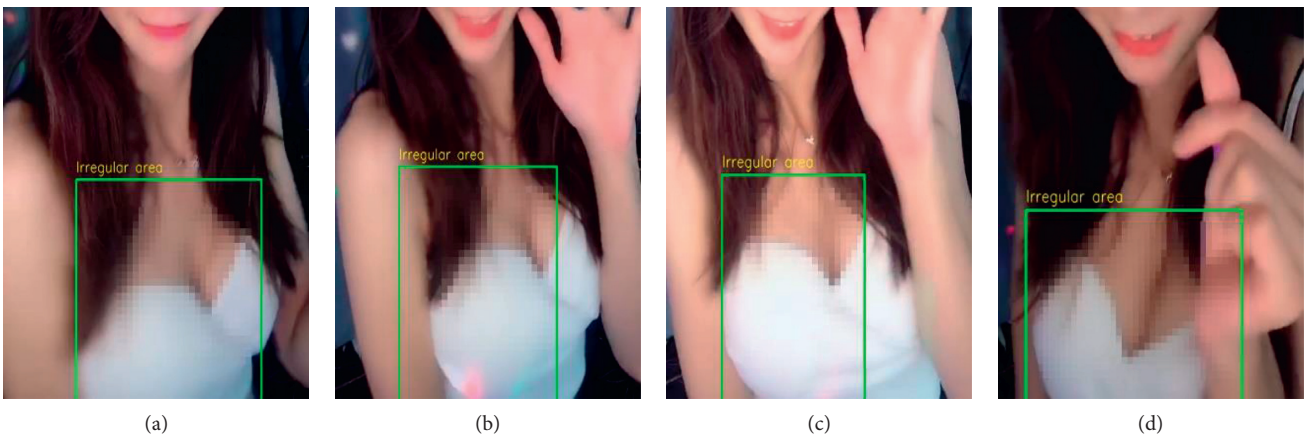
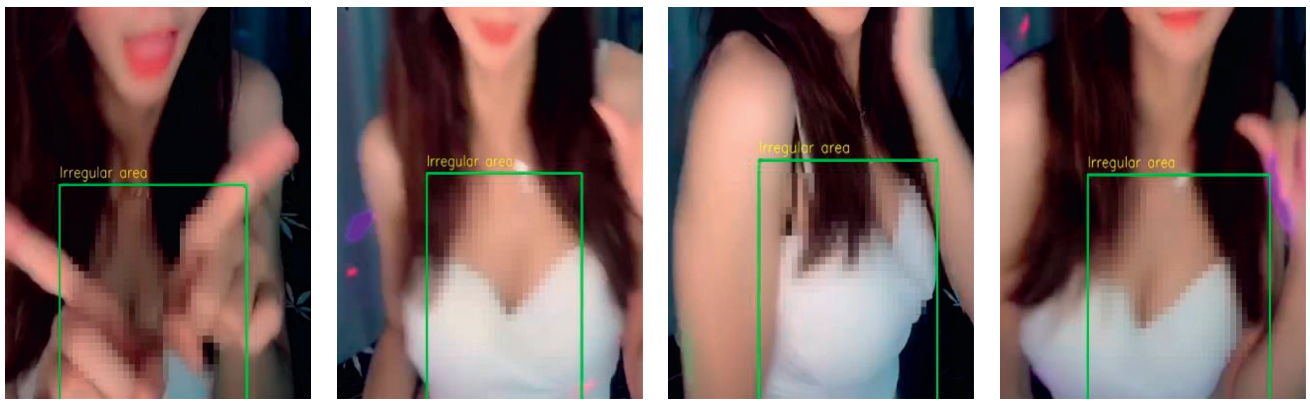


Figure 10: Continued.



(e) (f) (g) (h)

Figure 10: Example 3 of violation results detected.

5.4. Experimental Environment. The compiler used in this experiment is Matlab and Pycharm; operating system is Windows 10, 64 bit system, Linux 16.5; version of language is Matlab2018a, python3.7, and running environment is Intel Core i7 CPU, 32 GB RAM.

5.5. Experimental Results and Analysis. According to the evaluation criteria of the abovementioned experiments, the three models are now tested separately, and the test results of 6 models are shown in Table 2.

It can be seen from the multiple indicators in Table 2 that ResNet-50 has better recognition effect than other algorithms. In this paper, three illegal videos are selected and continuous frames are detected for the three videos. The specific video detection results are as follows.

The detection result of violation video 1 is shown in Figure 8.

The detection result of violation video 2 is shown in Figure 9.

The detection result of violation video 3 is shown in Figure 10.

6. Conclusions and Improvements

6.1. Conclusions. With the rapid development of online live broadcast, live broadcast has become an indispensable part of people's entertainment and enriches people's lives. In practical application scenarios, this research can be used to detect illegal video on live broadcast platforms and assist network police in intelligent detection of pornographic videos. Violation video detection can be added as a module to the network public security big data platform to capture real-time video from the network in real time and detect whether the video is illegal. If a video violation is detected, an alert will be issued immediately to assist the public security in maintaining the network environment. However, the quality and quality of live broadcast does not have a good detection platform. It currently stays at the stage of manual detection. Rely on manual testing. Then, the content and quality of the live broadcast will be uneven, especially the pornographic live broadcast, which will affect the user's visual experience and will endanger the health of children. Therefore, this paper proposes a new detection strategy based on the residual network (ResNet-50), which detects the live broadcast content of the anchor and improves the efficiency of detecting pornographic live broadcasts. As can be seen from the results, the accuracy of the detection of illegal and qualified pictures has reached more than 95%. The detection strategy proposed in this paper is better than traditional methods, especially for deep networks. The addition of residual network modules will increase the network's generalization level and reduce the complex calculation process of deep networks. ResNet-50 is a relatively mature network structure. According to the experimental results, it can be seen that the network model performs relatively well for various indicators of video detection.

