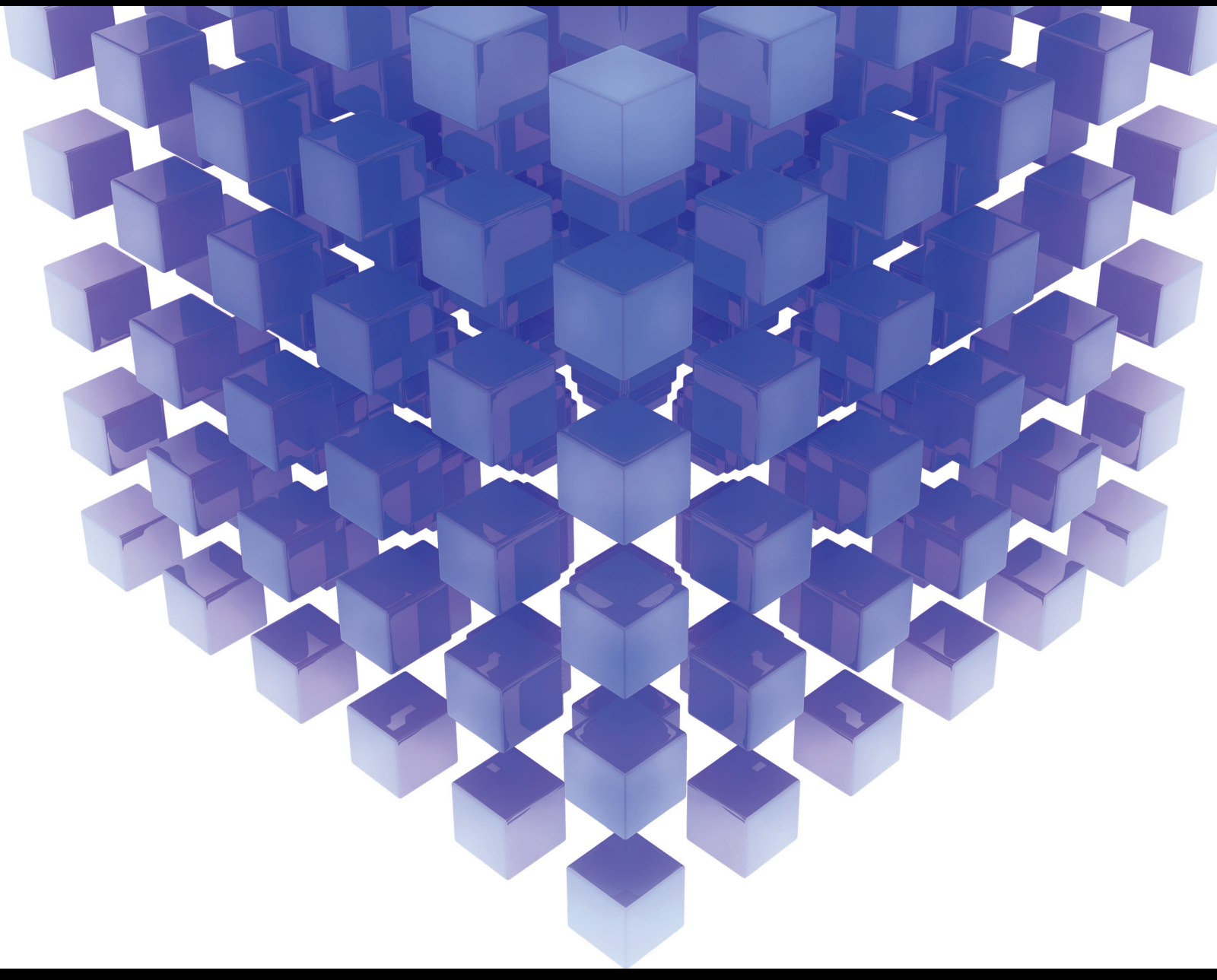
Mathematical Problems in Engineering

Mathematical and Intelligent Techniques for Data Analytics in Science and Engineering

Lead Guest Editor: William Guo
Guest Editors: Chih-Cheng Hung and Jun Shen



Mathematical and Intelligent Techniques for Data Analytics in Science and Engineering

Mathematical Problems in Engineering

Mathematical and Intelligent Techniques for Data Analytics in Science and Engineering

Lead Guest Editor: William Guo
Guest Editors: Chih-Cheng Hung and Jun Shen

Chief Editor

Academic Editors

Contents

ARIMA-FSVR Hybrid Method for High-Speed Railway Passenger Traffic Forecasting
Meng Ge, Zhang Junfeng, Wu Jinfei, Han Huiting, Shan Xinghua, and Wang Hongye
Research Article (5 pages), Article ID 9961324, Volume 2021 (2021)

Short-Term Master-Slave Forecast Method for Distributed Photovoltaic Plants Based on the Spatial Correlation
Jia Ning, Guanghao Lu, Sipeng Hao, Aidong Zeng, and Hualei Wang
Research Article (13 pages), Article ID 9922226, Volume 2021 (2021)

Experimental Research on Bearing Characteristics of the Asphalt Pavement Containing Buried Pipeline
Hailiang Xu, Jining Qin, Hehuan Ren, Jindou Sun, and Lian He
Research Article (10 pages), Article ID 6610003, Volume 2021 (2021)

An Unsupervised Intelligent Fault Diagnosis System Based on Feature Transfer
Nannan Lu, Songcheng Wang, and Hanhan Xiao
Research Article (12 pages), Article ID 6686057, Volume 2021 (2021)

Sentence Similarity Calculation Based on Probabilistic Tolerance Rough Sets
Ruiteng Yan, Dong Qiu, and Haihuan Jiang
Research Article (9 pages), Article ID 1635708, Volume 2021 (2021)

Dataset Denoising Based on Manifold Assumption
Zhonghua Hao, Shiwei Ma, Hui Chen, and Jingjing Liu
Research Article (14 pages), Article ID 6432929, Volume 2021 (2021)

An Improved Monte Carlo Method Based on Neural Network and Fuzziness Analysis: A Case Study of the Nanpo Dump of the Chengmenshan Copper Mine
Feng Gao, Xiaodong Wu, and LeWen Wu
Research Article (17 pages), Article ID 6685190, Volume 2021 (2021)

Study on Foundation Pit Construction Cost Prediction Based on the Stacked Denoising Autoencoder
Lanjun Liu, Denghui Liu, Han Wu, and Junwu Wang
Research Article (16 pages), Article ID 8824388, Volume 2020 (2020)

Traffic Flow Detection at Road Intersections Based on *K*-Means and NURBS Trajectory Clustering
Jun-fang Song, Shu-yu Wang, and Hai-li Zhao
Research Article (6 pages), Article ID 1383198, Volume 2020 (2020)

Research on Chinese Question-Answering for Gaokao Based on Graph
Zhizhuo Yang, Chunzhuan Li, Zhang Hu, Qian Yili, and Ru Li
Research Article (11 pages), Article ID 3167835, Volume 2020 (2020)

Modeling of Marine Asynchronous Shaft Generator and Simulation of Subsynchronization State
Yan Langtao, Tan Jiawan, Liu Yusheng, and Yang Hui
Research Article (11 pages), Article ID 3054969, Volume 2020 (2020)

Objective Evaluation of Drivability in Passenger Cars with Dual-Clutch Transmission: A Case Study of Static Gearshift Condition
Wei Zhou, Xuexun Guo, Xiaofei Pei, Chengcai Zhang, Jun Yan, and Jialei Xia
Research Article (13 pages), Article ID 2061083, Volume 2020 (2020)

An Intelligent Forensics Approach for Detecting Patch-Based Image Inpainting
Xinyi Wang, He Wang, and Shaozhang Niu
Research Article (10 pages), Article ID 8892989, Volume 2020 (2020)

Analyzing Machine Learning Models with Gaussian Process for the Indoor Positioning System
Yunxin Xie, Chenyang Zhu, Wei Jiang, Jia Bi, and Zhengwei Zhu
Research Article (10 pages), Article ID 4696198, Volume 2020 (2020)

Research on the Simulation of Wheelset Response Characteristic Identification of Railway Fastener Loosening
Wenbai Zhang, Lele Peng, Shubin Zheng, Xun Guo, and Yuling Wang
Research Article (15 pages), Article ID 4518624, Volume 2020 (2020)

An Intelligent Evaluation Method to Analyze the Competitiveness of Airlines
Jun Zhao and Xumei Chen
Research Article (9 pages), Article ID 8589346, Volume 2020 (2020)

Detection for Multisatellite Downlink Signal Based on Generative Adversarial Neural Network
Qing-yang Guan and Wu Shuang
Research Article (14 pages), Article ID 9765975, Volume 2020 (2020)

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9961324, 5 pages
https://doi.org/10.1155/2021/9961324

Research Article

ARIMA-FSVR Hybrid Method for High-Speed Railway Passenger Traffic Forecasting

Meng Ge, Zhang Junfeng, Wu Jinfei, Han Huiting, Shan Xinghua, and Wang Hongye

Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China

Correspondence should be addressed to Wang Hongye; llxxff_mg@126.com

Received 12 March 2021; Revised 21 April 2021; Accepted 22 May 2021; Published 30 May 2021

Academic Editor: Chih-Cheng Hung

In order to improve the prediction accuracy of railway passenger traffic, an ARIMA model and FSVR are combined to propose a hybrid prediction method. The ARIMA prediction model is established based on the known railway passenger traffic data, and then, the ARIMA prediction results are used as the training set of the FSVR method. At the same time, the air price and historical passenger traffic data are introduced to predict the future passenger traffic, to realize the mixed prediction of railway passenger traffic. The case study demonstrates that the hybrid prediction method can effectively improve the prediction performance of railway passenger traffic. Compared with the single ARIMA method, the hybrid prediction method improves the delay of the prediction results. Compared with the FSVR prediction result, the hybrid prediction method greatly reduces the errors in the extreme points of passenger traffic and long-term prediction. The relevant research results of this paper provide a useful reference for the prediction of railway passenger traffic.

1. Introduction

At present, commonly used passenger flow prediction methods are based on historical data including time-series methods, support vector machines, and neural networks [1–3]. For instance, Ni et al. [4] applied the autoregressive moving average (ARIMA) method to solve traffic flow prediction and proved that it can solve the problem of modeling about nonstationary time-series prediction. Xie et al. [5] designed the fuzzy time-series ARIMA method for long-term waterway traffic volume prediction. Li et al. [6] proposed a robust v-support vector regression (RSVR) method to forecast vessel traffic flow. Liu et al. [7] adopted a support vector machine- (SVM-) based regression prediction to predict the bus passenger flow in the target time window. Li et al. [8] put forward a backpropagation neural network (BPNN) model with population per distance band for traffic flow prediction of urban rail transit station. Hu et al. [9] developed a model re-sample recurrent neural network (RRNN) to forecast passenger traffic on mass rapid transit systems.

Due to the different advantages and disadvantages of various prediction methods, the prediction effect of a single mechanism prediction method is often not ideal. If two or more methods are organically combined to form a hybrid prediction method, it will overcome the deficiencies of a single prediction mechanism and improve the performance of passenger flow prediction [10, 11]. Khan et al. [12] combined wavelet transform (WT) with artificial neural network (ANN) and ARIMA into a hybrid model for meteorological drought forecasting, and the model inherits the merits of both WT and ANN-ARIMA. Wu et al. [13] created a hybrid model of ARIMA and wavelet neural network (WNN) combined with genetic algorithm to predict the river water quality. Yu et al. [14] built a novel SVR-ANN combined model with EEMD for rainfall prediction. Luo et al. [15] explored a combined prediction model based on the empirical mode decomposition, support vector regression, and wavelet neural network (EMD-SVR-WNN) to forecast the structural settlement and deformation. The above models achieved satisfactory results. It can be found that SVR and neural network are suitable for solving complex nonlinear problems, and the time-series model has great advantages for time-based prediction. However, there are still some inherent defects in the neural network model, such

as ease of sinking into local optimization and the overfitting. Therefore, the SVR and time-series method are selected for hybrid prediction.

In this thesis, a combination of differential integrated moving average autoregressive model (ARIMA) and fuzzy support vector regression machine (FSVR) is used to implement a mixed forecasting strategy for railway passenger flow. And, apply it to the actual passenger flow forecast of Shanghai-Guangzhou high-speed railway in order to obtain good forecast performance. Support vector regression (SVR) is a general learning method based on the statistical learning theory of limited samples (SLT) [16]. Fuzzy support vector regression (FSVR) is a new type of support vector regression machine that combines fuzzy mathematics and support vector regression. It introduces fuzzy membership and improves the generalization of machine learning ability. According to the theory of time-series analysis, the ARIMA model is suitable for the prediction and analysis of stationary time series, and the passenger flow data is generally non-stationary series, which needs to be smoothed by difference. Therefore, the differential autoregressive moving average model (ARIMA) is used to predict passenger flow.

2. ARIMA

Differential autoregressive moving average model (ARIMA) is an important method for studying time series. In ARIMA (p, d, q), AR is autoregressive and p is the number of autoregressive items, MA is the moving average, q is the moving average item number, and d is the number of differences made to make it a stationary sequence. The ARIMA (p, d, q) model is an extension of the ARMA (p, q) model.

The basic form of the ARMA model is

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \tag{1}$$

where c is the constant, $\phi_1, \phi_2, \ldots, \phi_p$, $\theta_1, \theta_2, \ldots, \theta_q$ is the coefficient, ε_t is the white noise sequence, p is the autoregressive order, and q is the moving average order.

After passing the difference, the basic form of the ARIMA model is

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t, \tag{2}$$

where L is the lag operator and d is the difference order, $d \in Z, d > 0$.

3. Fuzzy Support Vector Regression

The principle of FSVR is to find a function by minimizing the prediction error, use the nonlinear mapping function ϕ to map the data x_i in the input space to the high-dimensional space H, and perform linear regression calculation in H to achieve the effect of nonlinear regression in the original low-dimensional space [17].

In practical applications, different data points contribute differently to the training results, so FSVR solves the problem of overlearning due to the presence of noisy data by introducing fuzzy parameters to eliminate the influence of noise [18], that is, there is a fuzzy degree and each data point is connected so that a training set with fuzzy members will be generated.

For FSVR, let the training set be $S = \{(x_1, y_1, s_1), (x_2, y_2, s_2), \cdots (x_N, y_N, s_N)\}$, where $x_i \in R^n$, $y_i \in R$, and $s_i \in [0, 1]$. In the time-series problem, the membership degree s_i is a function of the time series $t_i (1 \le i \le N)$. In this thesis, the fuzzy membership function $f(t_i)$ is the quadratic function of the time series t_i, namely, $s_i = f(t_i)$:

$$s_i = (1 - \lambda)\left(\frac{t_i - t_1}{t_N - t_1}\right)^2 + \lambda. \tag{3}$$

The boundary conditions are

$$\begin{cases} s_1 = \lambda \ (0 < \lambda \le 1), \\ s_N = f(t_N) = 1. \end{cases} \tag{4}$$

FSVR is for solving quadratic programming problems:

$$\arg \min_{\omega, b, \xi, \xi^*} \phi_L(\omega, \xi, \xi^*) = \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{N} s_i(\xi_i + \xi_i^*),$$

$$\text{s.t.} \begin{cases} y_i - \omega^T \phi(x_i) - b \le \varepsilon + \xi_i, \\ \omega^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i^*, \\ \xi_i \ge 0, \xi_i^* \ge 0, \end{cases} \tag{5}$$

where ω is the regression hyperplane weight vector, b is the deviation coefficient, C is the penalty parameter (as a constant value), ε is the regression hyperplane bandwidth, ξ_i and ξ_i^* are the relaxation variable, and s_i is the fuzzy membership.

The dual form of equation (5):

$$\arg \max_{\alpha, \alpha^*} \phi_L(\alpha, \alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i (\alpha_i - \alpha_i^*),$$

$$\text{s.t.} \left\{ \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0 \alpha_i, \alpha_i^* \in [0, s_i C]. \right. \tag{6}$$

Solving the dual problem (6), we can get the FSVR regression function:

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b. \tag{7}$$

4. Experiments

Using the high-speed rail passenger flow between Shanghai and Guangzhou as experimental data, the passenger flow is

obtained by day, a total of 176 days of sample data are collected, the first 165 days of sample data are used to build the model, and the last 8 days of sample data are used as test samples to predict comparative analysis.

In order to reduce the computational complexity and accuracy of parameter selection, the raw data is normalized. Table 1 shows part of the passenger flow data.

Using the ARIMA model to predict the values, the results are as follows.

It can be seen from the prediction results shown in Figure 1, and the ARIMA model can realize the prediction and analysis of railway passenger traffic. The fluctuation of its prediction results is consistent with the actual passenger traffic curve, but there is a large delay phenomenon which causes a large prediction error and the prediction effect is not ideal.

Based on FSVR's passenger flow prediction, the results are as follows.

It can be seen from the prediction results shown in Figure 2, and the FSVR has a strong nonlinear approximation ability; it has shown good prediction performance in the process of railway passenger traffic forecast, especially in the short-term passenger traffic forecast; its prediction error is small, and the passenger traffic continues to increase or continue. The prediction error is small during the decrease, but at the extreme point, where the passenger traffic trend changes, that is, the passenger traffic changes from increasing to decreasing, or from decreasing to increasing, the prediction error is large. In other words, the dramatic fluctuations in passenger traffic reduce the generalization ability of FSVR and affect its prediction performance.

Using the above ARIMA forecast results as the input items of FSVR, the mixed forecast of railway passenger traffic is realized. The results are as follows.

It can be seen from the prediction results shown in Figures 3 and 4 that the hybrid prediction method can combine the advantages of the two prediction methods to obtain the best prediction results. Compared with the ARIMA method, the delay of the hybrid method prediction results is greatly improved; compared with the FSVR, the prediction effect at the extreme point is significantly improved, and the prediction error is greatly reduced.

In order to prove the performance of the proposed algorithm, it is compared with the ARIMA-WNN method and the EMD-SVR-WNN method. The results of the three hybrid prediction methods are shown in Figure 5.

It can be seen from the prediction results in Figure 5 that, though the ARIMA-WNN method is accurate in the early prediction, it gradually appears the phenomenon of delay after 4 days. The overall trend of the EMD-SVR-WNN method is consistent with the original data; however, the overall predicted value is small. Compared with the above two methods, the prediction results of the ARIMA-FSVR method are more accurate. The forecast error indexes of various methods are shown in Table 2.

Table 1: Part of the passenger flow data.

SN	Date	Passenger flow	Normalization
1	20190517	1431	0.4630
2	20190518	1421	0.4475
3	20190519	1466	0.5170
4	20190520	1779	1.0000
5	20190524	1213	0.1265
6	20190525	1219	0.1358
7	20190526	1138	0.0108
8	20190527	1131	0.0000

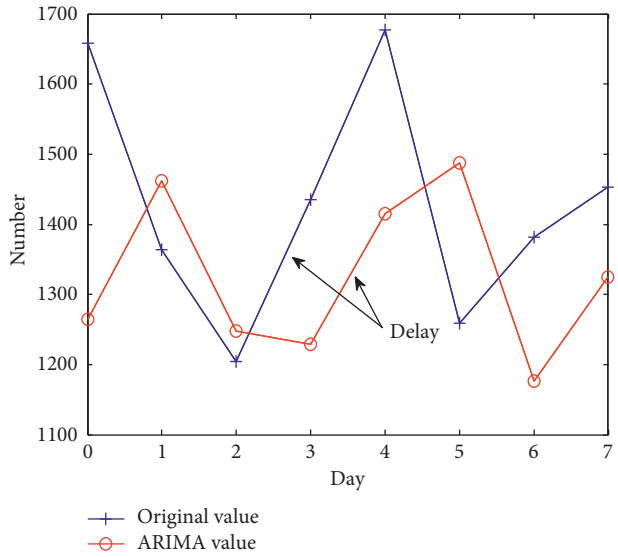


Figure 1: ARIMA results.

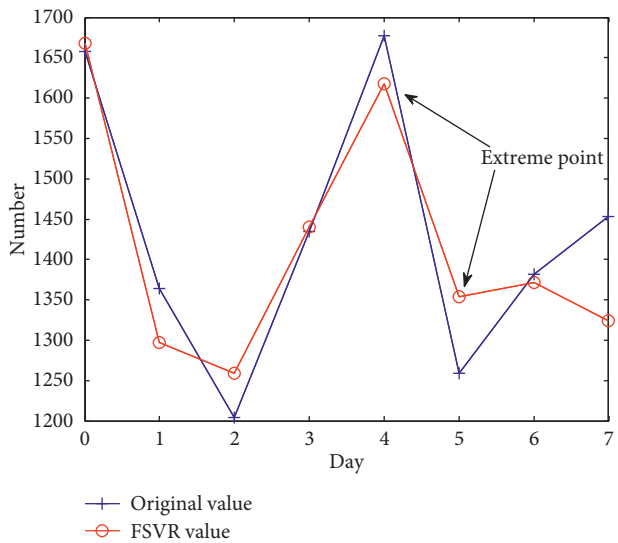


Figure 2: FSVR results.

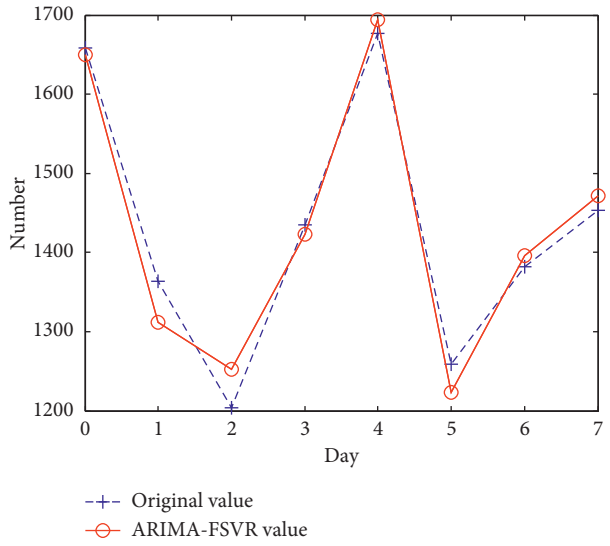


Figure 3: ARIMA-FSVR results.

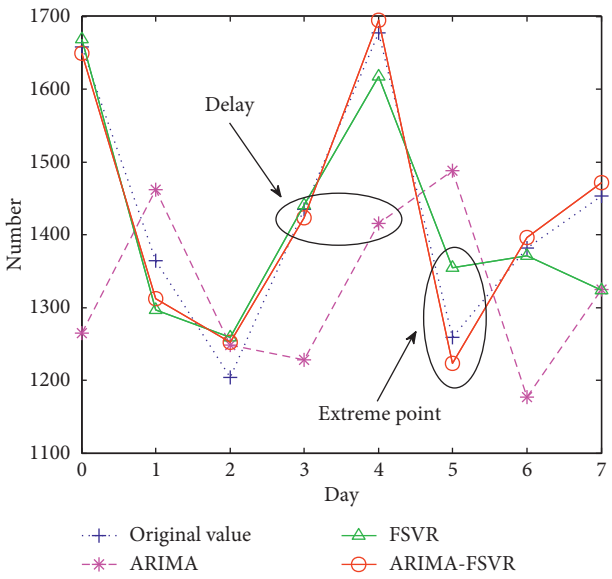


Figure 4: The ARIMA-FSVR, ARIMA, and FSVR predicted results vs. original values.

It can be seen from Table 2 that the standard error of the ARIMA-FSVR prediction is smaller than the ARIMA and FSVR methods. It is also smaller than the other two hybrid methods. The correlation coefficient of the ARIMA-FSVR method is less than 0.0001, and the P value is 0.9822. Compared with the other four methods, the correlation coefficient is larger and the P value is lower, which proves that the trend of the ARIMA-FSVR method is more accurate and can accurately predict the railway passenger traffic.

It can be found from the experimental results that the ARIMA-FSVR method can accurately predict the railway passenger traffic, handle complex nonlinear relationships, and obtain satisfactory prediction results.

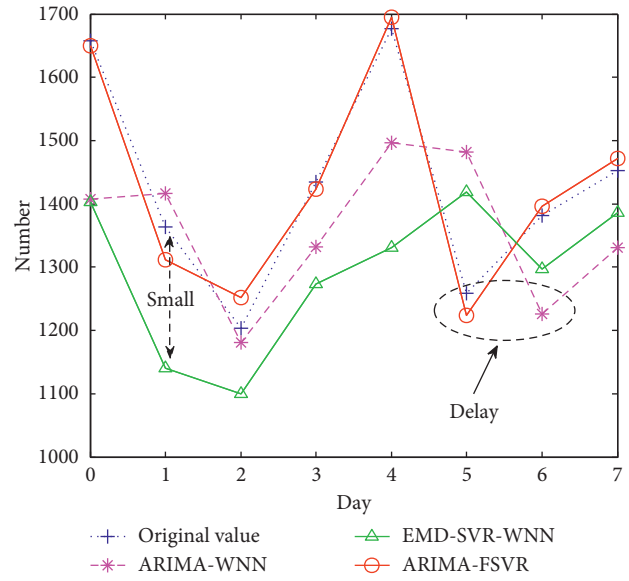


Figure 5: Three hybrid prediction results vs. original values.

Table 2: Forecast error indexes.

	ARIMA	FSVR	ARIMA-WNN	EMD-SVR-WNN	ARIMA-FSVR
RMSE	77.9183	24.0677	55.5990	69.5719	10.7347
Correlation coefficient	−0.0318	0.9065	0.4907	0.4849	0.9822
P value	0.9404	0.0019	0.2170	0.2233	0.0000

5. Conclusions

In this paper, a new hybrid method was successfully proposed which achieved great improvements regarding both the prediction accuracy and robustness of the single-item models:

(1) The ARIMA-FSVR hybrid prediction method overcame the shortcomings exposed in the single-item forecasting method, and it can improve the ARIMA delay phenomenon.

(2) The ARIMA-FSVR hybrid prediction method surmounts the extreme point problem of the FSVR method.

(3) Empirical studies on the realistic passenger flow data indicated that the ARIMA-FSVR hybrid method was clearly superior to other benchmark hybrid models. This hybrid method obtained the lowest prediction error and had higher accuracy and more reliable prediction results.

In conclusion, the ARIMA-FSVR hybrid method can accurately predict the railway passenger traffic, overcoming the shortcomings of the single-item forecasting method and, at the same time, merging the advantages of single-item forecasting and improving the accuracy of the forecast. This method effectively solves the nonlinear problem of railway

traffic data and provides a new and effective method for the nonlinear prediction problem in practical applications.

Data Availability

The case analysis data used to support this study are available from the railway passenger transport department upon request.

Conflicts of Interest

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

Acknowledgments

The project was supported by Science and Technology Research Project of Beijing-Shanghai High Speed Railway Co., Ltd. (Grant no. Beijing-Shanghai Scientific Research-2020-2), Scientific Research Projects of China Academy of Railway Sciences Co., Ltd. (Grant no. 2019YJ120), and Science and Technology Research and Development Plan of China Railway (Grant no. K2019X022).

References

[1] G. Ren and J. Gao, "Comparison of NARNN and ARIMA models for short-term metro passenger flow forecasting," in *Proceedings of the 19th COTA International Conference of Transportation Professionals g*, Nanjing, China, 2019.

[2] 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.

[3] T.-H. Tsai, C.-K. Lee, and C.-H. Wei, "Neural network based temporal feature models for short-term railway passenger demand forecasting," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3728–3736, 2009.

[4] L. Ni, X. Chen, and H. Qian, "ARIMA model for traffic flow prediction based on wavelet analysis," in *Proceedings of the 2nd International Conference on Information Science and Engineering IEEE*, pp. 1028–1031, Hangzhou, China, December 2010.

[5] Y. Xie, P. Zhang, and Y. Chen, "A fuzzy ARIMA correction model for transport volume forecast," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6655102, 10 pages, 2021.

[6] M.-W. Li, D.-F. Han, and W.-L. Wang, "Vessel traffic flow forecasting by RSVR with chaotic cloud simulated annealing genetic algorithm and KPCA," *Neurocomputing*, vol. 157, pp. 243–255, 2015.

[7] W. Liu, Q. Tan, and W. Wu, "Forecast and early warning of regional bus passenger flow based on machine learning," *Mathematical Problems in Engineering*, vol. 2020, Article ID 6625435, 11 pages, 2020.

[8] J. Li, M. Yao, and Q. Fu, "Forecasting method for urban rail Transit ridership at station level using Back propagation neural network," *Discrete Dynamics in Nature and Society*, vol. 2016, Article ID 9527584, 9 pages, 2016.

[9] R. Hu, Y. C. Chiu, C. W. Hsieh, T. H. Chang, and L. Liao, "Mass Rapid Transit system passenger traffic forecast using a Re-sample recurrent neural network," *Journal of Advanced Transportation*, vol. 2019, Article ID 8943291, 14 pages, 2019.

[10] S. Li, X. Liu, and A. Lin, "Fractional frequency hybrid model based on EEMD for financial time series forecasting," *Communications in Nonlinear Science and Numerical Simulation*, vol. 89, 2020.

[11] M. A. Jallal, A. González-Vidal, A. F. Skarmeta, S. Chabaa, and A. Zerouala, "A hybrid neuro-fuzzy inference system-based algorithm for time series forecasting applied to energy consumption prediction," *Applied Energy*, vol. 268, 2020.

[12] M. M. H. Khan, N. S. Muhammad, and A. El-Shafie, "Wavelet based hybrid ANN-ARIMA models for meteorological Drought forecasting," *Journal of Hydrology*, vol. 590, Article ID 125380, 2020.

[13] J. Wu, Z. B. Li, L. Zhu, and C. Li, "Hybrid model of ARIMA model and GAWNN for dissolved oxygen content prediction," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 48, pp. 205–210, 2017.

[14] X. Yu, G. Ling, L. He, S. Xia, and W. Wang, "A SVR-ANN combined model based on ensemble emd for rainfall prediction," *Applied Soft Computing*, vol. 73, 2018.

[15] X. Luo, W. Gan, L. Wang, Y. Chen, and X. Meng, "A prediction model of structural settlement based on EMD-SVR-WNN," *Advances in Civil Engineering*, vol. 2020, no. 4, Article ID 8831965, 11 pages, 2020.

[16] X. Luo, D. Li, and S. Zhang, "Traffic flow prediction during the holidays based on DFT and SVR," *Journal of Sensors*, vol. 2019, no. 10, Article ID 6461450, 10 pages, 2019.

[17] P. Huang, C. Wen, Li. P. Fu, Q. Y. Peng, and Z. C. Li, "A hybrid model to improve the train running time prediction ability during high-speed railway disruptions," *Safety Science*, vol. 122, 2019.

[18] T. Bahraini, S. Ghazi, and H. S. Yazdi, "Toward optimum fuzzy support vector machines using error distribution," *Engineering Applications of Artificial Intelligence*, vol. 90, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9922226, 13 pages
https://doi.org/10.1155/2021/9922226

Research Article

Short-Term Master-Slave Forecast Method for Distributed Photovoltaic Plants Based on the Spatial Correlation

Jia Ning[1,2] **Guanghao Lu**[1,2] **Sipeng Hao**[1,2] **Aidong Zeng**[1,2] **and Hualei Wang**[3]

[1]*School of Electric Power Engineering, Nanjing Institute of Technology, Nanjing 211167, China*
[2]*Jiangsu Collaborative Innovation Center for Smart Distribution Network, Nanjing 211100, China*
[3]*State Grid Lianyungang Power Supply Company, Lianyungang 222000, China*

Correspondence should be addressed to Jia Ning; ningjia@njit.edu.cn

Received 19 March 2021; Revised 17 April 2021; Accepted 10 May 2021; Published 20 May 2021

Academic Editor: Jun Shen

With the large-scale integration of distributed photovoltaic (DPV) power plants, the uncertainty of photovoltaic generation is intensively influencing the secure operation of power systems. Improving the forecast capability of DPV plants has become an urgent problem to solve. However, most of the DPV plants are not able to make generation forecast on their own due to the constraints of the investment cost, data storage condition, and the influence of microscope environment. Therefore, this paper proposes a master-slave forecast method to predict the power of target plants without forecast ability based on the power of DPV plants with comprehensive forecast system and the spatial correlation between these two kinds of plants. First, a characteristics pattern library of DPV plants is established with K-means clustering algorithm considering the time difference. Next, the pattern most spatially correlated to the target plant is determined through online matching. The corresponding spatial correlation mapping relationship is obtained by numerical fitting using least squares support vector machine (LS-SVM), and the short-term generation forecast for target plants is achieved with the forecast of reference plants and mapping relationship. Simulation results demonstrate that the proposed method could improve the overall forecast accuracy by more than 52% for univariate prediction and by more than 22% for multivariate prediction and obtain short-term generation forecast for DPV or newly built DPV plants with low investment.

1. Introduction

Electricity power consumption increases drastically in recent years, and with the decreasing supply of fossil fuels, the renewable generation, especially photovoltaic (PV) generation, has developed rapidly as well [1]. In the background of green energy strategy, the global PV installed capacity has reached 300 GW. However, the large utility-scale generation is typically deployed in rural areas, which are far from the load centers; thus, the generated power is not efficiently used. Integration of distributed PV generation with the distribution network could contribute to solving the unmatched location of generation and consumption [2]. However, distribution network is the terminal end of power system, with weak infrastructure and low reserve capacity. The increasing amount of highly intermittent and variant DPV generation will greatly affect the stability of power systems [3, 4]. Therefore, the accurate generation forecast of DPV is significant for the scheduling and stable operation of power systems. Generation forecasts could be categorized into short-term forecasts (0–72 hours ahead of the next day) and ultra-short-term forecasts (15 minutes–4 hours ahead) [5, 6]. Short-term generation forecast provides supportive data for decision-making of power system scheduling and helps improve operational reliability.

1.1. Literature Review. There is a considerable amount of scientific literature on renewable energy forecasting, and current research [7–10] on generation forecast of intermittent renewable energy has made great achievements, but the forecast methods are mostly focused on large capacity

wind and solar power plants, in which a single generating unit has an installed capacity at MW level. The renewable energy forecast methods could be classified into two major categories: time series forecast method and spatial distribution forecast method.

Time series forecast methods analyze the trends of the past to predict future events, with the assumption that future trends will hold similar to historical trends. Two numerical weather prediction models are utilized to forecast the weather variables used by the third module to predict the hourly energy production in the PV plant in [11]. Weather status pattern recognition model for short-term PV forecasting is presented using a solar irradiance feature extraction and support vector machine [12]. These references perform time series forecasting to predict weather and then obtain the short-term PV forecast power. Some other references on time series forecasting focus on different algorithms, e.g., traditional physical model prediction [13], BP-artificial neutral network (ANN) prediction with accurate numeric weather forecast [14], extreme learning machine (ELM) [15], and support vector machine (SVM) [16, 17]. In [18], a new model combines two well-known methods: the seasonal auto-regressive integrated moving average method and support vector machines method are proposed for short-term power forecasting of a grid-connected photovoltaic plant. A short-term forecasting method is presented for large-scale grid-connected PV plants using ANN in [19]. A genetic algorithm-based SVM model for short-term power forecasting of residential scale PV system is proposed in [20]. Reference [21] provides a review about the methods used to predict PV power, with the main focus being on the metaheuristic and machine learning methods. In general, these time series methods rely on a large amount of historical generation data and numerical weather forecasts and could obtain high forecast accuracy. However, the spatial characteristics of distributed PV systems are not considered. This paper focuses on distributed PV generations, which have the problem of deficient historical data, and considers their spatial distribution characteristics to realize their short-term power prediction.

Spatial distribution forecast method considers the geographic information and the spatial distribution characteristics of PV systems. The effect of spatial and spectral nonuniform irradiance distribution on multijunction solar cell performance is analyzed using an integrated approach [22], and the spatial dependence of variations for small residential PV system power output is investigated, indicating that the fluctuations are correlated up to a certain decorrelation length [23]. In [24], Karakaya applies the finite element method to forecast the diffusion of solar PV systems in time and space, in which the time-varying parameters are arduous to determine. Spatial clustering of PV systems and quantitative analysis of PV adoption drivers in the time dimension are investigated to propose a data-driven forecasting approach of PV diffusion in [25]. These references are studied to verify the spatial distribution characteristics or forecast the diffusion of PV systems. Our research is to utilize the spatial distribution characteristics for DPV power prediction.

These methods are not suitable to be applied to DPV prediction due to the data constraints and distributed characteristics of DPV [26]. In terms of data constraints, in actual DPV projects, most of the DPVs are not equipped with their own forecasting module and are not capable of storing a large amount of historical data or obtaining weather forecast data because of the limited investment.

1.2. Explanation of Spatial Correlation. In terms of distributed characteristics, the affecting factors of generation include not only natural factors such as radiation and temperature, but also the installed tilt angle, construction layout, vegetation, and microscope weather, which could vary widely even in a small range [27].

Figure 1 illustrates the spatiotemporal distribution characteristics of DPV. The DPVs are distributed in 6 areas across 3 time zones. The microscope environments in each area are different from each other, and the generation of DPV may be more closely related to its surrounding environment than the area it is in. For example, the generation pattern of the DPV in area A may be similar to that in area F, even if it is located far away and in a different time zone, because the microscope environments (shadow of obstacles, moisture, and building height) are similar. The installation details also vary, such as the tilt angle and direction. This similarity, regardless of time-space continuity, is revealed in data correlation, instead of physical connections [25]. We define this correlation as a spatial correlation as follows:

Spatial correlation refers to the numerical correlation of DPV generation at different locations. When analyzing the spatial correlation, eliminate the time difference of generation curve with data processing.

1.3. Contribution. The current technique bottleneck of DPV generation forecast is caused by data deficiency and complex influencing factors, making the traditional method of mathematically modelling infeasible in DPV forecast. A new method considering the data deficiency and spatiotemporal distribution characteristics is required to meet the need of DPV forecast. In the current installation, there are a few DPV plants with functional forecast system, which are used as reference plants in the following paper. Meanwhile, most of the DPV plants are not able to make generation forecasts on their own due to the economic and technological constraints. These plants are later referred to as target plants. According to this reality, this paper takes advantage of big data methodology and proposes a master-slave forecast technique based on spatial correlation between reference plants and target plants considering multiple affecting factors including radiation, temperature, time zone, etc., which were not studied before. The technique utilizes a master-slave forecast framework, matching the generation characteristics of target plants to reference plants using data correlation, forecasting the generation of slave target plants with the forecast data of spatially correlated master reference plants, and realizing DPV generation forecasting with data correlation relationship.

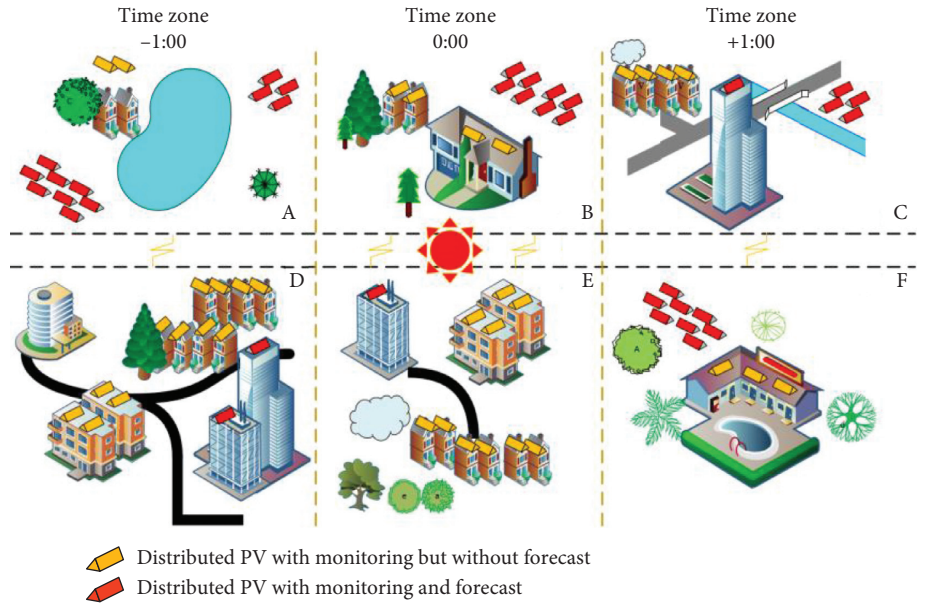


Figure 1: Spatiotemporal distribution of DPV plants considering the microscope environment.

Based on the bottleneck analysis of DPV generation forecast and the characteristics of DPV, the main contributions of this paper are listed as follows:

(1) A spatial correlation matching method is proposed to obtain the data correlation relationship across time and space between target plants and reference plants, in which the K-means clustering algorithm is utilized to cluster reference plants into groups with individual patterns on the basis of their generation characteristics. The clustering method could reduce the computation time of online matching and improve the matching accuracy.

(2) A master-slave forecast method is presented to make the generation forecast for a large number of target plants in short-term time scale, in which the LS-SVM algorithm is utilized to obtain the spatial correlation mapping relationship. Therefore, the power of target plants as slave could be predicted based on the power of reference plants as master and the spatial correlation between these two kinds of plants.

1.4. Article Organization. The following paper is composed as follows. Section 2 gives the introduction of the master-slave forecast framework. Section 3 describes the matching method for spatial correlation relationship and studies the time difference characteristics of DPV generation curves. Section 4 conducts a case study, validating the advantage of the proposed technique. Finally, Section 5 concludes the paper.

2. Framework of Short-Term Master-Slave Forecast Technique Based on Spatial Correlation

In this section, the framework of the proposed master-slave forecast technique is illustrated and explained. The forecast method is based on the spatial correlation between the generation characteristics of different DPV plants.

Data mining shows that the generation trajectories of different DPV plants in the same time dimension have a certain numerical correlation; that is, two or more numerical trajectories approximately fit in some correlation relationship. For example, Figure 2 shows the generation curves of some randomly chosen DPV plants in 3 different areas and the comparison of selected curves from all areas. It is seen that the generation curves in the same area have different shapes, while a curve might share more similarity with curves from other areas than the curves within the same area, although the DPV plants are geologically closer in one area. Therefore, the spatial correlation is defined as a numerical correlation between the generation data of different DPV plants, and the geological relationship is ignored.

The master-slave spatial correlation based forecast technique is to utilize short-term forecast of reference DPV plants (master plant) and spatial correlation relationship to forecast short-term generation of target DPV plants (slave plant) indirectly. The master-slave DPV generation forecast framework based on spatial correlation is shown in Figure 3.

As shown in Figure 3, the framework of master-slave prediction method consists of three parts, namely, left part, middle part, and right part. The left part is the forecast results

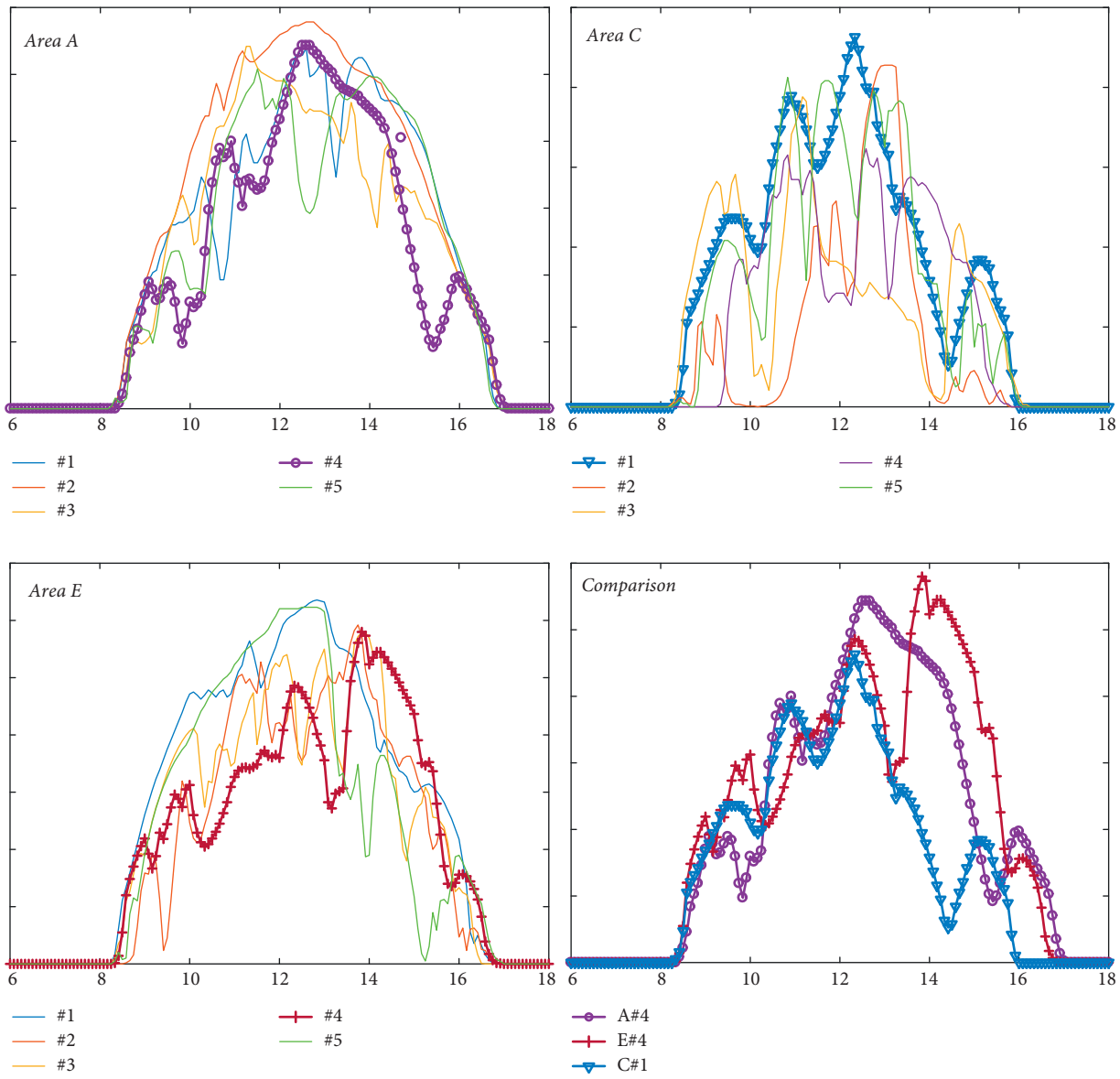


Figure 2: The principle of spatial correlation prediction.

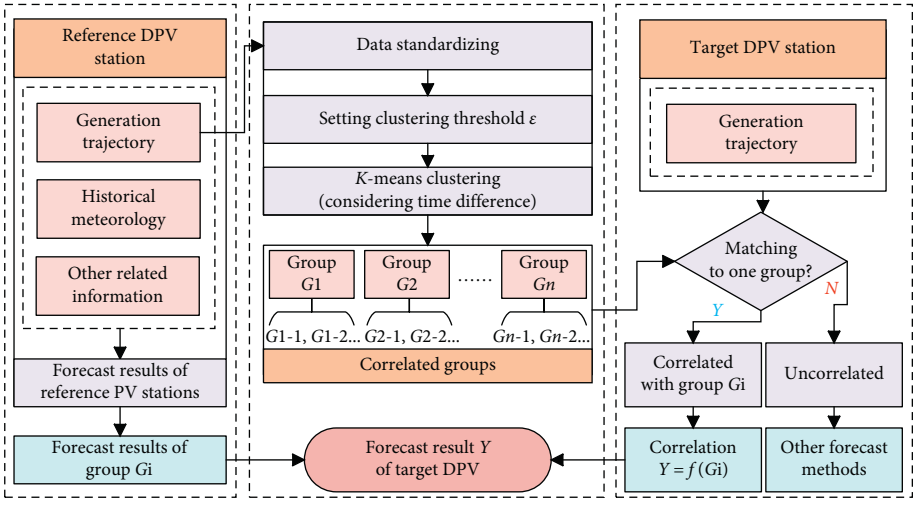


Figure 3: Framework of master-slave prediction based on spatial correlation.

of reference DPV stations. Based on the generation trajectory, historical meteorology, and other related information, the power of reference DPV plants is predicted to benefit the prediction of target DPV stations. The middle part is offline clustering of reference plants on the basis of their generation trajectory in history. Because there are a large number of reference plants, and many reference plants are spatial-correlated, a pattern library is established using offline K-means clustering to reduce the searching time for online matching [28]. The right part is an online matching process of target plants, to establish the mapping relationship of spatial correlation between target plants and patterns in library. If the matching is successful, the forecast results of target plants are obtained based on the forecast results of the correlated pattern and correlation relationship. If the matching fails, other forecast methods should be adopted.

3. Spatial Correlation Matching with K-Means Clustering

To utilize the spatial correlation between reference plants and target plants, the pattern matching method to find the correlated reference master plants with target slave plants is given in this section. K-means clustering algorithm is used to cluster master plants into groups with individual patterns according to their generation characteristics, thus constructing the standard pattern library. Next, the clustering significance index (CSI) is defined to set the cluster number, and standardized Euclidean distance (SED) is used to match the standardized data of DPV generation to the data patterns in the pattern library to establish the spatial correlation mapping. The spatial correlation matching process using K-means clustering is shown in Figure 4. The detailed corresponding algorithms are described in Sections 3.1–3.4.

3.1. Data Standardization. Data standardization is the process of data scaling and nondimensionalization so that the data could be compared. In this paper, the raw data of DPV generation A are processed row by row using normalization, with the following equation:

$$B_{ij} = \frac{\left(A_{ij} - \text{mean}\left(A_j\right)\right)}{\text{std}\left(A_j\right)}, \tag{1}$$

where A_j is the jth row of the matrix A, A_{ij} is the ith element in A_j, mean (A_j) denotes the mean value of vector A_j, std (A_j) denotes the standard deviation of vector A_j, and B_{ij} is the ith element of the jth row of matrix B. After standardization, A_j is converted to B_j. The mean of vector B_j is 0, and the variance of vector B_j is 1. Vector B_j is called a standard vector, and matrix B is the standardized matrix of A.

The data standardization could reduce the influence of DPV installed capacity difference on spatial correlation and preserve the characteristics of the trend of historical DPV generation data.

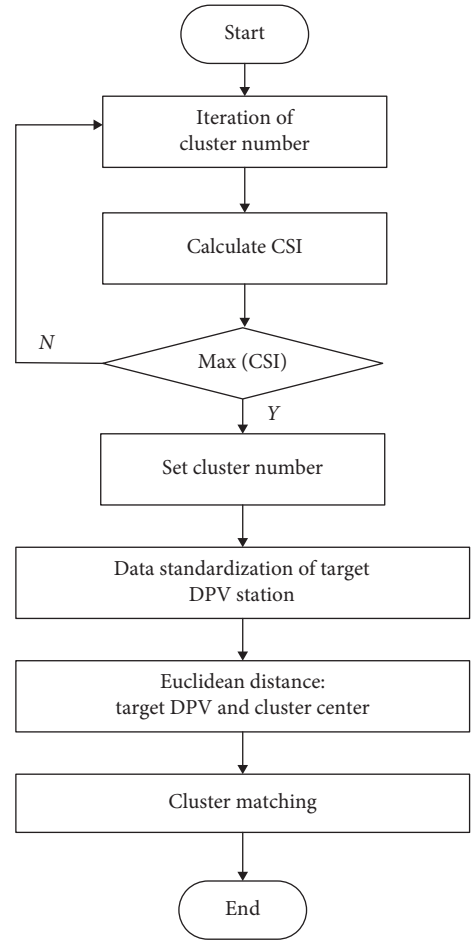


Figure 4: Flow chart of spatial correlation matching.

3.2. K-Means Clustering of Reference Plants. PV generation shows the characteristics of uncertainty and fluctuation, and the generation curve of a random day could not represent the general generation pattern of the plant. Therefore, the average of several generation days' data is used to establish the pattern library of reference plants.

In this paper, K-means algorithm is adopted for clustering, and several groups are formed of plants with similar generation pattern inside each group. The cluster number needs to be inputted when using K-means algorithm. The cluster significance index (CSI) is given in equation (2) to determine the group number; that is,

$$\text{CSI} = \frac{\sum_{j=1}^{N} \sum_{i=1}^{N} \left| X_{cj} - X_{ci} \right|}{\sum_{j=1}^{N} \sum_{i=1}^{n_j} \left| X_{ji} - X_{cj} \right|}, \tag{2}$$

where N is the number of clustering groups, n_j is the number of plants in the jth group, X_{cj} is the eigenvector of the jth group, and X_{ji} is the vector of the ith plant in the jth group.

The number of groups is determined through iteration. Different values of N are selected, and CSI is calculated for each N. The value N with the largest CSI is chosen as the input of the group number for K-means algorithm, and the PV generation patterns are obtained subsequently.

3.3. Online Matching of Spatial Correlation. The online pattern matching process is described as follows. Extract n monitoring points from recent historical data backward from the forecast point of the target DPV plant as a prediction window vector. Standardize the prediction window vector and add it to the pattern library as the $(T+1)$ cluster and perform a clustering process. If the current window vector could be put into the same cluster with the ith vector pattern, the target DPV plant is determined to have spatial correlation with the ith type of reference DPV plants.

Standard Euclidean Distance (SED) is more commonly adopted in actual application as the criterion of correlation. Therefore, in this paper, SED is used to quantify the correlation, and the optimal delay value Δt is determined by searching for the minimum SED. (3) gives the equation to calculate SED:

$$\rho(A,B) = \sqrt{\sum (a[i] - b[i])^2}, \quad (i = 1, 2, 3, \ldots, n), \tag{3}$$

where a $[i]$ and b $[i]$ are the ith element of vector A and vector B, respectively.

In theory, the probability of successful matching of spatial correlation is higher if the reference PV power plants are distributed more evenly and with larger number. For the target DPV plants that fail to match reference DPV plants, temporal correlation based forecast or other forecast methods are recommended.

3.4. Numerical Fitting Using LS-SVM. After spatial correlation matching, a single one or multiple reference DPV plants are chosen from the spatially correlated reference plant groups. The spatial correlation model is obtained by numerical fitting of the prediction window historical data of reference plants and target plant. Next, the short-term generation could be calculated with short-term forecast of reference DPV plants and the spatial correlation relationship.

least squares support vector machine (LS-SVM) regression is applied to perform numerical fitting, which could achieve better results of multivariate regression. The equation is shown as

$$P_{f2} = f(X) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K\left(X_i, X_j\right) + b, \tag{4}$$

where α_i and b are the coefficients to be determined, and K (X_i, X_j) is the kernel function. Radial basis function (RBF) is often used as the kernel function to solve a regression problem, which is given in

$$K\left(X_i, X_j\right) = \exp\left(\frac{-1}{2\sigma^2}\left\|X_i - X_j\right\|^2\right), \tag{5}$$

where σ is called the extension constant of RBF, which reflects the width of the function image. The smaller the width σ is, the more selective the function is.

3.5. Forecast Performance Evaluation. Although the prediction graph could show the results of all forecasting methods intuitively, it is arduous to quantitatively judge the pros and cons of each prediction method objectively. Therefore, this paper applies the root mean square error (RMSE) and mean absolute error (MAE) to compensate the shortcomings of the prediction graph. The two error formulas are as shown in equations (6) and (7):

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(P_p - P_r\right)^2}, \tag{6}$$

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|P_p - P_r\right|, \tag{7}$$

where P_p is the prediction value of PV power, P_r is the actual power, and m is the total number of prediction points.

3.6. Influence of Time Difference Characteristics on Spatial Correlation Matching. Considering the widespread distribution characteristics of DPV, the correlation relationship between the reference plant X and target plant Y may show some time and space difference characteristics; that is, Y (t) is more correlated to X $(t + \Delta t)$. This characteristic is referred to as time difference characteristics in the following paper. As shown in Figure 5, curves A and B have similar changing trend, but the starting and ending points are different. By moving the curve B to the right with a period of Δt, the distance between the curves is reduced, and the similarity of the trend is highlighted.

In references [10, 11], the Pearson product-moment correlation coefficient (PPMCC) is used to describe the correlation between vectors. The optimal value of Δt is determined by finding the value of PPMCC. Equation (7) gives the equation to calculate PPMCC:

$$r_P(X,Y) = \frac{\sum_{i=1}^{n}(x_i - x_{av})(y_i - y_{av})}{\sqrt{\sum_{i=1}^{n}(x_i - x_{av})^2 \sum_{i=1}^{n}(y_i - y_{av})^2}}, \tag{8}$$

where x_{av} and y_{av} represent the arithmetic mean of vector X and vector Y, respectively. PPMCC value close to 1 denotes strong correlation, while a value close to 0 denotes weak correlation.

Considering the fact that the DPV plants may be distributed in different time zones, a time shift method to improve pattern matching effects is given as follows. Set a unified reference time as 0 points, search from 0 points backward and forward with the time of Δt, and obtain the monitoring points within the range of $[-p, p]$.

For every iteration of spatial correlation matching, move the target plant vector backward or forward for one monitoring point and keep the other vectors in the pattern matrix unchanged. Calculate the correlation of the target plant and all other patterns and find the pattern with minimum SED, and the most spatially correlated DPV plant is found globally with consideration of time difference characteristics.

In summary, the advantages of considering Δt include the following:

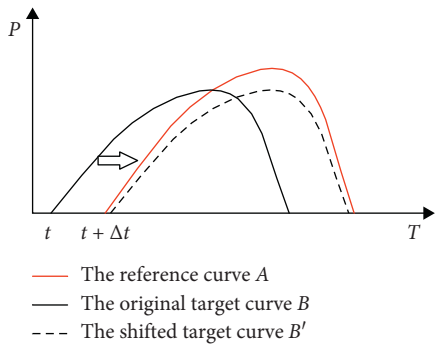


Figure 5: Time difference characteristics and time shift method for matching.

(1) Increasing the probability of successfully matching a target plant to the reference DPV power plants.

(2) Searching for the most correlated reference PV power plant globally; i.e., the global minimum is achieved rather than the local minimum, which could improve the forecast accuracy.

Therefore, the reference PV power plants matched with target PV plants in this paper are the most correlated plants globally considering the time difference characteristics.

4. Case Studies

The case used in this paper to demonstrate the forecast method is the actual historical generation data of 5166 DPV in the USA [29]. The DPV plants are located from 73° to 125°W, 25°–49°N, as shown in Figure 6. The monitored time is from 0 : 00 on 1 January 2006 to 23 : 45 on 31 December 2006, with a sampling interval of 15 minutes. The total number of sampling points is 30540. 1000 of the DPV plants (19.3%) are chosen randomly as reference DPV plants, and the remaining 4166 (81.7%) are regarded as target plants. The forecast target is the generation curve of day 307 in the year. The prediction window is set to be 3 days before the target forecast day, and the number of monitoring points is 288.

The preparation for online forecast is offline clustering. The 3616 reference plants are clustered into 50 spatial correlated groups; i.e., 50 patterns are generated. The calculated largest CSI is obtained to be 1.15485 when the N equals 50 based on equation (2). The clustering results are shown in Section 4.1.

Next, the online forecast process of master-slave short-term DPV generation forecast method is presented. In Section 4.2 and Section 4.3, two target plants T1 (3903#) and T2 (1346#) are chosen to show the forecast process. In Section 4.4, the forecasts of 1550 target plants are obtained, and the statistic error is compared. Section 4.5 discusses the situation in which multiple reference plants are used to make forecasts. Section 4.6 discusses the choice of prediction window size and its influence on forecast accuracy.

4.1. Clustering of Reference DPV Plants. 1000 DPV plants with forecast ability are chosen as reference plants to generate a pattern library. Using the clustering method in Section 3, a prediction window composed of the average of 10 days' generation data before the forecast target day is chosen as the pattern mining and clustering data, and the group number is set to be 50.

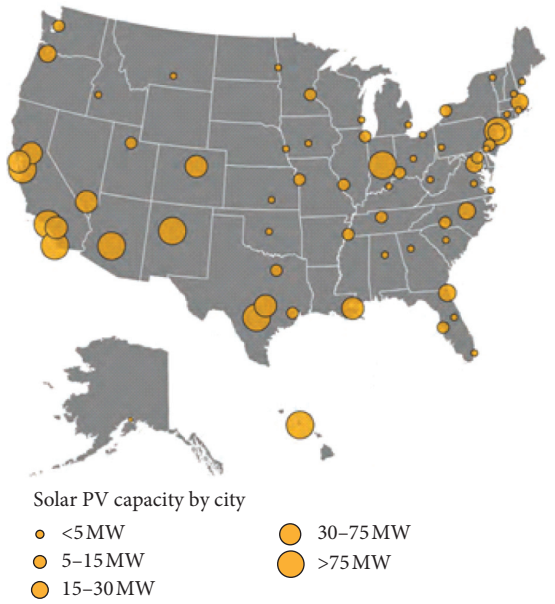


Figure 6: PV distribution in the USA [30].

Figure 7 shows the curves of 4 typical generation patterns in the pattern library. The plants in the same pattern group share similar generation characteristics, and the generation patterns between groups are extremely different. Therefore, the K-means clustering method could put the reference plants with similar generation patterns into the same groups and form a pattern library.

The choice of clustering group number should not only consider the CSI, which affects the clustering performance, but also the time consumption for matching target plants to reference plants. In short-term generation forecast, the forecast interval is 15 minutes. If the number of groups is too large, the matching process will be extremely time-consuming, and the forecast timeliness could not be guaranteed.

4.2. Searching for Most Correlated Plants considering Time Difference. This example shows the effect of the time shift

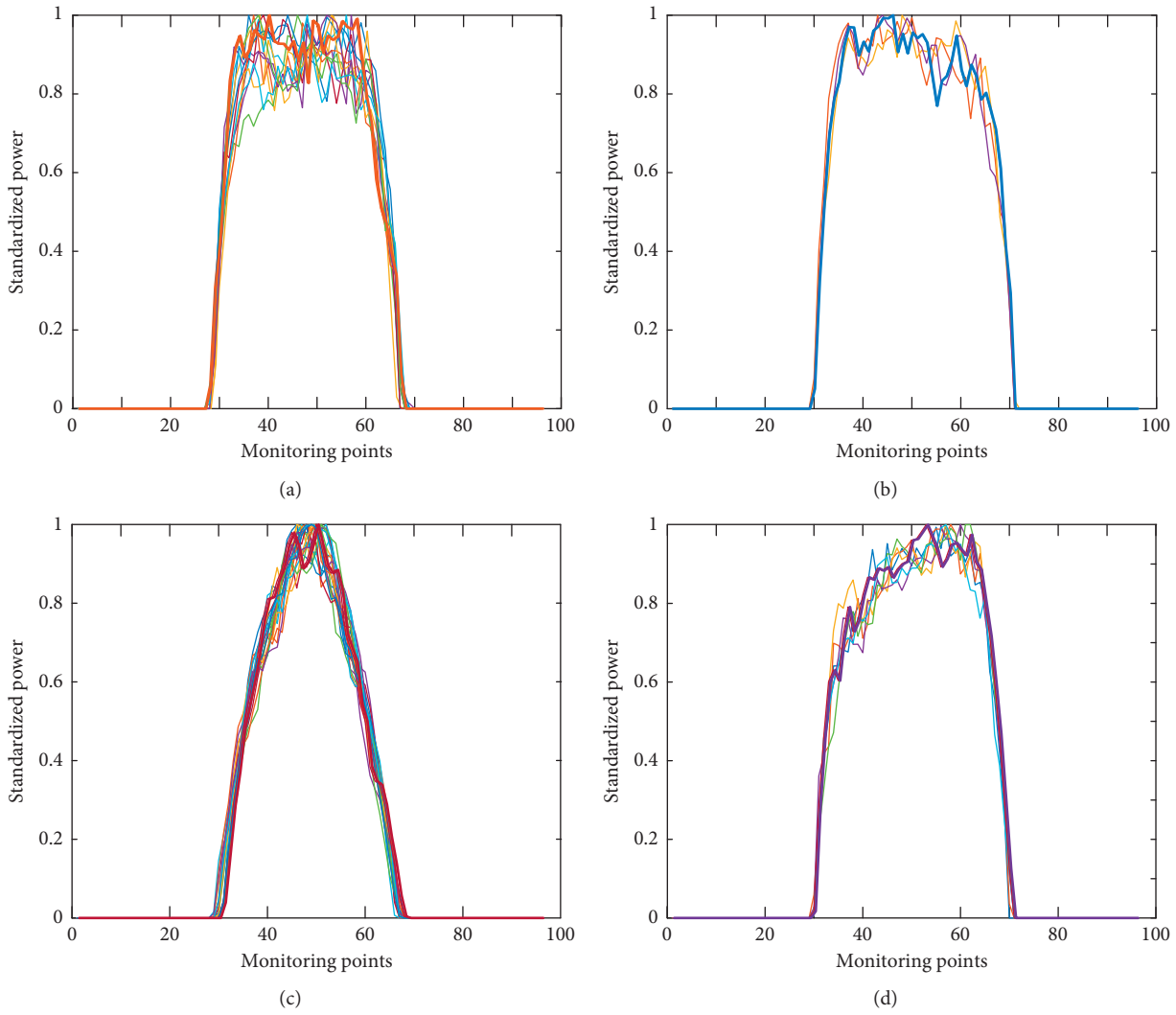


Figure 7: Examples of generation patterns obtained with K-means clustering. The numbers of plants in the pattern groups in subplots (a), b, (c), and d are, respectively, 24, 8, 35, and 7.

method given in Section 3. Following the proposed time shift method, the globally most spatially correlated reference plants to the target plant T1 are obtained. The time shift, minimum SED, PPMCC values, and chosen reference stations are listed in Table 1, and the search process is illustrated in Figure 8.

Several randomly chosen target plants are simulated, and the results show saddle-shaped curves similar to those in Figure 8, and the spatially correlated reference plants are usually located in time zones close to the target plants. There exists a minimum among the SED values achieved with different time shifts, and the most spatially correlated reference plant may not be synchronous with the target plants. Therefore, the most spatially correlated reference plants could be found globally with the time shift method, considering the time difference. In addition, the results in Table 1 show that the matched reference plants are different when different time shifts are applied, which means that the consideration of time difference could affect the matching results and further affect the forecast performance.

4.3. Spatial Correlation Matching considering Time Difference. Target plants T1 and T2 are added as two new patterns (patterns 51 and 52) into the pattern library. The clustering threshold is set as 1.40. K-means clustering is performed on the new library, and the results show that T1 is most strongly correlated with pattern 48. The reference plant R1 (785#), which has the highest correlation in that pattern group, is chosen as the master station, and the SED between the standard vectors of T1 and R1 is 1.1221. The spatial correlation results are shown in Figure 9.

In Figure 9, curves a and b are the real power of the reference PV plant and target PV plant, respectively, and c and d are the standard vectors of a and b, respectively. We compare the trajectory curves given that the nominal values of generation output of the two plants are quite different, which is the result of differences in installed capacity, converting efficiency, etc., but the overall changing trends are similar. Therefore, it is verified that the standardized trajectory curve could preserve the similarity of changing trend and could present the significant numerical correlation.

Table 1: Reference plant matching results for target plant T1 with different time shifts.

Time shift	Min.(SED)	|rp|	Ref_plant
−5	3.2552	0.97226	2057#
−4	2.3362	0.98571	2057#
−3	1.6388	0.99297	2057#
−2	1.2218	0.99609	2225#
−1	0.81687	0.99825	2239#
0	0.95176	0.99763	5141#
1	0.85250	0.9981	756#
2	0.88669	0.99794	2176#
3	0.84598	0.99813	4558#
4	1.3721	0.99507	452#
5	2.1725	0.98764	547#

However, the SED between T2's standard vector and the closed pattern's vector is 1.6794, which is higher than the clustering threshold. Thus, T2 will be regarded as a new pattern, and no match is found in the reference plant groups. The forecasting for unmatched DPV plants should adopt other forecast methods.

4.4. Univariate Prediction Based on Spatial Correlation. LS-SVM regression method is utilized to perform the numerical fitting of the prediction window generation data of R1 and T1, and the correlation relationship model is obtained. Considering that the actual generation in night time is 0, the following modification of the correlation relationship model is made to avoid human introduced error: if the reference plant generation is 0, the target plant generation should also be 0.

As the main purpose of the case study is to examine the forecast performance of the spatial correlation based method, the actual generation data of reference plants is utilized as the short-term forecast results to avoid the forecast errors of the reference plants. The short-term forecast generation is utilized as input of the correlation relationship model, and the entire day-ahead generation trajectory of target plant T1 with rolling calculation is obtained. Figure 10 shows the day-ahead forecast generation curve (green dotted line) and the actual generation curve (black line), with the comparison of forecast results using the temporal correlation method (blue broken line).

As shown in Figure 10, the predictive power of target PV plants with spatial correlation (green dotted line) is basically consistent with the predictive power of reference plants (red dotted line) and is closer to the real power of target plants (black line) compared with the predictive power of target PV considering timing correlation (blue broken line). It is obvious to know that the proposed spatial correlation method is effective and has high precision.

The forecast performance is evaluated with the forecast errors given in Section 3. The forecast errors are given in Table 2. It can be seen from Table 2 that both RMSE and MAE are smaller for the spatial correlation forecast method compared with the temporal correlation forecast method, which signifies that the proposed spatial correlation method achieves higher forecast accuracy.

4.5. Multivariate Prediction Based on Spatial Correlation. The spatial correlation matching is performed for 1550 target plants randomly chosen from all target plants, and 493 of the target plants fail to find a matching correlated pattern group, taking up 31.8% of all target plants. The short-term generation forecast for these plants should consider using temporal correlation forecast or other forecast methods. Among the rest 1057 target plants, which are matched to reference plant groups, 583 of them have 4 or more reference plants. Using the multivariate prediction function of LS-SVM, the generation forecasts for these 583 plants are obtained. The forecast statistic mean errors are shown in Table 3.

The longitudinal comparison of Table 3 shows that the more reference plants are matched, the more reference information is given, the less the forecast error is. Therefore, when there is more than one match of reference plants, the result of multivariate prediction is better than that of univariate prediction.

The horizontal comparison of Table 3 shows that the forecast based on spatial correlation is more accurate than the forecast based on temporal correlation. The reason is that the temporal forecast method only utilizes the historical generation data, and no information of future change is involved. The spatial correlated forecast method, on the other hand, uses the numeric weather forecast data (in the generation forecast of reference plants) and historical generation data, hence achieving higher forecast accuracy.

4.6. Influence of Prediction Window Size on Spatial Correlation Forecast. This part discusses the choice of prediction window size and its influence on forecast performance. Considering the limited data storage capability of target plants, we assume that only ten days of historical generation data is available. Use the first nine days' data to generate prediction window data and make a forecast, and the tenth day's data to examine the forecast performance. Figure 11 shows the forecast error of a randomly chosen target plant (plant 1000#) with different prediction window sizes, from 1 day to 9 days. It can be seen that, for example, the forecast error of spatial correlation method is larger than that of temporal correlation using MAE as a criterion if the prediction window size is 3 days or 4 days. The

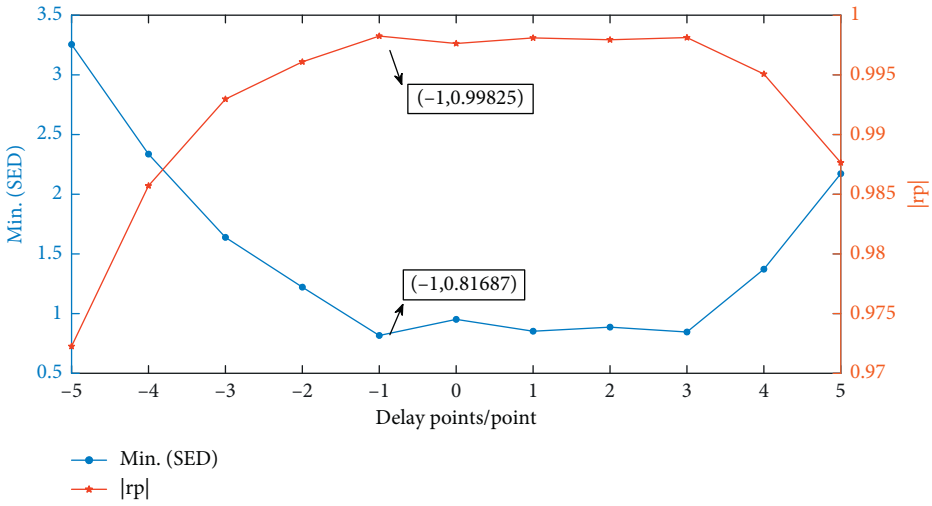


Figure 8: Min. (SED) and |rp| values with different time shifts.

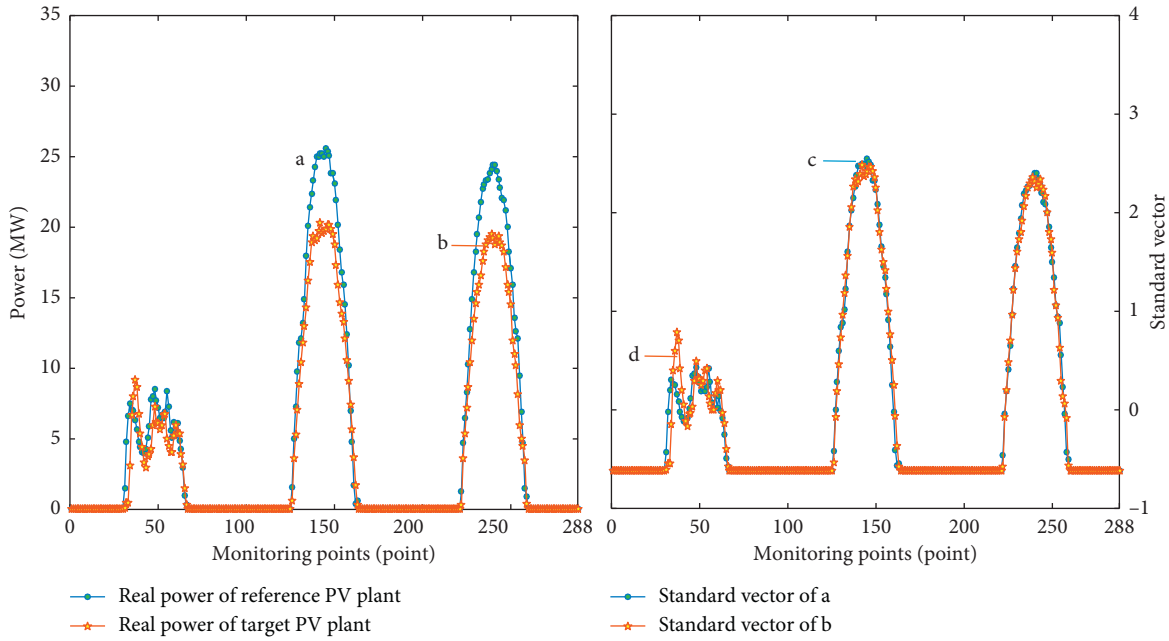


Figure 9: The spatial correlation of DPV plant T1 and R1.

forecast results of spatial correlation method with other prediction window sizes are better than those of the temporal correlation method. The optimal prediction window size is 2 days.

Next, 641 target plants are randomly chosen, and the optimal prediction window sizes are counted. As shown in Figure 12, it is noted that the optimal prediction window size is different for each plant, which is influenced by the characteristics of the plant and the surrounding environment. The majority of the plants could achieve good forecast performance with a prediction window of 3–7 days. Therefore, in practical application, the forecast scheme should be customized for each target plant according to its historical data, the prediction window size should be appropriately selected, and the value of prediction window size should be updated as time goes by. To reproduce the cases, there are four limitations including the data source, the number of reference/target DPV plants, the prediction window size, and offline clustering threshold value setting.

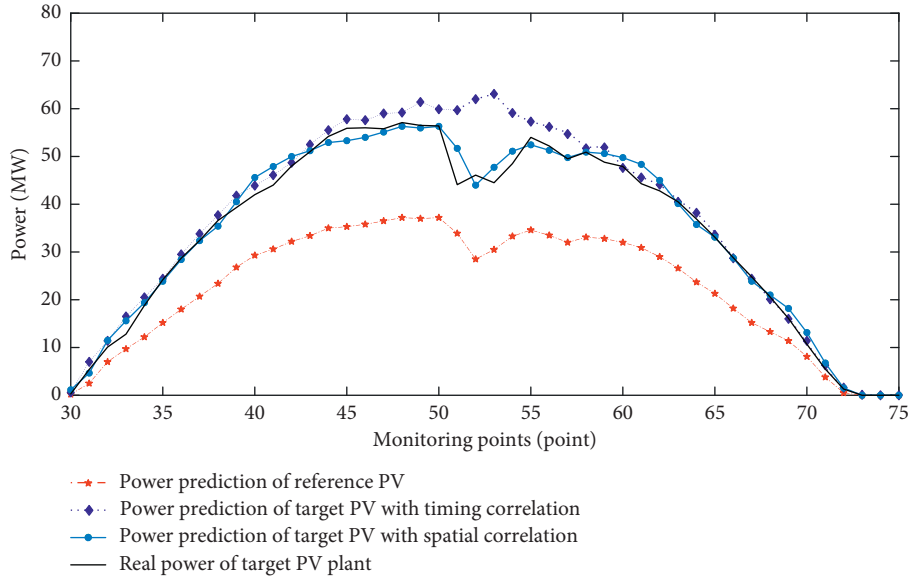


Figure 10: Univariate forecast based on spatial correlation and comparison with actual generation and forecast results using temporal correlation.

Table 2: Forecast errors of spatial and temporal correlation method (/MW).

Forecast method	RMSE	MAE
Temporal correlation forecast	3.4252	1.2927
Spatial correlation forecast	1.3898	0.6691

Table 3: Forecast mean errors of 583 plants using spatial and temporal correlation methods (/MW).

Number (referenced PV plants)	Temporal correlation forecast		Spatial correlation forecast	
	Avg. (RMSE)	Avg. (MAE)	Avg. (RMSE)	Avg. (MAE)
1	2.8990	0.9990	1.8478	0.7819
2	2.4093	0.9538	1.3656	0.7353
3	2.1938	0.9279	1.1381	0.6480
4	2.0283	0.9105	1.0271	0.6161

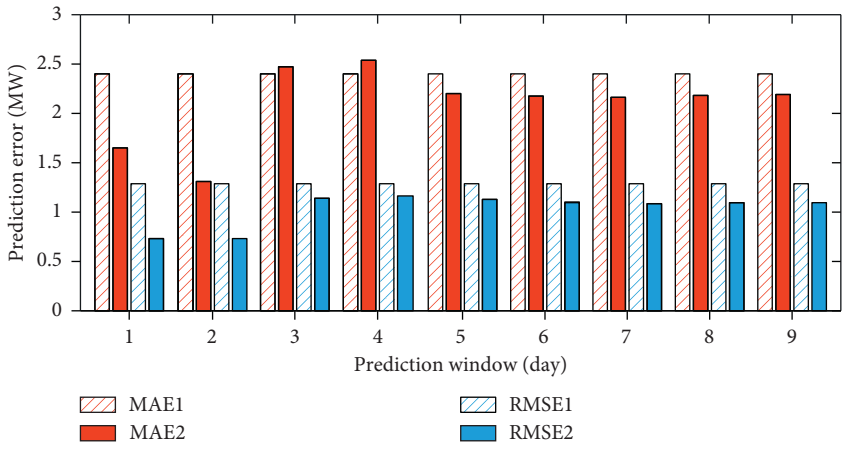


Figure 11: Forecast error with different prediction window sizes. The red and blue shady bars are the prediction errors of temporal forecast method, and the red and blue solid bars are the prediction errors of spatial forecast method.

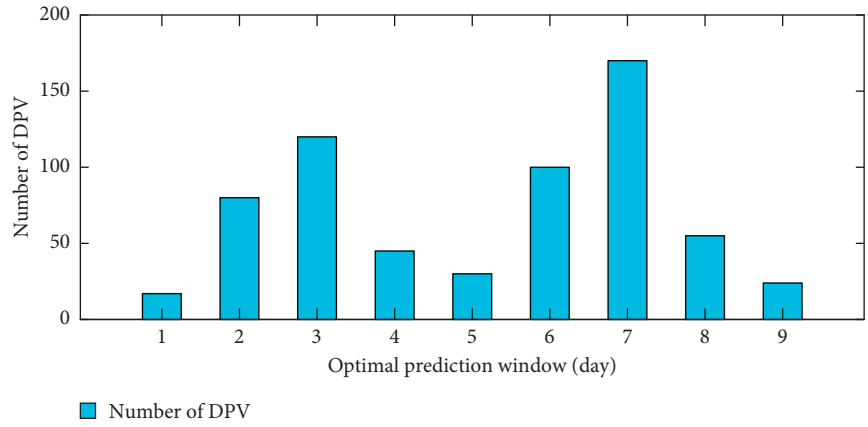


Figure 12: Statistical results of different optimal prediction window sizes.

5. Conclusions

Aiming to solve the technique bottleneck of small capacity DPV generation forecast caused by data deficiency and complex influencing factors, this paper proposes an indirect forecast method based on spatial correlation, using a master-slave structure and mapping the target plants incapable of making a forecast on their own to the reference plants, which could make the forecast with sophisticated method. The following conclusions are drawn:

(1) The historical generation data contain the complete background information such as meteorological data, so that the spatial correlated forecast method for DPV generation could make full use of historical data and achieve accurate short-term forecast.

(2) Adopting LS-SVM regression for numerical fitting of the spatial correlation relationship could improve the overall forecast accuracy, compared to prediction methods based on temporal correlation and least squares linear regression.

(3) The proposed spatial correlation forecast method could use the DPV plants that are already equipped with forecast systems and obtain short-term generation forecast for DPV or newly built DPV plants with low investment.

Data Availability

The data used to support the findings of this work 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 jointly supported by University Natural Science Research General Project of Jiangsu Province (No. 19KJB470004), the High-level Talent Introduction Scientific Research Foundation of Nanjing Institute of Technology (No. YKJ201820), and Open Research Fund of Jiangsu Collaborative Innovation Centre for Smart Distribution Network, Nanjing Institute of Technology (No. XTCX201906).

References

[1] A. Sangwongwanich, Y. Yang, F. Blaabjerg, and H. Wang, "Benchmarking of constant power generation strategies for single-phase grid-connected photovoltaic systems," *Institute of Electrical and Electronics Engineers Transactions on Industry Applications*, vol. 54, no. 1, pp. 447–457, 2018.

[2] F. Ding and B. Mather, "On distributed pv hosting capacity estimation, sensitivity study, and improvement," *Institute of Electrical and Electronics Engineers Transactions on Sustainable Energy*, vol. 8, no. 3, pp. 1010–1020, 2017.

[3] U. K. Das, K. S. Tey, M. Seyedmahmoudian et al., "Forecasting of photovoltaic power generation and model optimization: a review," *Renewable & Sustainable Energy Reviews*, vol. 81, pp. 912–928, 2017.

[4] N. Haghdadi, A. Bruce, I. MaCgill, and R. Passey, "Impact of distributed photovoltaic systems on zone substation peak demand," *Institute of Electrical and Electronics Engineers Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 621–629, 2018.

[5] B. Jing, Z. Qian, Y. Pei, and J. Wang, "Ultra short-term PV power forecasting based on ELM segmentation model," *The Journal of Engineering*, vol. 2017, no. 13, pp. 2564–2568, 2017.

[6] F. Liu, R. Li, Y. Li, R. Yan, and T. Saha, "Takagi-Sugeno fuzzy model-based approach considering multiple weather factors for the photovoltaic power short-term forecasting," *IET Renewable Power Generation*, vol. 11, no. 10, pp. 1281–1287, 2017.

[7] L. Gigoni, A. Betti, E. Crisostomi et al., "Day-ahead hourly forecasting of power generation from photovoltaic plants," *Institute of Electrical and Electronics Engineers Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 831–842, 2018.

[8] A. Bracale, G. Carpinelli, and P. De Falco, "A probabilistic competitive ensemble method for short-term photovoltaic power forecasting," *Institute of Electrical and Electronics Engineers Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 551–560, 2017.

[9] C. Wan, J. Lin, Y. Song, Z. Xu, and G. Yang, "Probabilistic forecasting of photovoltaic generation: an efficient statistical

approach," *Institute of Electrical and Electronics Engineers Transactions on Power Systems*, vol. 32, no. 3, pp. 2471-2472, 2017.

[10] A. Tascikaraoglu, B. M. Sanandaji, G. Chicco et al., "Compressive spatio-temporal forecasting of meteorological quantities and photovoltaic power," *Institute of Electrical and Electronics Engineers Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1295–1305, 2016.

[11] L. A. Fernandez-Jimenez, A. Muñoz-Jimenez, A. Falces et al., "Short-term power forecasting system for photovoltaic plants," *Renewable Energy*, vol. 44, no. 4, pp. 311–317, 2012.

[12] F. Wang, Z. Zhen, Z. Mi, H. Sun, S. Su, and G. Yang, "Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting," *Energy and Buildings*, vol. 86, pp. 427–438, 2015.

[13] C. Monteiro, T. Santos, L. Fernandez-Jimenez, I. Ramirez-Rosado, and M. Terreros-Olarte, "Short-term power forecasting model for photovoltaic plants based on historical similarity," *Energies*, vol. 6, no. 5, pp. 2624–2643, 2013.

[14] H. M. El-Helw, A. Magdy, and M. I. Marei, "A hybrid maximum power point tracking technique for partially shaded photovoltaic arrays," *Institute of Electrical and Electronics Engineers Access*, vol. 5, pp. 11900–11908, 2017.

[15] J. Liu, W. Fang, X. Zhang, and C. Yang, "An improved photovoltaic power forecasting model with the assistance of aerosol index data," *Institute of Electrical and Electronics Engineers Transactions on Sustainable Energy*, vol. 6, no. 2, pp. 434–442, 2015.

[16] K. Y. Bae, H. S. Jang, and D. K. Sung, "Hourly solar irradiance prediction based on support vector machine and its error analysis," *Institute of Electrical and Electronics Engineers Transactions on Power Systems*, vol. 32, no. 2, pp. 935–945, 2017.

[17] T. V. Da Silva, R. V. A. Monteiro, G. C. Guimaraes, F. A. M. Moura, M. A. Tamashiro, and M. R. M. C. Albertini, "Performance analysis of neural network training algorithms and support vector machine for power generation forecast of photovoltaic panel," *Institute of Electrical and Electronics Engineers Latin America Transactions*, vol. 15, no. 6, pp. 1091–1100, 2017.

[18] M. Bouzerdoum, A. Mellit, and A. Massi Pavan, "A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant," *Solar Energy*, vol. 98, pp. 226–235, 2013.

[19] A. Mellit, A. Massi Pavan, and V. Lughi, "Short-term forecasting of power production in a large-scale photovoltaic plant," *Solar Energy*, vol. 105, pp. 401–413, 2014.

[20] W. VanDeventer, E. Jamei, G. S. Thirunavukkarasu et al., "Short-term PV power forecasting using hybrid GASVM technique," *Renewable Energy*, vol. 140, pp. 367–379, 2019.

[21] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques," *IET Renewable Power Generation*, vol. 13, no. 7, pp. 1009–1023, 2019.

[22] M. Victoria, R. Herrero, C. Domínguez, I. Antón, S. Askins, and G. Sala, "Characterization of the spatial distribution of irradiance and spectrum in concentrating photovoltaic systems and their effect on multi-junction solar cells," *Progress in Photovoltaics: Research and Applications*, vol. 21, no. 3, pp. 308–318, 2013.

[23] B. Elsinga and W. Van Sark, "Spatial power fluctuation correlations in urban rooftop photovoltaic systems," *Progress in Photovoltaics: Research and Applications*, vol. 23, no. 10, pp. 1390–1397, 2015.

[24] E. Karakaya, "Finite Element Method for forecasting the diffusion of photovoltaic systems: why and how?," *Applied Energy*, vol. 163, pp. 464–475, 2016.

[25] T. Zhao, Z. Zhou, Y. Zhang, P. Ling, and Y. Tian, "Spatio-temporal analysis and forecasting of distributed PV systems diffusion: a case study of shanghai using a data-driven approach," *Institute of Electrical and Electronics Engineers Access*, vol. 5, no. 99, pp. 5135–5148, 2017.

[26] Y. Liu, Z. Li, K. Bai, Z. Zhang, X. Lu, and X. Zhang, "Short-term power-forecasting method of distributed PV power system for consideration of its effects on load forecasting," *The Journal of Engineering*, vol. 2017, no. 13, pp. 865–869, 2017.

[27] M. Lave and J. Kleissl, "Cloud speed impact on solar variability scaling - application to the wavelet variability model," *Solar Energy*, vol. 91, no. 3, pp. 11–21, 2013.

[28] R. Li, W. Wang, and M. Xia, "Cooperative planning of active distribution system with renewable energy sources and energy storage systems," *Institute of Electrical and Electronics Engineers Access*, vol. 6, pp. 5916–5926, 2017.

[29] National Renewable Energy Laboratory, *Solar Power Data for Integration Studies*, National Renewable Energy Laboratory, Golden, CO, USA, 2006, https://www.nrel.gov/grid/solar-power-data.html.

[30] A. Bradford, G. Weissman, R. Sargent, and B. Fanshaw, *Shining Cities 2017*, https://environmentamerica.org/sites/environment/files/cpn/AMN-033117-REPORT/shining-cities-2017.html, Environment America Research & Policy Center, Denver, CO, USA, 2017, https://environmentamerica.org/sites/environment/files/cpn/AMN-033117-REPORT/shining-cities-2017.html.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6610003, 10 pages
https://doi.org/10.1155/2021/6610003

Research Article

Experimental Research on Bearing Characteristics of the Asphalt Pavement Containing Buried Pipeline

Hailiang Xu, Jining Qin, Hehuan Ren, Jindou Sun, and Lian He

Department of Civil Engineering, North China University of Technology, Beijing, China

Correspondence should be addressed to Jining Qin; qin_1221@qq.com

Received 3 December 2020; Revised 20 February 2021; Accepted 6 March 2021; Published 22 March 2021

Academic Editor: William Guo

The bearing characteristics of the asphalt pavement structure are greatly influenced by buried pipelines. Improper treatment of buried pipelines would cause early damage to pavement structure. By the digital speckle correlation method (DSCM), the experimental research on bearing characteristics of the asphalt pavement containing buried pipe was carried out. The mechanical characteristics of the asphalt pavement structure are studied under four different pipeline burial conditions. The vertical displacement and strain values of each layer of the asphalt pavement structure are obtained under four operating conditions. The results showed that (1) the digital speckle observation test method can accurately obtain the displacement and strain values of each layer of asphalt pavement structure containing buried pipeline, and the application effect is good. Compared with the traditional contact strain measurement method, this method is simple and accurate and can provide effective analysis data for experimental research. (2) There exists an interlayer effect of the asphalt pavement structure. The vertical displacement value and the strain value are discontinuities and can suddenly change between two adjacent layers. At the same time, the vertical strain and the shear strain concentration phenomenon appear at the bottom of each layer, especially at the bottom of the upper layer and the subbase layer of asphalt. (3) Affected by the buried pipelines, the vertical displacement value of the asphalt pavement structure reduces, and the tensile and shear strain values of asphalt pavement structure increase. The subbase layer of asphalt is most affected by the buried pipelines, which accelerated the destruction of the asphalt pavement structure.

1. Introduction

Urban road is not only the carrier of transportation but also the main channel of municipal pipelines. With the existence of buried pipelines under roads, the stress state of road pavement structure will inevitably be affected (especially by shallow pipelines) [1]. A large number of existing projects show that the improper installation and disposal of buried pipelines often lead to the failure of pavement structure to reach the design life. Take the 2013 field survey of road surface damage in Tiancun Mountain road in Beijing as an example. Ninety-four cracks were found in the vertical and horizontal directions within a range of three kilometers, among which 51 occurred at or near the location of buried municipal pipelines, accounting for about 55% of the total cracks. The cracks of the asphalt pavement structure at the buried pipeline are shown in Figure 1.

At present, there is no clear control index for the buried pipeline under pavement structure from the point of view of pavement structure. The adverse effects of buried pipelines on pavement structure are not considered in the design process of pavement structure. The existing research mainly focuses on the stress state of pipelines under vehicle load [1–4].

In the experimental study of pavement structure stress, due to the limitation of measuring and testing means, sensors such as the pressure box, asphalt strain gauge, and axis gauge are mainly used [5, 6]. Although these testing methods can reflect the overall stress condition of pavement structure, there are still some unfavorable factors such as the limited number of locations, the inconvenience of testing the original parts, and the great influence of the embedded quality on the testing accuracy. In view of the above problems in the experimental study, this study uses a new

digital speckle correlation method to study the bearing characteristics of asphalt pavement structure under the coupling action of driving load and buried pipeline.

2. Basic Principle and Introduction of the Digital Speckle Correlation Method

The digital speckle correlation method (DSCM) was proposed in the early 1980s by a Japanese scientist Yamaguchi [7] and Peter and American scientists Rnaosn et al. [8]. They proposed their experimental method, respectively, mainly using optical observation and postimage processing. The method is based on gray scale analysis of object surface speckle to obtain displacement and the strain information measurement method. Advantages of this method include global observation, adjustable range accuracy, noncontact mode, simple operation, and reliable data.

The principle of DSCM is that the predeformed image is called "reference image" and the postdeformed image is called "target image." After loading, the speckle gray information of the reference image and the target image will be slightly different, the essence of which is caused by displacement and strain. DSCM carries out a correlation search on the speckle pattern before and after deformation, analyzes the gray distribution difference of each area, and finally realizes the measurement of displacement and strain.

The measurement process of DSCM is simple. Compared with the contact strain measurement method, the equipment does not need to contact with the specimen during the measurement, which omits the installation of the sensor and eliminates the errors caused in the installation process. The contact measurement can only reflect the strain information of the sensor position, while DSCM can obtain the full-field information of the speckle field photographed under the lens and can measure the continuous change process of the speckle field according to the continuous shooting of the camera.

In graphics processing, gray scale is the carrier of graphics data (gray scale is the speckle field formed by white spots on a black background or white spots on a white background). DSCM establishes the correlation function between the predeformation image gray level and postdeformation image gray level and converts graphic data into digital data by using computational software.

The density and diameter of the gray scale directly affect the accuracy of software numerical calculation. In the gray image taken by camera, the gray search algorithm is also very important. The classical gray search methods include the fineness search method, cross-search method, climbing search method, neighborhood search method, and Newton Lafayette partial differential correction method. The new search method includes the FFT search method in frequency domain. The search method used in the experiment is the cross-search method. The cross-search method reduces the search time and improves the search efficiency because it simplifies the two-dimensional search method to one-dimension.

The coordinate matching diagram of the calculation process of the digital speckle correlation method is shown in Figure 2. The reference subarea centered on P is calculated to realize the search. The region can be divided into $(2n+1)\times(2n+1)$ grids. By referring to the measurement point P in the image, the point P^* with the greatest similarity to P can be found by searching the target subarea in the target image through the limited threshold value.

Figure 1: Crack of asphalt pavement structure.

According to the continuity assumption of displacement field, the relationship between horizontal displacement and vertical displacement of any point Q in the image subarea can be expressed by equations (1) and (2), and the strain field relationship can be expressed by equation (3) [9]. The displacement field of the whole speckle image can be obtained by calculating multiple image subareas, and then, the strain field can be calculated. $(\Delta x, \Delta y)$ is the distance between point O and the center of the subarea. L is the length before deformation, and ΔL is its variation after deformation.

$$u = u_0 + \frac{\partial u}{\partial y} \cdot \triangle x + \frac{\partial u}{\partial y} \triangle y, \tag{1}$$

$$v = v_0 + \frac{\partial v}{\partial x} \cdot \triangle x + \frac{\partial v}{\partial y} \cdot \triangle y, \tag{2}$$

$$\varepsilon = \lim_{L \longrightarrow 0} \left(\frac{\Delta L}{L} \right). \tag{3}$$

The general standardized correlation function common to DSCM is given in the following equation [10–12].

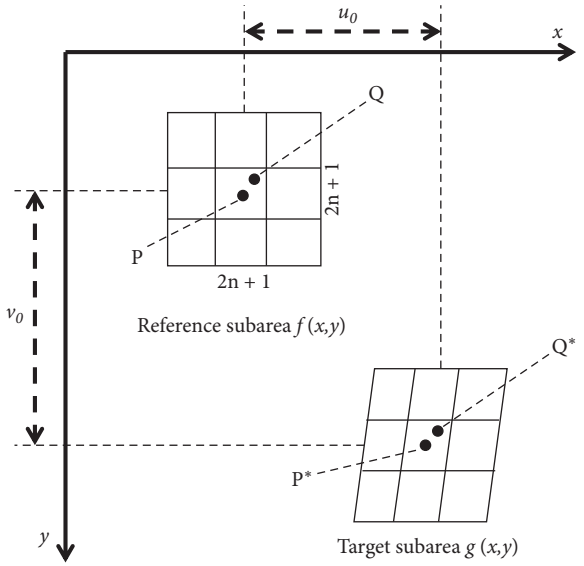


Figure 2: Coordinate matching diagram.

$$C(u,v) = \frac{\sum_{x=-2n+1}^{2n+1}\sum_{y=-2n+1}^{2n+1}[f(x,y)g(x+u,y+v)]}{\sqrt{\sum_{x=-2n+1}^{2n+1}\sum_{y=-2n+1}^{2n+1}f^2(x,y)}\sqrt{\sum_{x=-2n+1}^{2n+1}\sum_{y=-2n+1}^{2n+1}g^2(x+u,y+v)}}. \quad (4)$$

In the formula, $f(x,y)$ is the gray function of the reference image before deformation at a certain point (x,y), $g(x+u,y+v)$ is the gray function of the target image after deformation at a point $(x+u,y+v)$, and the range of the correlation coefficient $C(u,v)$ is [0, 1]. The larger the correlation coefficient is, the better the matching degree is.

In this study, the method of the digital speckle correlation is used to obtain the displacement and strain data of pavement structure directly through noncontact optical measurement. The changes of pavement structure can be observed and analyzed intuitively.

3. Manufacture of Asphalt Road Structural Specimens Containing Buried Pipeline

The preparation of structural specimens includes the asphalt layer (upper layer, middle layer, and lower layer), cement-stabilized macadam layer (upper base layer and lower base layer), and pipelines. The settings of structural parameters are given in Table 1.

For the preparation of asphalt structural layer materials, commercial asphalt materials AC-13, AC-20, and AC-25 were poured into the asphalt mold, respectively, and compacted. The cement-stabilized macadam layer is prepared at a 5% cement-stabilized macadam mixture ratio and is proportioned at the mass ratio of cement: coarse aggregate: fine aggregate: sand: water = 113 : 906 : 566 : 793 : 112. The combination of each layer of the structure mainly considers the flatness and tightness of the contact surface. Equipped with cement, the cement mortar with sand of 1 : 3 is used to smooth and connect the top surface of water-stable base course; the connection of the asphalt structural layer is first wetted with waterborne asphalt permeable oil at the bottom layer and then evenly coated with oily 70 # asphalt viscous oil, which combines the two layers before static pressing for 12 hours.

The PVC pipes are buried in the center of the road structure when the water-stabilized layer is poured. The center of the pipes is 0.22 m away from the surface of the specimens (shallow-buried pipes, where the upper edge of the pipes is at the bottom of the asphalt layer), 0.36 m (middle-buried pipes), 0.50 m (deep-buried pipes), and no pipes. The preparation of specimens meets the requirements of relevant specifications and mechanical indexes.

After the specimens are prepared, the artificial speckle field is sprayed on the observation surface, and the artificial speckle field is sprayed on the central area of the specimens at 20 cm × 56 cm. The black paint is used for priming. After the black paint is dried, white spots are evenly sprayed with white paint. Taking the middle-buried pipeline as an example, the structure of the specimen is shown in Figure 3.

The test system consists of WAW-600 electrohydraulic servo universal testing machine, CCD camera, and corresponding control acquisition system. Figures 4 and 5 represent the schematic diagram of the test system and the layout of the indoor model test, respectively. The experimental load is used in the "urban road design code" (CJJ 169–2012), a grounding pressure of 0.70 MPa.

4. Analysis of the influence of buried pipeline on pavement structure

Through the digital speckle correlation method, the influence of buried pipelines on the road structure is calculated

Table 1: Structure-layer parameters.

Horizon	Material	Length (m)	Width (m)	Thickness (m)
Asphalt upper layer	AC-13	0.8	0.3	0.04
Asphalt middle layer	AC-20	0.8	0.3	0.05
Asphalt lower layer	AC-25	0.8	0.3	0.07
Upper water-stabilization bases	Cement-stabilized macadam	0.8	0.3	0.2
Lower water-stabilization bases	Cement-stabilized macadam	0.8	0.3	0.2
Pipeline	PVC		External diameter, 110 mm Wall thickness, 10 mm	

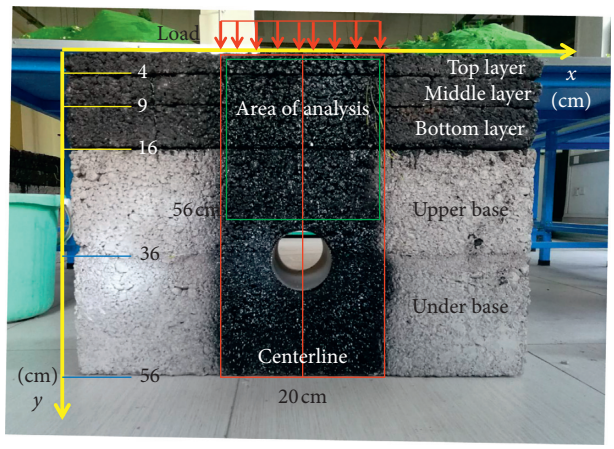


Figure 3: Middle-buried pipeline structure.

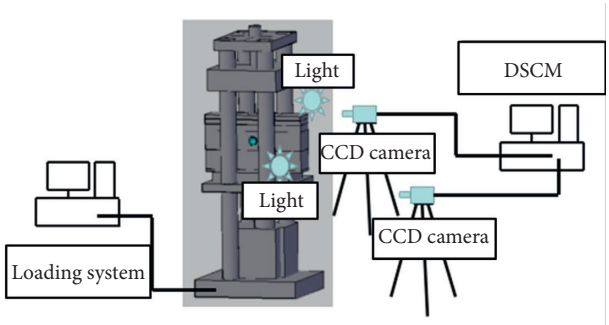


Figure 4: Schematic diagram of the experimental system.

under different buried pipelines. Aiming at the form of structural failure, the vertical displacement (deflection value), tension-compression strain, and shear strain of the road structural layer are selected as the main analysis indexes.

4.1. Analysis of the Effect of Buried Pipeline on the Vertical Displacement of the Structural Layer. The vertical displacement (deflection value) of road structure is an important reference value in road structure design. The vertical displacement nephogram of the structural analysis area under different buried depths of the pipeline obtained during the test is shown in Figures 6–9.

The pixel data of displacement cloud image are multiplied by a certain ratio (the ratio of the difference of pixel

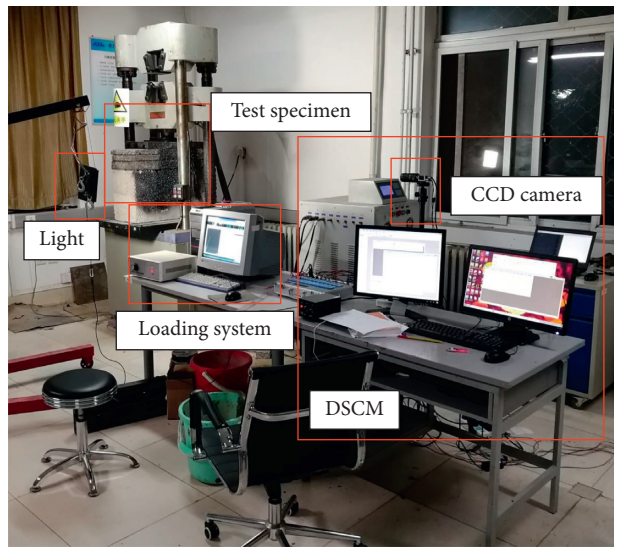


Figure 5: Laboratory model test.

points and the actual distance), before being converted it into the actual displacement unit millimeter. The vertical displacement values on the central line are selected for specific analysis, as shown in Figure 10.

From Figure 10, we can see that

(1) Pipeline embedment reduces the vertical displacement of each layer. With the decrease of pipeline embedment depth, the trend of reducing the displacement of each layer of pavement becomes more obvious. Taking the deflection value of the asphalt upper layer as an example, the deflection value is 26.09 (unit: less than 0.01 mm) under the condition of no pipeline embedding and 10.65 (0.01 mm) under the condition of shallow pipeline embedding, with a decrease of 59.2%. This is caused by the support of the pipe.

(2) As the road material is layered, the strength of each layer is different. In the road structure, the vertical displacement values are discontinuous, and the displacement values at the connections between layers of the structure have abrupt changes. This is shown in Table 2. This phenomenon is different from the continuity condition of displacement assumed in most current calculation models. Taking pipeline less asphalt as an example, the abrupt change values of the upper, middle, and lower surface layers of asphalt

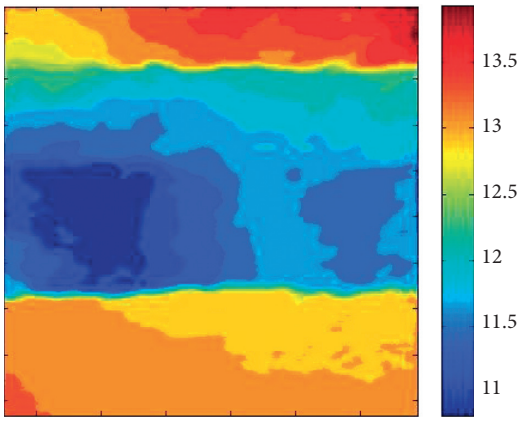


Figure 6: Vertical displacement of no buried pipe.

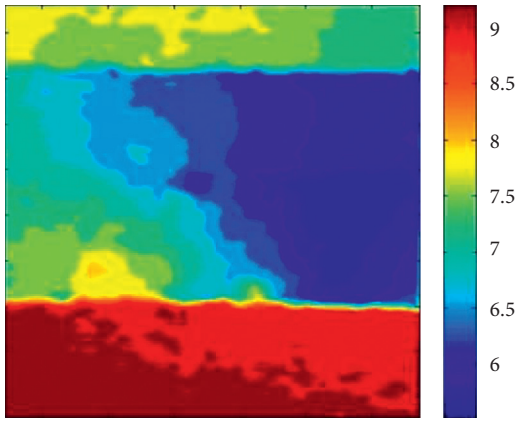


Figure 7: Vertical displacement of deep-buried pipe.

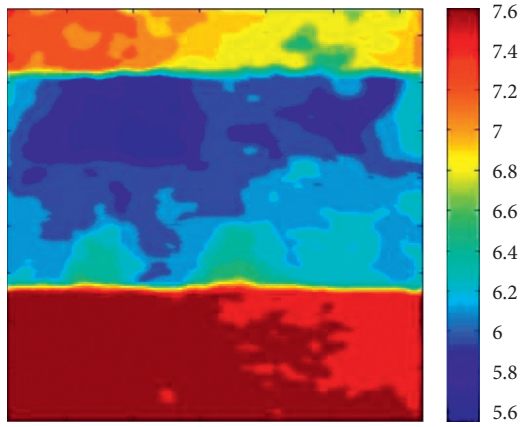


Figure 8: Vertical displacement of middle-buried pipe.

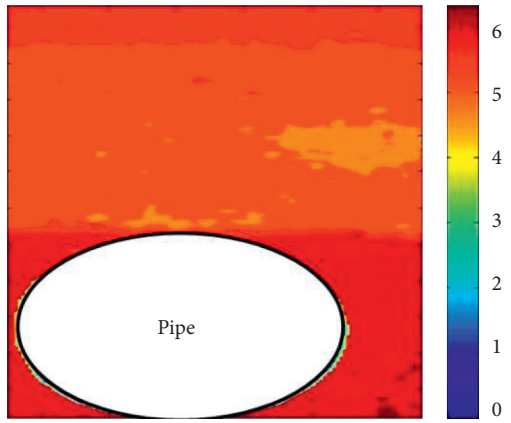


Figure 9: Vertical displacement of shallow-buried pipe.

are 1.98 (unit: 0.01 mm), 0.3 (unit: 0.01 mm), and 2.95 (unit: 0.01 mm), respectively. The abrupt change ranges are 7.58%, 0.13%, and 12.74%. The abrupt change between layers in other working conditions is shown in Table 3.

(3) Among the abrupt changes of interlayer displacement, the abrupt changes of the displacement of the upper layer and the bottom layer of the lower layer of asphalt are the most obvious. However, the two trends are opposite. The displacement between the upper layer of asphalt and the middle layer of asphalt decreases, while the displacement between the lower layer of asphalt and the water-stabilized macadam increases.

The buried pipeline supports the road and reduces the vertical displacement of each layer. The road material is layered, and the strength of each layer material is different, which makes the displacement value at the junction between layers abrupt. The abrupt displacement of the upper layer of asphalt and the bottom layer of asphalt is the most obvious.

4.2. Analysis of the Influence of Buried Pipeline on Tension and Compression Strain of Pavement Structure. The strain of tension and compression stress in each layer of pavement structure is an important research content of pavement structure stress. During the test, the tension and compression strain nephogram of the pavement structure analysis area under different buried depths is shown in Figures 11–14 .

The vertical strain values observed in Figures 11–14 are sorted out, and the strains at different buried depths of pipelines at the center line are selected for analysis, as shown in Figure 15.

From Figures 11–14, it can be seen intuitively that the pavement structure is mainly compressed, but the bottom of the asphalt layer appears at tension state. At the same time, it is shown in Figure 15 that

(1) The phenomenon of strain concentration appears at the bottom of each layer of pavement structure, that is, the maximum strain of each layer appears at the bottom of this layer. Among them, the strain concentration of the upper layer and the lower layer of asphalt is the most obvious. The maximum vertical strain and variation amplitude of pipeline under different burial depths are shown in Table 3. It can be seen from Table 3 that the influence of pipeline depth on the stress state of pavement structure cannot be ignored.

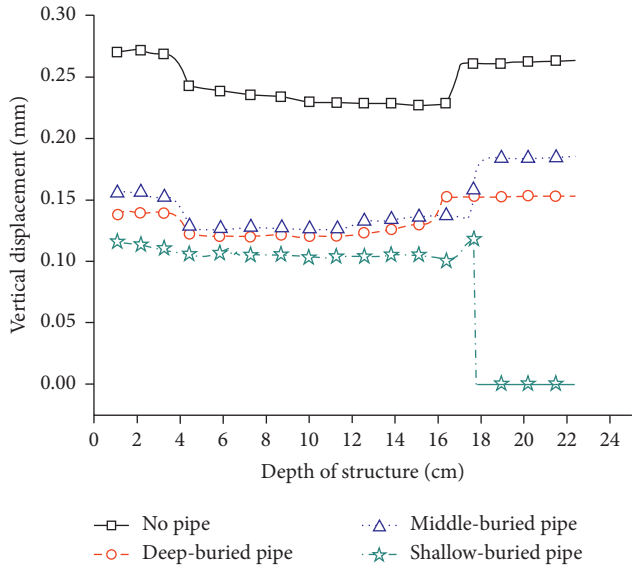


Figure 10: Vertical displacement curve of the center line.