6.2. Outlook and Improvement. Judging from the detection results and experimental accuracy, the algorithm in this article, based on the ResNet-50 network, shows good performance for video detection, but this algorithm also has some problems. Because many datasets are not public and the datasets are not complete, the accuracy of detection needs to be further improved. This article focuses on some detections of the video screen during live broadcast, and the target area is relatively small. It may also be possible to detect and recognize voice and text simultaneously in the future. For these problems, we can make related improvements in the future to improve the detection system. In traditional convolutional neural networks, convolution kernel size must be set in each convolution. In other words, if the size of the convolution kernel is variable, it may be useful to detect, rather than when it is fixed.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by the National Social Science Fund Project under Grant 17BGL102; Excellent Project of Jiangsu Province Social Science Union under Grant 15SYC-043; Soft Science Research of Wuxi Science and Technology Association under Grant KX15-B-01; and Fundamental Research Funds for the Central Universities under Grant 2015ZX18.

References

[1] H. Gao, W. Huang, and X. Yang, "Applying probabilistic model checking to path planning in an intelligent transportation system using mobility trajectories and their statistical data," *Intelligent Automation and Soft Computing (Autosoft)*, vol. 25, no. 3, pp. 547–559, 2019.

[2] H. Gao, W. Huang, Y. Duan, X. Yang, and Q. Zou, "Research on cost-driven services composition in an uncertain environment," *Journal of Internet Technology (JIT)*, vol. 20, no. 3, pp. 755–769, 2019.

[3] X. Ma, H. Gao, H. Xu, and M. Bian, "An IoT-based task scheduling optimization scheme considering the deadline and cost-aware scientific workflow for cloud computing," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, 2019.

[4] H. Gao, Y. Duan, L. Shao, and X. Sun, "Transformation-based processing of typed resources for multimedia sources in the IoT environment," *Wireless Networks*, 2019.

[5] H. Gao, Y. Xu, Y. Yin, W. Zhang, R. Li, and X. Wang, "Context-aware QoS prediction with neural collaborative filtering for internet-of-things services," *IEEE Internet of Things Journal*, 2019.

[6] J. Yu, L. Jing, Y. Zhou, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[7] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with and weighted triplet loss for place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[9] Y. Yin, J. Xia, Y. Li, y. Xu, W. Xu, and L. Yu, "Group-wise itinerary planning in temporary mobile social network," *IEEE Access*, vol. 7, pp. 83682–83693, 2019.

[10] Y. Chen, S. Deng, H. Ma et al., "Deploying data-intensive applications with multiple services components on edge," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 426–441, 2019.

[11] C. Zhang, H. Zhao, and S. Deng, "A density-based offloading strategy for IoT devices in edge computing systems," *IEEE Access*, vol. 6, pp. 73520–73530, 2018.

[12] L. Kuang, X. Yan, X. Tan, S. Li, and X. Yang, "Predicting taxi demand based on 3D convolutional neural network and multi-task learning," *Remote Sensing*, vol. 11, no. 11, p. 1265, 2019.

[13] J. Pei, J. Han, B. Mortazavi-Asl et al., "PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth," in *Proceedings of the International Conference on Data Engineering*, p. 215, IEEE Computer Society, Washington, DC, USA, 2001.

[14] J. Ayres, J. Flannick, J. Gehrke et al., "Sequential pattern mining using a bitmap representation," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Edmonton, Canada, pp. 429–435, July 2002.

[15] B. Le, M. T. Tran, and D. Tran, "A method for early pruning a branch of candidates in the process of mining sequential patterns, intelligent information and database systems, ACIIDS 2017," *Lecture Notes in Computer Science*, p. 10191, Springer, Berlin, Germany, 2017.

[16] M. Zihayat, Y. Chen, and A. An, "Memory-adaptive high utility sequential pattern mining over data streams," *Machine Learning*, vol. 106, no. 6, pp. 799–836, 2017.

[17] P. Qian, Y. Jiang, Z. Deng et al., "Cluster prototypes and fuzzy memberships jointly leveraged cross-domain maximum entropy clustering," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 181–193, 2015.

[18] P. Qian, Y. Jiang, S. Wang et al., "Affinity and penalty jointly constrained spectral clustering with all-compatibility, flexibility, and robustness," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 5, pp. 1123–1138, 2016.

[19] P. Qian, K. Zhao, Y. Jiang et al., "Knowledge-leveraged transfer fuzzy C-means for texture image segmentation with self-adaptive cluster prototype matching," *Knowledge-Based Systems*, vol. 130, pp. 33–50, 2017.

[20] P. Qian, C. Xi, M. Xu et al., "SSC-EKE: semi-supervised classification with extensive knowledge exploitation," *Information Sciences*, Information Sciences, vol. 422, , pp. 51–76, 2018.

[21] P. Qian, S. Sun, Y. Jiang et al., "Cross-domain, soft-partition clustering with diversity measure and knowledge reference," *Pattern Recognition*, vol. 50, pp. 155–177, 2016.

[22] P. Qian, J. Zhou, Y. Jiang et al., "Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes," *IEEE Access*, vol. 6, pp. 28594–28610, 2018.

[23] P. Qian, F. L. Chung, S. Wang, and Z. Deng, "Fast graph-based relaxed clustering for large data sets using minimal enclosing ball," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 672–687, 2012.

[24] Y. Jiang, D. Wu, Z. Deng et al., "Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 12, pp. 2270–2284, 2017.

[25] Y. Jiang, Z. Deng, F.-L. Chung et al., "Recognition of epileptic EEG signals using a novel multiview TSK fuzzy system," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 1, pp. 3–20, 2017.

[26] Y. Jiang, F. L. Chung, S. Wang, Z. Deng, J. Wang, and P. Qian, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 688–701, 2014.

[27] Y. Jiang, F. L. Chung, H. Ishibuchi, Z. Deng, and S. Wang, "Multitask TSK fuzzy system modeling by mining intertask common hidden structure," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 534–547, 2015.

[28] D. A. Forsyth and M. Fleck, "Finding naked people," in *Proceedings of the 4th European Conference on Computer Vision*, no. 2, pp. 593–602, Cambridge, UK, April 1996.

[29] M.-H. Yang and N. Ahujia, "Gaussian mixture model for human skin color and its application in image and video database," *SPIE Storage and Retrieval for Image and Video Database*, vol. 3656, pp. 45–466, 1999.

[30] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 1, no. 46, pp. 81–96, 2002.

[31] C. Srisaan, "A classification of Internet pornographic images," *International Journal of Electronic Commerce Studies*, vol. 7, no. 1, pp. 95–104, 2016.

[32] J. A. M. Basilio, G. A. Torres, G. S. Perez et al., "Explicit content image detection," *Signal & Image Processing*, vol. 1, no. 2, pp. 47–58, 2011.

[33] S. Karavarsamis, N. Ntarmos, K. Blekas, and I. Pitas, "Detecting pornographic images by localizing skin ROIs," *International Journal of Digital Crime and Forensics*, vol. 5, no. 1, pp. 39–53, 2013.

[34] X. Wang, C. Hu, S. Yao et al., "An adult image recognizing algorithm based on naked body detection," in *Proceedings of the ISECS International Colloquium on Computing, Communication, Control, and Management*, IEEE, Sanya, China, pp. 197–200, August 2009.

[35] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Proceedings of the 19th International Conference on Pattern Recognition*, IEEE, Tampa, FL, USA, pp. 1–4, December 2008.

[36] M. N. Moustafa, "Applying deep learning to classify pornographic images and videos," in *Proceedings of the Pacific-rim Symp on Image and Video Technology, PSIVT*, Auckland, New Zealand, November 2015.

[37] *NPDI Pornography Database*, https://sites.google.com/site/pornograohydatabase/, 2013.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classsificaton with deep convolution neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, NIPS, Lake Tahoe, NV, USA, 2012.

[39] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, pp. 1–9, June 2015.

[40] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, "Adult content detection in videos with convolutional and recurrent neural networks," *Neurocomputing*, vol. 272, no. 272, pp. 432–438, 2018.

[41] M. Perez, S. Avila, D. Moreira et al., "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, vol. 230, no. 230, pp. 279–293, 2017.

[42] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, Germany, 1995.

[43] L. V. Fulton, D. Dolezel, J. Harrop, Y. Yan, and C. P. Fulton, "Classification of alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50," *Brain Sciences*, vol. 9, no. 9, 2019.

[44] I. Amerini, C.-T. Li, and R. Caldelli, "Social network identification through image classification with CNN," *IEEE Access*, vol. 7, pp. 35264–35273, 2019.

[45] A. S. Hess and J. R. Hess, "Principal component analysis," *Transfusion*, vol. 58, no. 7, pp. 1580–1582, 2018.

[46] L. Xie, J. Z. Li, U. Zeng, and U. Kruger, "Block adaptive kernel principal component analysis for nonlinear process monitoring," *AIChE Journal*, vol. 62, no. 12, pp. 4334–4345, 2016.

[47] Y. Zhou, X. Wang, M. Zhang, J. Zhu, R. Zheng, and Q. Wu, "MPCE: a maximum probability based cross entropy loss function for neural network classification," *IEEE Access*, vol. 7, pp. 146331–146341, 2019.

[48] H. Nezamabadi-pour and E. Kabir, "Image retrieval using histograms of uni-color and bi-color blocks and directional changes in intensity gradient," *Pattern Recognition Letters*, vol. 25, no. 14, pp. 1547–1557, 2004.

[49] Z. Kolar, H. Chen, and X. Luo, "Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images," *Automation in Construction*, vol. 89, pp. 58–70, 2018.

Hindawi
Scientific Programming
Volume 2020, Article ID 4561834, 7 pages
https://doi.org/10.1155/2020/4561834

Research Article

Study on Evaluation Model of Chinese P2P Online Lending Platform Based on Hybrid Kernel Support Vector Machine

Shuang Pan,[1] **Jianguo Wei**,[1] **and Hao Pan**[2]

[1]*School of Economics, Wuhan University of Technology, Wuhan 430070, China*
[2]*School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China*

Correspondence should be addressed to Shuang Pan; panshuang@whut.edu.cn

Received 26 February 2020; Accepted 30 March 2020; Published 8 May 2020

Academic Editor: Chenxi Huang

Accurate evaluation of the risk level and operation performances of P2P online lending platforms is not only conducive to better functioning of information intermediaries but also effective protection of investors' interests. This paper proposes a genetic algorithm (GA) improved hybrid kernel support vector machine (SVM) with an index system to construct such an evaluation model. A hybrid kernel consisting of polynomial function and radial basis function is improved, specifically kernel parameters and the weight of two kernels, by GA method with excellent global optimization and rapid convergence. Empirical testing based on cross-sectional data from Chinese P2P lending market demonstrates the superiority of the improved hybrid kernel SVM model. The classification accuracy of credit risk level and operation quality is higher than the single kernel SVM model as well as the hybrid kernel model with empirical parameter values.