Table 2: Vertical displacement of structure interlayer.

Condition(mm)	Bottom of the upper layer/Top of the middle layer		Bottom of the middle layer/Top of the lower layer		Bottom of the lower layer/ Top of water-stabilization bases	
No pipeline	0.2609	0.2411	0.2293	0.2290	0.2313	0.2608
Rangeability	−7.58%		−0.13%		12.74%	
Deep-buried pipe	0.1928	0.1809	0.1818	0.1796	0.1938	0.2127
Rangeability	−6.14%		−1.16%		9.75%	
Middle-buried pipe	0.1448	0.1253	0.1161	0.1053	0.1400	0.1833
Rangeability	−13.46%		−9.27%		30.91%	
Shallow-buried pipe	0.1065	0.1059	0.1031	0.1022	0.1026	0.1179
Rangeability	−0.58%		−0.93%		14.87%	

Table 3: Maximum vertical strain of asphalt layer bottom.

Condition	Bottom of the upper layer (ε)	Bottom of the middle layer (ε)	Bottom of the lower layer (ε)
No pipeline	−0.0251	−0.0033	0.0304
Deep-buried pipe	−0.0252	0.0013	0.0320
Rangeability	0.24%	−138.18%	5.40%
Middle-buried pipe	−0.0263	−0.0013	0.0436
Rangeability	4.78%	−61.21%	43.69%
Shallow-buried pipe	−0.0156	−0.0078	0.0544
Rangeability	−38.03%	137.27%	79.24%

(2) Under the condition of different buried depths of pipelines, the maximum compressive strain appears above the asphalt layer and the maximum tensile strain appears below the asphalt layer.

(3) The bottom tension strain of the asphalt layer is most obviously affected by the buried depth of the pipelines. With the decrease of buried depth of pipelines, the bottom tension strain increases gradually. Taking the shallow-buried pipelines as an example, the tensile strain of the bottom of the asphalt layer increases by 79.24% compared with that of the non-pipeline condition.

The buried depth of pipeline has great influence on the stress state of pavement structure. The pavement structure as a whole is mainly under compression, but there is tension at the bottom of the asphalt underlayer. The vertical strain

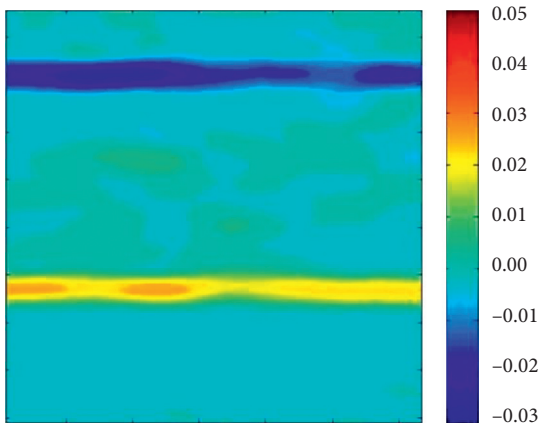


Figure 11: Vertical strain of no buried pipe.

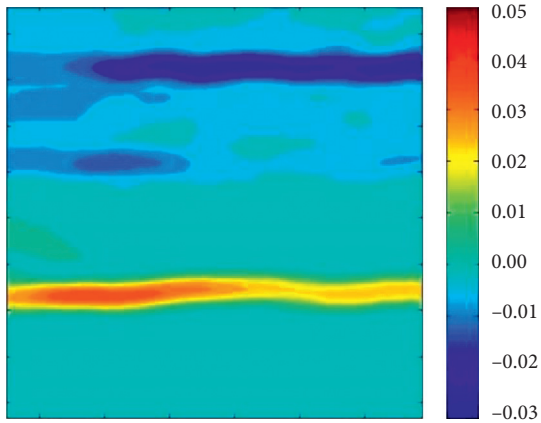


Figure 12: Vertical strain of deep-buried pipe.

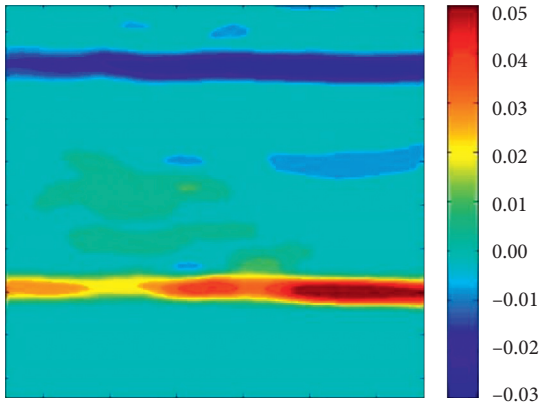


Figure 13: Vertical strain of middle-buried pipe.

value of pavement structure is discontinuous and abrupt. Strain concentration appears at the bottom of each layer. The maximum strain of each layer appears at the bottom of the layer.

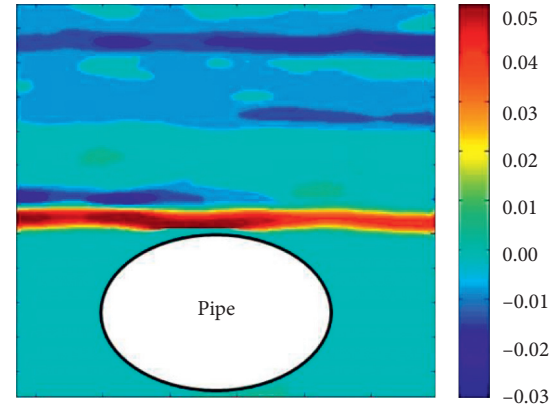


Figure 14: Vertical strain of shallow-buried pipe.

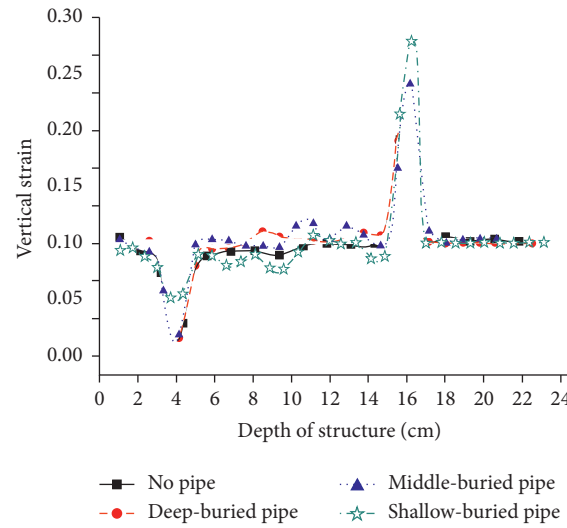


Figure 15: Vertical strain curve of buried pipe structure.

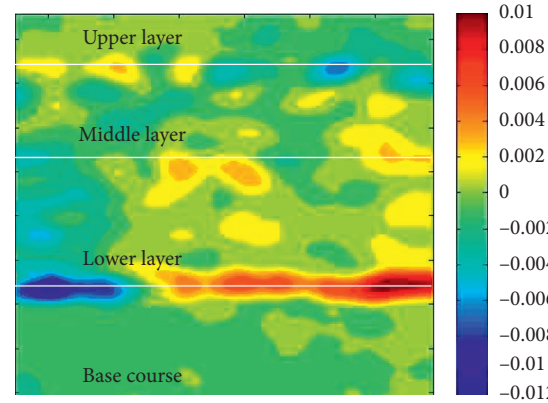


Figure 16: Shear strain of no buried pipe.

4.3. Analysis of the Influence of Buried Pipeline on Shear Strain of Pavement Structure. Shear stress can cause surface cracking, rutting, sliding, and other damages. Figures 16–19

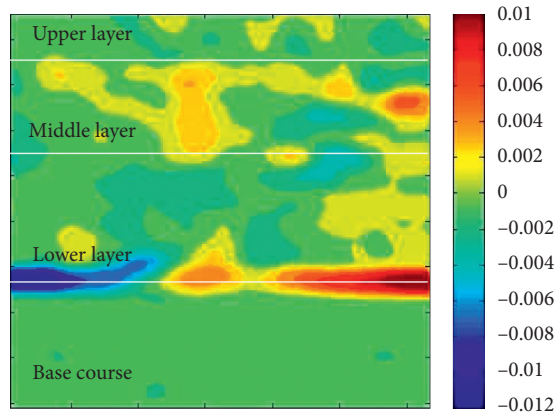


Figure 17: Shear strain of deep-buried pipe.

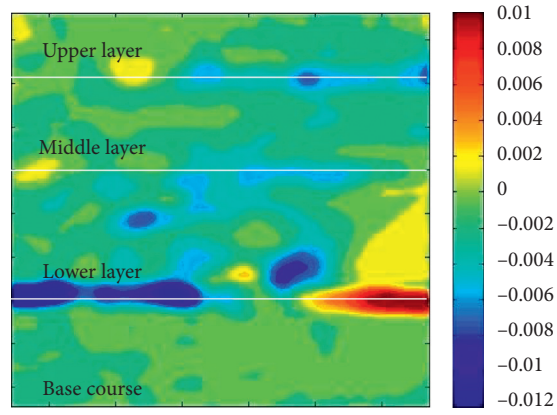


Figure 18: Shear strain of middle-buried pipe.

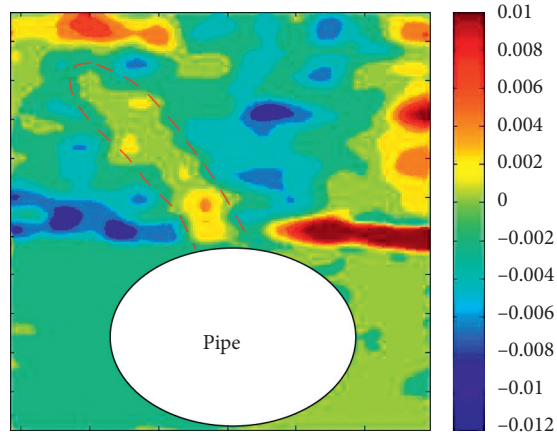


Figure 19: Shear strain of shallow-buried pipe.

show the distribution of shear strain in the pavement structure analysis area under different buried depths.

The maximum shear strain at the bottom of each layer of pavement structure is obtained by sorting out the shear strain data observed in Figures 16–19, as shown in Table 4.

From Figures 16–19 and Table 4, it can be seen that (1) in the pavement structure, the phenomenon of shear strain concentration also occurs at the bottom of each layer, that is, the maximum shear strain of each layer occurs at the bottom of this layer. (2) As shown in Table 4, the maximum shear strain of pavement structure appears at the bottom of the asphalt layer, the second is at the top layer, and the smallest is at the middle layer. The shear strain of the asphalt underlayer is 1.74–4.53 times of that of the other two layers. (3) With the

Table 4: Maximum shear strain of the asphalt layer bottom.

Condition	Bottom of the upper layer/ε	Bottom of the middle layer/ε	Bottom of the lower layer/ε
No pipeline	0.00636	0.00398	0.01108
Deep-buried pipe	0.00629	0.00379	0.01367
Rangeability	−1.24%	−4.72%	23.38%
Middle-buried pipe	0.00625	0.00466	0.01380
Rangeability	−1.82%	17.07%	24.62%
Shallow-buried pipe	0.00516	0.00572	0.02337
Rangeability	−18.91%	43.86%	110.97%

decrease of buried depth of the pipelines, the shear strain of the asphalt layer increases obviously. Taking the case of the shallowly buried pipelines as an example, the shear strain at the bottom of the asphalt subsurface increases by 110.97% compared with that without a pipeline. (4) Under the condition of the shallowly buried pipelines, the shear strain concentration trend of asphalt pavement under standard driving load has appeared, which is inclined to the upper left, as shown in Figure 19. Therefore, the burial of pipelines will accelerate the occurrence of shear failure of the pavement structure.

To sum up, the embedding of pipelines will accelerate the occurrence of shear failure of pavement structure. Similar to the vertical strain, the shear strain is also discontinuous and abrupt. The bottom shear strain of each layer is concentrated. With the decrease of the buried depth of the pipeline, the shear strain of the asphalt layer increases obviously.

5. Conclusion

The digital speckle correlation method (DSCM) has been used to study the load-bearing characteristics of asphalt pavement with buried pipeline under driving load. The experiment has been conducted in the study of the mechanical properties of the asphalt pavement structure of the unpaved and three different kinds of buried tunnels, studied under four different conditions of the asphalt structure and the vertical displacement and the strain value of each layer of asphalt pavement. The influence of the pipeline to the structural bearing properties of asphalt pavement is obtained by contrast analysis.

(1) The digital speckle observation test method can accurately obtain the displacement and strain values of each layer of asphalt pavement structure containing buried pipeline, and the application effect is good. Compared with the traditional contact strain measurement method, this method is simple and accurate and can provide effective analysis data for experimental research.

(2) Because the road structure material is layered, the strength of the material is different for each layer. There is an interlayer effect in the process of stress on asphalt pavement structure, which is mainly reflected in the discontinuity of vertical displacement and strain between layers, and there is a mutation. At the same time, vertical strain and shear strain concentrate at the bottom of each layer. The interlayer effect between the upper layer and the lower layer of asphalt is the most prominent.

(3) The supporting action of the pipeline reduces the vertical displacement of each layer and increases the tension and compression and shear strain of each layer. Pipeline embedding has the greatest influence on the asphalt underlayer, which accelerates the destruction of asphalt pavement structure.

Data Availability

The data used to support the findings of this study have not been made available because the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. Chen, *Digital Speckle Correlation Technique and its Application in Monitoring Structure*, Suzhou University, Jiangsu, China, 2005.

[2] J. Du, *Theoretical Lnvestigation Numerical Simulation and Experimental Research of Manhole Settlement under Traffic Load*, Zhejiang University, Hangzhou, China, 2006.

[3] L. Liu, *Study on the Technology Performance of Asphalt Pavement Layer Interfaces*, Chang'an University, Xi'an, China, 2008.

[4] X. L. Li, S. Z. Li, and Y. G. Shen, "Stress analysis and field testing of buried pipeline under traffic load," *Journal of Zhejiang University*, vol. 48, no. 11, pp. 1976–1982, 2014.

[5] W. H. Peters and W. F. Ranson, "Digital imaging techniques in experimental tress analysis," *Optical Engineering*, vol. 21, no. 3, pp. 427–431, 1982.

[6] Z. M. Wang, *Study on Mechanical Behaviors of Buried Pipelines under Traffic Loads*, Zhejiang University, Hangzhou, China, 2006.

[7] Y. H. Wang, H. Liang, S. Wang, H. Zhang, and L. X. Yang, "Advance in digital speckle correlation method and its applications," *Chinese Optics*, vol. 6, no. 4, pp. 470–480, 2013.

[8] X. D. Wang, "Design of pavement structure and material for full - scale test track," *Journal of Highway and Transportation Research and Development*, vol. 34, no. 6, pp. 30–37, 2017.

[9] B. Pan, H. Xie, and F. L. Dai, "AN investigation of SUB-pixel displacements registration algorithms IN digital image correlation," *Chinese Journal of Theoretical and Applied Mechanics*, vol. 7, no. 2, pp. 245–252, 2007.

[10] I. Yamaguchi, "A laser-speckle strain gauge," *Journal of Physics E: Scientific Instruments*, vol. 14, no. 11, pp. 1270–1273, 1981.

[11] J. T. Yang, *Research on Longitudinal Mechanical Characteristics of Pipelines Buried in Soft Soil under Vertical Loads*, Zhejiang University, Hangzhou, China, 2006.

[12] J. Zhao, *Research on Digital Speckle Correlation Method and its Applications in Mechanical Engineering Measurement*, Beijing Forestry University, Beijing, China, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6686057, 12 pages
https://doi.org/10.1155/2021/6686057

Research Article

An Unsupervised Intelligent Fault Diagnosis System Based on Feature Transfer

Nannan Lu, Songcheng Wang, and Hanhan Xiao

School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China

Correspondence should be addressed to Nannan Lu; lnn_921@126.com

Received 17 December 2020; Revised 26 February 2021; Accepted 10 March 2021; Published 18 March 2021

Academic Editor: Jun Shen

With the booming development of intelligent manufacturing in modern industry, intelligent fault diagnosis systems have become a necessity to equipment and machine, which have attracted many researchers' attention. However, due to the requirements of enough labeled data for most of the current approaches, it is difficult to implement them in real industrial scenarios. In this paper, an unsupervised intelligent fault diagnosis system based on feature transfer is constructed to extract the historical labeled data of the source domain, using feature transfer to facilitate the fault diagnosis of the target domain. The original feature set is acquired by EEMD time-frequency analysis. Then, the transfer component analysis algorithm is adopted to minimize the distance between the marginal distributions of the source and target domains which reduces the discrepancy of features between the different domains. Finally, SVM is used in multiclassification to identify different categories of faults. The performance of the fault diagnosis system under different loads is tested on the CWRU bearing data set, and the experiments show that the proposed system could effectively improve the recognition ability of unsupervised fault diagnosis.

1. Introduction

Rotating machinery is a crucial part of the mechanical system in industrial manufacturing. Its healthy condition seriously affects the safe and stable operations of equipment. It has been demonstrated that 30% of rotating machinery faults are caused by bearing faults [1]. Recently, the bearing fault diagnosis becomes a hot research topic to realize its intelligent surveillance and recognition.

The fault diagnosis methods of rotating machinery can be divided into a model-based method and a data-driven method [2]. The model-based fault diagnosis method is to achieve fault diagnosis by establishing a mathematical model and analyzing the residual error between the mathematical model and the actual signal. Because of the noise and other random factors in the working environment of equipment, the performance of the model-based rolling bearing fault diagnosis is seriously affected. However, data-driven methods collect representative data from signals and design simple models. The data is used to train the model to make it fit, so that we can get an ideal model. Comparatively, data-driven methods are more popular in recent years, owing to the amounts of available data collected from sensors.

Data-driven fault diagnosis methods consist of signal processing, feature extraction, and fault mode recognition [3]. The signal processing aims to obtain the original features by the signal transformation. But, different transformations may bring some redundant information, which will decline the diagnosis accuracy and make the calculation complex. The feature extraction is necessary to remove redundant information. Finally, machine learning methods are used to construct recognition models for fault diagnosis, such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), or Fuzzy Logic (FL) [4–6].

Fourier transformation is usually used to transform rolling bearing signals at the beginning [7]. Yet, since the signal has the features of nonstationary and nonlinear, it cannot get acceptable performance. Some short-time analysis methods such as short-time Fourier transform (STFT), wavelet transform (WT), empirical mode decomposition (EMD), and ensemble empirical mode decomposition

(EEMD) can explore the information hidden in the frequency domain [8]. Due to the fixed size of the window, the resolution of STFT is determined by the window size, so the frequency and time resolution cannot be optimized at the same time [9]. WT is easy to lose the high-frequency components of the signal [10]. EMD has the problem of mode mixing [11]. But, EEMD can remedy the defect of EMD when composing the vibration signals. Then, the features referring to the time domain, frequency domain, and time-frequency domain of vibration signals are extracted, which are taken as the input of the classifier to complete the training and fault diagnosis. The classifiers always use traditional statistical machine learning approaches [12]. However, statistical learning is based on mathematical statistics and requires that the learned knowledge should have the same statistical features as the applications [13]. Therefore, traditional statistical machine learning always assumes that the training and testing data come from the same distribution. However, actually, most of the cases do not obey the same distribution. Transfer learning relaxes the constraint that both training and testing data must obey the same distribution in traditional statistical machine learning [14]. It can learn the domain invariant features or structures between the different but related domains, so as to realize knowledge transfer and reuse between domains [15]. On the other hand, when the training and testing data do not satisfy the same distribution hypothesis, the training data will be out of date. Transfer learning can improve the learning ability of traditional statistical machine learning and greatly reduce the cost of labeling data [16].

Transfer learning is the approach that utilizes the learned knowledge from one domain to facilitate the learning tasks in the new domains [17]. Therefore, using transfer learning, we can learn new knowledge more easily through outdated experiences. Figure 1 shows the signals generated by the sphere fault (SF) and inner race fault (IF), respectively. Due to the different fault locations, the distributions are obviously different from each other. But, there still exist some similarities in the condition of fault occurrence, such as the bearing speed and fault diameter when the fault occurs. Thus, through the diagnosis of the SF, we can learn to recognize the IF.

Therefore, the distinctive characteristic of transfer learning is no requirement of the identical distribution between the training and testing data, which is more suitable for a rapid variation of sensor data [18–20]. Inspired by transfer learning, we try to construct an unsupervised intelligent fault diagnosis system for the real scenario with different distributions and without labeled data in the target domain. In the fault diagnosis system, the domain invariant feature representation must be learned from the extracted features. Unlike the high cost of feature learning in deep neural networks, we utilize EEMD to decompose the original signals and further extract the statistical features, which is used to learn the common feature space between the source and target domains by reducing the marginal distribution discrepancy. In this way, the proposed intelligent fault diagnosis system can uncover the hidden information in the signals and focus on learning the transferable mapping of the statistical features. Herein, we select transfer component analysis (TCA) [21] to transform the source domain and target domain features into a unified feature space, in which the maximum mean discrepancy (MMD) is used to minimize the distance between the source and target domain, so as to achieve accurate diagnosis task of the target without any labeled data. Then, the multiclassification-based SVM is used to identify the unseen faults that are different from the source domain.

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 introduces the proposed intelligent fault diagnosis system from signal processing, feature transferring, and classification. Section 4 describes the experiments, which mainly introduce the selected data set and show the experimental results and analysis. The conclusions are given in Section 5.

2. Related Works

Rotating machinery is often running under high speed and high pressure, where the rolling bearing of mechanical equipment is easy to be damaged and faults occur. Mechanical faults are a serious problem to the development of intelligent manufacturing in modern industry. In order to exactly identify the various categories of rotating machinery faults, many researchers try to propose approaches to improve the performance of intelligent fault diagnosis systems. Liu et al. [22] proposed an intelligent fault diagnosis model which is based on variational mode decomposition (VMD) and singular value decomposition. Yu et al. [23] proposed a deep inception net with atrous convolution (ACDIN) to realize bearing fault diagnosis. Besides, Chen et al. [24] proposed an integrated anomaly detection approach for seeded bearing faults, which use EMD and the Hilbert transformation to extract the feature set.

All the above studies utilize traditional machine learning approaches to implement intelligent fault diagnosis systems. However, once the training and testing data do not obey the same distribution, the performance will significantly decline. In real scenarios, most of the bearing faults happen randomly. It is impossible to label enough samples for training a new model. Therefore, transfer learning is necessary to implement intelligent fault diagnosis systems into real industrial scenarios. Among the current researches about transfer learning, Xu et al. [25] used TrAdaboost to transfer the knowledge of source domain to target after extracting features with WT. TrAdaboost assumed that there are a few labeled samples in the target domain and then constructed a mixed data set including the labeled data from the source and target domain to be the training data set [26]. More distinctively, the algorithm used the weight adjustment of AdaBoost, which determined the weights of samples by the feedback of the classification performance on the labeled target data. Thus, the method could make sure to learn an effective model for the source domain, while it might not obtain acceptable performance on the target task. Considering the corruption possibility of data during the collecting procedure, there exists some extent of uncertainty in both

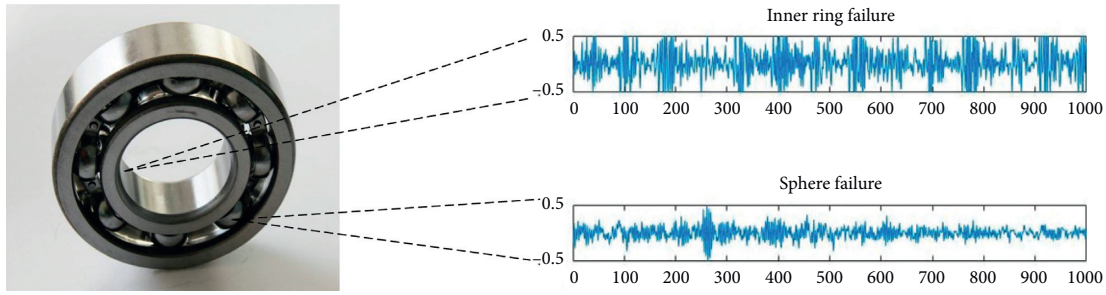


Figure 1: The discrepancy of signals collected from different faults.

the source and target domains. Thus, Xiao et al. [27] proposed to learn the proportions when transferring knowledge from source to target. With the explosive increase of data, transfer learning is combined with deep neural networks to improve the recognition performance of the transferring learning approaches. Prieto et al. [28] proposed a bearing fault diagnosis model based on statistical-time features and neural networks. Shao et al. [29] utilized the scaled exponential linear unit to improve the quality of the feature mapping, which compensated for the lack of labeled samples in the target domain. The good performance of all the deep transfer models benefits from the outstanding ability of the feature extraction of deep neural networks.

However, the training of the deep transfer learning model needs enough samples. Therefore, the study of shallow machine learning methods is still necessary for some real industrial scenarios. Unlike feature learning by some deep neural networks, the proposed intelligent fault diagnosis framework utilizes the statistical features and shallow transfer learning algorithm to learn the feature mapping that could reduce the marginal distribution discrepancy between the source and target domains. In this way, the proposed intelligent fault diagnosis can give another way to solve the data deficiency that may exist in real industrial scenarios. Herein, TCA is used to transform the source domain and target domain features into a unified feature space, in which the maximum mean discrepancy (MMD) is used to minimize the distance between the source and target domain, so as to achieve accurate diagnosis without labeled data [30, 31]. As to the features, we firstly use EEMD to process the signal and extract the feature set and then transfer the features through TCA to establish the unsupervised fault diagnosis model named EEMD-TCA-SVM. It was verified by Case Western Reserve University's (CWRU) public data set. The results show that our proposed system can obtain acceptable performance.

3. Transfer Learning-Based Intelligent Fault Diagnosis

In this paper, EEMD is used to decompose the vibration signals into multiple IMFs. Then, Hilbert envelope spectra (HES) and Hilbert marginal spectra (HMS) are calculated to acquire time and frequency features. After that, the unified feature space is learned by TCA to realize feature transfer from the source domain to the target domain. Finally, various faults are identified by the multiple classifications based on SVM. The specific procedure of the proposed transfer learning-based intelligent fault diagnosis system is described in Figure 2.

3.1. Fault Feature Extraction from Vibration Signals by EEMD. The data here used to extract features are vibration signals collected from accelerometers set on the rolling bearing. Then, it is segmented into short waves having several periods, which is useful to extract the features of time and frequency domains. We select EEMD to decompose the original signals into different IMF components, which improves EMD by adding white Gaussian noise to the signal to eliminate mode aliasing [32].

Before signal decomposition, white Gaussian noise is added to the original signal $x(t)$.

$$x_i(t) = n_i(t) + x(t), \tag{1}$$

where $n_i(t)$ $(i \in M)$ is the ith superimposed white Gaussian noise, and $x_i(t)$ is the corresponding signal with noise to be decomposed later. By subtracting the mean value $m_i(t)$ of the upper and lower envelope from $x_i(t)$, the signal component $h_i(t)$ could be obtained by the equation $h_i(t) = x(t) - m_i(t)$. Then, $h_i(t)$ is taken as a new signal to be decomposed and repeat the above operations till the termination criteria of equation (9) are satisfied.

$$S_D = \sum_{t=0}^{T} \left[\frac{h_{i(k-1)}(t) - h_{ik}(t)}{h_{i(k-1)}(t)} \right]^2, \tag{2}$$

where T denotes the length of the signal. Usually, the range of S_D is [0.2, 0.3]. When the requirements of IMF are satisfied, $h_{ik}(t)$ is the IMF component $c(t)$ we would like to obtain. And then, we can get the remaining subsequence $r(t)$, which is the residual component $c(t)$ from $x(t)$. Repeating the above process, the ultimate residual component $r_n(t)$ is obtained by

$$\begin{cases} x(t) - c_1(t) = r_1(t), \\ r_1(t) - c_2(t) = r_2(t), \\ \cdots \\ r_{n-1}(t) - c_n(t) = r_n(t). \end{cases} \tag{3}$$

Next, equation (3) can be rewritten as equation (4). Obviously, the original signal $x(t)$ can be decomposed into the IMF component and the residual subsequence $r_n(t)$, respectively.

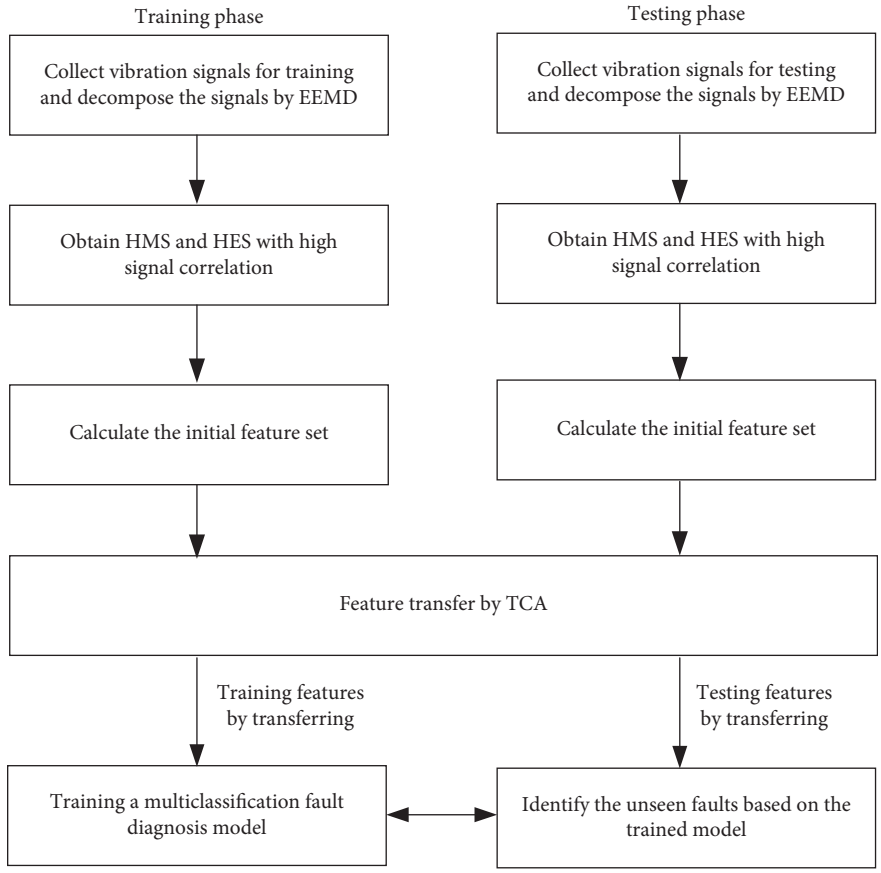


Figure 2: The general workflow of the proposed EEMD-TCA-SVM.

$$x(t) = \sum_{j=1}^{N} c_j(t) + r_n(t), \tag{4}$$

where $c_j(t)$ represents the jth IMF component by N decomposition, which is defined as equation (5). The implementation of EEMD is concretely described in Table 1.

$$c_j(t) = \frac{1}{M} \sum_{i=1}^{M} c_{ij}(t). \tag{5}$$

Additionally, an example is given in Figure 3 to show the decomposition performance of EEMD. The blue waveform is the original vibration signal, and the red ones are IMF1, IMF2, IMF3, IMF4, and residual component, respectively. Figure 3 shows that the original signal can be decomposed into IMF components with different frequencies and amplitudes, which efficiently extract features from the original signal. Through the decomposition, the redundant components can be removed, while preserving signal features.

However, not every IMF component can exactly represent the information of the original signal. The selection of IMF components is necessary after EEMD decomposition. In order to simplify the calculation, the first 4 IMF components are empirically used to do the feature extraction. After that, 9 statistical parameters are used to represent the original signal, HES, and HMS of EEMD decomposition. Table 2 shows the detailed formula of 9 statistical parameters.

In order to extract the features of the time-frequency domain, Hilbert transformation is used to extract the information of the variety of the vibration signal with time and frequency. At first, each IMF component $c_j(t)$ is transformed to $\widehat{c}_j(t)$ by Hilbert transformation of the following equation:

$$\widehat{c}_j(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{c_j(\tau)}{t-\tau} \mathrm{d}\tau. \tag{6}$$

Then, each IMF component is further analyzed to obtain an analytical signal $z_j(t)$ by the following equation:

$$z_j(t) = c_j(t) + j\widehat{c}_j(t) = a_j(t) e^{jf_j(t)}, \tag{7}$$

where $a_j(t)$ is amplitude function that is the spectra envelope actually, and $\phi(t)$ represents a phase function. Then, the Fourier transformation of $a_j(t)$ is HES $F(\omega)$ of the corresponding IMF component. Based on equation (7), Hilbert spectra are calculated by equation (8). After that, HMS can be obtained on the basis of Hilbert spectra, which is specifically shown in equation (9).

$$H(w,t) = \mathrm{Re} \sum_{j=1}^{N} a_j(t) e^{j\phi(t)} = \mathrm{Re} \sum_{j=1}^{N} a_j(t) e^{j \int w_j(t) \mathrm{d}t}, \tag{8}$$

$$h(w) = \int_0^T H(w,t) \mathrm{d}t, \tag{9}$$

Table 1: Pseudocode of EEMD.

Algorithm 1: ensemble empirical mode decomposition
Input: the original signal $x(t)$ and white noise $n_i(t)$ **Output**: IMF component $c_j(t)$ (1) Add white noise $n_i(t)$ to the original signal $x(t)$ to get the new signal $x_i(t)$ (2) Processing $x_i(t)$ with empirical mode decomposition (3) Calculated white noise interference by average the sum of each IMF components to get $c_j(t)$

where T is the length of the whole sequence. The pseudocode of HES and HMS calculation is shown in Table 3.

Figure 4 shows HES and HMS of the randomly selected vibration signals generated by the OF signal with a motor speed of 1797.

3.2. Unified Feature Space Learning between the Source and Target Domains. Different from the traditional machine learning approaches, we consider the real scenario where the training and testing data come from different distributions, $P(X_s) \neq P(X_t)$. If the training data is directly used to train a model for the test, the trained model will show a bad performance on the testing data. It is assumed that a feature mapping Φ lets the distributions of training and testing data approximate each other, $P(\Phi(X_s)) \approx P(\Phi(X_t))$. TCA is a classical transfer learning approach proposed by Pan et al. [31], which realizes transfer learning by mapping the data of the source and target domains into a High-dimensional Reproducing Kernel Hilbert (HRKH) space. It utilizes feature mapping to reduce the distribution discrepancy between different data sets, and we suppose that the conditional distributions can approximate each other by adjusting the marginal distributions. Specifically, when $P(\Phi(X_s)) \approx P(\Phi(X_t))$ is satisfied, there will be $P(Y_s|\Phi(X_s)) \approx P(Y_t|\Phi(X_t))$. Here, maximum mean discrepancy (MMD) is used to estimate the discrepancy between the training and testing data in the feature mapping space. Specifically, it can be calculated by the following equation:

$$\mathrm{DIS}(x_s, x_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \Phi(x_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} \Phi(x_j) \right\|_H, \tag{10}$$

where n_s and n_t are the number of samples in the training and testing set, respectively. $_H$ is the RKHS norm. Equation (11) cannot be calculated directly, which should transform the samples into the mapping space by some kernel method. In order to embed both the training and testing data into a shared low dimensional latent space, TCA introduces a kernel matrix K and a distribution discrepancy matrix L_{ij}. The kernel matrix contains the elements defined on the source domain, target domain, and cross-domain data in the feature mapping space, which are detailed in equation (11). The elements of L_{ij} are calculated by equation (12).

$$K = \begin{bmatrix} K_{ss} & K_{st} \\ K_{ts} & K_{tt} \end{bmatrix}, \tag{11}$$

$$L_{ij} = \begin{cases} \dfrac{1}{n_s^2}, & x_i, x_j \in X_s, \\ \dfrac{1}{n_t^2}, & x_i, x_j \in X_t, \\ -\dfrac{1}{n_s n_t}, & \text{otherwise.} \end{cases} \tag{12}$$

Then, the distance of equation (10) can be rewritten as $tr(KL) - \lambda tr(K)$, where the first term minimizes the distance between distributions, and the second term maximizes the variance in the feature space. λ ($\lambda \geq 0$) is a tradeoff parameter.

$$\begin{aligned} &\min_W tr\left(W^T KLKW\right) + \mu tr\left(W^T W\right), \\ &\text{s.t. } W^T KHKW = I_m, \end{aligned} \tag{13}$$

where $\mu > 0$ is a tradeoff parameter, and I_m is an $m \times m$ identity matrix. H is the centering matrix, which is defined as $H = I_n - (1/n)11^T$. n means the number of samples in training and testing sets. The values after dimension reduction are the mapped features.

3.3. Multicategory Fault Diagnosis. For the classification of possible errors, a penalty term $C\sum_{i=1}^{n} \xi_i$ is introduced. The following relation is obtained:

$$\varphi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i. \tag{14}$$

The objective function $(1/2)\|w\|^2$ of the optimal hyperplane can be replaced by $\varphi(w, \xi)$. And in general, the penalty factor C is a nonnegative real number; the solution formula of the optimal hyperplane can be expressed as follows:

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \dfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i \cdot x_j\right), \\ \text{s.t. } 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n, \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases} \tag{15}$$

The optimal hyperplane can be obtained by solving the above objective. To sum up, the decision function of SVM can be composed of the inner product and summation of the support vector. Therefore, the decision function of SVM is similar to neural networks in form. Each intermediate node

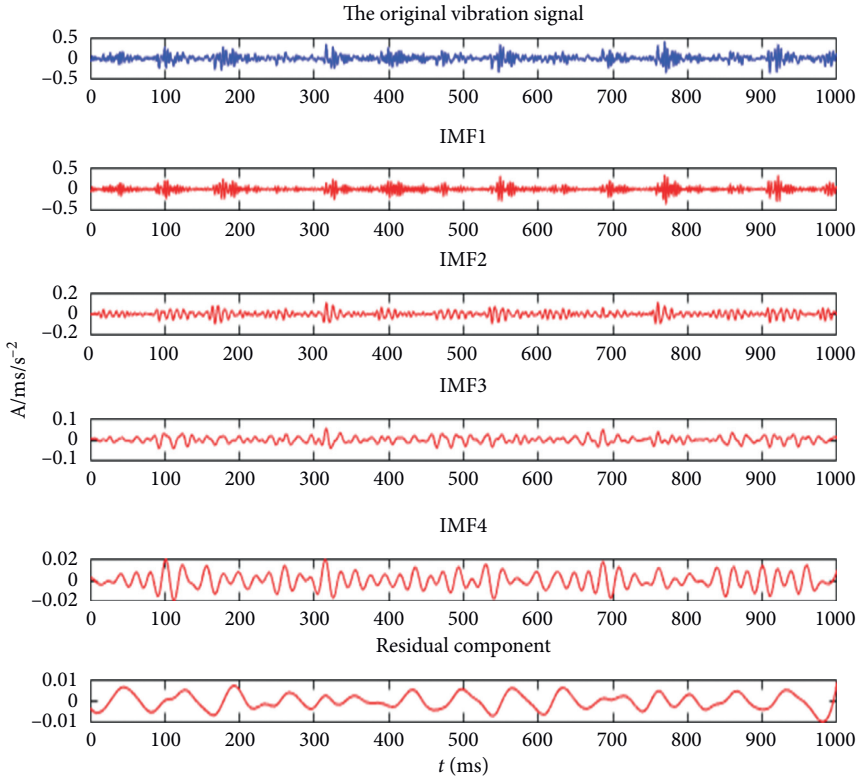


Figure 3: The IMF components of the signal decomposed by EEMD.

Table 2: Statistical parameters.

Feature	Expression
Mean	$T_1 = (1/n)\sum_{i=1}^{n} x(i)$
Standard deviation	$T_2 = \sqrt{(1/(n-1))\sum_{i=1}^{n}(x(i)-T_1)^2}$
Skewness	$T_3 = \sum_{i=1}^{n}(x(i)-T_1)^3/((n-1)T_2^3)$
Kurtosis	$T_4 = \sum_{i=1}^{n}(x(i)-T_1)^3/((n-1)T_2^3)$
Crest factor	$T_5 = \max\lvert x(i)\rvert/\sqrt{(1/n)\sum_{i=1}^{n} x(i)^2}$
Form factor	$T_6 = \sqrt{(1/n)\sum_{i=1}^{n} x(i)^2/(1/n)\sum_{i=1}^{n}\lvert x(i)\rvert}$
Impact factor	$T_7 = \max(\lvert x(i)\rvert)/\sqrt{(1/n)\sum_{i=1}^{n}\lvert x(i)\rvert}$
Latitude factor	$T_8 = \max(\lvert x(i)\rvert)/(1/n)\sum_{i=1}^{n}\lvert x(i)\rvert$
Range	$T_9 = \max(\lvert x(i)\rvert) - \min(\lvert x(i)\rvert)$

corresponds to the inner product of the input sample and support vector $x_1, x_2, \ldots, x_n$ completed by kernel function, and the output vector is a linear combination of intermediate nodes.

The fault diagnosis studied in the paper is a ten-class classification problem, but SVM is usually used to deal with binary classification. Thus, we combine multiple SVMs to construct a multiclass classifier. At first, one of the SVMs is used to identify the faults of category 1 from category 2 to 10. Likewise, the other 9 categories are classified by the binary classifier in the same way.

Table 3: Pseudocode of HES and HMS calculation.

Algorithm 2: Hilbert envelope spectra and marginal spectra
Input: IMF component $c_j(t)$ of the signal **Output**: HES $F(\omega)$ and HMS $h(\omega)$. (1) Transform each IMF component to get $\widehat{c}_j(t)$ by the Hilbert transformation. (2) Calculate each IMF component's analytical signal $z_j(t)$, obtain envelope function $a_j(t)$ and phase function $\phi_j(t)$ (3) Transform $a_j(t)$ by Fourier transformation to get HES $F(\omega)$ (4) Calculate the Hilbert spectra $H(\omega, t)$ (5) Calculate HMS $h(\omega)$ by $H(\omega, t)$

4. Experimental Analysis

4.1. Data Set. In this paper, the vibration signals of bearing faults are collected from the platform of Case Western Reserve University (CWRU) [33]. The bearing device is shown in Figure 5, which is composed of a three-phase induction motor, a torque sensor, and a dynamometer. Four kinds of motor loads of 0, 1, 2, and 3 HP are given in the database, referring to different categories of vibration signals. The sampling frequency is 12 kHz. The experimental data used in the following comes from the upper side of the drive end of the motor. The torque sensor collects the vibration signals in different fault conditions at the drive end. Moreover, SVM, TCA, and EEMD-SVM are used to be compared with our EEMD-TCA-SVM, which further demonstrates the feasibility of the proposed intelligent fault diagnosis system.

In the experiments, four data sets are prepared, which refers to different motor loads shown in Table 4. A, B, C, and

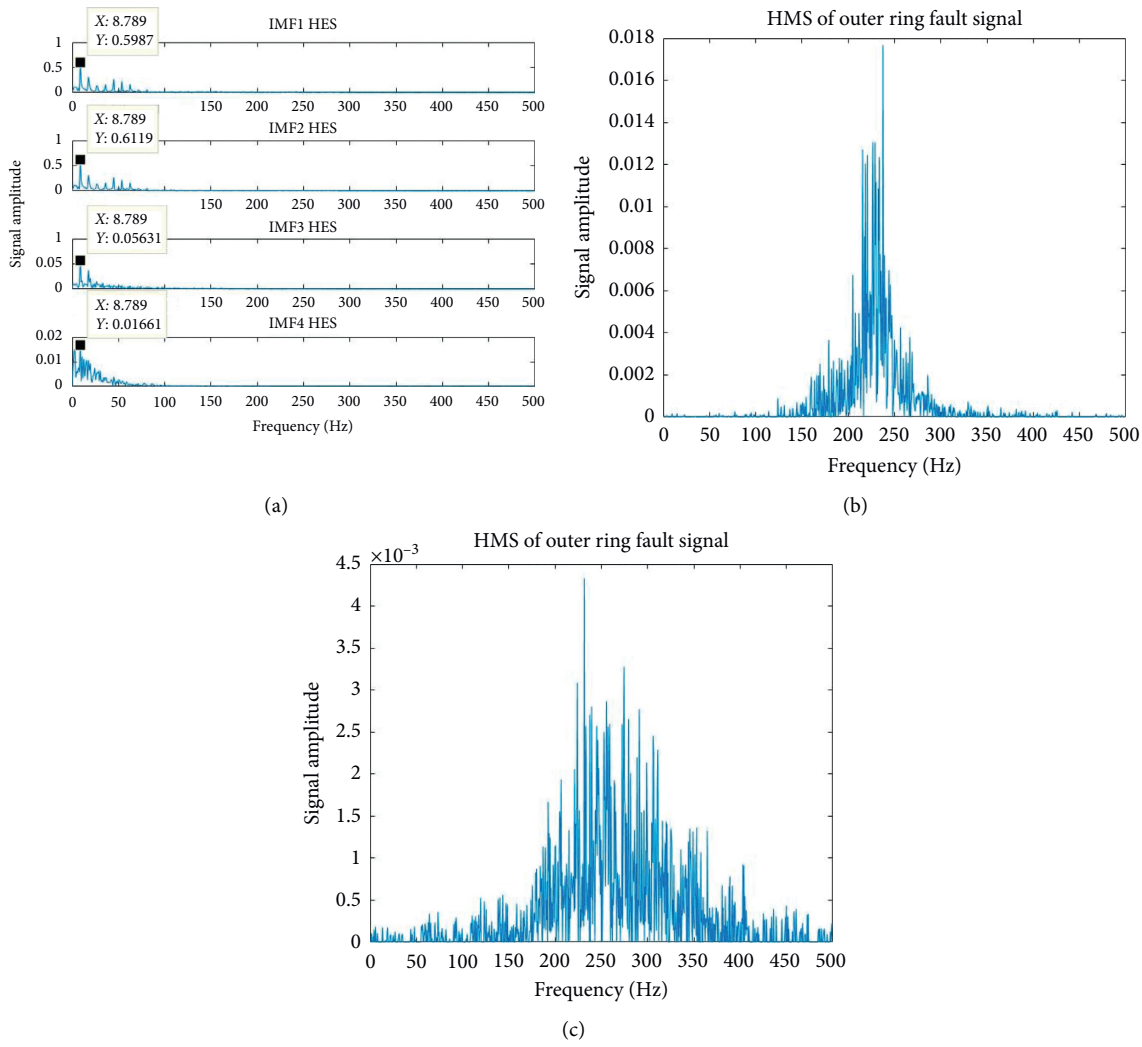


FIGURE 4: The examples of HES and HMS of the bearing fault signals. (a) HES of the OF signal. (b) HMS of the OF signal. (c) HMS of the IF signal.

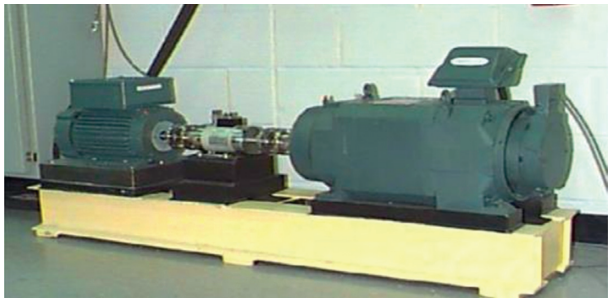


FIGURE 5: CWRU bearing fault test-bed.

D are the bearing fault data sets under the motor loads of 0, 1, 2, and 3 HP, respectively. There are ten fault types in total, including IF, SF, and OF with three kinds of diameters and health samples. For each category, the vibration signal with length 120,000 is selected, where the window and the step size are 2000 and 1000, respectively. Each experimental data set (such as A) has 10 categories. There are 1200 samples for each data set, where 960 samples are taken as training set, and the other 240 samples are testing set.

4.2. Experimental Steps and Result Analysis. In order to verify the performance of the proposed method, we use SVM, TCA-SVM, EEMD-SVM, and EEMD-TCA-SVM for comparing their classification performance on different transfer pairs among A, B, C, and D, respectively. Totally, 12 groups of experiments can be set, which is shown in Table 4.

When the data set is set up, the training and testing sets correspond to the source and target domain data in transfer learning. SVM is trained by the training set, and the testing set is then used to check the classification performance. As to TCA-SVM, both the training and testing sets are used to obtain the unified feature space by minimizing the distribution distance between the training and testing sets with TCA. Then, the training set is used to train SVM. The testing set is mapped to the unified feature space and then classified by SVM. As to EEMD-SVM and EEMD-TCA-SVM, the data sets are processed by EEMD. The first four IMFs are used to

Table 4: Experimental data sets.

Fault location	Diameter	Label	Data set A	Data set B	Data set C	Data set D
Motor load	—	—	0	1	2	3
Health	0	1	120	120	120	120
IF	0.07	2	120	120	120	120
	0.014	3	120	120	120	120
	0.021	4	120	120	120	120
SF	0.07	5	120	120	120	120
	0.014	6	120	120	120	120
	0.021	7	120	120	120	120
OF	0.07	8	120	120	120	120
	0.014	9	120	120	120	120
	0.021	10	120	120	120	120

calculate HMS and HES, where 9 statistical features are calculated. Considering 9 statistical features of the original signal, there are 81 features in total. And the following procedures are the same with SVM and TCA-SVM.

In order to verify the superiority of the EEMD-TCA-SVM model over the other methods, we give the accuracy, ROC curve, AUC value, and confusion matrix in the following.

4.2.1. Accuracy. Accuracy is an important standard to measure fault diagnosis systems, which denotes the ratio of correctly predicted samples to the total samples. Through accuracy, we can easily evaluate the diagnosis performance as a whole. Table 5 shows the accuracies of the methods on the different source-target pairs.

From the results of the first four groups of Table 5, EEMD-TCA-SVM can obtain a relatively higher average accuracy than other methods, where TCA shows the transferability from the average accuracy. In particular, for some cases such as C $\longrightarrow$ D, it can get an almost 20% increase. Compared with TCA-SVM, EEMD-TCA-SVM shows good performance on both average accuracy and each case, which is improved obviously. Thus, the process of the original signals by EEMD is necessary for the fault diagnosis system since the hidden information of different resolutions in time and frequency domains can be extracted through EEMD. In order to verify the reliability of the experiment, Random Forest (RF) is taken as an additional classifier to test the diagnosis performance of the transfer tasks. EEMD-TCA-RF can obtain a higher average accuracy than other methods. Comparing the results of TCA-RF with RF, TCA can effectively minimize the distribution discrepancy between the source and target domain, where the recognition accuracy is improved by about 16%. Comparing the results of TCA-RF with EEMD-TCA-RF, the accuracy is improved by about 30%. EEMD can effectively extract the important information from the original signal. Comparing the results of EEMD-TCA-RF with EEMD-RF, the accuracy is improved by about 5%. The reason is that the decomposition by EEMD and the calculation of the components' statistical features may alleviate the distribution discrepancy of the original signals to some extent, which does not improve the diagnosis performance so much. Overall, the classifier RF on the different tasks of Table 3 has identical conclusions with the classifier SVM.

4.2.2. Confusion Matrix. The confusion matrix represents the fact that the specific numbers of samples are classified into each category, and then the matrix is used to display the results [34]. The confusion matrix is mostly used to judge the quality of the classifier, which is applicable to the classification methods. It is the basic, intuitive, and simple way to further measure the accuracy of classification methods or systems.

The fault diagnosis is a multiclassification problem. The confusion matrix is a table with the size of $10 * 10$. Figure 6 shows the confusion matrix of SVM, TCA-SVM, EEMD-SVM, and EEMD-TCA-SVM, respectively.

Compared to the other methods, EEMD-TCA-SVM can identify most of the categories accurately. SVM shows the worst performance on the confusion matrix, where some of the categories cannot be recognized completely. In particular, for the healthy category (label 1), all the healthy data is identified as faults shown in Figure 6(a). Other fault categories are also easy to misclassify with each other. SVM does not have transferability, which is not used to do the fault diagnosis directly. When TCA is used to transfer features, the recognition performance in Figure 6(b) is improved to a certain extent but still shows very low classification accuracy. Although most of the healthy cases are identified correctly, the faults are misclassified between each other seriously. Therefore, it is infeasible to transfer the signals without any feature extraction. EEMD is a signal processing method that can separate the signals into different IMF components. In the procedure, the more distinguished information can be found. Based on the separation, the statistical features are calculated, which construct the new fault diagnosis data. Figures 6(c) and 6(d) truly show the improvement of the recognition performance by EEDM. But in the case B $\longrightarrow$ D, EEMD-SVM misclassifies all the healthy data to the 7^{th} category of faults in which the two categories of data may have more similarity in statistical features. Likewise, EEMD-TCA-SVM improves the recognition rate for almost all the categories by comparison with EEMD-SVM, especially for the healthy data. The domain adaptation is effective for the data with the

Table 5: Accuracy on the different source-target pairs.

	SVM	RF	TCA-SVM	TCA-RF	EEMD-SVM	EEMD-RF	EEMD-TCA-SVM	EEMD-TCA-RF
A-B	0.2581	0.1063	0.2709	0.3307	0.6212	0.4765	0.6914	0.5538
A-C	0.1745	0.1021	0.2060	0.2752	0.6595	0.5659	0.6529	0.6009
A-D	0.2340	0.1148	0.2667	0.2051	0.4808	0.5659	0.5316	0.5128
B-A	0.1106	0.1191	0.2342	0.2965	0.5447	0.4936	0.5529	0.5529
B-C	0.1277	0.1148	0.2222	0.2803	0.7446	0.7659	0.8283	0.7687
B-D	0.0809	0.1106	0.2410	0.2189	0.5361	0.3787	0.5923	0.4102
C-A	0.1064	0.0936	0.2897	0.2521	0.5532	0.5148	0.5274	0.5778
C-B	0.2341	0.1021	0.2863	0.2598	0.8851	0.7829	0.8847	0.7821
C-D	0.3362	0.1063	0.3471	0.3265	0.4597	0.3787	0.6521	0.5231
D-A	0.1957	0.1277	0.2623	0.2641	0.5974	0.5106	0.5728	0.4905
D-B	0.1574	0.1234	0.2940	0.2623	0.4978	0.4212	0.7103	0.6102
D-C	0.1914	0.1064	0.2623	0.2923	0.5829	0.5148	0.6128	0.5641
Average	0.1839	0.1106	0.2652	0.2719	0.5969	0.5307	0.6507	0.5789

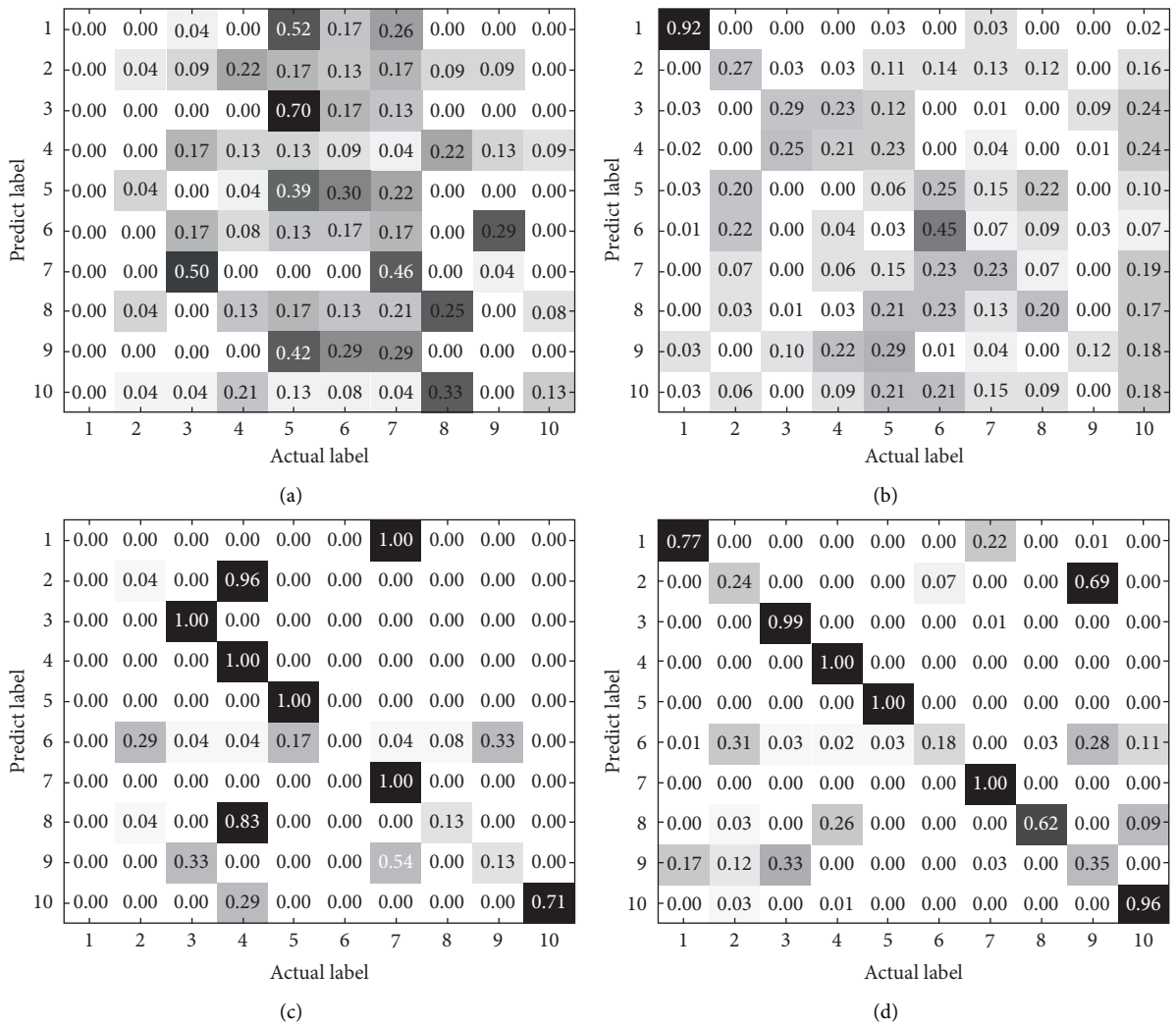


Figure 6: Confusion matrix. (a) SVM. (b) TCA-SVM. (c) EEMD-SVM. (d) EEMD-TCA-SVM.

distribution discrepancy. However, for SVM and TCA-SVM, EEMD-SVM and EEMD-TCA-SVM have more obvious improvement. So, we think that statistical features extracted by EEMD may alleviate the impact of the distribution discrepancy existing in the original signals on the fault diagnosis. It is very necessary to introduce the signal processing-based feature extraction into fault diagnosis systems.

4.2.3. ROC and AUC. Although the proportion of the correct classified samples to the whole testing set can be illustrated by the classification accuracy and confusion matrix, it neglects the relationship between false positive rate (the probability of negative samples wrongly categorized as positive) and true positive rate (the probability of positive samples correctly categorized as negative). Therefore, we further use Receiver Operating Characteristic (ROC) [35] curve and Area Under Curve (AUC) value [36] to evaluate the classification performance. ROC is the way to directly show the relations of FPR (False Positive Rate) and TPR (True Positive Rate). As shown in Figure 7, FPR and TPR are horizontal and vertical axis, respectively. AUC denotes the area under the ROC curve, which provides another way to evaluate the performance of the method. If the method is ideal, its AUC value equals 1. The AUC value of a random model equals 0.5.

Figure 7 illustrates the ROC curves and AUC values of SVM, TCA-SVM, EEMD-SVM, and EEMD-TCA-SVM. By the comparison, we can see that TCA can improve the unsupervised fault diagnosis performance. There are 10 categories in the fault diagnosis problem including healthy condition. All the ten categories are divided into two parts which are healthy and fault. As shown in Figure 7, the curves with different colors correspond to EEMD-TCA-SVM, EEMD-SVM, TCA-SVM, and SVM, respectively. EEMD-TCA-SVM obtains the best ROC curve and the highest AUC value among the four methods while SVM gets the worst ROC curve and AUC value, which are stochastic results. EEMD-SVM gets better performance than SVM and TCA-SVM, which further demonstrates that the feature quality seriously impacts the classification performance. Relatively, the impact of TCA is not so obvious from the comparison between EEMD-TCA-SVM and EEMD-SVM. The AUC values of the two methods are almost the same. In addition, the distributions of the extracted features by EEMD may have a stronger similarity than the distributions of the original vibration signals, which may be one of the reasons for the higher AUC value of EEMD-SVM.

Based on the above results, the feature selection is shown as a very important function in fault diagnosis. Traditional machine learning approaches cannot automatically mine the hidden information from sensor signals. Transfer learning can facilitate the unsupervised fault diagnosis and get promising classification results. The proposed transfer fault diagnosis system still has a bigger promotion space in the future. The ROC curve of the data after data preprocessing is obviously above the ROC curve without data preprocessing, and its AUC value is significantly increased compared with the value without data preprocessing. This shows that the performance of the model has been greatly improved after our data preprocessing; the ROC curve of the data processed by TCA is always at the upper end of the model without TCA processing, and the AUC value is also large. It shows that TCA can improve the performance of the unsupervised model.

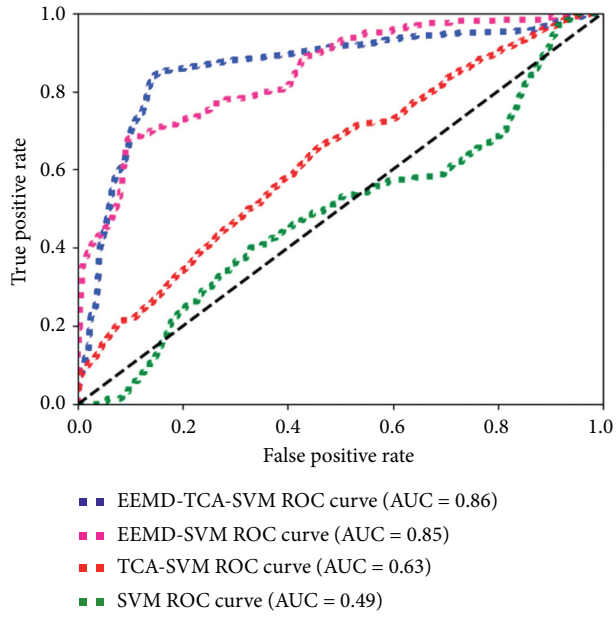


Figure 7: ROC curve and AUC value.

5. Conclusion

In this paper, we construct a transferable intelligent fault diagnosis system, which can transfer the statistical features across domains. In the proposed system, the original vibration signals are decomposed by the EEMD algorithm at first. And then, 81 statistical features are calculated to be the initial feature set, which are transferred by TCA to further obtain the sharable features between the different distributions. By minimizing the marginal distributions of the source and target domain, TCA does not need any extra knowledge to assist the transfer. Then, SVM is taken as the classifier to identify different categories of faults. The experiments on the bearing data set of CWRU show that the proposed system has good accuracy, confusion matrix, ROC curve, and AUC value among the four methods. From the specific results, EEMD can extract the hidden information from the signal, and TCA can calculate the common feature space of different domains for fault diagnosis.

Data Availability

The bearing data used to support the findings of this study have been deposited in the Bearing Data Center of Case Western Reserve University repository (https://csegroups.case.edu/bearingdatacenter/home).

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 (62006233, 51734009, U1710120, and 51504241).

References

[1] I. Attoui, N. Fergani, N. Boutasseta, B. Oudjani, and A. Deliou, "A new time-frequency method for identification and classification of ball bearing faults," *Journal of Sound and Vibration*, vol. 397, pp. 241–265, 2017.

[2] S. X. Ding, P. Zhang, T. Jeinsch, E. L. Ding, P. Engel, and W. Gui, "A survey of the application of basic data-driven and model-based methods in process monitoring and fault diagnosis," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 12380–12388, 2011.

[3] Z. Gao, S. X. Ding, and X. D. Steven, "A survey of fault diagnosis and fault-tolerant techniques-Part I: fault Diagnosis with model-based and signal-based approaches," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3757–3767, 2015.

[4] C.-f. Lin and S.-d. Wang, "Training algorithms for fuzzy support vector machines with noisy data," *Pattern Recognition Letters*, vol. 25, no. 14, pp. 1647–1656, 2004.

[5] R. P. Nikhil, P. Kuhu, and M. K. James, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, 2005.

[6] X. Xue and J. Zhou, "A hybrid fault diagnosis approach based on mixed-domain state features for rotating machinery," *ISA Transactions*, vol. 66, pp. 284–295, 2017.

[7] Z. Feng, M. Liang, and F. Chu, "Recent advances in time-frequency analysis methods for machinery fault diagnosis: a review with application examples," *Mechanical Systems and Signal Processing*, vol. 38, no. 1, pp. 165–205, 2013.

[8] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 108–126, 2013.

[9] Y. J. Guo, Z. W. Fang, and X. F. Chen, "A new improved synchrosqueezing transform based on adaptive short time Fourier transform," in *Proceedings of the IEEE Far East Forum on Nondestructive Evaluation/Testing*, Chengdu, China, June 2014.

[10] X. M. Ye and X. H. Liu, "The Harmonic detection based on wavelet transform and FFT for electric ARC furnaces," in *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, pp. 408–412, Baoding, China, July 2009.

[11] S. Gaci, "A new ensemble empirical mode decomposition (EEMD) denoising method for seismic signals," *Energy Procedia*, vol. 97, pp. 84–91, 2016.

[12] H. Y. Jiang, H. T. Liu, and X. Shu, "Multi-label transfer learning via maximum mean discrepancy," *Information and Control*, vol. 45, no. 4, pp. 463–470, 2016.

[13] A. K. Jain, P. W. Duin, and J. Jianchang Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[14] A. Margolis, "A literature review of domain adaptation with unlabeled data," Technical Report, 2011.

[15] J. Quionerocandela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*, MIT Press, Cambridge, MA, USA, 2009.

[16] L. Zhang, "Transfer adaptation learning: a decade survey," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[17] M. Baktashmotlagh, M. Harandi, and B. C. Lovell, "Unsupervised domain adaptation by domain invariant projection," in *Proceedings of the International Conference on Computer Vision*, Sydney, Australia, December 2013.

[18] L. Shao, F. Zhu, and X. D. Li, "Transfer learning for visual categorization: a survey," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[19] C. Jia and Y. Fu, "Low-rank tensor subspace learning for RGB-D action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4641–4652, 2016.

[20] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost, "Machine learning for targeted display advertising: transfer learning in action," *Machine Learning*, vol. 95, no. 1, pp. 103–127, 2014.

[21] D. D. Liu, Y. R. Li, and Z. B. Xu, "Application of selective ensemble transfer algorithms in the field of bearing fault diagnosis," *Machinery Design & Manufacture*, vol. 000, no. 5, pp. 28–31, 2020.

[22] C. Liu, Y. Wu, and C. Zhen, "Rolling bearing fault diagnosis based on variational mode decomposition and fuzzy c means clustering," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 35, no. 13, pp. 3358–3365, 2015.

[23] D. J. Yu, M. F. Chen, and J. S. Cheng, "A fault diagnosis approach for rotor systems based on empirical mode decomposition method and support vector machines," *Proceedings of the CSEE*, vol. 26, no. 16, pp. 162–167, 2006.

[24] Y. Chen, G. Peng, C. Xie, W. Zhang, C. Li, and S. Liu, "ACDIN: bridging the gap between artificial and real bearing damages for bearing fault diagnosis," *Neurocomputing*, vol. 294, no. 14, pp. 61–71, 2018.

[25] J. Xu, "Fault diagnosis of aircraft fuel pump based on wavelet packet and transfer learning," *Chinese Hydraulics & Pneumatics*, vol. 6, no. 29, pp. 183–188, 2020.

[26] W. Y. Dai, Q. Yang, G. R. Xue, and R. Yu, "Boosting for transfer learning," in *Proceedings of the International Conference on Machine Learning*, pp. 193–200, Corvalis, OR, USA, 2007.

[27] Y. Xiao, H. Wang, and B. Liu, "A new transfer learning-based method for label proportions problem," *Information Sciences*, vol. 541, pp. 391–408, 2020.

[28] M. D. Prieto, G. Cirrincione, A. G. Espinosa, J. A. Ortega, and H. Henao, "Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 8, pp. 3398–3407, 2013.

[29] H. D. Shao, X. Y. Zhang, J. S. Cheng, and Y. Yang, "Intelligent fault diagnosis of bearing using enhanced deep transfer auto-encoder," *Journal of Mechanical Engineering*, vol. 56, no. 9, pp. 84–91, 2020.

[30] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[31] G. Matasci, M. Volpi, and M. Kanevski, "Semisupervised Transfer component analysis for domain adaptation in remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3350–3564, 2015.

[32] Q. Fu, B. Jing, P. He, S. Si, and Y. Wang, "Fault feature selection and diagnosis of rolling bearings based on EEMD and optimized elman_adaBoost algorithm," *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5024–5034, 2018.

[33] Bearings data center seeded fault test data, https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website.

[34] L. Chen and H. L. Tang, "Improved computation of beliefs based on confusion matrix for combining multiple classifiers," *Electronics Letters*, vol. 40, no. 4, pp. 238-239, 2004.

[35] X. He, B. D. Gallas, and E. C. Frey, "Three-class ROC analysis--toward a general decision theoretic solution," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 206–215, 2010.

[36] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1635708, 9 pages
https://doi.org/10.1155/2021/1635708

Research Article

Sentence Similarity Calculation Based on Probabilistic Tolerance Rough Sets

Ruiteng Yan,[1] Dong Qiu ,[1,2] and Haihuan Jiang[1]

[1]*School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Nanan, Chongqing 400065, China*
[2]*College of Science, Chongqing University of Posts and Telecommunications, Nanan, Chongqing 400065, China*

Correspondence should be addressed to Dong Qiu; dongqiumath@163.com

Received 15 August 2020; Revised 1 December 2020; Accepted 15 January 2021; Published 28 January 2021

Academic Editor: Jun Shen

Sentence similarity calculation is one of the important foundations of natural language processing. The existing sentence similarity calculation measurements are based on either shallow semantics with the limitation of inadequately capturing latent semantics information or deep learning algorithms with the limitation of supervision. In this paper, we improve the traditional tolerance rough set model, with the advantages of lower time complexity and becoming incremental compared to the traditional one. And then we propose a sentence similarity computation model from the perspective of uncertainty of text data based on the probabilistic tolerance rough set model. It has the ability of mining latent semantics information and is unsupervised. Experiments on SICK2014 task and STSbenchmark dataset to calculate sentence similarity identify a significant and efficient performance of our model.

1. Introduction

With the rapid development of information technique, innumerable text data are continuously growing. Unlike digital data, the processing of text data is more complex and difficult. Sentence similarity aims at calculating the degree of resemblance or distance between two sentences. It plays an important role in the application of natural language processing (NLP), like text summarization [1, 2], machine translation [3], question answering systems [4], and information retrieval [5]. These applications are based on sentence similarity to a certain extent, whose development makes the research of sentence similarity become urgent.

Text data are characterized by uncertainty, inaccuracy, and incompleteness. Existing sentence similarity computation methods are almost all based on the relation among words and words in the sentences or based on the deep learning algorithms. Methods based on the relation among the words and words such as word cooccurrence mainly consider sentence semantics from the shallow level and cannot capture the latent semantics information behind the sentences. Methods based on the deep learning algorithms such as convolutional neural network (CNN) can capture deep semantics information, but most of them are with high time complexity and supervision. In addition, both of the classes of methods cannot commendably process the uncertainty and imprecision of text sentences. In this paper, we start with the uncertainty and imprecision of text data. We improve the tolerance rough set model [6] by Ho et al.and present a sentence similarity computation model based on the probabilistic tolerance rough set model. Our model can not only process the uncertainty and imprecision of text data, but also overcome the shortcomings mentioned before.

This paper is organized as follows. Some related works on sentence similarity measures are reviewed in Section 2. Section 3 presents our proposed probabilistic tolerance rough sets-based model for sentence similarity computation in detail. Section 4 demonstrates the experimental results and discusses on sentence similarity tasks. In Section 5, some conclusions are made.

2. Related Work

The main work is to improve the traditional tolerance rough set model, and then establish a sentence similarity computation model based on the probabilistic tolerance rough set model. In this section, we discuss some related works about sentence similarity calculation methods and tolerance rough set models in NLP.

2.1. Sentence Similarity Calculation. Traditional works about sentence similarity are generally categorized into two classes, methods based on shallow semantics and methods based on deep learning algorithms. The idea of shallow semantics methods is to calculate the similarity between words. Methods based on words' cooccurrence and on corpus are two representatives. Methods based on words' cooccurrence are mentioned in [7–9]. Han et al. used the Bag-of-Words (BoW) technique [8], and Jones et al. [7] applied the term frequency inverse document frequency (TF-IDF) technique to represent sentences, and then the cosine distance or Euclidean distance was utilized to calculate the similarity between sentences. A keyword-based approach was proposed [9], which calculates the keywords' ranking score extracted in the sentences. Methods based on corpus such as WordNet, HowNet are mentioned in [10, 11]. In [12], Prasad et al. combined common words and semantic features for measuring sentence similarity. They extracted both syntactic features by searching for common words between sentences and semantic features by utilizing information content of sentences. Methods based on shallow semantics can only obtain the literal meaning of sentences, and fail to capture high-level semantics information behind sentences.

Nowadays, neural network and deep learning have been widely used in NLP and have made great achievements. By training sentences with deep learning algorithms, deep semantics information can be captured in the computation of sentence similarity. In [13], a CNN-based parallel semantic matching model was established; two parallel CNNs were built to train two sentences, respectively. Then, the two CNNs were cascaded into one multilayer neural network for matching the similarity of sentences. An elaborate convolutional network (ConvNet) variant was presented [14], which inferred sentence similarity by integrating differences of convolutions at different scales. For the problem of variable length sentences and complex sentences, Mueller et al. proposed a Siamese Network on the basis of the long short-term memory (LSTM) model [15]. Methods mentioned before mainly concentrate on the similar information of two sentences; on the bias, methods concentrated on the dissimilar information of two sentences were proposed. Wang et al. developed a sentence similarity learning model by decomposing and composing lexical semantics which considered both the similar information and dissimilar information between sentences [16]. In [17], a context-aligned recurrent neural network (CA-RNN) model was put forward. In this model, the contextual information of the aligned words was integrated in the neural network. Liu et al. incorporated the shallow semantics and deep information to evaluate the sentence similarity [18]. The shallow part is represented by the lexical similarity based on keywords and sentence lengths; and the deep part is modeled by a parallel CNN which extracts both the whole sentence and their context as the features. However, most of the sentence similarity learning algorithms based on neural network and deep learning are supervised, which need to train the data set first. Jacob et al. [19] proposed an unsupervised Bidirectional Encoder Representations from Transformers (BERT) model, which has reached excellent results for language representation.

It is undeniable that text data possess uncertainty, imprecision, and incompleteness. However, methods mentioned above do not measure the similarity between sentences from the perspective of uncertainty and imprecision. Fuzzy set theory and rough set theory are created to process such uncertainty and imprecision. A fuzzy set and rough sets-based approach was developed for measuring cross-lingual semantic similarity [20]. In [1], Chatterjee et al. proposed a fuzzy rough sets-based model. Sentence similarity was computed according to the upper approximation and lower approximation of two sentences.

We improve the traditional tolerance rough set model and propose a sentence similarity computation model based on the probabilistic tolerance rough set model. With the model from the point of the uncertainty of text data to process text data, it can not only solve the problem of inability to obtain high-level semantics information on methods based on shallow semantics, but also overcome the drawback of supervision on methods based on deep learning algorithms, with the advantages of capturing more latent semantics information and nonsupervision.

2.2. Tolerance Rough Sets in NLP. Rough set theory was proposed by the Polish scholar Pawlak for handling uncertainty, imprecision, and fuzziness in 1982 [21]. It has been effectively applied in the field of machine learning, data mining, and NLP [22–24]. Rough sets partition a set X by using an equivalence relation. Whether one certain object belongs to a set X or not is represented by a pair of concepts called lower approximation space and upper approximation space. A possible part is the upper approximation except the lower approximation, called the boundary region. Researchers generalized rough set theory to some expanded models according to different requirements, including probabilistic rough set model [25], decision rough set model [26], and tolerance rough set model [27]. An equivalence relation contains three properties of reflexivity, symmetry, and transitivity, in which the limitation of transitivity leads to the inapplicability in some cases. The tolerance relation was introduced to replace the equivalence relation by Skowron et al. since some applications cannot achieve the condition of transitivity, and the corresponding model was tolerance rough set model [27].

With the tolerance rough set model applied in NLP, a search result clustering method was put forward [28], in which the tolerance relation was defined as the number of word cooccurrences in documents. In [29], a tolerance

rough sets-based semantic clustering algorithm is introduced by Meng et al. for web search results, extending the original text semantics and processing the limitation on the sparsity of data. A nonhierarchical document clustering algorithm was established by Ho et al. [6] for information retrieval based on a tolerance rough set model, which can capture more potential semantics information. Patra and Nandi developed a single-link clustering algorithm on the basis of tolerance rough set model to obtain a better clustering result [30]. In this paper, we adopt the tolerance rough set model via expressing each sentence as a pair of upper approximation and lower approximation to separately compute the upper approximation similarity and lower approximation similarity.

3. Proposed Method

In this section, we firstly describe the traditional tolerance rough set model briefly. Then, we introduce the probabilistic tolerance rough sets-based sentence similarity calculation model detailedly.

3.1. Tolerance Rough Set Theory. A tolerance space was defined as a quadruple $\mathbf{R} = (U, I, \nu, P)$ [6], where $U = \{x_1, x_2, \ldots, x_n\}$ is the universe of all the objects, $I(x)$ is an uncertainty function, I: $U \longrightarrow 2^U$, a set of tolerance classes, ν: $2^U \times 2^U \longrightarrow [0,1]$ is a vague inclusion, and P: $I(U) \longrightarrow \{0,1\}$ is a structural function. The uncertainty function I: $U \longrightarrow 2^U$ is defined as a tolerance class. If an object shares similar information with x, it is an element of $I(x)$. Any function satisfying reflexivity and symmetry can be defined as an uncertainty function $I(x)$, that is, for arbitrary $x, y \in U, x \in I(x)$ iff $x \in I(y)$. The vague inclusion ν is monotonous, i.e., for any $X, Y, Z \subseteq U$ and $Y \subseteq Z$, $\nu(X,Y) \leq \nu(X,Z)$. It measures the degree of inclusion of sets, whether a set X contains the tolerance class $I(x)$ of an object $x \in U$. The structural function P is defined as two classes—structural subsets $(P(I(x)) = 1)$ and nonstructural subsets $(P(I(x)) = 0)$—which are on functions of $I(x)$ for each $x \in U$ [6]. The upper approximation $\mathbf{U}(\mathbf{R}, X)$ and lower approximation $\mathbf{L}(\mathbf{R}, X)$ of any $X \subseteq U$ are defined as

$$\begin{aligned} \mathbf{U}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) > 0\}, \\ \mathbf{L}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) = 1\}. \end{aligned} \tag{1}$$

If the upper approximation and lower approximation are with parameters α and β, which are denoted as

$$\begin{aligned} \mathbf{U}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) > \alpha\}, \\ \mathbf{L}(\mathbf{R}, X) &= \{x \in U | P(I(x)) = 1 \,\&\, \nu(I(x), X) \geq \beta\}, \end{aligned} \tag{2}$$

where $\alpha \in [0,1)$, $\beta \in (0,1]$, $\alpha \leq \beta$, then it is called as the probabilistic tolerance rough set model [25].

3.2. Probabilistic Tolerance Rough Sets-Based Sentence Similarity Model. Firstly, we introduce the definition of the quadruple of tolerance rough sets in our model. Suppose that $W = \{w_1, w_2, \ldots, w_N\}$ is the set of all the words in the corpus, where N is the vocabulary size. Then, we define the universe as $U = W$. The determinations of tolerance relation and tolerance classes are the essential steps for formulating a tolerance rough set model. In the tolerance rough set model proposed by Ho et al. [6], the cooccurrence of terms in all the documents in the corpus was applied to construct the tolerance relation. However, it suffers from two disadvantages: (1) whenever the whole corpus gets some changes, even increasing or decreasing by only one document, all the procedures need to be recalculated; (2) the time complexity is relatively high. Hence, we choose the word similarity between words as the tolerance relation. Generally, the semantics similarity between two words is defined as the cosine similarity between the word vectors of the two words [31]. When the corpus increases or decreases one document, the number of cooccurrences of all the words will change and recalculation is needed, but the word similarity between words does not need to change. It provides the model employing the new tolerance relation to be incremental. According to the algorithm flow of the tolerance rough set model, the time complexity decreases from $O(N^3)$ to $O(N^2)$.

For a positive threshold θ, $0 \leq \theta \leq 1$, the uncertainty function I_θ of w_i is defined as follows:

$$I_\theta(w_i) = \{w_i\} \cup \{w_j | \mathrm{sim}(w_i, w_j) \geq \theta\}, \tag{3}$$

where $\mathrm{sim}(w_i, w_j)$ denotes the cosine similarity degree between the word w_i and w_j.

$$\mathrm{sim}(w_i, w_j) = \cos(w_i, w_j) = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\|\mathbf{w_i}\| \|\mathbf{w_i}\|}, \tag{4}$$

where $\mathbf{w_i}$ and $\mathbf{w_j}$ denote the word vectors of w_i and w_j, respectively. It is evident that the uncertainty function I_θ satisfies the condition of reflexive and symmetric.

Here, we give a counterexample to illustrate when I_θ does not satisfy the property of transitivity. Using the trained word2vec embeddings by Google [32], we can obtain similarity('beautiful', 'nice') = 0.5341, similarity('nice', 'pretty') = 0.5106, and similarity('beautiful', 'pretty') = 0.3299. Let the cosine similarity degree threshold $\theta = 0.5$; it is obvious that similarity('beautiful', 'nice') $> \theta$, similarity('nice', 'pretty') $> \theta$, and similarity('beautiful', 'pretty') $< \theta$. So, we conclude that the uncertainty function I_θ does not satisfy the condition of transitivity.

The vague inclusion function ν is defined the same as in [6]:

$$\nu(X, Y) = \frac{|X \cap Y|}{|X|}. \tag{5}$$

Let $S = \{S_1, S_2, \ldots, S_n\}$ be a collection of sentences, where S_i is represented by a group of words of the universe W, $1 \leq i \leq n$. Then, the fuzzy membership function μ for $w_j \in W, S_i \in S$ is expressed as

$$\mu(w_j, S_i) = \nu(I_\theta(w_j), S_i) = \frac{|I_\theta(w_j) \cap S_i|}{|I_\theta(w_j)|}. \tag{6}$$

Suppose that all the tolerance classes of words are structural subsets in the whole process, i.e., for any

$w_i \in W$, $P(I_\theta(w_i)) = 1$. Then, we define the upper approximation $\mathbf{U}(\mathbf{R}, S_i)$ and lower approximation $\mathbf{L}(\mathbf{R}, S_i)$ in R of any $S_i \in D$ as

$$\mathbf{U}(\mathbf{R}, S_i) = \left\{w_j \in W | \nu\left(\mathbf{I_\theta}\left(\mathbf{w_j}\right), \mathbf{S_i}\right) \geq \alpha\right\}, \tag{7}$$

$$\mathbf{L}(\mathbf{R}, S_i) = \left\{w_j \in W | \nu\left(\mathbf{I_\theta}\left(\mathbf{w_j}\right), \mathbf{S_i}\right) \geq \beta\right\}, \tag{8}$$

where $\alpha \in [0, 1)$, $\beta \in (0, 1]$, $\alpha < \beta$. $\mathbf{U}(\mathbf{R}, S_i)$ and $\mathbf{L}(\mathbf{R}, S_i)$ in R are also written as $\overline{S_i}$ and $\underline{S_i}$.

If S_i is regarded as one certain concept about the vague description of feature w_i; then $\mathbf{U}(\mathbf{R}, S_i)$ can be explained as a collection of concepts that share some semantics with S_i, and $\mathbf{L}(\mathbf{R}, S_i)$ can be explained as a collection of the core concepts of S_i. The probability values α and β can be used to adjust the accuracy of upper approximation and lower approximation.

Each sentence is denoted by two fuzzy sets on both upper approximation and lower approximation. Assume that one sentence S_1 is made up of a collection of words $\{w_1, w_2, \ldots, w_m\}$; then the upper approximation and lower approximation of S_1 are represented by

$$\overline{S_1} = \sum_{i=1}^{|\mathbf{U}(\mathbf{R}, S_1)|} \frac{\mu\left(w_j, S_1\right)}{w_j}, \tag{9}$$

$$\underline{S_1} = \sum_{j=1}^{|\mathbf{L}(\mathbf{R}, S_1)|} \frac{\mu\left(w_j, S_1\right)}{w_j}. \tag{10}$$

Considering the membership degree only, the upper approximation and lower approximation of S_1 can also be written as

$$\begin{aligned} \overline{S_1} &= \{u_{11}(w_1), u_{12}(w_2), \ldots, u_{1i}(w_k), \ldots, u_{1n}(w_n)\}, \\ \underline{S_1} &= \{l_{11}(w_1), l_{12}(w_2), \ldots, l_{1i}(w_i), \ldots, l_{1n}(w_n)\}, \end{aligned} \tag{11}$$

where $u_{11}(w_k)$ and $l_{11}(w_k)$ denote the membership degrees of w_k to $\overline{S_1}$ and $\underline{S_1}$, respectively. The upper approximation represents the expanded semantics of sentence S_1, capturing the latent semantics that S_1 contains. The similarity between two sentences can be measured by both the upper approximation similarity and lower approximation similarity of the two sentences. From the two different perspectives, both the expanded semantics similarity and the core semantics similarity can be captured sufficiently. For each sentence has been represented by two fuzzy sets, we employ two measurements to calculate the similarity between the two fuzzy sets, as defined as follows.

Measurement 1.

$$\mathrm{sim}_1\left(\overline{S_1}, \overline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(u_{1k}, u_{2k}\right)}{\sum_{k=1}^{N} \max\left(u_{1k}, u_{2k}\right)}, \tag{12}$$

$$\mathrm{sim}_1\left(\underline{S_1}, \underline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(l_{1k}, l_{2k}\right)}{\sum_{k=1}^{N} \max\left(l_{1k}, l_{2k}\right)}. \tag{13}$$

Measurement 2.

$$\mathrm{sim}_2\left(\overline{S_1}, \overline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(u_{1k}, u_{2k}\right)}{(1/2)\sum_{k=1}^{N}\left(u_{1k} + u_{2k}\right)}, \tag{14}$$

$$\mathrm{sim}_2\left(\underline{S_1}, \underline{S_2}\right) = \frac{\sum_{k=1}^{N} \min\left(l_{1k}, l_{2k}\right)}{(1/2)\sum_{k=1}^{N}\left(l_{1k} + l_{2k}\right)}. \tag{15}$$

On the basis of the upper approximations and lower approximations of the two sentences, except for representing each sentence by a pair of fuzzy sets, we propose another method to measure the similarity. Assume that the elements of the upper and lower approximation of sentence S_1 and sentence S_2 are

(i) $\mathbf{U}(\mathbf{R}, S_1) = \{w_1, w_2, \ldots, w_a\}$,
(ii) $\mathbf{L}(\mathbf{R}, S_1) = \{w_1, w_2, \ldots, w_b\}$,
(iii) $\mathbf{U}(\mathbf{R}, S_2) = \{w_1, w_2, \ldots, w_c\}$,
(iv) $\mathbf{L}(\mathbf{R}, S_2) = \{w_1, w_2, \ldots, w_d\}$, then a new similarity degree measurement is defined as follows.

Measurement 3. Consider

$$\mathrm{sim}_3\left(\overline{S_1}, \overline{S_2}\right) = \cos\left(\sum_{i=1}^{a} \mathbf{w_i}, \sum_{j=1}^{c} \mathbf{w_j}\right), \tag{16}$$

$$\mathrm{sim}_3\left(\underline{S_1}, \underline{S_2}\right) = \cos\left(\sum_{l=1}^{b} \mathbf{w_l}, \sum_{k=1}^{d} \mathbf{w_k}\right). \tag{17}$$

The lower similarity determines the degree to which two sentences are similar assuredly. Correspondingly, the upper similarity determines the degree to which two sentences are similar possibly. To measure the final similarity degree of the two sentences, we utilize the linear combination of the upper and lower approximation similarity, which is given as

$$\mathrm{sim}_i\left(S_1, S_2\right) = \lambda \cdot \mathrm{sim}_i\left(\overline{S_1}, \overline{S_2}\right) + (1 - \lambda) \cdot \mathrm{sim}_i\left(\underline{S_1}, \underline{S_2}\right), \tag{18}$$

where $i = 1, 2, 3$, λ is the linear coefficient. λ indicates the proportion of the upper approximation similarity degree and $(1 - \lambda)$ indicates the proportion of the lower approximation similarity degree. On account that the lower approximation is composed of the core semantics, the proportion of the lower approximation similarity degree is assigned a higher value than the upper approximation similarity degree. Generally, $0 \leq \lambda \leq 0.5$. (Algorithm 1)

Example 1. Here, we give an example of our proposed methods to calculate the sentence similarity. Assume that the corpus contained four sentences as follows:

(i) Three boys are jumping in the leaves.
(ii) Three kids are jumping in the leaves.

Algorithm 1: Probabilistic tolerance rough sets-based sentence similarity model
Input: A collection of sentences $S = \{S_1, S_2, \ldots, S_n\}$.
Parameters: The cosine similarity degree threshold: θ; the probabilistic value: α, β; the linear combination parameter: λ.
Output: The similarity degree between S_i and S_l,.

(1) Preprocess the sentence corpus $S = \{S_1, S_2, \ldots, S_n\}$, and generate the universe including all the distinct words of the corpus.
(2) Compute the uncertainty function $I_\theta(w_i)$ of each word in the universe according to equation (3).
(3) Suppose that the similarity degree between sentence S_i and S_l is to be calculated. Apply equation (6) to calculate the fuzzy membership degree $\mu(w_j, S_i)$ of each word in sentence S_i, $1 \le j \le N$, $1 \le i \le n$.
(4) Obtain the upper approximation $\mathbf{U}(\mathbf{R}, S_i)$ and lower approximation $\mathbf{L}(\mathbf{R}, S_i)$ of each sentence $S_i \in S$ according to equation (7) equation and (8). Similarly, acquire $\mathbf{U}(\mathbf{R}, S_l)$ and $\mathbf{L}(\mathbf{R}, S_l)$.
(5) Represent the upper approximation and lower approximation of S_i and S_l as fuzzy sets according to equation (9) and equation (10), which are written as $\overline{S_i}$, $\underline{S_i}$, $\overline{S_l}$ and $\underline{S_l}$.
(6) Calculate the upper approximation similarity $\mathrm{sim}(\overline{S_i}, \overline{S_l})$ between $\overline{S_i}$ and $\overline{S_l}$ and the lower approximation similarity $\mathrm{sim}(\underline{S_i}, \underline{S_l})$ between $\underline{S_i}$ and $\underline{S_l}$ according to equations (12)-(17) of the three measurements, respectively.
(7) Obtain the final sentence similarity degree $\mathrm{sim}(\overline{S_i}, \overline{S_l})$ between S_i and S_l utilizing the linear combination in equation (18).

Algorithm 1: The procedure of our proposed model in detail.

(iii) Three kids are sitting in the leaves.

(iv) Children in red shirts are playing in the leaves.

After preprocessing every sentence, 9 words are included in the corpus. Then, let the universe be the set of words $U = \{\text{boys, jumping, leaves, kids, sitting, children, red, shirts, playing}\}$. Then, we illustrate the proposed probabilistic tolerance rough sets-based sentence similarity model for computing the similarity degree of the following sentences:

(i) S_1: Three boys are jumping in the leaves.

(ii) S_2: Three kids are jumping in the leaves.

Here, we set the similarity degree threshold $\theta = 0.6$ and the probabilistic values $\alpha = 0$ and $\beta = 0.7$. Then, the upper and lower approximations of these two sentences are shown in Table 1.

The upper approximation similarity degrees and lower approximation similarity degrees by the proposed three measurements are listed in Table 2.

Let the linear combination coefficient $\lambda = 0.4$; then the final similarity degrees between S_1 and S_2 by three measurements are as follows:

(i) $\mathrm{sim}_1(S_1, S_2) = 0.967$,
(ii) $\mathrm{sim}_2(S_1, S_2) = 0.940$,
(iii) $\mathrm{sim}_3(S_1, S_2) = 0.987$.

It is apparent that our proposed probabilistic tolerance rough sets-based sentence similarity algorithm can reflect the similarity relation between sentences commendably. Firstly, from the sentences S_1 and S_2, it is evident that both of them express the core semantics of "jumping" and "leaves," just like the lower approximation obtained by our algorithm. Secondly, the lower approximation similarity degree is computed to be 1, which means that S_1 and S_2 share the same core meaning. Thirdly, from the upper approximation of S_2, it can be seen that the word "children" did not originally belong to S_2, but the meaning of "children" is mined through our method. The new meaning "children" comes from the tolerance class of the word "kid," so, in a sense, "children" is the explanation of "kids." Therefore, our proposed methods can capture some latent semantics behind texts from upper approximation, which can better distinguish whether two sentences are similar from a more general perspective. Analogously, our proposed algorithms can refine the core semantics of texts by the lower approximation, which can preferably analyze sentence similarity from a more accurate perspective.

Table 1: Approximations of each sentence.

	Upper approximation	Lower approximation
S_1	Boys, jumping, leaves	Jumping, leaves
S_2	Jumping, leaves, kids, children	Jumping, leaves

Table 2: Upper and lower approximation similarity degrees on three measurements.

	Upper approximation similarity	Lower approximation similarity
M_1	0.918	1.000
M_2	0.850	1.000
M_3	0.969	1.000

Example 2. We use the traditional tolerance rough set model [6] on Example 1 for comparison. The word cooccurrence degree is set as 2. Then, the upper and lower approximations can be seen in Table 3.

Table 4 displays the corresponding upper and lower approximation similarity degrees.

Then, the sentence similarity degrees of the three measurements are as follows:

(i) $\mathrm{sim}_1(S_1, S_2) = 0.517$,
(ii) $\mathrm{sim}_2(S_1, S_2) = 0.476$,
(iii) $\mathrm{sim}_3(S_1, S_2) = 0.799$.

From the results, we can see that it provides a worse performance in contrast with our methods.

Then, we discuss the condition that one sentence "Three boys are sitting in the leaves" is added to the corpus in

Table 3: Approximations of each sentence by traditional tolerance rough set model.

	Upper approximation	Lower approximation
S_1	Boys, jumping, leaves, kids	Boys, jumping
S_2	Jumping, leaves, kids	Jumping, leaves, kids

Table 4: Upper and lower approximation similarity degrees on three measurements by traditional tolerance rough set model.

	Upper approximation similarity	Lower approximation similarity
M_1	0.542	0.500
M_2	0.703	0.324
M_3	0.936	0.708

Example 1. The whole computational process and results by the probabilistic tolerance rough sets do not alter. However, the procedures have to repeat from the calculation of uncertainty function by the traditional tolerance rough sets; then the new upper and lower approximation of S_1 and S_2 are illustrated in Table 5. Thus, the applicability of the model [6] has been greatly reduced.

4. Experimental Results and Discussion

In this section, we take from SICK2014 task and STSbenchmark dataset to evaluate the performance of our methods.

4.1. Dataset and Preprocessing. SICK2014 [33] is a dataset for the similarity evaluation of sentence pairs, which contains the training set, trial set, and testing set for a total of 15000 sentence pairs. Since our proposed model is unsupervised, which do not require additional training on the dataset, we select the 5000 sentence pairs of the training set for the experiments. And each sentence pair has been assigned a similarity score from 0 to 5 by experts. Table 6 shows two examples in the SICK2014 dataset.

STS is the abbreviation for Semantic Textual Similarity. The SemEval STS datasets from 2012 [34] to 2017 [35] were selected for this dataset. Each sentence pair has been assigned a similarity score from 0 to 5 by experts. STS-train, STS-dev, STS-test, and MSRvid are chosen for the experiments.

For better comparison with our experimental results, we have normalized the similarity score. We take the word embedding trained by Google [23] as the word vector in the experiment.

4.2. Evaluation Metrics. We exploit the Pearson correlation coefficient (Pcc) [36] and mean square error (MSE) [37] to evaluate the performance of sentence similarity measurements.

Pcc is a linear correlation coefficient that reflects the linear correlation of two variables. As for two variants X and Y, the mathematical expression of Pcc is defined as

Table 5: New approximations of each sentence by traditional tolerance rough set model.

	Upper approximation	Lower approximation
S_1	Boys, jumping, leaves, kids, sitting	Boys, jumping
S_2	Boys, jumping, leaves, kids, sitting	Jumping, kids

Table 6: Example in dataset.

Sentence A	Sentence B	Relatedness score
A man is jumping into an empty pool	A man is jumping into a full pool	3
Two young girls are sitting on the ground	Two girls are sitting on the ground	4.4

$$\rho_{xy} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}}, \quad (19)$$

where $\mathrm{Cov}(X,Y)$ is the covariance of X and Y, $D(X)$ and $D(Y)$ denote the standard deviation of X and Y individually, and EX refers to the mathematical expectation of X. The greater the absolute value of Pcc, the stronger the correlation is.

MSE is a measure reflecting the degree of difference between estimator value and real value. The definition of MSE is

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y-\widehat{y})^2, \quad (20)$$

where N is the sample size, y is the real value, and $\widehat{y}$ is the estimator value. A smaller value of MSE demonstrates a smaller deviation between the estimator value and the real value.

4.3. Experimental Results and Analysis. We proposed three sentence similarity measurements based on the probabilistic tolerance rough set model. The performances on the SICK2014 dataset are displayed in Table 7. In the table, BERT-687 and BERT-1024 are two different BERT models for sentence representation, and the sentence similarity is calculated by the cosine similarity. Fuzzy rough is the model proposed in [12]. As can be seen in Table 7, on the whole, the three measures have much better performance than the other three models. Obviously, the results of Measurement 3 achieve at the optimal performance of Pcc as 0.725 and MSE as 0.033. Particularly for the value of MSE, it is evident that there is very small error between the sentence similarity degree calculated by our methods and the real value.

Tables 8 and 9 show the Pcc and MSE results on the STSbenchmark dataset. From the tables, we can see that all of the three measures have much better performance than the results by BERT on the four datasets of STSbenchmark. The reason is that more latent semantics behind sentences can be captured by our models. Therefore, the experimental results confirm the efficiency and applicability of our methods.

4.4. Cosine Similarity Degree Threshold. In our improved probabilistic tolerance rough set model, the cosine similarity degree threshold θ controls the accuracy of the uncertainty function. The higher the value of θ, the more precise the

Table 7: Experimental results of various measurements of sentence similarity on SICK2014.

	Pcc	MSE
BERT-687	0.611	0.104
BERT-1024	0.656	0.097
Fuzzy rough	0.609	0.863
M1	0.625	0.086
M2	0.572	0.137
M3	**0.725**	**0.033**

Table 8: PCC of various measurements of sentence similarity on STSbenchmark.

Dataset	Pcc				
	BERT-768	BERT-1024	M1	M2	M3
MSRvid	0.060	0.581	0.668	0.694	**0.819**
STS-train	0.514	0.597	0.628	0.655	**0.690**
STS-dev	0.569	0.620	0.655	0.637	**0.693**
STS-test	0.454	0.579	0.633	0.655	**0.714**

Table 9: MSE of various measurements of sentence similarity on STSbenchmark.

Dataset	MSE				
	BERT-768	BERT-1024	M1	M2	M3
MSRvid	0.349	0.327	0.020	**0.014**	0.019
STS-train	0.251	0.238	0.021	**0.015**	0.018
STS-dev	0.311	0.256	0.021	**0.015**	0.022
STS-test	0.279	0.283	0.024	**0.017**	0.021

uncertainty function. However, too high value of θ will result in inadequate semantics mining. Too small value of θ will lead to more redundancy and noisy information. The interaction of θ on Example 1 can be identified in Table 10.

Figures 1 and 2 reveal the interactions of different cosine similarity threshold value θ on Pcc and MSE, respectively. In this experiment, α is set as 0, β is set as 0.6, and λ is set as 0.3. θ ranges from 0.5 to 1. As shown in Figure 1, the value of Pcc increases from $\theta = 0.5$, achieves the peak at $\theta = 0.9$, then decreases. Similarly, the value of MSE decreases from $\theta = 0.55$, achieves the minimum at $\theta = 0.95$, then increases. We can conclude that the interaction of θ satisfies the regular analyzed above.

4.5. Probability Value. The values of α and β are used for adjusting the precision of upper approximation and lower approximation. α determines the range of upper approximation. The smaller the value of $\alpha$is, the more elements the upper approximate set has. In the traditional tolerance rough sets, α is 0, by which the upper approximation contains the most information. It can reduce the generation of redundant information and does not lose too much potential semantics information via adjusting the value of α. The influence of α on Example 1 can be observed in Table 11. Similarly, β determines the range of lower approximation. A larger value of β leads to fewer elements of the lower approximate set. When $\beta = 1$, the fewest elements are included in the lower approximation, which may cause the loss of some core semantics information. Adjusting β properly may better and more adequately mine core semantics information. The effect of β on Example 1 can be observed in Table 12.

Table 10: Approximations of each sentence on different θ.

	θ	Upper approximation	Lower approximation
S_1	0.5	Boys, jumping, leaves, kids, children	Jumping, leaves
S_2	0.5	Boys, jumping, leaves, kids, children	Jumping, leaves
S_1	0.7	Boys, jumping, leaves	Jumping, leaves
S_2	0.7	Jumping, leaves, kids, children	Jumping, leaves

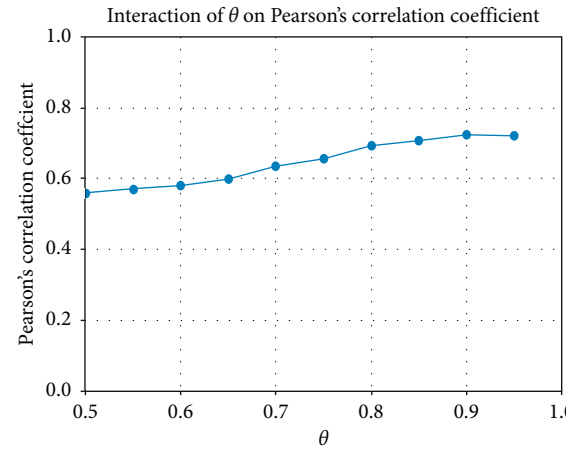


Figure 1: Interaction of cosine similarity threshold value θ on Pearson's correlation coefficient.

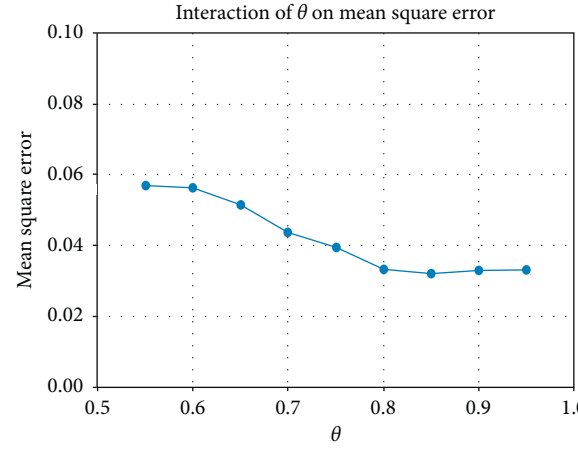


Figure 2: Interaction of cosine similarity threshold value θ on mean square error.

Table 11: Approximations of each sentence on different α.

	α	Upper approximation	Lower approximation
S_1	0	Boys, jumping, leaves, kids, children	Jumping, leaves
S_2	0	Boys, jumping, leaves, kids, children	Jumping, leaves
S_1	0.3	Boys, jumping, leaves	Jumping, leaves
S_2	0.3	Jumping, leaves, kids, children	Jumping, leaves
S_1	0.5	Jumping, leaves	Jumping, leaves
S_2	0.5	Jumping, leaves	Jumping, leaves

Table 12: Approximations of each sentence on different β.

	β	Upper approximation	Lower approximation
S_1	0.3	Boys, jumping, leaves, kids, children	Boys, jumping, leaves, kids, children
S_2	0.3	Boys, jumping, leaves, kids, children	Boys, jumping, leaves, kids, children
S_1	1	Boys, jumping, leaves, kids, children	Jumping, leaves
S_2	1	Boys, jumping, leaves, kids, children	Jumping, leaves

5. Conclusion

In this paper, owing to the property of uncertainty of text data, we incorporate the probabilistic tolerance rough sets to establish a novel sentence similarity computation model. For the reason that the traditional tolerance rough set model is not incremental and has high complexity, we make some improvement to it, making the model becoming incremental and reducing the time complexity. Through introducing the probability values α and β, the accuracy of the upper approximation and lower approximation can be adjusted. The upper approximation and lower approximation are served to represent every sentence. And on this basis, three sentence similarity calculation measurements are proposed. Upper approximation similarity and lower approximation similarity are individually calculated of each sentence pair. The linear combination of the upper approximation similarity and the lower approximation similarity is used to indicate the total sentence similarity. On the one hand, it can dig out more latent semantics information than the traditional methods based on shallow semantics. On the other hand, it is unsupervised, which relieves the defect of supervised deep learning-based methods. We carry out some experiments on the SICK2014 task to evaluate the performance of our proposed model. The results verify the efficiency and applicability of the proposed models.

The proposed model is established without considering the order of sentences, in which our future work will include it.

Data Availability

The SICK2014 task data used to support the findings of this study are available from clic.cimec.unitn.it/composes/sick.html.

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 the National Natural Science Foundation of China (Grant nos. 11671001 and 61876201).

References

[1] N. Chatterjee and N. Yadav, "Fuzzy rough set-based sentence similarity measure and its application to text summarization," *IETE Technical Review*, vol. 36, no. 5, pp. 517–525, 2019.

[2] S. Abujar, M. Hasan, and S. A. Hossain, "Sentence similarity estimation for text summarization using deep learning," in *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*, pp. 155–164, Springer, Comilla, Bangladesh, October 2019.

[3] K. Wu, X. Wang, and A. Aw, "Bilingual word embedding with sentence similarity constraint for machine translation," in *Proceedings of the 2017 International Conference on Asian Language Processing (IALP)*, pp. 119–122, IEEE, Singapore, December 2017.

[4] W. An, Q. Chen, W. Tao et al., "ECNU at 2017 LiveQA track: learning question similarity with adapted long short-term memory networks," in *Proceedings of the Twenty-Sixth Text REtrieval Conference*, pp. 1–9, TREC, Gaithersburg, MD, USA, 2017.

[5] Y. Wang, Q. Hu, Y. Song, and L. He, "Potentiality of healthcare big data: improving search by automatic query reformulation," in *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*, pp. 807–816, Morgan Kaufmann, San Mateo, CA, USA, December 2017.

[6] T. B. Ho and N. B. Nguyen, "Nonhierarchical document clustering based on a tolerance rough set model," *International Journal of Intelligent Systems*, vol. 17, no. 2, pp. 199–212, 2002.

[7] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 60, no. 5, pp. 493–502, 2004.

[8] L. Han, T. Finin, P. Mcnamee, A. Joshi, and Y. Yesha, "Improving word similarity by augmenting PMI with estimates of word polysemy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1307–1322, 2013.

[9] Y. Bi, K. Deng, and J. X. Cheng, *A Keyword-Based Method For Measuring Sentence Similarity*, pp. 379-380, ACM, New York, NY, USA, 2017.

[10] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

[11] P. D. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the European Conference on Machine Learning*, pp. 491–502, Springer, Freiburg, Germany, September 2001.

[12] M. K. Prasad and P. Sharma, "Combining common words and semantic features for sentence similarity," in *Proceedings of the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–4, IEEE, Bangalore, India, July 2018.

[13] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Advances in Neural Information Processing Systems*, pp. 2042–2050, 2014.

[14] H. He, K. Gimpel, and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1576–1586, Lisbon, Portugal, September 2015.

[15] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2786–2792, Phoenix, AZ, USA, February 2016.

[16] Z. Wang, H. Mi, and A. Ittycheriah, "Sentence similarity learning by lexical decomposition and composition," 2016, https://arxiv.org/abs/1602.07019.

[17] Q. Chen, Q. Hu, J. X. Huang, and L. He, "CA-RNN: Using context-aligned recurrent neural networks for modeling sentence similarity," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 265–273, New Orleans, LA, USA, February 2018.

[18] P. Liu, Z. Zheng, and Q. Su, "Sentence similarity computation by integrating shallow and deep information," in *Proceedings of the 2018 International Conference on Asian Language Processing (IALP)*, pp. 308–311, IEEE, Bandung, Indonesia, November 2018.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2019, https://arxiv.org/abs/1810.04805.

[20] H.-H. Huang and Y.-H. Kuo, "Cross-lingual document representation and semantic similarity measure: a fuzzy set and rough set based approach," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 6, pp. 1098–1111, 2010.

[21] Z. a. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[22] C. Luo, T. Li, H. Chen, H. Fujita, and Z. Yi, "Incremental rough set approach for hierarchical multicriteria classification," *Information Sciences*, vol. 429, pp. 72–87, 2018.

[23] R. K. Nowicki, M. Korytkowski, and R. Scherer, "Rough neural network ensemble for interval data classification," in *Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–7, IEEE, Rio de Janeiro, Brazil, July 2018.

[24] Y. Ma, X. Luo, X. Li, Z. Bao, and Y. Zhang, "Selection of rich model steganalysis features based on decision rough set α-positive region reduction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 336–350, 2018.

[25] Y. Yao, "Probabilistic rough set approximations," *International Journal of Approximate Reasoning*, vol. 49, no. 2, pp. 255–271, 2008.

[26] Z. Pawlak and R. Sɵwinski, "Rough set approach to multi-attribute decision analysis," *European Journal of Operational Research*, vol. 72, no. 3, pp. 443–459, 1994.

[27] A. Skowron and J. Stepaniuk, "Tolerance approximation spaces," *Fundamenta Informaticae*, vol. 27, no. 2, 3, pp. 245–253, 1996.

[28] C. L. Ngo and H. S. Nguyen, "A method of web search result clustering based on rough sets," in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 673–679, IEEE, 2005.

[29] X.-J. Meng, Q.-C. Chen, and X.-L. Wang, "A tolerance rough set based semantic clustering method for web search results," *Information Technology Journal*, vol. 8, no. 4, pp. 453–464, 2009.

[30] B. K. Patra and S. Nandi, "Fast single-link clustering method based on tolerance rough set model," in *Proceedings of the International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pp. 414–422, Springer, Delhi, India, December 2009.

[31] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 794–804, 2017.

[32] http://code.google.com/archive/p/word2vec.

[33] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, pp. 1–8, August 2014.

[34] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: a pilot on semantic textual similarity," in *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012)*, p. 385C393, Association for Computational Linguistics, Montréal, Canada, June 2012.

[35] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, September 2017, Article ID 670C680.

[36] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp. 1–4, Springer, Berlin, Germany, 2009.