1. Introduction

Chinese P2P online lending industry was once without supervision and regulation for more than five years so that most platforms act as credit intermediaries, providing credit enhancement measures such as principal guarantees and third-party guarantees [1, 2]. With increasing events in bankruptcy and disappearance of platforms, investors are more and more sensitive to platform characteristics in decision-making. Risk management focusing on platforms shall be a new trend for regulation of the P2P online lending industry [3, 4]. Interim Measures for the Administration of the Business Activities of Online Lending Information Intermediary Institutions issued jointly by four ministries and commissions of Chinese government in August 2016 clarified contents of P2P lending, regulatory system, and business rules; subsequently, a series of detailed rules and regulations on third-party depository, filing and registration, and information disclosure were promulgated to standardize the development of P2P online lending industry [5, 6]. Accurate evaluation of risk level and operation performances of platforms not only provides solid basis for practical measures adoption by regulation authorities but also acts as an important reference for investors' decision-making. Therefore, constructing an advanced evaluation model for P2P online lending platforms is of vital realistic significance [7].

Risk level and operation performance evaluation are hotspot issues in recent research studies given unstable market environment. Tsolas applied a new series two-stage DEA method while evaluating credit risk of enterprises [8]. Luo Sirong et al. introduced a regression spline-based discrete time survival model to assess comprehensive performance of credit card applicants [9]. Dahira et al. presented a feature selection-based hybrid-bagging algorithm (FS-HB) for improved credit risk evaluation [10]. With respect to Chinese P2P platforms, existing research studies usually adopt statistical methods such as factor analysis, principal component clustering, and analytic hierarchy process. Zhu Zongyuan and Wang Jingyu perform the analytic hierarchy process and data envelopment analysis to measure the technical, scale, and overall efficiencies of 22 P2P online lending platforms, finding those efficiencies to be generally low [11]. Shan Peng et al. successively apply the factor

analysis method to scoring and sequencing comprehensive strength and risk levels of the sample platforms [12]. Yan Xin et al. constructed a complex evaluation index system for P2P online loan platforms and utilized the two steps and Kohonen model to cluster 516 platforms for classification and providing references for investors' decision-making [13]. Liu Ao et al. determined optimal weights by means of the teaching and learning optimization algorithm and sorted efficiencies of 100 P2P online loan platforms [14].

There are mainly two defects in the existing research studies. Firstly, in most research studies, platforms are ranked according to certain criterion. The boundary of suitable platforms for investment is ambiguous, whereas an intuitive support for investors' decision-making is missing. Secondly, for researches adopting statistical models, data modeling is overemphasized so that accuracy of the model-based prediction will be affected, while data dimension is enlarged. Therefore, a machine learning algorithm integrating GA and hybrid kernel SVM is proposed in this study. The improved algorithm sets a clear boundary of whether the platform is credible that investors could trade on by classification of risk level and operation quality. Moreover, applying the GA method and hybrid kernel SVM will not only reach a higher classification accuracy than statistical and traditional machine learning models but also fit for large data volume analysis.

The rest of this paper is organized as follows. Section 2 discusses design of evaluation model for GA optimized hybrid SVM. Section 3 shows simulation experimental results, including the labeling process by principal component method and platform evaluation process by the optimized hybrid SVM method. Section 4 concludes the paper with summary and future research directions.

2. Principle of GA and Hybrid Kernel SVM Integrating Model

2.1. Establishment of SVM Hybrid Kernel. The principle of SVM as a classification algorithm is to find the separate hyperplanes with the maximum margin to maximize the distance between point x and hyperplane $(w^T x + b = 0)$. Slack variable, namely, a nonnegative parameter ξ, and penalty factor C are introduced to describe inseparability losses and penalty for sample misclassification. While the training samples are assumed as $\{x_i, y_i\}$ (x_i: input index and y_i: classification tag value), the basic model can be described as

$$\begin{aligned} &\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i, \\ &s.t.\ y_i\left(w^T x_i + b = 0\right) \geq 1 - \xi_i, \quad i = 1, \ldots, m. \end{aligned} \tag{1}$$

The kernel function is to map the data implicitly to the high-dimensional feature space so that linear inseparable issue in original low-dimensional space may be solved, whose form and parameter value significantly influence the classification accuracy of the SVM algorithm. The kernel function may generally be divided into two types (global and local kernels); the former has strong generalization capacity but weak learning ability, while the latter is opposite. Among common kernel functions, global kernel functions include the polynomial and Sigmod types and RBF type belongs to local kernel functions. The polynomial and RBF kernel functions were linearly combined in this study to obtain a hybrid kernel function which has both learning and generalization capacities to overcome limitations of the single kernel functions. Mathematical expressions are as follows:

Polynomial kernel function: $k(x_i, x_j) = (a{x_i}^T x_j + c)^d$

RBF kernel function: $k(x_i, x_j) = \exp(-\| x_i - x_j\|^2/\sigma^2)$

Polynomial-RBF hybrid kernel function: $k(x_i, x_j) = (1-\lambda)(a{x_i}^T x_j + c)^d + \lambda \exp(-\| x_i - x_j\|^2/\sigma^2)$

2.2. Optimization of SVM Parameters. While the hybrid kernel function is applied for classification, those parameters to be necessarily determined include λ (hybrid kernel weight coefficient), a, c, and d (polynomial kernel parameters), σ^2 (RBF kernel parameter), and C (penalty factor).

Firstly, the hybrid kernel weight coefficient is determined by principle of minimizing featured distances between similar samples and maximizing featured distances between dissimilar samples, which was put forward by Wang Xingfu and Yu Lu [2]. Evaluation function L (λ) is defined as the difference between spacing of any two dissimilar samples or any two similar samples; ϕ_1 and ϕ_2 represent the corresponding mappings of RBF and polynomial kernel functions, respectively. The distance between sample i and j may be expressed as

$$\begin{aligned} &d(\lambda) = A\lambda^2 + B\lambda + C, \quad 0 \leq \lambda \leq 1, \\ &A = S - 2k_1\left(x_i, x_j\right) + 2, \\ &B = -2S, \\ &C = S, \\ &S = k_2(x_i, x_i) - 2k_2\left(x_i, x_j\right) + k_2\left(x_j, x_j\right). \end{aligned} \tag{2}$$

Then,

$$\max L(\lambda) = \sum_{i=1}^{m}\sum_{j=1}^{i-1} d(\lambda)|_{y_i y_j = -1} - \sum_{i=1}^{m}\sum_{j=1}^{i-1} d(\lambda)\Big|_{y_i y_j = 1} = -\sum_{i=1}^{m}\sum_{j=1}^{i-1} d(\lambda) y_i y_j, \tag{3}$$

where x_i stands for the sample value and y_i stands for the sample type.

Plugging equation (2) in equation (3),

$$\lambda = \frac{\sum_{i=1}^{m}\sum_{j=1}^{i-1} S y_i y_j}{\sum_{i=1}^{m}\sum_{j=1}^{i-1}\left[S - k_1\left(x_i, x_j\right) + 2\right] y_i y_j}. \tag{4}$$

Secondly, GA with global optimization ability is used to optimize kernel parameters, and its basic principles are as follows:

(1) Initialization of SVM parameters and setting searching space for kernel parameters and the penalty facto and initialization of GA parameters, population size, encoding lengths, crossover and

mutation probability, and maximum number of iterations.

(2) Random selection of the number of individuals of the initial population for coding is based on the following equation:

$$M = \frac{(x-a)(2^l-1)}{b-a}, \quad (5)$$

where M represents a binary code string; x represents the independent variable, whose value range is [a, b]; and l represents the encoding length.

(3) Calculation of f (individual fitness) and marking the individual with highest fitness.

(4) Selection, crossing, and mutation: selection refers to selecting two parent individuals from the population in accordance with the principle that the greater the fitness, the higher the probability of being selected; crossing refers to offspring forming through random code exchanges of two parent individuals; mutation refers to flipping each bit of the parent's individual codes under a certain probability.

(5) Calculation of the fitness value of each individual according to the fitness function and decoding the individual with the highest fitness and output optimal SVM parameters. If the termination condition is not satisfied, Step (3) continues till the termination condition (the evolutional generation peaks or the individual fitness (f) converges to a certain value) is met.

3. Simulation and Tests

3.1. Construction of the Comprehensive Evaluation System and Index Preprocess. This study focused on evaluation of the monthly operation level of a P2P platform with reference to the industry average level by taking data availability and index stability into account, and the evaluation indexes were selected in the following four dimensions:

(1) Transaction level: it was decomposed in two subdimensions, trading scale and cost of capital, in which 3 indexes (namely, turnover, average reference rate of return, and net capital inflow) were examined.

(2) Platform popularity: it is primarily to examine the platform's attractiveness to investors and borrowers through the brand effects, public opinion communication, and other channels, and it is directly reflected by numbers of investors and borrowers, investment, and loan amount per capital.

(3) Loan decentralization: the explosive increasing in the trading volume and the high concentration of borrowing transactions leads to extensive payment pressure of platforms. This study focused on the degree of decentralization of borrowers; thus, two indexes (the per capita amount to be paid and the percentage of the amount to be paid by top ten borrowers) are selected for representation.

(4) Liquidity level: it refers to the ability of liquidating any assets at a reasonable price. As for any asset, the worse its liquidity is, the less active its transaction is. The average loan term is generally utilized to reflect the liquidity level, and the shorter the term is, the stronger the fund liquidity is.

Our platform and industry data are derived from statistics results for October 2017 of Website (http://www.wangdaizhijia.com), and 463 valid samples were obtained after deleting those samples whose data are incomplete. Software environment: WINDOWS 7/SPSS 19.0/Matlab R2016b. The statistical description of original indexes is shown in Table 1.

Original indexes are preprocessed in two steps: relativization and reversing negative indexes. Due to the imperfect supervision system of the Chinese P2P industry, regulatory authorities have bound neither cap nor floor for platform operation indexes. In this paper, the ratio of the absolute value and the industry average acts as input indexes for the kernel principal component analysis, which represents a relative level against the industry in an economic sense. Due to the lack of industry statistics of the index X9, a proportion of 50% is used here, which is the cap proportion authoritatively set for commercial banks in China.

Ten original indexes consist of positive and negative ones. The latter includes the per capita amount to be paid, the percentage of the amount to be paid by top ten borrowers, and the average loan term, whose absolute values have negative correlations with the operation level of a platform. Thus, the reciprocal of original negative indexes is adopted to unify the dependency of index value and the platform operation level.