[37] J. M. Lowerre, "On the mean square error of parameter estimates for some biased estimators," *Technometrics*, vol. 16, no. 3, pp. 461–464, 1974.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6432929, 14 pages
https://doi.org/10.1155/2021/6432929

Research Article

Dataset Denoising Based on Manifold Assumption

Zhonghua Hao,[1,2] **Shiwei Ma**,[3] **Hui Chen**,[4] **and Jingjing Liu**[5]

[1]*Qingdao University, College of Automation, Qingdao 266071, China*
[2]*Qingdao University, College of Electrical Engineering, Qingdao 266071, China*
[3]*School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, 200444, China*
[4]*College of Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China*
[5]*State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 201210, China*

Correspondence should be addressed to Shiwei Ma; masw@shu.edu.cn

Received 25 May 2020; Revised 1 November 2020; Accepted 22 December 2020; Published 18 January 2021

Academic Editor: Jun Shen

Learning the knowledge hidden in the manifold-geometric distribution of the dataset is essential for many machine learning algorithms. However, geometric distribution is usually corrupted by noise, especially in the high-dimensional dataset. In this paper, we propose a denoising method to capture the "true" geometric structure of a high-dimensional nonrigid point cloud dataset by a variational approach. Firstly, we improve the Tikhonov model by adding a local structure term to make variational diffusion on the tangent space of the manifold. Then, we define the discrete Laplacian operator by graph theory and get an optimal solution by the Euler–Lagrange equation. Experiments show that our method could remove noise effectively on both synthetic scatter point cloud dataset and real image dataset. Furthermore, as a preprocessing step, our method could improve the robustness of manifold learning and increase the accuracy rate in the classification problem.

1. Introduction

Since objects vary gradually in the real world, the manifold assumption indicates that the data points depict the state of an object should distribute on a smooth low-dimensional manifold embedded in high-dimensional observation space [1]. Dimensionalities of the manifold are key factors that control variation of the object state. For example, in Figure 1, the images of the rotational duck toy distribute on a one-dimensional manifold (a curve) embedded in high-dimensional pixel space. Each image depicts a particular state of the duck. Although the pixel values change dramatically at these images, humans could discover easily that they are controlled by one key factor: rotation of the duck.

Learning the knowledge hidden in the manifold-geometric distribution of a high-dimensional dataset is essential in many machine learning algorithms. For example, manifold learning algorithms aim to discover the nonlinear geometric structure dataset by preserving different local geometric properties [3–8]. The embedding results can be further used in data visualization, motion analysis, and classification [9, 10]. Moreover, much research takes manifold assumption as a constraint condition in its objective function [11, 12]. It is worth noting that manifold assumption is applied to explain why deep learning works well recently [13–15]. This research indicates deep learning could capture the manifold structure of one kind of knowledge by powerful nonlinear mapping.

However, noise is inevitable in data acquisition. For example, in Figure 1, the noiseless images of the rotational duck toy (red points) should lie on a curve embedded in the pixel space. However, due to the long exposure time and camera shake, the duck becomes "brighten" and "small" in the image. The corresponding noise data point, which is marked by "N" and green color in Figure 1, does not lie on the curve because pixel values change dramatically in the noise image.

Noise makes machine learning models fragile and hard to train. For example, the outlier points are difficult to handle in the classification and clustering task. Machine learning model needs to become more complex to get proper results [13]. In manifold learning algorithms, noise points make recovered embeddings difficult to capture the true

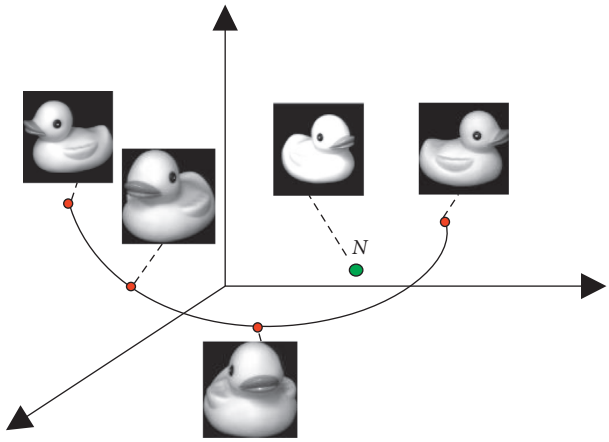


Figure 1: Image dataset of the rotational duck toy distributes on a one-dimensional manifold [2]. The red points correspond to noiseless images. The green point correspond to the noise image.

manifold-geometric distribution of the dataset. The reason is that the "short circuit" phenomenon arises easily in the noise dataset which destroys the local linear structure of the manifold [16].

In this paper, we propose a novel denoising method based on manifold assumption. Our aim is to obtain the data points that lie on the noiseless manifold through noise data points. Compared with the existing denoising methods, our method has two contributions worth being highlighted:

(1) Our method makes use of manifold-geometric distribution information of the dataset. Therefore, this method works for a dataset rather than a single data point.

(2) Our method improves the Tikhonov model to make the variational diffusion on the tangent space of the manifold for a high-dimensional nonrigid point cloud dataset.

Our method could capture the "true" geometric structure of the noise dataset. After denoising, the key factors that control the geometric distribution of the dataset are maintained and the characteristics of individual points are removed as noise. As a preprocessing step, our method could improve the robustness of manifold learning and increase the accuracy rate in the classification problem.

The rest of the paper is organized as follows: a brief review of the research on the manifold assumption is outlined in Section 1. Section 2 describes the motivation and details of the proposed method. In Section 3, experiments are conducted on both synthetic and real data to evaluate our method. Section 4 concludes remarks and a discussion of future work.

2. Related Work

Existing denoising methods always work for the noise in a single data point, such as "Gaussian noise" or "pepper noise" [17, 18] in an image. However, these methods could not deal with the noise that distorts the geometric distribution of the dataset, such as the noise duck toy image (green point) caused by longer exposure time and camera shake in Figure 1.

Only a few studies exist to deal with this problem. Gong et al. [19] proposed a local linear denoising method. This method removed noise by projecting noise data points to the tangent space of manifold which is estimated by the principal component analysis method firstly. Then, local denoised patches are aligned to get the global denoising dataset. However, the principal components may be distorted because they are calculated by the neighborhood of noise data points, which could lead to a wrong denoising result. Hao et al. [16] also utilized principal component analysis and projection method to find the noiseless data points. Therefore, it has the same problem. Moreover, many machine learning methods proposed the noise-resistant model for outliers but did not discuss denoising as an independent problem [7, 20]. For example, Zhang et al. [7] proposed an adaptive neighborhood selection method by the shrink and expand strategy to resist noise on the neighborhood of manifold.

In this paper, we propose a denoising method for the dataset. This method improves the Tikhonov method by adding a local structure term. The optimal solution is obtained by minimizing the objective function through a variational diffusion approach.

3. Proposed Approach

Let $\mathbf{F} = \{\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(\mathbf{m})\}$ be the noise dataset. $\mathbf{f}(\mathbf{x}) \in R^D$ is the x-th data point in F. D is the dimension number of $\mathbf{f}(\mathbf{x})$. Let $\mathbf{U} = \{\mathbf{u}(1), \mathbf{u}(2), \ldots, \mathbf{u}(\mathbf{m})\}$ be the noiseless dataset we want to obtain. $\mathbf{u}(\mathbf{x}) \in R^D$ is the x-th data point in U. $\mathbf{f}(\mathbf{x}) = \mathbf{u}(\mathbf{x}) + \xi(\mathbf{x}), \xi(\mathbf{x}) \in R^D$ is the noise of $\mathbf{f}(\mathbf{x})$. The goal is to recover U from F.

We illustrate our method in three steps: firstly, introduce to inspiration and motivation; then, construct the objective function by improving the Tikhonov model; and finally, optimize the objective function and get the solution by taking discrete operators.

3.1. Inspiration and Motivation. Manifold assumption claims that the noiseless data point $\mathbf{u}(\mathbf{x})$ that depicts the object state (the blue points in Figure 2) should lie on a smooth manifold $\mathcal{U}$ (blue surface in Figure 2) embedded in observation space. However, noise points $\mathbf{f}(\mathbf{x})$ (red points) distribute on the noise manifold $\mathcal{F}$. The denoising problem is how to obtain ux on $\mathcal{U}$from $\mathbf{f}(\mathbf{x})$ on $\mathcal{F}$.

3.2. Objective Function. The objective function is formulated in this part. Firstly, we illustrate the Tikhonov model briefly in image denoising which is similar to our problem. Then, the challenge of our problem is shown. Finally, we improve the Tikhonov model and construct the objective function for our problem.

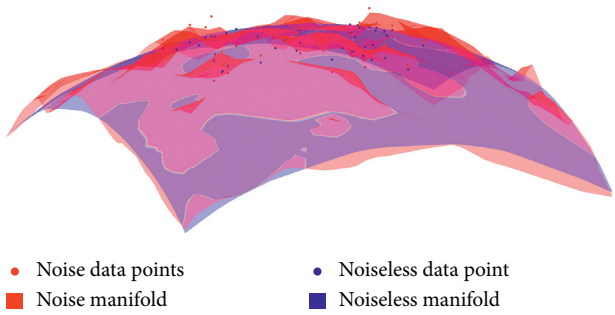


Figure 2: Illustration of the idea of our method: obtain the noiseless blue points that lie on smooth manifold (blue surface) from the noise red points that distribute on an irregular surface (noise manifold).

3.2.1. Tikhonov Model in the Image Denoising Problem. Our problem is similar to the image denoising problem $f(x, y) = u(x, y) + \xi(x, y)$, where x and y are row and column numbers of a pixel in an image. $f(x, y)$ and $u(x, y)$ are pixel values at row $x$and column y in noise and noiseless image, respectively. $\xi(x, y)$ is the noise. In Figure 2, if we regard the x, y, and z coordinate of $\mathbf{f}(\mathbf{x})$ as row number, column number, and pixel value, then the red manifold F depicts the pattern of noise image. Therefore, the image denoising problem is to find a noiseless image $\mathcal{U}$ from $\mathcal{F}$.

The Tikhonov model is one of the most classical variational models to deal with this problem [21]:

$$E(u) = \min_u \frac{1}{2}\int_{|\Omega} (u - f)^2 \mathrm{d}x + \frac{\alpha}{2}\int_{|\Omega} |\nabla \mathbf{u}|^2 \mathrm{d}x, \quad (1)$$

where Ω is the image domain and dx is the area element (pixel) in Ω. ∇u is the gradient of $u(x)$. The first term $\int_\Omega (u - f)^2$ is "data term" that measures the Euclidean distance between $\mathcal{F}$ and $\mathcal{U}$. The second term $\int_\Omega |\nabla \mathbf{u}|^2 \mathrm{d}x$ is "smooth term" that measures the noise strength of $\mathcal{U}$. Since these two terms have opposite effect, the parameter α balances these two terms. If α is small, $\mathcal{U}$ is close to $\mathcal{F}$ but the noise strength is large. On the other hand, the noise becomes small but the image pattern of $\mathcal{U}$ is "unlike" $\mathcal{F}$.

3.2.2. The Challenge of Our Problem. In the image denoising problem, the gradient operator is defined as [21]

$$\nabla \mathbf{u} = [u(x, y) - u(x - 1, y), u(x, y) - u(x, y - 1)]^T. \quad (2)$$

When minimizing the "smooth term" $\int_\Omega |\nabla \mathbf{u}|^2 \mathrm{d}x$ in (1), the pixel values in the image became the same, whereas the image area does not change since x and y are fixed.

However, in our problem, the dataset is nonrigid and high-dimensional cloud points. Let $\mathbf{u}(\mathbf{x}) = [u(x)^1, u(x)^2, \ldots, u(x)^D] \in R^D$ be a data point. D is the dimension number of $\mathbf{u}(\mathbf{x})$. Suppose $\mathcal{N}_{u(x)}$ is the neighborhood of $\mathbf{u}(\mathbf{x})$ which is determined by the KNN method:

$$\mathcal{N}_{u(x)} = \{\mathbf{u}(y_i) \in \mathcal{N}_{u(x)}\}, \quad i = 1, \ldots, k. \quad (3)$$

Naturally, the gradient operator is defined as

$$\nabla \mathbf{u} = [\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{y}_1), \mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{y}_2), \ldots, \mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{y}_\mathbf{k})]^T. \quad (4)$$

Therefore, the "smooth term" in (1) is

$$\int_\Omega |\nabla \mathbf{u}|^2 = \int_\Omega \int_{\mathcal{N}_{u(x)}} (u(x) - u(y_i))^2 \mathrm{d}y\, \mathrm{d}x. \quad (5)$$

When minimizing an objective function, the "smooth term" makes $\mathbf{u}(\mathbf{x})$ and $\mathbf{u}(\mathbf{y_i})$ become the same point. Therefore, the "cluster" phenomenon arises in the dataset—some points are brought close together and the other points are pushed away. Therefore, the geometric structure of the manifold $\mathcal{U}$ (blue surface in Figure 2) will shrink to a few point clusters rather than becoming smooth. Therefore, the Tikhonov model could not be applied directly to solve our problem.

3.2.3. Our Objective Function. To deal with this problem, we maintain the geometric distribution of $\mathcal{U}$ by keeping the tangent linear structure when minimizing the objective function. Since the neighborhood of the manifold could be regarded as tangent space (the blue plane in Figure 3), we make the neighborhood structure of $\mathcal{U}$ the same as $\mathcal{F}$.

The weight of local linear representation is utilized to depict the geometric structure of the neighborhood. The weight $\mathbf{W_f}$ of data point $\mathbf{f}(\mathbf{x})$ is defined as

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{k} W_{fi} \mathbf{f}(\mathbf{y_i}), \quad (6)$$

where $\mathbf{f}(\mathbf{y_i}) \in \mathcal{N}_{f(x)}$ and W_{fi} is the i-th component of $\mathbf{W_f}$ between $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}(\mathbf{y_i})$. Similarly, the linear representation weight of $\mathbf{u}(\mathbf{x})$ is defined as $\mathbf{W_u}$.

The local linear structure can be maintained if we set $\mathbf{W_u}$ the same as $\mathbf{W_f}$. Then, $\mathbf{f}(\mathbf{x})$ could only move along the normal space of manifold when minimizing the "smooth term" in the objective function because the tangent geometric structure is fixed by $\mathbf{W_u}$. Therefore, we add a "local structure term" in the Tikhonov model:

$$\int_\Omega (\mathbf{u}(\mathbf{x}) - \int_{\mathcal{N}_{u(x)}} W_{fi} \mathbf{f}(\mathbf{y_i}) \mathbf{dy})^2 \mathbf{dx}, \quad (7)$$

where $\int_{\mathcal{N}_{u(x)}} W_{fi} \mathbf{f}(\mathbf{y_i}) \mathbf{dy}$ is the linear reconstruction of $\mathbf{u}(\mathbf{x})$. Thus, our objective function is

$$E(\mathbf{u}) = \min \frac{1}{2}\int_\Omega (\mathbf{u} - \mathbf{f})^2 \mathbf{dx} + \frac{\alpha}{2}\int_\Omega |\nabla \mathbf{u}|^2 \mathbf{dx} + \frac{\beta}{2}\int_\Omega (\mathbf{u} - \mathbf{W}\mathbf{f}_{\mathcal{N}})^2 \mathbf{dx}, \quad (8)$$

where α and β are balance parameters.

3.3. Optimal Solution. In this part, we get optimal $\mathbf{u}$ by minimizing objective function (8). The solution in the continuous form is calculated firstly. Then, the discrete operator is defined and plugged to get a discrete solution.

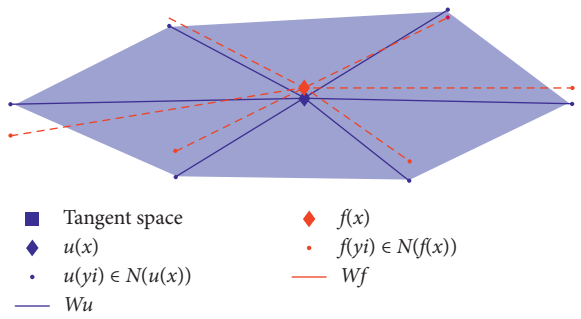


FIGURE 3: Local weights in the neighborhood. $\mathbf{W_u}$ is the weight of the linear representation of $\mathbf{u}(\mathbf{x})$ by $\mathbf{u}(\mathbf{y_i})$ that are in the neighborhood of $\mathbf{u}(\mathbf{x})$. $\mathbf{W_f}$ is the weight of the linear representation of $\mathbf{f}(\mathbf{x})$ by $\mathbf{f}(\mathbf{y_i})$ that are in the neighborhood of $\mathbf{f}(\mathbf{x})$.

3.3.1. Solution in Continuous Form. To get optimal $\mathbf{u}$, we calculate the derivative of (8) with respect to $\mathbf{u}$ by variational approach and set it to zero:

$$\begin{aligned} E'(\varepsilon) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} E(\mathbf{u}+\varepsilon\boldsymbol{\eta}, \nabla\mathbf{u}+\varepsilon\nabla\mathbf{u}) \\ &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(\frac{1}{2}\int_\Omega (\mathbf{u}+\varepsilon\boldsymbol{\eta}-\mathbf{f})^2\mathbf{dx} + \frac{\alpha}{2}\int_\Omega |\nabla\mathbf{u}+\varepsilon\nabla\boldsymbol{\eta}|^2\mathbf{dx} + \frac{\beta}{2}\int_\Omega (\mathbf{u}+\varepsilon\boldsymbol{\eta}-\mathbf{W}\mathbf{f}_{\mathcal{N}})^2\mathbf{dx}\right) \\ &= \int_\Omega \boldsymbol{\eta}(\mathbf{u}-\mathbf{f})\mathbf{dx} + \alpha\int_\Omega \nabla\boldsymbol{\eta}\nabla\mathbf{u}\,\mathbf{dx} + \beta\int_\Omega \boldsymbol{\eta}(\mathbf{u}-\mathbf{W}\mathbf{f}_{\mathcal{N}})\mathbf{dx} \\ &= \int_\Omega \boldsymbol{\eta}(\mathbf{u}-\mathbf{f})\mathbf{dx} + \alpha\int_{\partial\Omega} \vec{n}\boldsymbol{\eta}\nabla\mathbf{u}\,\mathbf{ds} - \alpha\int_\Omega \boldsymbol{\eta}\Delta\mathbf{u}\,\mathbf{dx} + \beta\int_\Omega \boldsymbol{\eta}(\mathbf{u}-\mathbf{W}\mathbf{f}_{\mathcal{N}})\mathbf{dx} \\ &= \eta\int_\Omega [\,(\mathbf{u}-\mathbf{f}) - \alpha\Delta\mathbf{u} + \beta(\mathbf{u}-\mathbf{W}\mathbf{f}_{\mathcal{N}})\,]\mathbf{dx} + \alpha\int_{\partial\Omega} \vec{n}\boldsymbol{\eta}\nabla\mathbf{u}\,\mathbf{ds}. \end{aligned} \tag{9}$$

Therefore, the Euler–Lagrange equation of $\mathbf{u}$ is

$$(\mathbf{u}-\mathbf{f}) - \alpha\Delta\mathbf{u} + \beta(\mathbf{u}-\mathbf{W}\mathbf{f}_{\mathcal{N}}) = 0. \tag{10}$$

Then,

$$\mathbf{u} = \frac{\mathbf{f}+\beta\mathbf{W}\mathbf{f}_{\mathcal{N}}+\alpha\Delta\mathbf{u}}{1+\beta}. \tag{11}$$

And the boundary condition is

$$\vec{n}\nabla\mathbf{u} = 0. \tag{12}$$

3.3.2. Solution in Discrete Form. To get the discrete solution, we define the discrete Laplacian operator in (11) by spectral graph theory [22]. Firstly, the gradient of $\mathbf{u}(\mathbf{x})$ is defined as

$$\nabla_{\mathbf{wG}}\mathbf{u}(\mathbf{x},\mathbf{y}) = \{(\mathbf{u}(\mathbf{y_i}) - \mathbf{u}(\mathbf{x}))\mathbf{W_d}(\mathbf{x},\mathbf{y})\}_{\mathbf{u}(\mathbf{y_i})\in\mathcal{N}_{u(x)}}, \quad i = 1,\ldots,k. \tag{13}$$

This gradient is a k-dimensional vector because there are k data points in $\mathcal{N}_{u(x)}$. The subscript "**wG**" is abbreviated to "weighted graph." $\mathbf{W_d}(\mathbf{x},\mathbf{y})$ is a weight vector. The component $\mathbf{W_d}(\mathbf{x},\mathbf{y_i})$ should be important if $\mathbf{u}(\mathbf{x})$ and $\mathbf{u}(\mathbf{y_i})$ are near. On the contrary, the component should be unimportant if $\mathbf{u}(\mathbf{x})$ and $\mathbf{u}(\mathbf{y_i})$ are far away. Therefore, we define $\mathbf{W_d}(\mathbf{x},\mathbf{y})$ as

$$\mathbf{W_d}(\mathbf{x},\mathbf{y}) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\mathbf{d}(\mathbf{x},\mathbf{y})}{2\sigma^2}\right), \tag{14}$$

where $\mathbf{d}(\mathbf{x},\mathbf{y})$ is the vector of Euclidean distance between $\mathbf{u}(\mathbf{x})$ and $\mathbf{u}(\mathbf{y_i}) \in \mathcal{N}_{u(x)}$. σ is the variance of $\mathbf{d}(\mathbf{x},\mathbf{y})$. For the convenience of calculations, we set $\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})} = \mathbf{W_d}(\mathbf{x},\mathbf{y})$. Therefore, the discrete gradient of $\mathbf{u}(\mathbf{x})$ is

$$\nabla_{\mathbf{wG}}\mathbf{u}(\mathbf{x},\mathbf{y}) = \left\{(\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{x}))\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\right\}_{\mathbf{u}(\mathbf{y})\in\mathcal{N}_{u(x)}}. \tag{15}$$

Consequently, the gradient of a vector $\mathbf{v}(\mathbf{x},\mathbf{y})$ is (the derivation procedure is listed at "Notice" at the end of this capture):

$$\nabla_{\mathbf{wG}}\mathbf{v} = -\sum_{\mathbf{y}\in\mathcal{N}(x)} (\mathbf{v}(\mathbf{y},\mathbf{x}) - \mathbf{v}(\mathbf{x},\mathbf{y}))\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}. \tag{16}$$

Let $\mathbf{v}(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{wG}}\mathbf{u}(\mathbf{x},\mathbf{y}) = (\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{x}))\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}$, therefore, the discrete Laplace operator of $\mathbf{u}(\mathbf{x})$ can be defined by

$$\begin{aligned}\Delta_{\mathbf{wG}}\mathbf{u} = \nabla_{\mathbf{wG}}\left(\nabla_{\mathbf{wG}}\mathbf{u}\right) &= \sum_{y\in\mathcal{N}(x)}\left(\left(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x})\right)\right.\\ &\left.\cdot\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}-\left(\mathbf{u}(\mathbf{x})-\mathbf{u}(\mathbf{y})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\right)\\ &= 2\sum_{y\in\mathcal{N}(x)}\left(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x})\right)\mathbf{w}(\mathbf{x},\mathbf{y}).\end{aligned} \tag{17}$$

We plug the discrete Laplace operator into (11). The solution of our object energy function (8) is

$$\mathbf{u}(\mathbf{x})^{k+1} = \frac{\mathbf{f}(\mathbf{x})+\beta\sum_{y\in\mathcal{N}(x)}\mathbf{W}(\mathbf{x},\mathbf{y})\mathbf{f}(\mathbf{y})+2\alpha\sum_{y\in\mathcal{N}(x)}\left(\mathbf{u}(\mathbf{y})^{k}-\mathbf{u}(\mathbf{x})^{k}\right)\mathbf{w}(\mathbf{x},\mathbf{y})}{1+\beta}, \tag{18}$$

where the superscripts k and k+1 are the iteration step. The initial value of $\mathbf{u}$ is set to $\mathbf{f}$. The optimal $\mathbf{u}$ is obtained by iteration, which ends up when $E(\mathbf{u}) < \varepsilon$, where $E(\mathbf{u})$ is the objective function value and ε is a small error we set. The boundary condition (12) could be ignored because the dataset is scattered and nonrigid cloud points.

Notice:

The gradient of a vector $\mathbf{v}$ could be derived as follows:

$$\begin{aligned}\sum_{x\in\Omega}\nabla_{\mathbf{wG}}\mathbf{u}\cdot\mathbf{v} &= \sum_{x\in\Omega}\sum_{y\in\mathcal{N}(x)}\left(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\mathbf{v}(\mathbf{x},\mathbf{y})\\ &= \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\mathcal{N}(x)}\left(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\mathbf{v}(\mathbf{x},\mathbf{y})\\ &\quad+\sum_{x\in\Omega}\sum_{y\in\mathcal{N}(x)}\left(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\mathbf{v}(\mathbf{x},\mathbf{y})\\ &= \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\mathcal{N}(x)}\mathbf{u}(\mathbf{x})\left(\mathbf{v}(\mathbf{y},\mathbf{x})-\mathbf{v}(\mathbf{x},\mathbf{y})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\\ &\quad+\frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\mathcal{N}(x)}\mathbf{u}(\mathbf{y})\left(\mathbf{v}(\mathbf{x},\mathbf{y})-\mathbf{v}(\mathbf{y},\mathbf{x})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\\ &= \sum_{x\in\Omega}\sum_{y\in\mathcal{N}(x)}\left(\mathbf{v}(\mathbf{y},\mathbf{x})-\mathbf{v}(\mathbf{x},\mathbf{y})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}\mathbf{u}(\mathbf{x})\\ &= -\sum_{x\in\Omega}\nabla_{\mathbf{wG}}\mathbf{v}\cdot\mathbf{u}.\end{aligned} \tag{19}$$

Therefore,

$$\nabla_{wG}\mathbf{v} = -\sum_{y\in\mathcal{N}(x)}\left(\mathbf{v}(\mathbf{y},\mathbf{x})-\mathbf{v}(\mathbf{x},\mathbf{y})\right)\sqrt{\mathbf{w}(\mathbf{x},\mathbf{y})}. \tag{20}$$

4. Experiments

In this section, we evaluate our algorithm on both the synthetic scatter point cloud dataset and real image dataset. Then, this method is utilized as a preprocess step for manifold learning and classification task. The major parameters of our algorithm include (1) the neighborhood size k; (2) the smooth term weight α; and (3) the local structure term weight β.

4.1. Experiments on Synthetic 3D Scatter Cloud Data. In this part, we test our algorithm on the classical "swiss roll" dataset. The data points are sampled from 2D manifold randomly embedded in the 3D space like a swiss roll cake. Figures 4(a) and 4(b) at first row are noiseless and noise dataset at [−8, 10] and [0, 0] viewpoint, respectively. It is obvious that noise data points distribute around the "swiss roll" manifold but do not lie on it exactly. Our goal is to recover the noiseless dataset in Figure 4(a) by the noise dataset in Figure 4(b). In this experiment, we set the number of data points $n = 1300$, KNN parameter $k = 12$, and the noise parameter NI = 1. The MATLAB code of the swiss roll dataset is listed in Table 1.

The second, third, and fourth rows in Figure 4 are denoising results by our method with α and β equal to (1, 1), (3, 1), and (0.3, 1), respectively. For ease of viewing, we set the denoising datasets at [−8, 10] and [0, 0] viewpoints in the left and right columns. In the right column, it is easy to see that the denoising data points are closed to the tangent space of manifold compared with (b), which show that our method is effective. Among them, (f) seems to be the best result because the denoising points are the nearest to manifold compared with (d) and (h). However, the "cluster" phenomenon arises in the denoising dataset; some points are close together and the other points are pushed away, which is easy to see in (e). The reason is that the large smooth parameter ($\alpha = 3$) makes geometric distribution distort when minimizing the objective function. Conversely, the "cluster" phenomenon in (g) is not serious when we set a small parameter $\alpha = 0.3$, but the noise is large.

To conduct a quantitative comparison between noise and denoising datasets, we assess the quality of the denoising datasets by mean square error (MSE) and tangent distance error (TE). MSE is a widely used index which measures the average squared Euclidean distance difference between two datasets:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(u_i - u_i^*\right)^2, \tag{21}$$

where N is the point number of the dataset. u_i and u_i^* are a noise data point and corresponding noiseless data point.

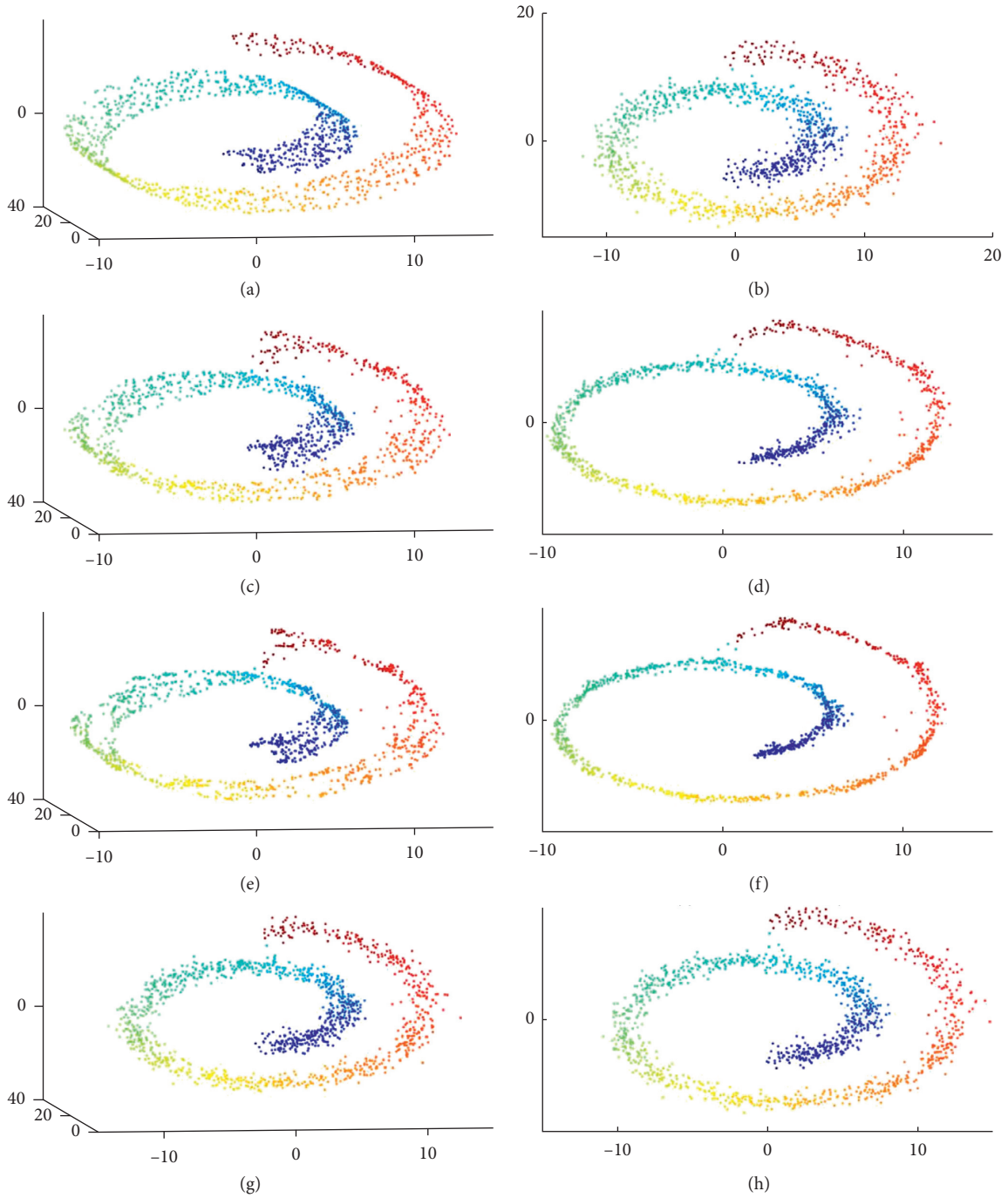


Figure 4: Denoising results with different parameters. (a) Noiseless dataset. (b) Noise dataset. (c) and (d) Denoising results with $\alpha = 1$ and $\beta = 1$. (e) and (f) Denoising results with $\alpha = 3$ and $\beta = 1$. Large α eliminates noise, but "cluster" phenomenon arises. (g) and (h) Denoising results with $\alpha = 0.3$ and $\beta = 1$.

Table 1: MATLAB code for the noise swiss roll dataset.

```
Input: the number of datasets: n; noise parameter: NI
Output: swiss roll dataset, noiseless and noise
t = (3*pi/2)*(1 + 2*rand(n, 1));
Height = 30*rand (n, 1);
Noiseless data = [t .*cos(t) height t .*sin(t)];
Noise data = [t .*cos(t) height t .*sin(t)] + NI*randn(n, 3);
```

The tangent distance error (TE) measures the distance of u_i to the tangent space of the manifold. A small TE indicates that u_i lies on the manifold and noise is weak. On the contrary, the noise strength is large if TE is big. For the convenience of calculations, we approximate TE as the Euclidean distance between u_i and its nearest data point in the noiseless dataset. The tangent distance error (TE) is defined as

$$\mathrm{TE} = \frac{1}{N}\sqrt{\sum_{i=1}^{N} u_i - \min_{u_i^*} d\left(u_i, u_i^*\right)^2}, \quad \text{s.t. } u_i^* \in U^*, \tag{22}$$

where N is the number of data points, u_i and u_i^* represent the denoising data point and noiseless data point, respectively. U^* is the noiseless dataset.

To evaluate our algorithm, we test seven sets of α and β ranging from 0 to 10. MSE and TE are listed in Tables 2 and 3. When α and β equal 0, the "data term" is the only term remaining in the objective function (8). Therefore, the denoising dataset is the same as the noise dataset and the value at $(\alpha = 0, \beta = 0)$ is the errors of the noise dataset. While α is small and β is large, the "data term" and "local structure term" maintain the geometric structure of the noise dataset. Therefore, the errors at the upper right of the table are close to the errors of the noise dataset. While α is large and β is small, the "smooth term" plays a major role. It could lead to a "cluster" phenomenon which distorts the geometric structure of the dataset and make errors large at the bottom left of the table. It is able to see that the errors near the diagonal of tables are much smaller than the others.

Table 2: MSE of our method (10^{-1}).

α β	0	0.2	0.5	0.8	1	3	10
0	2.53	2.53	2.53	2.53	2.53	2.53	2.53
0.2	2.36	2.30	2.25	2.27	2.30	2.35	2.49
0.5	3.00	2.67	2.44	2.34	2.33	2.28	2.40
0.8	3.77	3.18	2.76	2.55	2.46	2.26	2.34
1	4.28	3.52	3.00	2.73	2.60	2.30	2.33
3	9.63	7.29	5.54	4.67	4.22	2.82	2.30
10	28.4	21.5	15.3	11.9	10.6	5.37	2.91

Table 3: ET of our method (10^{-2}).

α β	0	0.2	0.5	0.8	1	3	10
0	2.01	2.01	2.01	2.01	2.01	2.01	2.01
0.2	1.70	1.73	1.74	1.80	1.81	1.91	1.95
0.5	1.68	1.67	1.69	1.70	1.70	1.80	1.91
0.8	1.74	1.69	1.66	1.67	1.68	1.72	1.89
1	1.80	1.72	1.69	1.65	1.66	1.72	1.87
3	2.42	2.12	1.93	1.82	1.80	1.66	1.71
10	4.55	3.73	3.14	2.60	2.47	1.89	1.65

4.2. Experiments on the Image Dataset. In this part, we test our method on two real image datasets: MNIST handwritten number dataset [23] and "LLE face" dataset. Image is regarded as a point in pixel space. For example, the image in the MNIST dataset could be regarded as a point in 784-dimensional space because it has 784 pixels. Therefore, the only difference between this part to experiment 3.1 is that the dimensionality of image-point is much higher than the synthetic scatter point in 3D space.

We analyze denoising images both from the subjective and objective aspects. Firstly, our method is applied to raw image datasets. Ideally, key factors that control the geometric distribution of the dataset could be maintained and the characteristics in individual images are removed as noise. Since there is no ground truth of the raw image dataset, we could only evaluate results by eyes subjectively. Secondly, we add several types of noise in an image and utilize MSE to measure the denoising images by our method and classical image denoising methods objectively.

4.2.1. Experiments on the Raw Image Dataset. We select "number 3" and "number 4" datasets in MNIST which contain 1010 and 982 images, respectively. The size of each image is $28 * 28$ pixels. The "LLE face" dataset contains 1965 face images with different expressions and shooting angles. The size of each image is $28 * 20$ pixels.

Figure 5 shows 110 images in the "handwritten number 3" dataset. The left side is original images and the right side is the corresponding denoising images by our method. In this experiment, $k = 15$, $\alpha = 0.8$, and $\beta = 1$. Four typical images are marked with a box and listed in Figure 5. It can be seen that the blurring strokes become clear and the posture of number in the image is maintained.

Figure 6 shows the 110 images in the "handwritten number 4" dataset. The left and right sides are original images and the corresponding denoising images by our method, respectively. In this experiment, $k = 15$, $\alpha = 8$, and $\beta = 1$. It can be seen that the denoising images maintain the main factors, such as the angularity of number "4." And the individual characteristics are removed after denoising; for example, the difference of stroke width becomes small after denoising. Four typical images are marked with a box and listed in Figure 6. It is obvious that the margin of "head" of number "4" becomes large in the first two images after denoising. In the third image, the stroke width becomes broad. In the fourth image, the "bend" at the upside of the stroke is removed.

Figure 7 shows the denoising result for the LLE face dataset. This dataset contains 1965 face images and the size of each image is $28 * 20$ pixels. In this experiment, $k = 15$, $\alpha = 3$, and $\beta = 0.8$. [4] shows that this dataset distributes on the manifold that spans by two key factors: head pose and expression, where the expression reflects by lip shape in images.

It can be seen that these two factors are maintained after denoising and the characters in the individual image are removed as noise. Four typical images are marked with a box and listed in Figure 7. In the first two images, the head twists to the left and right slightly in the original dataset whereas the head pose is fixed after denoising. In the third image, the original head seems to be smaller than the other images which may be caused by camera shake. The corresponding denoising image enlarges the face, and the cheek and chin became "fat." In the fourth image, the eyes are "open" after denoising.

4.2.2. Experiments on the Noise Image. In this part, we add several different types of noise to an LLE face image. Then, our method and three classical image denoising methods are

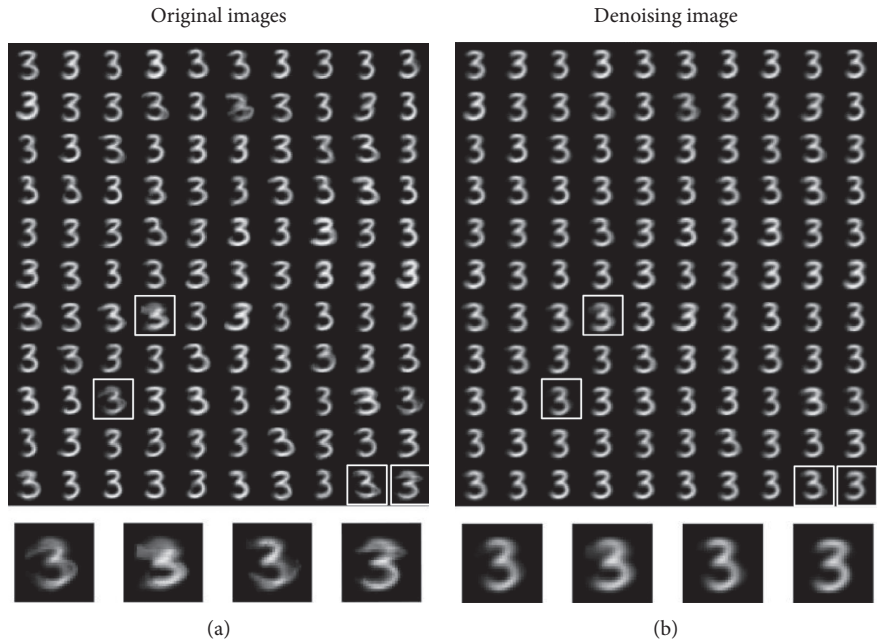


Figure 5: Denoising for the MNIST number 3 dataset. Original images and corresponding denoising images are listed in the left column and right column. The blurring strokes become clear.

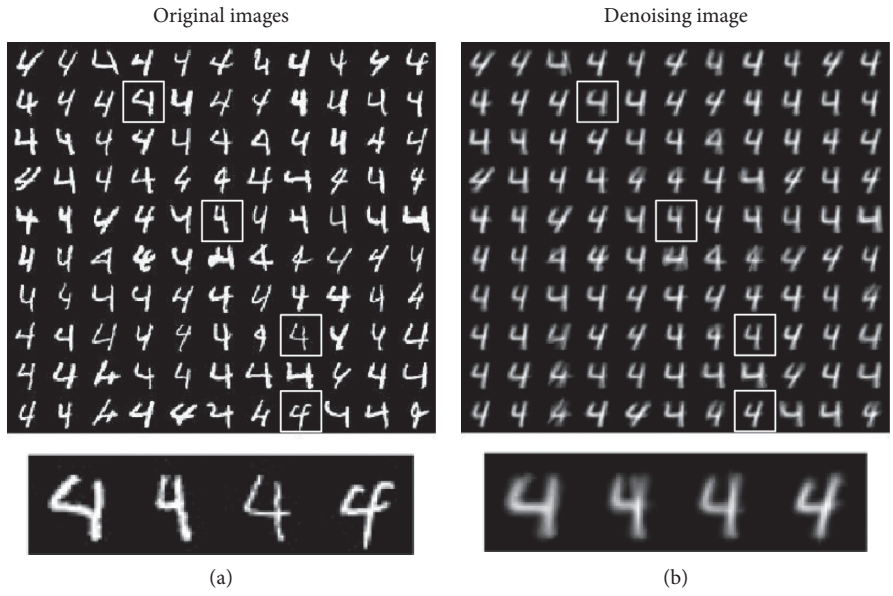


Figure 6: Denoising for the MNIST number 4 dataset. Original images and corresponding denoising images are listed in the left column and right column. The stroke widths become similar, and the posture of number 4 is maintained.

applied to these noise images. Finally, MSE is utilized to evaluate denoising images.

Figure 8 shows the denoising images by four denoising methods for five types of noise. The first column is a raw LLE face image. Brightness noise, Gaussian noise, salt and pepper noise, rotation noise, and scaling noise are added to the raw image which are shown in the second column, top to bottom row. The MATLAB code of noise model is listed in Table 4.

Three classical denoising methods, mean filtering, median filtering, and Tikhonov method are utilized to deal with these noise images. The corresponding denoising images are listed in the third, fourth, and fifth columns in Figure 8. The images in the last column are denoising results by our method. MSE is listed below each image. In this experiment, the size of the raw LLE face image is $28 * 20$ pixels. In mean filtering, the size of the filter is $2 * 2$ pixels. In median filtering, the size of the filter is $3 * 3$ pixels. In the Tikhonov

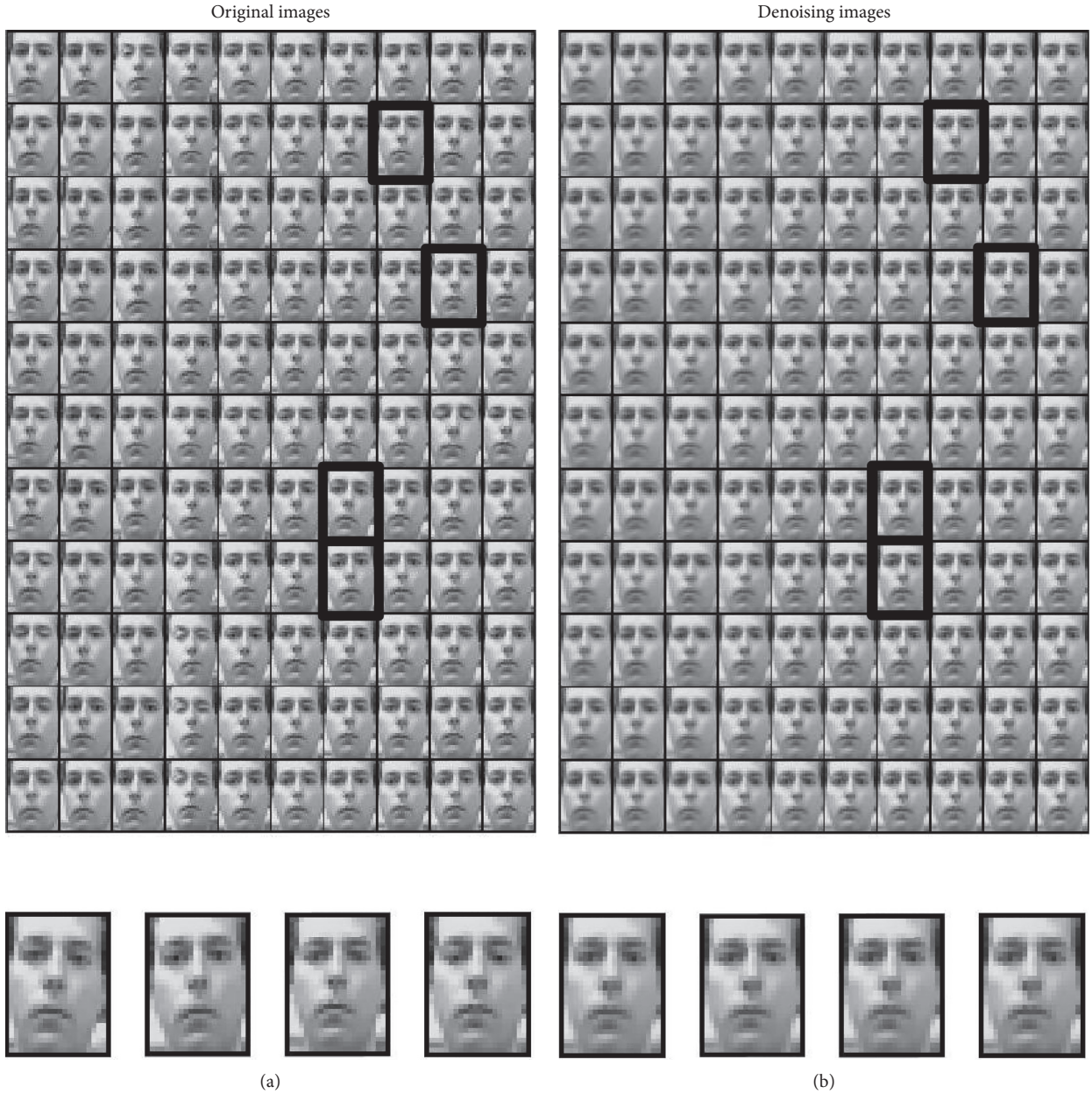


Figure 7: Denoising for the LLE face dataset. (a) Noise images. (b) Denoising images. Four corresponding typical images are listed. The posture of the head in the first two images is fixed after denoising. The head gets bigger in the third image. The eyes become open in the fourth image.

method, the smooth parameter is 0.3. The parameters in our method are set to $k = 15$, $\alpha = 3$, and $\beta = 3$.

It can be seen that three classical denoising methods have no effect on brightness noise, rotation noise, and scaling noise. These noises still exist in denoising images. The MSE even becomes larger after denoising in contrast to the noise image whereas our method has a good effect. For example, the rotation face is fixed at the fourth row and sixth column and MSE becomes smaller.

The reason is that classical image denoising methods make use of the pattern information in a single image. They could not "see" the geometric distribution information of the whole image dataset whereas our method removes noise by drawing noise data points back to the noiseless manifold-geometric distribution of the image dataset.

4.3. Denoising Dataset for Manifold Learning. In this part, we utilize our method as a preprocessing step and compare the recovered low-dimensional embeddings of noise and denoising datasets on several manifold learning algorithms. In this experiment, α, β, and k are 1, 0.8, and 13.

Figures 9(a) and 9(b) are noise "swiss roll" dataset and the ground truth of the noise dataset. Figures 9(c) and 9(d) are embeddings of the noise and denoising dataset by Isomap. Figures 9(e) and 9(f) are embeddings of the noise and

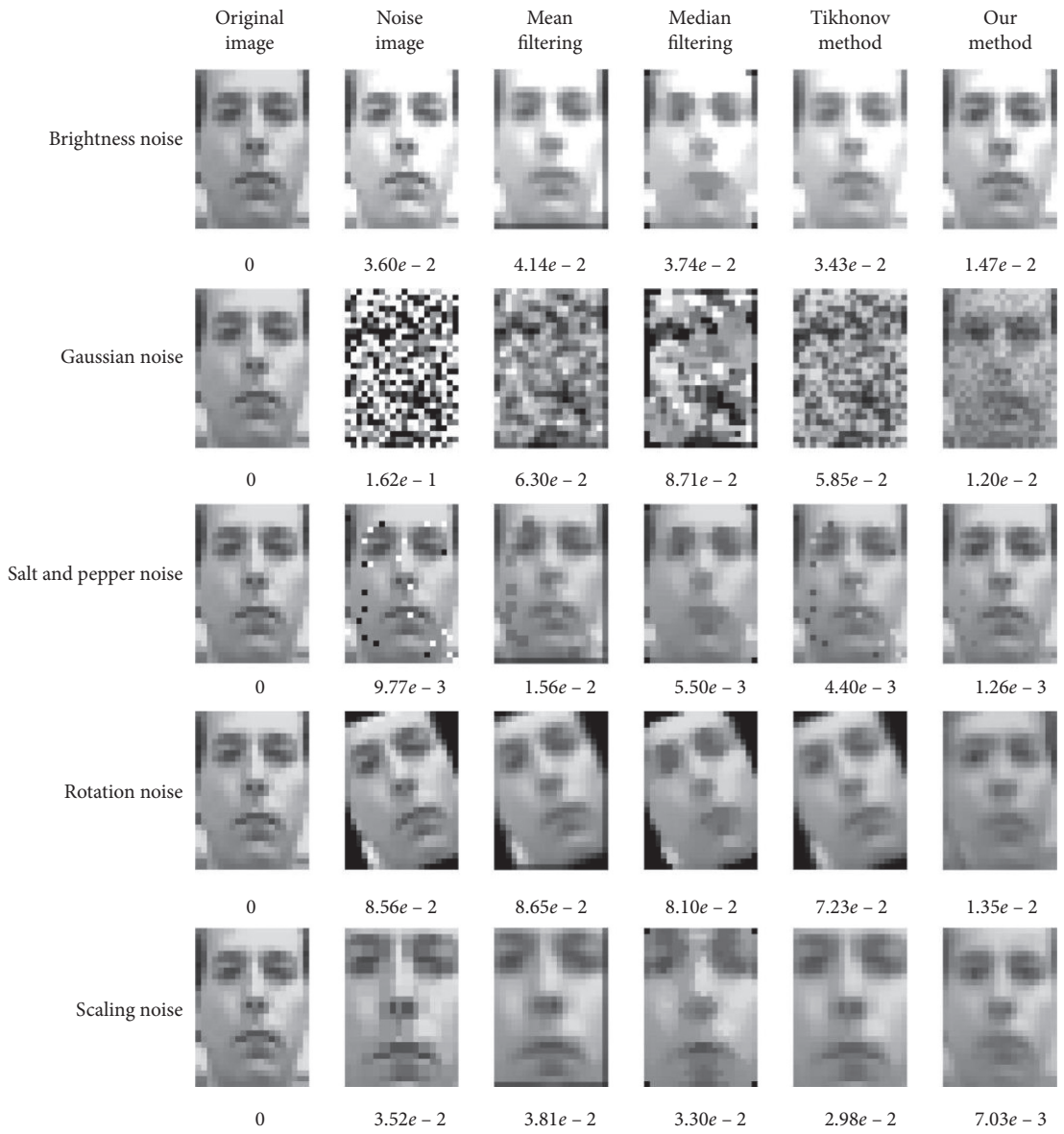


Figure 8: Denoising images' comparison. Three classical image denoising methods and our method are applied to image with five types of noise. Our method could eliminate this noise, whereas the classical image denoising methods could not deal with brightness noise, rotation noise, and scaling noise.

Table 4: MATLAB code for the noise model.

Brightness noise	NoiseImage = Image × 1.3
Gaussian noise	NoiseImage = imnoise (Image″, localvar″, size (Image)*0.5)
Salt and pepper noise	NoiseImage = imnoise (, ″salt & pepper″)
Rotation noise	NoiseImage = imrotate (Image, 20, ″bicubic″, ″*crop*″)
Scaling noise	NoiseImage = imresize (Image, 1.4, ″nearest″)

denoising dataset by LTSA. Figures 9(g) and 9(h) are embeddings of the noise and denoising dataset by HLLE. It is obvious that embeddings of the noise dataset could not reflect the geometric distribution of manifold since the neighborhoods easy to result in the "short circuit" phenomenon. By taking the denoising dataset, all the three manifold methods could get the proper embeddings. The results of Isomap result in the "hole" phenomenon because the calculated geodesic distance is always larger than it really is.

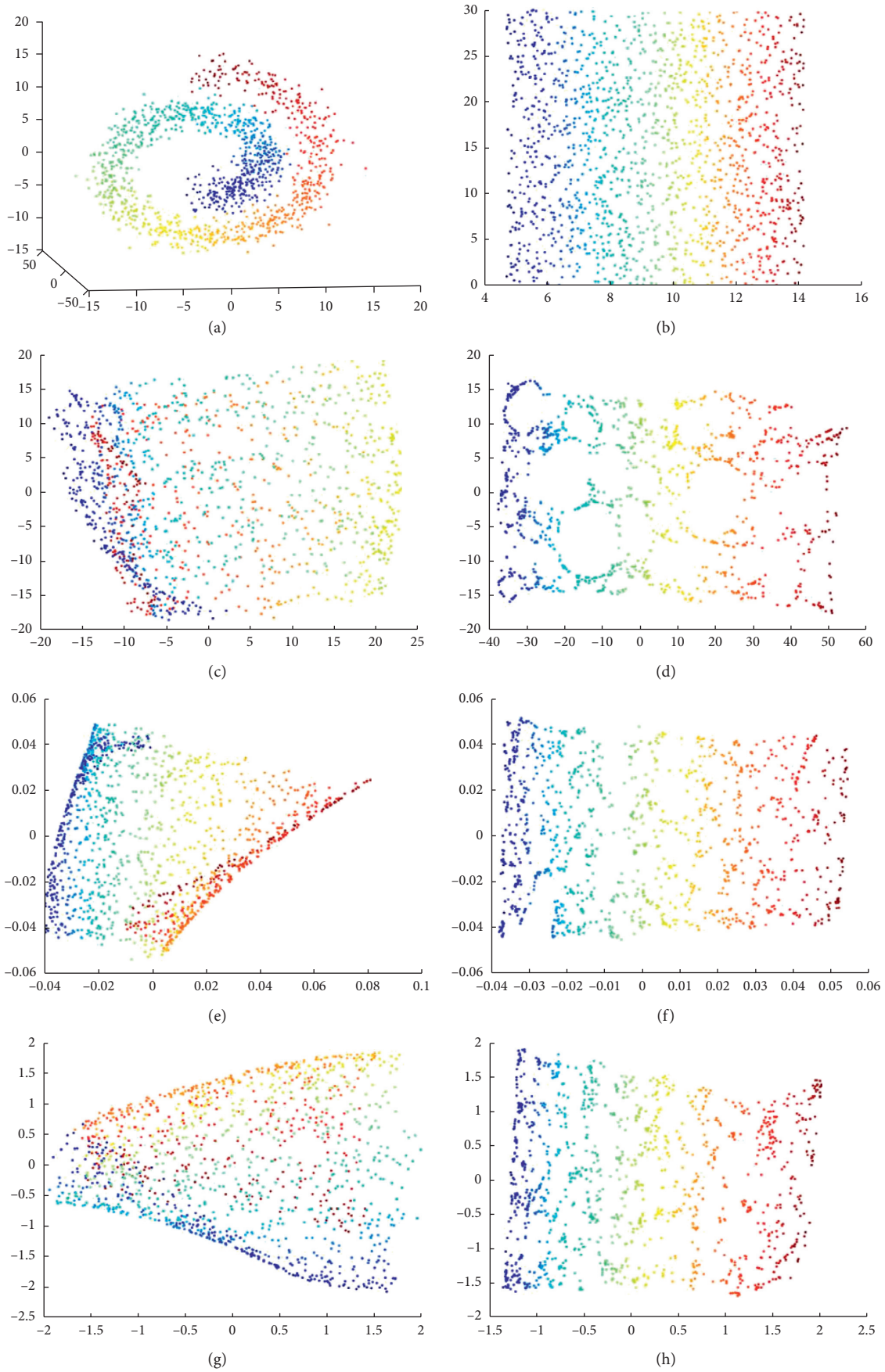


Figure 9: Embeddings of the noise dataset. (a) Noiseless dataset. (b) Ground truth. (c) and (d) Embeddings of the noise and denoising dataset by Isomap. (e) and (f) Embeddings of the noise and denoising dataset by LTSA. (g) and (h) Embeddings of the noise and denoising dataset by HLLE.

To conduct a quantitative comparison, we assess the quality of the embeddings by three indexes: embedding error, trustworthiness error, and continuity error [8]. The embedding error E measures the squared distance from the recovered low-dimensional embeddings to the ground truth coordinates which could be defined as

$$E = \sqrt{\sum_{n=1}^{N} y_n - y_n^{*2}}, \tag{23}$$

where N is the number of data points and y_n and y_n^* represent the embedding coordinates and ground true coordinates, respectively. This index tends to measure global structure distortion of the manifold.

The trustworthiness error T and continuity error C measure the local geometric structure distortion. The trustworthiness error measures the proportion of points that are too close together in the low-dimensional embedding and continuity error measures the proportion of points that are pushed away:

$$T(k) = 100 \times \frac{2}{Nk(2N - 3k - 1)} \sum_{n=1}^{N} \sum_{m \in U_n^{(k)}} (r(n,m) - k), C(k) = 100 \times \frac{2}{Nk(2N - 3k - 1)} \sum_{n=1}^{N} \sum_{m \in V_n^{(k)}} (\hat{r}(n,m) - k), \tag{24}$$

where k is the point number in the neighborhood, $r(n,m)$ is the rank of the point u_m in the ordering according to the pairwise distance from point u_n in the high-dimensional space, and $\hat{r}(n,m)$ is the rank of the point y_m in the ordering according to the pairwise distance from point y_n in low-dimensional embedding. The variables $U_n^{(k)}$ and $V_n^{(k)}$ denote the neighborhood points of u_m in low-dimensional embedding and high-dimensional space, respectively.

We test our method on several dimension reduction methods. The noise swiss roll dataset contains 1300 points. Here, we set α, β, and k to 1, 0.8, and 13. The best embedding results among several trials are selected in this experiment. The embedding error, trustworthiness error, and continuity error are listed in Tables 5–7, respectively. To show the effectiveness of our method, the errors of noise dataset, denoising dataset, and noiseless dataset are listed in three rows. It could be seen that the errors become small by taking the denoising dataset in Isomap, LLE, HLLE, LTSA, and AML. However, LE and LPP have a poor performance by taking denoising dataset.

4.4. Classification Experiment. In this part, we utilize our method as a preprocessing step and compare the accuracy rate of the original dataset and denoising dataset in the classification task. MNIST handwritten number dataset is selected which contains 60000 images with ten classes from numbers 0 to 9. Each class has about 6000 images and the size of each image is $28 * 28$ pixels. To get the denoising dataset, we utilize our denoising method for these ten classes, respectively.

In this experiment, we specify different numbers of images in each class as training data and utilize the remaining images as test data both in the original dataset and denoising dataset. A simple one-hidden-layer neural network is adopted as a classifier. The input layer has 784 units corresponding to the pixels in an image. The output layer has 10 units corresponding to ten categories from number zero to nine. We set 25 units in the hidden layer including a bias unit. The parameters of the network are trained by the BP method.

For each classification task, we repeat 10 times and list the mean accuracy rate in Figure 10. The labels "original dataset" and "denoising dataset" are raw MNIST dataset and denoising dataset with our method. The x-coordinate is the number of training images in each class and the y-coordinate is the accuracy rate. The blue and red lines are the accuracy rate of the original dataset and denoising dataset, respectively. It is obvious that the accuracy rate goes down as the number of training images decreases in each class. The performance of the denoising dataset is much better than the original dataset, especially when the training number is less than 50 in each class. The accuracy is above 96% even when there are only 10 training images in each class for the denoising dataset.

The reason is that the individual characters are removed in the denoising dataset, which is shown in Figures 5–7 in Section 3.2.1. The denoising datasets that distribute on a "clean" manifold expanded by key factors of the dataset could make machine learning algorithm easy to learn the geometric distribution knowledge of the dataset. It also illustrates that there is some kind of essential features to the classifier that is captured by our method.

Table 5: Embedding error.

	Isomap	LLE	HLLE	LTSA	LE	LPP	AML
Noise dataset	227.3	217.8	80.21	313.06	189.5	153.2	189.0
Denoising dataset	32.76	60.13	31.79	31.71	135.4	145.8	25.34
Noiseless dataset	28.79	54.63	13.42	25.80	87.29	147.9	24.70

Table 6: Trustworthiness error.

	Isomap	LLE	HLLE	LTSA	LE	LPP	AML
Noise dataset	12.78	11.46	4.43	13.91	27.48	8.39	12.84
Denoising dataset	2.99	3.29	0.94	1.23	4.12	8.34	1.09
Noiseless dataset	1.62	2.09	0.29	0.92	4.08	6.74	0.88

Table 7: Continuity error.

	Isomap	LLE	HLLE	LTSA	LE	LPP	AML
Noise dataset	5.96	3.34	1.14	4.87	5.89	2.51	4.44
Denoising dataset	1.83	2.11	0.52	0.67	2.29	2.30	0.60
Noiseless dataset	1.57	1.88	0.24	0.49	2.34	2.43	0.44

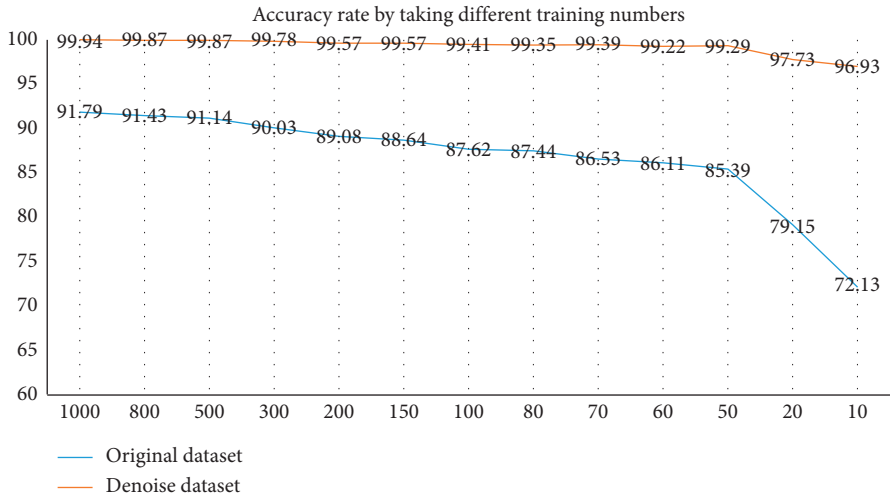


Figure 10: Accuracy rate by taking different numbers of training images.

5. Conclusion and Future Work

We propose a denoising method for the dataset rather than a single data point. This method is inspired by the manifold assumption. A local structure term is added in the Tikhonov model to make the noise points diffuse on the tangent space of the manifold. Our method could prominent the major factors hidden in the dataset and remove characteristics of the individual data point. Experiments show that our method could eliminate noise effectively on both synthetic scatter point cloud dataset and real image dataset. And as a preprocessing step, our method could improve the robustness of manifold learning and increase the accuracy rate of the classification problem. However, the parameters are sensitive in this model because the optimal solution is calculated by iteration. The geometric distribution of the dataset is distorted when the smooth term parameter is large. On the contrary, the noise intensity is still large after denoising. Our future work will focus on this problem.

Data Availability

Some or all data, models, or codes generated or used during the study are available from the first author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 51705304 and 61671285) and the Natural Science Foundation of Shanghai (Grant no. 19ZR1420800).

References

[1] H. S. Seung and D. D. Lee, "COGNITION: the manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268-2269, 2000.

[2] Columbia University Image Library (COIL-20). https://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[3] J. B. Tenenbaum, V. D. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[4] S. T. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, vol. 14, no. 6, pp. 585–591, 2002.

[6] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[7] Z. Zhang, J. Wang, and H. Zha, "Adaptive manifold learning," *Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 253–365, 2012.

[8] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao, "Maximal linear embedding for dimensionality reduction," *Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1776–1792, 2011.

[9] Bo Zhu, J. Z. Liu, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain transform manifold learning[J]," *Nature*, vol. 7697, no. 555, pp. 487–492, 2018.

[10] S. Rahimi, A. Ali, and M. Ezoji, "Human action recognition by Grassmann manifold learning," in *Proceedings of the 9th Iranian Conference on Machine Vision and Image Processing (MVIP)*, pp. 61–64, Tehran, Iran, November 2015.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *Machine Learning*, vol. 7, pp. 2399–2434, 2006.

[12] W. Liu, X. Ma, Y. Zhou, D. Tao, and J. Cheng, "p-Laplacian regularization for scene recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 2927–2940, 2019.

[13] P. B. Pratik, D. Wu, and Y. She, "Why deep learning works: a manifold disentanglement perspective," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 1997–2008, 2016.

[14] X. Gu, F. Luo, J. Sun, and S.-T. Yau, "Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge-Ampère equations," *Asian Journal of Mathematics*, vol. 20, no. 2, pp. 383–398, 2016.

[15] N. Lei, D. An, Y. Guo et al., "A geometric understanding of deep learning," *Engineering*, vol. 6, no. 3, pp. 361–374, 2020.

[16] Z. Hao, J. Liu, S. W. Ma, Xin Jin, and Xin Lian, "Noise-removal method for manifold learning," in *Proceedings of the International Conference on Life System Modeling and Simulation*, pp. 191–200, Phuket, Thailand, August 2017.

[17] Y. B. Tang, Y. Chen, N. Xu, A. Jiang, and L. Zhou, "Image denoising via sparse coding using eigenvectors of graph Laplacian," *Digital Signal Processing*, vol. 50, pp. 114–122, 2016.

[18] D. Wang, G. Song, and X. Tan, "Bayesian denoising hashing for robust image retrieval," *Pattern Recognition*, vol. 86, pp. 134–142, 2019.

[19] D. Gong, F. Sha, and G. Medioni, "Locally linear denoising on image manifolds," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 265–272, Sardinia, Italy, May 2010.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] W. Zhu, "Nonlocal variational methods in image and data processing," University of California, Los Angeles, CA, USA, Doctor of Philosophy in Mathematics, 2017.

[22] G. Gilboa and O. Stanley, "Nonlocal operators with applications to image processing," *Multiscale Model. Simul.*vol. 7, no. 3, pp. 1005–1028, 2008.

[23] The MNIST Database of Handwritten Digits. https://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6685190, 17 pages
https://doi.org/10.1155/2021/6685190

Research Article

An Improved Monte Carlo Method Based on Neural Network and Fuzziness Analysis: A Case Study of the Nanpo Dump of the Chengmenshan Copper Mine

Feng Gao, Xiaodong Wu, and LeWen Wu

School of Resources and Safety Engineering, Central South University, Changsha 410083, China

Correspondence should be addressed to Feng Gao; csugaofeng@126.com

Received 22 October 2020; Revised 30 November 2020; Accepted 28 December 2020; Published 15 January 2021

Academic Editor: Chih-Cheng Hung

The landslide of dump is a man-made geological disaster which will bring great harm to the surrounding people and environment, and probabilistic reliability analysis is commonly used to analyze the probability of slope landslide or whether protective measures should be taken. Monte Carlo simulation is the most commonly used method, but there are some problems, such as low efficiency, statistical ambiguity of small samples, and the fuzzy transition interval of the stability criterion. This paper proposes an improved Monte Carlo method that uses an improved bootstrap method to process small samples of geotechnical data, employs ELM (extreme learning machine) based on PSO (particle swarm optimization) to fit the limit equilibrium method function, and constructs the safety factor membership function of the dump site considering the fuzzy transition interval. This method was applied to an example slope of the dump site in Chengmenshan, Jiangxi. Comparing the analysis result with the result of the traditional MCS (Monte Carlo Search) method, it was found that after adding the safety factor membership function, the result was closer to the actual situation of the dump site, and the probability of failure and reliability index values were closer to those of the dangerous state; after the original function was replaced by the PSO-ELM model, the efficiency of the MCS method was greatly improved while the results maintained high consistency with the original results; the MCS method combined with the bootstrap method not only simulated the fuzzy uncertainty of the original sample statistics and distribution type but also expressed the reliability index and probability of failure as a two-sided confidence interval with a certain confidence level. The above conclusion proves the effectiveness and superiority of this method compared with the original MCS method.

1. Introduction

As a man-made geological disaster with the rise of mining industry, dump landslide not only threatens the life and property safety of mine personnel but also has a huge impact on the surrounding environment. Dump sites in mining areas have been studied by domestic and foreign scholars for stability analysis. At present, the most commonly used methods [1] in engineering are the limit equilibrium method and numerical analysis based on the finite element method, obtaining a certain safety factor value of the analyzed slope and clarifying the sliding surface or part that is most likely to be damaged. However, it often ignores widespread uncertainty of slopes' stability analysis (such as the uncertainty of the calculation parameters of the rock and soil, the geometric model, and the functional relationship between the slope-influencing factors and the safety factor), so the effect is not satisfactory. Reliability analysis introduces probability theory into the slope stability evaluation system and obtains the slope failure probability and reliability index based on consideration of the above uncertain factors, so as to achieve a more reasonable and scientific evaluation of the stability of the slope. In recent years, it has been widely valued and studied by scholars.

The reliability analysis methods of soil slopes can be roughly divided into the first-order second moment method [2–8], Monte Carlo method [9–14], and response surface method [11, 15–23]. Among them, the Monte Carlo method is favored in soil slope reliability analysis due to its accuracy and ease of operation (no matter how complicated the limit

state equations and calculations are, according to statistical principles, only enough simulation sampling times and random number sequences are needed to obtain an accurate failure probability value), but some problems also exist.

Firstly, the original limit equilibrium iterative calculation method that links the soil parameters and the slope safety factor requires a certain amount of time, and the Monte Carlo method has to repeat the process hundreds of thousands or even millions of times, so a huge amount of time is needed. Many scholars use models from other fields to simulate and approximate the functional relationship between slope safety factors and soil parameters to solve the problem of low efficiency. He [24] used sample data obtained by deterministic calculations to train a support vector machine (SVM) model to approximate the functional function in the reliability analysis and then combined it with the Monte Carlo method to calculate the probability of slope instability. After the application, it was proved that the accuracy of this method was better than the (first-order reliability method) FORM method, and the efficiency was higher than the ordinary MCS method. Tan et al. [25] proposed a method to replace the original functional function in reliability analysis with an SVM and (radial basis function) RBF model fitted based on two new sample selection methods and combined different examples to illustrate the accuracy and efficiency of the two. Su [26] constructed a Monte Carlo method based on Gaussian regression response surface and verified the effectiveness and efficiency of the method through three slope calculation examples. In addition, artificial bee colony evolution algorithm [27], artificial neural network [28], Kriging [29–31], and vector projection [32] have been used to construct the functional relationship between input parameters and output safety factors, alleviating the problem of low efficiency to a certain extent.

Secondly, the traditional MCS method also has the problem of fuzziness. Fuzziness is mainly reflected in two aspects: the traditional reliability analysis using $Z < 0$ as the criterion to judge slope failure is too arbitrary, ignoring the fuzzy interval of the intermediate transition between stability and instability, and the variability of the sample mean and standard deviation caused by insufficient sample test data will further cause fuzzy uncertainty of the soil parameter distribution types and statistics and ultimately affect the reliability calculation results. In response, some scholars try to solve these problems from the perspective of combining fuzzy mathematics theory and mathematical statistics methods. Habibagahi and Meidani [33], Xu et al. [34], Jia and He [35], Lou [36], Xu [37], Anvar et al. [38], and others successively established the membership function between safety factor and slope stability based on fuzzy mathematics theory. The membership function form has undergone an evolution from linear to nonlinear, triangular to trapezoidal and then to ridge distribution, and it has become closer to the reality of slope stability. However, the ways to determine the undetermined coefficients of the membership function are relatively scarce and subjective. It is necessary to collect as many corresponding slope examples as possible to help judge the rationality of the membership function. Most and Knabe [39], Luo et al. [40], and Tang [41] used the well-known bootstrap theory in statistics to expand the sample in order to reduce the turbulence of the reliability calculation caused by the variability of the parameter sample mean, standard deviation, and distribution, and the reliability index can be more reasonably characterized as a confidence interval with a certain confidence level. This will not only fully simulate the variability of sample statistics, distribution types, and reliability calculation results but also make the slope reliability calculation results more reasonable and true. However, the current combination of bootstrap and Monte Carlo methods is rarely applied, because each bootstrap subsample generated by sampling requires a Monte Carlo calculation, which brings a huge computational burden, so it is necessary to combine some measures to improve the efficiency of the Monte Carlo method.

The Chengmenshan Copper Mine in Jiujiang, Jiangxi Province, due to site selection constraints, had to construct a dump in the lake area near the mining area where the base bearing capacity is weak. In addition, the height of the heap load is relatively high, so it is very easy for progressive landslide instability to be produced over time. Wang [1] conducted a comprehensive reliability analysis on the traditional MCS used in the Chengmenshan dump site (natural unsupported state); the average value of the calculation results under different limit equilibrium functions is $P_f = 6.65\%$, $\beta = 1.71$, and the results of different methods are highly variable. According to the 1997 US Army Corps of Engineers Index [42], the situation is between unsatisfactory and poor, which is slightly different from the actual situation of the Nanpo dump site (support measures are needed to prevent landslide). The reason might be that the statistical uncertainty caused by insufficient rock and soil samples and the fuzziness of the traditional MCS method were ignored.

According to the question above, an improved Monte Carlo method was proposed that uses improved bootstrap methods to process small samples of geotechnical data, applies particle swarm optimization (PSO)-optimized extreme learning machine (ELM) to fit limit equilibrium method function, and constructs the safety factor membership function of the dump site considering the fuzzy transition interval. It was applied to the reliability analysis of the second-stage dump site of Chengmenshan Copper Mine, and the results were compared with those of the traditional MCS method for the purpose of verifying effectiveness and rationality of the improved Monte Carlo method.

2. Traditional Monte Carlo Analysis of Chengmenshan Dump

2.1. Overview of the Dump. Chengmenshan Copper Mine is located in Chengmen Town, Jiujiang, Jiangxi. The second-stage dump site is located at the Nanpo ravine and Dachengmen valley on the southeast side of the slope. It is divided into two parts: Nanpo (south slope) and Chengmengou dump sites. The climate in the mining area has four distinct seasons, humid, hot spring/summer, cold, and dry autumn/winter, the annual average temperature is 17°C, and the annual average precipitation is 1420 mm. The basement of the dump site on the south slope is generally a slippery stratum. When the overlying waste material accumulates to a certain extent, under the pressure of its own weight and

external factors, it is easy for a slip surface to be produced along the slippery bottom layer to the slope surface area and cause a landslide accident. So far, many bottom heaves of dump slope have been found.

The second-stage dump site of the mine has six characteristic sections within the delineated area. Section E-E in the eastern part of the Nanpo dump site was selected for analysis (natural unsupported state). A cross-sectional view is shown in Figure 1, and the statistical characteristics of the 20 experimental datasets and 14 expanded datasets are shown in Table 1. In the process of parameter sampling, the cohesion and internal friction angle of rock and soil mass meet the linear negative correlation, and the correlation coefficient is taken as −0.5.

2.2. Results of Traditional Monte Carlo Slope Reliability Analysis (Limit Equilibrium) Method. For geotechnical slope engineering, the state function of the slope's stability can be expressed as

$$Z = g(X_1, X_2, X_3, \ldots, X_m), \tag{1}$$

where the value of the state function Z is the safety factor and $X_1, X_2, X_3, \ldots, X_m$ are m-many random variables with a certain distribution, which are generally the key factors affecting the stability of a slope such as cohesion and severity of rock and soil. The number of times when the safety factor value $Z \leq 1$ is counted as M and the total number of simulations is N, and then the probability of failure can be obtained according to the law of large numbers:

$$P_f = \frac{M}{N}. \tag{2}$$

The reliability index can be expressed as

$$\beta = \Phi^{-1}(1 - P_f). \tag{3}$$

Using different limit equilibrium methods to calculate the safety factor, it was found that the values obtained by the Janbu simplified method were the lowest, so this method was used for subsequent calculation. Slide is a two-dimensional slope stability analysis software based on limit equilibrium method produced by Canadian Rocscience company and is highly praised in engineering applications due to its advantages of practicality, high efficiency, and accuracy. The reliability analysis results obtained by "slide" software using the MCS method are shown in Figure 2, the failure probability of the Nanpo slope dump of the copper mine is P_f = 15.25% and the reliability index is β = 1.026 (also expressed by RI), and the average value of safe factors is 1.045. The results showed that the slope is basically in a stable state at the end of dumping, but there is a lack of sufficient safety factor reserve, so landslide accidents are very easy to occur under static conditions. Compared with the actual situation, the calculation results tend to be conservative.

2.3. Problems

Problem 1. It was found that the corresponding probability of failure varies greatly when using different limit equilibrium methods for MCS calculation. By the Janbu simplified method, it was 15.25%, and by the Bishop simplification method and Spencer's method, it was about 1.1%, and the relative error reached 92.4%. The reason should be that the basis for judging the state of the slope based on whether the safety factor is greater than 1 does not conform to the actual situation. The safety factor calculated by the simplified method for the Nanpo slope dump of the Chengmenshan Copper Mine is basically around the critical value 1; although calculated by other methods (such as the Bishop simplified method), it is only 0.05 higher than the Janbu simplified method. The distribution is mostly on the right side of the critical value 1, so there is a problem of large differences in failure probability (as shown in Figure 3). And the accuracy of the calculation results under this criterion will also be affected. So, it is necessary to introduce a stable membership function $\mu(z)$ containing the safety factor of the intermediate transition state to improve the judgment method of the slope state in traditional reliability analysis.

Problem 2. The limit equilibrium program realized in MATLAB not only needs to input the sliding surface coordinates, number of blocks, trial radius, soil parameters, and boundary coordinates and other parameters in advance [43] but also needs to simplify the geometric model boundary to a neatly planned boundary and combine some optimized algorithms (such as genetic algorithm [44]) to help search for minimum safety factors. In this paper, the results of running the MCS method in MATLAB are P_f = 15.8% and β = 0.998, and the time consuming is 63005.44 s.

The calculation efficiency of the original MCS method has been criticized, and the bootstrap sampling method needs to repeat the whole process of Monte Carlo simulation tens of thousands or even hundreds of thousands of times, which will take a huge amount of time, so other fast and accurate methods are needed to replace the iterative calculation of the limit equilibrium method in MATLAB.

Problem 3. Due to various reasons in actual slope engineering, the experimental data of rock and soil parameters are often limited, so is this dump (there are only 20 sets of data). The statistical uncertainty caused by small sample may lead to deviations in the subsequent Monte Carlo calculation results.

The following part of this article will focus on methods and applications to solve the above problems.

3. Methodology

3.1. Improved Bootstrap Method. Bootstrapping is an effective method to solve the statistical uncertainty of small sample data. The idea of the bootstrap method is to randomly sample the initial samples with replacement to obtain a large number of bootstrap subsamples (the sample size is the same as the original sample) that contain the original sample information, then calculate the estimated value of statistics and Akaike information criterion (AIC) value for each subsample, get their optimal probability density distribution according to the AIC [45], and finally perform Monte Carlo calculation to

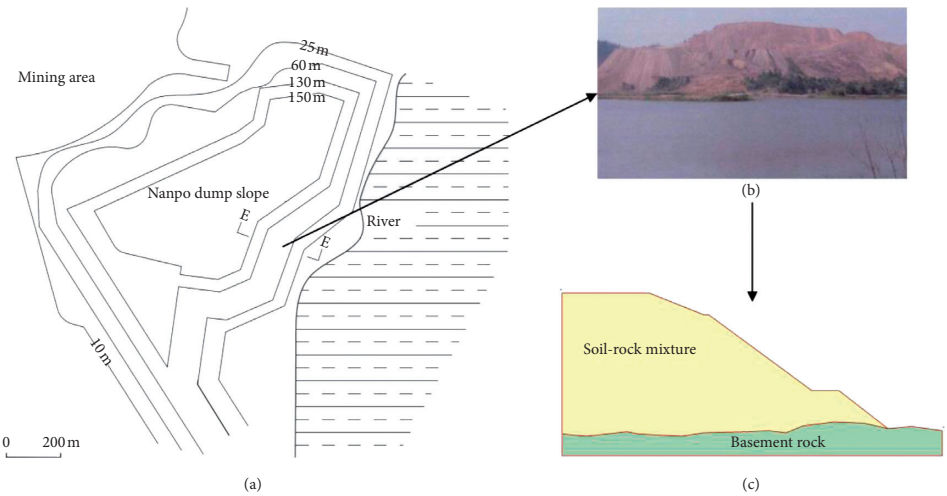


Figure 1: (a) Contour map of Nanpo dump site; (b) real view of E-E section of dump; (c) calculation model of E-E section of dump.

Table 1: Statistics of rock and soil parameters.

Number		Soil-rock mixture			Basement rock		
		Cohesion c (kPa)	Internal friction angle φ(°)	Bulk density γ(kN·m^{-3})	Cohesion c (kPa)	Internal friction angle φ(°)	Bulk density γ(kN·m^{-3})
Original 20 samples	1	38.3	23.5	17.3	913	17.1	26.5
	2	41.4	20.5	16.9	918.82	17.66	22.07
	3	39.9	26.2	17.1	914.5	17.27	22.24
	4	34.3	20.4	17.6	909.8	16.87	22.82
	5	38.1	25	18.3	904.1	14.94	19.16
	6	43.1	24	17.1	912.8	18.422	27.76
	7	41.5	24.1	17.0	915.5	18.21	25.31
	8	36.4	22	17.6	910.7	14.75	22.93
	9	45	25	19.7	909.2	16.78	27.61
	10	41.4	25.96	17.2	902	18.035	20.834
	11	43.4	26	17.83	914.7	14.807	27.83
	12	38.2	20.1	16.16	909.7	16.976	27.79
	13	40.5	19.8	17.92	907.9	19.55	18.98
	14	44.3	23.7	16.07	926	19.8	18.55
	15	32.4	23.3	17.9	882	17.1	24.37
	16	37.8	24	17.64	919.7	16.78	24.06
	17	35.6	23.5	14.7	900.6	13.07	25.3
	18	35.6	22	17.3	911.9	17.58	25.29
	19	35.8	20.5	17.6	917.2	13.59	22.69
	20	36.1	23.96	22.22	917.8	17.1	24.53
14 sets of samples after expansion	21	36	24.3	19.92	886	15.2	24.7
	22	46.9	19.8	16.4	892.2	13.3	27.9
	23	41.9	19.69	19.42	931.1	19.74	24.23
	24	36.4	27	18.8	908.5	18.87	18.08
	25	38.8	24.7	13.9	919.2	12.97	25.81
	26	32.2	22.7	22.1	923.4	12.87	19.6
	27	42.1	19.94	17.4	882.7	17.92	21.23
	28	33.8	23.28	14.62	918.3	15.81	17.81
	29	43.4	25.61	15.4	898.5	20.12	18.3
	30	37.3	19.84	23.1	902.9	17.38	23.5
	31	39.9	23.28	17.73	916.6	19.3	28.5
	32	39.5	26.31	22.54	913.8	17.2	18.7
	33	40.3	22.9	18.18	920.8	16.4	28.3
	34	42.7	25.4	16.81	881.2	20.03	26.9

Table 1: Continued.

Number		Soil-rock mixture			Basement rock		
		Cohesion c (kPa)	Internal friction angle φ (°)	Bulk density γ (kN·m^{-3})	Cohesion c (kPa)	Internal friction angle φ (°)	Bulk density γ (kN·m^{-3})
34 sets of samples after expansion	Mean	39.2	23.39	17.86	909.2	16.86	23.53
	Standard deviation	3.69	2.20	2.12	12.36	2.09	3.37
Original 20 samples	Mean	39	23.18	17.64	908.7	16.78	23.7
	Standard deviation	3.45	2	2.01	6.25	2.02	3.23

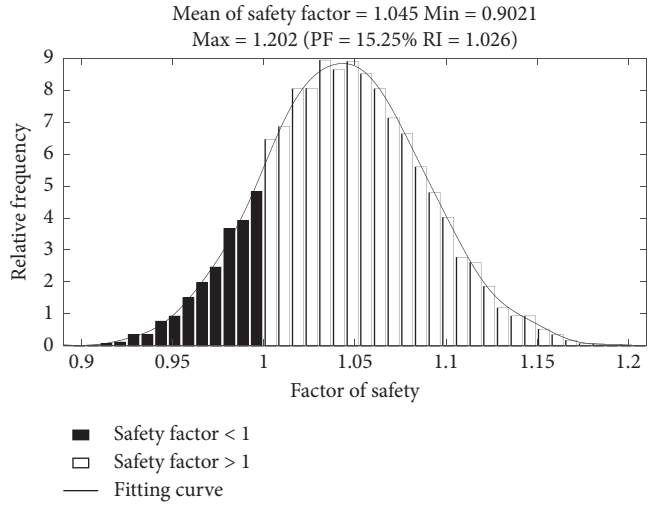


Figure 2: Monte Carlo sampling calculation results.

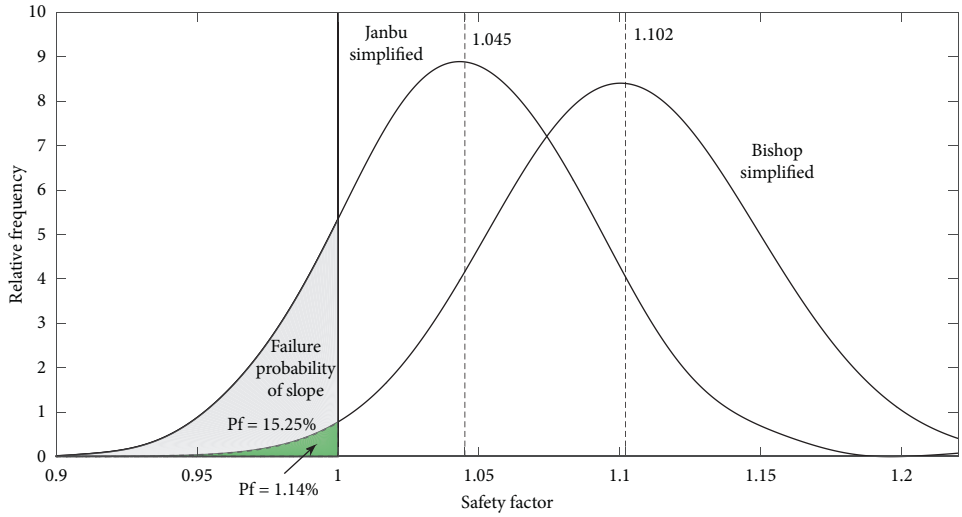


Figure 3: Differences in failure probability between limit equilibrium methods.

obtain their reliability index, probability of failure, and corresponding confidence interval.

Assume the sample of soil parameters $X = \{(c_i, \varphi_i, \gamma_i), i = 1, 2, \ldots, N\}$, then randomly sample with replacement N times as shown in Figure 4 to obtain a bootstrap subsample $B_j = \{B_{1,j}, B_{2,j}, \ldots, B_{N,j}\}$, and repeat this step M times to obtain M subsamples.

The theoretical basis and good convergence of the bootstrap method have long been proved by scholars, but it also has the problem that the sampling range is small and the probability distribution is concentrated on a small number of points for small samples (sample size of 10–30), which causes the calculation results to deviate from the true distribution. Some scholars [46–48] proposed an improved

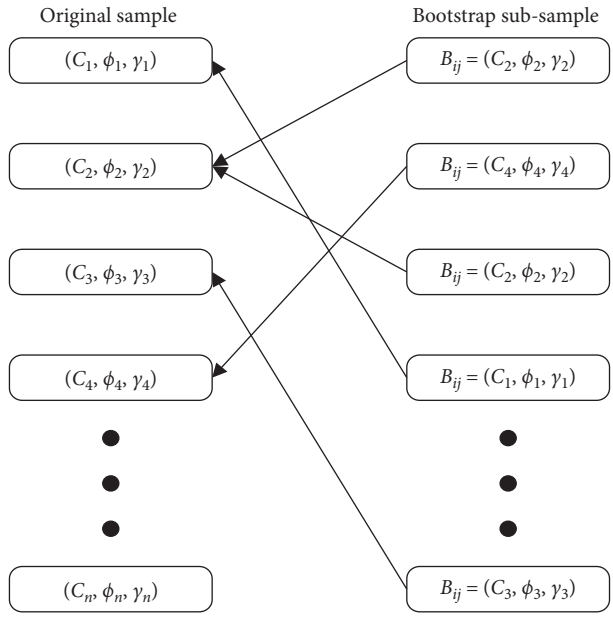


Figure 4: Bootstrap self-sampling method.

bootstrap method for this problem: first arrange the original sample in ascending order to get the order statistic $X = (x_1, x_2, x_3, \ldots, x_n)$, then, for each observation x_i, a larger and a smaller statistic are randomly selected in its neighborhood U_i according to the rule of uniform distribution to expand the sampling range and obtain information outside the observation point:

$$U = \begin{cases} U_1 = \left[x_1 - \dfrac{x_2 - x_1}{m}, x_1 + \dfrac{x_2 - x_1}{m}\right], \\ U_i = \left[x_i - \dfrac{x_i - x_{i-1}}{m}, x_i + \dfrac{x_{i+1} - x_i}{m}\right], \\ U_n = \left[x_n - \dfrac{x_n - x_{n-1}}{m}, x_n + \dfrac{x_n - x_{n-1}}{m}\right], \\ (i = 1, 2, 3, \ldots, n-1, n, \,)\,(m \geq 2). \end{cases} \tag{4}$$

The bootstrap method can continue to be used for parameter estimation based on the improved samples $X' = (x_1', x_2', \ldots, x_n')$ obtained by formula (4). Since the original sample size of 20 is less than 30, thus is a small sample, the size can be expanded by formula (4) before bootstrap sampling, but not every observation value needs to be expanded in this way, especially when it is very close to (or even the same as) an adjacent value. Use MATLAB for random permutation and combination after the six groups of data were, respectively, expanded, and the expanded sample statistics are shown in Table 1.

One of four types of functions commonly used in the distribution of rock and soil parameters, normal, lognormal, extreme value I-type, and Wilber distribution, is selected by calculate the AIC value of the original sample (the probability distribution function with the smallest AIC value can be considered as the one that best fits the probability distribution of the test data). It is found that the AIC value corresponding to the normal distribution is the smallest and the optimal probability distributions of the six types of data of the initial sample are all normal distributions, which conforms to the description in the literature [1].

The subsample mean, standard deviation, and optimal probability density function distribution type are recorded after each bootstrap sampling and used as the statistics and distribution type of random variables in the subsequent Monte Carlo calculation.

3.2. PSO-ELM Model. Extreme learning machine (ELM) is a new type of feedforward single hidden layer neural network. In addition to inheriting the good self-organization and self-adapting ability of general neural network algorithms, it also has fewer adjustable parameters, faster speed, better prediction accuracy, strong versatility, and other advantages. The input weight $\omega_{l\times n}$ and hidden layer threshold $b_{l\times n}$ are randomly selected, and then the input quantity $X_{n\times Q}$, output weight $\beta_{l\times m}$, and excitation function $g(x)$ are combined to obtain the expression of the output t_j:

$$H\beta = T, \tag{5}$$

where

$$H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1)g(w_2 \cdot x_1 + b_2)\cdots g(w_l \cdot x_1 + b_l) \\ g(w_1 \cdot x_2 + b_1)g(w_2 \cdot x_2 + b_2)\cdots g(w_l \cdot x_2 + b_l) \\ \vdots \\ g(w_1 \cdot x_Q + b_1)g(w_2 \cdot x_Q + b_2)\cdots g(w_l \cdot x_Q + b_l) \end{bmatrix},$$
$$\beta = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{l1} & \cdots & \beta_{lm} \end{bmatrix}, \tag{6}$$
$$T = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix}.$$

$$\beta = H^{+}T. \tag{7}$$

In ELM, the input weight hidden layer threshold is randomly selected, and the excitation function and input are known. Therefore, the training of the extreme learning machine is to calculate the output weight through equation (7), and then use it to perform predictions.

PSO is an intelligent optimization algorithm that simulates the foraging behavior of a bird colony to achieve swarm optimization. This method treats the solution of each problem as a particle. The distance between the particle's spatial position and the target is used as the fitness value, and each particle has a specific moving direction and speed. The PSO algorithm updates the speed and fitness in every search by tracking the individual's extreme values (Pbest, the optimal position of fitness values calculated from the positions of individuals) and the group's extreme values (Gbest, the optimal position of fitness of all particles in the population) until the termination requirements are met. The termination condition can be the number of iterations, the extreme value error of two consecutive searches below a specific tolerance value, or a mixture of both.

The extreme learning machine algorithm optimized by particle swarm optimization is shown in Table 2.

3.3. Membership Function of Safety Factor in Dump Sites. Monte Carlo reliability analysis considers that there are only three states in a slope: limit equilibrium state ($Z = 1$), unstable state ($Z < 1$), and stable state ($Z > 1$). However, the corresponding relationship between the safety factor and the stable situation in the actual soil slope is fuzzy and uncertain, which means the slope has a fuzzy interval of intermediate transition between completely unstable and completely stable. Even if the safety factor is greater than 1 or even 1.2, it may become unstable; less than 1 also has the possibility of stability. Therefore, the way the Monte Carlo method estimates slope failure probability based on the proportion of statistical failure times to total simulation times may be biased, so it is necessary to introduce the stability membership function $\mu(Z)$ of the safety factor to judge the stability of soil slope equation (2) can be improved as

$$P_f = \frac{\sum_{i=1}^{N} \mu_i(Z)}{N}. \tag{8}$$

As the name implies, the membership function is the degree to which the safety factor belongs to the stable state of the slope, $\mu(Z) \longrightarrow 0$. Slope becomes more unstable, $\mu(Z) = 0.5$. The probability of slope instability and stability is 0.5. At this time, the fuzziness is the strongest and the stability state is the most difficult to judge. At $\mu(Z) \longrightarrow 1$, the slope becomes more stable.

The membership functions commonly used to characterize the stability of geotechnical engineering structures include ridge distribution, quadratic parabolic distribution,

Table 2: Particle swarm optimization-extreme learning machine (PSO-ELM) algorithm program.

Algorithm: algorithmic flow of PSO-ELM
(1) Obtain the training and testing datasets
(2) Begin ELM training
(3) Set ELM parameters randomly
(4) Use the mean square error (MSE) as the fitness function
(5) Initialize PSO population (Inipop)
(6) Calculate the fitness value of each candidate solution
(7) S = global best solution
(8) For $i = 1$ to maximum iteration number do
(9) For $i = 1$ to P do
(10) Update the velocity and position of the ith particle
(11) Evaluate the fitness of the ith particle
1(2) Update the personal best solution of the ith particle
(13) S = current global best solution
(14) End for
(15) End for
(16) End
(17) Obtain the optimal input weights and hidden biases of hidden layer neurons using S
(18) Use the optimal input weights and hidden layer neurons for ELM test

and reduced semitrapezoidal distribution. Each of these membership functions can also be divided into three types: descending, rising, and intermediate. The membership function of the safety factor for stability should express the degree to which the safety factor belongs to the concept of stability, and additionally, the slope of the graph curve should reflect the strength of the fuzziness that affects the judgment of the steady state. The smaller the safety factor, the closer $\mu(Z)$ is to 0, the easier it is to judge that the slope is in an unstable state. Similarly, the larger the safety factor, the closer $\mu(Z)$ is to 1, and the slope is likely to be in a stable state. The fuzziness is the strongest when the safety factor and $\mu(Z)$ are in the middle, so the curve slope of the membership function first increases and then decreases, with the characteristics of slow at both ends and steep in the middle. Based on the above characteristics and engineering application principles, the raised-ridge-shaped distribution membership function was selected as the stable membership function of the slope safety factor. The function expression is as follows:

$$\mu(F_s) = \begin{cases} 0, & F_s \le a, \\ \dfrac{1}{2} + \dfrac{1}{2}\sin\left(\dfrac{\pi(F_s - b)}{2(b-a)}\right), & a < F_s \le b, \\ \dfrac{1}{2} + \dfrac{1}{2}\sin\left(\dfrac{\pi(F_s - b)}{2(c-b)}\right), & b < F_s \le c, \\ 1, & F_s > c. \end{cases} \tag{9}$$

The steps to determine the specific form in the membership function are as follows:

Step 1: collect the corresponding relationship between the actual stability of the dump site and the value of the safety factor. We collected 48 sets of statistical data on the stability factors of different dumps in China, as shown in Table 3. In the data, the highest value of safety factor of unstable slope is 1.21, and the lowest value is 0.92; therefore, it can be assumed that a dump slope with safety factor greater than 1.21 must be stable, and a dump slope with safety factor less than 0.92 can no longer remain stable. The median value of 1.065 is used to divide the safety factor into two intervals. The stable state of dump sites changes most obviously near 1.065, so the stability state of a dump slope with a safety factor of 1.065 is the fuzziest, and the corresponding membership degree should be 0.5.

Step 2: preliminarily determine the undetermined coefficient of the membership function. According to the analysis given in Step 1, $a = 0.92, c = 1.21, b = (a + c)/2 = 1.065$, and substituting into equation (9), the membership function of the safety factor of raised-ridge-shaped distribution can be expressed as

$$\mu F_s = \begin{cases} 0, & F_s \le 0.92, \\ 0.5 + 0.5\sin\left(\dfrac{\pi(F_s - 1.065)}{0.29}\right), & 0.92 < F_s \le 1.21, \\ 1, & F_s > 1.21. \end{cases} \tag{10}$$

Figure 5 shows a ridge-shaped distribution function diagram of the relationship between the safety factor and the steady state membership based on the collected data. The point with a membership value of 1 corresponds to the stable state of the dump, and the point with a membership value of 0 indicates an unstable state. The function diagram can fit the data points of the dump sites well and follow the slope membership function law mentioned above.

Step 3: use the degree of conformity between the statistical law of the dump sites and the safety factor interval, with the membership interval corresponding to the fuzzy language value, to judge whether the membership function determined in Step 1 and 2 is practical.

There are nine commonly used fuzzy linguistic values to describe a certain state; however, due to the small amount of collected samples, the number of samples in each fuzzy subset would be too small, which could cause fluctuations in statistical results, so it is necessary to reduce the fuzzy language subset describing slope stability. In this paper, five fuzzy linguistic values of stable π_1, basically stable π_2, critical π_3, basically unstable π_4, and unstable π_5 were selected to describe the stability state of the dump slope, and the membership interval calculation formula of fuzzy linguistic value in [37] was reduced to 5, as shown in equation (11), where a is a constant used to separate the interval, $a \in [0.5, 1]$, and here we take $a = 0.51$:

Table 3: Examples of safety factor and stability of soil slope of dump.

Actual state	Safety factor	Actual state	Safety factor	Actual state	Safety factor
Stable	1.877	Stable	1.242	Unstable	1.078
Stable	1.635	Stable	1.241	Stable	1.07
Stable	1.52	Stable	1.228	Stable	1.07
Stable	1.468	Stable	1.22	Unstable	1.06
Stable	1.421	Unstable	1.21	Unstable	1.05
Stable	1.404	Stable	1.2	Unstable	1.02
Stable	1.39	Stable	1.2	Unstable	1.015
Stable	1.39	Stable	1.19	Unstable	1.01
Stable	1.389	Stable	1.173	Unstable	1.003
Stable	1.356	Unstable	1.16	Unstable	0.98
Stable	1.351	Unstable	1.13	Unstable	0.96
Stable	1.337	Unstable	1.112	Stable	0.92
Stable	1.315	Unstable	1.11	Unstable	0.89
Stable	1.31	Unstable	1.1	Unstable	0.86
Stable	1.27	Unstable	1.08	Unstable	0.8
Stable	1.25	Stable	1.082	Unstable	0.480

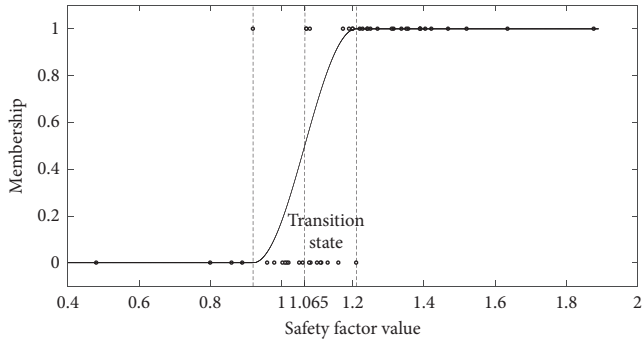


Figure 5: Membership function graph of safety factor.

$$\begin{cases} \pi_1 = \{1\}, \\ \pi_2 = [0.2806 + 0.7194a, 1], \\ \pi_3 = [0.7194 - 0.7194a, 0.2806 + 0.7294a], \\ \pi_4 = [0, 0.7194 - 0.7194a], \\ \pi_5 = \{0\}. \end{cases} \tag{11}$$

The safety factor interval corresponding to the fuzzy language value of the dump slope stability can be obtained by substituting each end value of the membership interval into equation (12). Table 4 lists the statistical results of the dump sites' samples, the membership degree interval, and the safety factor interval, where n_i represents the number of slopes in the dump in a stable state in the π_i safety factor interval and N_i is the number of slope instances in the corresponding safety factor interval. If the membership function is selected reasonably, the value of n_i/N_i increases to 1 as the membership interval approaches 1:

$$F_s = \frac{0.29 * \arcsin 2\left(\mu\left(F_s\right) - 0.5\right)}{\pi} + 1.065. \tag{12}$$

It can be seen from Table 4 that the n_i/N_i value is in good agreement with the membership interval corresponding to the ith fuzzy language value. The closer it is to the steady state, the closer its value is to 1. Therefore, it can be considered that the slope stability membership function proposed in this paper according to formula (10) can reflect the overall law of the slope stability of the dump site with the change of the safety factor, which can be used to improve the criterion of the subsequent Monte Carlo calculation state function.

The flowchart of specific application steps of methods is shown in Figure 6.

4. Application

4.1. Bootstrap Statistical Uncertainty Simulation. This paper used MATLAB to perform bootstrap sampling on the original sample and selected $N_B = 10^4$ as the number of subsamples (also the number of samples) based on ensuring good convergence and computational efficiency. Box-and-whisker plots of 20 randomly selected bootstrap subsamples and the original sample are shown in Figure 7. It can be seen that the generated subsamples perfectly retain the data information of the original sample.

For this procedure, 10,000 simulations generated 10,000 subsamples, corresponding to the mean, standard deviation, and AIC value of 10,000 simulated samples. Figures 8 and 9 and Table 5 show the distribution of the subsamples' mean and standard deviation, and the number of times that the four common probability density functions are identified as the optimal distribution. It is not difficult to see that except for bedrock cohesion, for the other indicators, the distribution of the mean and standard deviation of the bootstrap subsamples is close to that of the original sample, and the optimal probability distribution function of the subsamples obtained by the bootstrap method is mostly the same as the original sample, which is normal distribution, only a few of the optimal probability distributions of the subsample are the other indicators. This shows that the bootstrap subsamples can effectively reflect the basic characteristics of the original sample data, which means that the original limited data sample has statistical uncertainty and AIC value variability. This problem may affect the selection of subsequent

Table 4: Comparison of safety factor interval and statistical law of actual dump.

Fuzzy linguistic value	$a = 0.51$		
	Membership interval	Safety factor interval	n_i/N_i
Stable	{1}	$[1.21, +\infty)$	22/22
Basically stable	[0.653, 1]	[1.093, 1.21)	4/9
Critical	[0.316, 0.653]	[1.03, 1.093)	3/7
Basically unstable	[0, 0.316]	[0.92, 1.03)	1/6
Unstable	{0}	[0, 0.92)	0/4

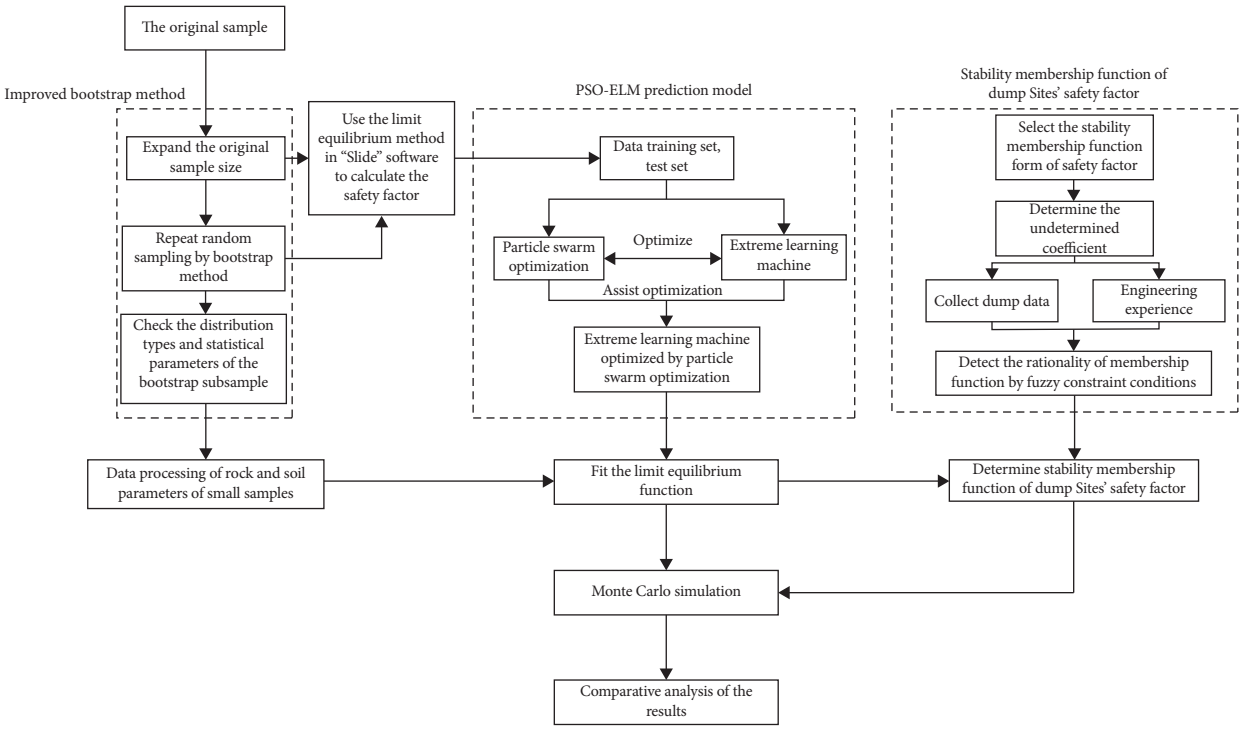


Figure 6: Operation flowchart.

distribution parameters and probability distribution function types, and the bootstrap method can simulate such variability.

4.2. Realizing the PSO-ELM-MCS Model in MATLAB. This paper uses the Janbu simplified method to calculate the safety factor of the sample to form the training and test sets and train the PSO-ELM model to achieve the purpose of replacing the original limit equilibrium function. The specific steps are as follows:

Step 1: determine training and test sets. In addition to the 34 sets of original samples, 44 sets of samples randomly selected from the bootstrap subsamples were substituted in the Slide software to calculate the safety factor to form 70 sets of training samples and 8 sets of test samples.

Step 2: set initial parameters. Six rock and soil parameters corresponded to one safety factor value in each group of samples, so there were six input layer units of the PSO-ELM model and one output layer unit. If the number of hidden layer units is too high, it will lead to excess performance, and if the number is too low, it will affect the prediction accuracy. After many calculations, we decided to set the number to 25. The learning factor $C_1 = C_2 = 1.49$, the sigma function was selected as the excitation function, and the linearly decreasing weight that can account for both local and global search was selected as the inertial weight.

Step 3: train and test models. The 70 sets of training samples obtained in Step 1 were combined with the PSO model to start training, and the weights and thresholds obtained after training were assigned to the ELM model, then the test samples were substituted into

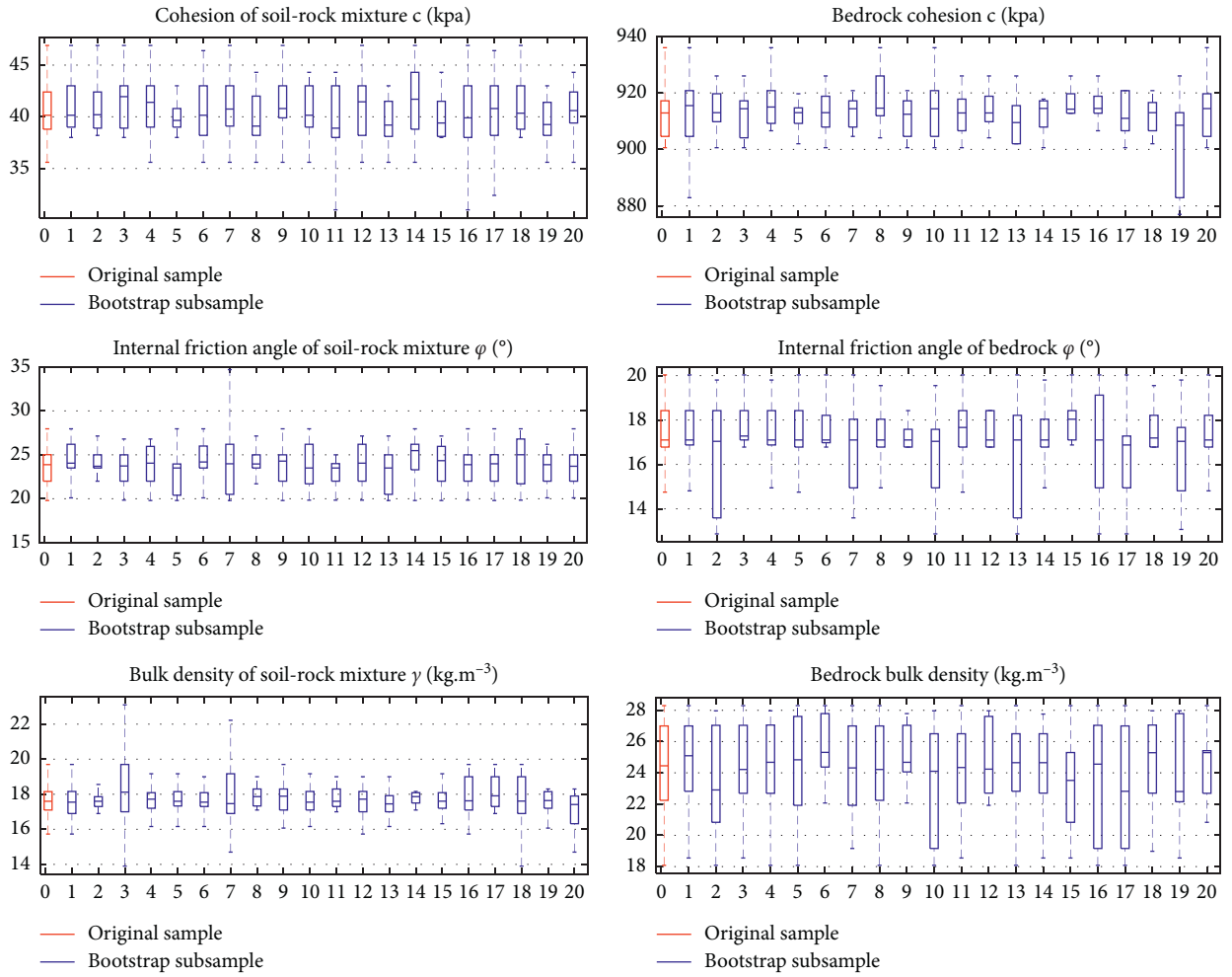


Figure 7: Box-and-whisker diagrams of original sample and bootstrap subsample data.

the predicted value of the safety factor to test the accuracy.

Figure 10 shows a comparison between the predicted and real values of the PSO-ELM model, as well as the predicted value of ordinary ELM. The predicted values of the safety factor obtained by the two models were not much different from the true value, but the PSO-ELM model was better than the ELM model in terms of the root mean square error (RMSE) of prediction accuracy or the coefficient of determination (R^2), which means particle swarm algorithm optimization of extreme learning machine does improve accuracy. The reliability analysis result obtained by running the PSO-ELM-MCS method in MATLAB was as follows: probability of failure $P_f = 15.02\%$, reliability index RI = 1.0356, and average safety factor 1.0454, which are highly consistent with the results obtained by the Slide software in the previous paper, indicating that the PSO-ELM model can perfectly fit the original Monte Carlo method to express the functional of the relationship between soil parameters and safety factors and reliability indicators. Meanwhile, the time to obtain the safety factor by the limit equilibrium method and the PSO-ELM model in MATLAB was statistically compared, and it was found that the efficiency of the latter was significantly better. The PSO-ELM model only required 0.0741 s, while the limit equilibrium method took 12.03 s.

4.3. The Analysis Result of MCS Method after Adding Membership Function. Figure 11 shows the calculation results of the improved PSO-ELM-MCS model after adding the slope stability membership function as the Monte Carlo calculation stable state criterion and expresses the relationship between the safety factor value and the membership degree. Comparing the probability of failure $P_f = 36.85\%$ and reliability index $\beta = 0.34$ with the calculation results of the PSO-ELM-MCS model ($P_f = 15.02\%$, $\beta = 1.0356$), it can be found that the failure probability of the dump site is large improved, and reliability is significantly decreased. Comparing the reliability calculation results obtained by different limit equilibrium methods (shown in Table 6), it is found that probability of failure and reliability indicators maintain good consistency under different limit equilibrium methods, which is obviously different from the previous results obtained by using ordinary Monte Carlo reliability analysis

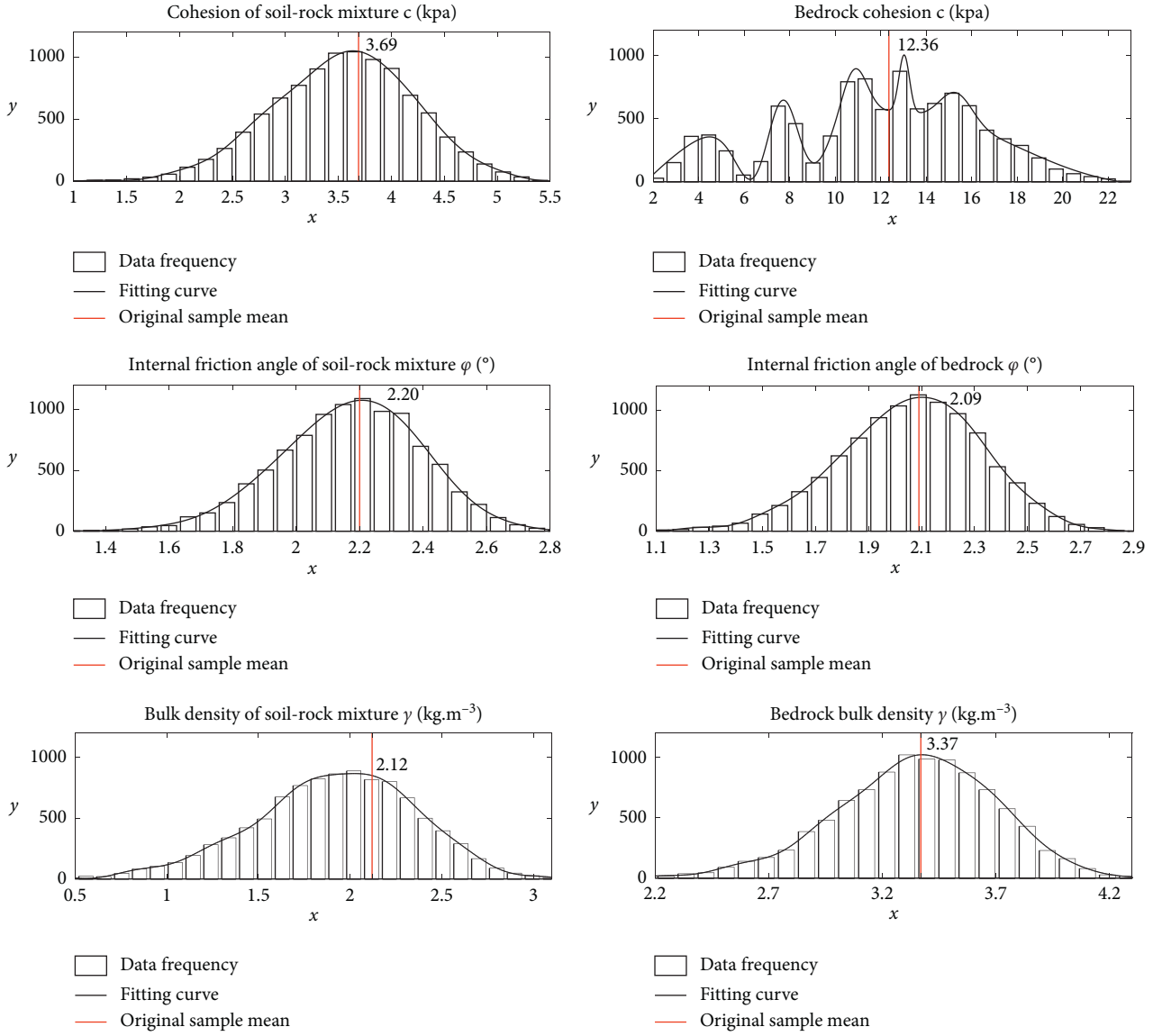


Figure 8: Standard deviation distribution of bootstrap subsamples.

(the relative error dropped from more than 90% to about 3%). That is to say, the results of Monte Carlo calculation with the stability membership function of the safety factor added are closer to reality than the traditional Monte Carlo reliability calculation, and it will not have the illusion of stability for simulators and engineers.

4.4. Monte Carlo Calculation Combined with PSO-ELM, Bootstrap Method, and Fuzzy Membership Function. This paper calculated the probability of failure $P_{f,i}$ and reliability index β_i $(i = 1, 2, 3, \ldots, N_B)$ of each bootstrap subsample and obtained the distribution probability density function and confidence interval (using the reliability index and the 5% and 95% quantile values of failure probability as the upper and lower limits of the 90% two-sided confidence interval), and the results as shown in Figure 12 and Table 7.

It can be seen that the reliability index of the dump site basically varies between −0.5 and 1, and the probability is the highest near 0.25, with an average value of 0.243, lower than the average value of 0.34 without bootstrap sampling. The probability of failure varies between 0.1 and 0.8, with the highest probability near 0.4, and the average value is 0.4075, slightly higher than the 0.3664 without bootstrap sampling. This shows that the reliability index and failure probability obtained by the ordinary MCS method may have variability, and the bootstrap method can simulate this variability. In addition, compared to the MCS method, which generally only obtains a single reliability index and failure probability value, the MCS method combined with bootstrap can express the reliability index and probability of failure as a two-sided confidence interval with a certain confidence level, which can reflect the actual reliability level of the dump slope.

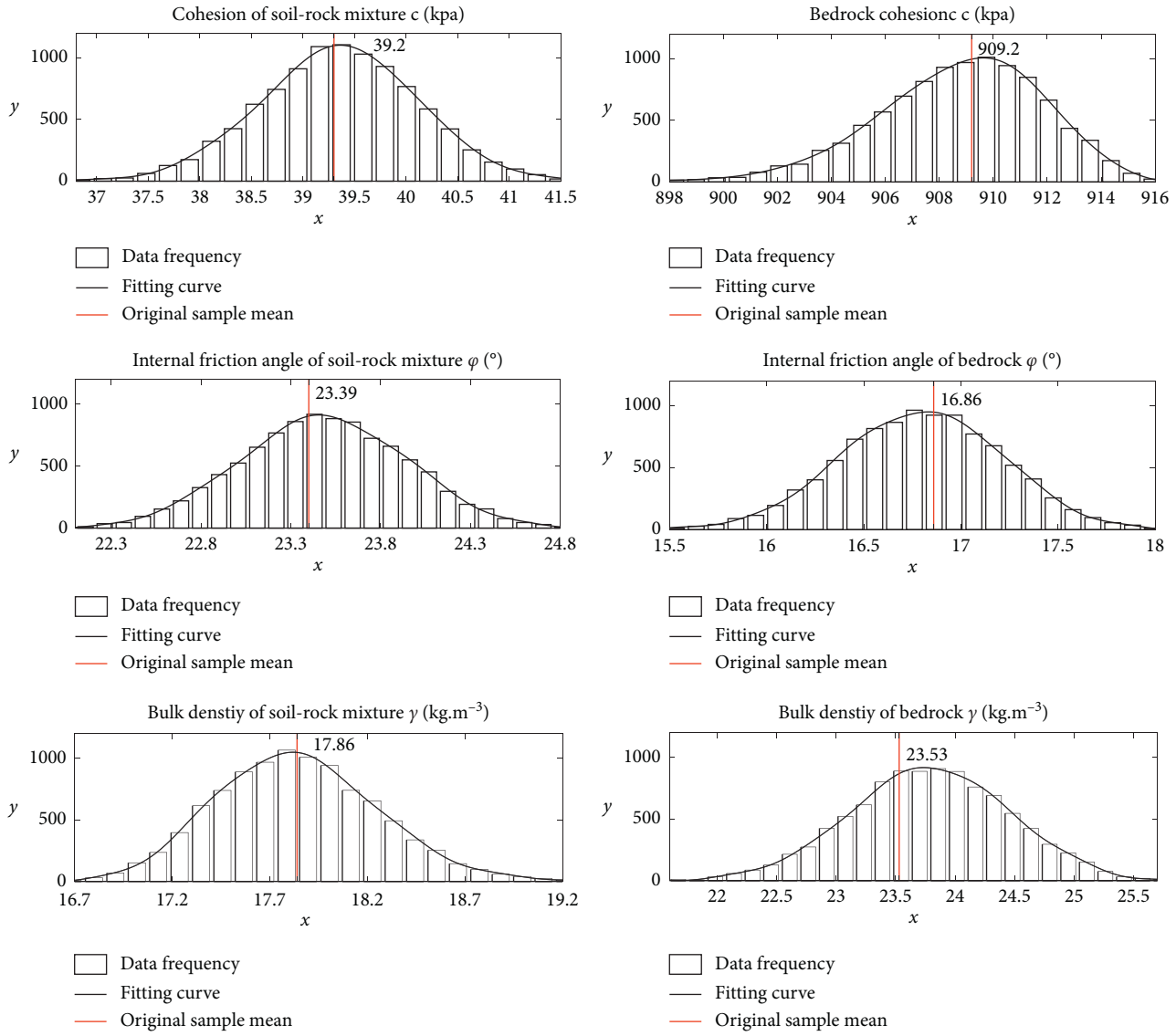


Figure 9: Mean distribution of bootstrap subsamples.

Table 5: Number of times that four common probability density functions are identified as optimal distribution.

Distribution type	Cohesion of soil-rock mixture (c (kPa))	Internal friction angle of soil-rock mixture (φ/(°))	Bulk density of soil-rock mixture (γ/(kN·m^{-3}))	Bedrock cohesion (c (kPa))	Internal friction angle of bedrock (φ/(°))	Bedrock bulk density (γ/(kN·m^{-3}))
Normal	9837	9918	9983	9107	9896	9943
Lognormal	163	82	17	625	104	57
Extreme value I	0	0	0	113	0	0
Wilber	0	0	0	155	0	0

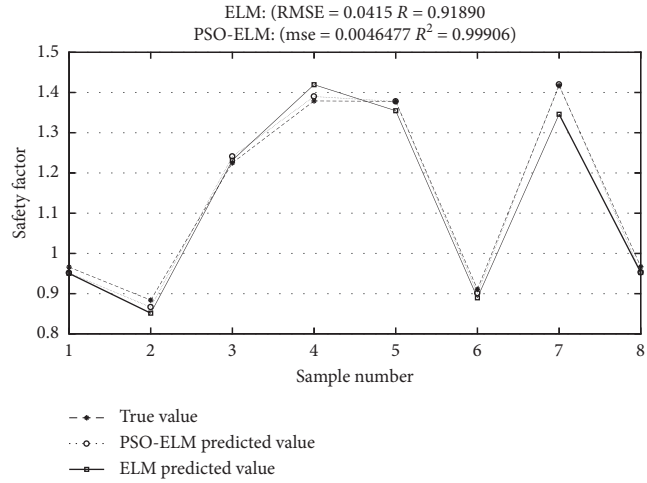


Figure 10: Comparison of prediction results of safety factor (true value and PSO-ELM and ELM).

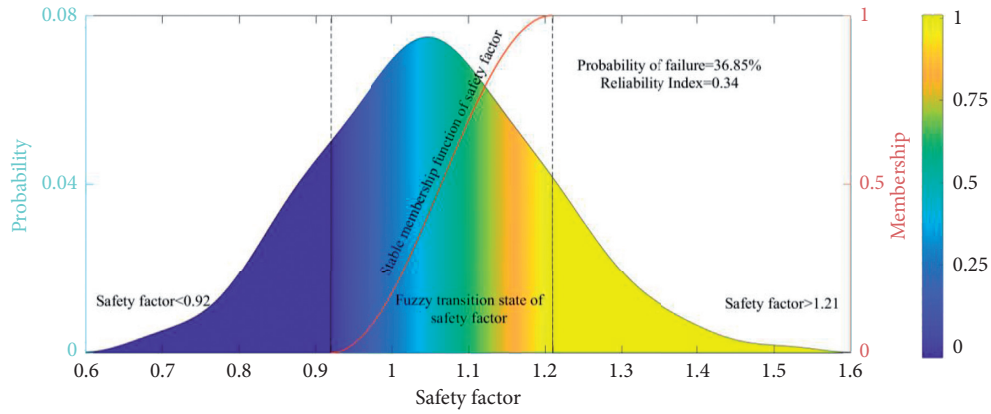


Figure 11: Monte Carlo calculation results after adding safety coefficient stable membership function.

Table 6: Reliability calculation results of different limit equilibrium methods.

Type of limit equilibrium method		Janbu simplified	Ordinary (Fellenius)	Bishop simplified	Spencer
Original MCS method	RI	1.035	1.288	2.266	2.246
	Pf	15.02%	9.74%	1.14%	1.3%
MCS method with stable membership function of safety factor added	RI	0.34	0.4067	0.4272	0.4097
	Pf	36.85%	34.21%	33.46%	34.1%

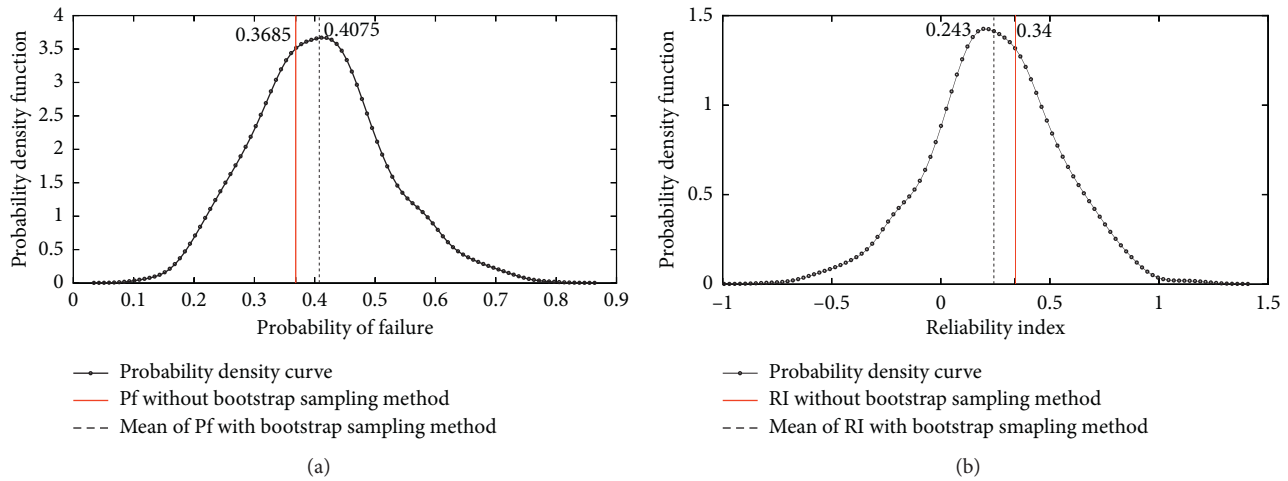


Figure 12: Probability density function of (a) failure probability P_f and (b) reliability index β of the dump site.

Table 7: Bootstrap method calculation results.

Probability of failure P_f			Reliability index β		
Mean	90% confidence interval of mean	Standard deviation	Mean	90% confidence interval of mean	Standard deviation
40.75%	[41.43%, 40.08%]	0.108	0.243	[0.2249, 0.2611]	0.291

5. Discussion and Conclusion

This paper proposed an improved Monte Carlo method that uses improved bootstrap methods to process small samples of geotechnical data, applies particle swarm optimization (PSO)-optimized extreme learning machine (ELM) to fit limit equilibrium method function, and constructs the safety factor membership function of the dump site considering the fuzzy transition interval, the results of each method combination are compared and analyzed, and the conclusions are as follows:

The MCS method using the PSO-ELM model to replace the original functional function can greatly improve efficiency on the basis of ensuring accuracy. The time consumed changed from the original 63005.44 s to 45.312 s, and the probability of failure (15.25% and 15.02%) and the reliability index (1.027 and 1.036) are almost the same, which proves that the artificial intelligence and neural network algorithms have good self-learning abilities and are effective at fitting implicit functional functions.

After the safety factor membership function of the dump site was added to the Monte Carlo method, the fluctuation of reliability index and probability of failure of the Nanpo dump site under different limit equilibrium methods was effectively improved, and the results showed an inclination toward slope instability (the failure probability of 36.64% and reliability index of 0.34 are significantly lower than the 15.02% and 1.036 without the membership function), which is closer to the situation at the actual dump site. This shows that the addition of a fuzzy transition zone of safety factor can take into account the uncertainty and gradualness of the slope in the process of failure, reduce the fluctuation of calculation results to a certain extent, and make it more real and effective.

Although the MCS method combined with the bootstrap method has a lot of repeated calculation work, the time changed from 44.519 s to 20275.23 s, but time was exchanged for accuracy, and the bootstrap subsample was generated to simulate the uncertainty brought by the small sample size of the experimental data (uncertainty of sample mean, standard deviation, optimal probability density distribution, reliability index, and failure probability), making the reliability calculation result change from a single value to a confidence interval with fuzzy characteristics, which is closer to reality. The average values of reliability analysis obtained by the improved Monte Carlo method based on bootstrap in the second-phase dump site of the mine are $P_f = 40.75\%$ and $\beta = 0.243$, which indicates a hazardous situation that requires immediate protective measures. The dump site immediately implemented preventive measures of antislip steel rail piles plus spoil grading and classified discharge to avoid the aggravation of displacement and the occurrence of landslides.

On summary, the method proposed in this paper can be used in the reliability analysis of the dump, and the effect is better than the traditional MCS method.

6. Limitation

However, the methods and research in this paper still have some shortcomings that could be improved. Such as the calculation process of MCS is not simplified and still consumes a lot of time, and the fuzzy membership function of the dump safety coefficient determined in this paper may still be subjective because of insufficient data collection and requires a professional authority to collect thousands of waste dump examples to reduce uncertainty. Follow-up studies will be carried out to improve the above deficiencies.

Data Availability

The soil and rock data used to support the findings of this study have been deposited in the (Research on Key Technology of Stability and Safety Control of Large Dump in Open-Pit Mine) repository (Wang L, Research on Key Technologies of Stability and Safety Control of Large Open-Pit Mine Dump, 2015, University of Science and Technology Beijing).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Wang, *Research on Key Technologies of Stability and Safety Control of Large Open-Pit Mine Dump*, University of Science and Technology, Beijing, China, 2015.

[2] R. Suchomel and D. Marin, "Comparison of different probabilistic methods for predicting stability of a slope in spatially variable c soil," *Computers & Geotechnics*, vol. 37, no. 1, pp. 132–140, 2010.

[3] B. K. Low and W. H. Tang, "Reliability analysis using object-oriented constrained optimization," *Structural Safety*, vol. 26, no. 1, pp. 69–89, 2004.

[4] J. Ji, "A simplified approach for modeling spatial variability of undrained shear strength in out-plane failure mode of earth embankment," *Engineering Geology*, vol. 183, pp. 315–323, 2014.

[5] B. K. Low and W. H. Tang, "Probabilistic slope analysis using janbu's generalized procedure of slices," *Computers and Geotechnics*, vol. 21, no. 2, pp. 121–142, 1997.

[6] B. K. Low, "Reliability analysis of rock slopes involving correlated nonnormals," *International Journal of Rock Mechanics and Mining Sciences*, vol. 44, no. 6, pp. 922–935, 2007.

[7] P. Zeng and R. Jimenez, "An approximation to the reliability of series geotechnical systems using a linearization approach," *Computers and Geotechnics*, vol. 62, pp. 304–309, 2014.

[8] B. K. Low, "FORM, SORM, and spatial modeling in geotechnical engineering," *Structural Safety*, vol. 49, pp. 56–64, 2014.

[9] S. E. Cho, "Effects of spatial variability of soil properties on slope stability," *Engineering Geology*, vol. 92, no. 3, pp. 97–109, 2007.

[10] S. E. Cho, "Probabilistic assessment of slope stability that considers the spatial variability of soil properties," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 136, no. 7, pp. 975–984, 2010.

[11] S. H. Jiang, "Efficient system reliability analysis of slope stability in spatially variable soils using Monte Carlo simulation," *Journal of Geotechnical & Geoenvironmental Engineering*, vol. 141, no. 2, Article ID 4014096, 2015.

[12] H. El-Ramly, N. R. Morgenstern, and D. M. Cruden, "Probabilistic slope stability analysis for practice," *Canadian Geotechnical Journal*, vol. 39, no. 3, pp. 665–683, 2002.

[13] H. El-Ramly, N. R. Morgenstern, and D. M. Cruden, "Probabilistic assessment of stability of a cut slope in residual soil," *Géotechnique*, vol. 55, no. 1, pp. 77–84, 2005.

[14] J. Huang, D. V. Griffiths, and G. A. Fenton, "System reliability of slopes by RFEM," *Soils and Foundations*, vol. 50, no. 3, pp. 343–353, 2010.

[15] F. S. Wong, "Slope reliability and response surface method," *Journal of Geotechnical Engineering*, vol. 111, no. 1, pp. 32–53, 1985.

[16] B. Xu and B. K. Low, "Probabilistic stability analyses of embankments based on finite-element method," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 132, no. 11, pp. 1444–1454, 2006.

[17] J. Zhang, L. M. Zhang, and W. H. Tang, "New methods for system reliability analysis of soil slopes," *Canadian Geotechnical Journal*, vol. 48, no. 7, pp. 1138–1148, 2011.

[18] J. Zhang, L. M. Zhang, and W. H. Tang, "Kriging numerical models for geotechnical reliability analysis," *Soils and Foundations*, vol. 51, no. 6, pp. 1169–1177, 2011.

[19] J. Zhang, H. W. Huang, C. H. Juang, and D. Q. Li, "Extension of hassan and wolff method for system reliability analysis of soil slopes," *Engineering Geology*, vol. 160, pp. 81–88, 2013.

[20] J. Zhang, H. W. Huang, and K. K. Phoon, "Application of the kriging-based response surface method to the system reliability of soil slopes," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 139, no. 4, pp. 651–655, 2013.

[21] D. Q. Li, "Reliability analysis of slopes considering spatial variability of soil parameters using non-intrusive stochastic finite element method," *Journal of Geotechnical Engineering*, vol. 35, no. 8, pp. 1413–1422, 2013.

[22] S.-H. Jiang, D.-Q. Li, L.-M. Zhang, and C.-B. Zhou, "Slope reliability analysis considering spatially variable shear strength parameters using a non-intrusive stochastic finite element method," *Engineering Geology*, vol. 168, pp. 120–128, 2014.

[23] D.-Q. Li, S.-H. Jiang, Z.-J. Cao, W. Zhou, C.-B. Zhou, and L.-M. Zhang, "A multiple response-surface method for slope reliability analysis considering spatial variability of soil properties," *Engineering Geology*, vol. 187, pp. 60–72, 2015.

[24] T. T. He, "Support vector machine method for slope reliability analysis," *Rock and Soil Mechanics*, vol. 11, pp. 232–239, 2013.

[25] X.-H. Tan, M.-F. Shen, X.-L. Hou, D. Li, and N. Hu, "Response surface method of reliability analysis and its application in slope stability analysis," *Geotechnical and Geological Engineering*, vol. 31, no. 4, pp. 1011–1025, 2013.

[26] G. S. Su, "Gaussian process dynamic response surface method for slope failure probability estimation," *Rock and Soil Mechanics*, vol. 12, pp. 3592–3601, 2014.

[27] F. Kang and J. Li, "Artificial bee colony algorithm optimized support vector regression for system reliability analysis of slopes," *Journal of Computing in Civil Engineering*, vol. 30, no. 3, Article ID 4015040, 2016.

[28] S. E. Cho, "Probabilistic stability analyses of slopes using the ANN-based response surface," *Computers and Geotechnics*, vol. 36, no. 5, pp. 787–797, 2009.

[29] Y. H. Su and H. B. Yang, "Slope stability reliability algorithm based on proxy model," *Journal of Applied Mechanics*, vol. 29, no. 6, pp. 705–710, 2012.

[30] Y. H. Su, "Active search method for slope stability reliability based on kriging," *Journal of Geotechnical Engineering*, vol. 35, no. 10, pp. 1863–1869, 2013.

[31] P. Yi, K. Wei, X. Kong, and Z. Zhu, "Cumulative PSO-kriging model for slope reliability analysis," *Probabilistic Engineering Mechanics*, vol. 39, pp. 39–45, 2015.

[32] Y. Wang, "Study and application of vector projection response surface for slope reliability evaluation," *Journal of Geotechnical Engineering*, vol. 33, no. 9, pp. 1434–1439, 2011.

[33] G. Habibagahi and M. Meidani, "Reliability of slope stability analysis evaluated using a fuzzy set approach," in *Proceedings of the 5th International Conference on Civil Engineering*, Ferdowsi University, Singapore, 2000.

[34] W. Y. Xu, Z. M. Jiang, and A. C. Shi, "Slope stability analysis based on fuzzy set theory," *Journal of Geotechnical Engineering*, vol. 25, no. 4, pp. 409–413, 2003.

[35] H. H. Jia and H. J. He, "Fuzzy random reliability analysis of slope stability," *Rock and Soil Mechanics*, vol. 24, no. 4, pp. 657–660, 2003.

[36] F. Lou, *Fuzzy Random Reliability Analysis of Slope Safety in Open Pit Mines*, Central South University, Changsha, China, 2008.

[37] H. Xu, *Fuzzy Random Reliability Analysis of Slope Stability Based on Fuzzy Set Theory*, Zhejiang University, Hangzhou, China, 2006.

[38] M. Anvar and M. Bahrami, "Uncertainty analysis of safety factor of embankment built on stone column improved soft soil using fuzzy logic alpha-cut technique," *Computers and Geotechnics*, vol. 14, pp. 101–104, 2016.

[39] T. Most and T. Knabe, "Reliability analysis of the bearing failure problem considering uncertain stochastic parameters," *Computers and Geotechnics*, vol. 37, no. 3, pp. 299–310, 2010.

[40] Z. Luo, S. Atamturktur, and C. H. Juang, "Bootstrapping for characterizing the effect of uncertainty in sample statistics for braced excavations," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 139, no. 1, pp. 13–23, 2013.

[41] X. S. Tang, "Joint distribution model identification of rock and soil parameters based on Bootstrap method," *Rock and Soil Mechanics*, vol. 4, pp. 913–922, 2015.

[42] H.,H. Tao, "Introduction to the hydroelectric engineering standard system of the U.S. army corps of engineers," *Hong-Shui River*, vol. 2, pp. 98–101, 2010.

[43] J. Wang, *Opportunity-constrained Programming Analysis of Soil Slope Stability Based on Limit Equilibrium Method*, Huazhong University of Science and Technology, Wuhan, China, 2018.
[44] Z. D. Wang, Y. H. Li, and F. Yun, "Application of annealing genetic algorithm in slope stability analysis," *Highways*, vol. 5, pp. 11–13, 2008.
[45] H. Akaike, "IEEE xplore abstract—a new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
[46] B. Efron and, "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
[47] L. Jian, W. Yi, and T. Lu, "Improvement of bootstrap method for self-help sampling," *Mathematical Theory and Application*, vol. 1, pp. 69–72, 2006.
[48] F. Luo, "Bootstrap estimation of rock and soil parameters and slope stability analysis for small samples," *Journal of Rock Mechanics and Engineering*, vol. 36, no. 2, pp. 370–379, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8824388, 16 pages
https://doi.org/10.1155/2020/8824388

Research Article

Study on Foundation Pit Construction Cost Prediction Based on the Stacked Denoising Autoencoder

Lanjun Liu,[1,2] **Denghui Liu**,[3] **Han Wu**,[2] **and Junwu Wang**[2]

[1]*School of Civil Engineering and Architecture, Wuhan Institute of Technology, Wuhan 430070, China*
[2]*School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China*
[3]*China Construction First Group Corporation Limited, Beijing 100161, China*

Correspondence should be addressed to Han Wu; wuhan20170620@163.com

Received 28 September 2020; Revised 20 November 2020; Accepted 24 November 2020; Published 3 December 2020

Academic Editor: Jun Shen

To accurately predict the construction costs of foundation pit projects, a model based on the stacked denoising autoencoder (SDAE) is constructed in this work. The influencing factors of foundation pit project construction costs are identified from the four attributes of construction cost management, namely, engineering, the environment, the market, and management. Combined with Chinese national standards and the practice of foundation pit project management, a method of the quantization of the influencing factors is presented. 60 deep foundation pit projects in China are selected to obtain 13 main characteristic factors affecting these project construction cost by using the rough set. Then, considering the advantages of the SDAE in dealing with complex nonlinear problems, a prediction model of foundation pit project construction costs is created. Finally, this paper employs these 60 projects for a case analysis. The case study demonstrates that, compared with the actual construction costs, the calculation error of the proposed method is less than 3%, and the average error is only 1.54%. In addition, three error analysis tools commonly used in machine learning (the determination coefficient, root mean square error, and mean absolute error) emphasize that the calculation accuracy of the proposed method is notably higher than those of other methods (Chinese national code, the multivariate return method, the BP algorithm, the BP model optimized by the genetic algorithm, the support vector machine, and the RBF model). The relevant research results of this paper provide a useful reference for the prediction of the construction costs of foundation pit projects.

1. Introduction

For developing countries, construction engineering not only provides the necessary infrastructure for economic and social development but is also the main driving force of GDP growth [1]. The capital investment of construction projects is large, and the investment recovery period is long. At present, there is a substantial difference between the predicted and actual construction costs in the construction industry, which hinders the efficient development of the entire industry. Around the world, a large number of construction projects have failed due to the cost overruns [2–4]. Accurate prediction of construction cost in the early stage is one of the main bases for project decision-making and investment cost control, and its research is of great significance.

Measures such as retaining structure, groundwater control, and environmental protection needed to ensure the safety and stability of underground space formed by underground excavation are called foundation pit engineering. Foundation pit projects are important subprojects of construction engineering that are characterized by a long construction period and many influencing factors with complex relationships [5]. The construction cost of foundation pit engineering accounts for a large proportion of the total cost, which is often between 20% and 50%. Compared with other partial projects, foundation pit projects are more likely to occur in the likelihood of large differences between the predicted and actual costs [6]. Therefore, it is of great significance to quickly and accurately predict the construction costs of foundation pit projects.

There are obvious differences in the definition of construction cost in different national codes and different construction contracts [7, 8]. In order to facilitate the follow-up research, the construction cost in this paper was the actual construction cost of the construction project. The research object of this paper, the foundation pit project construction cost, referred to all expenses incurred by the contractor during the construction of the foundation pit project. Total expenses generally include two parts, project productive expenses and enterprise nonproductive expenses [8]. Nonproductive expenses are the expenses incurred by construction enterprises for organizing and managing production and business activities, which are often closely related to the contractor's project management ability, while the project productive expenses are mainly the actual expenses incurred in the construction site of foundation pit engineering.

From the perspective of research objects, the related research is mainly focused on the prediction of construction engineering costs, but there has been little research on the prediction of the costs of foundation pit engineering or other construction subprojects. Forcada et al. [9] selected several important influencing factors, including the project type, organization type, and contract type, and predicted the rework cost of construction projects via the regression analysis method. Wang and Dai [10] selected only six indexes, including the building area, number of floors, building height, and type of roof, to effectively predict the construction cost of a 15-story building. However, they did not consider the relevant factors of project management, and the numbers of floors and building structures of the test set and sample set in the case analysis were basically the same. In Wang and Dai's case study, the construction cost of the project could be considered to have an approximately positive correlation with the building area, which was an important reason why the construction cost was accurately predicted via the use of only six indicators. Williams and Gong [11] carried out the text and data mining of project management materials to determine the main factors that affect construction costs. Although this method provided a new idea for the selection of indicators, the average calculation accuracy was found to be only 47.11%. The possible reason for this unideal result was that the main factors obtained by this method originated from project management data, but how these factors affect the project cost required more explanation and analysis. These scholars often selected only a few main influencing factors, but it is difficult for this strategy to reflect the complexity of construction projects.

In actual engineering practice, a construction project often includes several subprojects that are highly professional and varied [12]. Therefore, the basis of the scientific and effective prediction of construction costs is the scientific and effective prediction of the construction costs of divisional and subdivisional projects. To the best of the authors' knowledge, research on the prediction of foundation pit project construction costs based on deep learning has not yet been reported. Therefore, foundation pit engineering, which is a typical subproject, was selected as the research object in the present study.

The acquisition of key influencing factors of foundation pit project construction cost is another content that needs serious study [13]. By the means of questionnaire survey and subjective experience of experts, Lesniak and Juszczyk [14] determined that the project type, project geographical location, and construction period as key influencing factors and used the back propagation neural network (BP) to predict the indirect cost of the project. By the case-based reasoning and measurement similarity, Kim and Kim [15] determined the key influencing factors of construction cost. Dong et al. [16] interviewed eight experts in the field of engineering and construction and determined 16 indexes that affected the construction cost. Among the above typical research results, the selection of characteristic factors is mostly based on the empirical method, the proportional method, and questionnaire survey method. These traditional methods have the disadvantages of strong subjectivity and lack of scientificalness. In addition, the influence of the selection of characteristic factors on the prediction results has not been considered in the above studies. However, the variable selection method based on rough set has achieved good results in recent years, which could effectively overcome these shortcomings of traditional methods [17–20]. Su et al. [17] used the rough set method to effectively screen the key risk factors of subsea tunnel construction. Case study showed that the risk index system obtained by the rough set was more scientific than these traditional weight calculation methods. Zhang et al. [18] obtained the key factors and objective weights of landslide risks in mountain tunnel construction by using the rough set method. Xu et al. [19] used the variable precision fuzzy rough set (VPFRS) to screen the evaluation index system of the synergy effect of main and auxiliary industries in power grid. Research showed that rough set could eliminate redundant indicators and retain key indicators and effectively improve the efficiency and accuracy of evaluation. Barbagallo et al. [20] used the rough set method in the Rose Package to effectively identify and further screen out the index system for water supply reservoir management.

From the perspective of research methods, adopting appropriate methods to reflect the nonlinear relationship among influencing factors of foundation pit project construction cost and quickly and accurately predicting is the key to build an estimation model of foundation pit project construction cost. Because the influencing factors of building project construction cost are complex and the data collection of construction cost is not easy, the prediction of building project construction cost is a typical high-dimensional nonlinear problem. Moreover, the rapid prediction of building project construction cost often serves the cost management of enterprises and requires higher accuracy and time of prediction. At present, quota method is often used in engineering practice, such as Chinese National Standard (Standard method of measurement for public utilities works, GB 50857-2013; Standard for classification and measurement of construction cost index, GB/T 51290-2018; Code of bills of quantities and valuation for construction works, GB 50500-2013) and the construction contract of International Federation of Consulting

Engineers (FIDIC). Relevant researchers have already proposed some mathematical methods by which to predict the costs of construction projects. Trost and Oberlender [21] conducted a multiple regression analysis of 11 factors to predict project costs. Considering the poor calculation accuracy of multiple regression analysis, a score estimation program was also developed to evaluate the prediction accuracy. To accurately estimate the costs of construction projects, Ji and Ahn [22] proposed a prediction method based on the scenario-planning method. The results of their case study revealed that the estimation accuracy was between 4.23% and 4.86%. Cheng et al. [23] proposed a cost forecasting method based on the grey prediction model. The construction process was divided into three typical stages according to the cost prediction accuracy to take into account the complexity of construction cost prediction. Although these cost prediction methods implemented in these studies achieved certain results, it was difficult for them to scientifically and effectively deal with small sample numbers and nonlinear problems. In addition, these studies all assumed that the prices of labor, materials, and machines were static factors that did not change with time; thus, there was a substantial gap with actual engineering practice, in which the price-related indicators are dynamic. Therefore, sufficient prediction could not be achieved. Recently, some scholars have also applied artificial neural networks (ANNs) to construction cost prediction. Wang [24] divided construction engineering costs into three categories, namely, construction costs, structure costs, and outdoor engineering costs and used the BP network to predict the costs of construction projects. The results of a case study showed that, although the calculation error of this algorithm was large, it met the requirements of engineering practice. Gunduz et al. [25], respectively, used multiple regression analysis and the BP network, which is the most common ANN method, to predict the early costs of light rail transit and subway engineering development. In their research, 17 key factors that affect costs were selected, and 16 project datasets were selected as sample sets. The error of the multiple regression analysis was 2.32%, which was notably less than the error of the ANN (5.76%). The reason that the BP calculation accuracy of [24, 25] was not high, might be that the traditional BP model easily falls into the local extremum and has a diverse network structure [26].

Deep learning was developed from the traditional multilayer neural network, which has excellent nonlinear mapping and generalization abilities to represent complex high-dimensional functions [27]. The difference between the traditional multilayer neural network and deep learning mainly lies in the different training methods. Traditional ANNs are trained by supervised learning, while most deep learning algorithms combine unsupervised feature learning with supervised learning [28]. The denoising autoencoder (DAE) is a common deep learning network [29] that has two main forms. One is the stacked denoising autoencoder (SDAE), which has obvious advantages in dealing with high-dimensional nonlinear problems, and the other is the sparse encoder. When a neural network is used to classify and predict a large number of influencing factors of construction costs, a lot of noise is easily produced due to the data of the influencing factors. To avoid the interference of noisy data as much as possible [30], the SDAE was selected in the present research to study the prediction of construction engineering costs.

Liu et al. [31] accurately predicted short-term power load by the SDAE method. The case study showed that SDAE had higher calculation accuracy than the BP and the DAE. Dai et al. [32] applied the SDAE model to data processing of dissolved gas concentration in transformer oil and transmission line temperature. The results showed that this method could effectively identify and repair outliers and missing information. However, the performance of SDAE algorithm was not further analysed in that paper. Dong et al. [33] used the SDAE to predict the short-term wind speed in wind power generation. The results of case study showed that SDAE had higher computational accuracy and faster computational efficiency than the artificial intelligence algorithms such as the BP. Chen et al. [34] extracted hyperspectral image features by the SDAE method. Compared with common methods such as the support vector machine (SVM), SDAE had better computing power. In order to deal with the variability and nonlinear correlation in the prediction of regional sharp power generation, Yan et al. [35] used the SDAE method to predict the regional wind power generation. Compared with other common wind power generation forecasting methods, SDAE had the highest forecasting authenticity.

By summarizing the existing research work, the following key issues deserve study in this paper. (1) Most of the current research results took architectural project construction cost as the research object. The research object was not meticulous enough, which led to the rough selection of influencing factors and large prediction error. It is a possible way to improve the construction cost prediction by selecting some subprojects such as foundation pit engineering for construction cost prediction. (2) In the current related research, the selection of characteristic factors was mostly based on the empirical method or the proportional method, and these selection methods were subjective and unscientific. These studies also did not consider the influence of the selection of characteristic factors on the selection of forecasting methods. (3) At present, the regression analysis, the grey theory, and the traditional artificial neural network are often used to predict the construction cost, but their prediction accuracies were not high and they took a long time. The SDAE based on deep learning network has strong learning and prediction ability. Training and learning sample data by establishing prediction model provides a new idea for accurately and quickly forecasting the construction cost of nonlinear engineering projects.

Based on the above analysis, this paper constructed the construction cost prediction model of foundation pit engineering by the SDAE. The main contributions of this paper are as follows. (1) According to the characteristics of the construction content and project management of foundation pit engineering, a more comprehensive system of the determination of influencing factors is constructed, and a corresponding quantification and normalization process is

put forward. This provides a research basis for the subsequent construction cost prediction of foundation pit engineering. (2) In order to overcome the shortcomings of strong subjectivity and lack of scientificity in traditional index selection methods, this paper uses rough set, a typical quantitative analysis method, to obtain the main characteristic factors affecting the project construction cost of foundation pit. In this paper, 60 deep foundation pit projects in Hubei Province of China are selected as cases, and 13 main characteristic factors affecting these project construction cost are screened out. In addition, the influence of key factors on prediction results is also preliminarily analysed. (3) The SDAE based on a deep learning network is selected, and a prediction model is established to train and learn the sample data to accurately predict the costs of foundation pit engineering projects with complex nonlinear characteristics. (4) A case study demonstrates that the proposed calculation model has good calculation accuracy. Compared with the Chinese national code, the BP, GA-BP, SVM, RBF, and multivariate return models, the calculation accuracy is higher and the prediction results are more stable.

The organizational structure of the remainder of this paper is as follows. Section 2 introduces the research materials and methods in detail, including the analysis of the influencing factors of the construction costs of foundation pit engineering and the prediction model based on the SDAE. Section 3 presents a case analysis, and the model and case analysis are discussed in Section 4. Finally, Section 5 presents the research conclusions and further research prospects.

2. Materials and Methods

2.1. Influencing Factors of Foundation Pit Project Construction Costs

2.1.1. Analysis of the Influencing Factors of Foundation Pit Project Construction Costs. There are many factors that affect the construction costs of foundation pit projects, and these factors can generally be divided into four categories, namely, factors related to engineering characteristics, the environment, the market, and management [36, 37].

The factors related to engineering characteristics reflect the structural characteristics of foundation pit engineering itself, including the building area, the type of pile foundation, the depth of the foundation pit, and the form of the foundation structure. Environment-related factors reflect the construction site environment, including the construction site conditions, the availability of construction water and electricity, and the difficulty of earthwork excavation.

The factors related to the market are the prices of materials, such as construction machinery, construction personnel, concrete, and steel bars [38]. To reflect the market fluctuations of these prices, these factors were quantified in the present research by the project cost index. The project cost index is the ratio that reflects the degree of change of the project cost in a certain period relative to the project cost in a certain fixed period. It reflects the changing trend of the market price in the current period relative to the market price in the base period. With reference to China's national standard ("Standard for classification and measurement of construction cost index," GB/T 51290-2018), the calculation method of the project cost index is as follows:

$$A = \frac{P_a}{P_j} \times 100, \tag{1}$$

where A is the cost index, P_a is the current cost index, and P_j is the reference period cost index. The cost index of the base period is 100.

The factors related to management mainly reflect the management level of the contractor of the foundation pit project [39]. Owners, design units, testing units, and equipment suppliers were not taken into consideration in this study, as they have little influence on the construction costs of a foundation pit project without major changes or engineering accidents. In addition, the construction period, one of the three goals of the construction project, directly reflects the management achievements [40].

2.1.2. Selection and Quantification of Influencing Factors. The factors that affect the construction costs of foundation pit projects can be divided into two categories, namely, quantitative and qualitative factors. In the present research, the data acquisition method of quantitative factors included field investigation and the consultation of project management data, while data on the qualitative factors were obtained by questionnaires [41].

According to the analysis results presented in Section 2.1.1, the quantification of the influencing factors used in this paper is as follows.

(1) The depth of the foundation pit: the depth of the foundation pit has the most direct influence on the design and construction of foundation pit engineering. This index is a quantitative index, and its unit is meters (m).

(2) The form of the foundation pit support: there are 8 common forms of foundation pit support, namely, the row pile support (1), the underground diaphragm wall support (2), the cement retaining wall (3), the soil nailing wall (4), the arch wall constructed by the reverse method (5), the undisturbed soil slope (6), the reinforced concrete row pile (7), and other foundation pit support forms (8). The numbers in parentheses reflect the respective scores of these forms of support.

(3) The form of the infrastructure: according to the stress characteristics, there are five types of commonly used foundation structures, namely, the beam foundation, strip foundation, raft foundation, box foundation, and pile foundation structures. The dimensionless index for the beam foundation structure is 1. Similarly, the indexes of the strip, raft, box, and pile foundation structures are 2, 3, 4, and 5, respectively.

(4) The type of pile foundation: a pile foundation is a deep foundation composed of multiple piles and pile caps connecting the tops of the piles, or a single pile foundation connected by columns and piles. The selection of the pile foundation has a great influence on the construction costs of foundation pit and building engineering projects. Common pile foundations mainly include prefabricated pipe piles, rotary bored piles, manually excavated piles, punched piles, various pile types, and nonengineering piles [42], the data scores of which are 1, 2, 3, 4, 5, and 6, respectively. The reason for the introduction of nonengineering piles is that the data for the pile foundation type cannot be filled in if the beam or raft foundation is adopted.

(5) The quantity of the pile foundation: according to the Chinese national code ("Code of bills of quantities and valuation for construction works," GB 50500-2013), the engineering quantity of a pile foundation is mainly subject to the designed engineering quantity of concrete pouring. The engineering quantity of a pile foundation is a quantitative index, and its unit is m^3.

It is important to note that, in engineering practice, the pile foundation engineering quantity of a beam foundation or raft foundation is 0. To avoid the presence of "0″ in the unsupervised learning stage, the pile foundation engineering quantities of 0 in the actual collected data were changed to 0.01.

(6) The engineering geological conditions: the geological factors that affect building engineering mainly include topography, stratum lithology, geological structures, earthquakes, hydrogeology, natural building materials, and unfavorable physical and geological phenomena such as karst, landslide, collapse, sand liquefaction, and foundation deformation. According to China's national code ("Code of bills of quantities and valuation for construction works," GB 50500-2013), the difficulty of Earth and rock excavation is the most obvious embodiment of engineering geological conditions and has the most direct impact on the engineering cost. Therefore, in this work, the difficulty of earthwork excavation is divided into three levels, namely, very difficult, difficult, and general, and the dimensionless qualitative indexes of which are, respectively, 1, 2, and 3.

(7) The construction area of the foundation pit: this is a quantitative index that can be calculated according to the design drawings, and its unit is m^2.

(8) The on-site construction conditions: site enclosure equipment, material stacking, temporary facilities, site water supply and drainage, temporary electricity utilization, etc., have significant impacts on the smooth progress of construction. In general, when the on-site construction conditions are favorable for the normal construction process, the construction costs will decrease. This indicator is qualitative, and there are three situations, namely, complete compliance, basic compliance, and temporary noncompliance, the dimensionless indexes of which are 1, 2, and 3, respectively.

(9) The meteorological characteristics: meteorological characteristics mainly refer to the influence of meteorology on the construction progress during the peak construction period of foundation pit projects. These characteristics can be divided into three situations, namely, those that have large, small, and no influence on the construction. The dimensionless indexes of these situations are 1, 2, and 3, respectively.

(10) The off-site traffic conditions: because the construction sites of foundation pit engineering are generally located in city centers, the traffic conditions around the construction site have a certain influence on the construction progress and transportation costs of the project site. This is a qualitative indicator, and experts were invited to comprehensively evaluate the traffic flow in and out of the construction site, the road conditions, and the transportation routes. There are three kinds of evaluation results, which have great, small, and no influence on the construction progression. The dimensionless indexes of these results are 1, 2, and 3, respectively.

(11) The labor cost index: this is a quantitative index. After selecting a suitable reference period, it is calculated with reference to equation (1).

(12) The material cost index: the commonly used and expensive materials in foundation pit engineering are steel bars and concrete. Therefore, the steel bar cost index and concrete cost index are, respectively, introduced. After selecting a suitable reference period, they are calculated with reference to equation (1).

(13) The machinery cost index: there are many machines used in foundation pit engineering, but the engineering costs are mainly affected by large-scale mechanical equipment. Therefore, the prices of only large-scale mechanical equipment, such as rotary bored pile machines, punching pile machines, and excavators, are considered in this study. After selecting a suitable reference period, the machinery cost index is calculated with reference to equation (1).

(14) The management level of contractors: in general, the higher the contractor's management level, the more effectively the cost will be controlled. The management level of the contractor is determined according to the qualification of the construction unit, namely, excellent, good, medium, and poor. The dimensionless indexes of these results are 1, 2, 3, and 4, respectively.

(15) The construction period: the construction period refers to the actual number of construction days from the commencement to the completion of a foundation pit engineering project. The unit is days (d).

According to the above analysis, influencing factors of foundation pit project construction costs can be summarized, as shown in Table 1.

2.1.3. Normalization of the Influencing Factors of Foundation Pit Project Construction Costs. To prevent the output of the self-coding network from reaching saturation or even prematurely falling into the local minimum due to the large differences in the absolute values of the input data, it is necessary to normalize the input and output vectors of the training data sample set $\mathbf{X}$ in advance [43].

In this study, the input data is transformed by linear normalization [43]:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{2}$$

where x_i^* represents the normalized data, x_i represents the collected data, $x_{\min}$ is the minimum value of this type of data, and $x_{\max}$ is the maximum value of this type of data.

2.1.4. Selection of Key Influencing Factors Based on the Rough Set. In the initial index system established in Section 2.1.2, there might be repetitiveness and redundancy among these 16 indexes. Therefore, this paper used the rough set to screen the index system.

Rough set theory was put forward by the Polish mathematician Pawlak in 1982 [17]. Attribute reduction is one of the core contents of rough set theory. To understand the importance of an attribute or attribute set, you can remove this attribute or attribute set from the decision table to observe the change of decision attributes. If the decision attribute changes greatly after removing one attribute from the conditional attribute, then the conditional index has a high degree of importance in the index system. Otherwise, the conditional index has a low degree of importance. The basic theory of rough set and the use of ROSETTA software refer to references [18–20].

According to rough set theory, this paper analysed 16 influencing factors in Section 2.1.2. See Section 3.1 in this paper, for the acquisition and processing of 60 foundation pit engineering data in Hubei, China. The data in Section 3.1 was brought into the ROSETTA, and the original data was discretized and normalized to get the decision table [19]. Without changing the relationship between decision attributes and conditional attributes in the decision table, redundant attributes were removed, and the best attribute reduction was obtained.

The results showed that the optimal attribute reduction was [X11, X12, X14, X15, X16, X17, X21, X23, X31, X32, X33, X34, X41, X42]. The redundant attribute was [X13, X15, X22]. 13 main characteristic factors affecting these project construction cost by using the rough set were X11, X12, X14, X16, X17, X21, X23, X31, X32, X33, X34, X41, and X42, which was the influencing factor system of case analysis in this paper.

2.2. Prediction Model of Foundation Pit Project Construction Costs Based on the SDAE

2.2.1. Introduction of the Automatic Encoder. The automatic encoder (AE) deep learning neural network algorithm and unsupervised algorithm are the theoretical basis of this paper. The AE algorithm adopts unsupervised learning and supervised fine-tuning. It uses the BP algorithm and makes the output value approximate to the input value to the greatest extent via layer-by-layer training.

The main steps of the self-coding neural network are as follows [29].

(i) *Step 1.* Find the activation value of each layer of the network.

The activation value of the neurons in each layer is calculated by forward conduction and is taken as the input value of the next layer and transmitted forward in turn. The activation function is expressed by $f(z)$, and a_i^l represents the activation value of the ith neuron in the lth layer [44]:

$$a_i^l = f\left(z_i^l\right). \tag{3}$$

Additionally, ω_{ij}^l represents the weight between the jth neuron in the $(l+1)$th layer and the ith neuron in the lth layer, b_j^{l+1} represents the offset term of the jth neuron in the $(l+1)$th layer, and z_i^{l+1} represents the weighted sum of all inputs of the jth neuron in the $(l+1)$th layer [44]:

$$z_i^{l+1} = \sum_{i=1}^{n} \omega_{ij}^l x + b_j^{l+1}, \tag{4}$$

where n represents the number of neurons in the lth layer and x represents the input value.

(ii) *Step 2.* Update ω and b.

The residual error between the neurons in each layer and the output layer is obtained by the BP model, and ω and b are updated continuously by the gradient descent method to make the output increasingly more similar to the input. In the proposed method, ω and b are updated by the gradient descent method [26], and the equations are as follows:

$$\begin{aligned} \omega_{ij}^{l+1} &= \omega_{ij}^l - \alpha \frac{\partial}{\partial \omega_{ij}^l} J(\omega, b), \\ b_j^l &= b_j^l - \alpha \frac{\partial}{\partial b_j^l} J(\omega, b), \end{aligned} \tag{5}$$

where $J(\omega, b)$ is the cost function [25]:

Table 1: The complete and universal system of influencing factors.

Primary factors	Secondary factors	Type	Acquisition and calculation of data
X1: engineering attribute	X11: the depth of the foundation pit	Quantitative	Design documents or field investigation
	X12: the form of the foundation pit support	Quantitative	Design documents or field investigation
	X13: the form of the infrastructure	Quantitative	Design documents or field investigation
	X14: the type of pile foundation	Quantitative	Design documents or field investigation
	X15: the quantity of the pile foundation	Quantitative	Project management documents and field investigation
	X16: the engineering geological conditions	Qualitative	Field investigation and questionnaire survey
	X17: the construction area of the foundation pit	Quantitative	Design documents or field investigation
X2: environment attribute	X21: the on-site construction conditions	Qualitative	Field investigation and questionnaire survey
	X22: meteorological characteristics	Qualitative	Meteorological data and questionnaire survey
	X23: the off-site traffic conditions	Qualitative	Field investigation and questionnaire survey
X3: market attribute	X31: the labor cost index	Quantitative	Market research and equation (1)
	X32: the steel bar cost index	Quantitative	Market research and equation (1)
	X33: the concrete cost index	Quantitative	Market research and equation (1)
	X34: the machinery cost index	Quantitative	Market research and equation (1)
X4: manage attributes	X41: the management level of contractors	Qualitative	Field investigation and questionnaire survey
	X42: the construction period	Quantitative	Project management documents

$$J_{\mathrm{AE}}(\omega, b) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}a_{\omega,b}(x)^{i} - y^{i2}\right), \tag{6}$$

where m is the number of samples, x is the input, and y is the output.

The AE is not very effective in dealing with some noisy data, such as text, and its accuracy can even be decreased. The DAE proposed by Vincent [45] effectively eliminates noise interference and increases the robustness of the learned features. The inspiration of the ANN originated from the biological neural network, and the DAE also draws inspiration from reality.

2.2.2. Introduction of the SDAE. Erhan et al. [46] proposed a layer-by-layer unsupervised greedy learning algorithm in 2010. A stacked self-encoder is a superposition of multiple self-encoders in which the hidden layer of the previous layer is taken as the input layer of the next layer, and the parameters of the deep network are initialized by adopting unsupervised layer-by-layer pretraining, thereby improving the convergence speed and obtaining higher-level features. In this paper, softmax regression is used to construct a classifier to classify the features learned by SAE.

Taking the construction of a self-coding N-layer stack with N automatic encoders as an example, the general steps of the SDAE are subsequently introduced.

(i) *Step 1*. The first AE corresponds to the first hidden layer Z_1, the input layer is the original training data X, the output layer Y_1 is the reconstruction of the input layer, and the parameters are obtained by minimizing the reconstruction error:

$$\theta = \{\omega_{11}, b_{11}, \omega_{12}, b_{12}\}. \tag{7}$$

(ii) *Step 2*. The second AE corresponds to the second hidden layer Z_2, in which the upper hidden layer Z_1 is taken as the input layer and the output layer Y_2 is taken as the reconstruction of the input layer Z_1. The parameters are obtained by minimizing the reconstruction error. In the same way, the ith AE corresponds to the ith hidden layer Z_i with the upper hidden layer Z_{i-1} as the input layer and the output layer Y_i as the reconstruction of the input layer Z_{i-1}, and the parameters are obtained by minimizing the reconstruction error.

(iii) *Step 3*. Stack self-coding refers to the training of each self-encoder layer-by-layer from left to right, and the trained optimal parameters are used as the initialization parameters of the neural network. After pretraining, the parameters of all layers can be adjusted by the BP algorithm.

In order to facilitate readers to understand the structure of SDAE, this paper uses two DAEs to construct a two-layer stack self-coding, as shown in Figure 1. First of all, the first automatic encoder corresponds to the first hidden layer Z, the input layer X is the original training data, and the output layer Y is the reconstruction of the input layer X by minimizing the reconstruction error. Then, the second automatic encoder corresponds to the second hidden layer T, which takes the hidden layer Z of the previous layer as the input layer, reconstructs the input layer Z as the output layer S, and obtains the parameters by minimizing the reconstruction error. Finally, the output layer S is discarded, and the computing tools or classification tools needed for research are connected to the hidden layer T for output O [47].

2.2.3. The Data Flow Graph and Pseudocodes of SDAE. The data flow graph based on SADE classification prediction application is shown in Figure 2.

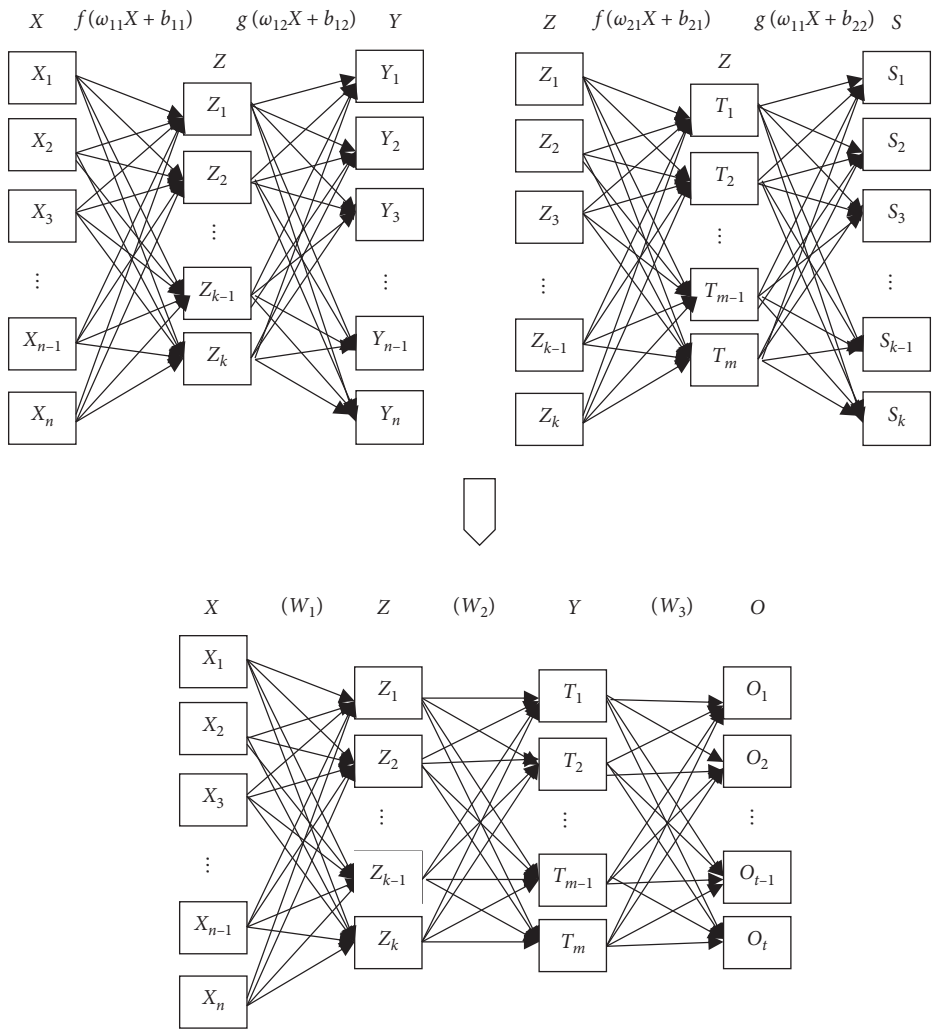


Figure 1: Structure diagram of SDAE composed of two DAEs.

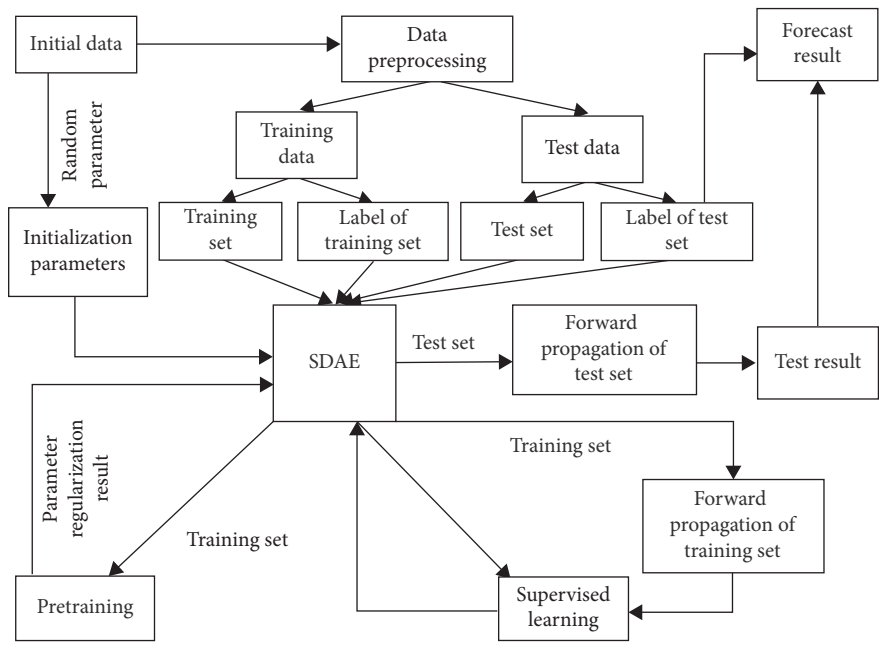


Figure 2: Data flow graph based on SADE classification prediction application.

(i) Step 1 (data collection and preprocessing): collect the original data $\mathbf{X}$ by various methods, preprocess the original data by equation (2) to get sample set $\mathbf{X}^*$, and divide the sample set $\mathbf{X}^*$ to get training set $\mathbf{X}_1^*$ and verification set $\mathbf{X}_2^*$. Because the deep learning algorithm belongs to the black box model, the predicted values in the training iteration process are random numbers. In this paper, the label data of the training set and the label data of the verification set are added to distinguish the prediction results.

(ii) Step 2 (initialize parameters): set the maximum training times, learning efficiency, number of DEA networks, and initial values of weights. And bring the initialized parameters and data stream into the SDAE.

(iii) Step 3 (unsupervised learning and forward flow of input data): its pseudocodes are in Table 2.

(iv) Step 4 (supervised learning): its pseudocodes are in Table 3. Output the prediction result after reaching the calculation termination condition.

To facilitate the readers' understanding of the data flow graph in Figure 2, this section also detailed the pseudocodes of noise reduction encoder (Table 4), the unsupervised learning in Step 3 (Table 2), and supervised learning in Step 4 (Table 3).

(1) Noise reduction encoder. The algorithm 1-1 in Table 4 is pseudocode of sae_train algorithm. The first line of the code defines the input layer as 28 nodes and the three hidden layers as 100 nodes. Lines 2–6 are the first autoencoder, which is equivalent to an encoder. The second autoencoder is in the 7th–11th lines. The third autoencoder is in the 12th–16th lines, which is equivalent to a decoder.

(2) Unsupervised learning. Algorithm 1-2 in Table 2 is pseudocode of unsupervised pretraining process.

The first line indicates that the unsupervised learning process is pretraining layer by layer, and each layer is pretrained independently. Lines 2–5 indicate that adding pretrained weights takes the weight matrix learned in the pretraining stage as the initial value of network weights. Line 6 of the code assigns sae trained weights to nn network as initial values, which covers the previous random initialization and prepares for the next supervised learning.

(3) Supervised learning. Supervised learning is the core of the algorithm. Algorithm 1–3 in Table 3 is pseudocodes of supervised learning algorithms.

The first line of algorithm 1–3 code indicates that 1000 iterations should be performed in the supervised training phase. The fourth line indicates that the prediction network extracts the data of influencing factors and affected factors (construction cost) in the stage of supervised learning and training. Lines 6–8 indicate that the input is propagated forward, and the results are obtained by the output layer after layer-by-layer weight assignment and feature extraction of all hidden layers.

3. Case Analysis

3.1. Acquisition and Normalization of Case Data. In this paper, 60 foundation pit projects in Hubei Province, China, were selected as a case study. The cost data of these projects were provided by two cooperative units (CCTEB Infrastructure Construction Investment Co., Ltd; China Construction First Group Corporation Limited). They provided the construction cost data of about 200 foundation pit projects, and the data of only 63 foundation pit projects was available. Finally, the authors randomly selected 60 foundation pit projects as case studies.

Only some engineering data of the projects are reported in Table 5 due to spatial constraints. In Table 5, y is the actual cost of each foundation pit project, and the unit is millions of RMB. The data of quantitative indicators were obtained by field research, market research, and the consultation of project management data. The reference period of the labor cost index, steel bar cost index, concrete cost index, and machinery cost index is January 1, 2018.

The data of the qualitative indicators were obtained by questionnaire surveys of 10 to 20 experts. The scoring result with the highest frequency was selected as the qualitative index score of the case project. In the questionnaire survey, experts were selected according to the criteria of being between 35 and 55 years old, holding a professional title above senior engineer or associate professor and having participated in the project construction for more than 6 months. With the assistance of SPSS 22 software, the reliability of the questionnaire survey results was analysed. The value of Cronbach's α was found to be 0.731; this exceeds the required minimum value of 0.6 [48], thereby indicating that the questionnaire survey results were reliable.

According to the content of Section 2.1.4, the 13 input vectors of the preliminary statistical training samples were normalized by equation (1) and were introduced into the SDAE via MATLAB software for calculation. The CPU of computer used in the case analysis was the Intel (*R*) Core (TM) i3-4170 @ 3.70 GHz, the memory was 6.00 GB, and the system was the Windows 7.

3.2. Prediction Results. In the process of modeling with the SDAE, the available data should be divided into two groups. The data of the training set is used for training, while the data of the test set is used for checking the model. Many researchers choose 90%: 10%, 80%: 20%, or 70%: 30% as the training and testing split ratio [49]. After normalization, the first 54 groups of data were taken as sample sets, and the remaining 6 groups were taken as test sets. Therefore, the ratio of training set data to test set data is 90%: 10%.

Table 2: The pseudocode of unsupervised pretraining process.

Algorithm 1-2: unsupervised pretraining algorithm pre_train
Entering: training data train_X Output: The result of parameter regularization 1. sae = saetrain (sae, train_x, opts); (%) Construction of pre-training network 2. nn = nnsetup ([28 100 100 1]); (%) Set the network structure 3. nn.activation_function = "sigm"; (%) Activate function 4. nn.output = "sigm";(%) Decoding 5. nn.learningRate = 1; (%) Learning rate in pretraining stage 6. nn.W{1} = sae.ae{1}.W{1}; (%) The weights trained by sae are assigned to nn network as initial values, covering the previous random initialization 7. nn.W{2} = sae.ae{2}.W{1};

Table 3: The pseudocode of supervised learning algorithms.

Algorithm 1-3: Nntrain
Input: training data and label of training data Output: Adjusted model 1. opts.numepochs = 1000; (%) Maximum number of iterations 2. opts.batchsize = 60; (%) Parameter training is carried out with all training samples in each training 3. (%) nn.weightPenaltyL2 = 10; 4. nn = nntrain (nn, train_x, train_y, opts); (%) Training data extraction 5. nn = nnff (nn, test_x, zeros (size (test_x, 1), nn.size (end))); (%) Test data extraction 6. str = sprintf "(Predicted value 1 is (%)f", nn.a{end}(1) ∗ (max1 − min1) + min1); (%) Denormalize the output predicted value 7. str1 = sprintf "(Predicted value 2is (%) *f*", nn.a{end}(2) ∗ (max1 − min1) + min1); 8. str2 = sprintf "(Predicted value 3 is (%) *f*", nn.a{end}(3) ∗ (max1 − min1) + min1);

Table 4: The pseudocode of sae_train algorithm.

Algorithm 1-1: sae_train
Input: Input (training set or the result of the previous layer noise reduction decoder) Output: The result of parameter regularization Method: sae_train 1. sae = saesetup ([28 100 100]); 2. sae.ae{1}.activation_function = "sigm"; (%) Activate function (encoding) 3. sae.ae{1}.output = "sigm"; (%) Decoding 4. sae.ae{1}.learningRate = 1; (%) Learning rate 5. sae.ae{1}.inputZeroMaskedFraction = 0.; (%) De-noising effect of automatic coding 6. (%) sae.ae{1}.weightPenaltyL2 = 10; Regularized L2 factor 7. sae.ae{2}.activation_function = "sigm"; (%) Coding 8. sae.ae{2}.output = "sigm"; (%) Decoding 9. sae.ae{2}.learningRate = 1; 10. (%) sae.ae{2}.weightPenaltyL2 = 10; 11. sae.ae{2}.inputZeroMaskedFraction = 0.; (%) The denoise autocoder trained in layers is equivalent to the dropout of hidden layers, which realizes the de-drying of automatic coder 12. sae.ae{3}.activation_function = "sigm"; (%) Coding 13. sae.ae{3}.output = "sigm"; (%) Decoding 14. sae.ae{3}.learningRate = 1; 15. (%) sae.ae{3}.weightPenaltyL2 = 10; 16. sae.ae{3}.inputZeroMaskedFraction = 0.; 17. opts.numepochs = 100; (%) Pre-training iterations 18. opts.batchsize = 60; (%) Parameter training is carried out with all training samples in each training

In view of the adjustability of the parameters in the SDAE model, the initial parameters were set as follows [30]: the maximum number of training iterations was 1000, the learning efficiency was 1.2, the number of DAEs was 3, and the initial weight value was 5. The error function diagram is presented in Figure 3.

According to Figure 3, the predicted data converged between 80 and 100 iterations of the training process.

Table 5: Data for 60 case projects.

Factors	1	2	3	4	5	6	⋯	58	59	60
X11	3.4	2.7	5.5	5.0	6.3	7.5	⋯	10.3	12.5	9.3
X12	2	3	4	2	7	5	⋯	5	2	2
X14	3	4	6	6	5	3	⋯	6	4	6
X16	1	1	2	2	3	1	⋯	1	2	1
X17	736	631	1781	1361	2061	3863	⋯	3651	6311	4378
X21	2	2	1	1	2	3	⋯	2	2	1
X23	3	3	2	2	3	3	⋯	2	3	3
X31	1.18	1.18	1.45	1.18	1.32	1.32	⋯	1.32	1.18	1.32
X32	0.96	0.96	0.85	0.85	0.89	0.89	⋯	0.89	0.85	1.03
X33	0.93	0.93	1.12	1.12	1.37	1.37	⋯	1.37	1.12	0.94
X34	1.25	1.25	1.16	1.39	1.16	1.39	⋯	1.16	1.39	1.16
X41	2	2	3	3	1	1	⋯	2	2	4
X42	60	82	73	60	94	103	⋯	144	230	126
Y	5.17	4.83	7.61	11.36	12.01	16.37	⋯	18.37	23.11	24.74

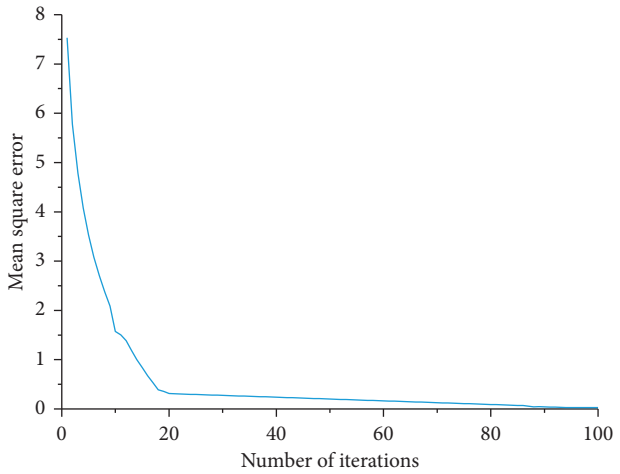


Figure 3: The chart of the error loss function.

MATLAB was then used to run the resulting code for a reverse data check, and the global error of pretraining (unsupervised learning) was found to reach 0.043 in the 90th training iteration and 0.025 in the 100th training iteration. In the supervised learning stage, it was found that the global error reached 0.00001 in the 653rd iteration, which met the prediction accuracy requirements.

The comparison between the predicted results and actual values of the foundation pit project construction costs is presented in Table 6. The average relative error of the six test sets was 1.54%.

In addition, 10-fold crossvalidation was conducted to test the accuracy of the algorithm [50]. The accuracy of 10 calculation results is exhibited in Figure 4 and was found to be very good. The errors of ten calculations were decreased. The average value of the maximum relative error was only 2.84%, and the average value of the minimum relative error was only 0.49%. In addition, the results of the 10 calculations were stable, which also proves the stability of the proposed algorithm.

4. Discussion

In this work, the SDAE was used to predict the construction costs of foundation pit projects. However, there were still two limitations in this study. (1) Different definitions of construction cost might have different influencing factors, which had a certain impact on the prediction results. If the cost definition was different from that, in the introduction of this paper, the influencing factors of foundation pit project construction cost would be likely to be different. (2) While the SDAE was successfully used to construct a prediction model of foundation pit project construction costs, many other deep learning methods could have been used.

4.1. Prediction Error Analysis of Different Forecasting Methods. At present, the commonly used cost forecasting methods are the calculation method based on national standard, the multivariate return analysis [21, 25], BP [25], GA-BP [51], SVM [34], and REF [52] models. In this study, the first 54 groups of data were selected as sample sets and the last 6 groups of data were selected as test sets and were also introduced into the models for calculation.

In this paper, 17 engineers were invited to calculate the construction cost of 60 foundation pit projects in the case analysis by using the Chinese national code (Code of bills of quantities and valuation for construction works, GB 50500-2013). The calculation took 24 days, and the results are shown in Figure 5. It could be seen that the calculation error of the GB 50500-2013 was very large, and the maximum error was 57.69%. The main reason might be that the calculation method based on the GB 50500-2013 roughly estimated that the construction cost was linear with the engineering quantity, while ignoring the influence of engineering changes on the construction cost. In addition, too long calculation time was another important deficiency of the calculation method based on the GB 50500-2013.

Using the return analysis function in Microsoft Excel 2016 software, the expression of multivariate return was calculated as follows:

Table 6: The prediction results of the proposed model.

Number of test set	55	56	57	58	59	60
Actual cost	7.38	4.19	31.48	18.37	23.11	24.74
Forecast cost	7.25	4.12	32.35	18.53	23.45	24.57
Relative error (%)	1.76	1.67	2.76	0.87	1.47	0.69

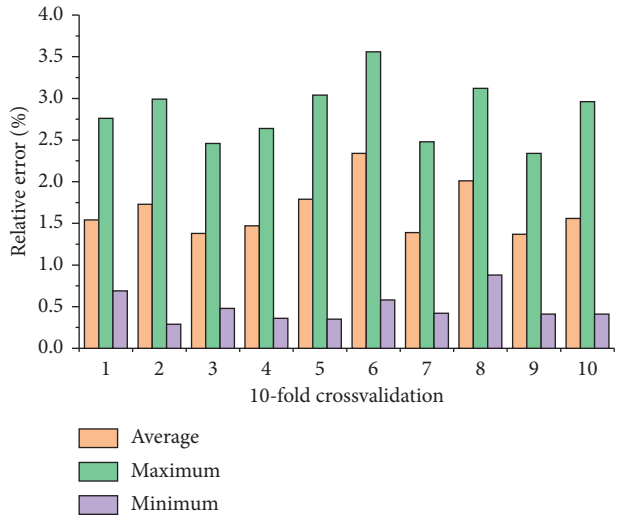


Figure 4: Error results of 10-fold crossvalidation.

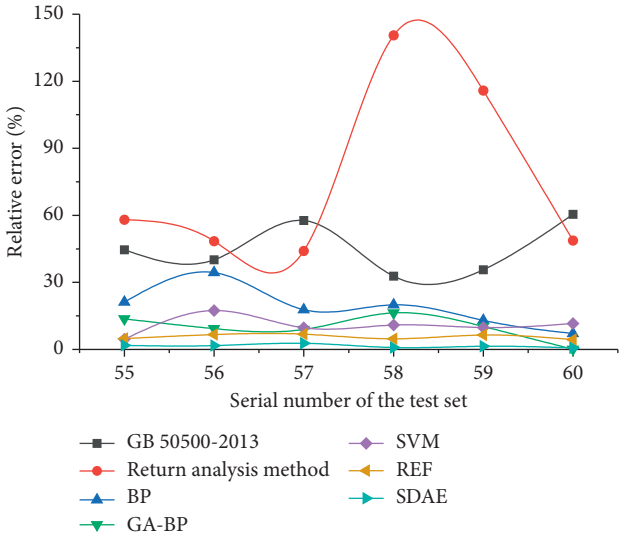


Figure 5: Relative errors of different computational models.

$$y = 4.278 + 1.375X_{12} - 0.000034X_{15} + 0.00493X_{17} + 3.04X_{41} - 0.1915X_{42}. \quad (8)$$

The data of the test set were introduced into equation (8), and the prediction results are presented in Figure 5. According to equation (8), it can be known that only five factors were related to the prediction results when using multivariate return analysis. The failure to make full use of all index data is one of the important reasons for the low accuracy of the calculation results of this method [21].

In the BP algorithm, the selected training function was "traingda," the activation function of the hidden layers was "logsig," and the activation function of the output layers was "purelin." The target error of training was set as 1×10^{-6}, and the maximum number of iterations was set as 1000. The learning rule of the network was the error gradient descent method. In the GA-BP algorithm, the number of individuals in the population was 50, the maximum genetic algebra was 1000, the number of binary digits was 20, and the generation gap was 0.9. In the cost prediction based on the SVM, the number of iterations was 100, and the population size and k value were 20 and 0.6, respectively. The calculation results are reported in Figure 5.

The maximum relative error of the SDAE model was only 0.0283, which is considerably less than the maximum relative errors of the other algorithms. The relative error is an important index in error analysis, and Table 7 presents the relative errors of several different calculation models. The relative error calculated by the proposed method was less than 3%, and the average error was only 1.54%. Among all the methods, the calculation error of the multivariate return analysis method was the largest, and the relative error was as high as 140.52%. The calculation errors of the BP, GA-BP, SVM, and RBF models were large, and the maximum relative errors were 34.43%, 16.39%, 13.60%, and 6.83%, respectively. These results also prove that the proposed method is effective and advanced in predicting the construction costs of foundation pit projects.

In addition, combined with the calculation results of other error analysis tools, it could be qualitatively considered that SDAE had the highest calculation accuracy in case analysis, and the calculation accuracy order of other methods was as follows: REF > SVM > GA-BP > BP > return analysis method. This sort of calculation accuracy was consistent with the previous research results [21, 25, 51], which also proved that the case analysis in this paper was scientific and correct.

In order to further compare and analyze the calculation errors of various calculation methods, the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE) were used to analyze the prediction error in the case study.

R^2 indicates the degree of correlation between the actual and predicted values. The closer R^2 is to 1, the higher the correlation; conversely, the closer R^2 is to 0, the lower the correlation [53]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i^{\exp} - y_i^{\mathrm{pre}}\right)^2}{\sum_{i=1}^{N} \left(y_i^{\exp} - \overline{y_i^{\exp}}\right)^2}, \quad (9)$$

where $y_i^{\exp}$ is the actual result, y_i^{pre} is the predicted result, and $\overline{y_i^{\exp}}$ is the average value of the actual results.

The RMSE is an important standard used to measure the prediction results of machine learning models [54]. Its calculation method is as follows:

Table 7: Comparison of the three error representations of different models.

Error representations	R^2	RMSE	MAE
GB 50500-2013	0.3798	2.1701	6.8911
Return analysis method	0.2311	3.9010	9.1758
BP	0.6410	1.3725	3.1433
GA-BP	0.8276	0.8130	1.6050
SVM	0.8531	0.6203	1.2604
REF	0.8970	0.6977	1.0432
SDAE	0.9743	0.4110	0.3050

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\exp} - y_i^{\mathrm{pre}}\right)^2}. \quad (10)$$

The MAE is the average of absolute errors, which can better reflect the actual situation of errors in predicted values [53]:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{\exp} - y_i^{\mathrm{pre}}\right|. \quad (11)$$

The error results of different methods are shown in Table 7.

According to the calculation results presented in Table 3, the R^2 value of the SDAE model was the highest, namely, 0.9743, which is very close to 1. In other words, the predicted values calculated by the SDAE model were very close to the actual values. Compared with the GB 50500-2013, the multiple return analysis method, BP, GA-BP, SVM, and REF models, the proposed model had better prediction results. The RMSE of the SDAE model was 0.1689, which is notably less than the RMSEs of the other algorithms. The MAE of the SDAE model was 0.305, which is notably less than the MAEs of the other algorithms. Compared with the other common methods, the SDAE model exhibited a superior calculation accuracy.

4.2. Stability Analysis of Different Computational Models. Stability determines the reliability and generalization of the model in engineering application. In this paper, the standard deviation was used as a measure of the stability of SDAE model. Among the 54 training samples in Section 3.1, 20, 30, 40, and 50 samples were randomly selected as training sets, and the last 6 samples were also used as test sets. The standard deviations of different models are shown in Table 8.

It can be seen from Table 8 that the SDAE model showed a low standard deviation of prediction, regardless of the size of the training sample. The SDAE had the strongest stability with the increase of the training sample size. The larger the training sample size, the stronger the stability of SDAE model. Compared with the Chinese national code, the diversified return method, the BP, the GA-BP, the SVM, and the RBF, the standard deviation of SDAE was lowest, which showed that this model had stronger stability than other models.

4.3. The Influence of the Number of Input Variables on the Prediction Results. According to previous research results [13], when an artificial intelligence method is applied to the prediction of construction costs, the number of input variables has a notable influence on the accuracy of the prediction results. Therefore, the influence of the number of input variables on the prediction results was analysed. Considering that many factors affect the construction costs of foundation pit projects, only the following situations were analysed. Plan A was the use of the 16 influencing factors identified in Section 2.2.1. In Plan B, the influencing factor X11 (*the depth of the foundation pit*) was deleted. In Plan C, the influencing factor X12 (*the form of the foundation pit support*) was deleted. In Plan D, the influencing factors X11 and X12 were deleted. In Plan E, the influencing factor X13 (*the form of the infrastructure*) was deleted. In Plan F, the influencing factors X11, X12, and X13 were deleted. In Plan G, the X13, X15, and X22 were deleted. The index system of the Plan G was the same as that in case analysis. In Plan H, the influencing factors X11, X12, X13, and X21 (*the on-site construction conditions*) were deleted. In Plan I, the influencing factors X11, X12, X13, X21, and X22 (*meteorological characteristics*) were deleted. Finally, in Plan J, the influencing factors X11, X12, X13, X21, X22, X31 (*the labor cost index*), and X32 (*the steel bar cost index*) were deleted. The calculation results of these plans are shown in Table 9.

When impact factor X11 (Plan B) or X12 (Plan C) was deleted, the error of the calculation results increased obviously, whereas this did not occur when other single factors (such as Plan E) were deleted. In the example of reducing two influence factors at the same time (Plan D), the error of the calculation results increased obviously when X11 and X12 were deleted. However, when other influencing factors in addition to X11 and X12 were deleted, the calculation error did not increase obviously. For example, the maximum relative error of Plan I, in which influencing factors X11, X12, X13, X21, X22, X31, and X32 were deleted, was 6.0%, which is only slightly larger when only influencing factors X11 and X12 were deleted based on this analysis, it can be preliminarily considered that the influencing factors X11 and X12 have a substantial influence on the calculation accuracy. Comparing Plan A and Plan G, the calculation errors of the two index systems were very close. This could explain the rationality and efficiency of the index screening results in Section 2.1.4 of this paper. It should be emphasized that the analysis and discussion on

Table 8: Stability of calculation results under different sample numbers.

Model	20	30	40	50
GB 50500-2013	4.890	3.283	2.840	2.594
Return analysis method	13.370	7.292	6.382	4.900
BP	10.063	7.063	5.985	3.801
GA-BP	3.266	2. 006	0.922	1.312
SVM	2.650	1.985	1.298	0.578
REF	1.783	1.357	1.034	1.231
SDAE	0.931	1.047	0.461	0.058

the number of input variables in this section was preliminary, not complete. The main reason was that there were too many input variables.

5. Conclusion

Foundation pit project construction costs are an important component of building project construction costs. The prediction of foundation pit project construction costs is the basis of not only cost planning but also of the cost decisions and planning of construction projects. In this paper, beginning from the four attributes of construction cost management (engineering, the environment, the market, and management), the influencing factors of foundation pit project construction costs were identified. Combined with China's national standards and the practice of foundation pit project management, a method of the quantization of the influencing factors was provided. Then, the SDAE was utilized to construct a prediction model of foundation pit project construction costs. Finally, 60 foundation pit projects in Hubei Province, China, were selected for a case analysis. The case study results demonstrated that, compared with the actual construction costs, the calculation error of the proposed method was less than 3%, and the average error was only 1.54%. In addition, three error analysis tools commonly used in machine learning (the determination coefficient, root mean square error, and mean absolute error) emphasized that the calculation accuracy of the proposed method was superior to those of the Chinese national code, the multivariate return method, the BP model, the BP model optimized by the genetic algorithm, the SVM model, and the RBF model. For 60 foundation pit projects in case analysis, deleting X13, X15, and X22 did not affect the prediction results. The result also proved the rationality and efficiency of the key impact indicators obtained by the rough set. On the basis of the research results in this paper, relevant researchers are encouraged to further find a complete and universal system of influencing factors affecting the project construction cost of deep foundation pit.

Data Availability

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

Table 9: Error analysis of the calculation results with different numbers of input variables.

Error representations	R^2	RMSE	MAE
Plan A	0.979	0.402	0.295
Plan B	0.943	0.473	0.323
Plan C	0.952	0.486	0.347
Plan D	0.901	0.520	0.483
Plan E	0.969	0.415	0.309
Plan F	0.894	0.423	0.334
Plan G	0.974	0.411	0.305
Plan H	0.863	0.569	0.518
Plan I	0.871	0.597	0.510
Plan J	0.858	0.566	0.528

Conflicts of Interest

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

Acknowledgments

This study was supported by the Science and Technology Project of Wuhan Urban and Rural Construction Bureau, China (201943).

References

[1] S. Durdyev and S. Ismail, "Role of the construction industry in economic development of Turkmenistan," *Energy Education Science and Technology Part a: Energy Science and Research*, vol. 29, no. 2, pp. 883–890, 2012.

[2] M. Mukuka, C. Aigbavboa, and W. Thwala, "Effects of construction projects schedule overruns: a case of the Gauteng Province, South Africa," *Procedia Manufacturing*, vol. 3, pp. 1690–1695, 2015.

[3] I. Zafar, T. Yousaf, and D. S. Ahmed, "Evaluation of risk factors causing cost overrun in road projects in Terrorism Affected Areas Pakistan - a case study," *KSCE Journal of Civil Engineering*, vol. 20, no. 5, pp. 1613–1620, 2016.

[4] P. E. D. Love and D. J. Edwards, "Curbing rework in offshore projects: systemic classification of risks with dialogue and narratives," *Structure and Infrastructure Engineering*, vol. 9, no. 11, pp. 1118–1135, 2013.

[5] J. P. Xu, Y. Tu, and Z. Q. Zeng, "A nonlinear multiobjective bilevel model for minimum cost network flow problem in a large-scale construction project," *Mathematical Problems in Engineering*, vol. 2012, Article ID 463976, 40 pages, 2012.

[6] J. D. Wu, "Selection of the scheme of foundation pit support based on value engineering," *Construction Economy*, vol. 353, no. 3, pp. 88–90, 2012.

[7] A. Gilchrist and E. N. Allouche, "Quantification of social costs associated with construction projects: state-of-the-art review," *Tunnelling and Underground Space Technology*, vol. 20, no. 1, pp. 89–104, 2005.

[8] L. F. Cabeza, L. Rincon, V. Vilarino et al., "Life cycle assessment (LCA) and life cycle energy analysis (LCEA) of buildings and the building sector: a review," *Renewable & Sustainable Energy Reviews*, vol. 29, pp. 394–416, 2013.

[9] N. Forcada, M. Gangolells, M. Casals et al., "Factors affecting rework costs in construction," *Journal of Construction Engineering and Management*, vol. 134, no. 8, Article ID 04017032, 2017.

[10] B. Wang and J. Dai, "Discussion on the prediction of engineering cost based on improved BP neural network algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 5, pp. 6091–6098, 2019.

[11] T. P. Williams and J. Gong, "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers," *Automation in Construction*, vol. 43, pp. 23–29, 2014.

[12] Y. M. Wei, "Earned value method in the total cost variance analysis of land development and consolidation project," *China Land Science*, vol. 25, no. 5, pp. 73–78, 2011.

[13] M. Attalla and T. Hegazy, "Predicting cost deviation in reconstruction projects: artificial neural networks versus regression," *Journal of Construction Engineering and Management*, vol. 129, no. 4, pp. 405–411, 2003.

[14] A. Lesniak and M. Juszczyk, "Prediction of site overhead costs with the use of artificial neural network based model," *Archives of Civil and Mechanical Engineering*, vol. 18, no. 3, pp. 973–982, 2018.

[15] K. J. Kim and K. Kim, "Preliminary cost estimation model using case-based reasoning and genetic algorithms," *Journal of Computing in Civil Engineering*, vol. 24, no. 6, pp. 499–505, 2010.

[16] J. C. Dong, Y. Chen, and G. Guan, "Cost index predictions for construction engineering based on LSTM neural networks," *Advances in Civil Engineering*, vol. 2020, 14 pages, Article ID 6518147.

[17] M. Su, P. Wang, Y. Xue et al., "Prediction of risk in submarine tunnel construction by multi-factor analysis: a collapse prediction model," *Marine Georesources & Geotechnology*, vol. 37, no. 9, pp. 1119–1129, 2019.

[18] G.-H. Zhang, Y.-Y. Jiao, L.-B. Chen, H. Wang, and S.-C. Li, "Analytical model for assessing collapse risk during mountain tunnel construction," *Canadian Geotechnical Journal*, vol. 53, no. 2, pp. 326–342, 2016.

[19] X. M. Xu, Q. Wang, D. X. Niu et al., "Synergistic effect evaluation of main and auxiliary industry of power grid based on the information fusion technology from the perspective of sustainable development of enterprises," *Sustainability*, vol. 10, no. 2, Article ID 457, 2018.

[20] S. Barbagallo, S. Consoli, N. Pappalardo, S. Greco, and S. M. Zimbone, "Discovering reservoir operating rules by a Rough Set approach," *Water Resources Management*, vol. 20, no. 1, pp. 19–36, 2006.

[21] S. M. Trost and G. D. Oberlender, "Predicting accuracy of early cost estimates using factor analysis and multivariate regression," *Journal of Construction Engineering and Management*, vol. 129, no. 2, pp. 198–204, 2003.

[22] S. H. Ji and J. Ahn, "Scenario-planning method for cost estimation using morphological analysis," *Advances in Civil Engineering*, vol. 2019, 10 pages, Article ID 4962653, 2019.

[23] Y.-M. Cheng, C.-H. Yu, and H.-T. Wang, "Short-interval dynamic forecasting for actual S -curve in the construction phase," *Journal of Construction Engineering and Management*, vol. 137, no. 11, pp. 933–941, 2011.

[24] X. J. Wang, "Forecasting construction project cost based on BP neural network," in *Proceeding of the 10th International Conference on Measuring Technology And Mechatronics Automation*, pp. 420–423, Changsha, China, February 2018.

[25] M. Gunduz, L. O. Ugur, and E. Ozturk, "Parametric cost estimation system for light rail transit and metro trackworks," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2873–2877, 2011.

[26] K. M. Kiani and T. L. Kastens, "Testing forecast accuracy of foreign exchange rates: predictions from feed forward and various recurrent neural network architectures," *Computational Economics*, vol. 32, no. 4, pp. 383–406, 2008.

[27] R. H. Mu and X. Q. Zeng, "A review of deep learning research," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, pp. 1738–1764, 2019.

[28] H. A. Owolabi, M. Bilal, L. O. Oyedele, H. A. Alaka, S. O. Ajayi, and O. O. Akinade, "Predicting completion risk in PPP projects using big data analytics," *IEEE Transactions on Engineering Management*, vol. 67, no. 2, pp. 430–453, 2020.

[29] G. Zhang, Y. Liu, and X. Jin, "A survey of autoencoder-based recommender systems," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 430–450, 2020.

[30] E. K. Wang, X. Zhang, and L. Pan, "Automatic classification of CAD ECG signals with SDAE and bidirectional long short-term network," *IEEE Access*, vol. 7, pp. 182873–182880, 2019.

[31] P. Liu, P. J. Zheng, and Z. Y. Chen, "Deep learning with stacked denoising auto-encoder for short-term electric load forecasting," *Energies*, vol. 12, no. 12, Article ID 2445, 2019.

[32] J. Dai, H. Song, G. Sheng, and X. Jiang, "Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders," *IEEE Access*, vol. 5, pp. 22863–22870, 2017.

[33] W. Dong, H. Sun, Z. Li, J. Zhang, and H. Yang, "Short-term wind-speed forecasting based on multiscale mathematical morphological decomposition, K-means clustering, and stacked denoising autoencoders," *IEEE Access*, vol. 8, pp. 146901–146914, 2020.

[34] X. Chen, M. Li, and Y. Q. Xiao, "Stacked denoise autoencoder based feature extraction and classification for hyperspectral images," *Journal of Sensors*, vol. 2016, Article ID 3632943, 10 pages, 2016.

[35] J. Yan, H. Zhang, Y. Liu, S. Han, L. Li, and Z. Lu, "Forecasting the high penetration of wind power on multiple scales using multi-to-multi mapping," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3276–3284, 2018.

[36] Y.-M. Cheng, "An exploration into cost-influencing factors on construction projects," *International Journal of Project Management*, vol. 32, no. 5, pp. 850–860, 2014.

[37] V. Sharma, C. H. Caldas, and S. P. Mulva, "Identification and prioritization of factors affecting the overall project cost of healthcare facilities," *Journal of Construction Engineering and Management*, vol. 146, no. 2, Article ID 04019106, 2020.

[38] J. Ngo, B. G. Hwang, and C. Y. Zhang, "Factor-based big data and predictive analytics capability assessment tool for the construction industry," *Automation in Construction*, vol. 110, p. 2020, Article ID 103042.

[39] M. T. Hatamleh, M. Hiyassat, G. J. Sweis, and R. J. Sweis, "Factors affecting the accuracy of cost estimate: case of Jordan," *Engineering, Construction and Architectural Management*, vol. 25, no. 1, pp. 113–131, 2018.

[40] Y. Wang and Q. J. Li, "Study on influencing factors of cost control of construction project based on structural equation," *Construction Economy*, vol. 41, no. 02, pp. 63–68, 2020.

[41] X. Gao and P. Pishdad-Bozorgi, "A framework of developing machine learning models for facility life-cycle cost analysis," *Building Research & Information*, vol. 48, no. 5, pp. 501–525, 2020.

[42] M. Yang and J. W. Jin, "Research progress on interaction of pile foundation with nearby existing subway tunnel," *Journal of Building Structures*, vol. 37, no. 8, pp. 90–100, 2016.

[43] H. Wu and J. W. Wang, "Assessment of waterlogging risk in the deep foundation pit projects based on projection pursuit

model," *Advances in Civil Engineering*, vol. 2020, Article ID 2569531, 11 pages, 2020.

[44] W. Liu, T. Ma, Q. Xie, D. Tao, and J. Cheng, "LMAE: a large margin Auto-Encoders for classification," *Signal Processing*, vol. 141, pp. 137–143, 2017.

[45] J. Zhang, Y. Zhang, L. Bai, and J. Han, "Lossless-constraint denoising based auto-encoders," *Signal Processing: Image Communication*, vol. 63, pp. 92–99, 2018.

[46] D. Erhan, Y. Bengio, A. Courville et al., "Why does unsupervisedpre-training help deep learning," *Journal of Machine Learning Research*, vol. 11, no. 3, pp. 625–660, 2010.

[47] F. Xu and P. W. Tse, "Automatic roller bearings fault diagnosis using DSAE in deep learning and CFS algorithm," *Soft Computing*, vol. 23, no. 13, pp. 5117–5128, 2019.

[48] M. H. P. Passos, H. A. Silva, A. C. R. Pitangui, V. M. A. Oliveira, A. S. Lima, and R. C. Araújo, "Reliability and validity of the Brazilian version of the pittsburgh sleep quality index in adolescents," *Jornal de Pediatria*, vol. 93, no. 2, pp. 200–206, 2017.

[49] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[50] J.-H. Kim, "Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap," *Computational Statistics & Data Analysis*, vol. 53, no. 11, pp. 3735–3745, 2009.

[51] O. Kisi, M. Alizamir, and M. Zounemat-Kermani, "Modeling groundwater fluctuations by three different evolutionary neural network techniques using hydroclimatic data," *Natural Hazards*, vol. 87, no. 1, pp. 367–381, 2017.

[52] R. Sonmez, "Range estimation of construction costs using neural networks with bootstrap prediction intervals," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9913–9917, 2011.

[53] M. A. Ahmadi, R. Haghbakhsh, R. Soleimani, and M. B. Bajestani, "Estimation of H2S solubility in ionic liquids using a rigorous method," *The Journal of Supercritical Fluids*, vol. 92, pp. 60–69, 2014.

[54] H. H. Li, Y. D. Lu, C. Zheng et al., "Groundwater level prediction for the arid oasis of Northwest China based on the artificial bee colony algorithm and a back-propagation neural network with double hidden layers," *Water*, vol. 11, no. 4, Article ID 860, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 1383198, 6 pages
https://doi.org/10.1155/2020/1383198

Research Article

Traffic Flow Detection at Road Intersections Based on *K*-Means and NURBS Trajectory Clustering

Jun-fang Song, Shu-yu Wang, and Hai-li Zhao

School of Information Engineering, Xizang Minzu University, Xianyang, Shaanxi 712082, China

Correspondence should be addressed to Jun-fang Song; 284786635@qq.com

Received 2 June 2020; Revised 16 October 2020; Accepted 28 October 2020; Published 17 November 2020

Academic Editor: William Guo

In view of the variety and occlusion of vehicle target motion on the urban intersection, it is difficult to accurately detect the traffic flow parameters in all directions and categories of the intersection, so an improved *k*-means trajectory clustering method based on NURBS curve fitting is designed to obtain the traffic flow parameters. Firstly, the B-spline quadratic interpolation function is used to fit the smooth NURBS curve of vehicle trajectory; secondly, *K*-means clustering is used to measure the minimum distance, and the location of the first and last end points of the vehicle trajectory is used to realize the automatic division of the intersection area; finally, according to the intersection area where the start and end points of vehicle trajectory belong, respectively, the moving mode of a vehicle is determined, and the traffic flow parameters are classified and counted. Experiments show that the method has high accuracy and simple algorithm, which can meet the application requirements of intelligent transportation. It can provide effective data for traffic congestion analysis and lane occupancy estimation, and it is an important parameter for dynamic time setting of intersection information lights.

1. Introduction

The perception and prediction of traffic scene mainly includes the acquisition of vehicle flow parameters, the identification of traffic abnormal behavior, and the prediction of traffic congestion status, which is one of the hot areas of traffic scholars in various countries. It is an important task in the traffic scene perception system to obtain the traffic flow parameters in the video sequence. Real-time and accurate vehicle flow parameters can provide important data support for urban traffic control and management organizations. In addition, vehicle flow parameters on urban roads also reflect the current traffic conditions to a certain extent, for example, road traffic congestion status, lane occupancy rate, and road status; these information are conducive to traffic accident early warning, road congestion prediction and automatic planning of travel path [1]. Statistics of the traffic flow information of urban road intersections in different periods and in all directions can also be used as an important basis for the setting of dynamic timing parameters of intersection signal lights. It can effectively improve the traffic efficiency of the road, improve the traffic condition of the urban road network, and bring significant economic benefits to the society [2]. The traffic scene of urban intersection is complex, the vehicle motion mode is changeable, and the occlusion is serious, which make it difficult to detect and track the moving target accurately and continuously. In order to obtain the traffic flow parameters of urban intersections, it is very important to accurately identify the vehicle target and determine its moving direction.

According to the different calculation methods and principles, there are currently four categories of traffic flow statistics: the target detection method [3–6], feature point motion trajectory clustering method [7–9], regional regression method [10, 11] and density estimation method [12–14]. Seenouvong et al. [3] completed the vehicle count in the set area by combining background subtraction with morphological filtering, with an accuracy of 96%; Memon et al. [4] introduced the Gaussian mixture model (GMM) to quickly detect and count vehicles in the field of view and realized the self-classification of moving targets by contour comparison; Chen [5] proposed to use the B-spline curve

method to obtain the vehicle area, so as to realize the vehicle count; Leonel et al. [6] proposed a video sequence vehicle counting method based on augmented quantum space learning. The accuracy rate of counting vehicles at an average speed of 26 frames per second is 96.6%, and it can well deal with camera shake and sudden illumination changes caused by the environment and automatic camera exposure. Rabbouch et al. [7] designed a pattern recognition system based on unsupervised clustering of tracks, which realized intelligent detection, counting, and recognition of traffic targets; Mehboob et al. [8] built a real-time system through analyzing the temporal and spatial characteristics of vehicle trajectory, which it can deduce and track the target behavior online, and joint Hungarian tracking algorithm to count the number of vehicles; in terms of vehicle flow statistics by different types, Rauf et al. [9] proposed a method based on target tracking and convolutional neural network transfer learning; the regional regression method is well reflected in the literature [10, 11]. Liang et al. [10] constructed a regression model by extracting the edge features and gradient features of vehicles on the highway to obtain the vehicle flow parameters; Chen [11] suggested a hierarchical classification-based regression model for accurate vehicle counting in view of the complex and changeable characteristics of the actual traffic scene; Lempitsky and Zisserman [12] proposed the target counting algorithm framework based on density estimation in 2010, which is the most common and concerned counting framework at present. The method first generates the truth value image set of the target density distribution by using the target center graph manually annotated. Zhang et al. [13] introduced a vehicle density estimation method based on rank constraint regression and Fully Convolutional Networks (FCN), so as to realize accurate statistics of the number of vehicles. Based on the convolutional neural network (CNN), Sindagi and Patel [14] used deep learning to estimate vehicle density in crowded situations.

In a word, in order to obtain accurate traffic flow parameters, in addition to establishing a reasonable mathematical model, it is particularly important to generate features that can truly reflect the running status of vehicle targets. However, the scene of urban intersection is complex, the moving state of the target is changeable, there are many occlusions, and rich target feature information is needed. The target trajectory can provide a wealth of spatiotemporal data such as direction and speed, and it has good stability in all-weather operation. This manuscript takes the trajectory of vehicle as an important feature, firstly, NURBS is used to fit the trajectory curve; then, k-means algorithm is used to complete the clustering of the trajectory curve; finally, based on the clustering results, the short-term traffic flow in each region and direction of the intersection is predicted.

2. *K*-Means Trajectory Clustering Based on NURBS Curve Fitting

Trajectory clustering is to use the appropriate similarity measurement method to complete the best classification of the obtained trajectory features. To explore suitable similarity measurement criteria to ensure the best clustering results, we choose K-means clustering algorithm, which has fast calculation speed and high efficiency and is suitable for finding convex shape clusters. Aiming at the defects of the randomness of initial mean and the sensitivity of noise outliers, this manuscript introduces NURBS curve fitting to solve the problem of noise points when the target pairs to form a trajectory. The NURBS curve can be regarded as the sum of the product of a series of control points $\{V_i\}$ $(i = 0, 1, 2, \ldots, n)$ and the base function $N_{i,k}(u)$ determined by the known node sequence $\{u_i\}$, $(i = 0, 1, 2, \ldots, n)$. Combined with a series of weight factors that can affect the shape of the curve, the mathematical expression of the vector function is as follows:

$$P(u) = \frac{\sum_{i=0}^{n} \omega_i d_i N_{i,k}(u)}{\sum_{i=0}^{n} \omega_i N_{i,k}(u)} = \sum_{i=0}^{n} d_i B_{i,k}(u),$$
$$B_{i,k}(u) = \frac{\omega_i N_{i,k}(u)}{\sum_{j=0}^{n} \omega_j N_{j,k}(u)}. \tag{1}$$

In the formula, $B_{i,k}(u)$ is the basis function of the k-th normal B-spline; d_i is the control vertex, also known as the De Boor point; ω_i is the weight factor; and u is the parameter value.

The weight factor introduced by the NURBS function can realize the local correction of curve flexibly. Before the k-means algorithm clusters the trajectory, first, our method fits the NURBS curve through B-spline quadratic interpolation and, then, calculates the first-order guide vector of all the trajectory points, which is equal to the tangent slope passing through the point in 2D plane. If the curve is smooth and continuous, then the tangent slopes of all trace points on the curve will change between the slopes of the two control points; otherwise, it indicates that there are irregular points on the curve. According to the nearest point search algorithm of k-d Tree [15], all irregular points can be found by traversing the whole trajectory, and the best trajectory can be obtained by NURBS curve fitting again.

2.1. NURBS Curve Fitting of Track Points. Based on detection results of vehicles, their track matching was used to eliminate irregular measurement points by the piecewise straight-line fitting method. However, the trajectory fitted by this method is not smooth, and there are many inflexion points; secondly, due to the influence of perspective projection transformation of image, the position of many points has been distorted, so it is unable to predict the macromotion trajectory of the vehicle intuitively. In this paper, the NURBS function in the theory of curve geometry [16], is applied to the fitting of vehicle trajectory. It has good invariance to perspective projection transformation. It can effectively eliminate irregular points, improve the representation ability of trajectory, and ensure that the traffic flow at intersections can be accurately predicted through the processing of the trajectory subregion and subdirection. The essence of NURBS curve fitting based on trajectory points is to interpolate or fit the measurement points into smooth curves.

The trajectory curve fitting steps are as follows:

(1) We calculate the parameter values and node vectors in the U or V direction of the curve from the measurement point data of the vehicle target trajectory

(2) The node vector interval of each parameter value is determined, and all basis functions are obtained

(3) A set of equations with curve control point (type value point) as unknown quantity is established, and the coordinates of the curve control point are obtained by solving the equations

We assume that the measurement point (track point) of the track is

$$\{B_k\}, \quad k = 0, 1, \ldots, n. \tag{2}$$

The NURBS curve is fitted by the interpolation function to the measuring points:

$$C(u) = \sum_{i=0}^{n} N_{i,p}(u) P_i,\ U = \left\{ \underbrace{0, \ldots, 0}_{p+1}, u_{p+1}, \ldots, u_{m-p-1}\ ,\ \underbrace{1, \ldots, 1}_{p+1} \right\}. \tag{3}$$

Among them, P_i is the type value point, the parameter $\overline{u}_k$ corresponding to measurement point B_k is calculated by the chord length parameterization method, and the node vector U is determined by the mean value method. In this way, after the basis function is found, it is substituted with the type value point into the following formula for calculation:

$$B_k = C(\overline{u}_k) = \sum_{i=0}^{n} N_{i,p}(\overline{u}_k) P_i. \tag{4}$$

After that, a system of equations with the point P_i as an unknown is obtained, and it is composed of $n+1$ equations but $n+3$ unknowns, belonging to the overdetermined system of equations with no linear solution, so we need to establish other two equations to solve it. Considering the particularity of the first and end points, the mathematical relationship between the first-order derivative vector D_0, D_n and the adjacent type value points P_0, P_1 and P_{n-1}, P_n is as follows:

$$-P_0 + P_1 = \frac{u_{p+1}}{p} D_0, \tag{5}$$

$$-P_{n-1} + P_n = \frac{u_{m-p-1}}{p} D_n. \tag{6}$$

Combining equations (5) and (6) with equation (3), a system of linear equations whose coefficient matrix is $(n+3) * (n+3)$ is obtained, and then, we can solve this system and get the 2D coordinates of $n+3$ control points. Then, the NURBS curve is fitted with a nonrational quadratic B-spline interpolation function, and the derivative of the vector is obtained at any point as follows:

$$C'(u) = -2(1-u)P_{j-1} + 2(1-2u) \times \frac{-(1-u)^2 P_{j-1} + -u^2 P_{j+1}}{2u\left(1 - uP_j\right)} + 2uP_{j+1}. \tag{7}$$

Among them, P_{j-1}, P_j, and P_{j+1} are three adjacent measurement points in the NURBS curve, and the parameter $u = (\sqrt{(x_{j-}x_{j-1})^2 + (y_{j-}y_{j-1})^2})/(\sqrt{(x_{j-}x_{j-1})^2 + (y_{j-}y_{j-1})^2} + \sqrt{(x_{j+1} - x_j)^2 + (y_{j+1} - y_j)^2})$.