3.2. Classification Evaluation Mechanism Based on Principal Component Analysis. To begin with, sample data are scored and labeled using the principal component analysis method to generate output results of the supervised learning of the SVM algorithm. The corresponding results are shown in Table 2.

The top six components whose accumulated variance contribution rates are up to 85% are extracted as the principal components, namely, as F1, F2, ..., and F6 in sequence. The score matrix is shown in Table 3.

Each component is expressed as a linear combination of index (X) according to the following equation, whose coefficient matrix is the score matrix of principal components in Table 3:

$$F(i) = \sum_{n=1}^{10} V_{ni} X(n), \quad i = 1, 2, \ldots, 6. \quad (6)$$

The comprehensive score function was established as follows, which is a weighted sum of scores of all principal components; and the weight is the corresponding variance contribution rate for each principal component:

Table 1: Original indexes and descriptive statistics.

Evaluation dimension	Name of index	Average	Standard deviation	Maximum	Minimum
Transaction level	X1: turnover (RMB 10,000 yuan)	31323.84	95231.63	896003.60	50.74
	X2: average reference rate of return (%)	10.73	2.49	21.72	4.68
	X3: net capital inflow (RMB 10,000 yuan)	379.96	3089.46	18542.88	−46473.09
Platform popularity	X4: investors (persons)	9202.03	40772.65	688387	8
	X5: borrowers (persons)	21862.80	161150.30	2986704	1
	X6: per capita investment amount (RMB 10,000 yuan)	21.73	309.09	6653.67	0.41
	X7: per capita loan amount (RMB 10,000 yuan)	585.02	3408.46	51960	0.13
Loan decentralization	X8: per capita amount to be paid (RMB 10,000 yuan)	5879.13	47507.22	632460	0.10
	X9: percentage of the amount to be paid by top ten borrowers (%)	33.01	33.99	100	0.01
Liquidity level	X10: average loan term (month)	5.80	5.32	35.82	0.55

Table 2: Principal component analysis results.

Component	Eigenvalue	Variance contribution rate (%)	Accumulated variance contribution rate (%)
1	2.601	26.464	26.464
2	1.766	17.659	44.123
3	1.170	11.696	55.819
4	1.110	11.100	66.919
5	0.996	9.956	76.875
6	0.891	8.910	85.785
7	0.592	5.916	91.702
8	0.584	5.842	97.543
9	0.136	1.363	98.906
10	0.109	1.094	100

Table 3: Principal component score matrix.

Index	Principal component					
	F1	F2	F3	F4	F5	F6
X1	0.315	0.044	−0.041	−0.045	0.080	−0.127
X2	−0.037	0.014	0.086	0.687	−0.186	0.643
X3	0.110	−0.430	0.345	−0.075	0.191	0.014
X4	−0.050	0.215	0.605	−0.205	0.190	0.236
X5	0.354	−0.081	0.026	−0.019	−0.071	0.109
X6	0.200	0.412	−0.232	0.123	−0.068	−0.139
X7	0.068	0.379	0.346	−0.081	0.142	0.073
X8	0.021	−0.010	−0.110	0.390	0.874	−0.141
X9	0.302	−0.113	0.055	0.031	−0.089	0.246
X10	−0.018	0.023	−0.414	−0.452	0.299	0.720

$$F = 0.265F1 + 0.17659F2 + 0.117F3 + 0.111F4 + 0.099F5 + 0.089F6. \quad (7)$$

While X(i) is taken as 1 for any i, the industry average score is calculated as $\overline{F} = 0.812$. While the comprehensive score $F \in (-\infty, \overline{F})$, the platform shall be below the industry average level and it belongs to the "ALERT" type platforms and is labeled with "−1". In contrast, while X(i) is taken as 10 for any i, its "EXCELLENT" type score is calculated as $F^* = 9.973$. While the comprehensive score $F \in (F^*, +\infty)$, it shall be labeled with "1". While the comprehensive platform score $F \in (\overline{F}, F^*]$, it belongs to the "GENERAL" type platforms and it shall be labeled with "0". The principal component analysis was performed to gain the results: 107 "EXCELLENT" type platforms, 334 "GENERAL" type platforms, and 22 "ALERT" type platforms.

In addition, in order to assess the ability of early warning of optimized evaluation model, a second classification standard is constructed. A binary classifier gives a definite answer to whether investors could trade on the platform based on its risk level, which is different from the ternary classifier we built before aiming at choosing the most outstanding platforms. "EXCELLENT" and "GENERAL" platforms are collectively called "NONALERT" platforms, labeled "1" and "0" for "ALERT" platforms. Accordingly, there will be 22 "ALERT" platforms and 441 "NONALERT" platforms.

3.3. Evaluation Model for Optimization of Hybrid Kernel SVM by GA

3.3.1. Classification Evaluation Results for Determining SVM Parameters Based on Empirical Values. The empirical value parameters were first selected to test the accuracies of the single and hybrid kernel SVM models. By taking $\lambda = 0.5$, $a = c = 1$, $d = 3$, $\sigma^2 = 10$, and $C = 1$, the 5-fold cross validation binary classification and ternary classification results are shown in Table 4.

As shown in Table 4, the classification accuracy of the polynomial-RBF hybrid kernel support vector machine evaluation model with empirical parameters is slightly better than that of the four common single-core models both in binary and ternary classification. However, ternary

Table 4: Empirical SVM classification accuracies.