2.2. K-Means Trajectory Clustering. There is a big deviation in the k-means algorithm due to the uncertainty of the initial clustering center. In our manuscript, the minimum distance is proposed as a measurement criterion to select the initial clustering center. In view of the fact that the start and end points of each moving vehicle's trajectory at the intersection must be clustered in each intersection area, it is accurate and simple to predict the trajectory trend with them. All the vehicle's movement trajectories obtained by fitting NURBS are represented by sets $\{P_{\text{start}}\}$and $\{P_{\text{end}}\}$ of start and end points. We calculate the normalized Euclidian distance between any two points in set $\{P_{\text{start}}\}$and $\{P_{\text{end}}\}$; find the samples with the shortest distance by the quicksort method and write it as P_{d0} and P_{dn}; calculate the Euclidean distance between them and residual sample points in the set $\{P_{\text{start}}\}$and $\{P_{\text{end}}\}$; find out the sample points with the smallest distance from the center of P_{d0} and P_{dn}; increase them to P_{d0} and P_{dn}, and repeat this process until we have k sets of objects. Finally, k object sets are averaged numerically to form a group of clustering centers, which are used as the initial clustering centers when k-means algorithm is used to complete the trajectories clustering, and k clustering centers J_k are obtained. The centroid is taken as the central location of the intersection J_e, and k is equal to the number of intersection areas.

3. Regional Traffic Flow Prediction

In order to facilitate the prediction of partitioned traffic flow, in the polar coordinate system, the motion direction angle (relative to J_e) of the trajectory point is introduced, which is as follows:

Table 1: Experimental scene data.

Feature	Scenario		
	Scenario 1	Scenario 2	Scenario 3
Area number	4	4	3
Direction number	12	12	6
Time quantum	PM 5:00 to 5:45	AM 7:00 to 8:00	AM 7:00 to 8:00
Intersection property	Intersection	Intersection	T-junction

(a) (b) (c)

Figure 1: Vehicle statistical results of each intersection scene by region. (a) Scenario 1. (b) Scenario 2. (c) Scenario 3.

$$\begin{cases} \dfrac{180^\circ}{\pi * \arctan\left((y_i - y_e)/(x_i - x_e)\right)}, & \text{if: } x_i > x_e, y_i > y_e, \\ 90^\circ, & \text{if: } x_i = x_e, y_i > y_e, \\ \dfrac{180^\circ + 180^\circ}{\pi * \arctan\left((y_i - y_e)/(x_i - x_e)\right)}, & \text{if: } x_i < x_e, y_i > y_e, \\ 180^\circ, & \text{if: } x_i < x_e, y_i = y_e, \\ \dfrac{270^\circ + 180^\circ}{\pi * \arctan\left((y_i - y_e)/(x_i - x_e)\right)}, & \text{if: } x_i < x_e, y_i < y_e, \\ 270^\circ, & \text{if: } x_i = x_e, y_i < y_e, \\ \dfrac{360^\circ + 180^\circ}{\pi * \arctan\left((y_i - y_e)/(x_i - x_e)\right)}, & \text{if: } x_i > x_e, y_i < y_e, \\ 360^\circ, & \text{if: } x_i > x_e, y_i = y_e, \\ 0^\circ, & \text{others.} \end{cases} \tag{8}$$

Taking the intersection traffic scene as an example, four θ values can be obtained, and the ascending order is $0^\circ \le \theta_1 < \theta_2 < \theta_3 < \theta_4 \le 360^\circ$; we calculate $\theta' = \theta_1 + \theta_2/2, \theta_2 + \theta_3/2, \theta_3 + \theta_4/2, \theta_4 + \theta_1/2$ and arrange it in ascending order as $0^\circ \le \theta_1' < \theta_2' < \theta_3' < \theta_4' \le 360^\circ$; then, we call (θ_1', θ_2') A-block, (θ_2', θ_3') B-block, (θ_3', θ_4') C-block, and $(\theta_4', 360^\circ)$ and $0, \theta_1'$ the D-block. Other intersection scene intersection area division methods can be analogized.

Calculating the direction angle θ_k and θ_k' of motion of the points P_{start} and P_{end}, the vehicle's movement pattern can be known by judging their respective regions. Assuming that θ_k belongs to zone A and θ_k' belongs to zone C, it is believed that there are vehicles in the direction from A to C. After traversing all the trajectories, traffic flow prediction results in each direction of the classification vehicle can be obtained. Vehicle types can also be obtained based on vehicle target detection technology, so each vehicle trajectory can be marked with category, so that the flow parameters of different types in each region can be counted.

4. Experiment and Result Analysis

The experiment only counts the vehicle flow. Based on the statistical data, more traffic flow parameters such as queue length and road proportion can be calculated. Three different intersection scenes including morning and evening rush hours are selected for experiments about 7 days, and the characteristics of experimental data are shown in Table 1.

For each scene, intersection partitions were obtained by combining k-means trajectory clustering through NURBS curve fitting, and vehicle trajectories in all directions were statistically analyzed to obtain specific results of vehicle flow prediction of different types.

Figure 1 is the vehicle statistical results of each intersection scene on a certain day by region using our system. Figure 2 is the traffic flow of the scene separately counted every day. Tables 2–4 are the comparison over 7 days between the average results obtained by our algorithm and the manual statistical results based on the absolute error. Among

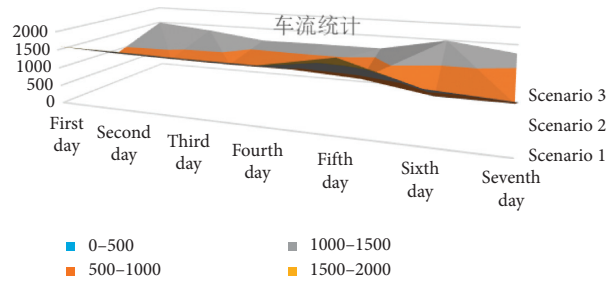


Figure 2: Vehicle statistical results of each day.

Table 2: Error analysis of average flow counting results in scenario 1.

From		To				
	Traffic flow	*A*	*B*	*C*	*D*	Total
A	Manual statistics	0	201	93	123	417
	Detection	0	292	90	114	379
	Error	0%	4.9%	3.3%	7.3%	9.1%
B	Manual statistics	187	0	45	601	833
	Detection	179	0	44	546	762
	Error	4.3%	0	2.2%	9.2%	8.5%
C	Manual statistics	75	88	0	43	206
	Detection	71	83	0	42	188
	Error	5.3%	5.7%	0	2.3%	8.7%
D	Manual statistics	78	189	46	0	313
	Detection	74	179	45	0	300
	Error	5.1%	5.3%	2.2%	0	4.2%

Table 3: Error analysis of average vehicle flow counting results in scenario 2.

From		To				
	Traffic flow	*A*	*B*	*C*	*D*	Total
A	Manual statistics	0	169	16	53	238
	Detection	0	154	16	52	224
	Error	0	8.9%	0%	1.9%	5.9%
B	Manual statistics	27	0	21	29	77
	Detection	27	0	21	28	77
	Error	0%	0	0%	3.4%	0%
C	Manual statistics	273	229	0	39	541
	Detection	267	208	0	39	496
	Error	2.3%	4.8%	0	0%	8.3%
D	Manual statistics	34	38	37	0	109
	Detection	34	37	37	0	102
	Error	0%	2.6%	0%	0	6.4%

them, N_E is the experimental value and N_T is the manual monitoring value.

$$\text{Error} = \left| \frac{N_E - N_T}{N_T} \right|. \tag{9}$$

The experimental results show that the statistical error of traffic flow is small, no matter for the *T*-junction or the intersection, and the maximum error is less than 10% and it can meet the needs of intelligent transportation applications. In view of the traffic condition unblocked scene, the target is shielded less, and the relative accuracy is slightly higher. For the scene of intersection, the vehicle target has more occlusion, the trajectory caused by it has fault discontinuity, which is effectively bridged through adoption of NURBS curve fitting to irregular points. Therefore, the vehicle flow error of *k*-means trajectory clustering statistics is much lower than that of the single-lane or multilane statistical method, and the new idea effectively solves the problems of incomplete tracking trajectory and complex calculations.

Table 4: Error analysis of average vehicle flow counting results in scenario 3.

From		To			
	Traffic flow	*A*	*B*	*C*	Total
A	Manual statistics	0	49	682	731
	Detection	0	49	628	664
	Error	0	0%	7.9%	9.1%
B	Manual statistics	313	0	72	385
	Detection	303	0	70	360
	Error	3.2%	0	3%	6.5%
C	Manual statistics	535	15	0	550
	Detection	512	15	0	521
	Error	4.3%	0%	0	5.3%

5. Conclusions

The real-time and accurate parameters of traffic flow at urban road intersections are the important basis for setting dynamic timing parameters of traffic signals at intersections. This paper mainly studies the vehicle flow counting method at urban intersections based on vehicle trajectory analysis. On the basis of the traditional method to obtain the tracking trajectory, the smooth track curve is fitted based on the NURBS algorithm, and a minimum distance measurement criterion is designed to select the initial clustering center for improving the *k*-means trajectory clustering method, then through subordinating the first and end points of the track to the intersection area, the vehicle movement mode is determined, and the traffic flow statistics of each area at the intersection are realized. Experimental results show that the method effectively solves the changeable mode and severe occlusion of moving vehicle problems, under the complex scene at the intersection, traffic statistical error is less than the classical method, and the algorithm is simple and can provide reliable data automatically for the traffic accident warning, road congestion prediction, and route planning.

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 by the Key Cultivation Program of Tibet under Grant 19MDZ03 and by the Science Research Program of Shaanxi Education Department under Grant 19JK0887.

References

[1] J. Zhang, F.-Y. Wang, K. Wang, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[2] A. M. Lin, *[IEEE 2018 Smart City Symposium Prague (SCSP)-Czech Republic (2018.5.24-2018.5.25)] 2018 Smart City Symposium Prague (SCSP)–Traffic*, Czech Technical University, Prague, Czech Republic, 2018.

[3] N. Seenouvong, U. Watchareeruetai, C. Nuthong et al., "A computer vision based vehicle detection and counting system," in *Proceedings of the 2016 8th International Conference on Knowledge and Smart Technology (KST)*, pp. 224–227, Chiangmai, Thailand, November 2016.

[4] S. Memon, S. Bhatti, L. A. Thebo et al., "A video based vehicle detection, counting and classification system," *International Journal of Image, Graphics & Signal Processing*, vol. 10, no. 9, 2018.

[5] L. H. Chen, *Intelligent Detection System Of Intersection Traffic Flow Based On Video Image Processing*, Anhui University of Science and Technology, Huainan, China, 2019.

[6] R.-A. Leonel, P.-P. Jose, A. Hernandez-Suarez et al., "Vehicle counting in video sequences: an incremental subspace learning approach," *Sensors*, vol. 19, no. 13, 2019.

[7] H. Rabbouch, F. Saadaoui, and R. Mraihi, "Unsupervised video summarization using cluster analysis for automatic vehicles counting and recognizing," *Neurocomputing*, vol. 260, 2017.

[8] F. Mehboob, M. Abbas, R. Jiang, S. A. Khan, and S. Rehman, "Trajectory based vehicle counting and anomalous event visualization in smart cities," *Cluster Computing*, vol. 21, no. 1, pp. 443–452, 2018.

[9] X. Rauf, X. Xiong, and H. Cheng, "Multi-vehicle flow detection method based on target tracking and transfer learning," *Journal of Guilin University of Electronic Science and Technology*, vol. 39, no. 2, pp. 119–123, 209.

[10] M. Liang, X. Huang, C.-H. Chen, and A. Tokuta, "Counting and classification of highway vehicles by regression analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2878–2888, 2015.

[11] L. Chen, *Research on Target Counting Method in Video Surveillance*, University of Science And Technology of China, Hefei, China, 2018.

[12] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Proceedings of Advances in Neural Information Processing Systems*, vol. 1, pp. 1324–1332, 2010.

[13] S. Zhang, G. Wu, P. Costeira, and J. M. F. Moura, "Understanding traffic density from large-scale web camera data," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[14] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, no. 5, pp. 3–16, 2018.

[15] Yu Liu and X. You-Lun, "Algorithm for searching nearestneighbor based on the bounded k-d tree," *Journal Of Huazhong University of Science and Technology (Nature Science Edition)*, vol. 36, no. 7, pp. 73–76, 2008.

[16] B. Sun, J.-H. Wang, D.-F. He, D. U. Hu-Bing, and B. Li, "Identification of aero-engine blade surface geometric defects with laser measurement," *Acta Automatica Sinica*, vol. 46, no. 3, pp. 594–599, 2020.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3167835, 11 pages
https://doi.org/10.1155/2020/3167835

Research Article

Research on Chinese Question-Answering for Gaokao Based on Graph

Zhizhuo Yang,[1] **Chunzhuan Li**,[1] **Zhang Hu**,[1] **Qian Yili**,[1] **and Ru Li**[1,2]

[1]*School of Computer and Information Technology of Shanxi University, Taiyuan, Shanxi 030006, China*
[2]*Key Laboratory of Computation Intelligence and Chinese Information Processing, Taiyuan, Shanxi 030006, China*

Correspondence should be addressed to Zhizhuo Yang; yangzhizhuo@sxu.edu.cn

Received 17 August 2020; Revised 24 October 2020; Accepted 26 October 2020; Published 6 November 2020

Academic Editor: Jun Shen

Reading comprehension Question-Answering (QA) for College Entrance Examination (Gaokao in Chinese) is a challenging AI task because it requires effective representation to capture complicated semantic relations between the question and answers. In this paper, a novel method of Chinese Automatic Question-Answering based on a graph is proposed. The method first uses the Chinese FrameNet and discourse topic (paragraph topic sentence and author's opinion sentence) to construct the affinity matrix between the question and candidate sentences and then employs the algorithm based on the graph to iteratively calculate the importance of each sentence. At last, the top 6 candidate answer sentences are selected based on the ranking scores. The recall on Beijing College Entrance Examination in the recent twelve years is 67.86%, which verifies the effectiveness of the method.

1. Introduction

Teaching the computer to pass the entrance examination of different education levels, which is an increasingly popular artificial intelligence challenge, has been taken up by researchers in several countries in recent years [1–3]. The Todai Robot Project [3] aims to develop a problem-solving system that can pass the University of Tokyo's entrance examination. China has launched a similar project "key technology and system for language question solving and answer generation," focusing on studying the human-like QA system for College Entrance Examination (commonly known as Gaokao). Gaokao is a national-wide standard examination for all senior middle school students in China and has been known for its large scale and strictness.

Although deep learning methods have achieved good results in many natural language processing tasks [4–7], they usually rely on a large scale of the dataset for effective training. However, the Gaokao task cannot receive sufficient training data under the current conditions. Different from previous typical QA tasks such as SQuAD [8], DuReader [9], and CMRC2018 [10] which can enjoy the advantage of holding a very large known QA pair set, the concerned task is equal to retrieving a proper answer from background article with guidelines of a very limited number of known QA pairs. In addition, the questions are usually given in an implicit way to ask students to dig the exactly expected meaning of the concerned facts. If such kind of meaning fails to fall into the feature representation for either question or answer, the retrieval will hardly be successful.

Generally speaking, for the Gaokao challenge, knowledge sources are extensive and no sufficient structured dataset is available, while the most existing work on knowledge representation focused on structured and semistructured types [11–14]. With regard to the answer retrieval, there are models based on semantic resources such as HowNet [15], WordNet [16], and Synonym Cilin [17]. Reference [18] proposed a sentence semantic relevance calculation method based on the multidimensional voting algorithm. This method considers the semantic relevance of different dimensions as a metric and uses the idea of the voting algorithm to select the best option for the problem. Reference [19] proposed a title selection method based on a correlation matrix between the title and the main points of the chapter. Reference [20] proposed a method for extracting candidate sentences based on frame matching and

frame relationship matching and then used manifold ranking to sort the candidate sentences.

This work focuses on reading comprehension question-answering in Gaokao Chinese examinations, which accounts for a large proportion of total scoring and is extremely difficult in the exams. Reference [2] made a preliminary attempt to take up the Gaokao challenge and proposed a three-stage approach that exploits and extends information retrieval techniques. Differently, this task is to solve reading comprehension questions and has to be based on deep semantic representation and computation rather than word matching in the previous work. Table 1 shows an example question in Chinese exams, consisting of a question and answer to the question. Some answer sentences are difficult to retrieve through literal matching, and these answer sentences are not distributed in a paragraph, but in different paragraphs of different articles. For instance, the question sentence would be confusing without knowing about the background article making cultural relics "live." In addition, some answers summarize the article from different paragraphs, while other answers summarize the author's point of view. How to retrieve those answers hidden in scattered paragraphs is a large challenge, and it is also the key to improving the effect of the system for Gaokao.

The challenge of our task would call for a new problem-solving framework for automatically answering comprehensive questions in exams. We propose a graph-based framework as shown in Figure 1. Firstly, we preprocess the articles and questions, and the evidence is drawn. Secondly, the Chinese FrameNet and discourse topic are used to construct the affinity matrix, which preserves the results of the semantic analysis of the question and each sentence. Finally, reasoning is performed by a graph-based ranking algorithm to check each candidate sentence, and the most relevent candidate sentence to the question will be returned as the answer.

Our contribution is threefold: (1) after showing Gaokao's difficulty and its difference from the existing research problems, we propose a new framework for reading comprehension QA in Gaokao. It is the first time to apply a graph-based algorithm in reading comprehension QA. (2) To the best of our knowledge, the relationship between candidate sentences has not been taken into account in the QA task. The relationship between candidate sentences is considered as a factor in our method, and the answer sentences are extracted by the unified model to improve the answering effect of the QA system. (3) Our approach achieves encouraging results on a set of real-life questions collected from recent Chinese examinations. We also release a Chinese comprehensive deep question-answering dataset to facilitate the research.

2. Reading Comprehension QA Method Based on Graph

2.1. Method Framework. The graph-based model [21] was firstly used by search engines to calculate the importance of webpages. It has been successfully used in many tasks, such as object retrieval [22], keyword extraction [23], and automatic summarization [24]. The algorithm is based on the following two assumptions. (1) Quantity assumption: in the web graph model, if a web page A is linked by a lot of other webpages, then page A is more important. (2) Quality assumption: if a page node A is linked by other higher-quality pages, then the A page is more important. The reading comprehension QA graph proposed in this paper is derived from the PageRank model. This model makes full use of the correlation between the question and candidate sentences. The global optimization ranking model is used to extract and sort the answer candidate sentences. The model is based on the following three hypotheses. (1) Quantity hypothesis: if an answer candidate sentence is associated with more other sentences, then the answer candidate sentence is more likely to be an answer sentence. (2) Quality hypothesis: if an answer candidate sentence is associated with other sentences of higher quality, then the answer candidate sentence is more likely to be an answer sentence. (3) Link weight hypothesis: the higher the degree of correlation between the question and the answer candidate sentence is, the more likely the answer candidate sentence is the answer sentence.

This paper makes use of the"voting" or "recommendations" between the question and sentences in the QA problem. The graph for reading comprehension QA is shown in Figure 2. The squares represent the candidate sentences $\{S_1, S_2, \ldots, S_n\}$ in the background article, and the edges between the squares represent the relationship between the candidate sentences, which is represented by the affinity matrix W_{ij}. The upper round node represents the question S_0. Usually, the College Entrance Examination has 1 or 2 questions. If there are 2 questions, they are merged into 1 sentence. The dotted line indicates the relationship between the question and candidate sentence nodes and is represented by the relationship matrix W_{0i} or W_{i0}. In the graph, the initial value of S_0 is set to 1, and the initial value of other candidate sentence nodes is 0. The importance of S_0 is passed to the candidate sentence node through the matrix W_{0i}. At the same time, the importance of the candidate sentences will also be strengthened with each other through the matrix W_{ij}. The importance of the candidate sentence nodes converges to a fixed set of values, and then the candidate sentence nodes are sorted according to the importance score. Finally, the top 6 sentences are selected as the final answer sentences. The difference between reading comprehension QA graph and PageRank graph is that, in PageRank network graph, the type of edge connecting nodes is the same, which indicates the recommendation of two website nodes; while the type of edge of reading comprehension graph is different, one is the edge between question and candidate sentence, which represents an association of answer or explanation. The other is the edge between candidate sentence nodes, which represents an association of similar contents between candidate sentences.

In this paper, the function $f: X \longrightarrow R$ is defined as a ranking function, which assigns a ranking score value f_i to each node S_i. f can be seen as a vector $f = [f_0, f_1, \ldots, f_n]^T$. The definition vector $y = [y_0, y_1, \ldots, y_n]^T$ represents the initial value of each node, where $y_0 = 1$, and the remaining $y_i = 0$. The algorithm is as shown follows:

Table 1: Example of reading comprehension QA in College Entrance Examination.

2017 Beijing College Entrance Examination question
Question: 请结合上述三则材料, 简述让文物“活”起来的含义与作用 Please combine the above three materials to briefly describe the meaning and function of making cultural relics “live.” 答案, 利用博物馆、各种现代技术让参观者近距离感悟文物的魅力。发挥它们在公众知史爱国, 鉴物审美, 以及技艺传承、文化养心的作用, 实现学术、趣味性统一, 以新鲜时尚的方式提供给观众审美与求知、娱乐与鉴赏的多元文化体验, 借助计算机等生成三维环境, 调动多感官, 带来沉浸感, 使用现代技术使得文物呈现方式灵活, 让更多的人喜欢上古文化, 更好地实现文物走近大众的作用。解决了展出空间有限、文物损毁等问题, 起到更好地保护文物的作用 Answer: use museums and various modern technologies to make visitors feel the charm of cultural relics up close. Play their role in public knowledge of history, patriotism, appreciation of objects, as well as technical inheritance, and cultural cultivation; achieve the unity of academic and interesting; provide audiences with a multicultural experience of aesthetics and knowledge, entertainment, and appreciation in a fresh and fashionable way; and use computers to generate a three-dimensional environment, mobilize multiple senses, and bring immersion; the use of modern technology makes the presentation of cultural relics flexible, so that more people like ancient culture, and better realize the role of cultural relics reaching the public. **(paragraph topic sentence)** It solves the problems of limited exhibition space and damage to cultural relics and plays a better role in protecting cultural relics. **(author's opinion sentence)**

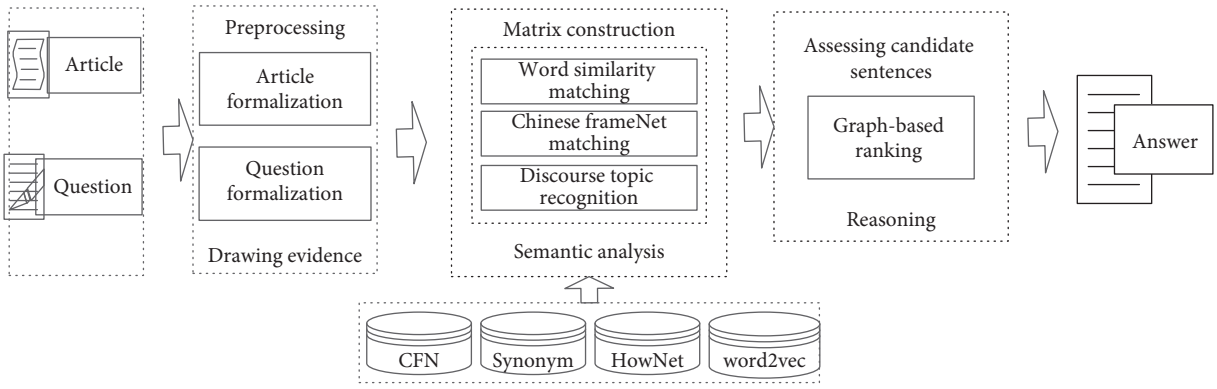


Figure 1: Overview of the approach.

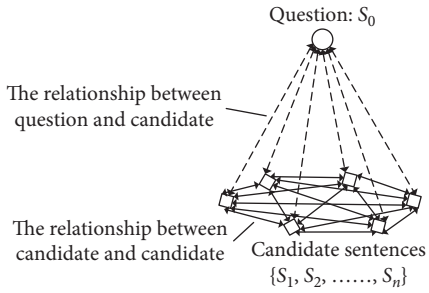


Figure 2: Reading comprehension QA graph for Gaokao.

In the first step of the algorithm, the relationship between the question S_0 and each candidate sentence $\{S_1, S_2, \ldots, S_n\}$ is calculated by the method based on word similarity matching, frame matching, and discourse topic. How to measure the relationship between the question and candidate sentences is the key step of the automatic QA method. For details, see Section 2.2.

In the second step of the algorithm, since the task of this paper is automatic QA, and the answer candidate sentences need to be extracted. The importance transmitted between candidate sentences should be related to the question, and the importance not related to the question should not be transmitted to each other. Therefore, the following formula is used to calculate the relationship between candidate sentences:

$$W_{ij} = \frac{\left(e_{i0} + e_{j0}\right)}{2}, \tag{1}$$

here $i, j \in [1, n]$ and e_{i0} and e_{0j} represent the similarity between the candidate sentences S_i and the question sentence S_0, respectively. The similarity of sentences is calculated by formula (5).

In the third step of the algorithm, the high-quality answer sentences are all explanations and answers to the question. The extraction effect depends largely on the relationship between the candidate sentences and the question, and it is less affected by the relationship between the candidate sentences. Therefore, different weights should be set for the affinity matrix of the two parts. $(\eta_1 > \eta_2)$ means that the relationship between the question and the candidate answer plays a larger role, and the relationship between the candidate sentences plays a smaller role. Previous studies have only focused on the relationship between the question sentence and answer sentences while ignoring the relationship between candidate sentences, but we believe that introducing the relationship between candidate sentences can improve the effect of the QA system. For example, a candidate sentence S_i is not only related to the question sentence but also related to other

Input: question S_0 and answer set $\{S_1, S_2, \ldots, S_n\}$, sentences initial value vector y.
Output: top 6 answer candidate sentences.

(1) Calculate the relationship between the question S_0 and each candidate sentence $\{S_1, S_2, \ldots, S_n\}$ by methods based on word similarity matching, frame matching, and discourse topic. If the degree of relationship between two nodes is greater than 0, the nodes are connected by an edge. Construct the affinity matrix $W_{0i} = W_{i0} = \text{relation}(S_0, S_i)$. In order to prevent the self-reinforcement of each node, let $W_{ii} = 0$.
(2) Calculate the relationship between each candidate sentence through the word similarity. If the degree of relationship between two nodes is greater than 0, the nodes are connected by an edge. Construct the affinity matrix W_{ij}, while $W_{ii} = 0$.
(3) Combine the affinity matrix and normalize it. Define $W = \eta_1(W_{0i} + W_{i0}) + \eta_2 W_{ij}$. Define the diagonal matrix D, where D_{ii} represents the sum of the i-th row of the W, and the W is normalized to $S = D^{(1/2)} W D^{-(1/2)}$.
(4) Iterate $f(t+1) = \alpha S f(t) + (1-\alpha) y$ until convergence, where $\alpha \in [0, 1]$.
(5) Use f_i^* to represent the convergence sequence $\{f_{i(t)}\}$, so that each sentence gets its ranking score.
Return top 6 candidate sentences with the highest score.

Algorithm 1: QA algorithm for reading comprehension based on a graph.

candidate sentences in the background article; then this candidate sentence S_i can represent other candidate sentences to a certain degree, so the candidate sentence is more likely to be the answer sentence. In this step, the affinity matrix is normalized to ensure the convergence of the iterative algorithm.

In the fourth step of the algorithm, α is a key parameter of the graph-based algorithm. This parameter can balance the impact of neighboring nodes and the initial scores of other nodes: the closer α is to 1, the greater the influence of neighboring nodes on the score; the closer α is to 0, the greater the influence of the initial score of nodes on the score. When the affinity matrix S satisfies the Markov process convergence conditions, the importance of the nodes converges. Usually, the convergence of the iteration algorithm is achieved when the difference between the scores computed at two successive iterations for any point falls below a given threshold (0.0001 in this paper).

By counting the suggested answers of the examination papers over several years, it is found that the average number of answer sentences is 6. If the number of outputs is less than 6 sentences, it is not enough to cover all answer points; if the number of outputs is greater than 6 sentences, the redundancy of the output answers is high. Finally, the top 6 candidate sentences are selected as answer sentences by the algorithm.

2.2. Calculation of the Relationship between the Question and Candidate Sentences. The calculation of the relationship between the question S_0 and each candidate sentence $\{S_1, S_2, \ldots, S_n\}$ directly affects the final answer. This paper uses a novel method based on CFN [20] and discourse topic to calculate the relationship of the question and candidate sentences. Our method takes into account sentence similarity, sentence frame matching, and discourse topic matching. The affinity matrix is used to record the relationship between the question and candidate sentences. The affinity matrix is a symmetric matrix. The calculation formula is as follows:

$$W_{0,i} = W_{i,0} = \lambda_1 * W_1 + \lambda_2 * W_2 + \lambda_3 * W_3 + \lambda_4 * W_4, \tag{2}$$

where $i \in [1, n]$, W_1 represents the sentence similarity matrix, W_2 represents the sentence frame matching matrix, W_3 represents the paragraph topic sentence matrix, W_4 represents the author's opinion sentence matrix, λ_k is the weight of the k-th dimension, $k \in [1, 4]$, and $0 \le \lambda_k \le 1$, $\sum_{k=1}^{4} \lambda_k = 1$. λ_k is used to adjust the weight of each matrix, and the value of the weight is set in the experiment.

2.2.1. Answer Sentence Extraction Based on Similarity Measure. First, preprocess the sentence, including word segmentation and removal of stop words. $S_0 = \langle k_1, k_2, \ldots, k_m \rangle$, $S_i = \langle w_1, w_2, \ldots, w_m \rangle$, and k_i and w_j represent the keywords of the question and candidate sentences, respectively; then, we combine HowNet [11] and word2vec [25] to calculate the similarity as follows:

$$\begin{aligned} \text{sum word} = {} & 0.4 \times \max_{1 \le i,j \le n} \left(\text{simHownet}\left(K_i, W_j\right)\right) \\ & + 0.6 \times \text{Cos}\left(K_i^v, W_j^v\right), \end{aligned} \tag{3}$$

where $\text{simHownet}(k_i, w_j)$ means calculating the similarity between the keyword k_i and w_j by HowNet. We use word2vec to calculate the cosine similarity of a word vector as follows:

$$\text{Cos}\left(k_i^v w_j^v\right) = \frac{k_i^v \cdot w_j^v}{\left(\left\|k_i^v\right\| \times \left\|w_j^v\right\|\right)}, \tag{4}$$

where k_i^v and w_j^v represent the word vectors of k_i and w_j. Finally, normalize (sumword_i) and the final calculation formula is

$$\begin{aligned} W_1 = W_{0i} = W_{i0} &= \text{Score}_{\text{sumWord}} \\ &= \frac{\text{sumword}_i}{\left\{\max_{1 \le i \le n}(\text{sumword}_i) - \min_{1 \le i \le n}(\text{sumword}_i)\right\}}. \end{aligned} \tag{5}$$

2.2.2. Answer Sentence Extraction Based on Frame Matching. Since the method based on similarity measure cannot mine the deep semantic information of the sentences in Gaokao, this paper uses the Chinese Frame Network (CFN) [26] to capture the semantic information in the semantic scene.

CFN is a Chinese vocabulary semantic knowledge base established by Shanxi University; it is based on FrameNet [27] of the University of California, Berkeley.

(1) Frame semantic matching: when the frame evoked by the target word of the question S_0 is the same frame evoked by the target word of the sentence S_i, the matching number is increased by one. (2) Frame semantic relationship matching: when the distance between the frame evoked by the question S_0 and the frame evoked by the sentence S_i is less than or equal to 2, the matching number is increased by one. Then, the frame matching number of the candidate sentence and the question is obtained. Finally, normalize it and the score based on frame matching is

$$W_2 = W_{0i} = W_{i0} = \text{Score}_{\text{sumFrame}} = \frac{\text{sumframe}_i}{\{\max_{1\le i\le n}(\text{sumframe}_i) - \min_{1\le i\le n}(\text{sumframe}_i)\}}. \tag{6}$$

An example of candidate sentence extraction based on frame matching is shown in Figure 3. The frame aroused by the target word "*development*" in question is the same as that aroused by the target word "*enhance*" in the candidate sentence; there is a relationship between the frame aroused by the target word "*development*" in the question and the frame aroused by the target word "*carry out*" in the candidate sentence. The involved scenes are relevant and the distance is less than or equal to 2. Therefore, the sentence S is extracted as an answer candidate sentence based on frame matching.

2.2.3. Answer Sentence Extraction Based on Discourse Topic. Through the study of the examination outline, it is found that College Entrance Examination often inspects the ability of students to summarize the main idea of the article. This paper proposes a method of extracting candidate sentences based on the discourse topic, which includes paragraph topic sentences and author opinion sentences.

2.2.4. Paragraph Topic Sentence Extraction. Through researching a large number of examination papers, it is found that the topic sentences are usually located at the beginning or end of the paragraph, and the sentence is usually related to the topic of other sentences in this paragraph. As shown in Table 1, "Use computers to generate a three-dimensional environment, mobilize multiple senses, and bring immersion," is located at the beginning of the paragraph and it is a topic sentence in the paragraph.

(1) Position Information. Paragraph topic sentence is a summary of the paragraph, which reflects the main idea of the paragraph. It is generally distributed at the beginning or end of the paragraph. Therefore, each sentence is calculated according to the position of the paragraph:

$$\text{score}_i = \begin{cases} 1, & i = 1, n, \\ 1 - \dfrac{\log i}{\log n}, & \text{others}, \end{cases} \tag{7}$$

where i is the sentence number and n is the total number of sentences in each paragraph.

For different paragraphs, in general, the first and last paragraphs of the article can reflect the topic of the article, so the weight of the first and last paragraphs should be greater, and the topic sentence of each paragraph is calculated according to the position of the paragraph:

$$\text{score}_{\text{loc}} = \begin{cases} 0.7 \times \text{score}_i, & i = 1 \text{ or } i = m, \\ 0.3 \times \text{score}_i, & \text{others}, \end{cases} \tag{8}$$

where m is the total number of paragraphs in the article.

(2) Semantic Similarity between Sentences Based on Paragraph. The keyword of sentence A is A_i, with p in total, and the keyword of sentence B is B_j, with q in total.

HowNet is used to calculate the similarity of sentences. The similarity of two words based on HowNet is $S(A_i, B_j)$. Let $a_i = \max\{S(A_i, B_1), S(A_i, B_2), \ldots, S(A_i, B_p)\}$, $b_j = \max\{S(B_j, A_1), S(B_j, A_2), \ldots, S(B_j, A_q)\}$; then, the similarity of sentences based on HowNet [11] is

$$\text{sim}(A, B) = \frac{\left(\sum_{i=1}^{p} a_i/p\right) + \left(\sum_{j=1}^{q} b_j/q\right)}{2}. \tag{9}$$

Then, the semantic similarity of sentence A based on paragraphs is

$$\text{score}_{\text{sim}} = \frac{\sum_{x=1}^{n} \text{sim}(A, B_x)}{n}, \tag{10}$$

where n is the total number of sentences in each paragraph.

Finally, the above factors are weighted to obtain the calculation formula as follows:

$$W_3 = W_{0i} = W_{i0} = \text{Score}_{\text{topic}} = \beta_1 * \text{score}_{\text{loc}} + \beta_2 * \text{score}_{\text{sim}}, \tag{11}$$

where $\beta_1 + \beta_2 = 1$.

2.2.5. Author's Opinion Sentence Extraction. It is found that the author's opinions and attitudes often appear in the suggested answers. The opinion sentences mainly indicate the author's viewpoint and attitude in the article, which are the overall grasp of the content and the topic of the whole discourse. Position information, similarity between sentences, and suggestive words are important features of author opinion sentences. As shown in Example 1, sentence S is the first sentence of the last paragraph in the article, and, secondly, the sentence is semantically related to other sentences, indicating the author's attitude in the whole article.

Example 1. 2018 Beijing College Entrance Examination.

Question: 根据材料一、材料二, 简要说明人类对人工智能的认识是如何不断深化的。

According to material one and material two, briefly explain how humans have continuously deepened their understanding of artificial intelligence.

Sentence: 面对人工智能可能带来的种种冲击, 上世纪50年代美国科幻小说家阿西莫夫提出的机器人三大定律, 今天对我们依然有借鉴意义。

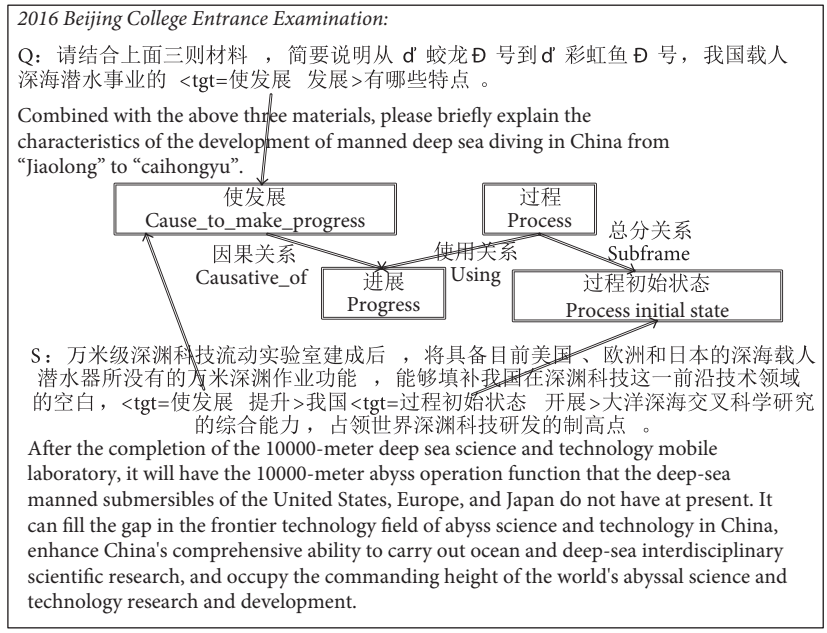


Figure 3: An example of candidate sentence extraction based on frame matching.

Faced with the various impacts that artificial intelligence may bring, the three laws of robotics proposed by the American science fiction novelist Asimov in the 1950s still have reference significance for us today.

(1) Position information

By analyzing the examination papers, it is found that the author's point of view is generally distributed at the end of the article, and the calculation is based on the different positions of the sentence in the last paragraph. The calculation formula is as formula (7), which is recorded as (score_i).

(2) Semantic similarity between sentences based on paragraph

The semantic similarity between sentences is calculated when extracting the author's opinion sentences. The calculation formula is as formula (10).

(3) Heuristic rules based on suggestive words

Candidate sentences are extracted based on whether the sentence contains suggestive words. If the sentence contains suggestive words, $\text{score}_{\text{Word}} = 1$; otherwise, $\text{score}_{\text{Word}} = 0$. This article expands the suggestive vocabulary through the CILIN [28]. Examples of suggestive words are shown in Table 2.

Finally, the above three factors are weighted to obtain the score of the author's opinion sentence:

$$\begin{aligned} W_4 &= W_{0i} = W_{i0} = \text{Score}_{\text{opinion}} \\ &= \gamma_1 * \text{score}_i + \gamma_2 * \text{score}_{\text{sim}} + \gamma_3 * \text{score}_{\text{Word}}, \end{aligned} \tag{12}$$

where $\gamma_1 + \gamma_2 + \gamma_3 = 1$.

3. Experiment and Result Analysis

3.1. Experimental Data. In the experiment, the language technology platform LTP [29] was used for word segmentation and part-of-speech tagging. The CFN [26] provided by Shanxi University was used for frame matching, and the HowNet [11] platform was used for word similarity calculation.

Due to the small proportion of questions in the College Entrance Examination, the dataset used in this paper includes the College Entrance Examination in each province, the simulation examination questions, and the questions transformed from multiple-choice questions. Finally, 132 questions were collected on the College Entrance Examination test in each province for the past 12 years, and 511 questions were collected on the simulation examination questions in each province. Each question consists of 1 or 2 questions. On average, each material contains 40 sentences and each sentence contains 30 Chinese words.

3.2. Experimental Results and Analysis

3.2.1. Comparison of Results of Different Experimental Methods. At present, the answers to the examination are graded according to the key points of the questions. In this paper, recall and accuracy are used as evaluation. A fivefold experiment was used to divide the corpus into five parts, one of which was used as the test set and the other four as the training set. The experiment was repeated five times, and the average value was taken as the final result. We manually find several answer sentences from the article according to the suggested answers, and mark them as the set A. S_A is the set of the top 6 sentences sorted by our method:

$$\begin{aligned} R &= \frac{\text{total sentences of correct sentences in } S_A}{\text{total sentences of } A} \times 100\%, \\ P &= \frac{\text{total sentences of correct sentences in } S_A}{\text{total sentences of } S_A} \times 100\%. \end{aligned} \tag{13}$$

Table 2: Examples of suggestive words.

看来、由此可见、由此看来、可见、无论如何、不管怎样、综上所述、由上述可知、如上所述、总的来看、总的来说、总之、总而言之、总体而言、首先、其次、表明、所以
It seems, thus it can be seen, it can be seen, anyway, no matter what, in summary, from the above, as mentioned above, in general, in general, in short, all in all, in general, first, second, show, so

To verify the effectiveness of the method in this paper, the method in this paper is compared with multiple baseline methods on Beijing College Entrance Examination questions for the past 12 years. The baseline methods include the following:

(1) Use frame matching [20] as baseline 1.

(2) We use the BERT model [30] as baseline 2. The model classifies answer candidates into two categories; in other words, it judges whether the candidate sentences in the article are answer sentences. College Entrance Examinations in all provinces except for Beijing and simulation examination questions (including 122 College Entrance Examination questions and 511 simulation examination questions, and we manually mark the answer sentences in the article according to the suggested answers) were used to train the model.

(3) The direct ranking method is used as baseline 3: the scores of each sentence are calculated by linear interpolation of word similarity matching, frame matching, and discourse topic, and the formula is as follows:

$$S = \phi_1 * \text{Score}_{\text{sumWord}} + \phi_2 * \text{Score}_{\text{sumFrame}} + \phi_3 * \text{Score}_{\text{topic}} + \phi_4 * \text{Score}_{\text{opinion}}, \quad (14)$$

where ϕ_k is the weight of the k-th dimension, $k \in [1, K]$, and $0 \le \phi_k \le 1$, $\sum_{k=1}^{K} \phi_k = 1$. In the experiment, we set $\phi_1 = 0.3, \phi_2 = 0.2, \phi_3 = 0.3, \phi_4 = 0.2$.

The experimental results are shown in Table 3.

There are many parameters in the method proposed in this paper, and these parameters are all based on experimental tests. Specifically, fix other parameters, take a parameter value from 0.0 to 1.0 in steps of 0.1, and test it, respectively. When the answer effect is the best, the parameter value is the final value. In Algorithm 1, $\eta_1 = 1.0$, $\eta_2 = 0.1$, $\alpha = 0.6$; when calculating the relationship between the question and candidate sentences, λ_k is set to 0.4, 0.2, 0.2, and 0.2. In the method of extracting the answer sentences based on discourse topic, $\beta_1 = 0.7$, $\beta_2 = 0.3$, $\gamma_1 = 0.3$, $\gamma_2 = 0.1$, and $\gamma_3 = 0.6$.

To compare with the international popular methods in reading comprehension for QA tasks, our method is compared with the deep learning method. It can be found that the recall of the BERT model is only 39.50%, which shows that the application of BERT in the College Entrance Examination is not good. The College Entrance Examination questions are more difficult than ordinary reading comprehension questions, and in the current scale of training data, we cannot train an efficient model which can capture complicated semantic relations between the question and answer. Moreover, the structure of the BERT model is very complex, and it is not easy to add rich linguistic knowledge to the model, which makes the model unable to adapt to the task in specific field.

When the direct ranking method is used, the recall and accuracy are 63.69% and 50.00%, respectively. When the QA method based on the graph is adopted, the recall and accuracy have been further improved. It should be noted that these two methods use the same external knowledge, but different algorithms to extract candidate sentences. The direct ranking method calculates the scores of each candidate sentence on each dimension and then performs a weighted sum of the scores of each dimension; the method based on the graph calculates the scores of each candidate sentence iteratively. The importance of the candidate sentences is transferred in the graph until it is finally stable. The experimental results show that our method can calculate the importance of each candidate sentence more reasonably and accurately.

3.2.2. Comparison of Results of Direct Ranking Method. To prove the advantages of the graph-based method, we use different methods in Section 2.2 to establish the affinity matrix of the question and candidate sentences and then use different methods to perform ranking. The experiment was carried out in the last 12 years of College Entrance Examination in Beijing, and the results are shown in Table 4.

It can be seen from Table 4 that when PageRank ranking is adopted, the experimental results are improved compared to the direct ranking method. The experimental results show the effectiveness of the iterative ranking method. Our method considers not only the relationship between the question and candidate sentences but also the relationship between candidate sentences. The algorithm based on graph can extract candidate sentences with both high degree of relevance to the question and high similarity with other candidate sentences. In addition, the experimental results also show that when four different methods (sord similarity + frame matching + paragraph topic sentence + author's opinion sentence) are used at the same time to extract candidate sentences, the experimental results have reached the optimal value whether it is direct ranking or PageRank ranking. It shows that various methods can make up for each other and jointly improve the effect of the system. λ_k of our method is set to 0.4, 0.2, 0.2, and 0.2, indicating that the word similarity method plays a greater role, while other methods play a smaller role. The last three methods can extract some answers that are not literally similar.

3.2.3. The Effect of $\eta_1 : \eta_2$ on the Experimental Results. $\eta_1 : \eta_2$ indicates the proportion of the relationship between candidate sentences and the question and the relationship between candidate sentences. The experiment was carried out on Beijing College Entrance Examinations, College

Table 3: Comparison results of different methods.

Method	R (%)	P (%)
Baseline 1 (frame matching)	50.48	35.00
Baseline 2 (BERT)	39.50	35.30
Baseline 3 (direct ranking method)	63.69	50.00
Automatic QA method based on graph	67.86	51.67

Table 4: The experimental results are compared with the direct ranking method.

Method		R (%)	P (%)
Word similarity	Direct ranking	48.57	33.33
	PageRank ranking	58.69	43.33
Paragraph topic sentence	Direct ranking	39.41	33.33
	PageRank ranking	50.48	36.67
Author's opinion sentence	Direct ranking	38.45	31.67
	PageRank ranking	51.79	38.33
Word similarity + frame matching	Direct ranking	52.74	36.67
	PageRank ranking	60.60	45.00
Word similarity + frame matching + paragraph topic sentence + author's opinion sentence	Direct ranking	63.69	50.00
	Automatic QA method based on graph	67.86	51.67

Entrance Examinations in other provinces, and simulation examination questions. The experiment fixed $\alpha = 0.6$. The results are shown in Figure 4. It can be found that when $\eta_1 : \eta_2 = 1 : 0.1$, the effect is the best. It proves that the relationship between candidate sentences is beneficial to QA in the College Entrance Examination, and the relationship between candidate sentences plays an auxiliary role, so η_1 is set larger and η_2 is set smaller.

3.2.4. The Effect of α on the Experimental Results. Experiments were carried out on Beijing College Entrance Examination questions, College Entrance Examination questions in other provinces, and simulated examination questions. $\eta_1 : \eta_2 = 1 : 0.1$ was fixed in the experiment. The results are shown in Figure 5. It can be found that the value of α has little effect on Beijing College Entrance Examination questions, while the best results are obtained when $\alpha = 0.6$ on College Entrance Examination questions and simulated questions in other provinces. The experiments show that neighboring nodes have a greater influence on candidate sentence scores, and initial score nodes have less influence on candidate sentence scores.

3.2.5. Differences between Real Questions in Different Provinces and Simulated Questions. It can be seen from Figures 4 and 5 that the method proposed in this paper has the best effect on Beijing College Entrance Examinations, but slightly worse on College Entrance Examinations in other provinces and simulation examinations. Because there are differences between them: the articles in Beijing College entrance examination are usually scientific and technological papers, while the articles in other provinces are mostly papers, academic papers, current reviews, book reviews, news, biographies, reports, popular science, and so on. In addition, most of the questions in Beijing College Entrance Examination are examined to select and integrate the information in the article, while most of the questions in other provinces are examined to understand the important words and sentences and grasp the structure and overall idea of the article. The recall of real and simulated questions in different provinces is shown in Figure 6.

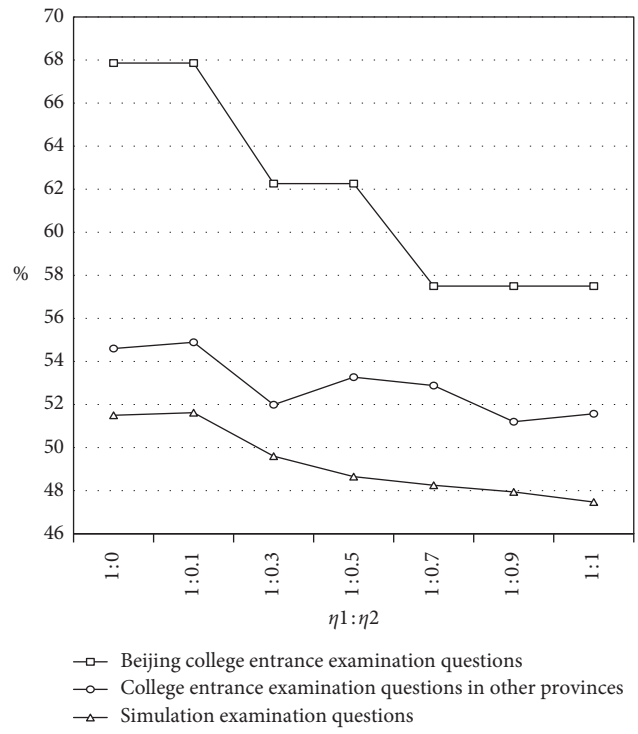


Figure 4: The effect of $\eta_1 : \eta_2$ on the experimental results.

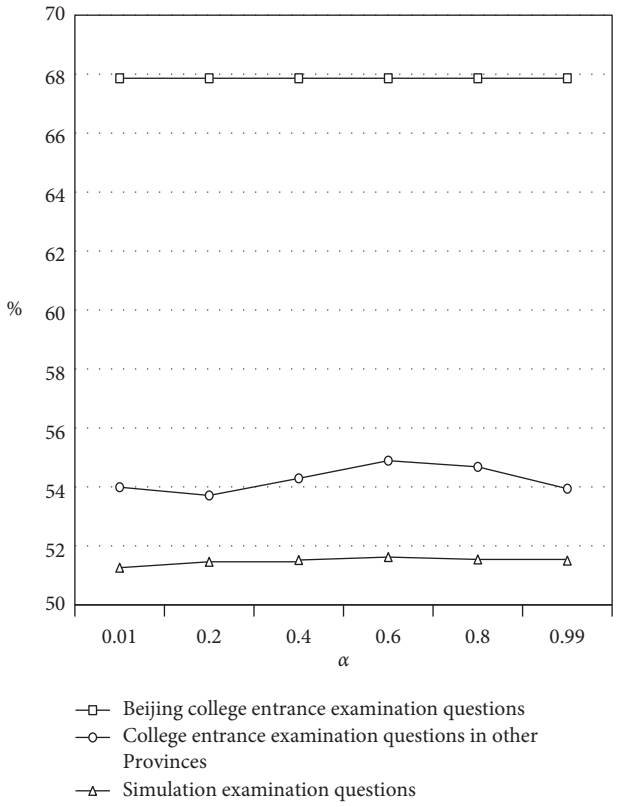


Figure 5: The effect of α on the experimental results.

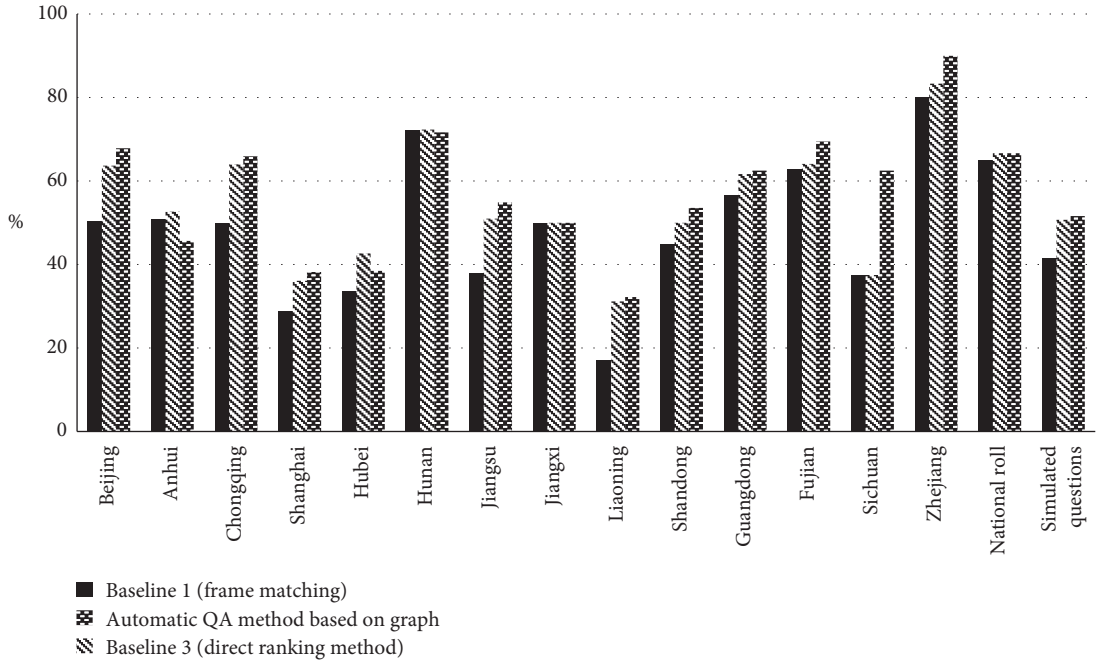


Figure 6: Recall rates of real and simulated questions in different provinces.

It can be seen from Figure 6 that our method can improve the experimental effect on real and simulated questions in different provinces. At the same time, it can be found that the recall of some provinces is relatively low, such as Example 2.

Example 2. 2009 Liaoning College Entrance Examination questions.

Question: "通俗历史热"在当今出现的原因是什么?

What is the reason for the emergence of "popular history fever" today?

Answer: "通俗历史热"是商品经济和文化教育发展到一定程度后定会出现的一种现象。

当商品经济趋于发达、文化教育发展迅速的时候, 人们在从事赖以谋生的职业活动之外, 带有文化色彩的业余需求会随之增长, 对作为文化存在常见形态之一的历史知识, 其"求解"欲望也会趋于强烈。

在当今市场经济逐步成熟、文化教育普及程度大为提高、高等教育开始走向大众化的时代, 人们的业余文化需求显著增长, 久远的尘封旧事引起了人们日益浓厚的兴趣。

对于广大民众而言, 在古奥难懂的传统史著和"学术模式"的现代史书皆难"卒读"的情况下, 通俗化的历史几乎成为他们"探寻过去"的唯一选择。

"Popular history fever" is a phenomenon that will surely appear after the development of the commodity economy and cultural education to a certain extent.

When the commodity economy tends to develop and cultural education develops rapidly, in addition to the professional activities that people rely on to make a living, the demand for culturally colored amateurs will increase accordingly. For historical knowledge as one of the common forms of cultural existence, its desire to "solve" will also become stronger.

In today's era, when the market economy is gradually maturing, the popularity of cultural education has greatly increased, higher education has begun to become popular, people's amateur cultural needs have increased significantly, and the dusty old things have aroused people's growing interest.

For the general public, under the circumstances that traditional historical books are difficult to understand in ancient times and modern history books of "academic mode" are difficult to "read," popularized history has almost become their only choice for "exploring the past."

Analyze the reasons and find the following: (1) the background material is discussed through the concept of "popular history fever" and many candidate sentences related to the question are not answer sentences, which need deep semantic understanding and reasoning technology. (2) It is found that there is a big semantic gap between "原因" in the question and the words such as "desire," "demand," "interest," and "choice" in the answer sentence. It is difficult for us to make semantic matching with existing tools such as HowNet, Word2Vector, and CFN.

The accuracy of extracting paragraph topic sentence and author's opinion sentence.

Annotate the paragraph topic sentences and author's opinion sentences on the Beijing 12 years College Entrance Examination. There are 19 materials, 89 paragraph topic sentences, and 26 author's opinion sentences. The experimental results are shown in Table 5.

Through the analysis of College Entrance Examination papers, it is found that, compared with the general news articles, it is more difficult to extract the topic sentence of the paragraph. As shown in Example 3, the topic sentence of the paragraph is "*Singing Kunqu Opera is something in the hall*" which is a concise summary of the paragraph. However, the similarity between topic sentences and other sentences is small, so it needs deeper semantic reasoning technology. The difficulty of extracting the author's opinion sentences is that some articles do not have a clear author's opinion. As shown in Example 3, the full text consists of four paragraphs. The first paragraph introduces "*Kunqu Opera*," and the next three paragraphs illustrate the strengths and limitations of "*Kunqu Opera*" from different perspectives, but there is no obvious general view and attitude.

Table 5: Experimental results of the paragraph topic sentence and author's opinion sentence.

Method	P (%)
Paragraph topic sentence recognition	80.62
Author's opinion sentence recognition	75.00

Example 3. 2009 Beijing College Entrance Examination

演唱昆曲是厅堂里的事情。地上铺了一方红地毯, 就算是剧中的境界, 唱的时候, 笛子是主要的乐器, 声音当然不会怎么响, 但是在一个厅堂里, 也就各处听得见了。搬上旧式的戏台去, 即使在一个并不宽广的戏院子里, 就不及平剧那样容易叫全体观众听清。如果搬上新式的舞台去, 那简直没有法子听, 大概坐在第五六排的人就只看见演员拂袖按鬓了。

Singing Kunqu Opera is something in the hall. There is a red carpet on the ground, even if it is the realm in the play; when singing, the flute is the main instrument, of course, the sound is not very loud, but in a hall, it can be heard everywhere. Moving on to an old-style theater, even in a theater that is not as wide as a theater, it is not as easy for the entire audience to hear. If you go to a new style stage, there is simply no way to listen. Perhaps the people sitting in the fifth and sixth rows will only see the actor's sleeves and temples.

4. Conclusion

After showing Gaokao's difficulty and its difference from the existing research problems, we propose a new framework for reading comprehension QA in Gaokao. The method first uses word similarity matching, frame matching, and discourse topic to construct the affinity matrix, which includes not only the relationship between the question and candidate sentences, but also the relationship between candidate sentences and then uses a graph-based algorithm to calculate the score of each sentence. Finally, the top 6 sentences are chosen as the answer sentences. At present, the deep reasoning ability of our method is not strong enough. In addition, the method in this article is extractive and cannot automatically generate some answers, so the score rate of the system is not high. In the next step, we will conduct a deep semantic understanding and reasoning on the background article and study a more efficient method. At the same time, we will further collect the relevant corpus to expand the scale of data and improve the answering effect of the system.

Data Availability

The data used to support the findings of this work 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 Key R&D Projects (2018YFB1005103), the National Natural Science Foundation of China (61772324), and the 1331 Engineering Project of Shanxi Province of China.

References

[1] S. Guo, X. Zeng, S. He, K. Liu, and J. Zhao, "Which is the effective way for Gaokao: information retrieval or neural networks?" in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics EACL*, pp. 111–120, Valencia, Spain, April 2017.

[2] C. Gong, W. Zhu, Z. Wang, J. Chen, and Y. Qu, "Taking up the Gaokao challenge: an information retrieval approach," in *Proceedings of the 2016 International Joint Conference on Artificial Intelligence IJCAI*, pp. 2479–2485, New York, NY, USA, July 2016.

[3] A. Fujita, A. Kameda, K. Ai, and Y. Miyao, "Overview of Todai robot project and evaluation framework of its Nlp-based problem solving," in *Proceedings of the International Conference on Learning Representations ICLR*, pp. 2590–2597, Banff, Canada, April 2014.

[4] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: a study and an open task," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding ASRU*, pp. 813–820, Scottsdale, AZ, USA, December 2015.

[5] J. Chen, Qi Zhang, P. Liu, X. Qiu, and X. Huang, "Implicit discourse relation detection via a deep architecture with gated relevance network," in *Proceedings of the ACL*, pp. 1726–1735, Berlin, Germany, August 2016.

[6] L. Qin, Z. Zhang, H. Zhao, Z. Hu, and E. P. Xing, "Adversarial connective-exploiting networks for implicit discourse relation classification," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pp. 1006–1017, Vancouver, Canada, July 2017.

[7] J. Cai, S. He, Z. Li, and H. Zhao, "A full end-to-end semantic role labeler, syntacticagnostic or syntactic-aware?" in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, NM, USA, August 2018.

[8] P. Rajpurkar, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the EMNLP 2016*, pp. 2383–2392, Association for Computational Linguistics, Austin, TX, USA, November 2016.

[9] W. He, "DuReader: a Chinese machine reading comprehension dataset from real-world applications," in *Proceedings of the MRQA 2018*, pp. 37–46, Melbourne, Australia, July 2018.

[10] Y. Cui, "A span-extraction dataset for Chinese machine reading comprehension," in *Proceedings of the EMNLP 2019 and 9th ICNLP*, pp. 5883–5889, Association for Computational Linguistics, HongKong, China, November 2019.

[11] D. Xiong, "A similarity calculation method of community question and answer based on LDA," *Journal of Chinese Information Processing*, vol. 26, no. 5, pp. 40–46, 2012.

[12] Z. Ye, "Research on open domain question answering system," in *Proceedings of the NLPCC 2015*, pp. 527–540, Springer, Nanchang, China, October 2015.

[13] L. T. Le, C. Shah, and E. Choi, "Assessing the quality of answers autonomously in community question–answering," *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 351–367, 2019.

[14] C. Li, "Syntactic analysis and deep neural network in answer extraction of Chinese question answering system," *Journal of Chinese Mini-Micro Computer Systems*, vol. 38, no. 6, pp. 1341–1346, 2017.

[15] Q. Liu, "Semantic similarity of vocabulary based on HowNet," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 7, no. 2, pp. 59–76, 2002.

[16] W. T. Yih, "Question answering using enhanced lexical semantic models," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1744–1753, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[17] Y. Zhou, "A method of sentence semantic similarity based on synonym forest and its application in question answering system," *Computer Applications and Software*, vol. 36, no. 8, pp. 65–68, 2019.

[18] S. Guo, "Sentence semantic relevance for college entrance examination reading comprehension," *Journal of Tsinghua University (Science and Technology)*, vol. 57, no. 6, pp. 575–579, 2017.

[19] Y. Guan, "A study on the selection of text titles for Chinese reading comprehension in college entrance examination," *Journal of Chinese Information Processing*, vol. 32, no. 6, pp. 28–35, 2018.

[20] G. Li, "The extraction of answer sentences from Chinese reading comprehension of college entrance examination based on frame semantics," *Journal of Chinese Information Processing*, vol. 30, no. 6, pp. 164–172, 2016.

[21] L. Page, "The PageRank citation ranking: bringing order to the web," Technical Report, Stanford InfoLab, Stanford University, Stanford, CA, USA, 1999.

[22] C. Fan, *Research on PageRank Algorithm in Web Structure Mining*, Soochow University, Suzhou, China, 2nd edition, 2009.

[23] J. Liu, "Keyword extraction based on language network," in *Proceedings of the 3rd National Conference on Information Retrieval and Content Security NCIRCS 2007*, pp. 711–715, Soochow University, Suzhou, China, November 2007.

[24] X. Wan, "An exploration of document impact on graph-based multi-document summarization," in *Proceedings of the EMNLP 2008*, Association for Computational Linguistics, Honolulu, HI, USA, October 2008.

[25] M. Liu, *Sentence Similarity Calculation Based on Word Vector and Its Application in Case-Based Machine Translation*, Beijing Institute of Technology, Beijing, China, 2nd edition, 2015.

[26] R. Li, *Research on the Semantic Structure Analysis Technology of Chinese Sentence Frame*, Shanxi University, Taiyuan, China, 2nd edition, 2012.

[27] C. F. Baker, "The berkeley framenet project," in *Proceedings of the 17th ICCL*, pp. 86–90, Association for Computational Linguistics, Chicago, IL, USA, May 1998.

[28] HIT IR-Lab Tongyici Cilin (Extended), http://www.ir-lab.org/.

[29] W. Che, "Ltp: a Chinese language technology platform," in *Proceedings of the International Conference on Coling*, pp. 13–16, Association for Computational Linguistics, Beijing, China, August 2010.

[30] J. Devlin, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, MN, USA, June 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3054969, 11 pages
https://doi.org/10.1155/2020/3054969

Research Article

Modeling of Marine Asynchronous Shaft Generator and Simulation of Subsynchronization State

Yan Langtao, Tan Jiawan, Liu Yusheng, and Yang Hui

School of Shipping and Naval Architecture, Chongqing Jiaotong University, Chongqing 400074, China

Correspondence should be addressed to Tan Jiawan; 6181141@qq.com

Received 16 August 2020; Revised 29 September 2020; Accepted 15 October 2020; Published 3 November 2020

Academic Editor: William Guo

The new type of marine asynchronous shaft generator has the advantages of adjustable excitation and power factor, compared to the traditional synchronous shaft generator, and has been widely used. However, the traditional synchronous shaft generator simulation system is still used in domestic ship power station simulators, which seriously restricts the renewal of crew training. In order to overcome the shortage of the simulation system of doubly fed shaft generator for ship power plant simulator, in this paper, the mathematical model of marine doubly fed shaft system is established for the first time, according to the characteristics of doubly fed machine and marine shaft generator. This paper realizes power decoupling by stator flux orientation and simulates and analyses the asynchronous shaft generator under subsynchronous working conditions. The changing trend of each physical quantity in the simulation waveform meets the mathematical relationships of the actual physical quantity, which proves the correctness of the mathematical model and lays a theoretical foundation for the development of the simulation system of asynchronous shaft generator.

1. Introduction

The marine asynchronous shaft generator is different from the traditional synchronous shaft generator. It has many remarkable characteristics, but the traditional synchronous shaft generator has no such characteristics. These remarkable characteristics include adjustable power factor, a wide range of speed variation, and small converter-capacity. The capacity of the converter is only slip power. It is suitable for the variable speed and constant frequency power generation system, such as the speed of the main engine of the ship, which frequently changes due to the complex and changeable sea conditions. Therefore, the asynchronous generator is more suitable for the marine shaft generator than the synchronous generator. The stator of the marine asynchronous shaft generator is connected to the marine power grid, and the rotor winding usually provides a three-phase low-frequency excitation current with adjustable amplitude, frequency, phase sequence, and phase through the dual PWM converter [1].

The marine asynchronous shaft generator is a new type of marine shaft generator, which has the function of energy saving and environmental protection and will gradually replace the traditional synchronous shaft generator. It will promote the renewal of the marine shaft generator system and gradually occupy the leading position of the marine shaft generator system. However, the renewal of marine mechanical and electrical equipment will bring the challenge of knowledge renewal and operation skill renewal to the marine engineers or electrical engineers. It is also a great difficulty to improve the management level and operation experience of the management of the marine asynchronous shaft generator system in a short time. The effective method is to use a ship power station simulator to train them without operating the real ship.

But at present, there is no marine asynchronous shaft generator system in the marine power plant simulator. The present situation makes the crew's training on the marine asynchronous shaft generator only conducted on the basis of synchronous shaft generator in the form of oral, without real operation on the marine asynchronous shaft generator. It

makes the updating of crew operation technology far behind the updating of existing electromechanical equipment. The building of the marine asynchronous shaft generator simulation system is of great significance to promote the development of ship electronic and electrical technology and even the development of the shipping business.

The necessary prerequisite for the development of the marine asynchronous shaft generator simulation system is the correct mathematical modeling of the system. The correctness and validity of the mathematical model are very important because only the correct and complete mathematical model can accurately reflect the typical characteristics and real-time dynamic process of the ship's shaft generator system [2, 3].

2. Mathematical Model

2.1. The Structure Diagram of the Marine Asynchronous Shaft Generator. According to the working principle of marine main engine and shaft generator, the structure diagram of the marine asynchronous shaft generator system is designed, as shown in Figure 1. The system uses the surplus power of the main engine driving the propeller to drive the asynchronous generator to generate electricity. The rotating speed on the main engine shaft of the ship will change with the sea state and waterway, which caused the rotor speed (represented by the letter n as shown in Figure 1) of the asynchronous generator to accordingly be changed. Therefore, the marine asynchronous shaft generator may work in different working states, such as subsynchronization, synchronization, and supersynchronization.

In order to ensure that the frequency of the stator output voltage of the marine asynchronous shaft generator is constant under different working conditions, the system adopts the dual PWM converter to realize AC excitation. The dual PWM converter rectifies the AC of the marine power grid into DC, and then inverts it into suitable AC with a certain amplitude, a certain frequency, and a certain phase, and sends it to the rotor of the marine asynchronous shaft generator for excitation.

It is assumed that the rotor winding and stator winding are symmetrical, and the number of pole pairs is p. When the voltage of frequency f_1 coming from the ship power grid is applied to the stator of the asynchronous generator, the stator winding will flow through the three-phase symmetrical alternating current, which creates a rotating magnetic field on the stator. The rotating speed of the rotating magnetic field is expressed as letter n_1. The relationship between the rotation speed of the rotating magnetic field and the pole number (expressed as letter p) can be shown as in the following formula:

$$n_1 = \frac{60 f_1}{p}. \tag{1}$$

Similarly, when the excitation current of the rotor with a certain frequency (expressed as letter f_2) is applied to the three-phase symmetrical rotor winding, a rotating magnetic field relative to the rotor itself will be generated. The corresponding speed and frequency are expressed as n_2 and f_2, respectively, which also meets the expression shown as formula (1).

It can be seen from formula (1) that the frequency (expressed as letter f_2) of excitation current coming from rotor determines the speed (expressed as letter n_2) of the corresponding magnetic field rotation [4], and the phase sequence of the excitation current determines the rotation direction of the corresponding magnetic field [5].

When the frequency of the ship power grid is 50 Hz, the speed of the stator rotating magnetic field of the marine asynchronous shaft generator is expressed by the letter n_1. To ensure that the frequency (expressed as letter f_1) of stator voltage of the marine asynchronous shaft generator is constant under different rotor speeds, the stator synchronous speed (expressed as letter n_1) should be constant.

The following expression can be obtained from the principle of Figure 1 and formula (1):

$$\begin{aligned} n_1 &= n \pm n_2, \\ \text{or } f_1 &= f \pm f_2. \end{aligned} \tag{2}$$

That is, when the rotor speed (expressed as letter n) is not equal to the synchronous speed (expressed as letter n_1) of the stator, the rotor excitation current frequency (expressed as letter f_2, the corresponding rotating magnetic field speed expressed as letter n_2) and the phase sequence of the rotor excitation current can be adjusted to ensure that the stator synchronous speed (expressed as letter n_1) or the corresponding frequency (expressed as letter f_1) is constant. There are three different relationships between the actual speed of the rotor and the synchronous speed of the stator. The three relationships are greater, equal, and smaller. According to these three different speed relations, the marine asynchronous shaft generator will work in three different states. They are supersynchronization, synchronization, and subsynchronization. In this paper, the subsynchronization working state of the marine asynchronous shaft generator is only discussed and analyzed.

2.2. The Modeling Assumptions. Before establishing the mathematical model of the marine asynchronous shaft generator, there are some assumptions as follows [6]:

(1) The stator winding and rotor winding are completely symmetrical, and the space position is 120° to each other. The magnetic motive force is sinusoidally distributed along the air gap circumference, and the harmonics are ignored.

(2) The saturation effect of the magnetic path is to be ignored. The mutual inductance and self-inductance of windings are assumed to be linear.

(3) The influence of frequency and temperature on the resistance of stator winding and rotor winding will be ignored.

(4) The hysteresis loss and eddy current loss will be ignored.

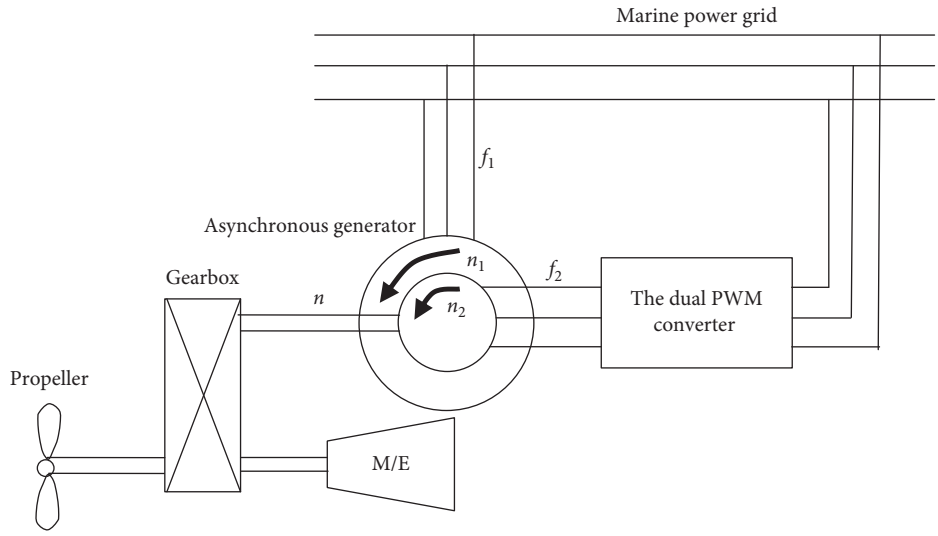


Figure 1: The structure of the marine asynchronous shaft generator system.

(5) The friction and power consumption between propeller main shaft and gearbox will not be considered; the friction and power consumption between gearbox and rotor will not be considered yet.

(6) The motor convention will be adopted on the direction of various physical quantities, including current, voltage, and the flux linkage of stator and rotor.

2.3. Mathematical Model

2.3.1. The Mathematical Model in d-q Coordinate System

(1) The Expression of Energy Balance. It can be seen from Figure 1 that the energy absorption of the marine asynchronous shaft generator consists of two parts. One part of the energy comes from the main engine of the ship, which is represented by the letter P_{SG}. The other part comes from the power converter, which is represented by the letter P_C. Two parts of energy are transferred to the ship power grid through the marine asynchronous shaft generator. In the case of ignoring the loss of the ship power grid, the total output power of the marine asynchronous shaft generator is equal to the total power consumed by all loads in the ship power grid [7, 8]. The energy balance equation of the system is as follows:

$$\begin{aligned} P_M &= P_P + P_{SG}, \\ (P_L)_{SG} &= (P_{SG} + P_C) \times \eta_{SG} = \sum_{i=1}^{n} P_i. \end{aligned} \tag{3}$$

In the energy balance equation shown as formula (3), the letter P_M is used to represent the power output of the main engine of the ship; the letter P_P stands for the power absorbed by the ship's propeller; the letter P_{SG} is used to represent the power transmitted by the main engine to shaft generator through gearbox; the letter $(P_L)_{SG}$ stands for the total power delivered by the marine asynchronous shaft generator to the ship's power grid; the letter P_C represents the electric power supplied by the dual PWM converter to rotor side of the asynchronous generator; the formula $\sum_{i=1}^{n} P_i$ is used to represent the electric power consumed by all electric loads on the ship under the current navigation conditions; the letter η_{SG} stands for the efficiency of the marine asynchronous shaft generator.

(2) The Expression of Flux Linkage. The flux-linkage expression of stator windings and rotor windings in the *d-q* coordinate system is described as follows:

$$\begin{aligned} \psi_{1d} &= L_1 i_{1d} + L_m i_{2d}, \\ \psi_{1q} &= L_1 i_{1q} + L_m i_{2q}, \\ \psi_{2d} &= L_2 i_{2d} + L_m i_{1d}, \\ \psi_{2q} &= L_2 i_{2q} + L_m i_{1q}. \end{aligned} \tag{4}$$

The footmarks of physical quantities of stator and rotor are represented by 1 and 2, respectively, in expression (4). The letter L represents inductance, so the letter L_1 represents the reactance of stator winding. The letter R represents resistance, so the letter R_1 represents the resistance of stator winding. Other letters have similar physical meanings.

(3) The Expression of Voltage. The voltage expression of stator windings and rotor windings in the *d-q* coordinate system is described as follows:

$$\begin{aligned} u_{1d} &= R_1 i_{1d} + p\psi_{1d} - \omega_1 \psi_{1q}, \\ u_{1q} &= R_1 i_{1q} + p\psi_{1q} + \omega_1 \psi_{1d}, \\ u_{2d} &= R_2 i_{2d} + p\psi_{2d} - \omega_s \psi_{2q}, \\ u_{2q} &= R_2 i_{2q} + p\psi_{2q} + \omega_s \psi_{2d}. \end{aligned} \tag{5}$$

In formula (5), the formula described as $\omega_s = \omega_1 - \omega_2$ is used to represent the slip electric angular velocity; the letter D stands for the differential operator.

Combiningg formula (4) and formula (5), the voltage expression can be rewritten as follows:

$$\begin{aligned} u_{1d} &= R_1 i_{1d} + \left(L_1 p i_{1d} + L_m p i_{2d}\right) - \omega_1 \left(L_1 i_{1q} + L_m i_{2q}\right), \\ u_{1q} &= R_1 i_{1q} + \left(L_1 p i_{1q} + L_m p i_{2q}\right) + \omega_1 \left(L_1 i_{1d} + L_m i_{2d}\right), \\ u_{2d} &= R_2 i_{2d} + \left(L_2 p i_{2d} + L_m p i_{1d}\right) - \omega_s \left(L_2 i_{2q} + L_m i_{1q}\right), \\ u_{2q} &= R_2 i_{2q} + \left(L_2 p i_{2q} + L_m p i_{1q}\right) + \omega_s \left(L_2 i_{2d} + L_m i_{1d}\right). \end{aligned} \tag{6}$$

(4) The Expression of Stator Power. The expression of stator active power and stator reactive power can be described as the following formula:

$$\begin{aligned} P_1 &= \frac{3}{2}\left(u_{1d} i_{1d} + u_{1q} i_{1q}\right), \\ Q_1 &= \frac{3}{2}\left(u_{1q} i_{1d} - u_{1d} i_{1q}\right). \end{aligned} \tag{7}$$

It can be seen from expression (7) and formula (6) that the voltage component and current component on the *d-q* axis are coupled and need to be decoupled.

2.3.2. The Mathematical Model in M-T Coordinate System. According to the power expression shown as formula (7), the active power and reactive power of the stator are determined by the voltage and current components on the *d-q* axis. As long as the voltage and current of the *d-q* axis can be controlled independently, the decoupling of active power and reactive power can be realized. In order to realize decoupling, the above mathematical model should be rewritten with stator flux orientation in *M-T* coordinate system. The vector relationship between various parameters of stator and rotor in different coordinate systems is shown in Figure 2.

According to the relationship between the parameters in Figure 2, the power expression (7) can be rewritten in *M-T* coordinate system as formula (8) and the relationship between the components of stator current and rotor current on the T-axis can be described as expression (9).

$$\begin{aligned} P_1 &= \frac{3}{2}\left(u_{M1} i_{M1} + u_{T1} i_{T1}\right), \\ Q_1 &= \frac{3}{2}\left(u_{M1} i_{T1} - u_{T1} i_{M1}\right), \end{aligned} \tag{8}$$

$$\begin{aligned} i_{T1} + i_{T2} &= -i_{d1}\sin\delta + i_{q1}\cos\delta - i_{d2}\sin\delta + i_{q2}\cos\delta \\ &= \left(i_{q1} + i_{q2}\right)\sin\delta - \left(i_{d1} + i_{d2}\right)\sin\delta \\ &= \left[\left(\frac{\psi_0}{L_M}\right)\sin\delta\right]\cos\delta - \left[\left(\frac{\psi_0}{L_M}\right)\cos\delta\right]\sin\delta = 0. \end{aligned} \tag{9}$$

In equation (9), the letter ψ_0 represents the flux linkage, shown as follows:

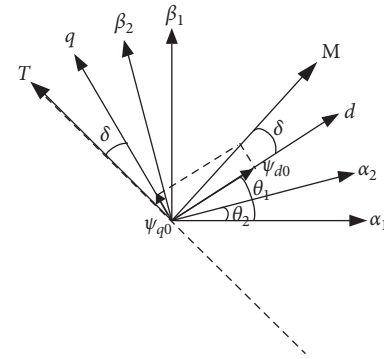


Figure 2: Vector diagram in the *M-T* frame.

$$\psi_0 = L_m\left(i_{M1} + i_{M2}\right). \tag{10}$$

A conclusion can be drawn from equation (9) that the sum of the two current components is equal to zero, described as follows:

$$i_{T1} = -i_{T2}. \tag{11}$$

If the flux linkage of the stator is oriented to the *M*-axis, the flux linkage (expressed as the letter ψ_0) and the component of stator voltage on *T*-axis will be constant; the component of stator voltage on *M*-axis is zero, which can be described as follows:

$$\begin{aligned} u_{M1} &= 0, \\ u_{T1} &= u_1. \end{aligned} \tag{12}$$

According to formula (9), formula (10), formula (11), and formula (12), the active power and the reactive power of stator shown as equation (8) can be rewritten as follows:

$$\begin{aligned} P_1 &= \frac{3}{2}\left(u_{M1} i_{M1} + u_{T1} i_{T1}\right) \\ &= \frac{3}{2}\left[-u_{M1} i_{M2} + \left(\frac{\psi_0}{L_m}\right)\cdot u_{M1} - u_{T1} i_{T2}\right] \approx \frac{3}{2} u_{T1} i_{T1} \\ &= -\frac{3}{2} u_{T1} i_{T2}, \end{aligned} \tag{13}$$

$$\begin{aligned} Q_1 &= \frac{3}{2}\left(u_{M1} i_{T1} - u_{T1} i_{M1}\right) = \frac{3}{2} u_{T1} i_{M1} \\ &= \frac{3}{2}\left[-u_{M1} i_{T2} - \left(\frac{\psi_0}{L_m}\right)\cdot u_{T1} + u_{T1} i_{M2}\right] \\ &\approx \frac{3}{2}\left[u_{T1} i_{M2} - \left(\frac{\psi_0}{L_m}\right)\cdot u_{T1}\right]. \end{aligned} \tag{14}$$

It can be seen from equation (13) that the stator active power (expressed as letter P_1) is proportional to the stator voltage component on the *T*-axis (expressed as the letter u_{T1}) and the rotor current component on the *T*-axis (expressed as the letter i_{T2}). And it can be seen from equation (14) that the stator reactive power Q_1 is related to three physical quantities which includes the stator voltage

component on the T-axis (expressed as the letter u_{T1}), the rotor current component on the M-axis (expressed as the letter u_{T1}), and the constant described as (ψ_0/L_m).

In order to realize decoupling, based on expression (13), the active-power control is taken as the outer control loop of the rotor current component on the M-axis, and the reactive power control is used as the outer control loop of the rotor current component on the T-axis. In other words, the output signal generated by the power controller is used as the given signal of rotor current control.

In this paper, the component of rotor input current on the M-axis and T-axis, represented by i_{M2} and i_{T2}, is used to control the output current of the stator side, denoted by i_{M1} and i_{T1}, respectively. Combined with literature 9 [9], IP control is adopted. In order to make the expression more concise, let $y_1 = i_{M1}$, $y_2 = i_{T1}$, $x_1 = i_{M2}$ and $x_2 = i_{T2}$. The following formula can be obtained according to expression (5):

$$\dot{y}_1 = \frac{L_2R_1}{L_m^2 - L_1L_2}y_1 + \frac{\omega_s L_m^2 - \omega_1 L_2L_2}{L_m^2 - L_1L_2}y_2 - \frac{L_mR_2}{L_m^2 - L_1L_2}x_1 - \frac{(\omega_1 - \omega_s)L_2L_m}{L_m^2 - L_1L_2}x_2 + \frac{L_mu_{M2} - L_2u_{M1}}{L_m^2 - L_1L_2}, \tag{15}$$

$$\dot{y}_2 = -\frac{\omega_s L_m^2 - \omega_1 L_2L_1}{L_m^2 - L_1L_2}y_1 + \frac{L_2R_1}{L_m^2 - L_1L_2}y_2 + \frac{(\omega_1 - \omega_s)L_2L_m}{L_m^2 - L_1L_2}x_1 - \frac{L_mR_2}{L_m^2 - L_1L_2}x_2 + \frac{L_mu_{T2} - L_2u_{T1}}{L_m^2 - L_1L_2}, \tag{16}$$

$$\dot{x}_1 = -\frac{L_mR_1}{L_m^2 - L_1L_2}y_1 + \frac{\omega_1 L_1L_m}{L_m^2 - L_1L_2}y_2 + \frac{L_1R_2 - \omega_s L_1L_m}{L_m^2 - L_1L_2}x_1 + \frac{\omega_1 L_m^2 - \omega_s L_1L_2}{L_m^2 - L_1L_2}x_2 + \frac{L_mu_{M1} - L_1u_{M2}}{L_m^2 - L_1L_2}, \tag{17}$$

$$\dot{x}_2 = -\frac{(\omega_1 - \omega_s)L_1L_m}{L_m^2 - L_1L_2}y_1 - \frac{L_mR_1}{L_m^2 - L_1L_2}y_2 - \frac{\omega_1 L_m^2 - \omega_s L_1L_2}{L_m^2 - L_1L_2}x_1 + \frac{L_1R_2}{L_m^2 - L_1L_2}x_2 + \frac{L_mu_{T1} - L_1u_{T2}}{L_m^2 - L_1L_2}. \tag{18}$$

From formula (15) to (18), there are four unknowns and four equations [10, 11]. After solving the four unknowns, they are carried into formula (13) and formula (14). According to formula (12), the decoupling expression of stator power can be obtained, shown as follows:

$$P = -\frac{3}{2}u_{T1}x_2 = -\frac{3}{2}ui_{T2},$$

$$Q = \frac{3}{2}\left[u_{T1}x_1 - \left(\frac{\psi_0}{L_m}\right)\cdot u_{T1}\right] = \frac{3}{2}\left[ui_{M2} - \left(\frac{\psi_0}{L_m}\right)\cdot u\right]. \tag{19}$$

3. Simulation

According to the above mathematical model, described as formula (9) to formula (16), the change process of torque, speed, current, voltage, reactive power, and active power of the marine asynchronous shaft generator in subsynchronous state is obtained by using simulation software named Matlab/Simulink.

3.1. System Simulation Module. According to the working mechanism of the marine asynchronous shaft generator, the functional characteristics of the rotor side converter and network side converter, the system part simulation module as shown in Figure 3 is built in Simulink environment.

In the simulation, the value of DC bus voltage (represented by the letter U_{DC}) is 460 V. The line voltage of the ship power grid is set at 190 V, and the transformation ratio of the transformer is 380/190. The three-phase symmetrical pure resistance load ($R = 50\,\Omega$) is star connected. Assuming that the speed of the ship's main engine is constant, the torque to the marine asynchronous shaft generator is 5 N·m, which makes it work in subsynchronous state. The parameters of the marine asynchronous shaft generator are shown in Table 1.

3.2. Simulation Waveform

3.2.1. Rotor Speed. When the ship sails stably at full speed on the sea, it is sent to the marine asynchronous shaft generator with constant torque.

The marine asynchronous shaft generator will start under this drive-torque, and the rotor speed will gradually increase and reach a stable state. The corresponding speed waveform is shown in Figure 4.

In Figure 4, after about 1.6 s, the rotor speed can be basically stable, about 1340 r/min, which is lower than the synchronous speed (represented by the letter n_1). The synchronous speed is 1500 r/min, because the pole-airs (represented by the letter p) are two, shown in Table 1. The marine asynchronous shaft generator will work in a subsynchronous state.

3.2.2. The Torque Waveform. When the ship sails steadily at full speed on the sea, the torque transmitted to the marine asynchronous shaft generator will be constant, and the corresponding electromagnetic torque change process of the marine asynchronous shaft generator is shown in Figure 5.

When time (represented by the letter t) is at zero, the marine asynchronous shaft generator is started by the main

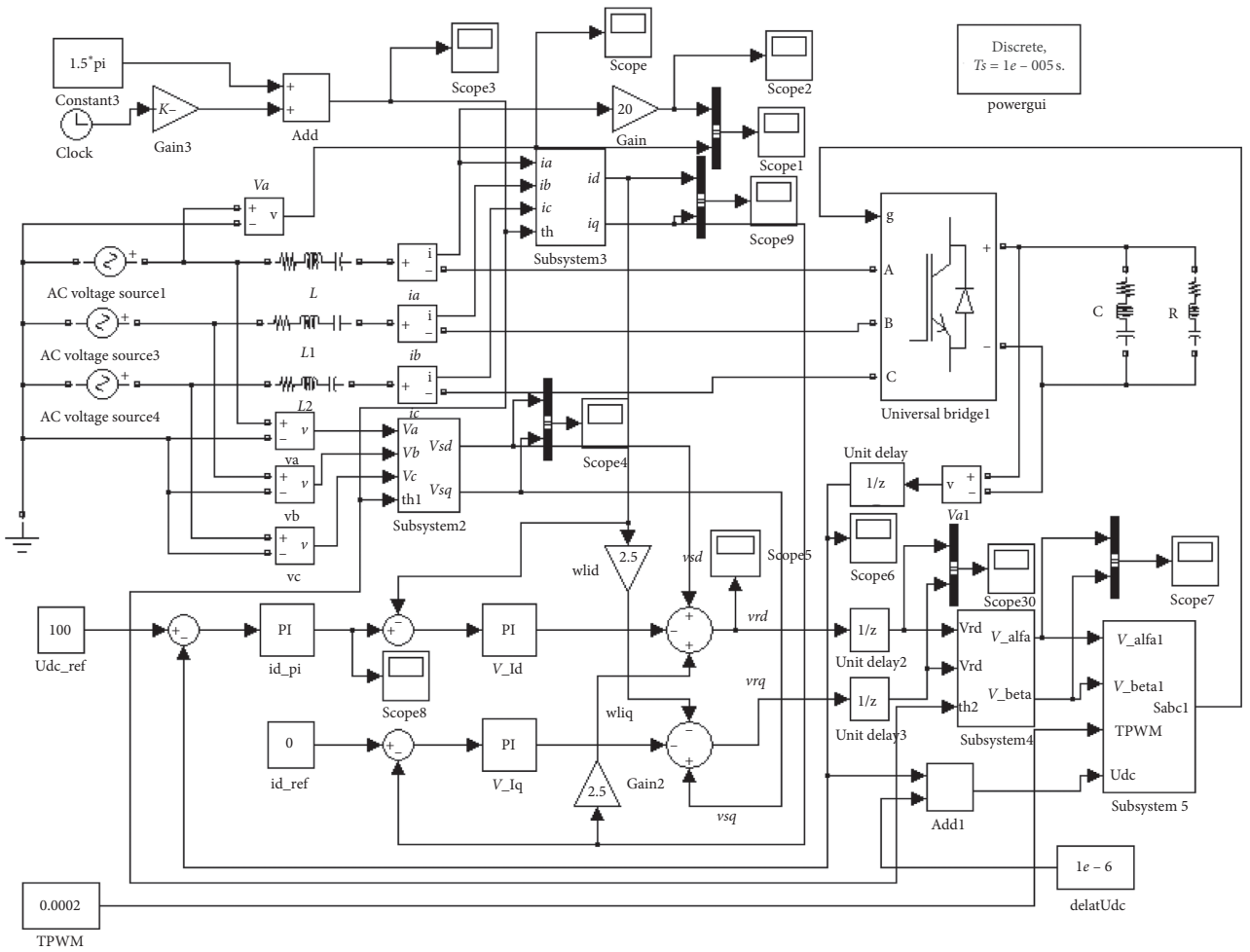


Figure 3: Simulation module of subsynchronization.

Table 1: Parameters of the marine asynchronous shaft generator.

Name	Value
U_e (V)	380
P_e (kW)	6
f_1 (Hz)	50
P	2
J (kg·m^2)	0.05
R_1 (Ω)	1.37
L_1 (H)	0.1613
R_2 (Ω)	1.65
L_2 (H)	0.9955
L_{12} (H)	0.1588

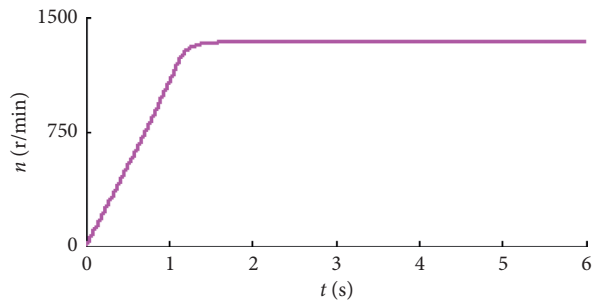


Figure 4: Rotor speed at subsynchronization.

engine driving torque. During the starting period of 0–1.6 s, the torque waveform changes greatly. After 1.6 s of stability, the electromagnetic torque is about −5 N·m. The electromagnetic torque is negative, which means the direction of electromagnetic torque is opposite to that of the rotor speed.

3.2.3. The Line Voltage of Rotor Side Converter. The line voltage between phase A and phase B of the rotor side converter is shown in Figure 6, which is about 460 V and basically stable, and will provide excitation for the rotor. The voltage value shown in Figure 6 is consistent with the voltage setting value of 460 V of the double PWM intermediate DC bus. It can be seen from Figure 6 that the voltage frequency is relatively high before 1.6 seconds in the starting stage of the marine asynchronous shaft generator, and after 1.6 seconds, the frequency becomes lower and almost unchanged. The voltage waveform tends to be stable. And the marine asynchronous shaft generator is in a stable subsynchronous state.

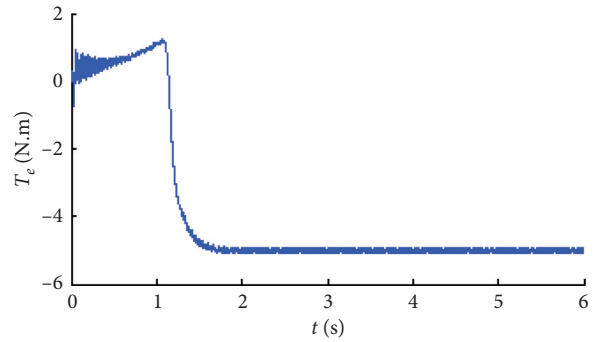


Figure 5: The electromagnetic torque.

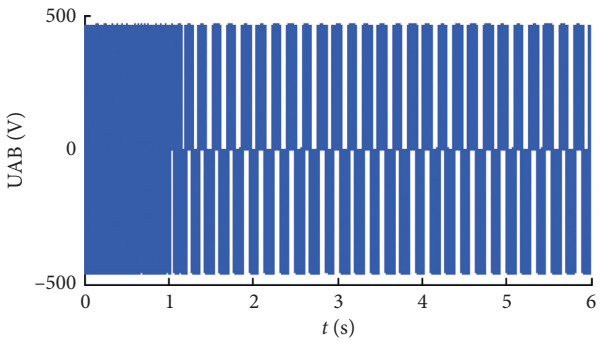


Figure 6: Output line voltage of the rotor side converter.

3.2.4. The Voltage Component and Current Component of Rotor on d-Axis and q-Axis. The components of rotor voltage and rotor current on the d-axis and q-axis are shown in Figures 7 and 8, which tend to be stable at about 1.6 seconds. The unstable state before 1.6 seconds is defined as the starting process of the marine asynchronous shaft generator.

It can be seen from Figures 7 and 8 that the components of rotor voltage and rotor current on the d-axis and q-axis are constant after stability, without alternating, and the frequency is zero.

3.2.5. Three-Phase Current of Rotor. The Rotor current waveform of phase A is shown in Figure 9(a), and the rotor current waveform of phase ABC is shown in Figure 9(b). The marine asynchronous shaft generator enters the stable state about 1.6 seconds after starting, and the magnitude and frequency of rotor excitation current remain unchanged after stability.

3.2.6. Active Power and Reactive Power. The simulation waveforms of active power and reactive power in the sub-synchronous state are shown in Figure 10. The negative sign shown in Figure 10 means the power of the generator is output. The reactive power and active power are stable after 1.6 seconds, and the reactive power (represented by the letter Q_1) is about zero after stabilization, the active power (represented by the letter P_1) is approximately constant.

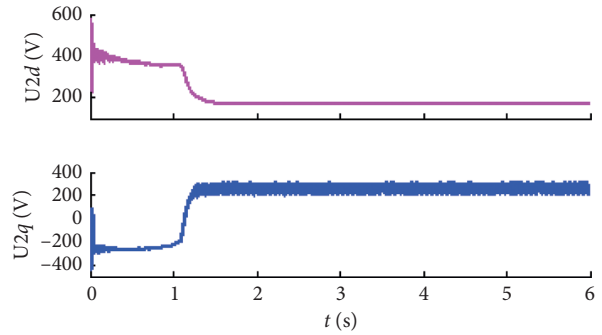


Figure 7: Component of rotor voltage on the d-axis and q-axis.

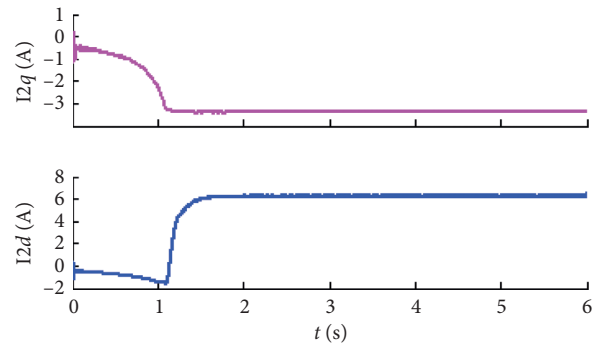


Figure 8: Component of rotor current on the d-axis and q-axis.

3.2.7. Stator Voltage and Stator Current. In order to display the changing trend of stator phase A current and phase A voltage in the same coordinate, the current is amplified by −10 times. The magnification is negative because the motor convention is used in the modeling process; the phase of voltage and current is opposite. Otherwise, because the current value is too small, the changing trend of current cannot be clearly seen in the same coordinate system with voltage. The simulation waveform of the stator voltage waveform and stator current is shown in Figure 11(a), with the negative ten times current waveform.

It can be read from Figure 11(a) that the period of stator A-phase voltage and A-phase current is 0.02 s. That is, the frequency is 50 Hz. Since the given load is pure resistive load, the voltage and current in the simulation results are in the same frequency and phase, and the actual current waveform of stator A-phase is shown in Figure 11(b).

3.3. Result Analysis of Simulation Waveform. The voltage and current of stator A-phase in Figure 11(a) are in the same frequency and phase. It can be deduced that the reactive power is zero, which is consistent with the simulation waveform of zero reactive power described in Figure 10 and consistent with the simulation premise that the load is pure resistance load.

The stator phase A current (the peak-value is about 3 A) is shown in Figure 11(b), and the stator voltage (the peak-value is about 160 V) is shown in Figure 11(a), and the corresponding power including the active power and reactive power can be calculated; the power factor also can be

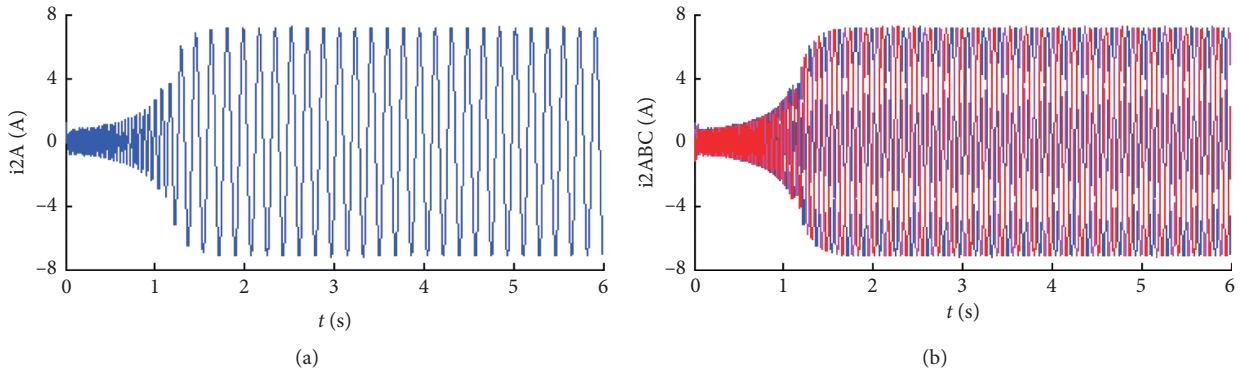


Figure 9: Rotor current in the subsynchronous state. (a) Phase A. (b) Phase ABC.

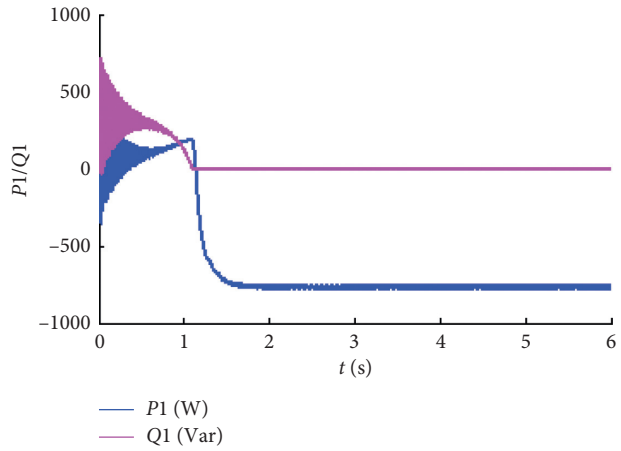


Figure 10: Active power and reactive power.

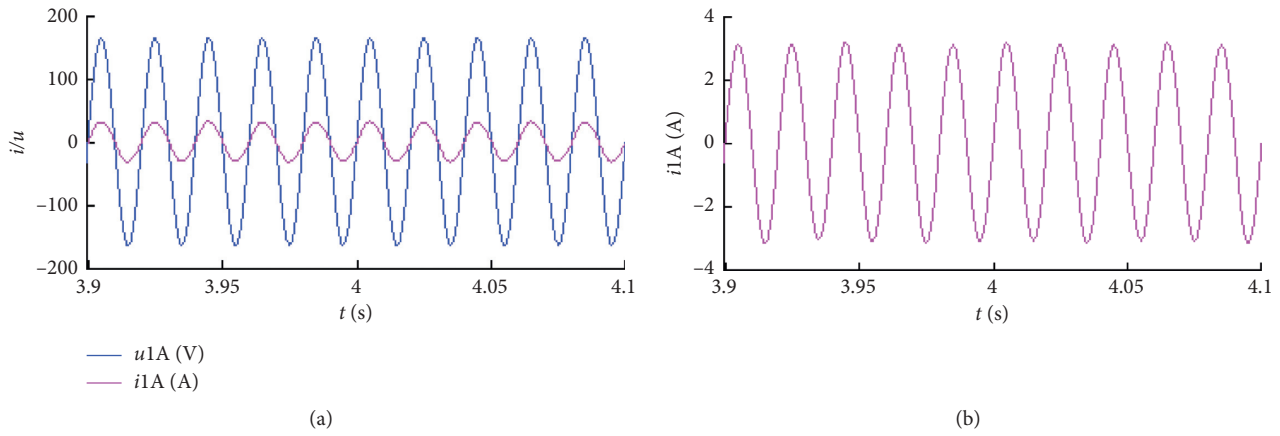


Figure 11: Stator voltage and stator current in subsynchronous state. (a) Voltage and current (–10 times). (b) Actual stator current.

calculated. The active power is calculated to be about 720 W, and the reactive power is zero, the corresponding power factor is 1. The calculated results are consistent with the simulation waveforms described in Figure 10.

According to the calculated active power of 720 W and the speed of 1340 r/min described in Figure 4, the corresponding torque can be calculated under the premise of modeling, which is about 5 N·m. The torque calculation results are consistent with the electromagnetic simulation waveform shown in Figure 6.

According to the speed waveform of 1340 r/min shown in Figure 4, the corresponding frequency of rotor speed can

be calculated, which is about 44.7 Hz. The sum of this frequency with a value of 44.7 Hz and the rotor excitation current frequency is shown in Figure 9 (about 5.5 Hz) is 50.17 Hz, which is basically consistent with the corresponding frequency (50 Hz) of stator synchronous speed because there may be some errors in reading the pictures.

4. Experiment

4.1. Overview of Experimental Platform. The hardware part of the marine asynchronous shaft generator experimental platform in the laboratory mainly includes the main circuit and control circuit, which is mainly composed of variable frequency motor used to simulate the main engine of a ship, asynchronous generator, transformer, frequency converter, back-to-back dual PWM converter, reactors, and grid-connected relay. The schematic diagram is shown in Figure 12.

In Figure 12, the city power on the laboratory wall is used to simulate the 380 V ship power grid. The frequency converter drives the variable frequency motor to simulate the main engine of the ship. Different frequencies are set to simulate the speed of the main engine under different working conditions. The variable frequency motor drives the asynchronous generator to generate electricity. The above equipment together constitutes the marine asynchronous shaft generator experimental platform. The stator output voltage of the doubly fed shaft generator can be connected to the power grid of the ship through the main switch shown in Figure 12 after the transformer step-down. The transformation ratio of the three-phase transformer is 2 (380/190). At the same time, the voltage is sent to the rectifier stage of the dual PWM converter. And the inverter stage of the dual PWM converter outputs sinusoidal AC to the rotor of the doubly fed shaft generator for excitation.

The hardware layout of the experimental platform is shown in Figure 13. The power of the frequency converter, variable frequency motor, and the asynchronous generator is 3.7 kW, 7.5 kW, and 6 kW, respectively, and the pole pairs of the variable frequency motor and the doubly fed generator are both 2. The maximum speed of the variable frequency motor and the doubly fed generator is 1800 r/min.

4.2. The Experimental Results. In the experiment, the grid side converter is powered on and the input voltage is 190 V. The voltage with 190 V is obtained by reducing the voltage of 380 V through a transformer, with a transformation ratio of 2, as shown in Figure 12. The dual PWM converter is put into operation for rectification. The voltage on the intermediate DC bus will increase gradually and finally stabilizes at 460 V, which is prepared for rotor excitation.

The speed of the simulated main engine is set by frequency conversion, and at the same time, the rotor of the asynchronous generator is driven to rotate. The frequency of the frequency converter is changed to realize the speed change of the marine main engine simulated by the variable frequency motor. The rotor speed is stable at 1340 r/min by changing the frequency of the frequency converter. The marine asynchronous shaft generator will work in subsynchronous state. The load in the experiment is a three-phase symmetrical pure resistance load (represented by the letter R_L, $R_L = 50\,\Omega$), and star connected.

The waveforms of the experimental results are shown in Figures 14– 16.

4.2.1. The Waveform of DC Bus Voltage. The stable DC bus voltage waveform is shown in Figure 14, and it is 460 V after stabilization, which lays the foundation for providing a suitable rotor excitation current. The experimental waveform is consistent with the 460 V of simulation waveform shown in Figure 6.

4.2.2. Rotor Excitation Current Waveform in Subsynchronous State. The experimental waveform of the three-phase current of rotor winding during subsynchronous stable operation is shown in Figure 15.

The period of rotor excitation current under subsynchronous stable state can be read out from Figure 15 as follows:

$$T \approx \frac{30\,\text{ms}}{\text{grid}} \times 6\,\text{grid} = 180\,\text{ms} = 0.18\,\text{s}. \tag{20}$$

From formula (20), the period (represented by the letter T) of rotor excitation current is 0.18 seconds. It can be calculated that the rotor excitation current frequency (represented by the letter f_2) is as follows:

$$f_2 = \frac{1}{T} = \frac{1}{0.18} \approx 5.55. \tag{21}$$

Because of the existence of error in reading graphs, the calculated frequency described as formula (21) is basically consistent with the slip frequency corresponding to the rotor speed of 1340 r/min. And the slip frequency is represented by the letter f_s, shown in the following formula:

$$f_s = \frac{\Delta n \times p}{60} = \frac{(1500 - 1340) \times 2}{60} \approx 5.33. \tag{22}$$

4.2.3. Stator Voltage and Stator Current. The voltage and current of stator phase A are shown in Figure 16.

In order to be able to see the changing trend and phase relationship of voltage and current in the same coordinate system, the output waveform of experimental current shown in Figure 16 was amplified by 4 times of actual current. It can be seen from Figure 16 that the stator voltage and current are sinusoidal AC with the same period and same frequency, the amplitude of the voltage is 160 V, and the amplitude of the current is about 3 A. The period (represented by the letter T) can be calculated as follows:

$$T = \frac{10\,\text{ms}}{\text{grid}} \times 2\,\text{grid} = 20\,\text{ms} = 0.02\,\text{s}. \tag{23}$$

The phase between the stator current of phase A and the stator voltage of phase A is the opposite. The reason is that the current direction in the experimental measurement is

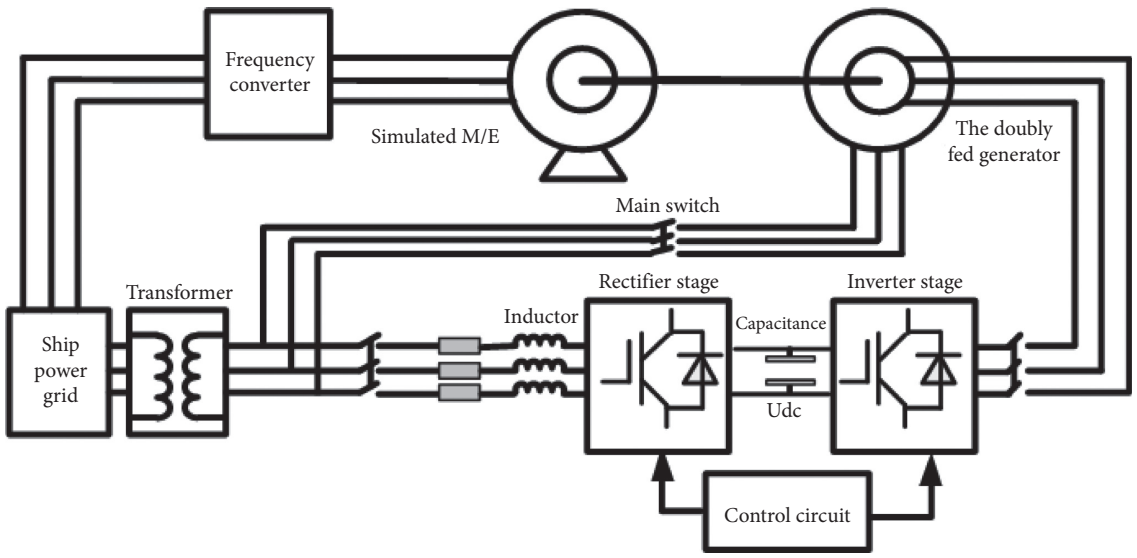


Figure 12: Schematic diagram of the experimental platform.

(a) (b) (c)

Figure 13: Experimental platform photos. (a) Prime mover and generator. (b) Transformer and reactor. (c) The dual PWM converter.

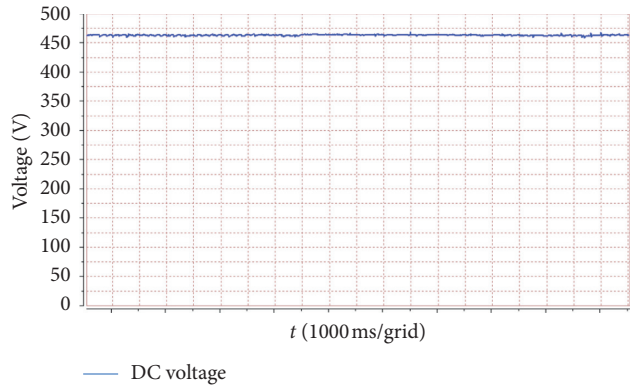


Figure 14: DC bus voltage 460 V.

opposite to the actual current direction of the generator. The actual current waveform direction is the reverse of the current waveform in Figure 16. The phase difference is 180°, as is shown in Figure 16. Therefore, in Figure 16, the variation trend of stator output voltage and stator current is consistent with that of the simulation waveform shown in Figure 11.

4.3. Summary. The following conclusions can be drawn from the comparison between the simulation waveform and the experimental waveform:

(1) The value, frequency, and changing trend of stator and rotor parameters in the simulation diagram are consistent with the experimental results.

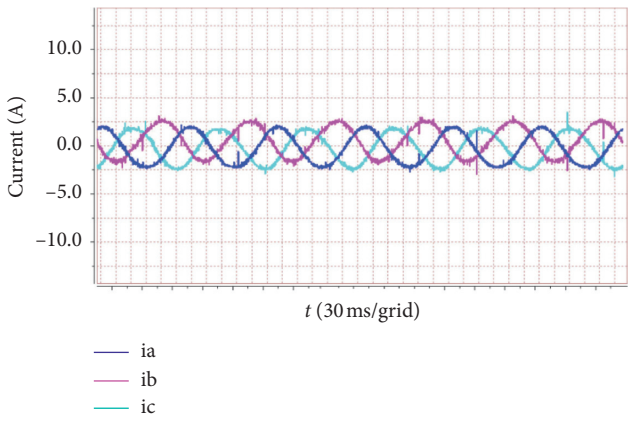


Figure 15: Rotor current in subsynchronous stable operation.

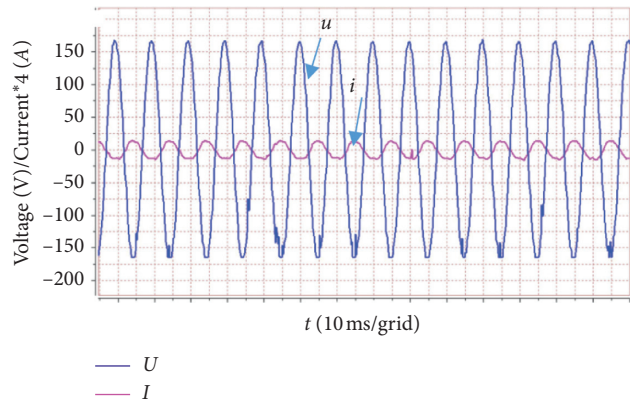


Figure 16: Stator voltage and stator current.

(2) Because the influence of temperature, frequency, and other physical parameters on the stator and rotor winding resistance is ignored in the modeling process, the simulation waveform results are generally smoother than the experimental waveform, and the harmonics are less.

(3) The above simulation results are in good agreement with the actual situation, which proves the correctness of the mathematical model, which is helpful to the development of the marine asynchronous shaft generator simulator.

5. Conclusion

In this paper, the modeling and subsynchronous simulation of a new marine asynchronous shaft generator are carried out. The relationship between the physical quantities in the simulation waveform is consistent with the actual situation, which is verified by experiments. The consistency of simulation results and experimental waveforms proves the correctness of the mathematical model. It lays a theoretical foundation for the development of a virtual simulation system and fault diagnosis system of the marine asynchronous shaft generator, which is helpful for crew training.

Data Availability

All data generated or analyzed during this study are included in this article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Zhu, Z. Chen, Y. Tang, F. Deng, and X. Wu, "Dual-loop control strategy for DFIG-based wind turbines under grid voltage disturbances," *IEEE Transactions on Power Electronics*, vol. 31, no. 3, pp. 2239–2253, 2016.

[2] S. Lekhchine, T. Bahi, I. Abadlia, Z. Layate, and H. Bouzeria, "Speed control of doubly fed induction motor," *Energy Procedia*, vol. 74, no. 74, pp. 575–586, 2015.

[3] T.-K. T. Layate, N.-V. Nguyen, An efficient four-state zero common-mode voltage PWM scheme with reduced current distortion for a three-level inverter," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1021–1030, 2018.

[4] F. Xiong, X.-f. Wang, and B. Hua, "D-q axis mathematical model of wound-rotor brushless doubly-fed machine," *Electric Machines and Control*, vol. 19, no. 5, pp. 81–88, 2015.

[5] P. Ma, W. Liu, and S. Mao, "Dead-time compensation method of single-phase AC excitation for three-stage brushless synchronous machines," *Zhongnan Daxue Xuebao (Ziran Kexue Ban)*, vol. 47, no. 12, pp. 4048–4055, 2016.

[6] K. Yu and P. Tang, "Equivalent circuit model and characteristic analysis of brushless douly-fed machine," *Proceeding of the CSEE*, vol. 38, no. 14, pp. 4222–4232, 2018.

[7] G. Liu, X. Wang, and F. Xiong, "A "Π" –type equivalent circuit of wound rotor brushless doubly-fed machines," *Proceedings of the CSEE*, vol. 36, no. 20, pp. 5632–5639, 2016.

[8] T. Trivedi, R. Jadeja, and P. Bhatt, "Improved direct power control of shunt active power filter with minimum reactive power variation and minimum apparent power variation approaches," *Journal of Electrical Engineering and Technology*, vol. 12, no. 3, pp. 1124–1136, 2017.

[9] Y. Lang-tao, D. Wang, and S.-l. Wang, "Modeling and simulation of DFIG decoupling control based on IP control," *Power System Protection and Control*, vol. 40, no. 20, pp. 113–118, 2012.

[10] S.-H. Gan, S.-M. Bi, W. Gu, and C. Jian-xin, "Improved Z-source grid-connected inverter of ship shaft generator," *Dianji Yu Kongzhi Xuebao*, vol. 22, no. 12, pp. 68–76, 2018.

[11] D. Liang, D. Wang, and Z. Peng, "Synchronization of ship shaft DFIG based on linear active disturbance rejection control," *Small & Special Electrical Machines*, vol. 45, no. 2, pp. 83–87, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2061083, 13 pages
https://doi.org/10.1155/2020/2061083

Research Article

Objective Evaluation of Drivability in Passenger Cars with Dual-Clutch Transmission: A Case Study of Static Gearshift Condition

Wei Zhou,[1,2] **Xuexun Guo**,[1,2] **Xiaofei Pei**,[1,2] **Chengcai Zhang**,[1,2] **Jun Yan**,[3] **and Jialei Xia**[3]

[1]*Hubei Key Laboratory of Advanced Technology of Automotive Parts, Wuhan University of Technology, Wuhan 430070, China*
[2]*Hubei Collaborative Innovation Centre for Automotive Components Technology, Wuhan University of Technology, Wuhan 430070, China*
[3]*Powertrain Development Department, Dongfeng Motor Corporations Technical Centre, Wuhan 430058, China*

Correspondence should be addressed to Chengcai Zhang; zhangchc@whut.edu.cn

Received 5 July 2020; Revised 18 September 2020; Accepted 8 October 2020; Published 30 October 2020

Academic Editor: Jun Shen

This paper is aimed at the problem that the subjective drivability evaluation by experienced test drivers is limited in time efficiency and is of high cost and poor repeatability. In this article, an intelligent drivability objective evaluation tool (I-DOET) for passenger cars with dual-clutch transmission (DCT) is developed and verified by real vehicle testing. First, the signal denoising method and its key parameters, which are suitable for drivability evaluation, are selected based on analytic hierarchy process (AHP) and technique for order preference by similarity to ideal solution (TOPSIS). Besides, combined with the uncertainty characteristics of subjective judgment, a mathematical model of the objective drivability evaluation FARODE (fuzzy AHP-RS based on objective drivability evaluation) is proposed by using the fuzzy comprehensive assessment (FCA) method. The AHP and rough set (RS) method are used to calculate the subjective and objective weights of the drivability evaluation, respectively, and the proportion of subjective and objective weights is determined by the principle of minimum relative information entropy. The fuzzy matrix is built by membership function of the evaluation indexes. Finally, the static gearshift condition focused on by the subjective evaluation experts is taken as a case study. The predictability score is obtained by combining the drivability quantization lever vector, comprehensive weight, and fuzzy matrix. The experimental results indicate that the proposed method is applicable for objective drivability evaluation in passenger cars with DCT.

1. Introduction

Drivability is essentially needed as one of important performances for vehicle manufacturers, and it is often evaluated based on subjective feelings of experienced test drivers [1]. However, these evaluations are time-consuming, costly, and nonrepeatable, and the reliability of the results is closely related to the psychological state of the evaluator and the test environment [2]. As an important supplement to subjective evaluation, the objective test method can not only express subjective feelings through objective indicators but also obtain more system information that is not subjectively perceptible, and it is easier to optimize vehicle performance [3]. Hence, the development of a reliable and efficient objective evaluation system plays an important role in improving the drivability of vehicles.

Several researchers have studied the objective evaluation of drivability for vehicles. As early as 2000, drivability was defined as positive and negative acceleration response, comfortable pedal response, and accurate gear shifting to meet the normal performance expectations of drivers [4]. In 2007, a series of drivability indicators for engine start conditions were proposed by AVL, but only a few evaluation indicators were published [5]. Jayaraman et al. [6] constructed a drivability index system for 9 working conditions such as engine start, shut-off, starting, and full-throttle acceleration; however, only some evaluation indexes are given. Khodabakhshian et al. [7] proposed that the

drivability of the vehicle can be improved by optimizing the shift strategy and established the mapping relationship between the evaluation index and the control parameters.

The objective evaluation system is mainly composed of a signal acquisition module and a data analysis module. The signal acquisition module is used to obtain vehicle status signals synchronously, including vehicle acceleration, vehicle speed, engine speed, engine torque, transmission gear, brake pedal signal, accelerator pedal signal and air conditioner switch signal, etc., depending on the determination of working conditions and the selection of objective indicators [8–10]. In the data analysis module, the method of multisource data fusion and sliding window is employed to convert the original data into a single working condition fragment, which can greatly overcome the shortcoming of relying on vague professional experience and inefficient manual operation [11]. In [8], the objective evaluation can be divided into 12 operation conditions, such as idle, engine start, tip-in, tip-out and static gearshift, etc.

In recent years, with the development and popularization of passenger cars with DCT, better shift quality has been demanded by drivers [12]. The purpose of drivability evaluation in static gearshift condition is to evaluate the comfort during the manual switching of P/R/N/D by the driver [13]. Because the power transmission system is a multirotational inertia system, it is difficult to complete the static gearshift process instantaneously. The disturbance of the engine and the unreasonable control logic of the gearbox may cause different degrees of shock on the vehicle. When the vibration is severe, drivers and passengers will be left with an unbearable driving experience, which decreases the drivability evaluation. Therefore, it is necessary to find the main factors that affect the shift quality through the objective drivability evaluation and provide certain suggestions to calibration engineers.

In the objective evaluation of drivability, vehicle longitudinal acceleration is the main parameter of response characteristics and comfort evaluation, which has a great influence on the extraction of feature indexes [10, 14]. Several researchers have studied the method of acceleration signal denoising [8, 15–22]. However, the evaluation indexes and analysis methods are rarely studied.

Although Liu et al. [20, 21] proposed the selection of threshold rules for wavelet filtering by quantitative indicators such as root-mean-square error (RMSE), signal-to-noise ratio (SNR), and smoothness of signal (SS), the effect of denoising composite weights was not considered. Zhou et al. [8] also developed an evaluation model of denoising, combined with SNR, mean error (ME), RMSE, and SS, but they only focused on evaluating the denoising effect based on a single factor; it is difficult to choose a suitable denoising method and its key parameters. Hence, a comprehensive evaluation method is indispensable to be applied for the selection of denoising method and its key parameters.

Generally, reasonable indicators and evaluation methods are the prerequisites for constructing an objective evaluation system for drivability. The jerk, shock, engine speed fluctuation, stabilization duration, shift delay, and engine speed undershoot are proposed by Jauch et al., Koprubasi et al., Lakshmanan et al., and Winter et al. [12, 23–25]. The research on drivability evaluation has been discussed using various methods [26–30]. The TOPSIS method, neural networks (NN), FAHP, multihierarchical grey relational analysis, and support vector regression are applied for the development of the objective evaluation for drivability. However, the rules for formulating the weights are not considered. This is insufficient for the objective drivability evaluation model to guarantee the accurate for predict result.

The rest of the paper is organized as follows. The structure of objective evaluation for drivability and the test platform of I-DOET are designed in Section 2. The principle for determining the longitudinal acceleration signal denoising method and its key parameters are introduced in Section 3. The mathematical model of the objective drivability evaluation FARODE is developed in Section 4. In Section 5, combined with the static gearshift condition in real vehicle testing, the prediction accuracy of the proposed drivability evaluation model is analysed. Finally, the conclusions are summarized inSection 6.

2. Structure of Objective Evaluation for Drivability

Objective evaluation for drivability is a precise and effective evaluation of driving performance through scientific quantitative methods, which involves three main tasks, namely, parameter acquisition, index extraction, and performance prediction, as illustrated in Figure 1. The parameter acquisition layer collects vehicle state information through an acceleration sensor and an on-board CAN network. Using signal smoothing and denoising to process the original signal, the effective information could be obtained, which reflects the actual state of the vehicle. The index extraction layer reflects the driver's subjective feelings (e.g., comfort and response) through quantitative indicators. And then, based on operating condition characteristics and the multisource signal (e.g., engine speed, pedal, and gear) provided by the parameter acquisition layer, the index extraction layer identifies the current conditions (e.g., gear shift, engine start, idle, constant speed, and tip-in) and calculates the corresponding objective characteristic index. Finally, in order to evaluate the objective drivability of the vehicle, the performance prediction layer uses an evaluation model to calculate the index weights, and at the same time, the drivability score and its optimization items can be obtained objectively and steadily.

In order to achieve the above tasks, a set of I-DOET is developed, which includes 4 ADC acquisition modules, a vehicle CAN signal acquisition module, and a software analysis module. In [23], vehicle longitudinal dynamics theory and calibration engineer experience are combined to show that longitudinal acceleration, vehicle speed, engine speed, engine torque, transmission gear, accelerator pedal opening, and braking signal are the main parameters of drivability. The longitudinal acceleration signal is obtained through ADC module in the I-DOET platform, and the vehicle speed, engine speed, engine torque, transmission gear position, accelerator pedal opening, and braking signal

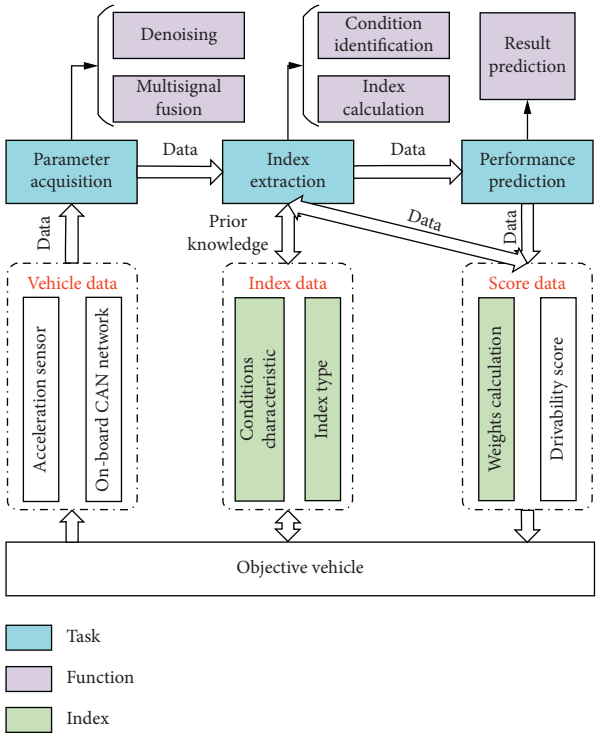


Figure 1: The structure of objective evaluation for drivability.

can be obtained through CAN module. The functional modules of the software platform are described in Sections 3–5.

3. The Denoising Quality Evaluation Based on the AHP-TOPSIS Model

The longitudinal acceleration signal of the vehicle is one of the main parameters that determine the accuracy of the feature point extraction of the objective evaluation. Due to the random vibration caused by uneven road surface and vehicle suspension, and the fact that the acceleration sensor is easily interfered by the electromagnetic field, the acceleration signal shows nonlinearity and instability, which contains a certain amount of noise, making it difficult to take both denoising and smoothing into account. Therefore, multiple indicators need to be selected to analyse the quality of denoising.

3.1. Evaluation Indexes of Denoising Quality. The denoising method of longitudinal acceleration signal suitable for the drivability objective evaluation not only needs to retain the true value of the signal but also needs to have smooth features to ensure the accuracy and reliability of the feature point recognition. The ME, RMSE, SNR, and SS are commonly used evaluation indicators [8]. Among them, the ME is the average value of the difference between the filtered signal and the ideal acceleration signal, which is obtained by

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n} |r(i) - s(i)|, \tag{1}$$

where n is the number of signal points and $r(i)$ and $s(i)$ are the ideal acceleration signal and the filtered acceleration signal, respectively.

The RMSE can be used to express small deviations and significant deviations, and the precision of the filtering effect can be given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (r(i) - s(i))^2}. \tag{2}$$

The SNR is the ratio of the average power for signals and the average power for noises. The larger the value, the more effective the components of the signal characteristic. It is represented by

$$\mathrm{SNR} = 10\lg\left(\frac{\sum_{i=1}^{n} f^2(i)}{\sum_{i=1}^{n} [f(i) - s(i)]^2}\right), \tag{3}$$

where $f(i)$ is the acceleration signal with noise.

The sudden changes and spikes of the signal will greatly interfere with the accuracy of the evaluation index extraction. The SS is obtained by the ratio of the sum for the signal after denoising and the variance of the original signal. It is calculated by

$$\mathrm{SS} = \frac{\sum_{i=2}^{n} [s(i) - s(i-1)]^2}{\sum_{i=2}^{n} [f(i) - f(i-1)]^2}. \tag{4}$$

In order to evaluate the quality of denoising more accurately, the signal-to-noise ratio gain (SNRG) and correlation coefficients (CC) have been proposed [31, 32]. The SNRG is an important indicator to measure random vibration. It can be given by

$$\mathrm{GSNR} = \log\left(\frac{\sum_{i=1}^{n} f^2(i)}{\sum_{i=1}^{n} [f(i) - s(i)]^2}\right)\left[\frac{\sum_{i=1}^{n} f^2(i)}{\sum_{i=1}^{n} [f(i) - r(i)]^2}\right]. \tag{5}$$

The CC is used to characterize the approximate relationship between the filtered acceleration signal and the ideal acceleration signal, which is obtained by

$$\mathrm{CC} = \frac{\sum_{i=1}^{n} (s(i) - \bar{s})(r(i) - \bar{r})}{\sqrt{\sum_{i=1}^{n} (s(i) - \bar{s})^2 \sum_{i=1}^{n} (r(i) - \bar{r})^2}}, \tag{6}$$

where $\bar{s}$ and $\bar{r}$ are the mean value of filtered acceleration signal and the mean value of the ideal acceleration signal, respectively.

According to the requirement of denoising, the filtered signal should be closer to the ideal signal and smoother. The quantitative index system for denoising of the longitudinal acceleration signal is shown in Table 1.

3.2. AHP-TOPSIS Evaluation Model. AHP is a method of dealing with complex and constrained multifactor systems through expert experience. It can be used to plan the hierarchical structure clearly and assign the weight of factors scientifically. TOPSIS is a method for solving multiattribute

Table 1: The quantitative index system for denoising.

Macro index layer	Micro index layer	Index type
Signal error	ME	I
	RMSE	I
	SNR	II
	SNRG	II
	CC	II
Smoothing	SS	I

Index type division rules: I is the smaller the better type; II is the larger the better type.

decision-making schemes. It often uses ideal and anti-ideal solutions to measure the order of evaluation schemes. Integrating the evaluation model of AHP-TOPSIS can find the goal that meets the intention of the decision maker from the decision object group scientifically and effectively. The flowchart is shown in Figure 2.

First, based on the experience of experts, a binary comparison method is used to compare and assign different indicators at the same level. The assignment criteria are listed in Table 2, and the judgment matrix X is constructed:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{nn} \end{bmatrix}, \tag{7}$$

$$X_{ij} = \frac{1}{X_{ji}}, \quad X_{ji} \neq 0. \tag{8}$$

The square root method is used to obtain the normalization and maximum feature root of the judgment matrix. The relative weight of each factor is obtained from the maximum feature root and its feature vector:

$$\begin{cases} A_i = \prod_{j=1}^{n} X_{ij}, \\ B_i = \dfrac{\sqrt[n]{A_i}}{\sum_{i=1}^{n} \sqrt[n]{A_i}}, \\ \lambda_{\max} = \sum_{i=1}^{n} \dfrac{(XB)_i}{nB_i}, Xw = \lambda_{\max} \times w, \end{cases} \tag{9}$$

where A and B are the result of multiplication of the comparison matrix by row and the normalized vector, respectively, and $\lambda_{\max}$ and w are the largest feature root and the weight of each factor, respectively.

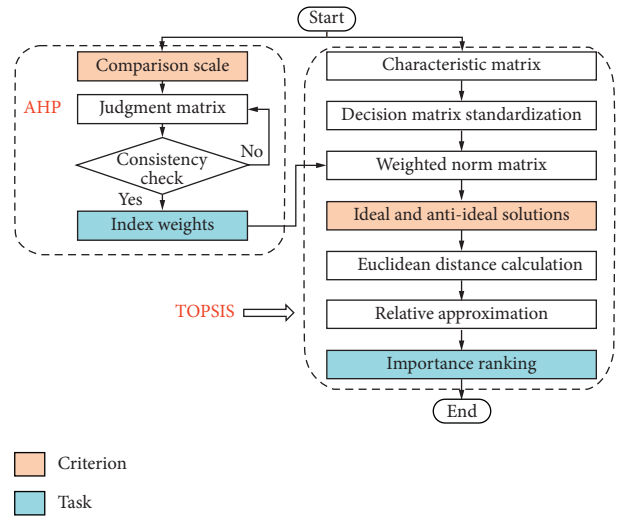


Figure 2: Flowchart of AHP-TOPSIS evaluation.

The consistency check of AHP is used to analyse the rationality of judgment matrix, which directly affects the reliability of the AHP output. The random index (RI) can be searched in Table 3, and the consistency index (CI) and consistency ratio (CR) can be obtained by

$$\begin{cases} \mathrm{CI} = \dfrac{\lambda_{\max} - n}{n}, \\ \mathrm{CR} = \dfrac{\mathrm{CI}}{\mathrm{RI}}. \end{cases} \tag{10}$$

The upper limit for CR is 0.1. If CR exceeds 0.1, the judgment matrix should be reestablished and the index weights should be calculated until the consistency check is satisfied.

TOPSIS method can incorporate important weights of criteria into the comparison procedures to provide an understandable and rational ranking result, which consists of the following seven steps.

Step 1. An initial matrix D for evaluation is established by

$$D = \begin{bmatrix} D_{11} & \cdots & D_{1l} & \cdots & D_{1q} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ D_{k1} & \cdots & D_{kl} & \cdots & D_{kq} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ D_{p1} & \cdots & D_{pl} & \cdots & D_{pq} \end{bmatrix}, \tag{11}$$

where p and q are the number of schemes and indexes, respectively, and D_{kl} is a crisp value indicating the performance rating of each alternative D_k with respect to each criterion D_l, $k = 1, 2, \ldots, p$; $l = 1, 2, \ldots, q$.

Step 2. The standardized decision matrix S is adopted by

Table 2: Comparable scale of 1~9.

Scale	Definition	Notation
1	Equally important	X_i is equally important to X_j
3	Slightly important	X_i is moderately more important than X_j
5	Strongly important	X_i is strongly more important than X_j
7	Very important	X_i is very strongly more important than X_j
9	Extremely important	X_i is extremely more important than X_j
2, 4, 6, 8	Intermediate values of the above judgment	

Table 3: Values of average stochastic coincidence indicators.

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

$$S = (S_{kl})_{p\times q} = \begin{bmatrix} S_{11} & \cdots & S_{1l} & \cdots & S_{1q} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ S_{k1} & \cdots & S_{kl} & \cdots & S_{kq} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ S_{p1} & \cdots & S_{pl} & \cdots & S_{pq} \end{bmatrix}, \tag{12}$$

where the S_{kl} value is calculated by

$$S_{kl} = \frac{D_{kl}}{\sqrt{\sum_{k=1}^{p} D_{kl}^2}}. \tag{13}$$

Step 3. The weighted standardized decision matrix R is given by

$$R = (R_{kl})_{p\times q} = \begin{bmatrix} R_{11} & \cdots & R_{1l} & \cdots & R_{1q} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ R_{k1} & \cdots & R_{kl} & \cdots & R_{kq} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ R_{p1} & \cdots & R_{pl} & \cdots & R_{pq} \end{bmatrix}, \tag{14}$$

where the R_{kl} value is calculated by

$$R_{kl} = S_{kl} * w_l, \tag{15}$$

where w_l represents the weight of the lth criterion.

Step 4. The index attributes applicable to TOPSIS are divided into benefit criteria (B) and cost criteria (C), and the ideal solution R_l^+ and anti-ideal solution R_l^- are obtained by

$$\begin{cases} R_l^+ = \{(\max R_{kl} \mid l \in B), (\min R_{kl} \mid l \in C)\}, \\ R_l^- = \{(\max R_{kl} \mid l \in C), (\min R_{kl} \mid l \in B)\}. \end{cases} \tag{16}$$

Step 5. The separation measures of each alternative from the ideal solution and anti-ideal solution are calculated by

$$\begin{cases} E_k^+ = \sqrt{\sum_{l=1}^{q} (R_{kl} - R_l^+)^2}, E_k^- = \sqrt{\sum_{l=1}^{q} (R_{kl} - R_l^-)^2}. \end{cases} \tag{17}$$

where E_k^+ and E_k^- are the Euclid distance of each alternative from the ideal solution and anti-ideal solution, respectively.

Step 6. The relative approximation to the ideal solution E_k is obtained by

$$E_k = \frac{E_k^-}{E_k^- + E_k^+}. \tag{18}$$

Step 7. The important ranking of the performance order is given by E_k index value, which lies between 0 and 1. The larger the index value, the better the performance of the alternatives.

3.3. Application of Longitudinal Acceleration Signal Denoising Based on AHP-TOPSIS. According to the previous work [4], firstly, the acceleration signal with noise is taken as the research objective, and then the ME, RMSE, SNR, SNRG, SS, and CC are selected as index layers to take the best scheme of denoising. AHP-TOPSIS is used to select the appropriate denoising method and its key parameters. In order to calculate the weight of each characteristic index more reliably, three experts with rich experience in signal denoising are invited in the questionnaire survey to discuss the importance of different evaluation indexes on the signal denoising effect. The 1–9 scale method to compare the importance of evaluation indicators for experts is selected, which is described in Table 2. Table 4 is the judgment matrix of the target layer.

According to equations (7)–(10), $\lambda_{\max} = 6.125$, CI = 0.0249, RI = 1.24, CR = 0.020 < 0.1, which satisfies the test requirements for judging the consistency of the matrix. The weight of each index is w = [0.036, 0.044, 0.428, 0.275, 0.150, 0.067]. Then, TOPSIS is used to rank the schemes, and the proposed AHP-TOPSIS decision model is used to select the appropriate denoising method and its key parameters. After obtaining the local index weights through AHP, the decision matrix is established.

In the research, 5 Hz low-pass filtering, 21-point smoothing filter, and Db6, Coif5, Sym3, Sym4, and Sym6 wavelet threshold as the basis function are selected as the research objective and calculated by equations (1)–(6), the calculation results of ME, RMSE, SNR, SNRG, SS, and CC are shown in Table 5. And these values are used as the initial

Table 4: The judgment matrix of the denoising.

	ME	RMSE	SNR	GSNR	SS	CC
ME	1	1	1/9	1/7	1/5	1/3
RMSE	1	1	1/9	1/7	1/4	1
SNR	9	9	1	2	3	7
GSNR	7	7	1/2	1	2	5
SS	5	4	1/3	1/2	1	2
CC	3	1	1/7	1/5	1/2	1

Table 5: Denoising effect analysis table with different methods.

Denoising method	Key parameter	ME	RMSE	SNR	GSNR	SS	CC
Db6 wavelet	1 layer	0.1116	0.1413	22.0485	1.159	0.1955	0.9968
Db6 wavelet	2 layers	0.0794	0.1	20.2787	1.066	0.0284	0.9984
Db6 wavelet	3 layers	0.0558	0.0705	19.597	1.03	0.0038	0.9992
Db6 wavelet	4 layers	0.0393	0.0498	19.2912	1.014	0.0006	0.9996
Db6 wavelet	5 layers	0.028	0.0391	19.1149	1.005	0.0001	0.9998
Coif5 wavelet	1 layer	0.112	0.1413	22.0481	1.159	0.1898	0.9968
Coif5 wavelet	2 layers	0.0792	0.0999	20.2753	1.066	0.027	0.9984
Coif5 wavelet	3 layers	0.0559	0.0704	19.5964	1.03	0.0036	0.9992
Coif5 wavelet	4 layers	0.0395	0.0501	19.2905	1.014	0.0006	0.9996
Coif5 wavelet	5 layers	0.028	0.0378	19.1288	1.006	0.0001	0.9998
Sym3 wavelet	1 layer	0.1111	0.1413	22.0504	1.159	0.2085	0.9968
Sym3 wavelet	2 layers	0.0783	0.0999	20.2762	1.066	0.0332	0.9984
Sym3 wavelet	3 layers	0.055	0.0701	19.5914	1.03	0.0047	0.9992
Sym3 wavelet	4 layers	0.0389	0.0501	19.2925	1.014	0.0008	0.9996
Sym3 wavelet	5 layers	0.0276	0.0381	19.1225	1.005	0.0002	0.9998
Sym4 wavelet	1 layer	0.1112	0.1412	22.0463	1.159	0.2016	0.9968
Sym4 wavelet	2 layers	0.0786	0.0999	20.2735	1.066	0.0303	0.9984
Sym4 wavelet	3 layers	0.0553	0.0702	19.5929	1.03	0.0041	0.9992
Sym4 wavelet	4 layers	0.0391	0.0499	19.2927	1.014	0.0006	0.9996
Sym4 wavelet	5 layers	0.0277	0.0373	19.1307	1.006	0.0002	0.9998
Sym6 wavelet	1 layer	0.1117	0.1412	22.0465	1.159	0.1953	0.9968
Sym6 wavelet	2 layers	0.0794	0.1001	20.2813	1.066	0.0285	0.9984
Sym6 wavelet	3 layers	0.0555	0.0701	19.5903	1.03	0.0037	0.9992
Sym6 wavelet	4 layers	0.039	0.0495	19.2903	1.014	0.0006	0.9996
Sym6 wavelet	5 layers	0.0277	0.0372	19.1307	1.006	0.0001	0.9998
Low-pass filter	5Hz	0.0529	0.0675	18.6924	0.983	0.0042	0.9993
Smoothing filter	21	0.0343	0.0436	19.2237	1.011	0.0024	0.9997

decision matrix of the TOPSIS method. After the decision matrix is normalized according to Table 5 and equation (15) fd15, the standard weight is obtained from the AHP method. Ideal solution and anti-ideal solution are obtained by equation (16)fd16: the ideal solution is represented by $R+ = \{0.0129, 0.0028, 0.0908, 0.0583, 4.23*10-5, 0.0129\}$, and the anti-ideal solution is represented by $R- = \{0.0114, 0.0139, 0.0796, 0.0494, 0.0698, 0.1286\}$. Finally, equation (17)fd17 is used to calculate the distance between the ideal solution and the anti-ideal solution, and the last equation, equation (18) fd18, is used for the final ranking. Based on the E_k values shown in Table 6, the application results show that Coif5 wavelet with 3 layers is the best; the E_k value is 0.8439.

4. Mathematical Model of the FARODE

This section introduces the analysis steps of FARODE method in objective evaluation for drivability. As shown in Figure 3, it includes the preparation of research objects and indicators, AHP-RS combined weighting method, and multiple hierarchical fuzzy comprehensive evaluation.

4.1. Determination of Research Objects and Objective Indicators. It is important to control the synchronizer, which directly affects the quality of gear shifting performance in the static gearshift process for passenger car with DCT. The optimized shifting force strategy can coordinate with the changes of the synchronizer in real time. After the optimized shifting force is adopted in real vehicles, the shifting time is extended due to the addition of damping force, but it can ensure that the entire shifting time is within the required range. By consulting the expert group, the expected static gearshift quality includes the following attributes: fast, smooth, and comfortable. Therefore, from the three aspects of shifting comfort, response characteristics, and stability, the determination of the drivability evaluation index for static shifting conditions is considered. Figure 4 is the dynamic curve of the subdivided working condition N-D

Table 6: Weight ranking.

Scheme	E_k	Rank	Scheme	E_k	Rank	Scheme	E_k	Rank
Db6 (1 layer)	0.2029	25	Coif5 (5 layers)	0.8325	14	Sym4 (4 layers)	0.8384	9
Db6 (2 layers)	0.8006	19	Sym3 (1 layer)	0.1879	27	Sym4 (5 layers)	0.8326	13
Db6 (3 layers)	0.8437	2	Sym3 (2 layers)	0.7858	22	Sym6 (1 layer)	0.2031	24
Db6 (4 layers)	0.8384	6	Sym3 (3 layers)	0.8429	5	Sym6 (2 layers)	0.8002	20
Db6 (5 layers)	0.8318	16	Sym3 (4 layers)	0.8384	8	Sym6 (3 layers)	0.8437	3
Coif5 (1 layer)	0.2132	23	Sym3 (5 layers)	0.8321	15	Sym6 (4 layers)	0.8384	10
Coif5 (2 layers)	0.805	17	Sym4 (1 layer)	0.1943	26	Sym6 (5 layers)	0.8326	12
Coif5 (3 layers)	0.8439	1	Sym4 (2 layers)	0.795	21	Low-pass (5 Hz)	0.8028	18
Coif5 (4 layers)	0.8383	7	Sym4 (3 layers)	0.8434	4	Smoothing (21)	0.8348	11

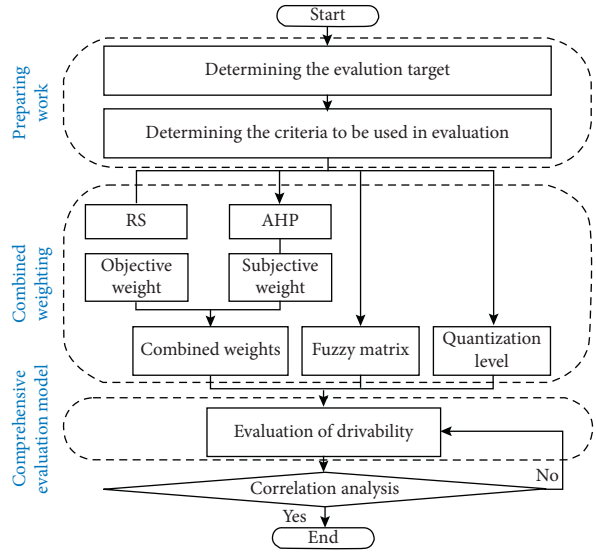


Figure 3: Flowchart of FARODE

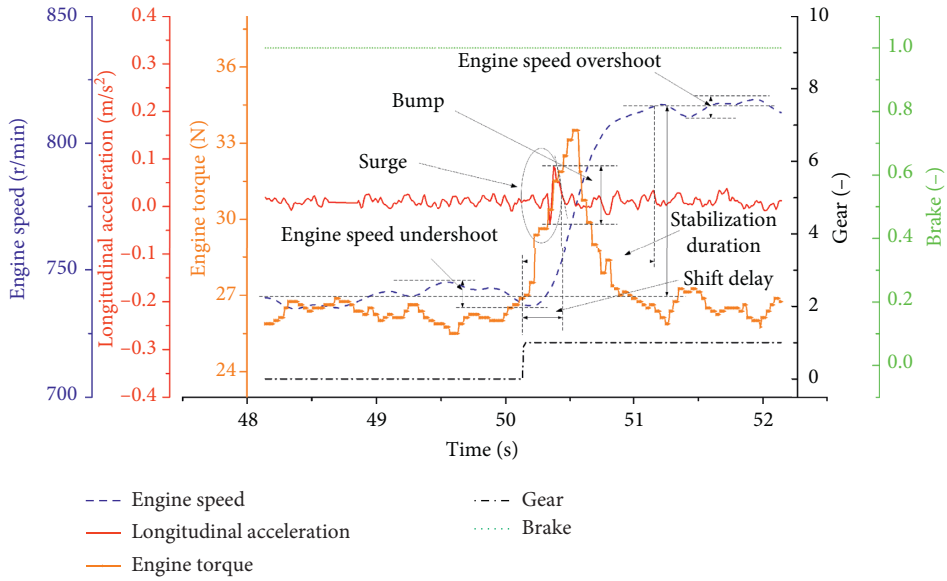


Figure 4: Dynamic curves of vehicle performance parameters in static N-D condition.

in the static gearshift processes, and gear position 0 represents N gear, and gear position 1 represents *D* gear. In order to achieve the goal of energy saving and smooth clutch integration, OEMs will set different target speeds and target torque outputs when designing EMS and TCU control strategies, which will affect the driving experience of drivers

and passengers. In the N gear, the clutch is in a disconnected state, no transmission function is needed, and a lower target engine speed and engine torque is usually set. However, the clutch is engaged in D gear, and the engine speed and engine torque are compensated by EMS or TCU control strategy, which is used to ensure that the surge for vehicle is slight.

As shown in Figure 4, the response characteristics of the vehicle consider the effects of shift delay (SD) and shift stabilization duration (SSD); the stability characteristics of the vehicle consider engine speed overshoot (ESO), and engine speed undershoot (ESU). Longitudinal acceleration bump and surge of static gearshift are used to describe the shift comfort of the vehicle.

According to the SAE (J1441) scoring standard and the characteristics of static gear shift conditions, the quantization level score is shown in Table 7. The quantization level of drivability in static gearshift conditions is set to five levels. The indicator classification criteria are shown in Table 8.

Table 7: Drivability quantification level.

Subjective description	Rating interval	Grade rating
Well	(8, 10]	9
Better	(7, 8]	7.5
General	(5, 7]	6
Poor	(4, 5]	4.5
Difference	(2, 4]	3

4.2. AHP-RS Combined Weighting Method. The index weight and rationality in the objective evaluation of vehicle drivability have an important influence on the reliability of the prediction results, which is affected by the cognition of the evaluator and the dynamic characteristics for vehicles. The perception for evaluator is strongly subjective and usually used for subjective empowerment. The vehicle dynamic characteristics are reflected in real and objective driving data, which are usually used for objective weight calculation. In order to be able to take into account both the cognition for evaluator and the vehicle dynamic characteristics, it is necessary to establish a subjective and objective weight calculation model. The principle of minimum information entropy is used to determine the subjective and objective allocation weights, which are applied to objective evaluation model of drivability for vehicles.

Subjective weights are based on prior knowledge and experience. The subjective weight $s(i)$ of the lowest-level indicator relative to the highest-level indicator is determined based on AHP model in the principle of Section 3.2.

Objective weights are determined by RS principle based on the information from original data and the logic of knowledge system. The traditional RS method can only express the information of current data on decision and ignore the prior knowledge of decision-makers. Based on Table 8, an improved RS method is developed to calculate the weight of drivability evaluation index. The objective weight $o(i)$ of each evaluation index is calculated by

$$\Phi_A(B) = \frac{1}{m}\sum_{i=1}^{n}\left|\Phi_A(B_i)\right|,$$

$$\Phi_{A-A_i}(B) = \frac{1}{m}\sum_{i=1}^{n}\left|\Phi_{A-A_i}(B_i)\right|, \tag{19}$$

$$o(i) = \frac{\Phi_A(B) - \Phi_{A-A_i}(B)}{\sum_{i=1}^{n}\left|\Phi_A(B) - \Phi_{A-A_i}(B)\right|},$$

where m is the number of samples, $\Phi_A(B)$ and $\Phi_A(B_i)$ are the dependence and the number of compatible samples in the decision table, respectively, and $\Phi_{A-A_i}(B)$ and $\Phi_{A-A_i}(B_i)$ are the dependence and the number of compatible samples after excluding the evaluation index A_i, respectively.

In order to seek the balance between subjective and objective weights, the principle of minimum relative information entropy is used to reduce the deviation between subjective and objective weights, and a more scientific and reliable AHP-RS combination optimization weight w is obtained by

$$w(i) = \frac{\sqrt{s(i)o(i)}}{\sum_{i=1}^{n}\sqrt{s(i)o(i)}}. \tag{20}$$

4.3. Multilevel Fuzzy Comprehensive Evaluation Model. In order to solve the uncertainty of subjective evaluation results, fuzzy mathematics theory is applied to build the comprehensive evaluation model. According to the review standard (Table 7) and the quantitative grade of the static gearshift drivability evaluation index (Table 8), the trapezoid distribution membership function is selected to determine the membership grade. The distribution membership function is shown in Figure 5; the shaded area in the figure is the fuzzy interval. The limits of fuzzy interval c_i, $i = 1, 2, \ldots, 10$, are obtained by analysing the real vehicle test results. The membership function is expressed by

$$\mu_1 = \begin{cases} 1, & c_1 \le x \le c_2, \\ \dfrac{c_3 - x}{c_3 - c_2}, & c_2 < x < c_3, \\ 0, & x \ge c_3, \end{cases}$$

$$\mu_k = \begin{cases} \dfrac{x - c_{2(k-1)}}{c_{2k-1} - c_{2(k-1)}}, & c_{2(k-1)} < x < c_{2k-1}, \\ 1, & c_{2k-1} \le x \le c_{2k}, \\ \dfrac{c_{2k+1} - x}{c_{2k+1} - c_{2k}}, & c_{2k} < x < c_{2k+1}, \\ 0, & x < c_{2(k-1)}, x \ge c_{2k+1}, \end{cases} \tag{21}$$

$$\mu_5 = \begin{cases} 0, & x < c_8, \\ \dfrac{x - c_8}{c_9 - c_8}, & c_8 \le x \le c_9, \\ 1, & x > c_9, \end{cases}$$

Table 8: Indicator classification criteria.

Level	SD	SSD	ESO	ESU	Bump	Surge
Well	[0, 0.3)	[0, 0.5)	[0, 20)	[0, 20)	[0, 0.05)	[0, 0.03)
Better	[0.3, 0.5)	[0.5, 1.0)	[20, 30)	[20, 30)	[0.05, 0.075)	[0.03, 0.045)
General	[0.5, 0.8)	[1.0, 1.5)	[30, 50)	[30, 50)	[0.075, 0.125)	[0.045, 0.075)
Poor	[0.8, 1.0)	[1.5, 2.0)	[50, 60)	[50, 60)	[0.125, 0.15)	[0.075, 0.09)
Difference	[1, +∞)	[2, +∞)	[60, +∞)	[60, +∞)	[0.15, +∞)	[0.09, +∞)

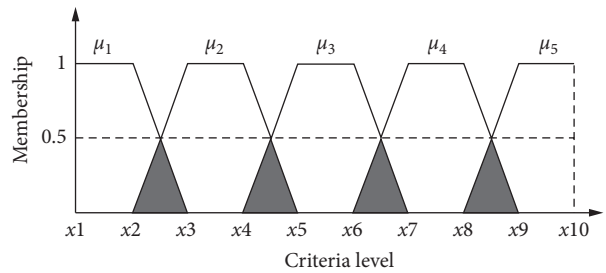


Figure 5: The membership function of the evaluation indexes relative to all levels.

where $k = 2, 3, 4$; $\mu_i \in [0,1]$ $(i = 1, 2, 3, 4, 5)$. It is the probability of different membership function.

The fuzzy matrix $M = (\mu_{kj})_{5\times n}$ is obtained, and the score is quantization level $S = [\,9 \;\; 7.5 \;\; 6 \;\; 4.5 \;\; 3\,]$. According to M, S, and the optimization weight w, the objective score can be calculated as $T = S \cdot M \cdot w^T$.

5. Case Study of Objective Drivability Evaluation

5.1. Case Study. In this section, I-DOET is used to collect and process data on 15 static gearshift conditions of passenger cars with DCT on the Chinese market, as shown in Figure 6, including I-DOET, an acceleration sensor, and a laptop. In order to test the signals of the vehicle's dynamic response accurately, the acceleration sensor is installed on the driver's seat rail through the F clip, and its position is close to the centre of mass of the vehicle. Multisource signal is obtained by using different signal sources in parallel acquisition control method, and the signal sampling rate, signal type, data acquisition, and data storage can be realized by the designed and written digital acquisition system. The static gearshift condition of real vehicle test is designed as follows: windows, air condition, and other electrical equipment are turned off, engine is started and the engine speed is idling, hand brake is released while brake system is depressed, and gear is shifted at norm speed, including P-D, D-N, N-R, R-N, N-D, D-R, R-D, D-P, P-R, and R-P. Meanwhile, gearbox gears, engine speed, engine torque, vehicle longitudinal acceleration, and braking signals are obtained by I-DOET hardware and laptop.

In this real vehicle test, the sampling frequency of each signal is set as 100 Hz, and the static gearshift conditions of 15 vehicles are collected and given the subjective scores by 3 experts. According to the analysis results ofSection 3, the 3-layer Coif5 wavelet threshold method is used for denoising of the longitudinal acceleration signal. Sliding window method and D-S semantic segmentation method are used for working condition recognition and objective index extraction, respectively. 43 of the 45 sets of effective objective indexes are used for model training and weighting. Table 9 lists the remaining two sets of indicator data (N_1 and N_2) used to verify the accuracy of proposed FCA model.

5.2. Discussion. In order to obtain the subjective weights more accurately, in this study, an expert group composed of two scholars from Wuhan University of Technology and three vehicle calibration engineers from Dongfeng Motor Corporations Technical Centre is formed. A hierarchical model for objective evaluation of drivability in static gearshift condition is constructed on the basis of prior knowledge of the expert group, as shown in Figure 7.

The expert group determined the indicators to be used in the static gearshift evaluation model and compared the criteria in pairs. Table 2 is used to compare the relative advantages and disadvantages of the evaluation indicators by the expert group, and the judgment matrix of the target layer and the subtarget layer is established in Tables 10 and 11.

The subjective weights of the target layer and subtarget layer in static shifting conditions are obtained by the AHP method, and the objective weights of the subtarget layer are calculated by the RS method. The optimization weight is calculated by equation (20)fd20.The subjective weight of the target layer and the optimized weight distribution of the subtarget layer are shown in Table 12.

Figure 8 compares the predicted results of the drivability evaluation based on FARODE model and FAHP method, including the objective scores of FARODE and FAHP with IDs N_1 and N_2 and the real scores given by the expert group. From the subjective and objective score analysis in Figures 8(a) and 8(b), the prediction results of FARODE and FAHP for static gearshift subworking conditions are demonstrated. The scoring results show that the drivability score

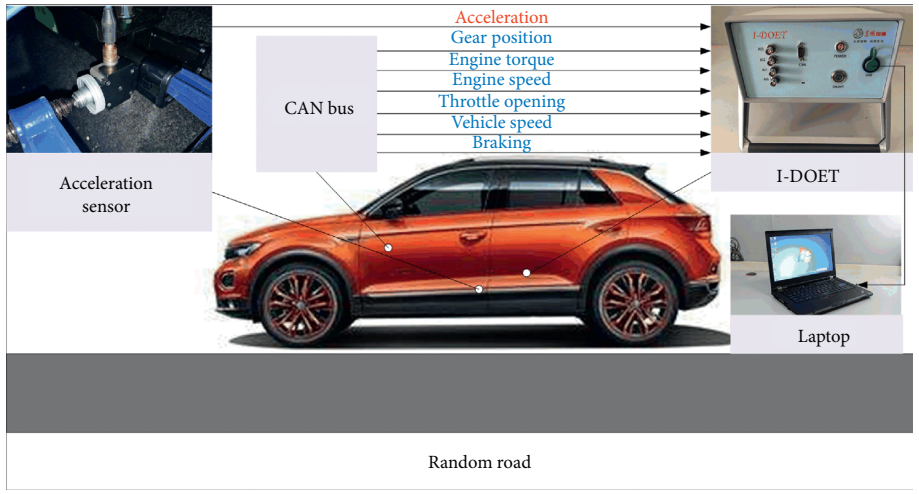


Figure 6: Real vehicle test.

Table 9: Evaluation indexes values for ID N_1 and ID N_2.

Indicator data	Operation model	SD	SSD	ESO	ESU	Bump	Surge
N_1	P-D	0.32	0.53	27.158	29.132	0.062	0.043
	D-N	0.48	0.87	30.51	25.802	0.065	0.042
	N-R	0.37	0.69	23.289	23.097	0.073	0.047
	R-N	0.39	0.53	28.283	23.612	0.059	0.037
	N-D	0.32	0.53	27.796	23.447	0.062	0.043
	D-R	0.56	0.89	24.798	27.178	0.086	0.06
	R-D	0.43	0.81	23.796	22.289	0.072	0.046
	D-P	0.22	0.51	19.191	23.75	0.062	0.048
	P-R	0.35	0.62	21.197	16.763	0.075	0.043
	R-P	0.36	0.61	24.987	18.02	0.071	0.05
N_2	P-D	0.35	0.89	28.671	20.467	0.076	0.048
	D-N	0.28	0.51	27.767	24.559	0.054	0.045
	N-R	0.54	0.92	32.3	21.453	0.072	0.05
	R-N	0.47	0.66	24.453	18.842	0.075	0.038
	N-D	0.34	0.55	27.849	22.954	0.058	0.041
	D-R	0.62	0.99	10.079	18.194	0.096	0.062
	R-D	0.51	0.85	17.241	16.91	0.078	0.05
	D-P	0.27	0.64	24.25	24.237	0.062	0.048
	P-R	0.38	0.71	13.941	18.493	0.084	0.047
	R-P	0.43	0.76	29.586	28.875	0.064	0.044

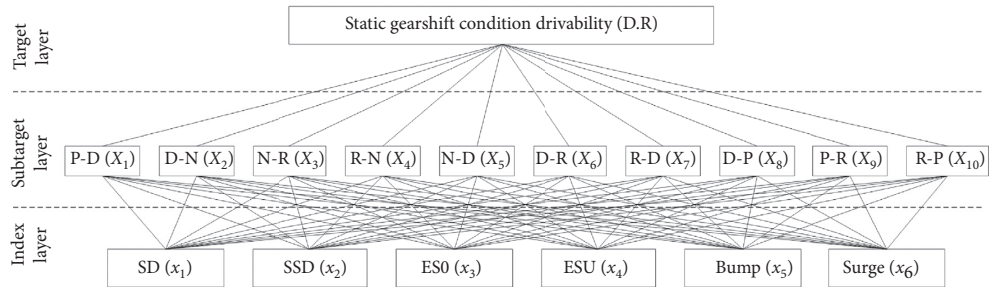


Figure 7: The hierarchical model of drivability in static gearshift condition.

of D-R subdivision operating conditions in static gearshift is lower than other suboperating conditions, which is closer to the expert group score and consistent with the actual vehicle response. This is related to the driver needs to control the filling and discharging of oil during the clutch disengagement and coupling action during the D-R shift operation. The formulation of the TCU control strategy needs to consider the shift completion time and shift shock

Table 10: The judgment matrix for the target layer.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1	2	2	2	2	1/2	1/2	1	1	1
X_2	1/2	1	1	1	1	1/3	1/3	1/2	1/2	1/2
X_3	1/2	1	1	1	1	1/3	1/3	1/2	1/2	1/2
X_4	1/2	1	1	1	1	1/4	1/4	1/2	1/2	1/2
X_5	1/2	1	1	1	1	1/4	1/4	1/2	1/2	1/2
X_6	2	3	3	4	4	1	1	2	2	2
X_7	2	3	3	4	4	1	1	2	2	2
X_8	1	2	2	2	2	1/2	1/2	1	1	1
X_9	1	2	2	2	2	1/2	1/2	1	1	1
X_{10}	1	2	2	2	2	1/2	1/2	1	1	1

Table 11: The judgment matrix for the subtarget layer.

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1	1	2	2	1/3	1/3
x_2	1	1	2	2	1/3	1/3
x_3	1/2	1/2	1	1	1/7	1/7
x_4	1/2	1/2	1	1	1/7	1/7
x_5	3	3	7	7	1	1
x_6	3	3	7	7	1	1

Table 12: Weight distribution of the target layer and subtarget layer.

Operation model	Subtarget weight						Target weight
	w_1	w_2	w_3	w_4	w_5	w_6	w
X_1	0.1447	0.1390	0.1033	0.0998	0.2470	0.2662	0.1015
X_2	0.1385	0.1281	0.1022	0.0883	0.2634	0.2795	0.054
X_3	0.1383	0.1320	0.0954	0.0995	0.2566	0.2782	0.054
X_4	0.1403	0.1330	0.0812	0.0865	0.2733	0.2857	0.0508
X_5	0.1362	0.1274	0.0939	0.0879	0.2707	0.2839	0.0508
X_6	0.138	0.1293	0.0891	0.0823	0.2606	0.3007	0.1923
X_7	0.1457	0.1372	0.0949	0.0886	0.2589	0.2747	0.1923
X_8	0.1449	0.1326	0.0957	0.0999	0.2686	0.2583	0.1015
X_9	0.1306	0.1221	0.1008	0.0901	0.2718	0.2846	0.1015
X_{10}	0.1399	0.1328	0.1012	0.0916	0.2608	0.2737	0.1015

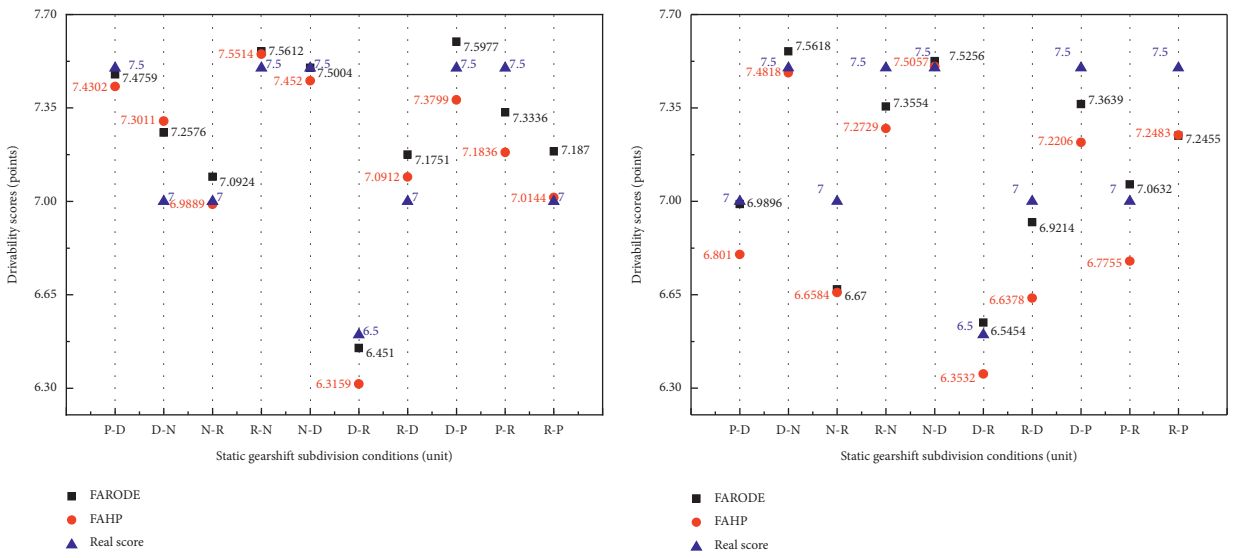


Figure 8: Scoring results and real score of different drivability evaluation model. (a) ID N_1; (b) ID N_2.

Table 13: Comparison indexes for different evaluation models.

	ID N_1		ID N_2	
	FARODE	FAHP	FARODE	FAHP
Maximum deviation	0.2576	0.3164	0.33	0.36
Correlation coefficient	0.9323	0.8893	0.921	0.8548
Errors more than 0.25/time	1	2	2	4
Pass rate	90%	80%	80%	60%

simultaneously. The calibration of D-R subdivision conditions has high requirements. Thus, the effectiveness of the FARODE model and the FAHP method in the evaluation of the drivability under static gearshift conditions is verified.

The accuracy and stability of the drivability evaluation model are analysed by introducing comparative indicators including maximum deviation, Pearson correlation coefficient, and pass rate. Looking at Table 13, the maximum deviation of the FARODE comprehensive evaluation model in the ID N_1 and ID N_2 static gearshift conditions in actual vehicle testing is 0.2576 and 0.33, respectively. The maximum deviation of the AHP method is 0.3164 and 0.36, respectively. The correlation between the model prediction results and the real score of the expert group is researched through Pearson correlation coefficients to determine the model accuracy. The Pearson correlation coefficients of the FAHP method are 0.83 and 0.8548, respectively, and the FARODE model can increase the Pearson correlation coefficients to 0.9323 and 0.921. With reference to the scoring rules of the subjective evaluation engineers of Dongfeng Motor Centre and the sensitivity test results in [23], it is assumed that when the predicted result and the actual score exceed 0.25, the predicted result of the model needs to be improved. The adoption rate of FARODE mathematical model for ID N_1 and ID N_2 is 90% and 80%, respectively. However, the passing rate through FAHP can only reach 80% and 60%. The analysis results show that drivability evaluation of the FARODE model in static gearshift conditions is more accurate and reliable than the simple FAHP method. FARODE model is used to obtain the drivability evaluation scores in static gearshift with ID N_1 and ID N_2, and the scores are 7.1641 and 7.0234, respectively. Analysing the subdivision conditions that affect the static gearshift and their objective indicators can be used to guide the improvement of drivability for automobiles.

6. Conclusions

This paper developed an intelligent drivability objective evaluation tool (I-DOET) for passenger cars with dual-clutch transmission (DCT) and verified by real vehicle testing. Based on the analytic hierarchy process (AHP) and technique for order preference by similarity to ideal solution (TOPSIS), the signal denoising method and its key parameters suitable for drivability evaluation are selected. Thereafter, combined with the uncertainty characteristics of subjective judgments, a mathematical model of the objective drivability evaluation FARODE (fuzzy AHP-RS based on objective drivability evaluation) is proposed by using the fuzzy comprehensive assessment (FCA) method. Furthermore, the static gearshift condition is taken as a case study to verify the accuracy and stability of evaluation model including FARODE and FAHP. The conclusions of this investigation are given as follows:

(1) Based on AHP-TOPSIS model, the ME, RMSE, SNR, SNRG, SS, and CC are selected as index layers to take the best scheme of denoising. The application results show that Coif5 wavelet with 3 layers is the most suitable de-noising method for objective evaluation of drivability.

(2) FARODE method is introduced in objective evaluation for drivability; it includes the preparation of research objects and indicators, AHP-RS combined weighting method, and multiple hierarchical fuzzy comprehensive evaluation. The AHP and rough set (RS) method are used to calculate the subjective and objective weights of the drivability evaluation, respectively. The principle of minimum relative information entropy is used to solve the unscientific problem of subjective and objective weight distribution.

(3) The static gearshift condition focused on by the subjective evaluation experts is taken as a case study, and maximum deviation, Pearson correlation coefficient, and pass rate are used as comparative indicators to compare the FARODE evaluation model and the FAHP method. The real vehicle test results show that the FARODE model proposed in this paper is more accurate and reliable.

Given the recommendations for the follow-up investigations, the fuzzy membership function of objective indicators should be further studied with respect to the stability and accuracy of the consistency of subjective and objective evaluations. Moreover, future investigations are needed to find a more suitable drivability evaluation method in mathematics.

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 express their deep appreciation to Hubei Key Laboratory of Advanced Technology for Automotive Components for the continuous support. The authors also acknowledge the support of Hubei Collaborative Innovation Centre for Automotive Components Technology and Dongfeng Motor Corporations Technical Centre. This work was financially supported in part by the National Natural Science Foundation of the People's Republic of China (grant number: 51505354).

References

[1] W. S. Chang, C. Jongdae, W. C. Suk, and L. Wonsik, "An objective method of drivability evaluation using a simulation model for hybrid electric vehicles," *International Journal of Precision Engineering and Manufacturing*, vol. 15, no. 2, pp. 219–226, 2014.

[2] K. Chandrasekaran, N. Rao, S. Palraj, and C. Kurella, "Objective drivability evaluation on compact suv and comparison with subjective drivability," SAE, Pittsburgh, PA, USA, SAE Technical Paper 2017-26-0153, 2017.

[3] Z. F. Chen and X. M. Shi, *The Subjective Evaluation of Vehicle Dynamic Performance*, China: People's Communication Press, Beijing, China, 2011.

[4] P. Schoeggl and E. Ramschak, *Vehicle Drivability Assessment Using Neural Networks For Development, Calibration And Quality Tests*, SAE Technical Paper, Pittsburgh, PA, USA, 2000.

[5] AVL-DRIVE Function Description, Advanced Software Version 3x, 2007.

[6] H. Jayaraman, N. Rao, S. Muthiah et al., "Optimization of tip-in response character of sports utility vehicle and verification with objective methodology," *Neuropeptides*, vol. 18, no. 2, pp. 87–91, 2015.

[7] M. Khodabakhshian, L. Feng, and J. Wikander, *Optimization of Gear Shifting and Torque Split for Improved Fuel Efficiency and Drivability of HEVs*, SAE World Congress and Exhibition, Detroit, MI, USA, 2013.

[8] W. Zhou, X. X. Guo, X. F. Pei, and C. C. Zhang, "Research on objective drivability evaluation with multi-source information fusion for passenger car," SAE, Pittsburgh, PA, USA, SAE Technical Paper 2020-01-1044, 2020.

[9] T. Deng, C. Lin, J. Luo, and B. Chen, "NSGA-II multi-objectives optimization algorithm for energy management control of hybrid electric vehicle," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 233, no. 4, pp. 1023–1034, 2019.

[10] W. S. Chang, K. Hyungkyoon, K. K. Mun, L. Wonsik, and W. C. Suk, "Development of an evaluation method for quantitative drivability in heavy-duty vehicles," *Journal of Mechanical Science and Technology*, vol. 28, no. 5, pp. 1615–1621, 2014.

[11] E. Galvagno, D. Morina, A. Sorniotti, and M. Velardocchia, "Drivability analysis of through-the-road-parallel hybrid vehicles," *Meccanica*, vol. 48, no. 2, pp. 351–366, 2013.

[12] S. Lakshmanan, A. Palaniappan, and V. Chekuri, "Methodology for evaluation of drivability attributes in commercial vehicle," SAE, Pittsburgh, PA, USA, SAE Technical Paper 2015-01-2767, 2015.

[13] X. Zhao, "Research on gearshift control for dual clutch transmission based on objective evaluation," Ph D. thesis, Jilin University, Changchun, China, 2015.

[14] S. G. Pickering and C. J. Brace, "Automated data processing and metric generation for driveability analysis," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 221, no. 4, pp. 429–441, 2007.

[15] C. H. Jiang, S. Chen, Y. W. Chen, and B. Y. Zhang, "A MEMS IMU de-noising method using long short term memory recurrent neural networks (LSTM-RNN)," *Sensors*, vol. 18, no. 10, pp. 1–14, 2018.

[16] A. S. EI-Wakeel, A. Osman, and N. Zorba, "Robust positioning for road information services in challenging environments," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3182–3195, 2020.

[17] W. Seo, S. Hwang, J. Park, and J.-M. Lee, "Precise outdoor localization with a GPS-INS integration system," *Robotica*, vol. 31, no. 3, pp. 371–379, 2013.

[18] C. Zhang and C. Y. Xu, "A de-noising method of acceleration signal for vehicle based on kalman filter and average filter," *Advanced Materials Research*, vol. 225-226, no. 1-2, pp. 605–608, 2011.

[19] Q. K. Zhao, C. G. Liu, X. R. Gao, and L. Luo, "Improved wavelet de-noising method of rail vibration signal for wheel tread detection," *Seventh International Symposium on Precision Engineering Measurements and Instrumentation*, vol. 8321, pp. 1–7, 2011.

[20] H. J. Liu, M. Li, W.. Huang, and R. H. Tong, "Signal de-noising method for whole vehicle drivability evaluation based on wavelet transform," *Noise and Vibration Control*, vol. 38, no. 1, pp. 103–108, 2018.

[21] H. J. Liu, S. G. Liu, and M. Li, "EMD and wavelet threshold de-noising method of gear-shift acceleration signals," *Noise and Vibration Control*, vol. 38, no. 2, pp. 198–203, 2018.

[22] F. Xiao, G. S. Chen, W. Zatar, and J. L. Hulsey, "Signature extraction from the dynamic responses of a bridge subjected to a moving vehicle using complete ensemble empirical mode decomposition," *Journal of Low Frequency Noise, Vibration and Active Control*, pp. 1–17, 2019.

[23] C. Jauch, S. Tamilarasan, and K. Bovee, "Modeling for drivability and drivability improving control of HEV," *Control Engineering Practice*, vol. 70, pp. 50–62, 2018.

[24] C. Jauch, S. Tamilarasan, and K. Bovee, "Design and verification of drivability improving control fro the Eco-CAR 2 hybrid electric vehicle," in *Proceedings of the 2016 American Control Conference*, pp. 631–636, Boston, MA, USA, July 2016.

[25] K. Koprubasi, E. R. Westervelt, and G. Rizzoni, "Experimental validation of a model for the control of drivability in a hybrid-electric vehicle," in *Proceedings of the ASME International Mechanical Engineering Congress and Exposition*, vol. 16, pp. 105–114, Seattle, WA, USA, November 2007.

[26] W. Huang, H. J. Liu, and Y. F. Ma, "Drivability evaluation model using principal component analysis and optimized extreme learning machine," *Journal of Vibration and Control*, vol. 25, no. 6, pp. 2274–2281, 2019.

[27] L. Y. Liang, S. Chen, and P. Li, *The Evaluation of Vehicle Interior Impact Noise Inducing by Speed Bumps Based on Multi-Features Combination and Support Vector Machine*, , pp. 1–21, Applied Acoustics, 2020.

[28] Y. Fu, Y. Lei, S. H. Shao, and H. Zeng, "Shift quality evaluation of DCT based on TOPSIS model," SAE, Pittsburgh, PA, USA, SAE Technical Paper 2014-01-1166, 2014.

[29] P. Schoeggl and E. Ramschak, "Vehicle drivability assessment using neural networks for development, calibration and quality tests," SAE, Pittsburgh, PA, USA, SAE Technical Paper 2000-01-0702, 2000.

[30] W. Huang and H. J. Liu, "Application of fuzzy dynamic weights drivability evaluation model in tip-in condition," *Journal of Vibration and Control*, vol. 25, no. 4, pp. 1–9, 2018.

[31] K. S. Leung, H. B. Ji, and L. Yee, "Adaptive weighted outer-product learning associative memory," *IEEE Transactions on System, Man, and Cybernetics, Part B, Cybernetics*, vol. 27, no. 3, pp. 533–543, 1997.

[32] L. Zhou, Q. X. Yu, and D. Z. Liu, "Compressive sensing-based vibration signal reconstruction using sparsity adaptive subspace pursuit," *Advances in Mechanical Engineering*, vol. 10, no. 8, pp. 1–12, 2018.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8892989, 10 pages
https://doi.org/10.1155/2020/8892989

Research Article

An Intelligent Forensics Approach for Detecting Patch-Based Image Inpainting

Xinyi Wang,[1] **He Wang**,[2] **and Shaozhang Niu**[1]

[1]*Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China*
[2]*State Grid Jiangsu Electric Power Co. Ltd., Marketing Service Center, Nanjing, China*

Correspondence should be addressed to Shaozhang Niu; szniu@bupt.edu.cn

Received 17 September 2020; Revised 10 October 2020; Accepted 18 October 2020; Published 28 October 2020

Academic Editor: William Guo

Image inpainting algorithms have a wide range of applications, which can be used for object removal in digital images. With the development of semantic level image inpainting technology, this brings great challenges to blind image forensics. In this case, many conventional methods have been proposed which have disadvantages such as high time complexity and low robustness to postprocessing operations. Therefore, this paper proposes a mask regional convolutional neural network (Mask R-CNN) approach for patch-based inpainting detection. According to the current research, many deep learning methods have shown the capacity for segmentation tasks when labeled datasets are available, so we apply a deep neural network to the domain of inpainting forensics. This deep learning model can distinguish and obtain different features between the inpainted and noninpainted regions. To reduce the missed detection areas and improve detection accuracy, we also adjust the sizes of the anchor scales due to the inpainting images and replace the original nonmaximum suppression single threshold with an improved nonmaximum suppression (NMS). The experimental results demonstrate this intelligent method has better detection performance over recent approaches of image inpainting forensics.

1. Introduction

With the popularity of digital cameras and smartphones, digital images have become increasingly widely used. However, a large amount of powerful, user-friendly image editing and processing software is making it easier to edit and modify digital images. People can easily modify digital images without having to learn professional and complex techniques and even use computers to synthesize realistic digital images, resulting in a large number of forgeries and composite images propagating on the network. It is difficult for ordinary people to visually observe the traces of these forged image modifications. Once those forged photos are used by people with bad purposes, there is no doubt that it will seriously threaten the stability and development of society. Therefore, the digital image forensics technology [1–4] has become essential. Passive image forensics methods designed to detect tamper traces without using prior knowledge, especially, have generated much research interest because they do not need any auxiliary information, for example, watermarks or signatures [5–8].

Passive image forensics is mostly targeted at specific tampering methods, such as double JPEG compression [9–11] and median filtering [12, 13]. Among them, the object removing operation is one of the most concerned malicious tampering methods. Since the reduction of image objects can mask important targets, the content of the image is changed to a large extent and affects the viewer's cognitive understanding. Currently, object removing operations are mainly implemented in two ways: copy-move [14, 15] and image inpainting methods [16–18].

This paper focuses on image tampering forensics based on inpainting techniques. Image inpainting is a significant study domain in computer vision and has attracted many researchers over the years [19–22]. Its main purpose is to use the information of the known area of the image to repair the damaged or removed area and to make the inpainted image keep the consistency on texture and structure as much as possible. We

can obtain a realistic visual effect so that the observer is unable to detect that the image was once edited. It can also be utilized to eliminate image semantic objects for malicious purposes. The basic symbols of image inpainting can be seen in Figure 1. In this sketch map, I is the original image and Λ represents the undamaged part, where Ω indicates the area to be repaired. Many image inpainting approaches fill the damaged area Ω using the undamaged part Λ. The patch-based methods [23, 24] are the representative studies of image restoration and many methods have been improved on this basis. However, these methods are often used to remove objects to change the semantics of images. As shown in Figure 2, this is an example of an image tampering operation using the inpainting algorithm, in which Figure 2(a) is an original image and Figure 2(b) is an inpainted image.

At present, there is much conventional research on inpainting forensics algorithms, which have various limitations and deficiencies. As a first attempt, Wu et al. [25] were the first to propose an image detection algorithm based on sample synthesis restoration. In their paper, the idea of zero connectivity to select suspicious regions was presented. Moreover, the block ambiguity level was used to distinguish the repaired patches. Since their approach required prior selection of the areas and used a full search strategy to find suspicious blocks, the computational complexity was high and could not meet the needs of applications. On the basis of this, Bacchuwar et al. [26] presented a jump patch-block matching method. Although some simplifications can save a certain amount of time, it was still a half-automatic detection method. While Zhang et al. [27] presented a faster forensics method using central pixel mapping (CPM). However, the rate of misrecognition was still high. These are conventional methods of inpainting detection, which rely on approximate characteristics between image blocks to exploit the difference between inpainting and noninpainting areas. They all have a large time complexity to extract features and low robustness to postprocessing operations. To complete the forensic task of the repair area, we need to extract features from the images and then distinguish the pixels in the image into two categories: inpainting and noninpainting. However, the inpainting operation often leaves no obvious traces, so it is difficult to obtain features with high discrimination by conventional methods. To overcome these issues, we employ an improved network based on the Mask Regional Convolutional Neural Network (Mask R-CNN) [15] to detect the inpainting manipulation and identify the inpainting localization.

In this paper, our main contributions are as follows. First, the backbone of the Mask R-CNN is applied to detect and locate the manipulated areas under complex backgrounds successfully. Use the prior information at the pixel level of the inpainted area to guide and supervise the training of this deep neural network. Then, we adjust and improve the network according to the shape of the inpainted data, including the size of the anchor scales and an improved method using the threshold of nonmaximum suppression, so that the model can generate the more accurate area of interest. Finally, the deep neural network works well in our self-made dataset.

The rest of this paper is organized as follows. Section 2 reviews the background of the patch-based inpainting method. In Section 3, we describe the deep learning architecture for our inpainting forensics. Section 4 gives the extensive experimental results on the datasets. Finally, we conclude the paper in Section 5.

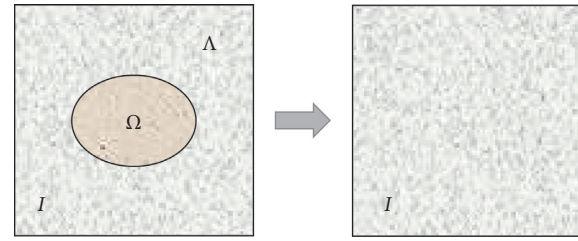


Figure 1: The schematic diagram of image inpainting.

2. Patch-Based Inpainting

Criminisi et al. [28] proposed the inpainting approach, combined both structure and texture features, which do not divide the image, and simultaneously processed the image texture and structure. So far, it is the most popular exemplar-based image inpainting algorithm. This algorithm first searches for the pixel block that best matches the area to be repaired and then fills the damaged area with the obtained pixel block. This inpainting method can obtain better results. Many of the subsequent studies on these image inpainting techniques were improved under the framework of the Criminisi algorithm. Therefore, we use the Criminisi algorithm as a representative example to introduce the patch-based inpainting principle in detail.

Figure 3 shows the patch-based inpainting process of the Criminisi algorithm. First, the user specifies a target area that needs to be repaired or removed as shown in Figure 3(a). There has been an image with the damaged area Ω and the known area Φ, and the aim of Criminisi's inpainting is to restore the target area (damaged area Ω) with the image information of the source area (known area Φ). The boundary area is represented by ∂Ω. The primary steps of the Criminisi algorithm for image restoration are as follows:

Step 1: compute the priorities of the points on the interior ∂Ω and find the point p with the highest priority. Then, the image patch Ψp centered at the point p is chosen as a target patch to be inpainted, which is shown in Figure 3(b).

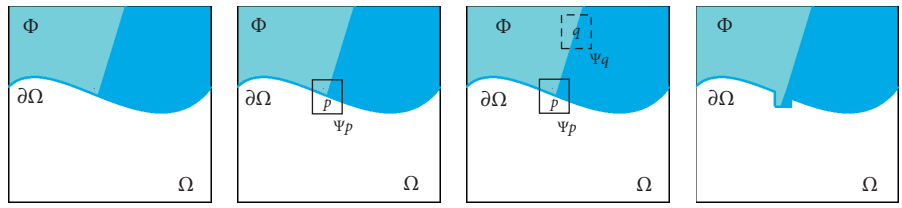
Step 2: search the whole known region Φ for the reference block Ψq which is the most similar to Ψp as shown in Figure 3(c). Minimizing the perceived distance d (Ψp, Ψq), which is used to measure the similarity between Ψp and Ψq, it is defined as follows:

$$d(\Psi p, \Psi q) = \sum_i \sum_j |\Psi p(i,j) - \Psi q(i,j)|^2. \quad (1)$$

Step 3: search for the corresponding pixels of Ψq to repair the damaged area in Ψp, and keep the priority between Ψp and ∂Ω constantly refreshed, which can be seen in Figure 3(d).

(a) (b)

Figure 2: An example of semantic object removal using image inpainting method: (a) original image and (b) inpainted image.



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

Figure 3: Inpainting by Criminisi algorithm. (a) Determine target area Ω, (b) select target patch Ψp, (c) search for reference patch Ψq and (d) update priority and contour.

Step 4: cycle from the first to the third step until the damaged area is completely inpainted.

The Criminisi algorithm and the improved algorithm based on it have some dissimilarities in the methods of calculating the priority and finding the closest block that can match, but using these methods to fill the damaged block introduces the abnormal similarity. They differ from texture-like blocks in nature [12] because they are not generated by normal imaging algorithms. This will cause an unusual distribution of pixel values. Moreover, the block-filled area based on the patch will be different from the area of natural imaging. This leaves traceable traces for image forensics. We can mine these features to detect patch-based inpainting tampering.

3. Methodology

The backbone network in this paper is based on the deep learning framework: Mask R-CNN [29]. It is a small and flexible general object instance segmentation framework that can achieve the best experimental results. Therefore, Mask R-CNN can be thought of as a Faster R-CNN [30] boundary box detection model with a small Fully Convolution Network (FCN) [31]. Mask R-CNN is an extension above Faster R-CNN, adding a layer for predicting the segmentation mask on each region of interest (ROI), called the mask branch. It can effectively detect the object in the image and can also generate a high-quality segmentation mask for each instance, so it is equivalent to multitask learning. Since the mask layer only adds a small amount of computation to the entire system, this method can simultaneously obtain object detection and instance segmentation.

This deep neural network convolves the entire input image to obtain the feature map. After the candidate region is generated by the RPN network, the candidate frame is filtered by the method of improved nonmaximum suppression (NMS), and the candidate region corresponds to the high-dimensional feature vector on the feature map. Then, we use the full-convolution network to obtain the category of pixels in the image, calculate the score in the detection network, and finally output the detection and location results of the inpainted regions.

3.1. Architecture of the Proposed Method. This section describes the proposed method for patch-based image inpainting detection. The architecture of it is given in Figure 4. In this structure, the backbone network utilizes the 101-layer deep residual network ResNet and Feature Pyramid Network (FPN). ResNet is a framework for remnant learning that reduces the burden of network training. The network has a deeper level than the networks used before. It uses this layer to associate with the input layer to learn the residual function. FPN extracts ROI features from different levels of features based on feature size and corrects each ROI using ROI Align. After getting the feature map of each ROI region, the classification and bounding box of each ROI are predicted. Each ROI uses the designed FCN framework to predict the category of each pixel in the ROI region. Finally, a segmentation result of the image inpainted region is obtained.

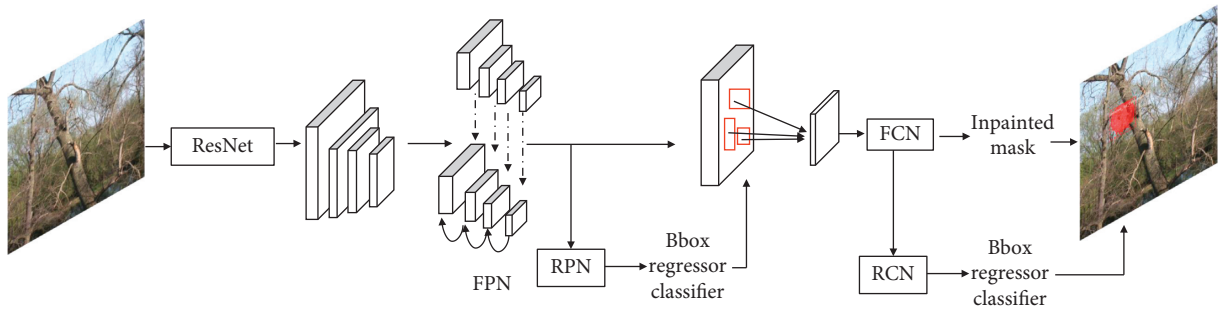


Figure 4: The architecture in our method.

The designed idea of the algorithm is as follows: manipulated localization, manipulated recognition, and semantic segmentation work together to achieve the purpose of enhancing object detection. Feature extraction is performed by the first backbone convolutional network ResNet, which obtains RoIAlign and is different from the RoIPool obtained by Faster R-CNN for the next object detection process. Then, the obtained RoIAlign is divided into two branches: one branch gains the boundary box and the object class through the fully connected network and the other branch gains the pixel-level instance segmentation result of the inpainted area through the calculation of the convolutional layer. The two branches are combined to gain the ultimate manipulated detection result. The adjunction of instance segmentation also improves the precise manipulated recognition, and the obtained detection information is more useful.

This process is simulated by a fully convoluted convolutional neural network. The network of RPN [30] is very small and it is shown in Figure 5, and the red box is a window that slides on the feature pyramid of the above trunk. Fully concatenate with the input convolutional feature map and map the sliding window to a low-dimensional vector, usually to 256 or 512 dimensions. The vector is put into two fully connected layers for classification and border regression. When n is chosen to be 3, the effective area of the input image is the largest. At the position of each sliding window, we simultaneously predict k proposal regions, and the box regression layer will generate $4k$ coordinate codes of k boxes. The object recognition layer for each recommended area or the nonobject likelihood frame recognition layer will generate $2k$ scores. The k proposals are parameterized and are referred to as anchors. Each anchor is concentrated in the middle of the sliding window in the identification problem. Usually, 3 scales and 3 aspect ratios are used, resulting in $k = 9$ anchors at each sliding window position. For size $W \times H$ (usually less than 2400), there are typically WHk anchors. An important feature of this method is that it has translation invariance and has this property in both the computational anchor and the function of calculating correlations and anchors. In contrast, multiframes using the K-means method to generate anchors do not have translation invariance. If the object position in an image changes, the proposed object should also change, so the prediction of the proposal object could be at any position.

3.2. Improved NMS. The sizes of anchor scales are adjusted to (16, 32, 64, 128, 256) due to the inpainted image data. We use Soft-NMS [19] instead of the original NMS. Because it uses a scoring system to score a lower score than the threshold of IOU, instead of deleting it directly.

After the detection frame is generated in the object detection, the traditional NMS [28] can eliminate the repeated bounding box. Generally, the detection box is sorted according to the score, the box with the highest score will be selected, and other boxes with the coincidence degree exceeding the threshold will be removed. However, this method of uniformly zeroing the detection frames that exceed the threshold is easy to make objects that are closer and have occlusions not detected. Therefore, we use an improved nonmaximum suppression (NMS) [32] instead, which uses a scoring system to evaluate scores below the Intersection over Union (IoU) threshold, rather than directly deleting them.

The improved NMS is shown in the following equation:

$$s_i = \begin{cases} s_i, & \mathrm{IoU}\,(M, b_i) < N_i, \\ s_i\left(1 - \mathrm{IoU}\left(M, b_i\right)\right), & \mathrm{IoU}\left(M, b_i\right) \geq N_i. \end{cases} \tag{2}$$

Among them, N is the threshold, M is the box with the highest current score, s_i is the corresponding score, and b_i is the box to be processed. The improved NMS does not directly remove the detection frame that exceeds the threshold with the current highest score frame but reduces the score of the detection frame. When the degree of coincidence of the detection frame is greater, its score will also decay more. If there is only a small part of the overlap, it will not be much impact on the original detection score. In this way, the prediction frame will not be deleted by mistake when the objects of the same type are highly overlapped, and the improved NMS algorithm does not introduce any hyperparameters during the training. The hyperparameters used to adjust the improved NMS algorithm will only appear in the test or demonstration phase and will not cause an increase in computational complexity.

3.3. Loss Construction. The loss function of the RPN is defined as the proposed object should also change, so the prediction of the proposal object could be at any position:

$$L\left(\{s_i\}, \{t_i\}\right) = \frac{1}{N_{\mathrm{cls}}} \sum_i L_{\mathrm{cls}}\left(s_i, s_i^*\right) + \lambda \frac{1}{N_{\mathrm{reg}}} \sum_i s_i^* c\left(t_i, t_i^*\right). \tag{3}$$

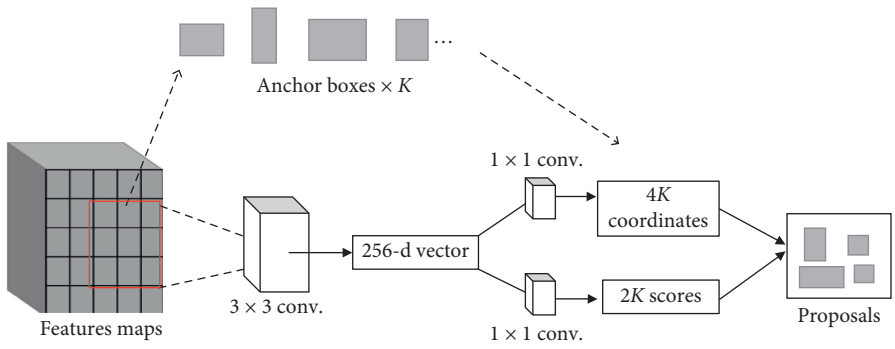


Figure 5: The architecture of the RPN.

The RPN network consists of two parts: one is to judge whether the bounding box is the foreground or the background and the other is to predict the regression of the bounding box. The loss function of the corresponding RPN network is also composed of two parts, as $L(\{s_i\}, \{t_i\})$, where L_{cls} is the partial loss function of the classification, L_{reg} is the regression loss function of the bounding box, N_{cls} is the minibatch size, N_{reg} is the anchor location, and the weighted sum of the two partial loss functions is the total loss function of the RPN.

Here, i is the anchor index, s_i and s_i^* are the probability that the bounding box is predicted to be the foreground and the probability that the prediction box is the foreground, t_i and t_i^* are the coordinates of the predicted border and the coordinates of the ground-truth border, respectively, and $s_i^* L_{\text{reg}}$ indicates that the frame is the foreground. It performs regression calculations.

The model initialized by ImageNet training is finetuned end-to-end by region proposal. Then, it uses the proposal generated before to train the detection network through Fast R-CNN segmentation. It is also initialized by the ImageNet training model. The third step is to initialize the regional proposal training network through the object detection network, but only the shared convolution layer and finetuned region proposal network. Now the two networks share the convolution layer.

Finally, the shared convolution layer is fixed, so that the two networks share the convolution layer to form a unique network. For pixel-level segmentation, it is parallel to the above box regression and object recognition. The multitask loss function is defined for the region of interest:

$$L_{\text{Total}} = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}. \tag{4}$$

Due to the fact that the prediction of masks is depending on the region proposals, the loss of mask must be added to the total loss function. This can make region proposals more precise. In general, the prediction of masks and the region proposals complement each other and ultimately improve the accurate positioning of the inpainted boundary.

4. Experimental

In this paper, all experiments are performed on Ubuntu 16.04 of NVidia GeForce GTX 2080 Ti, operating in an Intel Core CPU i7-9700K. The sizes of anchor scales are adjusted to (16, 32, 64, 128, 256) due to the inpainted image dataset.

4.1. Dataset and Evaluation Metrics

4.1.1. Dataset. We select two typical image databases for experiments in our inpainting detection network. First, in the COCO [33] dataset, we randomly selected 2×10^4 color images and selected all 2,000 color images in the UCID [34] dataset; the size is cropped into 256×256. The inpainted images generated in the dataset are all repaired using the Criminisi algorithm of [28], which have different tampering areas in sizes and shapes. Half of the images are selected with some regular masks, resulting in tampering areas with tampering rates of 5%, 10%, and 20% but randomly selected for tampering. For the remaining half of the image, the inpainting area has an irregular shape, and the inpainting ratio varies between 1% and 50%. Several sample images are shown in Figure 6 (masked area is marked in green). A ground-truth tag matrix for repairing an image is formed depending on the tamper region used. Finally, the inpainted images with the corresponding ground-truth are split into two datasets: the part containing 80% of the images are training and validation dataset and 20% of the images are the testing dataset. Separate the training and the testing set to make sure that equal background and inpainted operation do not occur between them.

4.1.2. Evaluation Metrics. We choose True Positives Rate (TPR), False Positives Rate (FPR), and Accuracy Precision (AP) as evaluation metrics standard. Compare the performance of their detection in the same dataset with the approaches in [35, 36]. One of the algorithms for selecting contrast is that the traditional method has a better effect, and the other is the latest method of using the deep learning algorithm.

4.2. Forgery Detection. Figure 7 shows the detection results of the inpainted images (the first two column of Figure 6) using our method. We can see the detection bounding box, class and confidence score in the resulting image. This proves that the method of this paper can accurately detect and locate the tampering area (inpainted by the method in [28]).

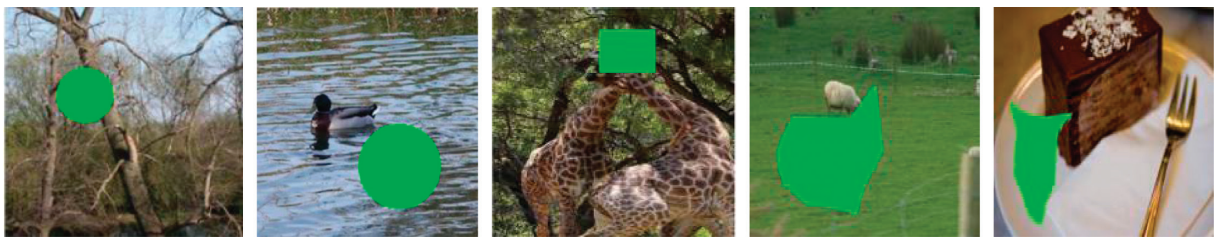


Figure 6: Sample images with the damaged areas (marked in green) of the regular and irregular mask shapes.

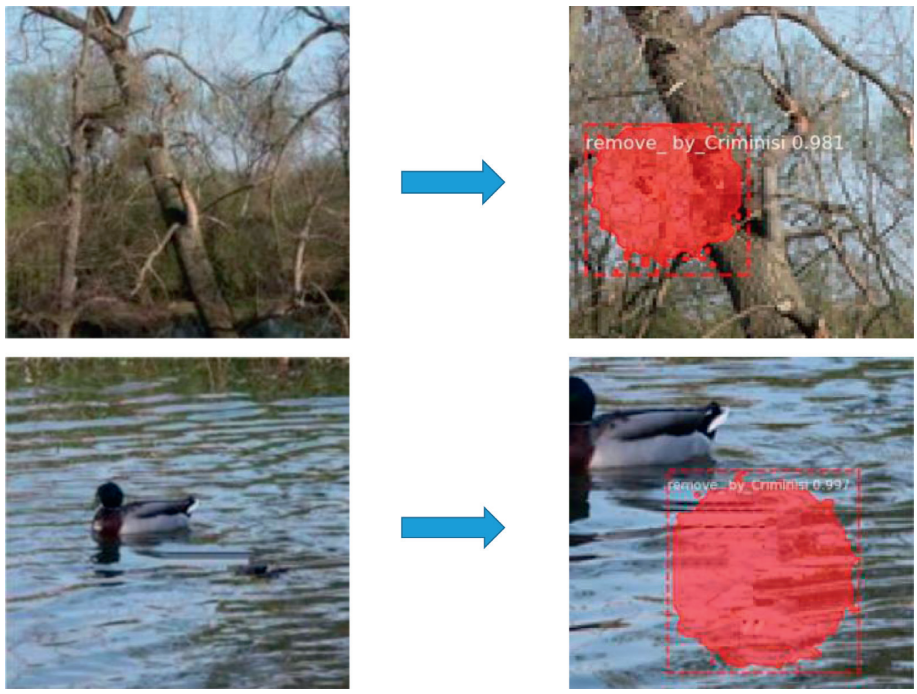


Figure 7: The detection results of the first two sample images in Figure 6 after regular inpainting using the Criminisi method [28].

Some examples of image inpainting detection can be seen in Figures 8 and 9, we find these three test algorithms from left to right: [35, 36] and our method can basically distinguish the inpainted area which has been listed in Figure 6. The last column of this figure shows that we can get more accurate pixel-level positioning in the detection using our method in different inpainted shapes. The other two methods have different degrees of false alarm pixels, and our method basically does not show false alarm pixels. Compared with [35, 36], our deep neural network provides complete and accurate detection on each measured image (closer to the ground truth in Figure 6).

For these three forensic methods, Table 1 summarizes three evaluation metrics averaged on the test datasets. The experimental results show that our network performs the highest True Positives Rate of 96.7%, the lowest False Positives Rate of 1.9%, and the tampering rate of more than 30% on the test images with irregular tampering areas. At the same time, we can also see that the detection performance decreases at the slow pace as the tampering ratio decreases. Moreover, our method still achieves less than 1.6% FPR with the images without restoration (extreme case). This demonstrates that our intelligent neural network can definitely capture the inpainting features. Similar experiments are carried out on circular or rectangular inpainted areas, our network shows slightly less effect under the regular shape mask. We suspect this is because our network has obtained the shape characteristics of the inpainted areas, which may have an adverse effect. Among those different shapes of inpainted areas we tested in the experiments, the performance of our method in FPR and AP is obviously better than that of other methods [35, 36].

4.3. Experimental Results. Considering tampering images are often attacked by JPEG compression and image scaling, we test the robustness of our proposed approach in Tables 2 and 3. Actually, we randomly select 3000 images from COCO [33] dataset. Then, the images with irregular tamper area and tamper rate greater than 30% are generated. In order to get the tampered images for robustness testing, we compress those images with 65%, 70% 80%, 90%, and 95% quality factor (QF), respectively. For the image scaling, these selected images are scaled in the range of 0.5, 0.75, and 1.5, respectively. The comparisons of ROC curve under two different attacks are shown in Figure 10.

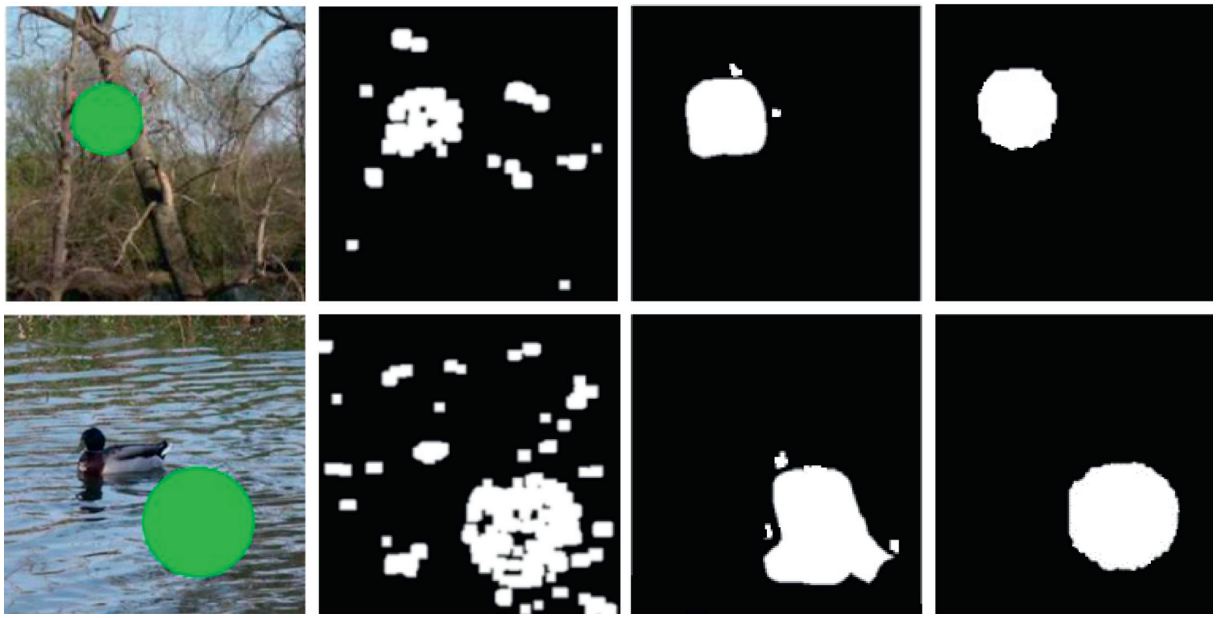


Figure 8: The above regular inpainted images (1st column) and the corresponding tampered region detection results obtained by the methods proposed in [35] (2nd column) and [36] (3rd column), and our method (4th column).

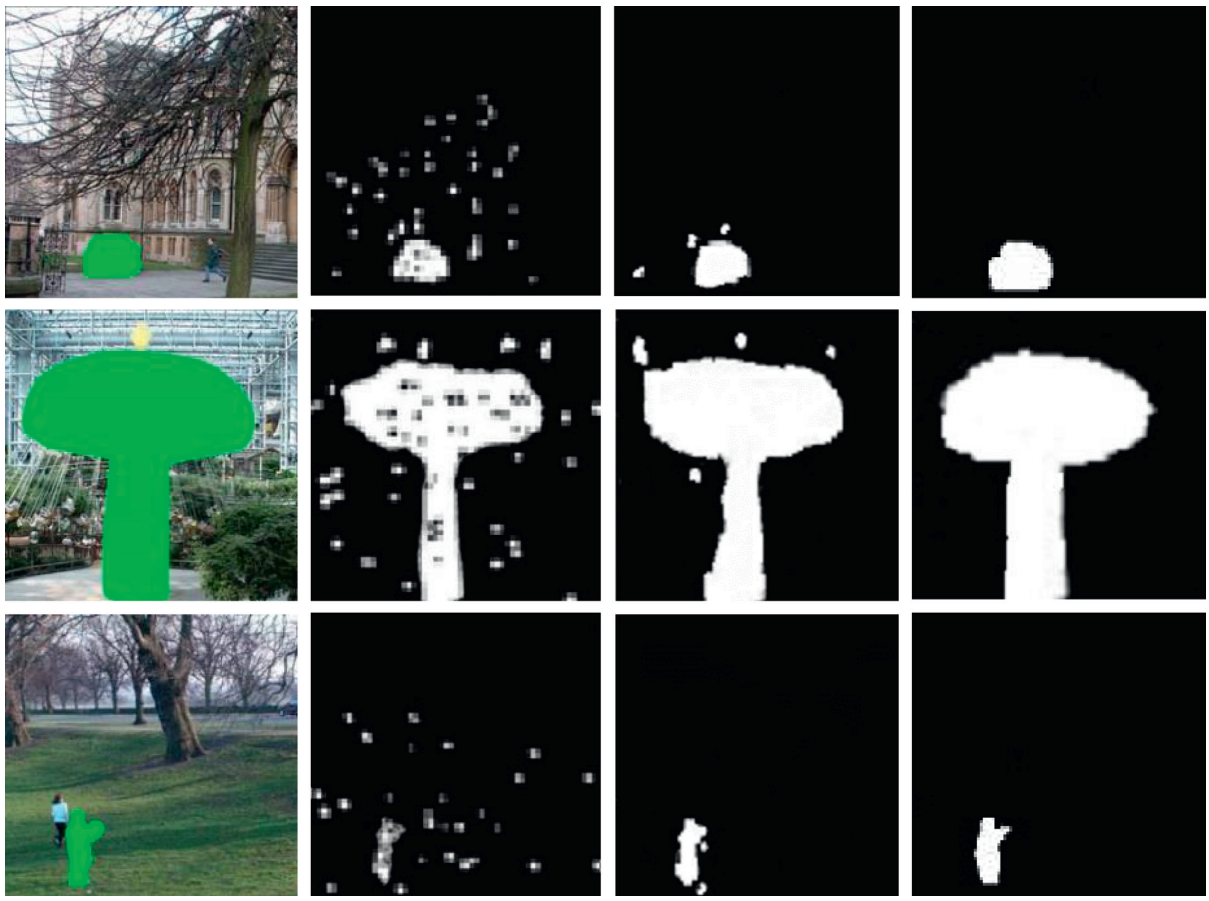


Figure 9: Some irregular inpainted images (1st column) and the corresponding tampered region detection results obtained by the methods proposed in [35] (2nd column), and [36] (3rd column), and our method (4th column).

Table 1: Performance experiments of three methods with different evaluation metrics standards.

Region shape		Regular			Irregular		
Tamper ratio		5	10	20	[0, 10)	[10, 30)	[30, 50)
Method [35]	TPR	81.2	87.8	90.5	73.5	84.6	86.5
	FPR	3.1	8.5	16.4	7.5	15.2	52.6
	AP	94.5	91.4	82.6	92.5	82.2	56.4
Method [36]	TPR	85.3	82.5	81.6	87.6	90.4	95.5
	FPR	2.3	2.5	2.8	1.2	2.3	3.2
	AP	94.5	95.2	94.8	93.1	95.7	93.5
Proposed	TPR	89.3	86.5	85.1	90.5	93.4	96.7
	FPR	1.2	1.5	1.7	0.9	1.4	1.9
	AP	97.5	96.4	96.9	98.5	97.4	96.4

Table 2: The accuracy performance of different methods under five JPEG compression (with quality of 65, 70, 80, 90, and 95).

QF (%)	Method [35]	Method [36]	Proposed
65	44.5	76.8	82.73
70	53.8	85.4	87.45
80	69.5	89.7	92.11
90	82.4	91.3	94.2
95	92.4	93.8	96.7

Table 3: The accuracy performance of three methods under different scaling conditions (with scaling factors 0.5, 0.75, and 1.5).

Scaling (%)	Method [35]	Method [36]	Proposed
0.5	61.2	87.9	90.1
0.75	75.3	90.6	93.4
1.5	84.6	91.3	92.7

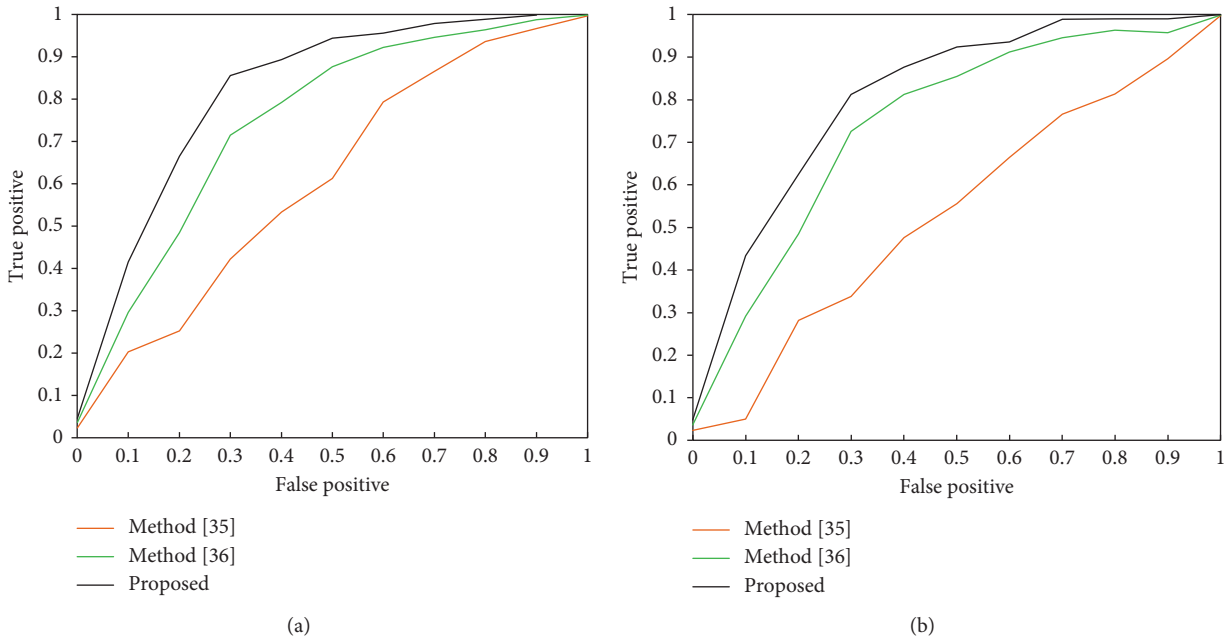


Figure 10: ROC curve comparison of three forensic methods under different attack operations. (a) ROC curve comparison (95% JPEG compression quality) and (b) ROC curve comparison (scaling ratio 1.5).

According to the experimental results (Table 2 and the ROC curve in Figure 10), compared with the test image without JPEG compression, the detection accuracy of the method used in the paper is slightly less than 96.7%, and the low-quality compression strength at 65% also has an accuracy rate of more than 80%, which is better than method

[36] under different compression attacks. The traditional detection method [35] is greatly affected by JEPG compression. When the compression factor is lower than 80%, the accuracy rate does not exceed 60%. Experimental results show that our proposed network can maintain the robustness of JPEG compression. Both Table 3 and the ROC curve show that when the scaling factor is gradually away from 1, AP is less affected in our proposed method, method [36] has a certain effect, and method [35] has the greatest effect. Therefore, our method is insensitive to scaling. In conclusion, our method based on the deep neural network can keep robust to JEPG and scaling attacks.

Although the traditional detection method [35] uses a search algorithm based on weight transformation to speed up the detection and reduces the search time for matching blocks, it usually takes several minutes. The intelligent detection algorithms using the deep learning method, our proposed method, and method [36], in addition to consuming some time during training, generally only take a few seconds during detection and can basically be detected in real time.

5. Conclusion

This paper proposes an effective and intelligent image inpainting detection approach using the deep neural network. In order to distinguish and obtain different features between the inpainted and noninpainted regions, this paper applies the improved region detection network here, converts the object detection problem to a pixel-level segmentation, and uses the mask to create the bounding box. The nonmaximum suppression, especially, is defined by using the overlap ratio of the mask area instead of the bounding box to filtrate the output result. Adjust the anchors scale and the step size and use the improved NMS of the area suggestion network according to the object morphological in order to generate a more accurate region of interest. This paper proves that the state-of-the-art instance segmentation model can obtain the differences of features in inpainted images. In the future, we will optimize the network and increase the tampering dataset, which may improve the effectiveness of different image tampering forensics.

Data Availability

The tables and figures' data used to support the findings of this study are included within the article. Previously reported tables and figures' data were used to support this study and are available in the relevant references. These prior studies (and datasets) are cited at relevant places within the text as references [31–36]. The data supporting this research are from previously reported studies and datasets, which have been cited. The processed data 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 National Natural Science Foundation of China (nos. 61370195 and U1536121).

References

[1] X. Wang, S. Niu, and J. Zhang, "Digital image forensics based on CFA interpolation feature and Gaussian mixture model," *International Journal of Digital Crime and Forensics*, vol. 11, no. 2, pp. 1–12, 2019.

[2] S. Y. Li, "Based on digital image forensics edge characteristic morphological filtering technology," *Advanced Materials Research*, vol. 912–914, pp. 1181–1184, 2014.

[3] S. Xiaoting, L. Yezhou, N. Shaozhang, and H. Yanli, "The detecting system of image forgeries with noise features and EXIF information," *Journal of Systems Ence & Complexity*, vol. 28, no. 5, pp. 1164–1176, 2015.

[4] J. Y. Xu and Y. T. Su, "Smoothing filtering detection for digital image forensics," *Journal of Electronics & Information Technology*, vol. 35, no. 10, pp. 2287–2293, 2014.

[5] H. Guo, A. Y. Li, and S. Jajodia, "A fragile watermarking scheme for detecting malicious modifications of database relations," *Information Sciences*, vol. 176, no. 10, pp. 1350–1378, 2006.

[6] I. J. Cox and J.-P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 4, pp. 587–593, 1998.

[7] T. I. Williams, K. L. Toups, D. A. Saggese et al., "Optimization methods for the insertion, protection, and detection of digital watermarks in digitized data," *Journal of Proteome Research*, vol. 6, no. 8, pp. 2936–2962, 2003.

[8] S. St George and E. Nielsen, "Signatures of high-magnitude 19th-century floods in *Quercus macrocarpa* tree rings along the red river, Manitoba, Canada," *Geology*, vol. 28, no. 10, pp. 889–899, 2000.

[9] A. Taimori, F. Razzazi, A. Behrad, A. Ahmadi, and M. Babaie-Zadeh, "A novel forensic image analysis tool for discovering double JPEG compression clues," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7749–7783, 2017.

[10] Y. Cao, T. Gao, G. Sheng, L. Fan, and L. Gao, "A new anti-forensic scheme—hiding the single JPEG compression trace for digital image," *Journal of Forensic Sciences*, vol. 60, no. 1, pp. 197–205, 2015.

[11] X. Huang, S. Wang, and G. Liu, "Detecting double jpeg compression with same quantization matrix based on dense cnn feature," in *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, October 2018.

[12] Y. Niu, Y. Zhao, and R. Ni, "Robust median filtering detection based on local difference descriptor," *Signal Processing: Image Communication*, vol. 53, no. 2, pp. 65–72, 2017.

[13] C. Chen, J. Ni, and J. Huang, "Blind detection of median filtering in digital images: a difference domain based approach," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 22, no. 12, pp. 4699–4710, 2013.

[14] S. Bayram, H. Taha Sencar, and N. Memon, "An efficient and robust method for detecting copy-move forgery," in *Proceedings of the IEEE International Conference on Acoustics*, Taipei, Taiwan, April 2009.

[15] X. Y. Wang, L. X. Jiao, X. B. Wang, H. Y. Yang, and P. P. Niu, "A new keypoint-based copy-move forgery detection for color image," *Applied Intelligence*, vol. 48, no. 10, pp. 3630–3652, 2018.

[16] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 882–889, 2003.

[17] M. Bertalmio, G. Sapiro, V. Caselles et al., "Image inpainting," in *Proceedings of the SIGGRAPH Conference*, pp. 417–424, New Orleans, LA, USA, July 2000.

[18] X. N. Tang, J. G. Chen, C. M. Shen, and G. X. Zhang, "Modified exemplar-based image inpainting algorithm," *Journal of East China Normal University*, vol. 135, no. 6, pp. 24–28, 2016.

[19] Y. Li, D. Jeong, J.-i. Choi, S. Lee, and J. Kim, "Fast local image inpainting based on the Allen-Cahn model," *Digital Signal Processing*, vol. 37, pp. 65–74, 2015.

[20] P. Arias, G. Facciolo, and G. V. Sapiro, "A variational framework for exemplar-based image inpainting," *International Journal of Computer Vision*, vol. 93, no. 3, pp. 319–347, 2011.

[21] C. Qin, C.-C. Chang, and Y.-P. Chiu, "A novel joint data-hiding and compression scheme based on SMVQ and image inpainting," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 969–978, 2014.

[22] C. Qin, Q. Zhou, F. Cao et al., "Flexible lossy compression for selective encrypted image with image inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 3341–3355, 2019.

[23] K. A. Vahid and Y. Farzin, "Exemplar-based image inpainting using SVD-based approximation matrix and multi-scale analysis," *Multimedia Tools Applications*, vol. 76, no. 5, pp. 7213–7234, 2017.

[24] H. Wang, L. Jiang, R. Liang, and X. L. Xiao, "Exemplar-based image inpainting using structure consistent patch matching," *Neurocomputing*, vol. 269, pp. 90–96, 2017.

[25] Q. Wu, S. J. Sun, W. Zhu, H. L. Guo, and D. Tu, "Detection of digital doctoring in exemplar-based inpainted images," in *Proceedings of the International Conference on Machine Learning Cybernetics*, Kunming, China, July 2008.

[26] K. S. Bacchuwar, K. Aakashdeep, and K. R. Ramakrishnan, "A jump patch-block match algorithm for multiple forgery detection," in *Proceedings of the International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, Kottayam, India, March 2013.

[27] D. Zhang, Z. Liang, G. Yang, Q. Li, L. Li, and X. Sun, "A robust forgery detection algorithm for object removal by exemplar-based image inpainting," *Multimedia Tools and Applications*, vol. 77, no. 10, pp. 11823–11842, 2018.

[28] A. Criminisi, P. Pérez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, WI, USA, June 2003.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 42, no. 2, pp. 1–13, 2020.

[30] 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, 2017.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentati," in *Proceedings of the International Conference Computer Vision Pattern Recognition CVPR*, pp. 3431–3440, Boston, MA, USA, June 2015.