Kernel function	Linear kernel (%)	Polynomial kernel (%)	RBF kernel (%)	Sigmoid kernel (%)	Polynomial-RBF hybrid kernel (%)
Two-classification accuracy	92.704	93.272	92.762	91.248	94.690
Three-classification accuracy	75.162	75.594	75.162	72.138	76.890

classification results are not satisfactory especially. GA is introduced to optimize the hybrid kernel weight coefficient and kernel parameters to achieve higher classification accuracy.

3.3.2. Optimization of SVM Parameters Based on GA. Parameters are optimized by LIBSVM toolkit, and gamma = $a = 1/\sigma^2$ is taken when applying the hybrid kernel function. SVM parameters are optimized by the GA algorithm in accordance with the specific steps as follows:

Input: inputting 463 sample data after feature extraction.

Step 1: parameters are encoded in binary mode to construct a population (pop size: 50; individual chromosome length: 10). The ranges of polynomial kernel parameters are$c \in [0, 100] d \in [1, 4]$,gamma $\in [0.01, 500]$, and $C \in [0.01, 100]$. A $50 * 40$ matrix is generated randomly as the initial population.

Step 2: λ is solved based on the characteristic distance method.

Step 3: SVM classification accuracy based on the 5-fold test method is calculated and defined as the fitness function of GA.

Step 4: selection was performed by the roulette wheel selection method so that the greater the fitness of individuals, the higher the probability of being selected. The generation gap is set as 0.9, which means that 90% individuals are copied to the next generation. The probability of an individual being selected is

$$F(x_i) = \frac{f(x_i)}{\sum_{i=1}^{\text{num}} f(x_i)}. \tag{8}$$

Step 5: crossing was performed by the two-point crossover method. Two crossover points were set randomly in two paired individual encoded strings, between which some genes were exchanged. The crossover probability is pc = 0.7.

Step 6: mutation was performed by the discrete mutation method, where the mutation probability is taken as pm = 0.01.

Step 7: keep the current optimal solution and the filial generation was inserted again into the parent to generate a new population. If the number of iterations is not up to the maximum which is 100, operation shall be performed again from Step 2; otherwise, Step 8 shall be performed.

Step 8: decoded outputs $(\lambda, a, c, d, \sigma^2, C)$ and classification accuracy.

The warning capability of the optimized model for "ALERT" platforms is investigated firstly. The best binary classification accuracy (fitness) during the evolution process is shown in Figure 1. When the iteration goes to the fiftieth generation, the accuracy reaches 98.9201% and finally converges to the value, which is significantly higher than that with empirical parameters in Table 4. The ROC curve of the binary classifier is shown in Figure 2, from which we can see that the AUC value reaches 0.9817. This shows that the evaluation model of hybrid kernel SVM method optimized by genetic algorithm has outstanding warning ability for "ALERT" platforms. Optimal parameter values of binary classifier are shown in Table 5.

The fitness curve of optimized ternary classifiers is shown in Figure 3. When it evolves to the 26th generation, the ternary classification accuracy reaches 96.7603% and finally converges to the value. The accuracy is significantly higher than that of single kernel (72.14%–75.59%) and hybrid kernel support vector machines with empirical parameters (76.89%) are presented in Table 4. This shows that the GA optimized hybrid kernel SVM algorithm is effective in accurate classification of risk level and operation quality of Chinese P2P online lending platforms. Parameter values of ternary classifiers during evolution are shown in Table 6.

4. Conclusions

How P2P platforms operate is closely related to investors' fund safety and their investment decisions, which creates requirements for rating and classification of platforms. An improved hybrid kernel SVM evaluation model is put forward to effectively increase the accuracy of traditional SVM algorithm. A hybrid kernel function is introduced in which the weight is solved by the characteristic distance method and the parameter value is determined by the GA algorithm. The transaction data test indicates that this improved model has strong learning ability and generalization ability, and the prediction accuracy is significantly higher than of either single kernel SVM models or the hybrid kernel model with empirical value parameterization, which enables evaluation and classification of Chinese P2P online lending platforms to be more accurate and more objective.

Nonetheless, the premature defect of GA algorithm is not solved in this study. The improved hybrid kernel model has limited ability while exploring an unknown space as well

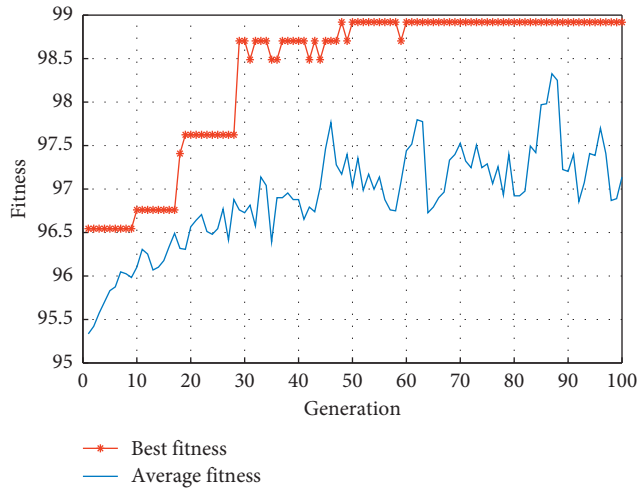


Figure 1: Fitness curve of the binary classifier.