[32] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—improving object detection with one line of code," in *Proceedings of the IEEE International Conference Computer Vision Pattern Recognition CVPR*, Venice, Italy, October 2017.

[33] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, September 2014.

[34] G. Schaefer and M. Stich, "Ucid: an uncompressed color image database," *The International Society for Optical Engineering*, vol. 5307, pp. 472–480, 2004.

[35] C. I. Chang, J.-C. Yu, and C.-C. Chang, "A forgery detection algorithm for exemplar-based inpainting images using multi-region relation," *Image and Vision Computing*, vol. 31, no. 1, pp. 57–71, 2013.

[36] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun, "A deep learning approach to patch-based image inpainting forensics," *Signal Processing: Image Communication*, vol. 67, pp. 90–99, 2018.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4696198, 10 pages
https://doi.org/10.1155/2020/4696198

Research Article

Analyzing Machine Learning Models with Gaussian Process for the Indoor Positioning System

Yunxin Xie,[1] Chenyang Zhu,[2] Wei Jiang,[2] Jia Bi,[3] and Zhengwei Zhu[2]

[1]*School of Petroleum Engineering, Changzhou University, Changzhou 213100, China*
[2]*School of Information Science and Engineering, Changzhou University, Changzhou 213100, China*
[3]*Electronics and Computer Science, University of Southampton, University Road, Southampton SO17 1BJ, UK*

Correspondence should be addressed to Chenyang Zhu; cz4g16@soton.ac.uk

Received 18 August 2020; Revised 26 September 2020; Accepted 14 October 2020; Published 24 October 2020

Academic Editor: William Guo

Recently, there has been growing interest in improving the efficiency and accuracy of the Indoor Positioning System (IPS). The Received Signal Strength- (RSS-) based fingerprinting technique is essential for indoor localization. However, it is challenging to estimate the indoor position based on RSS's measurement under the complex indoor environment. This paper evaluates three machine learning approaches and Gaussian Process (GP) regression with three different kernels to get the best indoor positioning model. The hyperparameter tuning technique is used to select the optimum parameter set for each model. Experiments are carried out with RSS data from seven access points (AP). Results show that GP with a rational quadratic kernel and eXtreme gradient tree boosting model has the best positioning accuracy compared to other models. In contrast, the eXtreme gradient tree boosting model could achieve higher positioning accuracy with smaller training size and fewer access points.

1. Introduction

Wireless indoor positioning is attracting considerable critical attention due to the increasing demands on indoor location-based services. Examples of this service include guiding clients through a large building or help mobile robots with indoor navigation and localization [1]. However, the global positioning system (GPS) has been used for outdoor positioning in the last few decades, while its positioning accuracy is limited in the indoor environment. Moreover, the GPS signals indoor are also limited so that it is not appropriate for indoor positioning. Generally, the IPS is classified into two types, namely, a radiofrequency-based system and infrared-based system. The radiofrequency-based system utilizes signal strength information at multiple base stations to provide user location services [2]. The infrared-based system uses sensor networks to collect infrared signals and deduce the infrared client's location by checking the location information of different sensors [3]. As the coverage range of infrared-based clients is up to 10 meters while the coverage range of radiofrequency-based clients is up to 50 meters, radiofrequency has become the most commonly used technique for indoor positioning.

Estimating the indoor position with the radiofrequency technique is also challenging as there are variations of signals due to the motion of the portable unit and dynamics of the changing environment [4]. Moreover, the traditional geometric approach that deduces the location based on the angle and distance estimates from different signal transmitters is problematic as the transmitted signal might be distorted due to reflections and refraction and the indoor environment [5]. Machine learning approaches can avoid the complexity of determining an appropriate propagation model with traditional geometric approaches and adapt well to local variations of indoor environment [6]. Thus, we use machine learning approaches to construct an empirical model that models the distribution of Received Signal Strength (RSS) in an indoor environment. The model can determine the indoor position based on the RSS information in that position. The model-based positioning system involves offline and online phases. The RSS readings from different AP are collected during the offline phase with the machine learning

approach, which captures the indoor environment's complex radiofrequency profile [7]. The model is then trained with the RSS training samples. During the online phase, the client's position is determined by the signal strength and the trained model. Moreover, there is no state-of-the-art work that evaluates the model performance of different algorithms. No guidelines of the size of training samples and the number of AP are provided to train the models.

In this paper, we compare three machine learning models, namely, Support Vector Regression (SVR), Random Forest (RF), and eXtreme Gradient Tree Boosting (XGBoost), with the Gaussian Process Regression (GPR) to find the best model for indoor positioning. Each model is trained with the optimum parameter set obtained from the hyperparameter tuning procedure. Besides, the GPR is trained with three kernels, namely, Radial-Basis Function (RBF) kernel, Matérn kernel, and Rational Quadratic (RQ) kernel, and evaluated with the average error and standard deviation. We used the hyperparameter tuning procedure to tune the parameter for each model and get the optimal parameter set for each model and then compare the performances. The prediction results are evaluated with different sizes of training samples and numbers of AP. Results show that the XGBoost model outperforms all the other models and related work in positioning accuracy. Moreover, the XGBoost model can also achieve high positioning accuracy with smaller training size and fewer APs. We design experiment and use results to show the optimal number of access points and the size of RSS data for the optimal model.

This paper is organized as follows. Section 2 summarizes the related work that constructs models for indoor positioning. Section 3 introduces the background of machine learning approaches as well as the kernel functions for GPR. Sections 4 and 5 describe procedure and experiment result we carried out for the indoor positioning with different approaches. Section 6 concludes the paper and outlines some future work.

2. Related Work

A great deal of previous research has focused on improving the indoor positioning accuracy with machine learning approaches. Brunato evaluated the k-nearest-neighbor approach for indoor positioning with wireless signals from several access points [8], which has an average uncertainty of two meters. Battiti et al. compared the neural network- (NN-) based model and k-nearest-neighbor model to determine the mobile terminal under the wireless LAN environment [9]. Results show that the NN model performs better than the k-nearest-neighbor model and can achieve a standard average of 1.8 meters. Wu et al. compared different kernel functions of the support vector regression to estimate locations with GSM signals [6]. Their results show that the SVR models have better positioning performance compared with NN models. As SVR has the best prediction performance in the current work, we select SVR as a baseline model to evaluate the performance of the other three machine learning approaches and the GPR approach with different kernels.

Besides machine learning approaches, Gaussian process regression has also been applied to improve the indoor positioning accuracy. Schwaighofer et al. built Gaussian process models with the Matérn kernel function to solve the localization problem in cellular networks [5]. Bekkali et al. compared the kernel functions for GPR and developed a location sensing system based on RSS data [7]. Alfakih et al. proposed a Gaussian Mixture Model to approximate the distribution of RSS for indoor localization [10]. Their approach reaches the mean error of 1.6 meters. Less work has been done to compare the GPR with traditional machine learning approaches. Our work assesses the positioning performance of different models and experiments on the size of training samples and the number of APs for the optimum model.

3. Machine Learning Models and Gaussian Process Regression

In the past decade, machine learning played a fundamental role in artificial intelligence areas such as lithology classification, signal processing, and medical image analysis [11–13]. More recently, there has been extensive research on supervised learning to predict or classify some unseen outcomes from some existing patterns. Given a set of data points $\{x_1, x_2, \ldots, x_n\}$ associated with set of labels $\{y_1, y_2, \ldots, y_n\}$, supervised learning could build a regressor or classifier to predict or classify the unseen y from x. Here each x_i is a feature vector with size n and each y_i is the labeled value. A model h_θ is built with supervised learning for the given input x_i and the predicted value is $h_\theta(x_i)$. The training process of supervised learning is to minimize the difference between predicted value $h_\theta(x_i)$ and the actual value y_i with a loss function $L(h_\theta(x_i), y_i)$. The model performance of supervised learning is usually assessed by $\sum_{i=1}^{M} L(h_\theta(x_i), y_i)$.

3.1. Support Vector Regression. The support vector machine (SVM) model is usually used to construct hyperplane to separate high-dimensional feature space and distinguish data from different classes [14]. Drucker et al. proposed a support vector regression (SVR) algorithm that applies a soft margin of tolerance ξ in SVM to approximate and predict values [15]. ξ is used to define the soft margin allowed for the model. The weights w of the model are calculated given that model function $h_\theta(x)$ is at most E from the target y; formally, $\forall i \cdot y_i - w * x_i - b \le \varepsilon + \xi_i \wedge w * x_i + b - y_i \le \varepsilon + \xi_i^*$. In SVR, the goal is to minimize the function in equation (1). Here, C is the penalty parameter of the error term $\sum_{i=1}^{n} (\xi_i + \xi_i^*)$:

$$\frac{1}{2}\|w\|^2 + C * \sum_{i=1}^{n} (\xi_i + \xi_i^*). \quad (1)$$

SVR uses a linear hyperplane to separate the data and predict the values. However, in some cases, the distribution of data is nonlinear. Thus, kernel functions map the nonlinear separable feature space to linear separable feature space with kernel functions [16]. Equation (2) shows the

Radial Basis Function (RBF) kernel for the SVR model, where σ defines the standard deviation of the data. $\|x - y\|^2$ defines the squared Euclidean distance between feature vectors x and y:

$$K(x, y) = \exp\left(-\frac{1}{2 * \sigma}\|x - y\|^2\right). \tag{2}$$

3.2. Random Forest. In supervised learning, decision trees are commonly used as classification models to classify data with different features. Classification and Regression Trees (CART) [17] are usually used as algorithms to build the decision tree. However, using one single tree to classify or predict data might cause high variance. Thus, ensemble methods are proposed to construct a set of tree-based classifiers and combine these classifiers' decision with different weighting algorithms [18]. Random Forest (RF) algorithm is one of the ensemble methods that build several regression trees and average the result of the final prediction of each regression tree [19]. Algorithm 1 shows the procedure of the RF algorithm. Given the feature space and its corresponding labels, the RF algorithm takes a random sample from the features and constructs the CART tree with randomly selected features. During the procedure, N trees are built to generate the forest. During the test phase, test data are fed into the forest, and each CART tree in the forest predicts a value $\hat{y}_i$ based on the tree structure. The final output of the forest is the average vote of all the predicted values. During the training process, the number of trees and the trees' parameter are required to be determined to get the best parameter set for the RF model.

3.3. Boosting Approaches. Besides SVR and RF, boosting is also useful in supervised learning to reduce bias and variance of the model by constructing strong models from weak models step by step [20]. In each step, the model's weakness is obtained from the data pattern, and the weak model is then altered to fit the data pattern. In recent years, there has been a greater focus placed upon eXtreme Gradient Tree Boosting (XGBoost) models [21]. Friedman et al. proposed to use gradient descent in the boosting approach to minimize the loss function $\sum_{i=1}^{n} L(y_i, h_{(x_i)})$ [22] and refined the boosting model with regression trees in [23]. In their approach, the first-order Taylor expansion is used in the loss function to approximate the regression tree learning. In Chen and Guestrin's approach, they proposed to use a higher-order approximation to get a better regression tree structure [21]. The XGBoost algorithm works as Algorithm 2. The model is initialized with a function $f_0(x) = a_0$ which minimizes the loss function $\sum_{i=1}^{n} L(y_i, a_0)$. In each boosting step, the multipliers p_{ik} and q_{ik} are calculated as first-order Taylor expansion and higher-order Taylor expansion of loss function $L(y_i, f(x_i))$ to calculate the leaf weights which build the regression tree structure. Then the current model $f_k(x)$ is updated with the previous model $f_{k-1}(x)$ with the shrunk base model $\rho a_k h_k(x)$. At last, the weak models $f_k(x)$ are combined to generate the strong model $f(x)$. In XGBoost, the number of boosting iterations and the structure of regression trees affect the performance of the model. Thus, these parameters are tuned to with cross-validation to get the best XGBoost model.

3.4. Gaussian Process Regression. Gaussian process (GP) is a distribution over functions with a continuous domain, such as time and space [24]. In recent years, Gaussian process has been used in many areas such as image thresholding, spatial data interpolation, and simulation metamodeling. A GP $g(x)$ is usually parameterized by a mean function $\mu(x)$ and a covariance function $K(x, x')$, formalized in equations (3) and (4). Given a set of data points $\{x_1, x_2, \ldots, x_n\}$ associated with set of labels $\{y_1, y_2, \ldots, y_n\}$, each label y_i can be seen as a Gaussian noise model as in equation (5). Here, $\mathrm{GP}((\mu(x), K(x, x'))$ defines the stochastic map for each data point and its label and ε_i defines the measurement noise assumed to satisfy the Gaussian noise with σ standard deviation:

$$\mu(x) = \mathrm{E}[g(x)], \tag{3}$$

$$K(x, x') = \mathrm{E}[(g(x) - \mu(x))(g(x') - \mu(x'))], \tag{4}$$

$$\begin{aligned} y_i &= g(x_i) + \varepsilon_i, \text{ where } g(x_i) \\ &= \mathrm{GP}(\mu(x), K(x, x')), \ \{\varepsilon\} \sim \mathrm{N}(0, \sigma_n^2). \end{aligned} \tag{5}$$

Given the training data x with its corresponding labels y as well as the test data x^* with its corresponding labels y^* with the same distribution, then equation (6) is satisfied. Here, $K(x, x)$ is the covariance matrix based on training data points x, $K(x^*, x)$ is the covariance matrix between the test data points and training points, and $K(x^*, x^*)$ is the covariance matrix between test points. Then, the conditional probability of y^* can be formalized as equation (7):

$$\begin{bmatrix} y \\ y^* \end{bmatrix} / x, x* \sim \mathrm{N}\left(\begin{bmatrix} \mu(x) \\ \mu(x^*) \end{bmatrix}, \begin{bmatrix} K(x,x) + \sigma_n^2 I & K(x^*,x)^T \\ K(x^*,x) & K(x^*,x^*) \end{bmatrix}\right), \tag{6}$$

$$p(y^* | y, x, x^*) \sim N(\hat{\mu}, \hat{\nu}), \tag{7}$$

where

$$\begin{aligned} \hat{\mu} &= K(x, x^*)^T \left(K(x, x) + \sigma_n^2 I\right)^{-1} (y - \mu x) + \mu(x^*), \\ \hat{\nu} &= K(x^*, x^*) - K(x, x^*)^T \left(K(x, x) + \sigma_n^2 I\right)^{-1} K(x, x^*) + \sigma_n^2 I. \end{aligned} \tag{8}$$

Maximum likelihood estimation (MLE) has been used in statistical models, given the prior knowledge of the data distribution [25]. Thus, given the training data points x with label y, the estimated y^* of target x^* can be calculated by maximizing the joint likelihood $\log(p(y^* | y, x, x^*))$ in equation (7).

(1) Given: $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in X$, $y_i \in \mathrm{IR}$ the number of CART trees: N
(2) For $n = 1$ to N:
  (a) Construct a CART tree CART(x) with randomly selected data $x \in x_i$ with randomly selected features
  (b) Get the prediction $\widehat{y}$ of each CART tree and add the CART tree to the forest $F = F \cup \mathrm{CART}(x)$
(3) Predict the final result y from the forest: $\widehat{y} = \sum_{i=1}^{N} \widehat{y}_i / N$

Algorithm 1: Random Forest algorithm.

(1) Given: $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in X$, $y_i \in \mathrm{IR}$
  The number of iterations: N
(2) Initialize $f_0(x) = a_0 = \mathbf{argmin} \sum_{i=1}^{n} L(y_i, a_0)$
(3) For $k = 1$ to N:
  (a) Compute
    $p_{ik} = -[\partial L(y_i, f(x_i))/\partial f\ (x_i)]_{f(x)=f_{k-1}(x)}$, for $i = 1, \ldots, n$
    $q_{ik} = -[\partial^2 L(y_i, f(x_i))/\partial f\ (x_i)^2]_{f(x)=f_{k-1}(x)}$, for $i = 1, \ldots, n$
  (b) Fit base model $h_k(x)$
    $h_k(x) = \sum_{j=1}^{J_k} b_{jk} I(x \in R_{jk})$
  (c) Determine the leaf weight for the learnt structure with p_{ik} and q_{ik}
  (d) Update current model $f_k(x)$ with previous model $f_{k-1}(x)$ and the constrained $\rho a_k h_k(x)$
    $f_k(x) = f_{k-1}(x) + \rho a_k h_k(x)(x \in R_{jk})$
(4) Calculate the final boosting model $f(x) = \sum_{k=0}^{N} f_k(x)$

Algorithm 2: XGBoost algorithm, modified from [21].

In GPR, covariance functions are also essential for the performance of GPR models. This paper mainly evaluates three covariance functions, namely, Radial Basis Function (RBF) kernel, Matérn kernel, and Rational Quadratic kernel. The RBF kernel is a stationary kernel parameterized by a scale parameter l that defines the covariance function's length scale. Equation (2) shows the kernel function for the RBF kernel. The Matérn kernel adds parameter ν that controls the resulting function's smoothness, which is given in equation (9). Equation (10) shows the Rational Quadratic kernel, which can be seen as a mixture of RBF kernels with different length scales. In the equation, the α parameter controls the mixture of the length scales:

$$K\left(x_i, x_j\right) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\gamma\sqrt{2\nu}\left(\frac{x_i}{l}, \frac{x_j}{l}\right)\right)^{\nu} \cdot K_{\nu}\left(\gamma\sqrt{2\nu}\left(\frac{x_i}{l}, \frac{x_j}{l}\right)\right), \tag{9}$$

$$K\left(x_i, x_j\right) = \left(1 + \frac{\left(x_i, x_j\right)^2}{2\alpha l^2}\right)^{-\alpha}. \tag{10}$$

4. Experiment with Offline Training

In this paper, we use the RSS-based modeling technique that explores the relationship between the specific location and its corresponding RSS. Figure 1 shows the procedure that builds indoor positioning model by comparing the performance of different machine learning models. In the offline phase, RSS data from several APs are collected as the training data set. There are two procedures to train the offline RSS-based model. In the first step, cross-validation (CV) is used to test whether the model is suitable for the given machine learning model. The CV can be used for feature selection and hyperparameter tuning. By using the 5-fold CV, the training data is split into fivefold. During the training process, the model is trained with the four folds of data and test with the left fold of data. The training procedure is repeated five times to calculate the average accuracy of the model with the specific parameter. After we get the model with the optimum parameter set, the second step of the offline phase trains the model with the RSS data. Then, we got the final model that maps the RSS to its corresponding position in the building. Later in the online phase, we can use the generated model for indoor positioning.

4.1. Data Collection. To construct the fingerprinting database and evaluate the machine learning models, we collect RSS data in an indoor environment whose floor plan is shown in Figure 2. In the building, we place 7 APs represented as red pentagram on the floor with an area of 21.6 M × 15.6 m. The RSS measurements are taken at each point in a grid of 0.6 m spacing between each other. The RSS data are measured in dBm, which has typical negative values ranging between 0 dBm and −110 dBm. The RSS data of seven APs are taken as seven features. The output is the

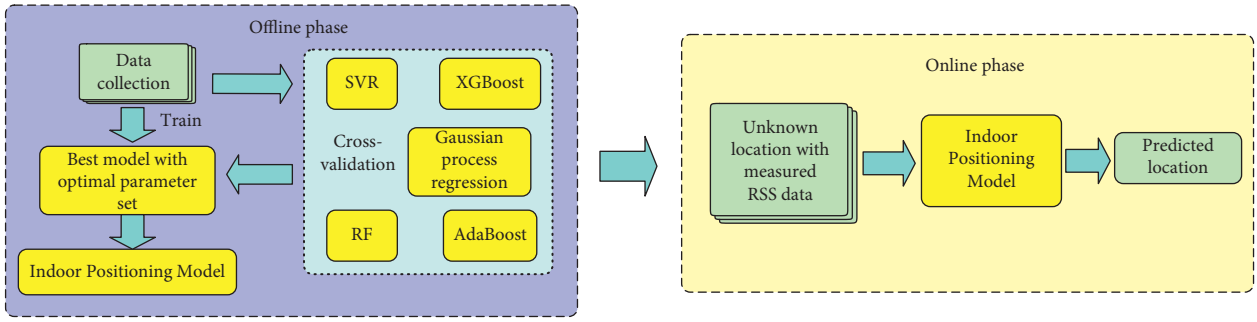


Figure 1: Indoor positioning modeling procedure with offline phase and online phase.

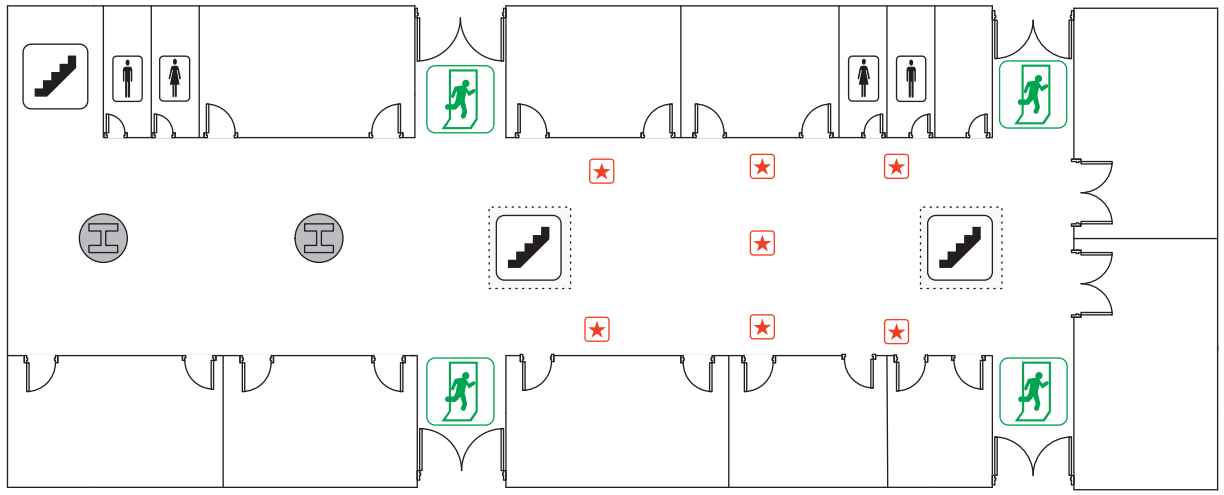


Figure 2: Indoor floor plan with access points marked by red pentagram.

coordinates of the location on the two-dimensional floor. We write Android applications to collect RSS data at reference points within the test area marked by the seven APs, whereas the RSS comes from the Nighthawk R7000P commercial router. During the field test, we collect 799 RSS data as the training set. The training set's size could be adjusted accordingly based on the model performance, which would be discussed in the following section. The 200 RSS data are collected during the day with people moving or environment changes, which are used to evaluate the model performance.

4.2. Hyperparameter Tuning. As is shown in Section 2, the machine learning models require hyperparameter tuning to get the best model that fits the data. Table 1 shows the parameters requiring tuning for each machine learning model. Tuning is a process that uses a performance matrix to rank the regressors with different parameters to optimize a parameter for each specific model [11]. In this paper, we use the distance error as the performance matrix to tune the parameters. Given the predicted coordinates of the location as $(\hat{x}, \hat{y})$ and the true coordinates of the location as (x, y), the Euclidean distance error is calculated as follows:

$$d = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2}. \quad (11)$$

Underfitting and overfitting often affect model performance. In this paper, we use the validation curve with 5-fold cross-validation to show the balanced trade-off between the bias and variance of the model. In the validation curve, the training score is higher than the validation score as the model will be a better fit to the training data than test data. The increasing of the validation scores indicates that the model is underfitting. When the validation score decreases, the model is overfitting. Thus, validation curves can be used to select the best parameter of a model from a range of values. Results show that nonlinear models have better prediction accuracy compared with linear models, which is evident as the distribution of RSS over distance is not linear. Table 1 shows the optimal parameter settings for each model, which we use to train different models.

Figure 3 shows the tuning process that determines the optimum value for the penalty parameter C and kernel coefficient parameter γ for the SVR with RBF and linear kernels. Results show that RBF has better prediction accuracy compared with linear kernels in SVR. It is evident, as the distribution of RSS over distance is not linear. Thus, linear models cannot describe the model correctly. Moreover, the selection of coefficient parameter γ of the SVR with RBF kernel is critical to the performance of the model. The validation curve shows that when γ is 0.01, the SVR has the best performance in predicting the position.

Table 1: Hyperparameter tuning for different machine learning models.

Machine learning model	Parameter	Optimal setting	Distance error
SVR with linear kernel	Penalty C Kernel coefficient	2 0.01	4.58
SVR with RBF kernel	Penalty C Kernel coefficient	2 0.01	1.75
Random Forest	Number of trees Max depth of individual tree Min samples split of individual tree Min samples leaf of individual tree	400 10 8 4	0.93
XGBoost	Number of boosting iterations Learning rate Max depth of individual tree Subsample	500 0.25 15 0.9	0.85

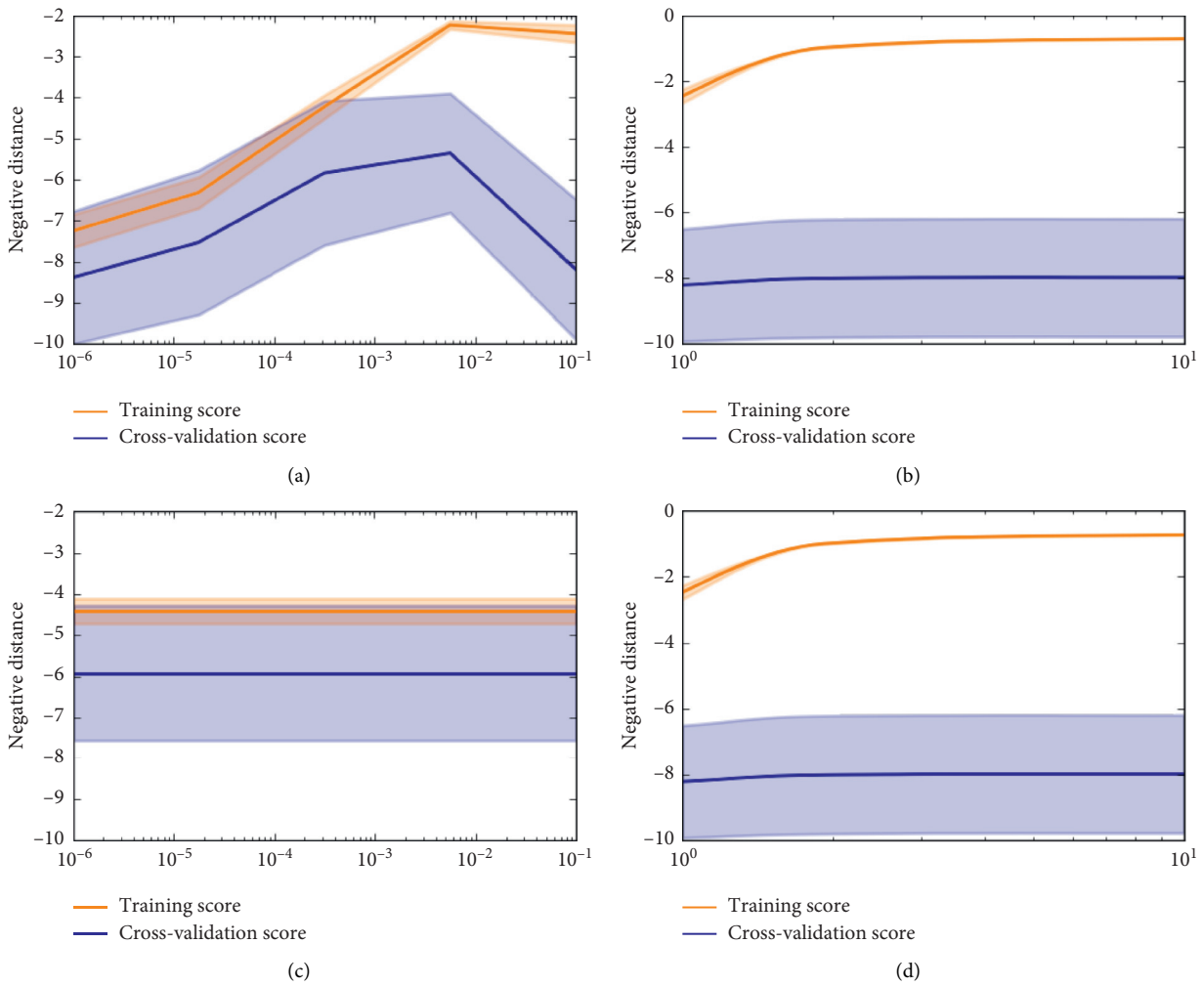


Figure 3: Hyperparameter tuning for SVR with linear and RBF kernel. (a) γ of RBF kernel. (b) C of RBF kernel. (c) γ of linear kernel. (d) C of linear kernel.

Figure 4 shows the tuning process that calculates the optimum value for the number of trees in the random forest as well as the tree structure of the individual tree in the forest. The validation curve shows that the maximum depth of the tree might affect the performance of the RF model. When the maximum depth of the individual tree reaches 10, the model comes to the best performance. The number of boosting iterations and other parameters concerning the tree structure do not affect the prediction accuracy a lot.

Figure 5 shows the tuning process that calculates the optimum value for the number of boosting iterations and the learning rate for the AdaBoost model. Results show that a

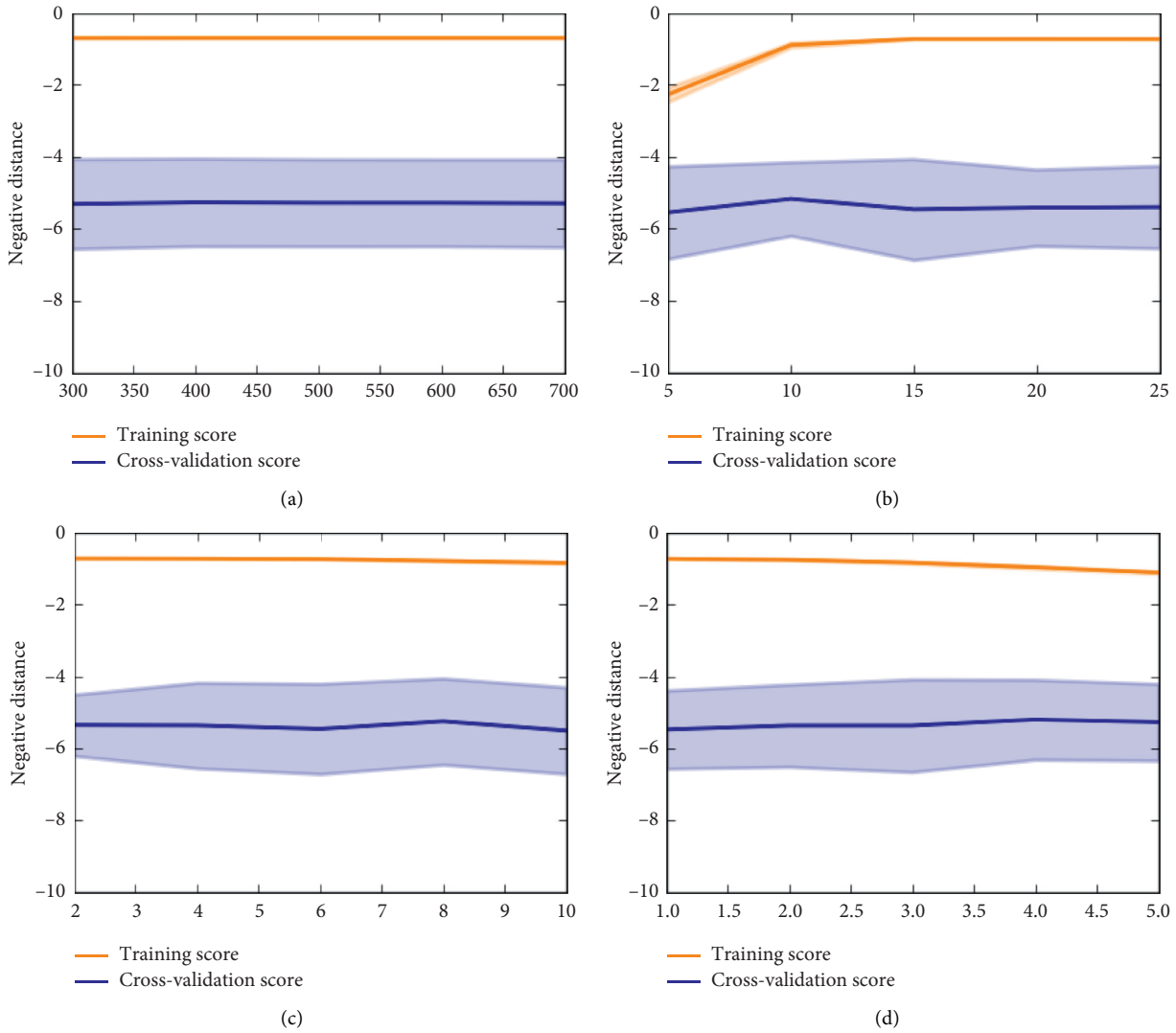


Figure 4: Hyperparameter tuning for Random Forest model. (a) Number of estimators. (b) Max depth. (c) Min samples split. (d) Min samples leaf.

higher learning rate would lead to better model performance. While the number of iterations has little impact on prediction accuracy, 300 could be used as the number of boosting iterations to train the model to reduce the training time.

Figure 6 shows the tuning process that calculates the optimum value for the number of boosting iterations, the learning rate, and the individual tree structure for the XGBoost model. To avoid overfitting, we also tune the *subsample* parameter that controls the ratio of training data before growing trees. The validation curve suggests that a higher learning rate and the number of boosting iterations could have a better model performance. The individual tree structure and the ratio of training data have less impact on prediction accuracy.

4.3. Kernel Selection for Gaussian Process Regression. Besides the typical machine learning models, we also analyze the GPR with different kernels for the indoor positioning problem. Table 2 shows the distance error with a confidence interval for different kernels with length scale bounds. In statistics, 1.96 is used in the constructing of 95% confidence intervals [26]. We calculate the confidence interval by multiplying the standard deviation with 1.96. Overall, the three kernels have similar distance errors. However, the confidence interval has a huge difference between the three kernels. The RBF and Matérn kernel have the 4.4 m and 8.74 m confidence interval with 95% accuracy while the Rational Quadratic kernel has the 0.72 m confidence interval with 95% accuracy. Rational Quadratic kernel is the most stable model for the GPR algorithm. Thus, we select this as the kernel of the GPR model to compare with other machine learning models.

5. Model Evaluation and Experiment Results

In the previous section, we train the machine learning models with the 799 RSS samples. In this section, we evaluate

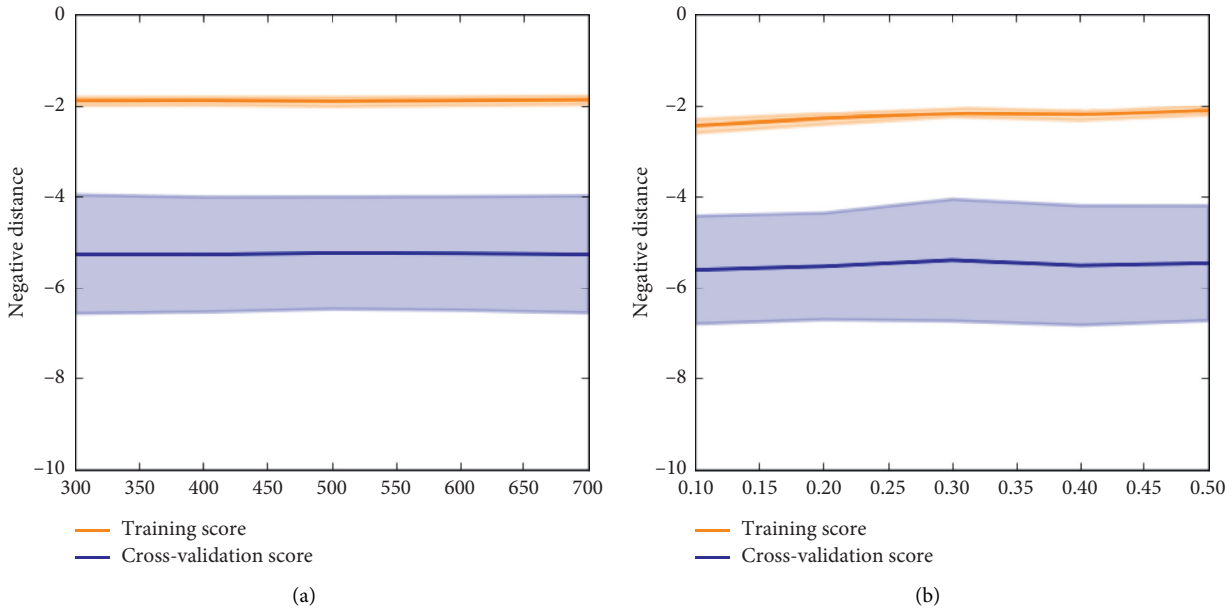


Figure 5: Hyperparameter tuning for AdaBoost model. (a) Number of estimators. (b) Learning rate.

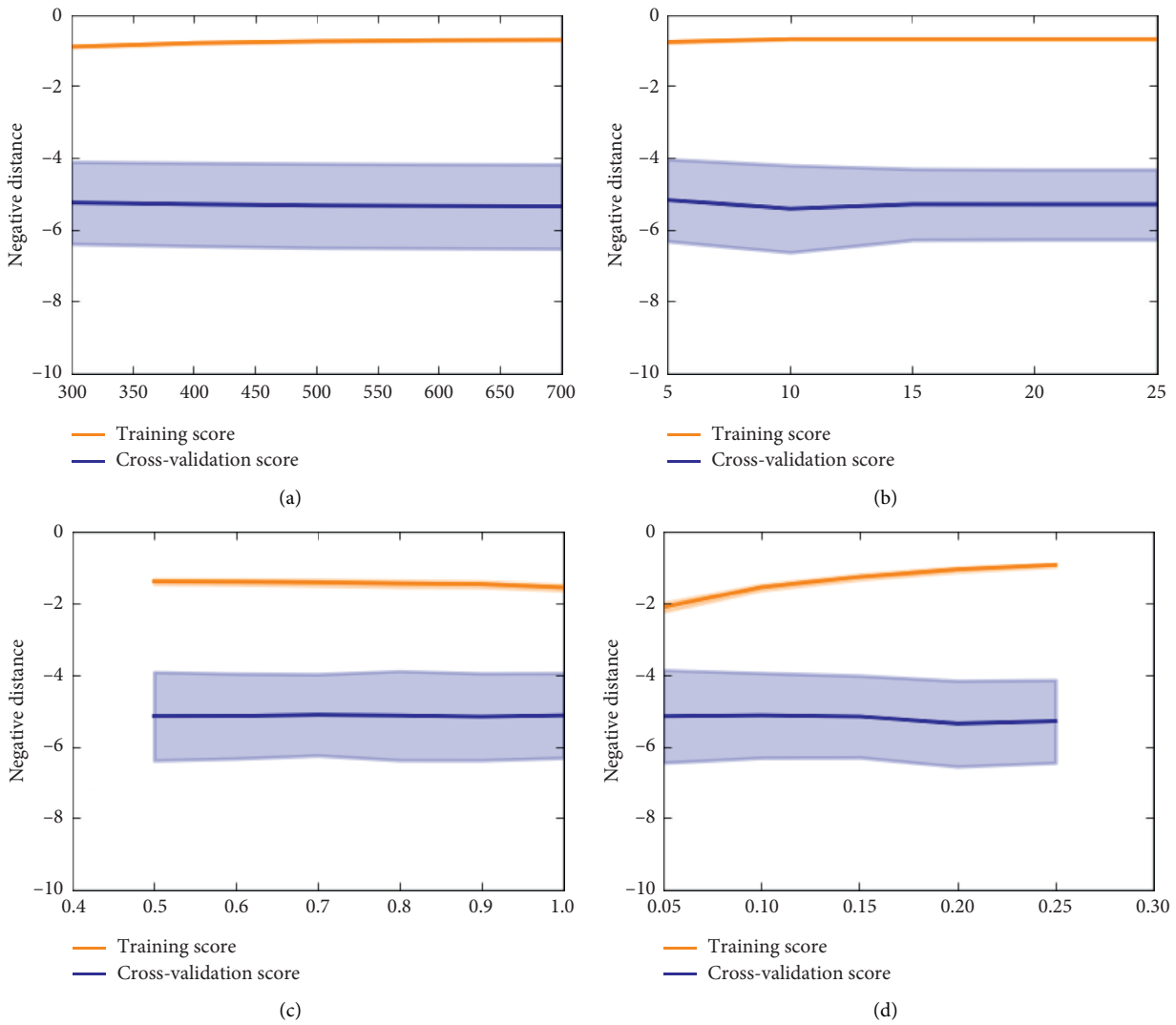


Figure 6: Hyperparameter tuning for XGBoost model. (a) Number of estimators. (b) Max depth. (c) Subsample. (d) Learning rate.

Table 2: Distance error with confidence interval for different Gaussian progress regression kernels.

Kernel	Length scale bounds	Distance error	Confidence interval with 95% probability
RBF	0.1–10	0.865	4.4
Rational Quadratic	0.1–10	0.864	0.72
Matérn	1	0.865	8.74
Hline			

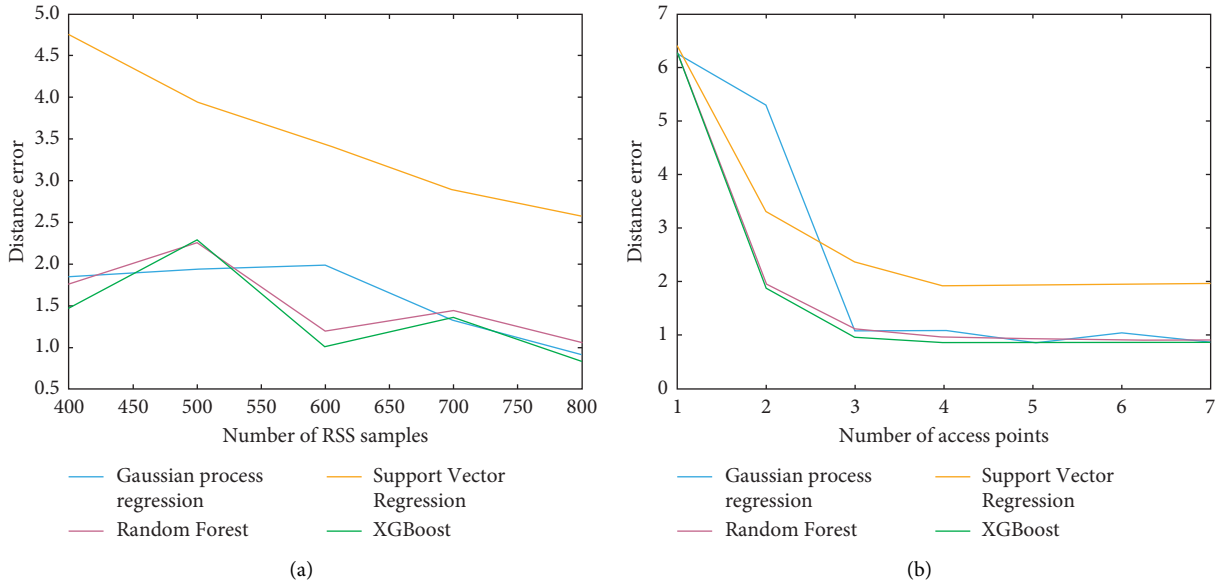


Figure 7: Features that affect model performance of indoor positioning. (a) Impact of the number of RSS samples. (b) Impact of the number of access points.

the result by evaluating the performance of the models with 200 collected RSS samples with location coordinates. Tables 1 and 2 show the distance error of different machine learning models. Overall, the GPR with Rational Quadratic kernel has the lowest distance error among all the GP models, and XGBoost has the lowest distance error compared with other machine learning models. Specifically, XGBoost model achieves a 0.85 m error, which is better than the RF model. XGBoost also outperforms the SVR with RBF kernel. However, the XGBoost and the GPR with Rational Quadratic have similar performance concerning the distance error. In this section, we evaluate the impact of the size of training samples and the number of APs to get the model with high indoor positioning accuracy but requires fewer resources such as training samples and the number of APs.

5.1. Size of Training Samples. Figure 7(a) shows the impact of the training sample size on different machine learning models. During the training process, we restrict the training size from 400 to 799 and evaluate the distance error of different trained machine learning models. Results reveal that there has been a gradual decrease in distance error with the increasing of the training size for all machine learning models. In all stages, XGBoost has the lowest distance error compared with all the other models. The RF model has a similar performance with a slightly higher distance error. With the increase of the training size, GPR gets the better performance, while its performance is still slightly weaker compared with the XGBoost model.

5.2. Number of Access Points. Figure 7(b) reveals the impact of the size of APs on different machine learning models. In the training process, we use the RSS collected from different APs as features to train the model. The size of the APs determines the size of the features. Results show that the distance error decreases gradually for the SVR model. The graph also shows that there has been a sharp drop in the distance error in the first three APs for XGBoost, RF, and GPR models. Then the distance error of the three models comes to a steady stage. This trend indicates that only three APs are required to determine the indoor position. More APs are not helpful as the indoor positioning accuracy is not improving with more APs. Overall, XGBoost still has the best performance among RF and GPR models.

6. Conclusion and Future Work

In this paper, we evaluate different machine learning approaches for indoor positioning with RSS data. The models include SVR, RF, XGBoost, and GPR with three different kernels. Hyperparameter tuning is used to select the optimum parameter set for each model. Then the performance of different models is evaluated using the Euclidean distance error between the predicted coordinates and real

coordinates. Results show that XGBoost has the best performance compared with all the other machine learning models. Also, 600 is enough for the RSS training size as the distance error does not change dramatically after the training size reaches 600. Results also reveal that 3 APs are enough for indoor positioning as the distance error does not decrease with more APs.

Indoor position estimation is usually challenging for robots with only built-in sensors. Accumulated errors could be introduced into the localization process when the robot moves around. However, based on our proposed XGBoost model with RSS signals, the robot can predict the exact position without the accumulated error. Thus, more work can be done to decrease the positioning error by using the extended Kalman filter localization algorithm to fuse the built-in sensor data and the RSS data.

Data Availability

The data are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the work.

References

[1] A. Serra, D. Carboni, and V. Marotto, "Indoor pedestrian navigation system using a modern smartphone," in *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, pp. 397-398, ACM, San Francsico, CA, USA, 2010.

[2] P. Bahl, V. N. Padmanabhan, V. Bahl, and V. Padmanabhan, "Radar: an in-building rf-based user location and tracking system," in *Proceedings IEEE INFOCOM 2000*, Tel Aviv, Israel, 2000.

[3] A. Harter and A. Hopper, "A distributed location system for the active office," *IEEE Network*, vol. 8, no. 1, pp. 62–70, 1994.

[4] H. Hashemi, "The indoor radio propagation channel," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 943–968, 1993.

[5] A. Schwaighofer, M. Grigoras, V. Tresp, and C. Hoffmann, "Gpps: a Gaussian process positioning system for cellular networks," *Advances in Neural Information Processing Systems*, pp. 579–586, 2004.

[6] Z. L. Wu, C. H. Li, J. K. Y. Ng, and K. R. Leung, "Location estimation via support vector regression," *IEEE Transactions on Mobile Computing*, vol. 6, no. 3, pp. 311–321, 2007.

[7] A. Bekkali, T. Masuo, T. Tominaga, N. Nakamoto, and H. Ban, "Gaussian processes for learning-based indoor localization," in *Proceedings of the 2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 1–6, IEEE, Xi'an, China, 2011.

[8] M. Brunato and C. Kiss Kallo, "Transparent location fingerprinting for wireless services," Technical Report, University of Trento, Trento, Italy, 2002.

[9] R. Battiti, A. Villani, and T. Le Nhat, "Neural network models for intelligent networks: deriving the location from signal patterns," in *Proceedings of AINS*, Kennesaw, GA, USA, 2000.

[10] M. Alfakih, M. Keche, and H. Benoudnine, "Gaussian mixture modeling for indoor positioning wifi systems," in *Proceedings of the 2015 3rd International Conference on Control*, pp. 1–5, IEEE, Tlemcen, Algeria, 2015.

[11] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, and M. Tu, "Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances," *Journal of Petroleum Science and Engineering*, vol. 160, pp. 182–193, 2018.

[12] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[13] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, pp. 155–161, 1997.

[16] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and testing low-degree polynomial data mappings via linear svm," *Journal of Machine Learning Research*, vol. 11, pp. 1471–1490, 2010.

[17] L. Breiman, *Classification and Regression Trees*, Routledge, Abingdon, UK, 2017.

[18] T. G. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the International workshop on multiple classifier systems*, pp. 1–15, Springer, Cagliari, Italy, 2000.

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

[20] R. E. Schapire, "The boosting approach to machine learning: an overview," in *Nonlinear Estimation and Classification*, pp. 149–171, Springer, Berlin, Germany, 2003.

[21] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, Newyork, NY, USA, 2016.

[22] J. H. Friedman, "machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[23] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[24] C. E. Rasmussen, *Gaussian Processes in Machine Learning in Summer School on Machine Learning*, pp. 63–71, Springer, Berlin, Germany, 2003.

[25] R. J. Rossi, *Mathematical Statistics: An Introduction to Likelihood Based Inference*, John Wiley & Sons, Hoboken, NJ, USA, 2018.

[26] L. J. Savage, *The Foundations of Statistics*, Courier Corporation, North Chelmsford, MA, USA, 1972.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4518624, 15 pages
https://doi.org/10.1155/2020/4518624

Research Article

Research on the Simulation of Wheelset Response Characteristic Identification of Railway Fastener Loosening

Wenbai Zhang,[1] **Lele Peng**,[1] **Shubin Zheng**,[1] **Xun Guo**,[1] **and Yuling Wang**[2]

[1]*School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai 201620, China*
[2]*R&D Center, Shanghai Aerospace Equipment Manufacturing Co., Ltd., Shanghai 200245, China*

Correspondence should be addressed to Wenbai Zhang; zhangwenbaipost@163.com and Lele Peng; 13661773128@139.com

Received 7 June 2020; Accepted 8 September 2020; Published 19 September 2020

Academic Editor: William Guo

Rail fastener is a crucial component equipment to ensure the safe operation of the train, and it is very paramount to detect the loose state of the fastener. In this paper, the vertical vibration acceleration signal of wheelset is taken as the research object, and the loose state of fastener is identified by separating and calculating the key IMF energy entropy. Firstly, based on the finite element theory and the principle of multibody dynamics, the rigid-flexible coupling simulation model of vehicle track is established. Then, the vertical vibration acceleration signals of the wheelset under the speed of 200 km/h are obtained by setting the different loosening degrees of the fastener. Finally, we use optimized HHT to process signals, and the orthogonal empirical mode decomposition method (OEMD) is proposed to optimize the orthogonality of the intrinsic mode function, to eliminate the IMF component having poor correlation with the original signal; the Hilbert time spectrum and information entropy theory are combined to calculate the energy entropy of the key IMF, and the HHT energy entropy evaluation algorithm of the vertical acceleration response signal of the train wheelset is proposed. The simulation results show that the HHT energy entropy of 100% fastener looseness is less than 25%, 50%, and 75%, decreasing trend. The algorithm can recognize the looseness of track fastener through the experiment under different working conditions.

1. Introduction

At present, the methods of detecting fasteners for railway lines in China mainly include manual inspection, track inspection car, and computer-aided visual inspection. When the vehicle is running on the rail, the interaction between wheel and rail will happen, and when the fastener is damaged or missing, the dynamic parameters of the structure will change to some extent [1]. These changes will be responded to by the vibration signals of some vehicle components in some forms [2, 3].

Liu et al. [4] used ANSYS/ LS-DYNA simulation software to establish the vertical coupling vibration model of vehicle-ballastless track-subgrade system, which verified that the sudden change of fastener stiffness had tiny effect on the vertical acceleration of bogie and had obvious effect on the wheelset vibration acceleration; Zhao and Tan [5] established a simulation model of metro vehicle-track flexible body and compared and analyzed the maximum dynamic responses of vehicle-track system under different conditions of fastener failure and different speeds. The research shows that fastener failure has a certain impact on the vibration response of track; Huang [6] analyzed the orthogonality of each intrinsic mode function by numerical simulation and applied the improved HHT method to the damage identification of large structure system; Zhang [7] combined HHT with various theories and applied it to identify rail vibration signals with different fastener looseness under moving load impact. The results show that the method can reflect the change of rail fastener looseness to a certain extent.

To sum up, there is no way to judge the loose state of fastener by separating the vibration acceleration of wheelset [8, 9]. This paper presents a method to identify whether the fastener is loose from the response signal of the wheelset before and after the change of the fastener state. It is of great practical significance to identify the fastener.

2. Vehicle-Track Rigid-Flexible Coupling Modeling and Dynamic Simulation

2.1. Flexible Track Modeling. In order to obtain the wheelset vibration signal, an accurate vehicle-track rigid-flexible coupling model needs to be established [10]. At first, the track system is considered as a flexible body on the basis of the existing rigid vehicle-track model [11, 12]. The slab ballastless track system is shown in Figure 1. The finite element model of rail and rail slab is established in ANSYS. In the multibody dynamics software SIMPACK, force element is used to simulate CA mortar and fastener and assembled into the whole ballastless track system.

2.1.1. Track System Dynamics Model Based on ANSYS. When the rail is modeled, the rail is regarded as a continuous elastic beam. China's standard 60 kg/section rail parameters are used for modeling and analysis. Because the number of marker points in SIMPACK is limited and if there are too many master DOF (degrees of freedom), it will lead to the incomputable situation in the SIMPACK; therefore, in order to ensure the calculation efficiency, and combined with the purpose of this paper, a 120-meter rail model in ANSYS is established. The distance between two fasteners is 0.6 m, and this length is divided into 20 small units. A total of 201 master DOF nodes are selected. The discrete model of the rail is shown in Figure 2(a). The solid model of the rail slab is shown in Figure 2(b).

2.1.2. Dynamics Analysis on Rail Subsystem. In this paper, the substructure modal analysis of rail and rail slab needs to reduce the main degree of freedom of the model, and it is reduced by Guyan reduction method in ANSYS. The substructure model of rail and rail slab in this paper is shown in Figure 3.

The first 100 modes of the rail and the top 20 modes of the rail slab are calculated, respectively. When selecting the model of rail and rail slab, it is indispensable to accurately reflect the vibration characteristics of rail and rail slab and consider the weight of each model [13]. When the train passes, the vertical vibration is the main vibration of rail, and Figure 4 shows the main mode shapes of the rail and rail slab.

2.2. Vehicle-Track Rigid-Flexible Coupling Model

2.2.1. FEMBS (Finite Element Multibody System) Interface Program. The basic process flow of flexible data transmission between ANSYS and SIMPACK is shown in Figure 5. After modal analysis of rail and rail slab, ANSYS generates geometric model file (.CDB), mass stiffness matrix file (.Sub), and modal mode file (.RST). The interface of FEMBS (finite element multibody system) converts the flexible body data to the standard code (SID) in the format of ASCII readable by SIMPACK, that is, inputting the characteristics of flexible body of rail and rail slab into motion equation to generate SID file. The input information for FEMBS includes DOF and coordinates of nodes, mass attribute, translational and rotational vibration, modal mass matrix, stiffness matrix, damping matrix, and geometric stiffening matrix for initial loads.

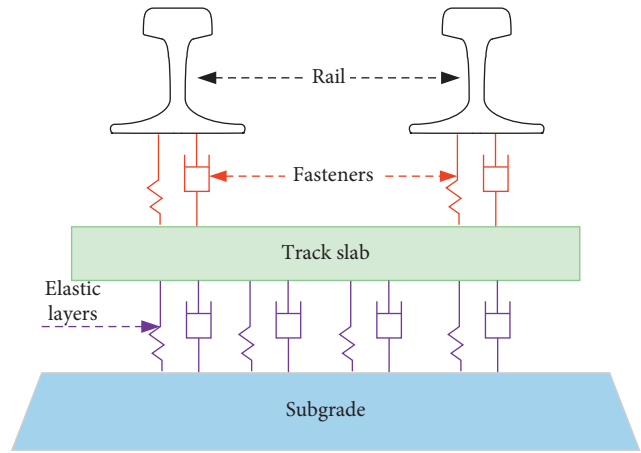
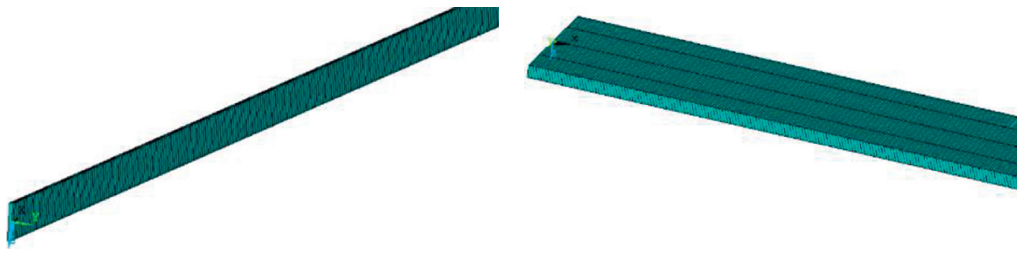


Figure 1: Slab ballastless track system, which consists of rail (black), fasteners (red), rail slab (green), elastic layer (purple), and concrete subgrade (blue).

2.2.2. Realization of Rigid-Flexible Coupling. When assembling the track system in SIMPACK, the fastener and the CA mortar are modeled by linear force element. The flexible rails and the rigid wheels are connected with Hertz spring; however, the contact conditions of wheel and rail in SIMPACK must be rigid wheelset and rigid track and longitudinal relative displacement cannot occur between wheel and rail, and the track model is flexible; the rail does not move with the wheel, so to define a virtual rail between each wheel and flexible rail and define a moving marker point allow the wheel rail force to be transmitted downward to the flexible rail system.

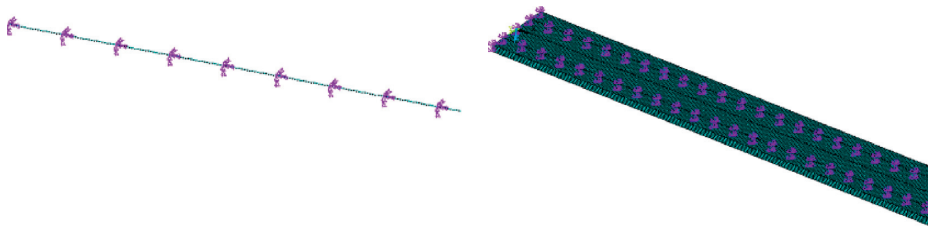
The virtual rail body is an object with zero mass and zero-moment of inertia, which only acts as a connection in the dynamic system, and does not affect the other connection structures, but the mass and the moment of inertia of the object in SIMPACK cannot be zero, so the relevant parameters of the virtual rail body are fetched as small as possible, which is assumed to be $1.0e-6$ in this paper, whose impact on wheel rail force transmission is negligible.

In order to ensure that each virtual rail body moves longitudinally along the rail with its corresponding wheelset, the moving marker points on the virtual rail body and the flexible rail are required to define the restraint to hold them, and the wheelset is assembled to the virtual rail body through the hinges, so that the virtual rail body is equivalent to a part of the rail and the vehicle can run along the rail. The wheel-rail rigid-flexible coupling model is shown in Figure 6(a). In simulation, the wheel rail force is calculated through Hertz contact between the virtual rail body and rigid wheel. Data exchange between the virtual rail body and the flexible rail is done through the deformation coordination condition and the force balance condition [14, 15]. According to the above method, the vehicle and the flexible track are assembled



(a) (b)

Figure 2: Finite element model of rail and rail slab: (a) model of rail, the distance between two fasteners is 0.6 m, and this length is divided into 20 small units; the total of main DOF nodes is 201. (b) Model of rail slab, continuous elastomer, having an elastic modulus of 3.6*e*10 Pa, a Poisson's ratio of 0.1, and a density of 2500 kg/m^3, is the solid-45 eight-node-space entity unit. (a) Rail. (b) Rail slab.



(a) (b)

Figure 3: Substructure model of rail and rail slab: (a) rail substructure model. (b) Rail board substructure model; the purple triangle mark in the figure is the location of the selected master DOF node; the total number of master DOF should be greater than the order of the subsequent modal analysis, or equal to twice. The predicted deformation direction of rail and rail slab is selected as the master DOF. The location of the load and constraint shall be selected as the main degree of freedom, master DOF is evenly distributed as far as possible, and the nodes at both ends of rail and rail slab and the nodes at fastener positions are selected as the main nodes of degrees of freedom. (a) Rail. (b) Rail slab.

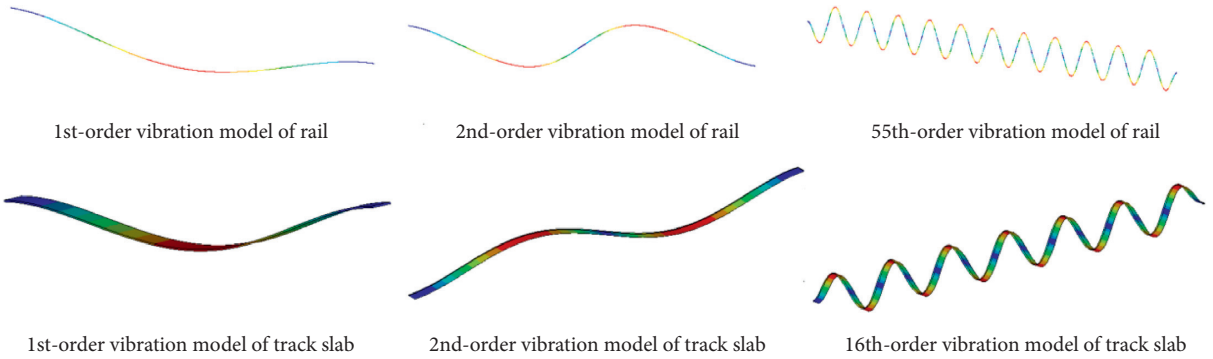


Figure 4: Main vibration model of rail and rail slab; vibration model of rail is shown in the 1st, 2nd, and 55th order; vibration model of rail slab is shown in the 1st, 2nd, and 16th order; the order is determined by the length of the model.

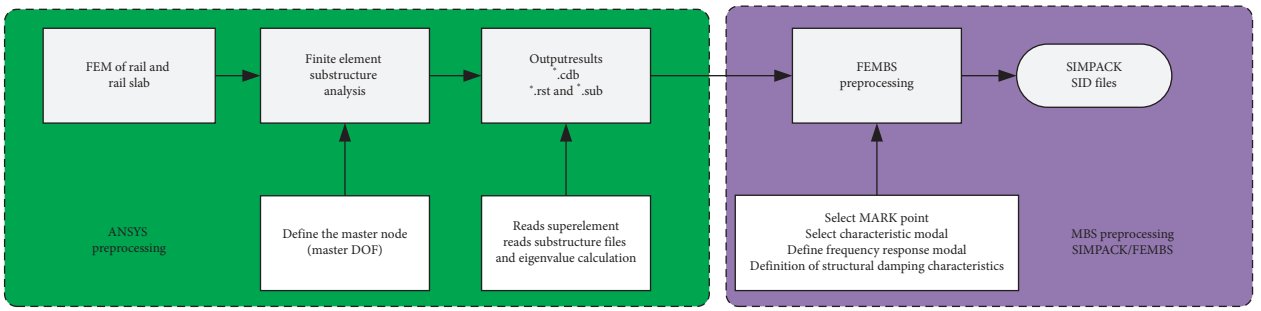


Figure 5: ANSYS and SIMPACK data transmission flow chart; the left is the flow of ANSYS preprocessing and the right is the flow of SIMPACK/ FEMBS preprocessing.

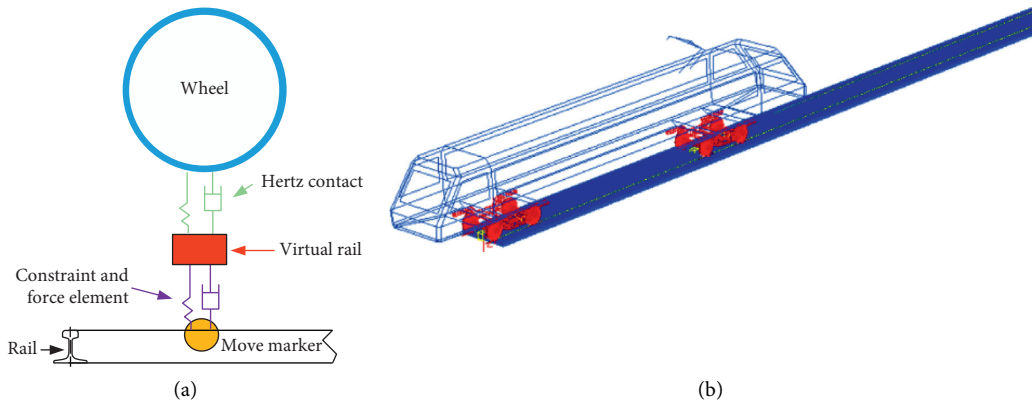


FIGURE 6: Coupling mode: (a) wheel (blue), hertz contract (green), virtual rail (red), constraint and force element (purple), move marker (yellow). (b) There is a motor vehicle with two bogies in the coupling model. (a) Wheel-rail coupling model. (b) Vehicle-track rigid-flexible coupling model.

together in the SIMPACK. The established vehicle-track rigid-flexible coupling model is shown in Figure 6(b).

2.2.3. Model Verification When Fasteners Are Not Loose. Before simulation, firstly we should verify whether the initial state of vehicle-track rigid-flexible coupling model is balanced, which is similar to the verification of rigid body vehicle model. In SIMPACK, we verify the model by calculating the nominal force maximum acceleration value of initial state. The maximum acceleration is calculated to be $6.811 * 10^{-6}\,\mathrm{m/s^2}$, which is less than the evaluation order of magnitude $10^{-4}\,\mathrm{m/s^2}$, so it can be determined that the built model is balanced in the initial state and the coupling model is correct.

In [16], the actual measured values of the vehicle vibration acceleration when the high-speed vehicle runs through under different conditions of track irregularity are given, and the simulation calculation is carried out with the VICT simulation software under the same conditions. In this paper, the same test conditions are set in the model to carry out the simulation calculation of vehicle dynamics, and the simulation results are compared with the above results. See Table 1 for comparison. We can see that the simulation results of this paper are not much different from the test results and the VICT simulation results in [17]. The reason for the slight difference is because the vehicle simulation parameters of this paper are different from the other article, and the vertical correctness of the simulation model is verified, which lays a foundation for the subsequent simulation.

2.3. Vehicle Dynamics Simulation

2.3.1. SIMPACK Simulation Excitation Model. SIMPACK can generate excitation by means of inputting the coefficient of the track spectral density formula and the method of converting the measured data. The software can be set up for two types of excitations: one is track-related; the other is rail-related. Due to the fact that there is no unified track spectrum standard in our country, the low-interference track spectrum of German high-speed railway is adopted in this paper. The vertical track irregularities generated in SIMPACK are shown in Figure 7.

2.3.2. Simulation Conditions. When two fasteners are loosened on the same section of the same track, and two fasteners are loosened continuously, the impact on the vehicle and track system is relatively large [18, 19]. Therefore, three typical simulation conditions are set up for the looseness of the fastener, as shown in the diagram. The cross line is the fastener loosening position in the picture, and there is one loose fastener on the first condition, there are two loose fasteners on the same section of the track in the second condition, and there are two loose fasteners on the same rail in the third condition.

Selecting the fastener classic stiffness of 50 kN/mm, the vehicle passes through the fastener in different degrees of looseness at 200 km/h speed (not loose, 25% loose, 50% loose, 75% loose, and completely loose), the looseness of the fastener is simulated by changing the stiffness value of the fastener force element in the SIMPACK, the corresponding stiffness value of the fastener loosening degree, and the equivalent stiffness value of the mortar unit length as shown in Table 2, and the looseness degree of fastener is the same in the three working conditions, shown in Figure 8. The vertical acceleration of wheelset and vehicle body is calculated by simulation.

2.3.3. Simulation Results of Wheelset Vertical Acceleration

(1) Wheelset vertical acceleration in condition 1

The vehicle response is the same before the vehicle arrives at the loosened part of the fastener [20, 21], so only the vertical acceleration response of the 0.6 s~1.7 s is given. The vertical acceleration comparison diagram of the wheelset in different degrees of fastener loosening is given, as shown in Figure 9; the acceleration at about 0.72 s~0.76 s starts changing; with the increase of the degree of fastener looseness, the vibration signal also changes

Table 1: The comparison of max value from experiment [16], VICT simulation [17], and current coupled model.

Test conditions	Longitudinal irregularity	Wavelength λ (m)	10	12	12	24	24
		Wave depth a (mm)	10	9	9	16	20
		Test speed v (km/h)	160	135	150	160	160
Max measured value in the reference (g)			0.12	0.06	0.08	0.12	0.13
Max value of the VICT simulation (g)			0.104	0.078	0.085	0.096	0.120
Value of simulation in the current model (g)			0.108	0.067	0.079	0.110	0.111

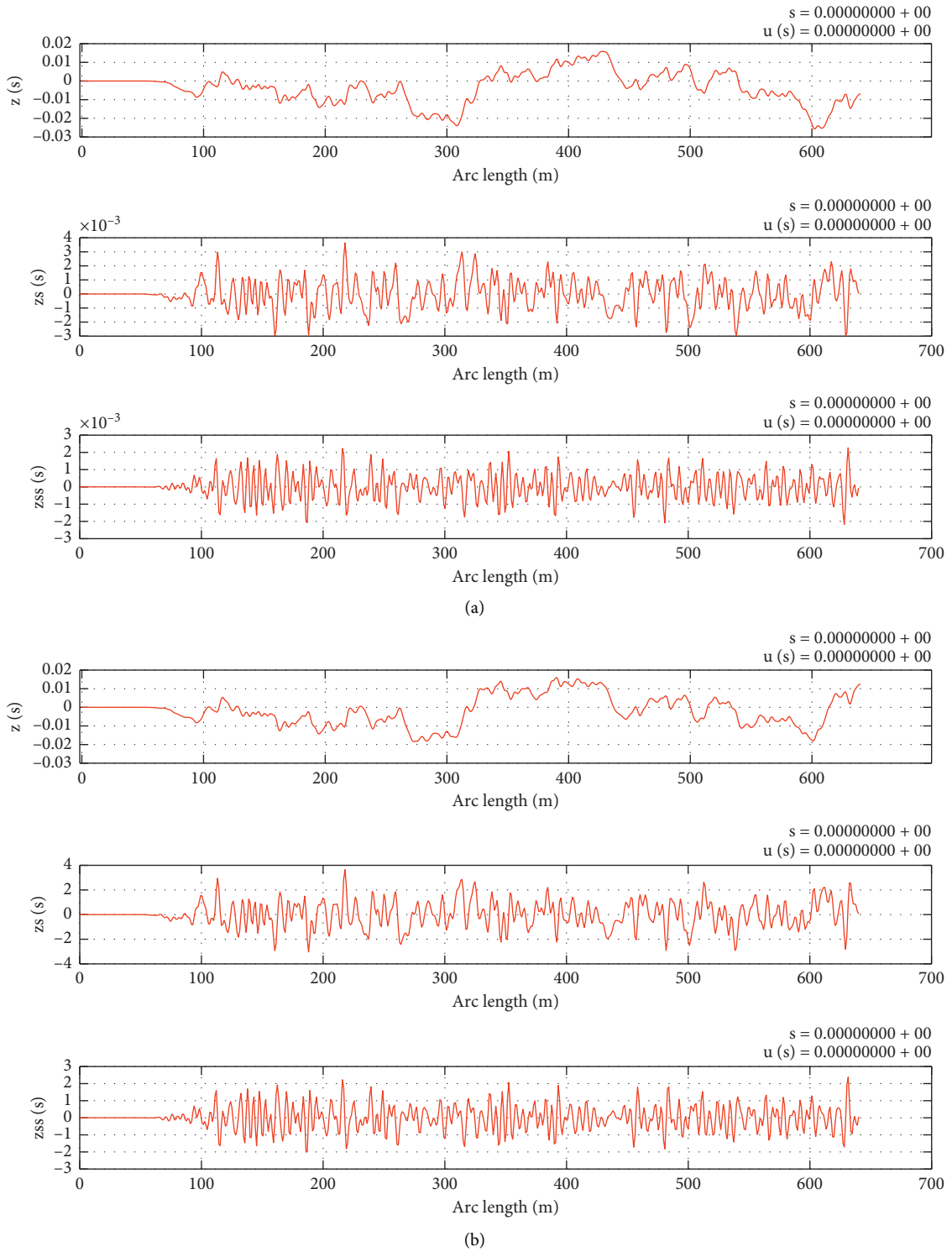


Figure 7: Track vertical irregularity. In (a) and (b), the first figure is the vertical irregularity imposed on the track, the second figure is the irregularity under the speed excitation, and the third figure is the irregularity under the acceleration excitation. (a) Left track irregularity. (b) Right track irregularity.

Table 2: The corresponding stiffness value of the loosening degree of the fastener and the equivalent stiffness value of the length of the mortar unit.

Degrees of fastener stiffness looseness (%)	Vertical, horizontal, and longitudinal stiffness value of the fastener (kN/mm)	Vertical, horizontal, and longitudinal damping value of the fastener (kN·s/mm)	Equivalent stiffness of CA mortar unit length (N/m^3)	Equivalent damping of CA mortar unit length (N·s/m)
0	50			
25	37.5			
50	25	75/60/60	$1.25 * 10^9$	$3.46 * 10^4$
75	12.5			
100	0			

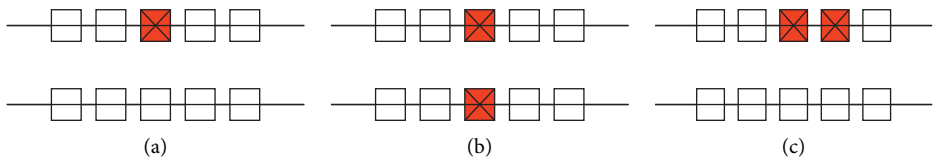


Figure 8: Loose fastener condition. The red square and cross is the position of loose fastener: one loose fastener in condition 1, two opposite in condition 2, and two consecutive in condition 3. The looseness of fastener could be set separately. (a) Condition 1. (b) Condition 2. (c) Condition 3.

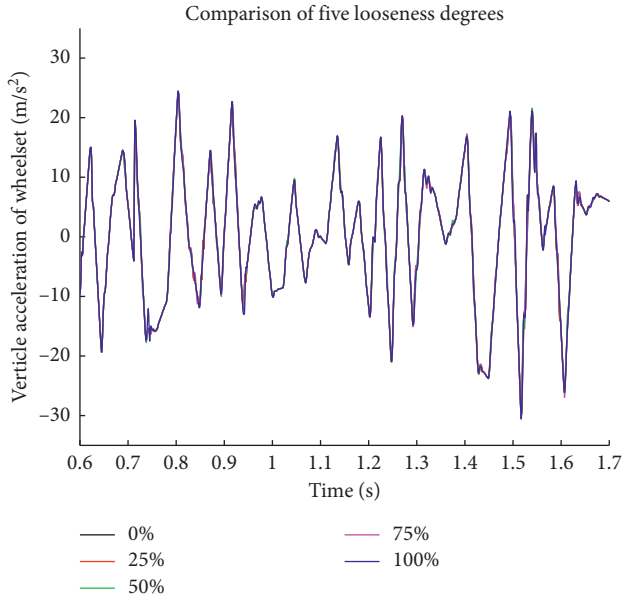


Figure 9: Wheelset vertical acceleration signal with different degrees of fastener looseness in condition 1, looseness 0% (black), 25% (brown), 50% (green), 75% (purple), and 100% (blue).

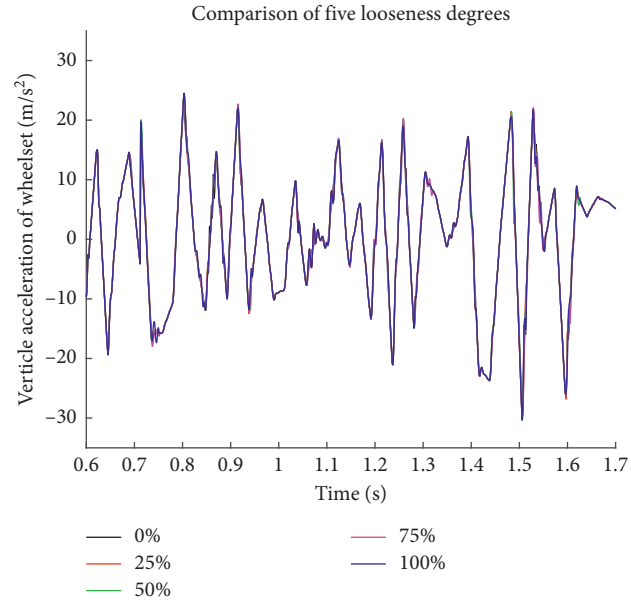


Figure 10: Wheelset vertical acceleration with different degrees of fastener looseness in condition 2.

obviously. The vertical acceleration of the wheelset with loosened fasteners has a variation of 1 m/s^2 to 6 m/s^2, when the fastener is completely loose, the change reaches the maximum. Wheelset vertical acceleration comparison of five looseness degrees in condition 2 and condition 3 is shown in Figures 10 and 11. The looseness of the fastener will increase the vibration and displacement of the rail, which is equivalent to the influence of the irregularity on the train in the running line.

(2) Wheelset vertical acceleration in condition 2

(3) Wheelset vertical acceleration in condition 3

3. Identification of Fastener Loosening Feature Based on Response Signal

3.1. Identification of Fasteners State in Varying Degrees. HHT has good reliability in dealing with nonstationary and nonlinear signals, detecting structural faults. However, it is not based on a complete theory and has energy leakage problems. In order to solve this problem, this paper optimizes the algorithm according to the theory of orthogonality and correlation and combines the improved Hilbert Huang algorithm with the theory of information entropy [22, 23], which is applied to the identification of rail fastener looseness.

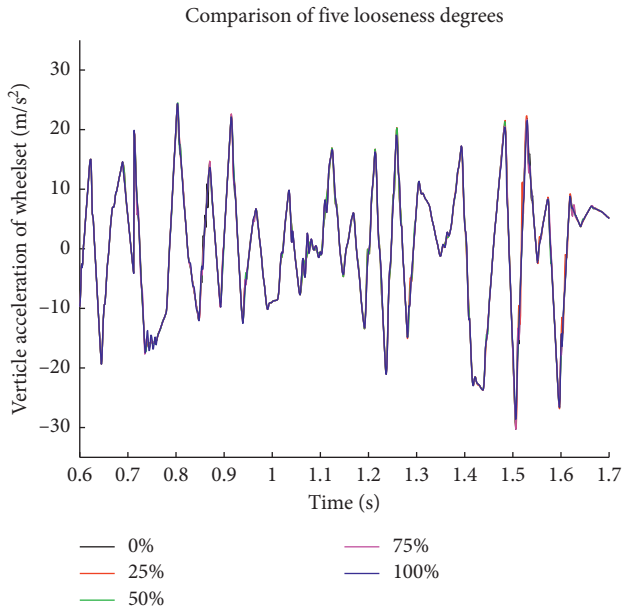


FIGURE 11: Wheelset vertical acceleration with different degrees of fastener looseness in condition 3.

3.1.1. Hilbert–Huang Transform. The crux of HHT algorithm is EMD decomposition. The result of EMD decomposition directly affects the accuracy of subsequent signal processing, so it is necessary to ensure that the IMF component of the EMD decomposition should possess completeness and orthogonality, so that no energy leakage in the decomposition can be guaranteed [24, 25]. From the practical point of view, Huang et al. consider that there exists orthogonality between all IMFs decomposed by EMD, but there is no rigorous theoretical derivation proving that they are rigorously orthogonal in the overall situation[29, 30].

Huang proposed two orthogonality indicators to measure the orthogonality between IMF components, namely, the overall orthogonality index (IOT) and the orthogonality index between two components (IO_{jk}) [31]. When all the IMF components are rigorously orthogonal to each other, the IOT and IO_{jk} should be zero.

$$\mathrm{IOT} = \sum_{j=1}^{n+1} \sum_{\substack{k=1 \\ k \neq j}}^{n+1} \frac{\sum_{i=1}^{N} \mathrm{imf}_{ji} \mathrm{imf}_{ki}}{\sum_{i=1}^{N} x_i^2}, \tag{1}$$

$$\mathrm{IO}_{jk} = \frac{\sum_{i=1}^{N} \mathrm{imf}_{ji} \mathrm{imf}_{ki}}{\sum_{i=1}^{N} \left(\mathrm{imf}_{ji}^2 + \mathrm{imf}_{ki}^2 \right)}.$$

In the formula, x_i is the ith original signal, imf_{ji} is the jth IMF component of x_i, imf_k is the kth IMF component of x_i, and j is not equal to k. In addition, the energy can be used to measure the degree of orthogonality between IMFs. The energy of the original signal $X_{(t)} = [x_1, x_2, \ldots, x_n]$ is

$$E_x = \int_0^T X^2(t) \mathrm{d}t = \sum_{i=1}^{N} x_i^2. \tag{2}$$

The energy of each IMF is

$$E_j = \int_0^T \mathrm{imf}_j^2(t) \mathrm{d}t = \sum_{i=1}^{N} \mathrm{imf}_{ji}^2(t), \quad j = 1, \ldots, n+1. \tag{3}$$

In the formula, $\mathrm{imf}_j(t)$ is the jth IMF component; if all the IMFs are strictly orthogonal to each other, $E_x = E_j$ and the leaking energy between the IMFs is zero, that is,

$$E_{\mathrm{IOT}} = \sum_{j=1}^{n+1} E_j(t) = E_x, \tag{4}$$

where E_{IOT} is the energy of IOT, total energy after decomposition which is constant.

$$\begin{aligned} E_{jk} &= \int_0^T \mathrm{imf}_j(t) \mathrm{imf}_k(t) \mathrm{d}t \\ &= \sum_{i=1}^{N} \mathrm{imf}_{ji} \mathrm{imf}_{ki} = 0, \quad j, k = 1, \ldots, n+1; \ j \neq k. \end{aligned} \tag{5}$$

E_{jk} is the leaking energy between imf_j and imf_k. From the processing steps of EMD decomposition, we can see that, in the actual decomposition process, the mean value curve is obtained by fitting the approximate value, and the decomposed IMF is a part of the original signal, so the different IMF components are approximately orthogonal [29, 30]. Huang's theory proves that the EMD orthogonality is in the order of 10^{-3} to 10^{-2}, and the IMFs are not strictly orthogonal, resulting in the phenomenon of energy leakage and modal aliasing during signal analysis, which will bring errors in the later analysis, making the result inaccurate. In order to eliminate energy leakage, it is necessary to ensure that the orthonormal mode functions are strictly orthogonal to each other [31]. Therefore, the traditional EMD algorithm is improved.

3.1.2. OEMD Algorithm. In order to ensure the strict orthogonality between the IMF components after EMD decomposition [32, 33], this paper orthogonalizes the IMF components; the basic process is shown as follows:

(1) The original signal $X(t)$ is decomposed into the form of the sum of multiple IMF components $c_i(t)$ ($i = 1, 2, \ldots, n$) and a residual component $r_n(t)$ through EMD:

$$X(t) = \sum_{j=1}^{n} c_j(t) + r_n(t). \tag{6}$$

(2) Assume $\mathrm{imf}_1(t) = c_1(t)$ and it is the first orthogonalized IMF component of the original signal, which is the highest frequency component of the IMF.

(3) In the process of decomposition to obtain the second IMF component $c_2(t)$, we can see that it is impossible to ensure the orthogonality between $c_2(t)$ and $\mathrm{imf}_1(t)$. Therefore, we need to remove $\mathrm{imf}_1(t)$ from $c_2(t)$; that is,

$$\mathrm{imf}_2(t) = c_2(t) - \beta_{21}\mathrm{imf}_1(t), \tag{7}$$

$\mathrm{imf}_2(t)$ is the second orthogonalized IMF component of the original signal $X(t)$, and β_{21} is the orthogonalization coefficient between $c_2(t)$ and $\mathrm{imf}_1(t)$. Each side of equations (2)–(7) is multiplied by $\mathrm{imf}_1(t)$ so as to integrate t. Since $\mathrm{imf}_1(t)$ and $\mathrm{imf}_2(t)$ are orthogonal, it can be concluded that

$$\begin{aligned}&\int_0^T \mathrm{imf}_1(t)\mathrm{imf}_2(t)\mathrm{d}t \\ &= \int_0^T c_2(t)\mathrm{imf}_1(t)\mathrm{d}t - \beta_{21}\int_0^T \mathrm{imf}_1^2(t)\mathrm{d}t = 0,\end{aligned} \tag{8}$$

$$\beta_{21} = \frac{\left[\overrightarrow{c_2}\right]^T\left[\overrightarrow{\mathrm{imf}_1}\right]}{\left[\overrightarrow{\mathrm{imf}_1}\right]^T\left[\overrightarrow{\mathrm{imf}_1}\right]}. \tag{9}$$

The numerator and denominator in equations (2)–(9) are inner product operations of two vectors.

(4) By analogy, the $j+1$th orthogonal component, $\mathrm{imf}_{j+1}(t)$, of the original signal can be obtained by eliminating the $\mathrm{imf}_i(t)$ $(i = 1, 2, \ldots, j)$ components in the $j+1$th IMF component $c_{j+1}(t)$. That is,

$$\mathrm{imf}_{j+1}(t) = c_{j+1}(t) - \sum_{i=1}^{j}\beta_{j+1,i}\mathrm{imf}_i(t). \tag{10}$$

Each side of the above equation is multiplied by $\mathrm{imf}_i(t)$ $(i \le j)$, so as to integrate t. Owing to the fact that $\mathrm{imf}_i(t)$ and $\mathrm{imf}_{j+1}(t)$ are orthogonal, it can be concluded that

$$\beta_{j+1,i} = \frac{\left[\overrightarrow{c_{j+1}}\right]^T\left[\overrightarrow{\mathrm{imf}_i}\right]}{\left[\overrightarrow{\mathrm{imf}_i}\right]^T\left[\overrightarrow{\mathrm{imf}_i}\right]}. \tag{11}$$

After the above calculation, the original signal $X(t)$ is decomposed into the form of the sum of the orthogonalized IMF components and the residual components, that is,

$$\begin{aligned}X(t) &= c_1(t) + c_2(t) + \cdots + c_n(t) + r_n(t) \\ &= \mathrm{imf}_1(t) + \left[\mathrm{imf}_2(t) + \beta_{21}\mathrm{imf}_1(t)\right] \\ &\quad + \left[\mathrm{imf}_3(t) + \beta_{31}\mathrm{imf}_1(t) + \beta_{32}c_2(t)\right] \\ &\quad + \cdots + \left[\mathrm{imf}_n(t) + \beta_{n1}\mathrm{imf}_1(t) + \beta_{n2}\mathrm{imf}_2(t)\right. \\ &\quad \left. + \cdots + \beta_{n,n-1}\mathrm{imf}_{n-1}(t)\right] + r_n(t) \\ &= \left(1 + \beta_{21} + \beta_{31} + \beta_{41} + \cdots + \beta_{n1}\right)\mathrm{imf}_1(t) \\ &\quad + \left(1 + \beta_{32} + \beta_{42} + \cdots + \beta_{n2}\right)\mathrm{imf}_2(t) \\ &\quad + \cdots + \left(1 + \beta_{n,n-1}\right)\mathrm{imf}_{n-1}(t) + \mathrm{imf}_n(t) + r_n(t) \\ &= \mathrm{imf}_1^*(t) + \mathrm{imf}_2^*(t) + \cdots + \mathrm{imf}_n^*(t) + r_n(t) \\ &= \sum_{j=1}^{n}\mathrm{imf}_j^*(t) + r_n(t) \\ &= \sum_{j=1}^{n} a_j\mathrm{imf}_j(t) + r_n(t).\end{aligned} \tag{12}$$

In equations (2)–(12), $a_j = \sum_{i=j}^{n}\beta_{i,j}$, $j = 1, 2, \ldots, n$, $\beta_{i,j} = 1$ $(i = j)$. From the above calculation, we can see that there is a strict orthogonality in $\mathrm{imf}_j(t)$, so the linear transformation on $\mathrm{imf}_j(t)$ will not change its orthogonality, and $\mathrm{imf}_j^*(t)$ is also strictly orthogonal.

The integration of the intrinsic mode components after the decomposition of the above EMD is called the orthogonal empirical mode decomposition (OEMD). According to the different orthogonal sequence of IMF components, the OEMD can be divided into three kinds. The way to orthogonalize the high frequency IMF component before orthogonalizing the low frequency IMF component is called OEMD1, the way to deal with IMF in reverse order is called OEMD2, and the way to perform orthogonalization from any IMF is called OEMD3. The essence of OEMD is still EMD; only the IMF components decomposed by EMD are orthogonalized and then reorganized to achieve strict orthogonality of each IMF component.

3.2. OEMD Decomposition of the Vertical Acceleration Signals of Wheelset. The EMD is used to decompose the vertical acceleration signal (Figure 9) of the wheelset without loosening of the fastener under the working condition 1 and obtain 7 IMF components and 1 residual component, as shown in Figure 12. The same signals are processed by OEMD1 and OEMD2, respectively; the results of 7 IMF components and 1 residual component can be obtained.

The overall orthogonal index obtained by the decomposition of EMD is 0.7546, which can be calculated by formula (1). The overall orthogonal indices obtained by OEMD1 and OEMD2 are 0.006463 and 0.007347, respectively. It can be seen that the overall orthogonal index obtained by the OEMD is 2 orders of magnitude higher than that of the overall EMD. There is little difference in accuracy of the orthogonality of the two orthogonal EMD algorithms. Since the 8th IMF is a residual component with a very small value, it has no orthogonality with the first 7 components.

The orthogonality between the 7 IMF components after the decomposition of EMD, OEMD1, and OEMD2 is calculated, respectively, by formula (1); the results of the calculation are shown in Tables 3 and 4. The upper triangular data of Table 3 are the orthogonal indexes between IMF components obtained by EMD decomposition, and the lower triangular data are the orthogonal indexes between IMF components obtained by OEMD1 decomposition. The upper triangular data of Table 4 are the orthogonal indexes obtained by EMD decomposition, and the lower triangular data are the orthogonal indexes obtained by OEMD2 decomposition.

It can be seen from Tables 3 and 4 that the orthogonality index between IMFs by the traditional EMD decomposition is up to 10^{-4}, and the IMF orthogonal index by orthogonalization can reach 10^{-19}, and the accuracy is improved by 15 orders of magnitude, which is less than 10^{-16}, the effective magnitude that the computer calculates in orthogonal index. The result shows that the IMF obtained by using the orthogonalized algorithm has strict orthogonality.

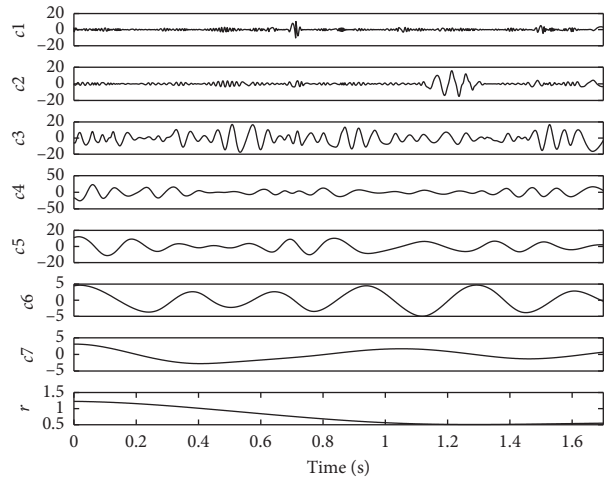


Figure 12: EMD decomposition results of simulation signal (no loose fasteners).

In addition, the energy index can also be used to analyze the orthogonality of IMF. Based on the total energy of EMD, OEMD1, and OEMD2 which can be drawn from formulas (2) and (3), and IMF-decomposed energy shown in Tables 5 and 6, it can be seen that the energy sum of IMF components obtained by EMD has a larger error than that of the original signal, which is 30.8%, and the energy sums of IMF components obtained by the decomposed algorithm of OEMD1 and OEMD2 have the errors of 0.76% and 1.14% against that of the original signal, so the application of OEMD1 and OEMD2 can effectively reduce the energy error, and the effect is obvious.

It is known from Table 5 that the energy of this signal is mainly concentrated in the first 5 orders, accounting for 96.05% of the total energy. Therefore, OEMD3 is used to process the signal in the orthogonal order of $c5$-$c4$-$c3$-$c2$-$c1$-$c6$-$c7$-$c8$. The decomposed IMFs are shown in Figure 13. The overall orthogonal index is 0.001121 drawn from formula (1)–IOT. The orthogonal index of each component is calculated according to formula (1)–IOT_{jk}, as shown in Table 7; the upper triangular data of this table are the orthogonal indexes between each IMF component obtained by EMD decomposition, and the lower triangular data are the orthogonal indexes by OEMD3 decomposition. From the table, it can be seen that the application of OEMD3 algorithm improves the IMF orthogonal order of magnitude as well as the first two algorithms. This algorithm ensures the strict orthogonality of each IMF component that is decomposed.

According to equations (2) and (3), the energy values of IMFs obtained by the decomposition of OEMD3 algorithm are shown in Table 8. The table shows that the energy error between IMFs decomposed by this algorithm is 0.38%, which is smaller than that of the first two algorithms. The energy errors between the IMF components reduce the energy leakage effectively.

To sum up, from the calculation of the wheelset vertical acceleration signals without fastener loosening, it can be seen that the orthogonal index of IMFs decomposed by OEMD1, OEMD2, and OEMD3 has a great improvement, and the strict orthogonality between the components is guaranteed compared with the orthogonal index of the IMFs decomposed by EMD. According to the results of energy index and orthogonality index calculated by the three algorithms, the total energy index error of each IMF component decomposed by OEMD3 and the error of each component energy index are less than the other two algorithms. Therefore, this algorithm is used to decompose the vertical acceleration of the wheelset with different degrees of fasteners loosening in condition 1, condition 2, and condition 3. The wheelset vertical accelerations decomposed by EMD and OEMD3 with fasteners loosening 50% and 100% in condition 1 are shown in Figure 14. It can be seen that, in condition 1, there are some certain differences in the IMF obtained from OEMD3 with the vertical acceleration signals that have different degrees of loose fasteners. In particular, there is one more orthogonal IMF component when the fastener is completely loosened than in the other loosening degrees. This shows that the different degrees of fastener loosening have different effects on the frequency components of the vertical acceleration signal; the difference is most obvious when the fastener is completely loosened, so the time-frequency analysis is considered in the follow-up analysis to identify the loosening features of the fastener.

3.3. Orthogonal IMF Selection Based on Correlation. The above analysis shows that the OEMD can well guarantee the strict orthogonality between the IMF components and ensure that the energy does not leak, but it can be seen from Figure 13 that some interfering signals are introduced by IMFs. These interfering signals have an impact on the accuracy of the subsequent analysis and should be eliminated. Orthogonalized IMF should have a good correlation with the original signal, so this paper uses the correlation coefficient to select a large correlation IMF. Setting the threshold to 0.5, when the correlation coefficient of IMF and the original signal is greater than 0.5, it is believed that the reliability of the signal is relatively higher. The correlation coefficients between the IMF components and the original signals decomposed by the OEMD3 in the first condition when the fastener is not loosed are shown in Table 9. It can be seen

Table 3: Orthogonal index of each component in EMD and OEMD2.

IMF	1	2	3	4	5	6	7	8
1	0.5	$5.60e-2$	$9.82e-3$	$1.17e-2$	$6.07e-3$	$2.16e-3$	$1.80e-2$	$3.63e-2$
2	$5.68e-19$	0.5	$6.29e-2$	$6.65e-3$	$1.13e-2$	$1.83e-2$	$7.31e-2$	$1.59e-2$
3	$2.45e-17$	$8.16e-17$	0.5	$3.80e-2$	$1.92e-2$	$1.02e-2$	$1.11e-2$	$4.79e-3$
4	$2.16e-17$	$9.92e-18$	$1.48e-16$	0.5	$5.84e-2$	$2.36e-3$	$6.89e-3$	$3.48e-3$
5	$3.31e-18$	$2.11e-17$	$1.91e-17$	$1.22e-16$	0.5	$6.19e-4$	$2.80e-3$	$4.35e-3$
6	$5.56e-18$	$2.65e-18$	$1.75e-17$	$4.80e-18$	$1.18e-17$	0.5	$8.86e-2$	$4.76e-3$
7	$4.90e-18$	$7.17e-18$	$1.78e-17$	$9.65e-18$	$3.10e-17$	$1.93e-16$	0.5	$1.53e-2$
8	$3.63e-2$	$1.94e-2$	$7.18e-3$	$3.62e-3$	$2.73e-4$	$4.38e-2$	$1.08e-2$	0.5

Table 4: Orthogonal index of each component in EMD and OEMD1.

IMF	1	2	3	4	5	6	7	8
1	0.5	$8.86e-2$	$2.80e-3$	$6.89e-3$	$1.11e-2$	$7.31e-2$	$1.80e-2$	$1.70e-2$
2	$4.76e-17$	0.5	$6.19e-4$	$2.36e-3$	$1.02e-2$	$1.83e-2$	$2.16e-3$	$1.07e-2$
3	$3.15e-18$	$2.76e-17$	0.5	$5.84e-2$	$1.92e-2$	$1.13e-2$	$6.07e-3$	$1.03e-3$
4	$3.10e-17$	$1.07e-17$	$4.52e-17$	0.5	$3.80e-2$	$6.65e-3$	$1.17e-2$	$3.37e-4$
5	$2.27e-17$	$1.17e-17$	$3.00e-17$	$8.13e-18$	0.5	$6.29e-2$	$9.82e-3$	$1.89e-3$
6	$1.64e-18$	$1.24e-17$	$3.46e-17$	$1.37e-17$	$2.66e-16$	0.5	$5.60e-2$	$1.81e-2$
7	$3.87e-20$	$1.97e-18$	$2.60e-18$	$4.42e-17$	$5.67e-17$	$1.23e-16$	0.5	$9.93e-2$
8	$1.53e-2$	$2.80e-3$	$3.90e-3$	$3.95e-3$	$3.71e-4$	$1.65e-2$	$4.25e-2$	0.5

Table 5: Energy index of each component in OEMD1 and EMD.

Method	E_X	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_{IOT}	Error
EMD	$2.63e5$	$4.6e3$	$1.5e4$	$1.1e5$	$1.5e5$	$4.6e4$	$1.3e4$	$4.3e3$	$1.2e3$	$3.44e5$	30.8%
OEMD1	$2.63e5$	$7.9e3$	$9.4e4$	$9.5e4$	$5.3e4$	$4.1e3$	$4.0e3$	$1.6e3$	$1.2e3$	$2.61e5$	0.76%

Table 6: Energy index of each component in OEMD2 and EMD.

Method	E_X	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_{IOT}	Error
EMD	$2.63e5$	$4.3e3$	$1.3e4$	$4.6e4$	$1.5e5$	$1.1e5$	$1.5e4$	$4.6e3$	$1.2e3$	$3.44e5$	30.8%
OEMD2	2.63e5	2.6e3	4.2e3	2.6e4	1.1e5	1.0e5	1.2e4	4.3e3	1.2e3	2.60e5	1.14%

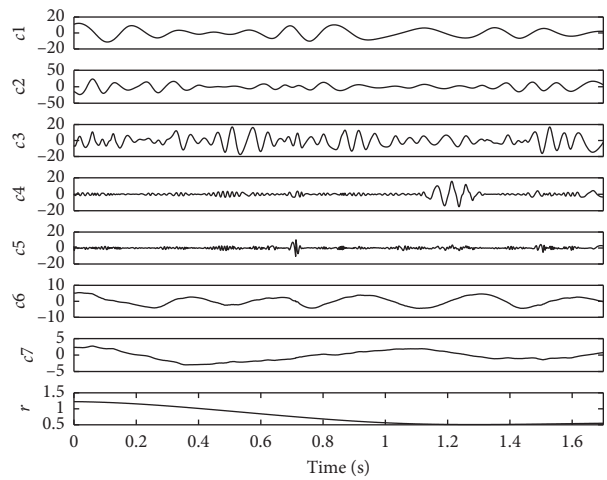


Figure 13: OEMD3 decomposes IMF components (no loose fasteners).

Table 7: Orthogonality index of each component in EMD and OEMD3.

IMF	1	2	3	4	5	6	7	8
1	0.5	$3.80e-2$	$1.97e-2$	$1.02e-2$	$1.11e-2$	$6.29e-2$	$9.82e-3$	$4.79e-3$
2	$1.05e-17$	0.5	$5.84e-2$	$2.36e-3$	$6.89e-3$	$6.65e-3$	$1.17e-2$	$3.48e-3$
3	$4.42e-17$	$1.93e-16$	0.5	$6.19e-4$	$2.80e-3$	$1.13e-2$	$6.07e-3$	$4.35e-3$
4	$1.23e-18$	$8.65e-18$	$8.68e-18$	0.5	$8.86e-2$	$1.83e-3$	$2.15e-3$	$4.76e-3$
5	$4.38e-17$	$7.20e-18$	$5.32e-17$	$1.39e-16$	0.5	$7.31e-3$	$1.80e-4$	$1.53e-3$
6	$1.62e-16$	$1.38e-17$	$1.10e-17$	$3.53e-19$	$1.87e-18$	0.5	$5.6e-2$	$1.59e-2$
7	$4.78e-18$	$8.10e-17$	$5.82e-18$	$6.86e-18$	$3.79e-18$	$1.20e-16$	0.5	$3.63e-2$
8	$4.79e-3$	$3.26e3$	$4.75e3$	$4.68e3$	$1.15e2$	$1.65e2$	$4.25e2$	0.5

Table 8: The energy values of IMFs decomposed by OEMD3 algorithm.

Method	E_X	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_{IOT}	Error (%)
EMD	$2.63e5$	$4.6e4$	$1.5e5$	$1.1e5$	$1.5e4$	$4.6e3$	$1.3e4$	$4.3e3$	$1.2e3$	$3.39e5$	29.0
OEMD3	2.63e5	2.3e4	1.1e5	9.5e4	1.1e4	4.5e3	1.5e4	4.2e3	1.2e3	2.64e5	0.38

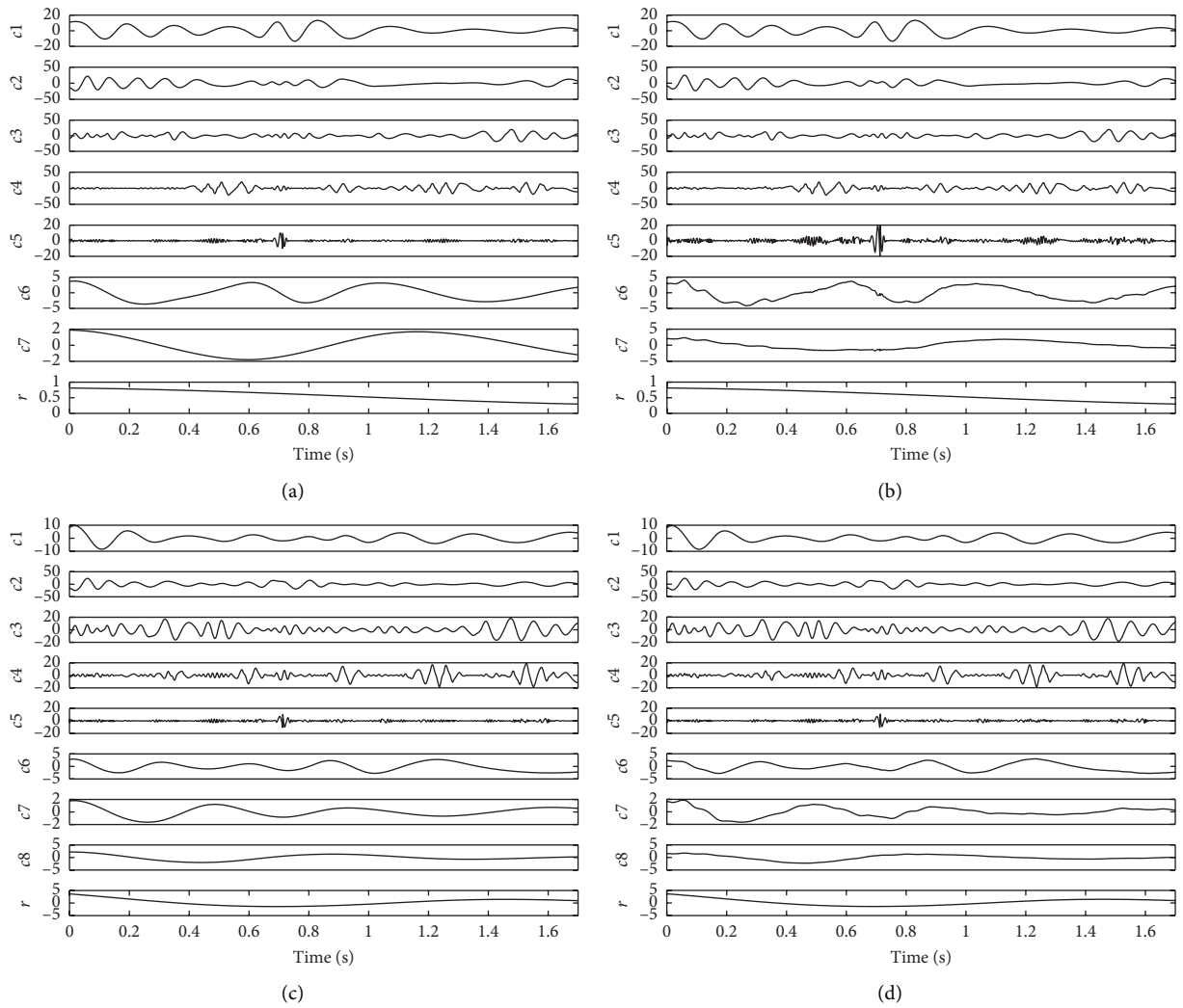


Figure 14: Decomposition results of EMD and OEMD3: (a, b) fastener is loosened by 50% in condition 1. (c, d) Fastener is loosened by 50% in condition 1. (a) EMD decomposition (50%). (b) OEMD3 decomposition (50%). (c) EMD decomposition (100%). (d) OEMD3 decomposition (100%).

Table 9: Correlation coefficient between IMF and original signal.

IMF	Correlation coefficient
1	0.1842
2	0.5661
3	0.6235
4	0.3881
5	0.0264
6	0.1003
7	0.0580
8	0.0839
9	0.0123

from the table that only IMF2 and IMF3 satisfy the conditions, and the rest of IMF components and residual components are discarded without any computing.

3.4. Fastener Looseness Signal Analysis Based on Hilbert–Huang Transform. Hilbert transform is applied to the IMF signal processed by OEMD3 to obtain the corresponding Hilbert time spectrum, and the loose feature is extracted by quantitative analysis. Time-frequency spectrum can reflect the distribution of energy in time and frequency and can reflect the change of signal amplitude with time and frequency. However, due to the complexity of wheelset vertical acceleration signal components and Hilbert spectrum, it requires a lot of experience to directly identify the feature information from Hilbert spectrum. In the time and frequency plane, the difference of different degrees of loosening signals is reflected in the different energy distribution in the corresponding area, that is, the uniformity of energy distribution in the same area. Therefore, the information entropy theory is applied to the quantitative recognition of Hilbert time spectrum, which simplifies the complexity of the algorithm in the calculation speed and arduousness.

3.5. Fastener Looseness Feature Recognition Based on Energy Entropy. In the information theory, supposing the sample space of discrete random variable X is $S = \{x_1, x_1, \ldots, x_n\}$ and the probability of random variable $X = x_j$ is P_j, then the information quantity is

$$I\left(x_j\right) = I\left(X = x_j\right) = \log\left(\frac{1}{p_j}\right) = -\log p_j. \tag{13}$$

The amount of information describes the relationship between the probability of occurrence and information content. The mean value of information quantity $I(x_j)$ in S is the information entropy of X, written as

$$H(X) = -\sum_{j=1}^{N} p_j \log p_j. \tag{14}$$

In this paper, the information entropy and the energy distribution of Hilbert time-frequency spectrum are combined to calculate the entropy of fastener looseness in each condition. This method is called HHT energy entropy. The specific method is as follows.

The energy of the entire Hilbert spectrum plane is denoted as A, which is equally divided into N equal areas, the energy in each small area is denoted as W_i $(i = 1, 2, \ldots, N)$, and the energy of each small area is normalized; thus, $q_i = W_i/A$. According to the formula of information entropy, the formula of HHT energy entropy of wheel acceleration signal is

$$E(q) = -\sum_{i=1}^{N} q_i \ln q_i. \tag{15}$$

According to the basic properties of energy entropy of HHT, the value of $E(q)$ can reflect the uniformity of the energy in the time-frequency plane. The more uniform the energy distribution in the plane, the larger the value of $E(q)$; otherwise, $E(q)$ becomes smaller.

According to the principle of correlation, the IMF components which are closely related to the original signal are selected for the Hilbert frequency spectrum analysis. Finally, the time-frequency spectrum is divided into 170 equal portions, and the entropy of HHT energy in time-frequency spectrum is calculated. The specific process is shown in Figure 15.

The values of HHT energy entropy and their trends under various conditions are shown in Table 10 and Figure 16. It can be seen that the HHT energy entropy of the first condition (one loose fastener) is higher than that of condition 2 and condition 3 (two loose fasteners); in case of 100% looseness, the energy entropy of condition 1 is 4.2672, and those of condition 2 and condition 3 are 4.2043 and 4.2339, respectively. For the same working condition, the HHT energy entropy is less than that without looseness; for example, in working condition 1, the energy entropy decreases from 4.5678 to 4.2672 as looseness increases. Therefore, the HTT energy entropy of wheelset vibration signals decreases with the increase of fastener loosening degree. When the fastener is completely loose, the HHT energy entropy is the smallest. This is because the stiffness of the fastener affects the vibration of various components of the vehicle and rail system. As the rigidity decreases, the rail vibration and the rail displacement will change, just like higher irregularity's impact on a vehicle in the running line, which will affect the vibration of the vehicle system, make the time-frequency distribution of the vibration signal inhomogeneous, and cause the value of energy entropy to drop. Therefore, calculating HHT energy entropy of wheelset vertical acceleration can reflect the looseness of fastener to a certain extent, and it can provide reference for the identification of fastener looseness.

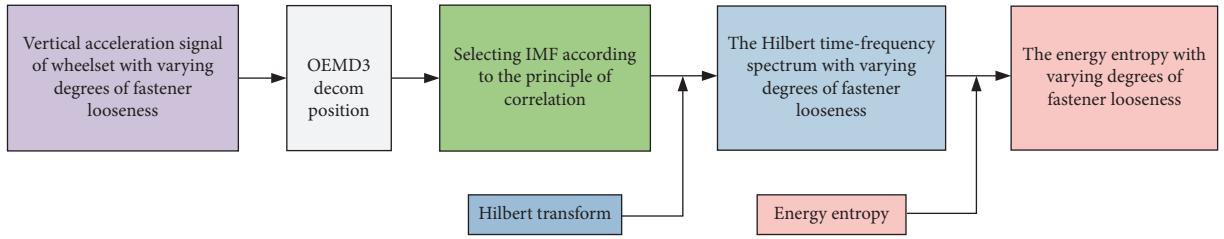


Figure 15: Calculation of the Hilbert energy entropy process. Firstly, varying vertical acceleration signals are gotten and then decomposed by OEMD3. Secondly, IMFS is selected based on correlation principle, which is calculated by Hilbert transform. Lastly, energy entropy theory is used to compute the energy in different looseness conditions.