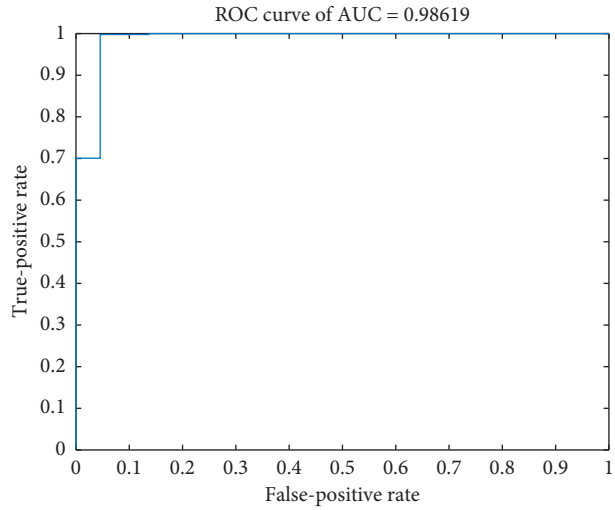


Figure 2: ROC curve of the binary classifier.

Table 5: Optimal parameter values of the binary classifier.

λ	A	c	D	σ^2	C
0.497	0.011	4.782	2.775	95.238	34.266

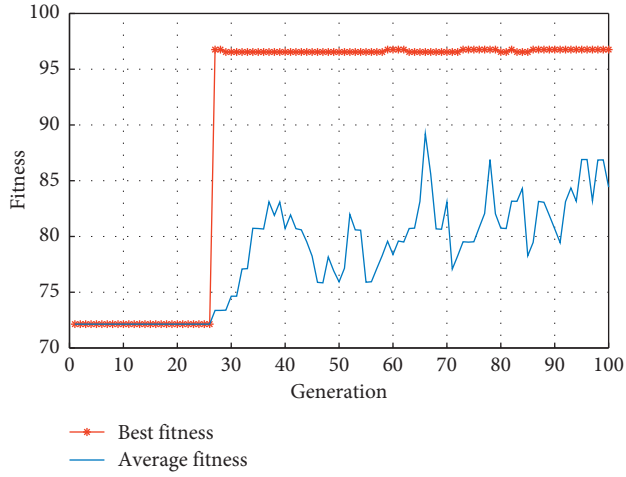


Figure 3: Fitness curve of the ternary classifier.

Table 6: Parameter values of GA optimized ternary classifier.

Generation	λ	A	c	d	σ^2	C
1	0.539	0.157	2.493	1.875	6.369	13.850
20	0.687	0.268	1.166	1.925	3.731	29.677
40	0.746	0.078	7.207	2.046	12.821	24.864
60	0.732	0.059	14.282	3.298	16.949	52.191
80	0.813	0.041	27.085	3.451	24.390	77.367
100	0.855	0.019	18.173	2.817	52.632	16.783

as the tendency to converge to a local optimal solution. Optimization could be further developed through these aspects.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Social Science Foundation of China (Grant no. 14BGL185).

References

[1] I. E. Tsolas, "Firm credit risk evaluation: a series two-stage DEA modeling framework," *Annals of Operations Research*, vol. 233, no. 1, pp. 483–500, 2015.

[2] X. Wang and L. Yu, "Weight solving method in hybrid kernel function," *Computer Systems & Applications*, vol. 24, no. 4, pp. 129–133, 2015.

[3] Y. Wang, E. Zio, X. Wei, D. Zhang, and B. Wu, "A resilience perspective on water transport systems: the case of Eastern Star," *International Journal of Disaster Risk Reduction*, vol. 33, pp. 343–354, 2019.

[4] Z. Zhu and J. Wang, "The comprehensive measurement of efficiency of P2P online loan platform-based on the method of AHP-DEA," *South China Finance*, vol. 4, pp. 31–38, 2016.

[5] W. Wei and Y. Qi, "Information potential fields navigation in wireless Ad-Hoc sensor networks," *Sensors*, vol. 11, no. 5, pp. 4794–4807, 2011.

[6] Y. Liu and H. Wang, "The risk evaluation and control of private lending in China from the perspective of Internet: taking P2P platform as an example," *Macroeconomics*, vol. 3, pp. 146–157, 2017.

[7] W. Wei, X. Fan, H. Song, X. Fan, and J. Yang, "Imperfect information dynamic stackelberg game based resource allocation using hidden markov for cloud computing," *IEEE Transactions on Services Computing*, vol. 11, no. 1, pp. 78–89, 2016.

[8] L. Dong, Q. Guo, and W. Wu, "Speech corpora subset selection based on time-continuous utterances features," *Journal of Combinatorial Optimization*, vol. 37, no. 4, pp. 1237–1248, 2019.

[9] S. Luo, X. Kong, and T. Nie, "Spline based survival model for credit risk modeling," *European Journal of Operational Research*, vol. 253, no. 3, pp. 869–879, 2016.

[10] S. Dahiya, S. S. Handa, and N. P. Singh, "A feature selection enabled hybrid-bagging algorithm for credit risk evaluation," *Expert Systems*, vol. 34, no. 6, 2017.

[11] Z. Chen, Y. Zhang, C. Wu, and B. Ran, "Understanding individualization driving states via latent dirichlet allocation model," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 41–53, 2019.

[12] P. Shan, Y. Wang, and Y. Deng, "The design and control of P2P online lending platforms' comprehensive strength rating," *Macroeconomics*, vol. 1, pp. 115–127, 2017.

[13] X. Yan, J. Sun, and L. Kang, "Inventor-oriented P2P lending platform selection: application of cluster-analysis methods," *Research on Library Science*, vol. 5, pp. 38–45, 2017.

[14] A. Liu, Z. Tong, X. Deng, and K. Liu, "An improved TOPSIS evaluation of Peer-To-Peer lending operational efficiency," *Journal of Systems Science and Mathematical Sciences*, vol. 7, pp. 1620–1632, 2017.