Table 10: Energy entropy of fastener loosening in various working conditions.

Level of fastener looseness	0%	25%	50%	75%	100%
Energy entropy (condition 1)	4.5678	4.5032	4.4672	4.3986	4.2672
Energy entropy (condition 2)	4.5678	4.4791	4.3627	4.3276	4.2043
Energy entropy (condition 3)	4.5678	4.4356	4.3914	4.3022	4.2339

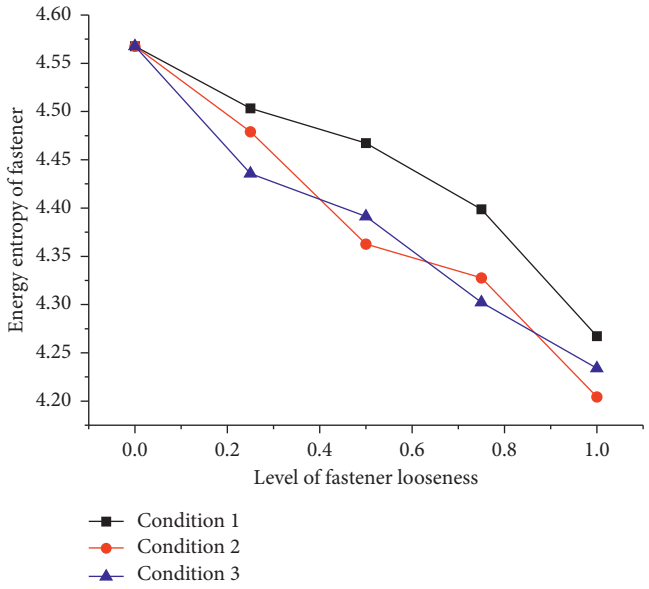


Figure 16: Energy entropy trend of vertical vibration of wheelset. Condition 1 (black square), condition 2 (red dot), and condition 3 (blue triangle); the looseness is from 0% to 100%, increasing gradually at an interval of 25%.

4. Conclusions

This paper starts with the idea of identifying the loosening of the fastener from the vibration response signal of the vehicle, and the traditional vehicle-track rigid-body model is improved. The vehicle-track rigid-flexible coupling model is established by using multibody dynamics software SIMPACK and finite element analysis software ANSYS. The rigid-flexible coupling model simulates the vibration response signals of the wheelsets and the vehicle body in three operating conditions. Combined with improved HHT algorithm and the information entropy theory, the energy entropy is applied to the recognition of the vertical acceleration signal of wheelset with fastener loosening; a certain degree of recognition effects had been achieved. The main conclusions of this paper are listed as follows:

(1) Firstly, the finite element model of rail and rail slab is established in ANSYS, and the model is discretized. The substructure modal analysis of rail and rail slab is carried out, and the modeling parameters are obtained. In order to solve the shortcoming that SIMPACK wheel rail module can only realize the physical contact between rigid body and rigid body, a virtual rail without mass and moment of inertia is added between the wheel and the rail, and the transfer of force is realized based on the balance of force and deformation compatibility condition, the fastener and CA mortar are simulated by force element, and the flexible track and rigid vehicle are assembled in the SIMPACK, and a vehicle-track rigid-flexible coupling model is established. Finally, the vertical accuracy of the model is verified by comparing the measured acceleration of quasi-high speed vehicles with different track irregularities.

(2) Three typical lines with loosened fasteners are set up and the five degrees of fastener loosening are set up. Based on the model built in this paper, the vehicle runs on the rails that the fasteners are loosened in varying degrees and under the effect of German low-interference spectrum at the speed of 200 km/h, and the corresponding vertical acceleration of wheelset and the vibration response of vehicle vertical acceleration are obtained. Finally, the vertical acceleration of the vehicle wheel with obvious vibration response characteristics is selected for the subsequent analysis.

(3) The Hilbert–Huang Transform is applied to the analysis of wheelset vibration signals. In order to solve the problems of low orthogonality and energy leakage in the first step EMD of HHT, the EMD is replaced by the orthogonal empirical mode decomposition in processing the vertical acceleration signal of wheelset. Based on this, the correlation

principle is applied to eliminate the components whose correlation with the original signal in the decomposition results is weak. Owing to the fact that Hilbert time-frequency spectrum of vertical acceleration signals under different working conditions is complicated, the algorithm of Hilbert energy-spectrum entropy is proposed by combining the theory of HHT time-frequency spectrum with the theory of information entropy and is applied to the analysis and identification of wheelset vertical acceleration signals in different conditions. The results show that the method of energy entropy can be used to analyze the vertical vibration signals of the wheelset under different degrees of fastener looseness, and it can realize the identification of the looseness of rail fasteners. The simulation results show that the HHT energy entropy of the wheelset vertical vibration signals decreases with the increase of the degree of fastener loosening; the algorithm can recognize the loose state of rail fastener.

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.

Authors' Contributions

W.Z. and X.G. conceptualized the study; W.Z., X.G., and S.Z. contributed to methodology; W.Z., Y.W., and L.P. provided software and validated the study; W.Z., L.P., Y.W., and S.Z. investigated the study; X.G. contributed to resources; W.Z., L.P., and S.Z. contributed to data curation; W.Z., L.P., S.Z., and X.G. wrote, reviewed, and edited the manuscript; W.Z. contributed to visualization, writing—original draft preparation, formal analysis, and project administration; L.P. supervised the study; S.Z. acquired funding. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (grant nos. 51907117 and 51975347).

References

[1] L. Peng, S. Zheng, X. Chai, and L. Li, "A novel tangent error maximum power point tracking algorithm for photovoltaic system under fast multi-changing solar irradiances," *Applied Energy*, vol. 210, pp. 303–316, 2018.

[2] Y. Sun, J. Xu, H. Qiang, and G. Lin, "Adaptive neural-fuzzy robust position control scheme for maglev train systems with experimental verification," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8589–8599, 2019.

[3] Y. Jia, S. Kwong, and J. Hou, "Semi-supervised spectral clustering with structured sparsity regularization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 403–407, 2018.

[4] X. Y. Liu, Z. W. Zhang, and Z. B. Wang, "Influence of fastener stiffness mutation in ballastless track on dynamic characteristics of high-speed trains," *Journal of Railway Engineering Society*, vol. 31, pp. 53–58, 2014.

[5] L. J. Zhao and D. Z. Tan, "Influence of dynamic behavior of traffic orbit system by fastener failure," *Anhui Architecture*, vol. 22, pp. 105-106, 2015.

[6] T. L. Huang, *Study on Some Methods for Identification of Structural System and Damage*, Tongji University, Shanghai, China, 2007.

[7] L. H. Zhang, *HHT Analysis Application in Rail Fastener Loosening Detection*, Dalian University of Technology, Dalian, China, 2014.

[8] H. Hong, H. Lee, N. Jeong, K. Baek, and M. Suh, "A study on an equivalent model of the threaded fasteners in complex structures through tightening and loosening analysis," *Journal of Mechanical Science and Technology*, vol. 34, no. 3, pp. 1195–1205, 2020.

[9] L. Baeza, J. Fayos, A. Roda, and R. Insa, "High frequency railway vehicle-track dynamics through flexible rotating wheelsets," *Vehicle System Dynamics*, vol. 46, no. 7, pp. 647–659, 2008.

[10] P. F. Weston, C. S. Ling, C. J. Goodman, C. Roberts, P. Li, and R. M. Goodall, "Monitoring lateral track irregularity from in-service railway vehicles," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 221, no. 1, pp. 89–100, 2007.

[11] Z. K. Zhan, H. Sun, X. D. Yu et al., "Wireless rail fastener looseness detection based on MEMS accelerometer and vibration entropy," *IEEE Sensors Journal*, vol. 6, pp. 3226–3224, 2019.

[12] H. Fan, Y. Xiong, and Y. Fei, "High-speed railway fastener detection using minima significant region and local binary patterns," *Journal of Physics: Conference Series*, vol. 1302, Article ID 042046, 2019.

[13] D. Gapinski, Z. Koruba, and I. Krzysztofik, "The model of dynamics and control of modified optical scanning seeker in anti-aircraft rocket missile," *Mechanical Systems and Signal Processing*, vol. 45, no. 2, pp. 433–447, 2014.

[14] Z. Tang, X. Yuan, X. Xie, J. Jiang, and J. Zhang, "Implementing railway vehicle dynamics simulation in general-purpose multibody simulation software packages," *Advances in Engineering Software*, vol. 131, pp. 153–165, 2019.

[15] T. Zhang, H. True, and H. Dai, "The influence of the perturbation of the wheel rotation speed on the stability of a railway bogie on steady curve sections of a track," *Vehicle System Dynamics*, vol. 57, no. 3, pp. 425–443, 2019.

[16] W. M. Zhai, *Vehicle-Track Coupling Dynamics*, Science Press, Beijing, China, 3rd edition, 2006.

[17] W. Zhai, "Excitation models of vehicle-track coupled system," *Vehicle–Track Coupled Dynamics*, pp. 151–202, Springer, Singapore, 2020.

[18] F. Liu, H. Zhang, X. He, Y. Zhao, F. Gu, and A. D. Ball, "Correlation signal subset-based stochastic subspace identification for an online identification of railway vehicle suspension systems," *Vehicle System Dynamics*, vol. 58, no. 4, pp. 569–589, 2020.

[19] F. Huda, I. Kajiwara, N. Hosoya, and S. Kawamura, "Bolt loosening analysis and diagnosis by non-contact laser excitation vibration tests," *Mechanical Systems and Signal Processing*, vol. 40, no. 2, pp. 589–604, 2013.

[20] K.-H. Baek, N.-T. Jeong, H.-R. Hong et al., "Loosening mechanism of threaded fastener for complex structures,"

Journal of Mechanical Science and Technology, vol. 33, no. 4, pp. 1689–1702, 2019.

[21] L. Farkas, D. Moens, S. Donders, and D. Vandepitte, "Optimisation study of a vehicle bumper subsystem with fuzzy parameters," *Mechanical Systems and Signal Processing*, vol. 32, pp. 59–68, 2012.

[22] S. Klus, I. Schuster, and K. Muandet, "Eigendecompositions of transfer operators in reproducing Kernel Hilbert spaces," *Journal of Nonlinear Science*, vol. 30, no. 1, pp. 283–315, 2020.

[23] M. Mehrpouya, E. Graham, and S. S. Park, "FRF based joint dynamics modeling and identification," *Mechanical Systems and Signal Processing*, vol. 39, no. 1-2, pp. 265–279, 2013.

[24] Y. Sun, H. Qiang, J. Xu, and G. Lin, "Internet of Things-based online condition monitor and improved adaptive fuzzy control for a medium-low-speed maglev train system," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2629–2639, 2020.

[25] Y. Sun, J. Xu, H. Qiang, C. Chen, and G. Lin, "Adaptive sliding mode control of Maglev system based on RBF neural network minimum parameter learning method," *Measurement*, vol. 141, pp. 217–226, 2019.

[26] N. E. Huang, Z. Shen, S. R. Long, et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-tationary time series analysis," Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, vol. 454, no. 1971, pp. 903–995, 1998.

[27] N. E. Huang, Z. Shen, and S. R. Long, "A new view of nonlinear water waves: the Hilbert spectrum," *Annual Review of Fluid Mechanics*, vol. 31, no. 1, pp. 417–457, 1999.

[28] N. E Huang, "New method for nonlinear and nonstationary time series analysis: empirical mode decomposition and Hilbert spectral analysis," *Proceedings of SPIE: The International Society for Optical Engineering*, vol. 4056, pp. 197–209, 2000.

[29] W. Zhang, L. Peng, and L. Li, "Design and implementation of a new algorithm for optimal route traversal searching in interlocking stations," in *Proceedings of the 13th Asia Pacific Transportation Development Conference*, pp. 353–361, Shanghai, China, May 2020.

[30] L. Liu, Z. Zuo, Y. Zhou, and J. Qin, "Insights into the effect of WJ-7 fastener rubber pad to vehicle-rail-viaduct coupled dynamics," *Applied Sciences*, vol. 10, no. 5, p. 1889, 2020.

[31] Y. Jia, S. Kwong, J. Hou, and W. Wu, "Convex constrained clustering with graph-Laplacian PCA," in *Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME)*, San Diego, CA, USA, July 2018.

[32] J. Yang, W. Tao, M. Liu, Y. Zhang, H. Zhang, and H. Zhao, "An efficient direction field-based method for the detection of fasteners on high-speed railways," *Sensors*, vol. 11, no. 8, pp. 7364–7381, 2011.

[33] M. Bocciolone, A. Caprioli, A. Cigada, and A. Collina, "A measurement system for quick rail inspection and effective track maintenance strategy," *Mechanical Systems and Signal Processing*, vol. 21, no. 3, pp. 1242–1254, 2007.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8589346, 9 pages
https://doi.org/10.1155/2020/8589346

Research Article

An Intelligent Evaluation Method to Analyze the Competitiveness of Airlines

Jun Zhao and Xumei Chen

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Xumei Chen; xmchen@bjtu.edu.cn

Received 6 June 2020; Revised 4 August 2020; Accepted 17 August 2020; Published 7 September 2020

Guest Editor: Jun Shen

An intelligent evaluation method is presented to analyze the competitiveness of airlines. From the perspective of safety, service, and normality, we establish the competitiveness indexes of traffic rights and the standard sample base. The self-organizing mapping (SOM) neural network is utilized to self-organize and self-learn the samples in the state of no supervision and prior knowledge. The training steps of high convergence speed and high clustering accuracy are determined based on the multistep setting. The typical airlines index data are utilized to verify the effect of the self-organizing mapping neural network on the airline competitiveness analysis. The simulation results show that the self-organizing mapping neural network can accurately and effectively classify and evaluate the competitiveness of airlines, and the results have important reference value for the allocation of traffic rights resources.

1. Introduction

Traffic rights are the most competitive resources for airlines to enter the international aviation market. According to the convention on the International Civil Aviation (Chicago Convention), civil aviation authorities of each country can sign a bilateral or multilateral air transport agreement, which stipulates the right of each designated airline to operate an agreed flight on a prescribed route. Based on the different agreements, some countries do not limit the number of designated carriers, route list, and capacity quota for airlines. Airlines can arrange flight plans according to route network planning and market demand, such as China and the Association of Southeast Asian Nations (ASEAN) countries. Other countries have set restrictions on the operation rights of airlines, and the resources of traffic rights are relatively scarce. There exists competition for the traffic rights among the airlines. The reasonable allocation of these traffic rights resources is of great significance to the aviation transport market. There are many experiences in the airline management of other countries. The most reasonable allocation basis for the traffic rights resources is based on the competitiveness ranking of the airlines.

The experts and scholars have studied the competitiveness of airlines deeply. The competitiveness of airlines can be measured in terms of some indexes, such as normality [1, 2], safety [3, 4], service [5, 6], productivity [7], on-time performance [8], and market share [9] [10–12]. However, these studies of airlines' competitiveness are mainly from the perspective of airlines to acquire and retain customers in such a highly competitive market [13–15]. There is little research on the levels and competitiveness from the perspective of management departments, so as to provide a support for the allocation of the traffic rights resources. It is thus the primary purpose of this paper to evaluate and analyze the competitiveness of airlines based on the management organization.

At present, the analysis and evaluation of airline competitiveness mainly focus on the use of the analytic hierarchy process (AHP), fuzzy comprehensive evaluation, factor analysis, Delphi technique, data envelopment analysis (DEA) model, grey clustering analysis, and other methods [16–18]. Among them, Delphi technique is a qualitative method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem [19].

Because the airlines competitiveness can be viewed as a complex problem, Delphi technique was utilized to reach a consensus among the airlines experts. The AHP technique is "a theory of measurement through pairwise comparisons and relies on the judgments of experts to derive priority scales." It can determine the priority of the key competitiveness indicators and drivers of airlines [20]. Delbari et al. investigated the competitiveness of airlines using Delphi and AHP techniques. The results revealed that the ranking of the key competitiveness drivers with respect to each indicator differs significantly [21]. Yu and Li [22] used the multilevel fuzzy comprehensive evaluation method to carry out the empirical analysis and the competitiveness analysis of network airlines, so as to solve the fuzzy problem of evaluation indexes and evaluation grades. Fu and Wu [23] established a simple fuzzy analytic hierarchy process of the competitiveness evaluation model of airlines and analyzed the changes of the competitiveness level of airlines in different periods. Li et al. [24] chose the entropy weight extension matter-element model to evaluate the competitiveness of airlines. He believed that airlines can cultivate the competitiveness according to the evaluation results. Li and Wu [25] established a two-stage analysis model reflecting the subjective and effective efforts of airline managers and selected 15 domestic airlines as empirical samples to analyze the competitiveness ranking of each airline. Bai et al. used the DEA model to establish an evaluation method to evaluate the airline's service quality. This method does not need explicit expression of the relationship between input and output, and the result has certain practical value [26, 27]. The abovementioned correlation analysis mainly relies on the model to evaluate the competitiveness of airlines, and the methods mainly rely on the experience and evaluation of experts themselves, which is subjective and lacks the direct reflection of relevant data.

In order to ensure the fairness of the analysis, we use a novel algorithm of self-organizing feature mapping neural network to model and analyze. The self-organization map network has good self-organization, self-adaption, and robustness [28]. It can learn or simulate the unknown environment or sample space without prior knowledge and can deal with quantitative and qualitative knowledge at the same time [29–31], which avoids the subjective evaluation of the competitiveness of airlines. Due to the use of self-learning without teachers, it is a recognition network based on small sample training, which is different from the traditional neural network that requires a large number of training samples to ensure the accuracy of the analysis [32–36]. It is of certain significance for the analysis of the small sample size of the competitiveness of airlines. In addition, there are many factors in the competitiveness of airlines. The network model has the characteristics of distribution [37, 38], which is quite suitable for the study of the classification mechanism of the competitiveness of the airlines. Therefore, this paper uses the index data of traffic rights as feature samples to input SOM neural network for competitiveness analysis and clustering and uses the index data of typical airlines as an example to verify, so as to determine the level of competitiveness of each airline.

In the subsequent sections, we first analyze the key performance indexes suitable for evaluating competitiveness of the airlines. Next, we develop a self-organizing mapping neural network model and design an algorithm process to evaluate the airlines. We then establish a standard sample library to train the SOM network and utilize the actual data to obtain the airline competitiveness results. Finally, we conduct a discussion to compare the SOM neural network method with the grey clustering analysis so as to verify the effectiveness of the method.

2. Determination of Traffic Right Indexes

In China, the civil aviation transportation industry is a fully competitive service industry. Management organizations mainly evaluate and choose airlines through the dimensions of safety, service, and normality. Here, we mainly stand in the perspective of management organizations to make the safety, normality, and service as the basic indexes to evaluate the level of competitiveness. Among them, safety is the foundation of civil aviation operation, and service and normality are the key links of airline competitiveness. These three dimensions constitute the basic indexes of airline competitiveness evaluation and analysis. Considering the perspective of safety, service, and normality, this paper selects four indexes, namely, the ten-thousand-hour rate of flight accidents, the abnormal rate of flights caused by the airline, the rate of passenger complaints, and the rate of flight plan execution, for the evaluation and analysis of the airline competitiveness.

The ten-thousand-hour rate of flight accidents refer to the events that affect or may affect flight safety in every ten-thousand-hour flight activities. It is the core evaluation index for the safe operation of airlines. The abnormal rate of flights caused by the airline refers to the percentage of abnormal flights caused by airline's own reasons in the number of scheduled flights in the assessment cycle. It objectively reflects the flight operation organization and enterprise service management level of the airline and is an important index to assess the operation quality of the airline. The passenger complaint rate refers to the percentage between the number of passenger complaints and passenger traffic volume, which is the main measure of passengers' satisfaction with the transportation services provided by airlines. The rate of flight plan execution refers to the percentage of the actual and planned flight volume of the airline, which reflects the core index of the efficiency of the airline's flight execution.

3. A Self-Organizing Mapping Neural Network

3.1. SOM Network Model. SOM is a new unsupervised competitive learning feedforward neural network model. Considering the unique attributes of SOM model, self-organization and feature mapping, SOM has its core characteristics, that is, fully and effectively retaining the neuron topology. Due to this feature, the winning neuron of the network model and its neighborhood neuron adjust the weight together, and the neighborhood of the neuron is more sensitive to the specified input. In the analysis competitiveness

of the airlines, the determination of index weight avoids the situation that the weight is easily influenced by personal subjective opinions when the weight is determined artificially. At the same time, it avoids the situation that the weight is easily inconsistent with the actual state when it is determined completely according to the numerical value. The analysis result is more stable and efficient.

Firstly, a self-organizing feature mapping neural network model is constructed, which is composed of input layer and competition layer (output layer). The number of neurons in the input layer is n. Set the characteristic input node, $x_1, x_2, \ldots, x_n$. The competition layer is the output layer, which is a two-dimensional plane array composed of $a \times b$ neurons. Each input neuron is connected with all the neurons in the two-dimensional plane array. The training process of SOM network is to constantly adjust the connection weight of network nodes, so that different input types correspond to different neurons in the two-dimensional plane array. Figure 1 shows the structure of the model.

As shown in Figure 1, each node of the competition layer is arranged into a neighborhood mode in some specific form, which specifies the neighborhood structure of each neuron and the location of characteristic nodes belonging to the domain or excluding from the neighborhood. According to the results of the network specific identification and classification, all nodes in the output layer are connected or partially connected. All input nodes and output nodes between input layer and competition layer are connected by weight. This ensures that in the process of SOM neural network training, not only the weights and thresholds corresponding to the winning neuron will be adjusted but also other neurons in the adjacent range will have the opportunity to adjust the weights and thresholds, which guarantees the good learning ability and generalization ability of SOM neural network.

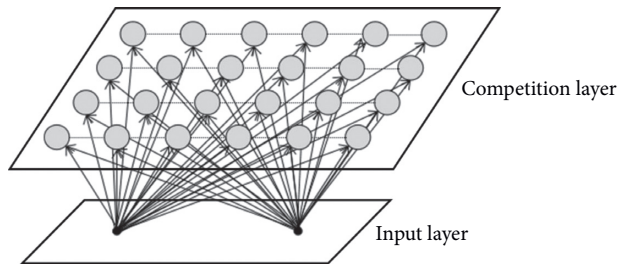


Figure 1: Network model of SOM.

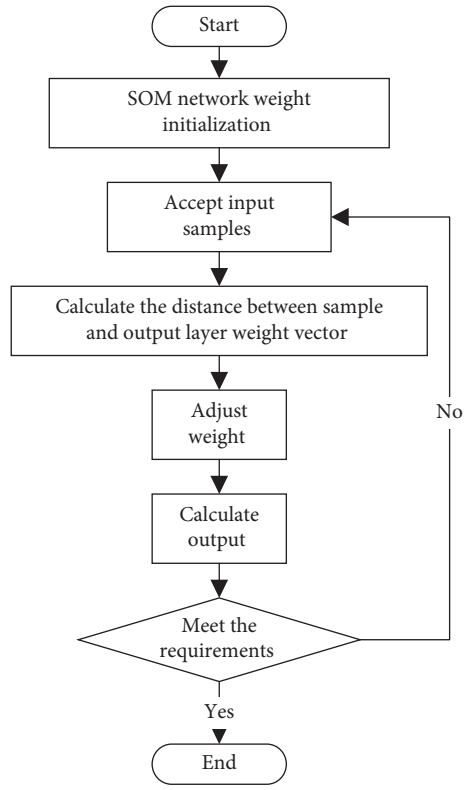


Figure 2: SOM network algorithm flowchart.

3.2. Algorithm Process. The input of the SOM is the competitiveness index value of each airline. Then, the input sample set is classified. The index component parameter vector corresponds to the neuron weight vector of the output layer one by one. In order to clearly describe the learning process of network model, the flowchart of SOM network model learning algorithm is constructed as shown in Figure 2.

3.2.1. Initialization. Generally, any value in the interval [0, 1] of the weight vector will be given, expressed in W. The weight vector $W = [w_{i1}, w_{i2}, \ldots, w_{iN}]^T$ $(i = 1, 2, \ldots, N)$. The learning rate is η, $\eta \in [0, 1]$.

3.2.2. Set the Input Vector. The input vector is composed of indicators, that is, the training samples of the network model are

$$X = [x_1, x_2, x_3, \ldots, x_{N-1}, x_N]^T. \quad (1)$$

3.2.3. Derive the Euclidean Distance. W_{ij} represents the weight of the neuron i in the input layer and the neuron j in the mapping layer. The Euclidean distance between the input vector and the weight vector is derived to obtain the specific location of the neuron. The Euclidean distance d_i is calculated by

$$d_i = \left\| X - W_j \right\| = \sqrt{\sum_{i=1}^{n} \left[x_i(t) - w_{ij}(t) \right]^2}. \quad (2)$$

3.2.4. Label the Winning Neurons. The position of the winning neuron is the minimum neuron position of Euclidean distance between the input vector and the weight vector. The Euclidean distance between the input vector and the weight vector is expressed by $\| \cdot \|$. When the input vector is X, the t-th neuron wins, which needs to meet the following requirements:

$$\|X - W_i\| = \min_i \|X - W_i\|, \quad i = 1, 2, 3, \ldots, N-1, N. \tag{3}$$

3.2.5. Cooperation of Topological Neighborhood Neurons. The winning neurons are located in the center of the topological neighborhood and consist of single or group neurons. According to the construction of network model, the shape of neighborhood is selected. Common neighborhood geometry includes linear neighborhood, square neighborhood, polygon neighborhood, etc. In order to improve the efficiency of cooperation, this paper chooses hexagon neighborhood as the neighborhood shape of airline competitiveness network model, as shown in Figure 3.

In the process of training, the neighborhood radius and the number of winning neurons gradually increase. As the only index to determine the size of neighborhood, the radius of topological neighborhood is recorded as $N_i(n)$, which represents the radius of topological neighborhood under n times of superposition, that is, the area of neighborhood. It changes with time. The rule is described by equation (4), and the neighborhood change under the specified conditions is described in Figure 4:

$$N_i(n) = INT\left(N_i(0)\left(1 - \frac{n}{N}\right)\right), \quad n = 1, 2, 3, \ldots, N, \tag{4}$$

where INT$(\cdot)$ is to round the function, $N_i(0)$ is the initial value of the topological neighborhood, and N is the total number of iterations.

3.2.6. Adjust the Weight. The weight vector of all winning neurons in the neighborhood is updated until the recognition results meet the specific requirements of the first set competitiveness level judgment. The updating adopts the Hebb learning method. According to equation (5), the connection weights of input neurons and neighboring neurons are modified:

$$\Delta w_{ij} = w_{ij}(t+1) - w_{ij}(t) = \eta(t)\left[x_i(t) - w_{ij}(t)\right], \tag{5}$$

where $\eta(t)$ is the learning rate at time t, $\eta(t) \in [0, 1]$. $\eta(t)$ decreases gradually with the increase of time and is inversely proportional to t, and its expression is

$$\eta(t) = \frac{1}{t}, \quad \text{or } \eta(t) = 0.2\left(1 - \frac{t}{1000}\right). \tag{6}$$

3.2.7. Calculate the Output Value O_k. The calculation output value is given as

$$O_k = f\left(\min\|X - W_j\|\right). \tag{7}$$

Then judge whether the output results meet the requirements of the preset competitiveness level. If the result meets the requirements of the competitiveness level, export the competitiveness level; if the result does not meet the requirements of competitiveness level, return to step (2) to continue learning until it meets the results of competitiveness level.

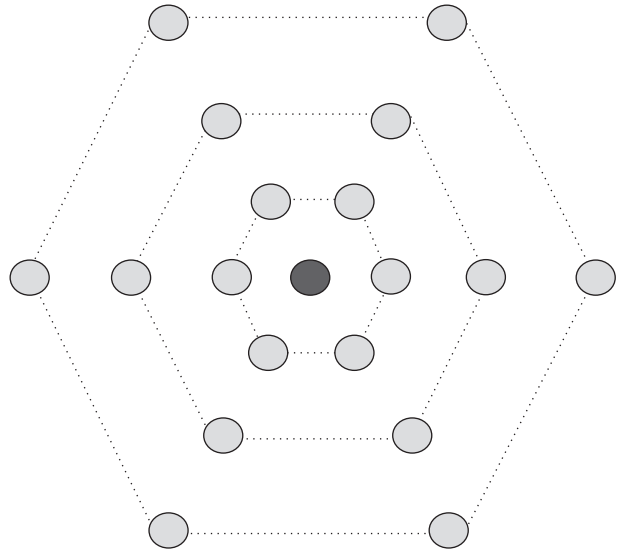


Figure 3: Neighborhood of airline competitiveness network model.

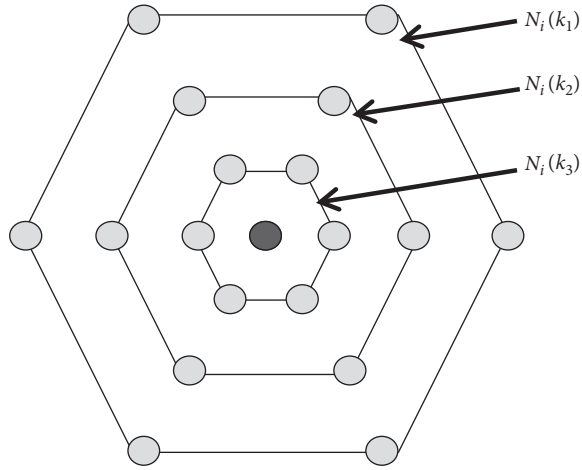


Figure 4: Topological neighborhood change of hexagon.

4. Analysis of Airline Competitiveness Results

4.1. Standard Sample Library. Using the factors of competitiveness introduced in Section 1, four main indexes are selected for testing: the ten thousand-hour rate of flight accidents, the abnormal rate of flights caused by the airline, the rate of passenger complaints, and the rate of flight plan execution. In this paper, four linguistic variables are selected, and the evaluation characteristics are divided into "excellent," "good," "general," and "poor." The index values of the 20th percentile, the 40th percentile, the 60th percentile, and the 80th percentile of each index of airlines are taken as the standard samples of four evaluation characteristics, as shown in Table 1.

Table 1: Standard sample.

Evaluation characteristics	Ten-thousand-hour rate of flight accidents	Abnormal rate of flights caused by airline	Rate of passenger complaints	Rate of flight plan execution
Excellent	0.0000	0.1493	0.2108	0.4049
Good	0.0072	0.1542	0.2273	0.4413
General	0.0446	0.1572	0.2444	0.6811
Poor	0.1382	0.1596	0.2832	0.8124

4.2. SOM Network Training. The evaluation grade and the four indexes of the corresponding index types (i.e., the ten thousand-hour rate of flight accidents, the abnormal rate of flights caused by the airline, the rate of passenger complaints, and the rate of flight plan execution in Table 1) is taken in the standard sample as the input vector of SOM network [$x1$, $x2$, $x3$, $x4$]. The competition layer of the network is set as $6 \times 6 = 36$ neurons; the number of training steps is set as 10, 50, 100, 200, 500, and 1000. The classification effect of different training steps in the SOM network training process is shown in Table 2, where different numbers represent different classification numbers. Figure 5 shows the topology of the winning neurons in different steps.

It can be seen from Table 2 and Figure 5 that, when the number of training steps is 10, excellent and good can be divided into one category and general and poor can be divided into one category. The SOM network only preliminarily classifies standard samples. When the number of training steps increases to 50 and 100, the classification accuracy will be further improved with the gradual increase of training steps, so as to distinguish excellent, good, and general. When the number of training steps reaches 200, the four evaluation levels are completely distinguished. At this time, if we continue to increase the number of training steps to 500 or even 1000, each sample is divided into one category, which has no practical significance. Therefore, 200 steps are the best number of training steps. According to the number of excellent, good, general, and poor neurons corresponding to the 200 step training in Table 2, we can see that the competitive winning neurons in four situations are 7, 26, 30, and 6, respectively, as shown in the grey hexagon in Figure 5 (4). Each neuron has a topological position, that is, (x, y) in the neuron coordinates. It can be seen that the corresponding states of four types of standard samples are clearly distinguished in the two-dimensional array.

4.3. Application and Validation. In order to validate the effect of SOM neural network on airline competitiveness analysis, 20 typical Chinese airlines index data of 2019 are selected. The data are from the transportation department of Civil Aviation Administration of China. The SOM neural network trained in Section 3.2 is input to get airline competitiveness analysis results.

Combining Table 3 and Figure 6, it can be seen that the category label of sample nos. 1, 7, 10, 16, and 19 is 7, which belongs to the excellent category label of clustering results; the category label of sample nos. 6, 8, 9, 11, and 20 is 26, which belongs to the good category label of clustering results; the category labels of sample nos. 2, 5, 13, 17, and 18 is 30, which belongs to the general category labels of clustering results; the category labels of sample nos. 3, 4, 5, 12, 14, and 15 is 6, which belongs to the poor category label of clustering results. The evaluation results above are mostly consistent with the actual situation. The above results can accurately predict the competitiveness of different airlines. The results have a certain reference value for the allocation of traffic rights.

Table 2: Clustering results for different training steps.

Train steps	Clustering results			
	Excellent	Good	General	Poor
10	7	7	30	30
50	7	7	30	6
100	7	26	30	30
200	7	26	30	6
500	8	26	23	6
1000	14	26	36	6

Similarly, we select 20 typical Chinese airlines index data of the past 5 years. The index data are input into SOM neural network to analyze the competitiveness evaluation results of the airlines. The evaluation accuracy TP is calculated as follows:

$$\mathrm{TP} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - E_i}{X_i}. \tag{8}$$

Among them, n represents the data year, X_i represents the total amount of the samples in the i-th year, E_i represents the error samples in the i-th year, and TP represents the evaluation accuracy. According to the calculation of the prediction results, TP = 94.8%. The evaluation accuracy is within a reasonable range. The above prediction results and evaluation accuracy have a good clustering and evaluation effect for the competitiveness analysis of airlines. It has a certain value for the aviation authority to allocate scarce traffic right resources, so that airlines pay more attention to safety, normality, and service, form a benign competition and ultimately enhance the overall competitiveness of the industry.

5. Discussions

The SOM neural network herein is utilized to conduct a cluster evaluation of the airlines. In the previous studies, the grey clustering analysis is an evaluation method to analyze

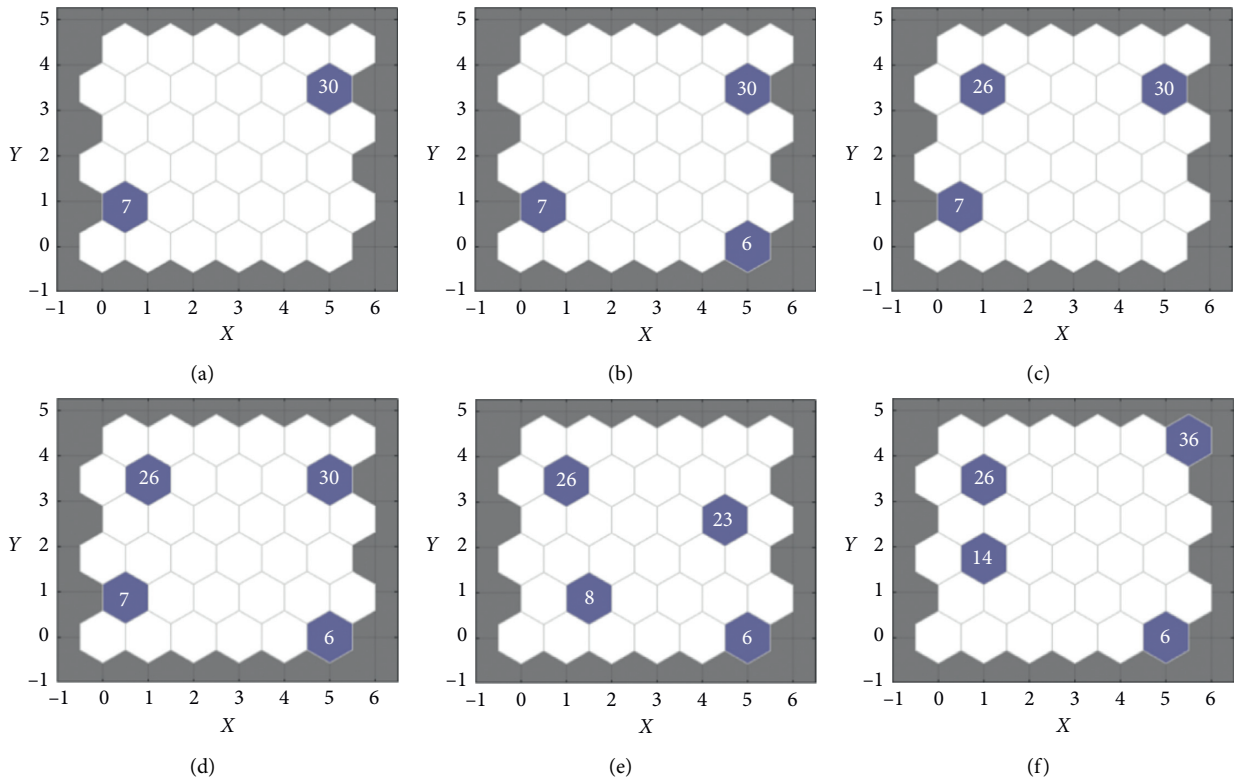


Figure 5: Topology of winning neurons with asynchronous number. (a) 10 steps. (b) 50 steps. (c) 100 steps. (d) 200 steps. (e) 500 steps. (f) 1000 steps.

Table 3: Sample airline data.

Sample label	Ten-thousand-hour rate of flight accidents	Abnormal rate of flights caused by airline	Rate of passenger complaints	Rate of flight plan execution
1	0.0000	0.1458	0.2438	0.9902
2	0.0230	0.1472	0.2042	0.4173
3	0.1310	0.1525	0.8897	0.3492
4	0.1430	0.1568	0.2470	0.4473
5	0.0000	0.1587	0.2290	0.3966
6	0.0000	0.1602	0.2206	0.6667
7	0.0090	0.1668	0.3074	0.8615
8	0.0680	0.1677	0.2859	0.7396
9	0.0060	0.1765	0.3897	0.8258
10	0.0180	0.1776	0.2832	0.8895
11	0.0000	0.1793	0.1958	0.6674
12	0.1320	0.1823	0.7352	0.5384
13	0.0000	0.1927	0.4982	0.8219
14	0.0000	0.1953	0.4810	0.5000
15	0.0000	0.1958	0.2594	0.3418
16	0.0390	0.2075	0.0000	0.8986
17	0.0000	0.2139	0.3597	0.7738
18	0.0160	0.2317	0.3415	0.6035
19	0.0000	0.2516	0.1886	0.7813
20	0.0210	0.1817	0.2958	0.8032

Note: data are from Civil Aviation Administration of China.

the airlines. As described in reference [39], the operation and service quality of the airlines was evaluated and analyzed based on the grey clustering analysis. The whitening weight function is determined the experience, which has a certain of subjectivity as shown in Figure 7. Through the analysis of index data of 20 typical domestic airlines from July to

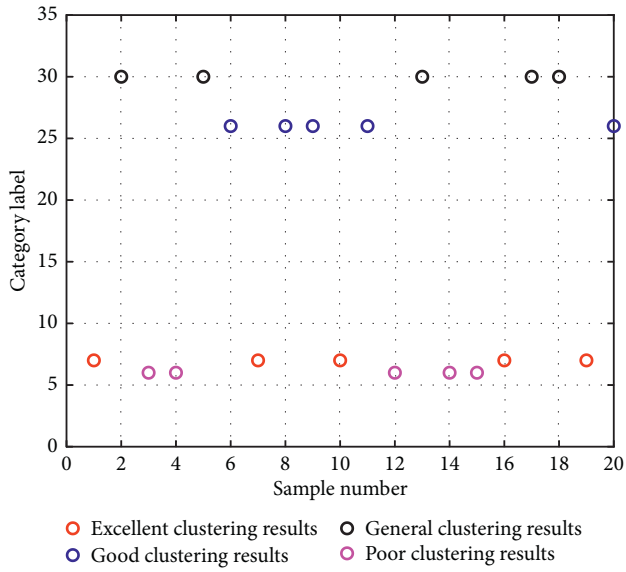


Figure 6: Sample forecast classification.

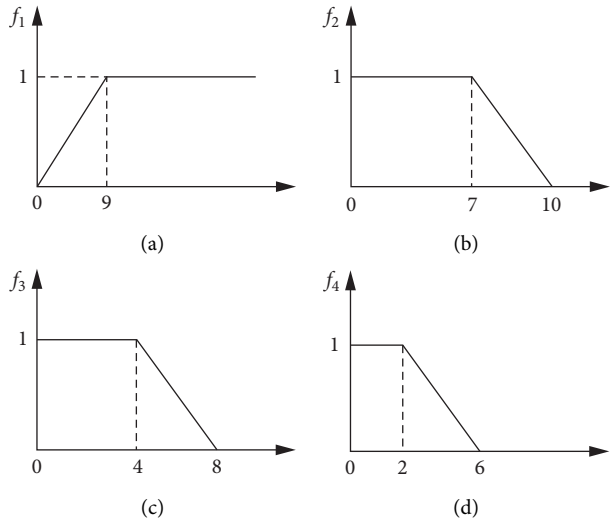


Figure 7: Whitening weight function.

December 2019, the rating of each airline's operating service quality is shown in Table 4.

From Table 4, we can see that the quantity of the excellent level is 9, while the quantity of the poor level is just 2. This is because that the weight of the grey class is determined by the subjective judgment of people. The situation leads to the cluster results where fluctuations exist so that resources cannot be allocated reasonably.

In addition, we also compare the clustering method of the principal component analysis (PCA) from reference [40]. In this existing related work, principal component analysis (PCA) was used to study the index data of samples. The sample index data were standardized. According to the standardized sample index data, the covariance matrix of the factor was obtained, which was the correlation matrix of the original sample index data. Then, the contribution rate of the selected principal components and the cumulative contribution rate of each principal component were calculated. The number of principal components was determined, so as to calculate the comprehensive evaluation value. Finally, the reference obtained the comprehensive score of competitiveness evaluation of five airlines in 2007–2010, as shown in Table 5.

Table 4: Evaluation level.

Evaluation level	Sample serial number
Excellent	10, 11, 12, 14, 15, 16, 17, 18, 19
Good	1, 2, 3, 4, 5, 9
General	6, 8, 20
Poor	7, 13

Table 5: Comprehensive score of competitiveness evaluation of five airlines in 2007–2010.

Year	Airline 1	Airline 2	Airline 3	Airline 4	Airline 5	Average score
2007	3.500	2.850	1.190	1.140	0.560	1.848
2008	4.090	2.560	1.350	0.920	0.450	1.874
2009	3.880	2.960	1.270	0.750	0.710	1.914
2010	3.900	2.430	1.420	1.020	0.450	1.844
Annual average score of the company	3.843	2.700	1.308	0.958	0.543	1.870

Comparing the PCA method with the proposed method, the PCA can obtain the scores of the airlines, which is more intuitive, while the selected principal components and their number are gained from the expert experience. It has certain subjectivity. The method of SOM neural network is to train and learn the standard sample library, which is more objective. And, the results will not be changed by the human factors. It is more suitable for clustering analysis of airline competitiveness.

6. Conclusions

This paper proposed an analysis method of competitiveness of the airlines based on SOM neural network. This method has a good judgment effect on the classification of competitiveness of the airlines. Considering the perspective of safety, service, and normality, the indexes are determined as the ten-thousand-hour rate of flight accidents, the abnormal rate of flights caused by the airline, the rate of passenger complaints, and the rate of flight plan execution. The SOM neural network is used to train and learn the standard sample library. The multistep setting is utilized to determine the number of training steps for high network convergence speed and high clustering accuracy.

The results show that the SOM neural network algorithm has a good clustering and evaluation effect on the analysis of the airline competitiveness through the verification of examples. Based on small samples of the airlines, the evaluation accuracy reached 94.8%, and it was a relatively accurate evaluation accuracy. It has an important reference application value for the allocation of scarce traffic rights resources and the allocation of other key resources.

In addition, further investigation of the airline competitiveness analysis is needed. On one hand, more comprehensive index parameters on airline competitiveness need to be considered. On the other hand, using scientific calculation methods to rank the airlines competitiveness needs to be explored further. For this purpose, it will conduct reasonable allocation of traffic rights resources effectively.

Data Availability

The data in this paper are on the topic of safety and operation of airlines. They are from Civil Aviation Administration of China. The processed data used to support the findings of this study are included within the article. The raw 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 funded by the National Natural Science Foundation of China (project no. 71871013).

References

[1] A. G. Assaf and A. Josiassen, "The operational performance of UK airlines: 2002-2007," *Journal of Economic Studies*, vol. 38, no. 1, pp. 5–16, 2011.

[2] T. J. Hannigan, R. D. Hamilton, and R. Mudambi, "Competition and competitiveness in the US airline industry," *Competitiveness Review*, vol. 25, no. 2, pp. 134–155, 2015.

[3] X. S. Li, "Research on evaluation of the competitiveness of domestic and foreign airlines based on matter-element model," *Communications in Computer & Information Science*, vol. 268, 2012.

[4] I. Vlachos and Z. Lin, "Drivers of airline loyalty: evidence from the business travelers in China," *Transportation Research Part E: Logistics and Transportation Review*, vol. 71, pp. 1–17, 2014.

[5] R. Merkert and J. Pearson, "A non-parametric efficiency measure incorporating perceived airline service levels and profitability," *Journal of Transport Economics and Policy*, vol. 49, pp. 261–275, 2015.

[6] W. Zhen and L. I. Xue-Gong, "Evaluation and research of port competitiveness based on FAHP," *Port & Waterway Engineering*, vol. 5, pp. 964–971, 2011.

[7] H. S. Jenatabadi and N. A. Ismail, "Application of structural equation modelling for estimating airline performance," *Journal of Air Transport Management*, vol. 40, pp. 25–33, 2014.

[8] S.-J. Joo and K. Fowler, "Exploring comparative efficiency and determinants of efficiency for major world airlines," *Benchmarking: An International Journal*, vol. 21, no. 4, pp. 675–687, 2014.

[9] C. Wu, X.-Y. Zhang, I.-C. Yeh, F.-Y. Chen, J. Bender, and T.-N. Wang, "Evaluating competitiveness using fuzzy analytic hierarchy process-A case study of Chinese airlines," *Journal of Advanced Transportation*, vol. 47, no. 7, pp. 619–634, 2013.

[10] T. Fischer and D. R. Kamerschen, "Measuring competition in the U.S. Airline industry using the rosse-panzar test and cross-sectional regression analyses," *Journal of Applied Economics*, vol. 6, no. 1, pp. 73–93, 2003.

[11] S. Borenstein, "An index of inter-city business travel for use in domestic airline competition analysis," in *Proceedings of the NBER Working Paper*, Cambridge, MA, USA, 2010.

[12] J. K. Brueckner, D. Lee, and E. S. Singer, "Airline competition and domestic US airfares: a comprehensive reappraisal," *Economics of Transportation*, vol. 2, no. 1, pp. 1–17, 2013.

[13] Y.-H. Chang and C.-H. Yeh, "Evaluating airline competitiveness using multiattribute decision making," *Omega*, vol. 29, no. 5, pp. 405–415, 2001.

[14] T. H. Oum and C. Yu, "Cost competitiveness of major airlines: an international comparison," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 6, pp. 407–422, 1998.

[15] P. J. G. Pineda, J. J. Liou, C. C. Hsu, and Y. C. Chuang, "An integrated MCDM model for improving airline operational and financial performance," *Journal of Air Transport Management*, vol. 68, pp. 103–117, 2018.

[16] J. Raheleh and Y. Wen, "Fuzzy modeling for uncertainty nonlinear systems with fuzzy equations," *Mathematical Problems in Engineering*, vol. 2017, pp. 1–10, 2017.

[17] L. Moir and G. Lohmann, "A quantitative means of comparing competitive advantage among airlines with heterogeneous business models: analysis of U.S. airlines," *Journal of Air Transport Management*, vol. 69, pp. 72–82, 2018.

[18] K. Yeh and C. W. Chen, "Stability analysis of interconnected fuzzy systems using the fuzzy lyapunov method," *Mathematical Problems in Engineering*, vol. 2010, pp. 23.1–23.10, 2010.

[19] H. A. Linstone and M. Turoff, *The Delphi Method: Techniques and Applications*, Addison-Wesley, Reading, MA, USA, 2002.

[20] T. L. Saaty, "Decision making with the analytic hierarchy process," *International Journal of Services Sciences*, vol. 1, no. 1, p. 83, 2008.

[21] S. A. Delbari, S. I. Ng, Y. A. Aziz, and J. A. Ho, "An investigation of key competitiveness indicators and drivers of full-service airlines using Delphi and AHP techniques," *Journal of Air Transport Management*, vol. 52, pp. 23–34, 2016.

[22] J. Yu and Y. Li, "Multi-level fuzzy comprehensive evaluation method for airline competitiveness," *Journal of Transportation Engineering*, vol. 8, no. 3, pp. 116–121, 2008, in Chinese.

[23] P. Fu and C. Wu, "Research on competitive evaluation of domestic airlines based on FAHP method," *Journal of Suzhou University*, vol. 1, no. 2, pp. 131–136, 2011, in Chinese.

[24] Y. Li, J. Yu, and Y. Wu, "Evaluation and empirical study of airline competitiveness," *Journal of Beijing Institute of Technology (Social Science Edition)*, vol. 11, no. 4, pp. 49–53, 2009, in Chinese.

[25] W. Li and C. Wu, "Research on airline competitiveness based on secondary relative evaluation," *Journal of Wuhan University of Technology (Information and Management Engineering Edition)*, vol. 1, no. 10, pp. 846–850, 2011, in Chinese.

[26] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European Journal of Operational Research*, vol. 2, no. 6, pp. 429–444, 1978.

[27] P. Andersen and N. C. Petersen, "A procedure for ranking efficient units in data envelopment analysis," *Management Science*, vol. 39, no. 10, pp. 1261–1264, 1993.

[28] B. Marian, "Gorzałczany and filip rudziński, evolution of SOMs' structure and learning algorithm: from visualization of high-dimensional data to clustering of complex data," *Algorithms*, vol. 13, no. 5, p. 109, 2020.

[29] H. Simon, *Neural Network Principle*, China Machine Press, Beijing, China, 2004.

[30] T. Kohinen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 4, no. 3, pp. 59–69, 1982.

[31] M. Y. Kiang, "Extending the Kohonen self-organizing map networks for clustering analysis," *Computational Statistics & Data Analysis*, vol. 38, no. 2, pp. 161–180, 2001.

[32] J. Zhang, "Transformer fault diagnosis based on SOM," *Journal of Electric Power*, vol. 29, no. 4, pp. 318–321, 2014.

[33] Y. S. Qian, M. Wang, H. X. Kang et al., "Study on the road network connectivity reliability of valley city based on complex network," *Mathematical Problems in Engineering*, vol. 2012, no. 9, pp. 430785.1–430785.14, 2012.

[34] Y. Liu and L. Chen, "Mechanical fault diagnosis of vacuum circuit breaker based on som," *Transaction of China Electrotechnical Society*, vol. 32, no. 5, pp. 49–54, 2017.

[35] K. Xie, Y. Yang, Y. Xin et al., "Cellular neural network-based methods for distributed network intrusion detection," *Mathematical Problems in Engineering*, vol. 2015, no. 3, pp. 343050.1–343050.10, 2015.

[36] Z. Duan, K. Zhang, Z. Chen et al., "Prediction of city-scale dynamic taxi origin-destination flows using a hybrid deep neural network combined with travel time," *IEEE Access*, vol. 7, no. 99, 2019.

[37] L. Lei, W. Shi, and M. Fan, "Water quality evaluation analysis based on improved SOM neural network," *Chinese Journal of Scientific Instrument*, vol. 30, no. 11, pp. 2379–2383, 2009.

[38] Y. Du and Q. Tian, "Performance evaluation for mechanical products based on fuzzy neural network," *Systems Engineering and Electronics*, vol. 27, no. 9, pp. 1583–1586, 2005.

[39] J. Zhao and X. Chen, "Evaluation and analysis of airlines' operation and service quality based on grey system theory," in *Proceedings of the 2020 IEEE International Conference on Civil Aviation Safety and Information Technology(ICCASIT 2020)*, Weihai, China, 2020.

[40] X. Xie, *Domestic Airline Competitiveness Evaluation Evaluation Based on Principal Component Analysis*, Harbin Institute of Technology, Harbin, China, 2011, in Chinese.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9765975, 14 pages
https://doi.org/10.1155/2020/9765975

Research Article

Detection for Multisatellite Downlink Signal Based on Generative Adversarial Neural Network

Qing-yang Guan [1,2] **and Wu Shuang** [1]

[1]*College of Engineering, Xi'an International University, Xi'an 710077, China*
[2]*College of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China*

Correspondence should be addressed to Qing-yang Guan; gqy_gqy@163.com

Received 31 May 2020; Revised 11 July 2020; Accepted 23 July 2020; Published 12 August 2020

Guest Editor: Jun Shen

A method for satellite downlink signal detection based on a generative adversarial network is proposed. The generator adversarial network and adversarial network are established, respectively. The generator network realizes the local generator of satellite signals, and the adversarial network is used for high-precision signal detection. The error network is generated by the error signal to form the satellite link downlink. The network reconstructs the optimal weights by generating errors, forms an error matrix for different satellite downlink, and then forms an adaptive matrix weight adjustment. Through the reconstruction of the optimal detection matrix, detection for the downlink signals of multiple satellites is completed. The proposed generative adversarial network can realize the high-precision detection for the downlink signal.

1. Introduction

1.1. Background Introduction. Due to the rapid development of AI technology, technologies are used in signal detection of the satellite to ground links. For the typical satellite to ground downlink, due to the low-orbit satellite, the relatively large-scale movement introduces a Doppler frequency shift. Doppler frequency will occur between the users and the low-orbit satellites. Downlink for low-orbit satellite mobile communication system is sensitive to carrier frequency offset, which seriously destroys the orthogonality between subcarriers and causes distortion to the receiver. A broadband signal detection technology is developed from AI technology. AI technology based on deep learning is applied in various fields, such as inserting and extracting knowledge [1], speech recognition [2], and language translation. Since deep learning has sufficient hidden layers, it can accurately simulate the target function and optimization direction [3]. In [4, 5], sparse coding is applied for data detection, including spectrum analysis of communication systems [6] and channel estimation [7]. Sparse coding has the following two important features, which are suitable for selective frequency channel and multiuser detection. Chen et al. [8] established a deep neural network, which uses fewer data to complete labeling and training and then optimizes the deep neural network by reverse tuning.

For communication systems, Gaur and Ingram [9] proposed a simple MMSE Interference Suppression for Real and Rate-1/2 Complex Orthogonal Space-time Block Codes. In [10–13], the multiuser detection methods were proposed, especially for massive MIMO systems. In [14, 15], data encoding method was proposed, which improved accuracy through precoding. Çelebi and H. Arslan [16] proposed the ML-SIC receiver for the theoretical analysis of the coexistence of the LTE-A system. Rusek et al. [17] proposed a broadband mass MU-MIMO detection based on the ZF equalization algorithm, which uses interpolation to improve accuracy. But it requires large-scale matrix operations to increase the complexity. Ahn et al. [18] proposed the Sparsity-Aware Ordered Successive Interference Cancellation for Massive Machine-Type Communications.

For the research of deep learning neural networks, Ghamisi et al. [19] proposed a classification algorithm to extract features and classify data. Then, the algorithm forms

classification templates to improve classification accuracy. A kernel-space algorithm is proposed in [20], which establishes a transformed network in order to further improve the data classification method.

Yuan et al. [21] combined sparse coding and Markov random fields to establish the classification network based on spatial data correlation, which is to improve classification accuracy. On this basis, Wang et al. [22] propose the subspace analysis method to further the accurate classification performance of the network.

For data classification and detection, the above methods do not use deep networks. Bengio et al. [23] proved that only classification methods such as SVM or logistic regression cannot effectively improve classification, such as decision trees or kernel-space transformation. The single-layer classification method could not obtain better classification accuracy. Depth models, including multiple hidden layers, are identified through acquiring target data features. Deep learning models are widely used in related research fields, such as image classification and speech recognition [2, 24]. Zhang et al. [25] also gave a summary analysis of deep learning. In 2014, a deep learning architecture based on SAE encoding was proposed for data classification [26]. Later, a deep learning architecture based on DBN was used for HSI data classification [27, 28].

Deep learning networks, such as deep CNN networks [29], were used in image recognition and classification. Li et al. [30] proposed a novel pixel-pair method for the classifier, which could use less training data and label data. A deep learning architecture based on CNN was proposed in [31, 32]. For example, Yue et al. [32] proposed a PCA-based analysis method. In [33, 34], deep learning architecture was proposed, combined with sparse coding and Gabor filter for deep feature extraction. On this basis, authors in [35, 36] proposed to use the CNN architecture deep learning network for classification and recognition for hyperspectral imagery.

1.2. Reasons for Proposed Algorithm. The biggest advantage of deep learning is that it can realize the complex nonlinear function approximation of massive data through nonlinear network architecture, then characterize the distribution, and form the ability to learn the essence of data features.

If the features of the data change or the types of data expand, the ability for deep learning describing massive data becomes weaker. For multisatellite downlink signal detection, the multisatellite downlink channel variety is complex. For multisatellite downlink service scenarios, the model established by deep learning is not universal, and the model of each communication type changes. So the number of variables that deep learning can provide is limited, and the number of layers of the deep network is also limited.

Secondly, deep learning requires excessively high-quality training data. The accuracy of data analysis increases as the training data increase. In satellite downlink, high-quality training data cannot be obtained in many communication scenarios, so poor quality data could not be formed for deep learning to obtain a general effective model.

How to use a small amount of training data and establish reliable multisatellite downlink detection in different scenarios is important. In this paper, we use the GAN network to overcome this difficulty in satellite downlink. Zhang et al. [37] proposed a GAN network based on game theory, which consists of the generator network and the discriminator network. The generator network is used to generate data to improve the recognition accuracy, and the discriminator network is used to distinguish the target. Wang et al. [38] adopted the BP method to train the generating network and the discriminating network separately.

GAN network is a model that includes the G generator network part and the D network part [39]. The generator network and the discriminator network part are formed in an adversarial manner. Given the advantages of adversarial networks, Chen et al. [40] applied adversarial networks to image data. Creswell et al. [41] applied GAN network to data analysis, feature extraction, and classification.

In order to improve the performance of the GAN network, in [42, 43], improved modes were proposed. These improved modes are divided into two categories, one based on structural optimization and the other based on objective optimization. Sun et al. [42] proposed the generative adversarial network, which introduced variables to improve the efficiency of GAN games. Based on generative networks, in addition, Yu et al. [43] proposed conditional information adversarial networks based on mutual information to improve the efficiency of generating networks.

In the optimization process, in [40, 44–46], the coding part for the GAN network was added. The coding part could improve the accuracy and the efficiency as the whole and improve the GAN network objective function optimization by adding a label classifier. In order to improve the convergence of the GAN network, in [47–51], an optimized objective function was proposed to improve the training process of the GAN network. Among them, authors in [47–49] used different models to achieve the loss objective function.

For multisatellite downlink signal detection, high-quality training data cannot be obtained in many downlink scenarios. The models established by traditional deep learning are not universal, so the model of the downlink data type varies with different channel models. But deep learning requires too much quality for training data. Therefore, we propose the GAN network. The main goal is using few data training information to form through adversarial generation network under poor channel conditions and then to achieve efficient and high-precision signal detection.

2. System Model and Problem Formulation

2.1. System for Satellite Downlink. Figure 1 shows the multisatellite downlink transmission network. The satellite network is divided into the space segment and the ground segment. The space segment is composed of multiple satellites, and the ground segment is composed of ground users. The space segment communicates with the ground segment through the downlink channel.

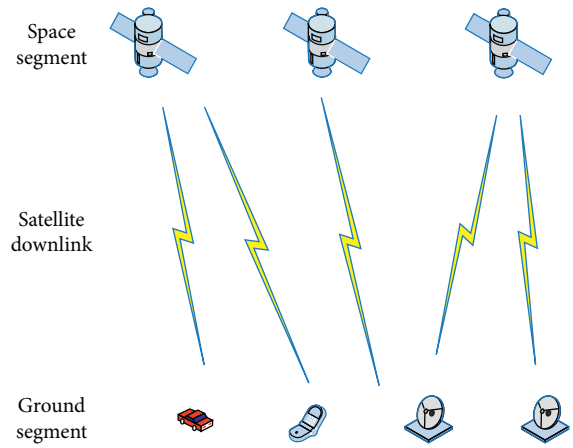


Figure 1: Diagram for multisatellite downlink transmission network.

Due to the high-speed movement, define multipath propagation delay as τ. The time channel impulse response function for frequency-selective mobile satellite channel is $h(t,\tau)$. For different τ, $h(t,\tau)$ is not related to each other. For definite delay τ, $h(t,\tau)$ is an average complex Gaussian random process. Simultaneously, impulse response $h(t,\tau)$ is according to the flat fading characteristics. Therefore, the time-varying impulse response of L multichannel can be expressed as follows:

$$h(t,\tau) = \sum_{l=0}^{L-1} b_l v_l(t)\delta(\tau - \tau_l), \tag{1}$$

where τ_l is the transmission delay of the lth path. $v_l(t)$ is the complex Gaussian process, which is the weight component of the lth path; the power spectrum is the Doppler power spectrum of the lth path, which also can represent the fading rate of the lth path. b_l is the time delay coefficient, and the value is the square root of the average delay power for the lth path, which can be expressed as the weighting of unresolved paths arriving at different incidence angles within a certain delay interval.

$d(t)$ and $r(t)$ represent the system input and output, respectively, so that the channel can be expressed as follows:

$$r(t) = \sum_{l=0}^{L-1} b_l v_l(t) d(\tau - \tau_l). \tag{2}$$

The amplitude $b_0(t)$ of the direct path can be considered as Rician distribution, while the amplitude of other paths can be considered as $b_l(t)$, and $l = 1, 2, 3, \ldots, L-1$ can be considered as Rayleigh distribution. The downlink satellite channel simulation model proposed in this paper is essentially based on Rician and Rayleigh fading signals with specific Doppler spectrum. Simulation for Rician and Rayleigh processes requires two Gaussian processes. Sinusoidal superposition is generally used to simulate these Gaussian processes.

$v_l(t)$ is defined as the complex Gaussian process, which is satisfied as follows:

$$v_l(t) = v_{r,l}(t) + jv_{i,l}(t), \tag{3}$$

where $v_{r,l}(t)$ is the real part and $v_{i,l}(t)$ is the imaginary part, which are independent of each other and have the same mean and autocorrelation function. $v_{i,l}(t)$ can be expressed as follows:

$$v_{i,l}(t) = \sum_{n=1}^{K_{i,l}} c_{i,n,l} \cos\left(2\pi f_{i,n,l} t + \theta_{i,n,l}\right), \tag{4}$$

where $K_{i,l}$ is the number of sine waves of the lth path, $f_{i,n,l}$ is the nth Doppler shift of the lth path, $\theta_{i,n,l}$ is the nth Doppler phase of the lth path, and $c_{i,n,l}$ is the nth Doppler coefficient of the lth path.

$\theta_{i,n,l}$ is uniformly distributed within the interval $[0, 2\pi)$; get it by taking the random number in. $c_{i,n,l}$ and $f_{i,n,l}$ can be calculated using the MEA method and the Monte Carlo method.

In the tapped delay line method to establish a channel model of a satellite-ground link, the multipath channel impulse response is composed of multiple paths with different delay characteristics, and each path has specific signal amplitude fading and power spectrum characteristics. The proposed model is measured with broadband satellite downlink channels in the wilderness, rural, and urban environments, with the signal carrier frequency of 1.02 GHz. Tables 1–3 show the measured channel characteristic parameters of the wilderness environment, rural environment, and urban environment, respectively.

Figures 2–7 show the signal amplitude and Doppler power spectrum for each tap of the direct component for satellite mobile channel in the wilderness environment, rural environment, and urban environment, respectively, established by using tapped delay lines.

Due to different satellite simulation, multipath delays of the channels are different, and the signal amplitude fading is completely different. The first path of each simulation scenario has a large impulse response corresponding to the amplitude. This is because the first path has a direct component, and its envelope corresponds to the Rician channel density distribution. For rural simulation scenarios, the reflection of signals from more buildings and the diffraction effect cause a larger number of multipaths, and the extended delay of the received signal is larger, which makes the signal get more severe fading. For the rural simulation scene, compared with the urban scene, the number of buildings is small. The reflection and refraction phenomena are reduced compared to the urban simulation environment, so the number of multipaths is reduced, and the channel fading is flat compared to the urban environment. For urban simulation, because the number of buildings is smaller than the number of urban simulation, and the number of vegetation is reduced compared to the rural environment, the reflection and refraction phenomena are reduced compared to urban and rural areas, and channel fading is also more difficult than urban environments.

Table 1: Channel model parameter in the wilderness environment.

Tap	Distribution function	Parameter	Parameter distribution	Value (dB)	Delay (ns)
1	LOS Rician	Rician factor	K	6.3	0
	NLOS Rayleigh	Average multipath power	$2\sigma_l^2$	−9.5	
2	Rayleigh	Average multipath power	$2\sigma_l^2$	−24.1	100

Table 2: Channel model parameter in the rural environment.

Tap	Distribution function	Parameter	Parameter distribution	Value (dB)	Delay (ns)
1	LOS Rician	Rice factor	K	5.3	0
	NLOS Rayleigh	Average multipath power	$2\sigma_l^2$	−12.1	
2	Rayleigh	Average multipath power	$2\sigma_l^2$	−17.0	60
3	Rayleigh	Average multipath power	$2\sigma_l^2$	−18.3	100
4	Rayleigh	Average multipath power	$2\sigma_l^2$	−19.1	130
5	Rayleigh	Average multipath power	$2\sigma_l^2$	-22.1	250

Table 3: Channel model parameter in the urban environment.

Tap	Distribution function	Parameter	Parameter distribution	Value (dB)	Delay (ns)
1	LOS Rician	Rician factor	K	9.7	0
	NLOS Rayleigh	Average multipath power	$2\sigma_l^2$	−7.3	
2	Rayleigh	Average multipath power	$2\sigma_l^2$	−17.6	30
3	Rayleigh	Average multipath power	$2\sigma_l^2$	−18.3	180
4	Rayleigh	Average multipath power	$2\sigma_l^2$	−19.3	60
5	Rayleigh	Average multipath power	$2\sigma_l^2$	−22.1	100
6	Rayleigh	Average multipath power	$2\sigma_l^2$	−25.3	190
7	Rayleigh	Average multipath power	$2\sigma_l^2$	−28.1	250
8	Rayleigh	Average multipath power	$2\sigma_l^2$	−29.1	270

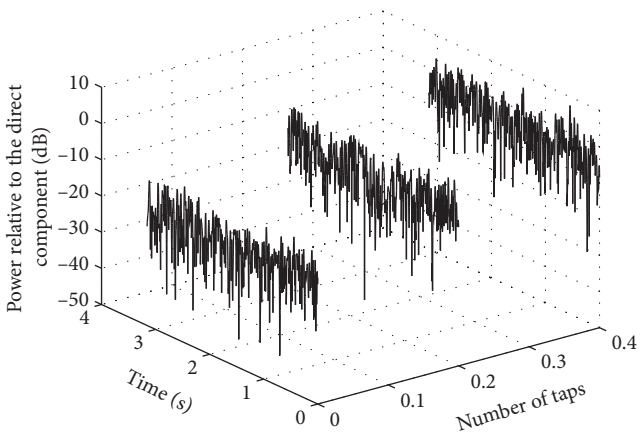


Figure 2: Signal amplitude for each tap (wilderness environment).

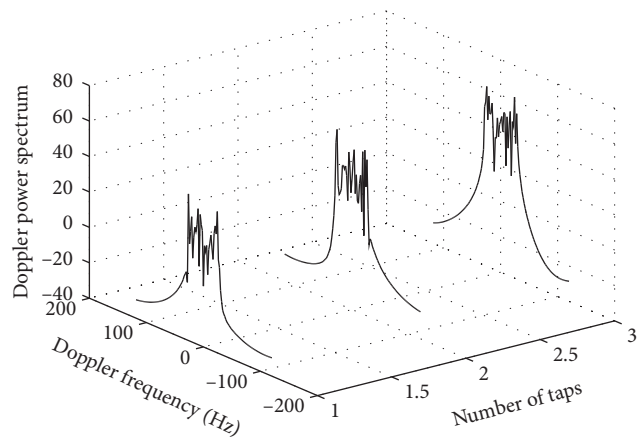


Figure 3: Doppler power spectrum for each tap (wilderness environment).

2.2. Problem Formulation for Detection of Satellite Downlink. Figure 8 shows a block diagram of the multisatellite for the downlink communication system. The subcarrier allocation method uses IFDMA; after the user obtains the carrier allocation, the subcarriers of each user do not overlap with each other. And the satellite receives superimposed interference. It is assumed that the signals of each user can be separated by the carrier mapping of the signal. Due to the high-speed movement, the carrier frequency offset is introduced into the transmission signal, and its value depends on the speed of the satellite and the maximum elevation angle.

Because of receiving multiple satellite signals, the relative frequency offset factors of satellites are also different. When the number of system subcarriers is determined, access interference will be introduced due to differences of multiple satellites. Figure 8 shows a block diagram of multisatellite for the downlink DFT-S OFDM system. The subcarrier allocation method uses IFDMA; after allocating the carrier, the subcarriers of each user do not overlap with each other. The satellite receives each user with overlapping interference. It is assumed that the multisatellite signal can be separated by the carrier mapping of the signal.

Define the relative frequency offset factor ξ^i of the ith satellite, and the number of satellites received by the downlink ground user is N. The frequency domain signal is modulated by N orthogonal subcarriers, and the time domain signal of multiple satellites received by the ground user can be expressed as follows:

$$y_i(m) = \frac{1}{N}\sum_{k=0}^{N-1} S(k)\sum_{l=0}^{L-1} h(m,l)\exp\left(j2\pi k\frac{(m-l)}{N}\right) \cdot \exp\left(j2\pi\xi^i\frac{m}{N}\right) + z(k), \tag{5}$$

where $z(k)$ is. The receiver performs N-point modulation conversion; it can be obtained as

$$Y_i(k') = \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=0}^{N-1} S(k)H(k)\exp\left(j2\pi m\frac{k}{N}\right) \cdot \exp\left(j2\pi\xi^i\frac{m}{N}\right)\cdot \exp\left(-j2\pi m\frac{k'}{N}\right) + Z(k'), \tag{6}$$

where the ground user receives the signal of the ith satellite with interference $Y_i(k\prime)$, which can be obtained by separating the interference items:

$$\begin{aligned} Y_i(k') &= \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=0}^{N-1} S(k)H(k)\exp\left(j2\pi m\frac{k}{N}\right)\exp\left(j2\pi\xi^i\frac{m}{N}\right) \\ &\quad\cdot \exp\left(-j2\pi m\frac{k'}{N}\right) + Z(k'), \\ &= \frac{1}{N}\sum_{m=0}^{N-1}\sum_{k=0}^{N-1} S(k')H(k')\exp\left(j2\pi\xi^i\frac{m}{N}\right) \\ &\quad + \sum_{m=0}^{N-1}\sum_{\substack{k=0,\\ k\neq k'}}^{N-1} X(k)H(k)\exp\left(j2\pi m\frac{(k-k')}{N}\right) \\ &\quad\cdot \exp\left(j2\pi\xi^i\frac{m}{N}\right) + Z(k'), \quad k' = 0,1,\ldots,N-1. \end{aligned} \tag{7}$$

The first term is the interference between communication symbols, the second term is the interference introduced by communication access, and the third term is the interference introduced by Gaussian white noise. The multiple satellite signals received by the downlink ground user can be expressed as follows:

$$\begin{aligned} Y(k\prime) &= \frac{1}{N}\sum_{i=1}^{T-1}\sum_{m=0}^{N-1}\sum_{k=0}^{N-1} S(k)H(k)\exp\left(j2\pi m\frac{k}{N}\right)\exp\left(j2\pi\xi^i\frac{m}{N}\right)\cdot \exp\left(-j2\pi m\frac{k'}{N}\right) + Z(k\prime), \\ &= \frac{1}{N}\sum_{m=0}^{N-1}\sum_{k=0}^{N-1} S(k')H(k')\exp\left(j2\pi\xi^i\frac{m}{N}\right) + \sum_{\substack{i\in T\\ j\neq i}}\sum_{m=0}^{N-1}\sum_{\substack{k=0,\\ k\neq k'}}^{N-1} S(k)H(k)\exp\left(j2\pi m\frac{(k-k')}{N}\right)\exp\left(j2\pi\xi^i\frac{m}{N}\right) \\ &\quad + \sum_{\substack{j\in T\\ j\neq i}}\sum_{m=0}^{N-1}\sum_{\substack{k=0,\\ l\neq k'}}^{N-1} X(l)H(l)\exp\left(j2\pi m\frac{(l-k')}{N}\right)\exp\left(j2\pi\xi^j\frac{m}{N}\right) + Z(k'), \quad k' = 0,1,\ldots,M-1, \end{aligned} \tag{8}$$

where $Y(k\prime)$ is the satellite signals received by ground users. The last term in the above formula is the reception interference introduced by receiving multiple satellites.

3. Generative Adversarial Network Algorithm for Satellite Signal Detection

3.1. Architecture of GAN Network for Satellite Signal Detection. Figure 9 gives the information processing flow based on the GAN algorithm. The generator network realizes the local signals, and the discriminator network is used for signal detection.

We define the generalized loss function model as follows:

$$\begin{aligned} &\min \quad \|Y(k) - WS(k)\|_F^2 + \alpha f(W), \\ &\text{s.t.} \quad \left\|W_j\right\|_2^2 \le 1,\ \forall j, \end{aligned} \tag{9}$$

with adversarial network generator; we could achieve to obtain $\alpha f(W)$.

In order to obtain the signal detection of the satellite downlink, we implement the optimal weight W by introducing the GAN network. The following discussion is aimed at using the adversarial network to obtain the optimal

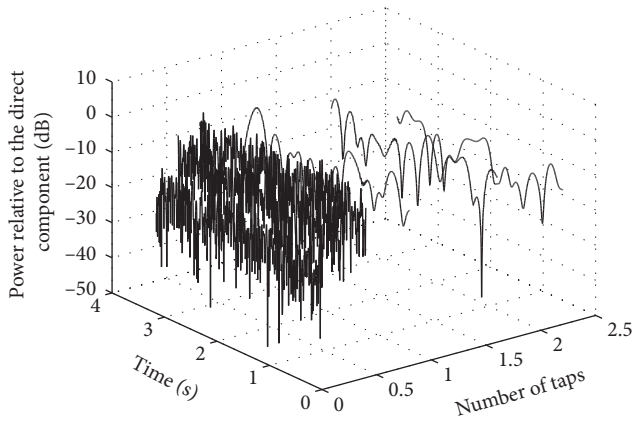


Figure 4: Signal amplitude for each tap (rural environment).

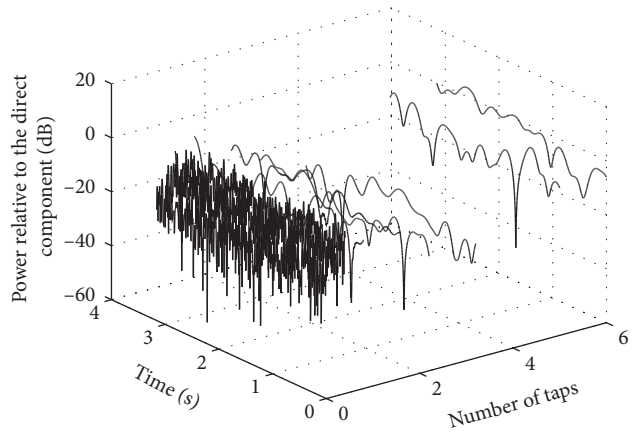


Figure 6: Signal amplitude for each tap (urban environment).

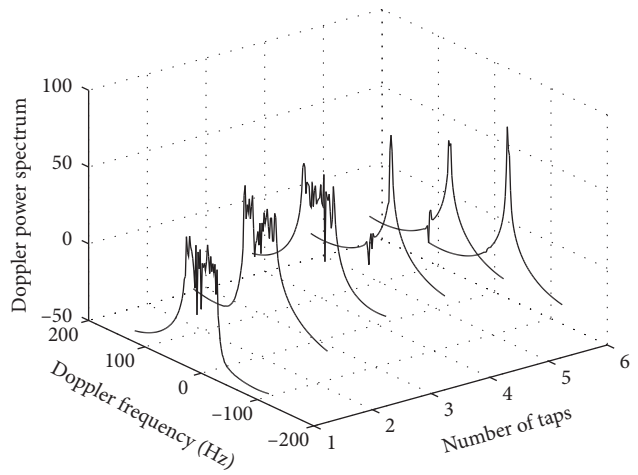


Figure 5: Doppler power spectrum for each tap (rural environment).

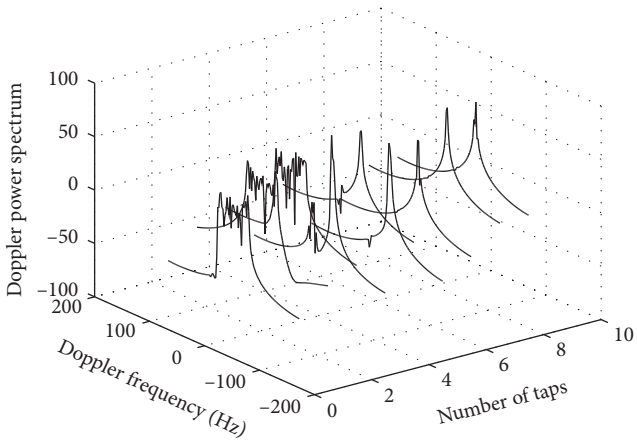


Figure 7: Doppler power spectrum for each tap (urban environment).

elimination matrix. In order to achieve optimal weights, we established the loss function of GAN network.

We introduce a generative adversarial network loss cost function to identify multisatellite downlink data. The cost function of the GAN can be expressed as follows:

$$L_{\mathrm{GAN}} = E_{x\sim p_t}[\log D(s)] + E_{x\sim p_t}[\log(1 - D(G(s)))]. \tag{10}$$

From formula (10), we could obtain GAN which consists of a generator network and a discriminator network, where $G(s)$ is defined as generator network and $D(s)$ is defined as discriminator network.

The generator network $G(s)$ produces the output as the target domain as shown in Figure 10. However, combined with a wide range of capacity, the target is to minimize adversarial losses. It can ensure that the learning network is directed to the input s which corresponds to the ideal output.

Figure 10 gives the adversarial network model for satellite signal processing flow. In order to further reduce the GAN function, we propose a reconstruction of generating error weights W. In order to reduce the number of iterations of reconstructed features, a method of closely cost function is proposed. Then, we propose an iterative cost function L_{GAN}:

$$L_{\mathrm{GAN}} = L_G + L_D, \tag{11}$$

where L_G represents the cost function for the generator network. L_D represents the cost function for the discriminator network. The generator network includes forward and backward features. That is to say, the separately generated network can be reconstructed according to the original generated data and then form bidirectional iteration.

Figure 11 gives the signal processing flow based on the adversarial network. From Figure 11, we also give the processing for GAN network establishment. The goal is to optimize the loss function, including two parts. That is G network, and the second is the D network. In the following, we establish the generative network and the confrontation network separately.

3.2. Establishment for Generator Network. The process of generating regularization terms is as follows. Define $\{S_i, Y_j\} = \{(s_1, y_1), (s_2, y_2), \ldots, (s_i, y_j)\}$ as the generation set, where s_i is the data feature after the ith generation, s_j is

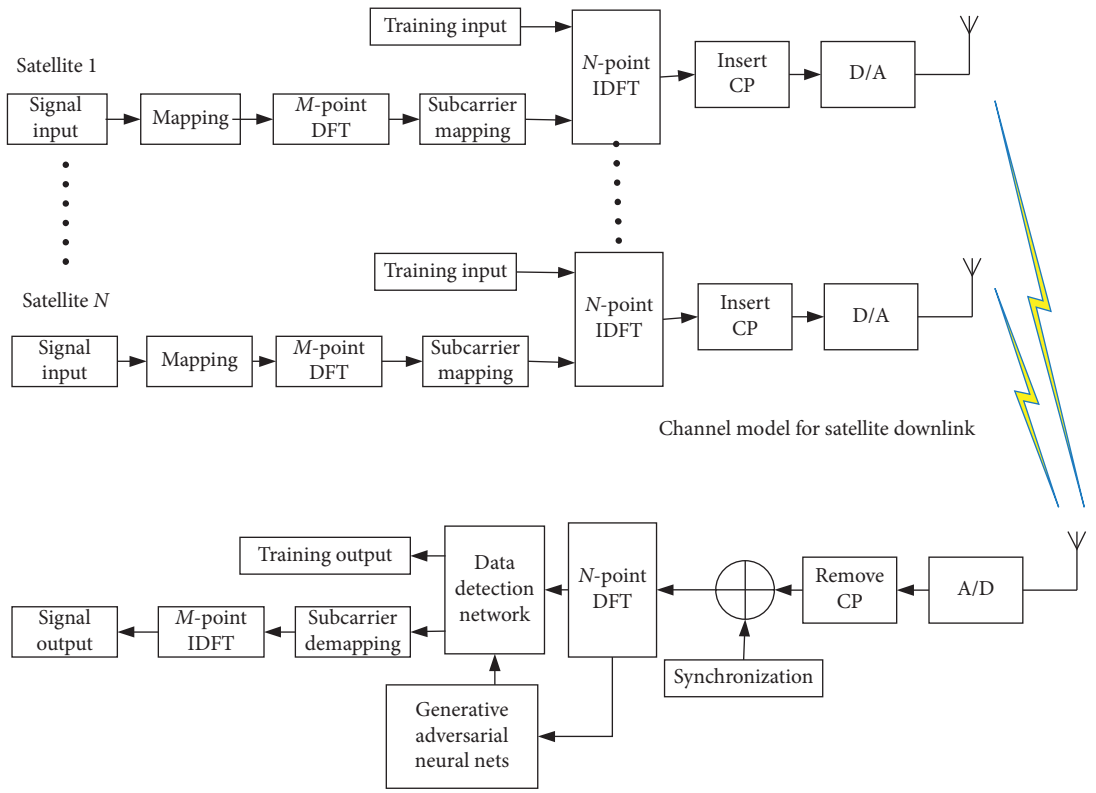


Figure 8: Multisatellite model for the downlink communication system.

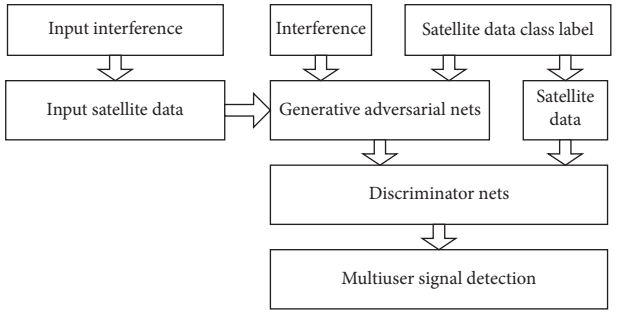


Figure 9: Information processing flow based on the GAN algorithm.

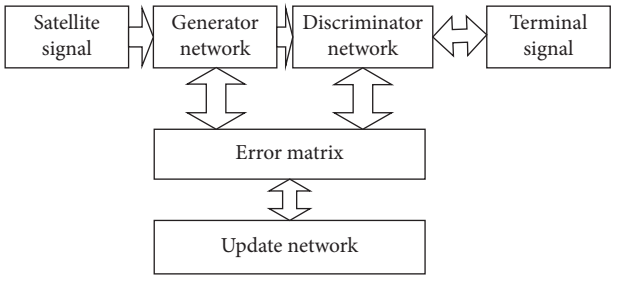


Figure 10: GAN network model information processing flow for the satellite signal.

the data feature after the jth generation, and y is the expected detection data. We establish a positive multidimensional space distance M between the feature space and the generating space, so we can classify the data in the target space. The spatial distance is formed in different target classifications.

The first generator network is also satisfied for formula (11), and we define the generator network loss L_G as follows:

$$L_G = \min_M \sum_{(x_i,x_j)\in S} \left\| s_i - s_j \right\|_M^2. \tag{12}$$

The error of the generator network is satisfied, which is to satisfy

$$\min_M \sum_{(x_i,x_j)\in D} \left\| s_i - s_j \right\|_M^2 \geq 1. \tag{13}$$

The spatial distance can be expressed as follows:

$$L_G = \min_M \sqrt{\left(s_i - s_j\right)^T M \left(s_i - s_j\right)}. \tag{14}$$

Further simplification is as follows:

$$L_G = \min_M \sqrt{\left(s_i - s_j\right)^T W W^T \left(s_i - s_j\right)}, \tag{15}$$

where W is the weight matrix of the GAN network.

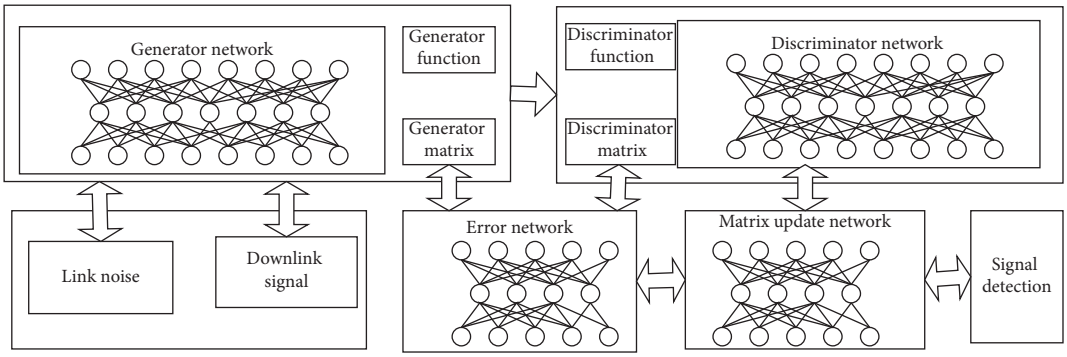


Figure 11: Signal processing flow based on the adversarial network.

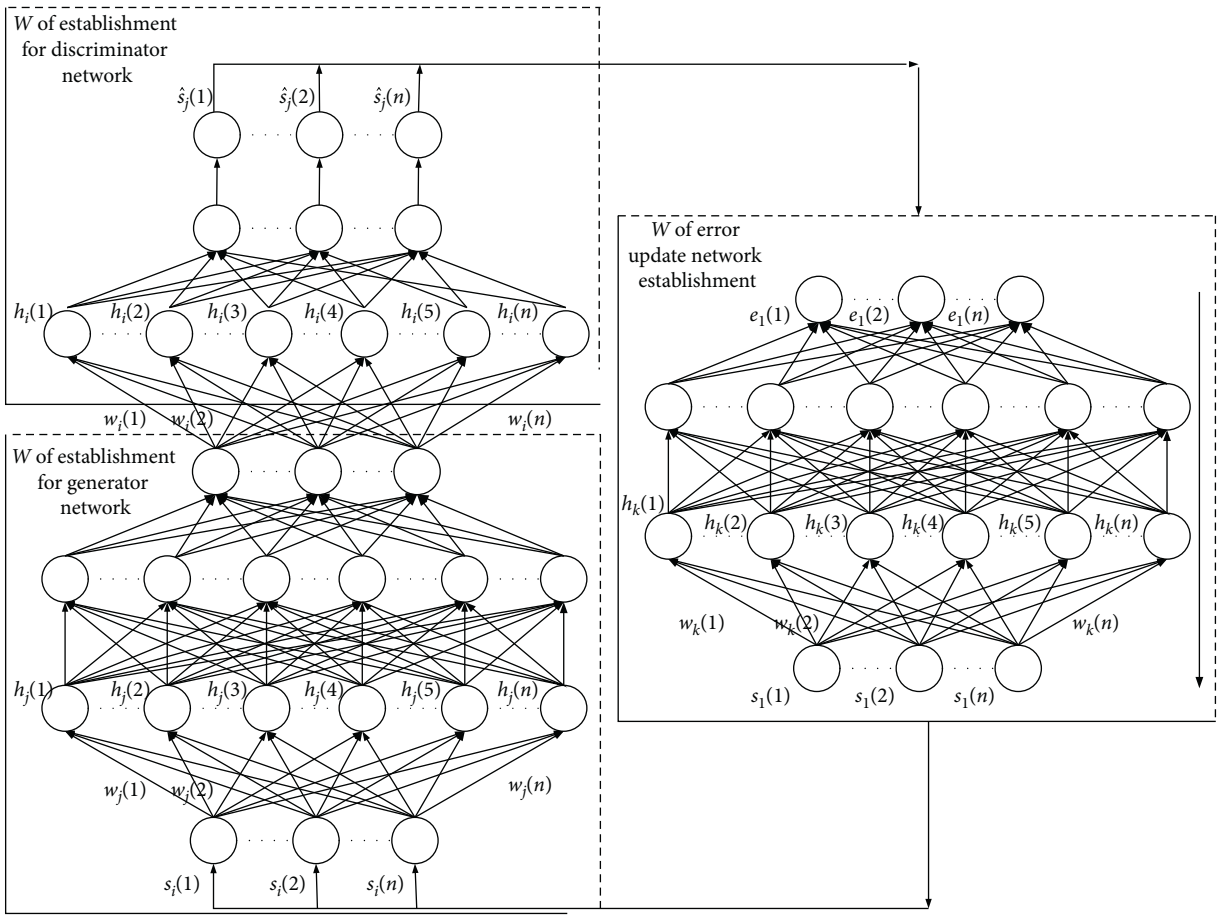


Figure 12: Flow for weight W establishment of the GAN network.

3.3. Establishment for Discriminator Network. Establish the extraction direction of the discriminator feature space, and we could obtain as follows:

$$L_{\mathrm{GAN}} = L_G - \alpha L_D, \tag{16}$$

where L_G represents the cost function for the generator network and L_D represents the cost function for the discriminator network. We also define the elimination coefficients α for generator and adversarial networks:

$$\alpha = \frac{\mathrm{tr}\left(W^T S L_s S^T W\right)}{\mathrm{tr}\left(W^T S L_D S^T W\right)}. \tag{17}$$

Further, the weight can be obtained as follows through GAN network:

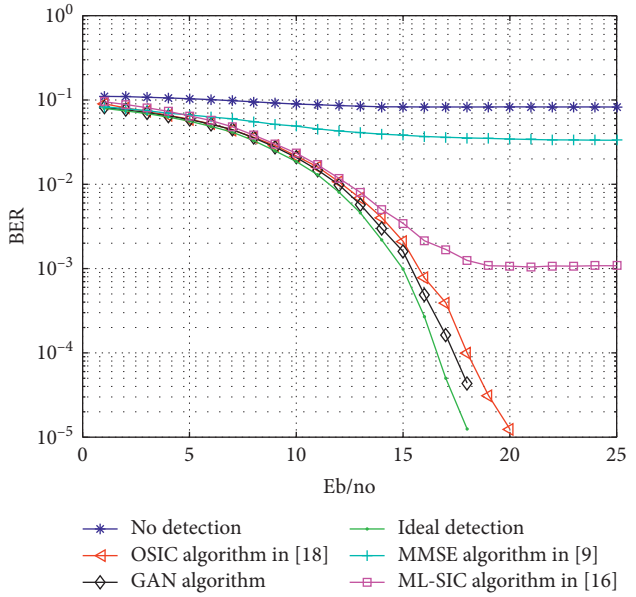


Figure 13: BER performance of QPSK signal in the rural environment channel model.

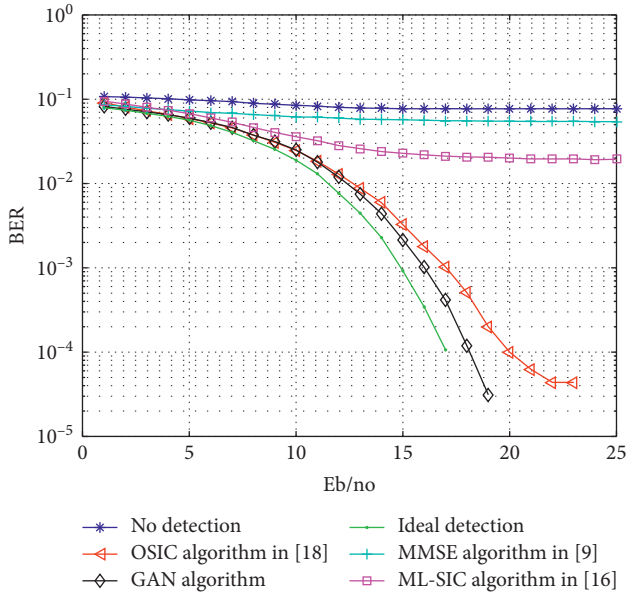


Figure 14: BER performance of QPSK in the urban environment channel model.

$$W = \arg\min_{W^TW=1} \mathrm{tr}\left(W^T S(L_{\mathrm{GAN}})S^T W\right). \tag{18}$$

3.4. Establishment for Error Update Network. Furthermore, we establish an error network as follows:

$$W = \arg\min_{W^TW=1} \mathrm{tr}\left(W^T S(L_s - \alpha L_D)S^T W\right), \tag{19}$$

where $\mathrm{tr}(\cdot)$ represents the operation of taking matrix traces, in which α is defined as the distance measurement parameter

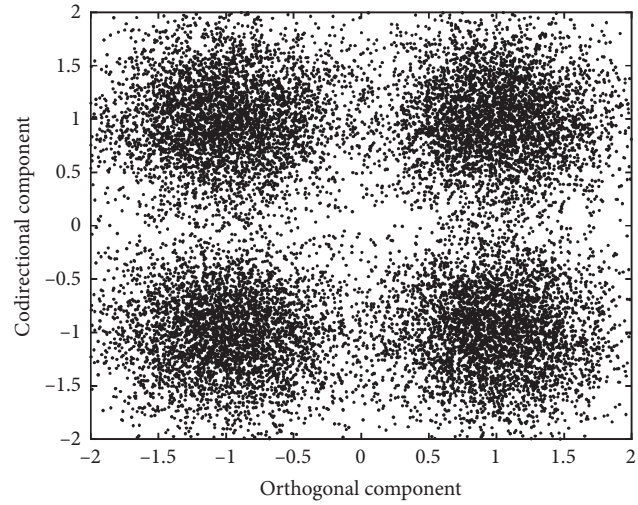


Figure 15: QPSK signal constellation (after the MMSE algorithm in [9]).

of the two adversarial matrix feature spaces, expressed as follows.

Through extensibility analysis, further calculations can be obtained:

$$\begin{aligned} W &= \arg\min_{W^TW=1} \mathrm{tr}\left(W^T S(L_s - \alpha L_D)S^T W\right) \\ &= \arg\min_{W^TW=1} \mathrm{tr}\left(W^T S\left(U \sum U^T\right)S^T W\right) \\ &= \arg\min_{W^TW=1} \mathrm{tr}\left(W^T S\left(U \sum^{1/2}\sum^{1/2} U^T\right)S^T W\right) \\ &= \arg\min_{W^TW=1} \mathrm{tr}\left(W^T S\left(U_{\sum} U_{\sum}^T\right)S^T W\right). \end{aligned} \tag{20}$$

To further obtain an optimized representation of the weights for GAN network,

$$W = \arg\min_{W^TW=1} \left\| W^T S U_{\sum} \right\|_F^2. \tag{21}$$

Figure 12 gives the flow for weight W establishment of the GAN network, which shows the three processing model implemented with the GAN network through the weight $\{W_{ijk}(n)\} = \{w_{ijk}(1), w_{ijk}(2), \ldots, w_{ijk}(n)\}$.

The meaning of the parameters in Figure 12 is as follows.

The weight of the GAN network includes the process of generator weight $w_j(n)$ transfer, the process of discriminator weight $w_i(n)$ transfer, and the process of error weight $w_k(n)$ update through the network.

$\{h_{ijk}(n)\} = \{h_{ijk}(1), h_{ijk}(2), \ldots, h_{ijk}(n)\}$ is defined as the activation function of the GAN network, including the generator network activation function $h_j(n)$, the discriminator network activation function $h_i(n)$, and the error updating network activation function $h_k(n)$.

$\{s_i, \hat{s}_j\} = \{(s_1, \hat{s}_1), (s_2, \hat{s}_2), \ldots, (s_i, \hat{s}_j)\}$ is defined as the generation set during the generation process given by formula (12). During this processing, $\{s_i\} = \{s_1, s_2, \ldots, s_i\}$ is

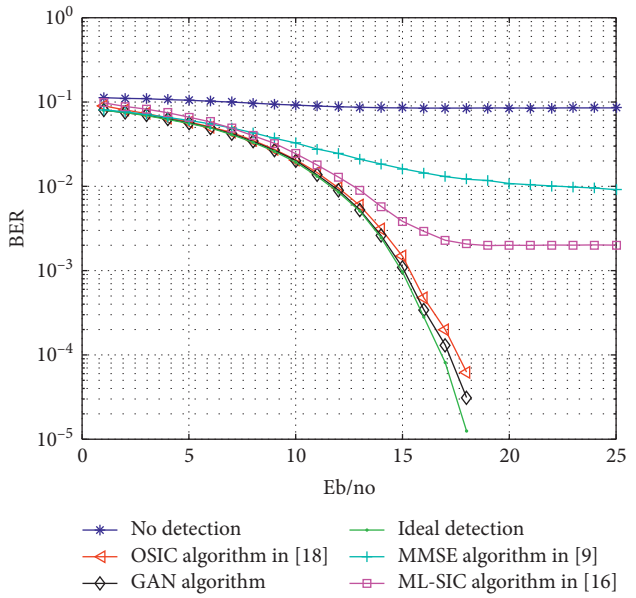


Figure 16: BER performance of QPSK signal in the wilderness environment channel model.

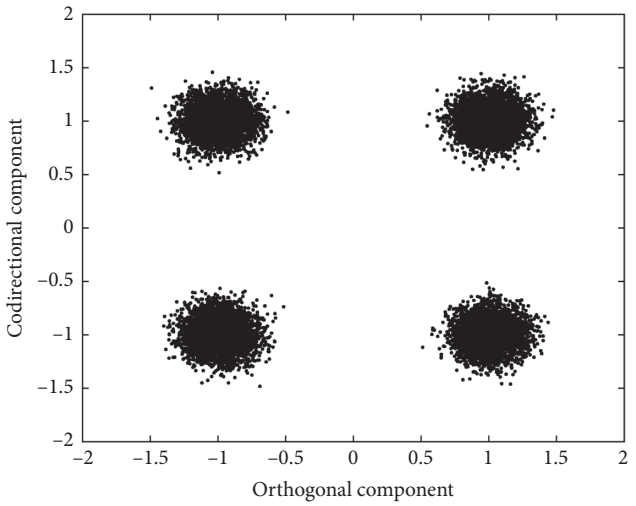


Figure 17: QPSK signal constellation (after the ML-SIC algorithm in [16]).

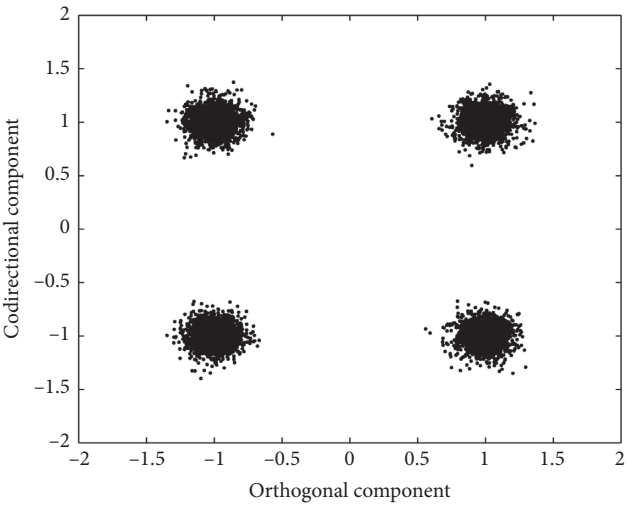


Figure 18: QPSK signal constellation (after the OSIC algorithm in [18]).

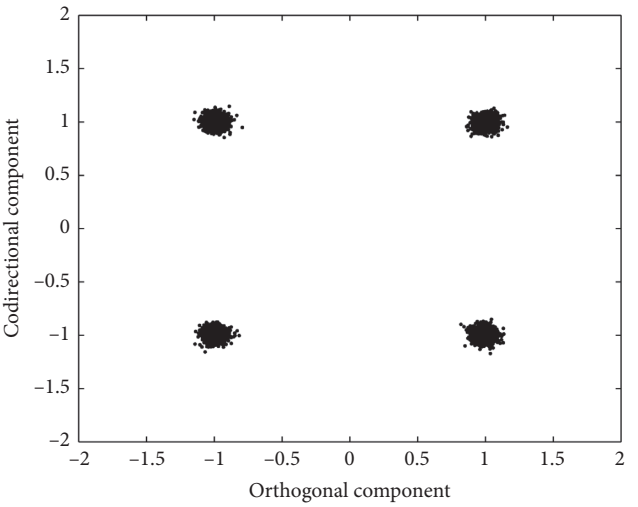


Figure 19: QPSK signal constellation (after the GAN algorithm proposed).

defined as the local data. $\{\hat{s}_j\} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_j\}$ is defined as the generation set. $\{e_j(n)\} = \{e_j(1), e_j(2), \ldots, e_j(n)\}$ is the jth error between the jth generator network and the jth discriminator network, which is produced by the processing between the generator network and the discriminator network.

The generator space $w_j(n)$ is established by optimizing the weights as shown in the generator part and then in the discriminator process given in formula (16). The discriminator space $w_i(n)$ is established as shown in the discriminator part. Then, the optimal weight $w_k(n)$ is obtained by updating the error network.

nonlinear processing and sparse coding introduce the nonlinearization to the learning target, especially due to the noise interference.. Regularization becomes complicated, especially for the feature transformation of the projection space, and it is difficult to regularize in the feature space. Therefore, we use a simplified method to replace and reduce the impact of noise. The sparse distance loss function adopted can obtain better target features.

3.5. Process for GAN Network Optimization. Further, optimization is available as follows:

$$\min_{W} \sum_{(x_i, x_j) \in D} \|Y - WS\|_F^2 + \alpha \operatorname{rank}(WSU_\Sigma), \tag{22}$$

where we define as

$$\text{s.t.} \quad \|W\|_2^2 \leq 1. \tag{23}$$

Further, rank$(\cdot)$ is the representation matrix which takes a rank operation. WXU_Σ is to identify feature limitations for

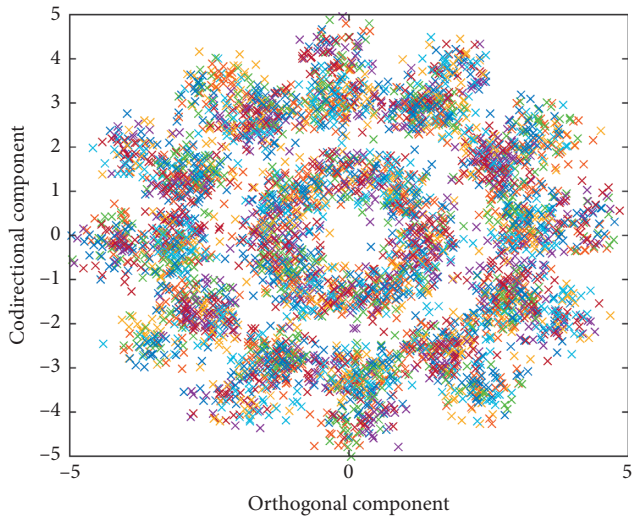


Figure 20: 16QAM signal constellation (undetected multisatellite signal).

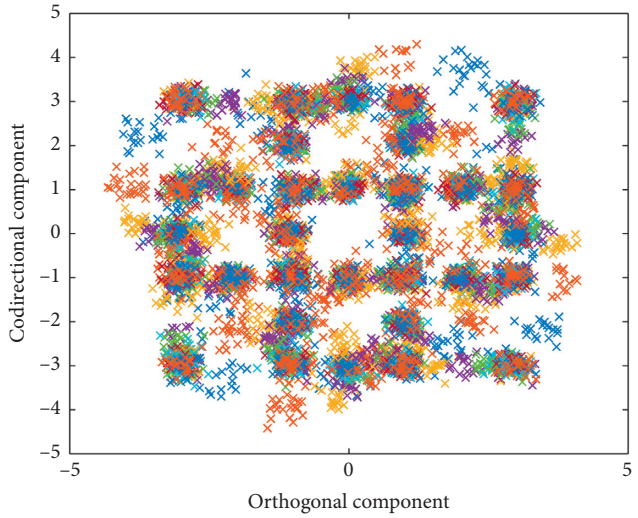


Figure 21: 16QAM signal constellation (detected with no generator network).

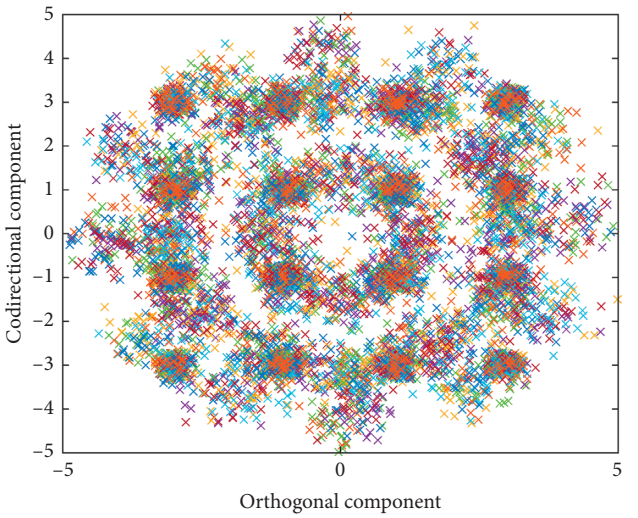


Figure 22: 16QAM signal constellation (detected with no error update network).

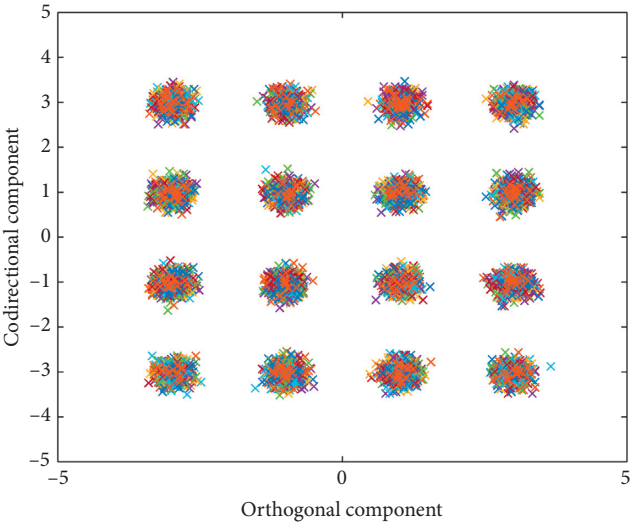


Figure 23: 16QAM signal constellation (detected with GAN network).

different user signals in the process of matrix confrontation. For example, the modulation mode of the received data detection signal can be QPSK, BPSK, 16QAM, etc.

Because rank minimization is an NP problem, it is necessary to obtain optimal convergence in the feature space. In particular, the loss function can be expressed by minimizing the rank of the matrix. By completing the operation of the minimum rank of the feature transformation matrix, the objective loss of the function is to obtain the optimal recognition. In order to improve the structural consistency of the cost function and reduce the influence of sampling interference, the cost loss function gives the low-rank optimization method.

Discussion: we should optimize the objective solution. Among them, $\|\cdot\|_1$ is considered as the modular matrix. The minimum optimization equation is established to solve W as shown below:

$$J = \|Y\|_1 + \frac{\mu}{2}\left\|W^T S U_\Sigma\right\|_1^2, \quad \text{s.t.} \quad W^T W = I. \tag{24}$$

To construct an optimization method, convex theory seeks optimization.

4. Experimental Classification Results and Analysis

In order to verify the detection performance of the proposed GAN algorithm, the average altitude for the satellite is defined as 1450 km, the number of low-orbit satellites is defined as 10, and the spot beams were defined as 7. The parameters for low-orbit satellites are set in accordance with the wilderness model, rural model, and the urban model

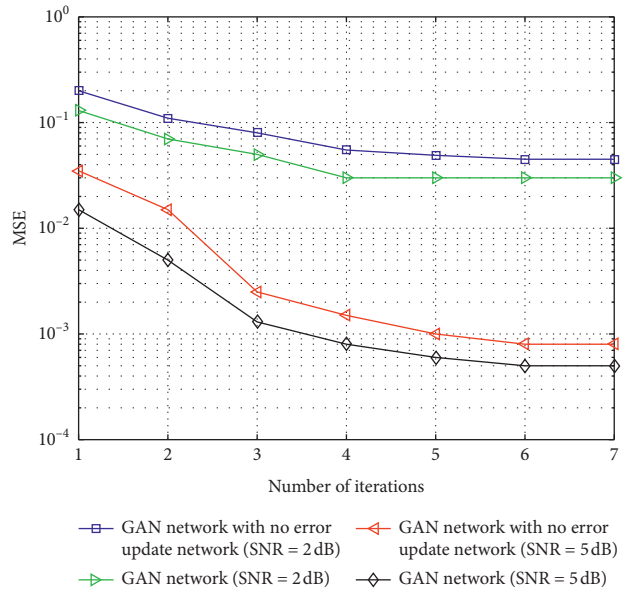


Figure 24: Analysis of stability based on the proposed algorithm.

which are established in the simulation scenarios. Define the maximum working elevation angle of satellite as 35°. Define the downlink for low-orbit satellite transmission bit rate as 160 Mbit/s. We set that the cyclic prefix is greater than the spread delay. We also set the modulation as QPSK.

4.1. BER Analysis. The proposed adversarial network algorithm can be verified by Matlab. The maximum frequency deviation range allowed as the interval of subcarrier. Figures 13–15 show the bit error rate curves under different rural scenarios of low-orbit satellite downlinks, when ground users receive multiple satellite signals. The related algorithms include the frequency domain equalization algorithm based on the MMSE criterion proposed in [9], the ML-SIC algorithm criterion proposed in [16], the iterative order SIC algorithm proposed in [18], and the generative adversarial network proposed in this paper.

The conditions in Figures 13, 14 and 16 assume that each subchannel is independent. It can be obtained from the BER curve that the curve without the elimination is larger. Due to the large frequency offset interference, with the increase in SNR, the BER performance is not increasing. The MMSE algorithm can eliminate part of the interference. The ML-SIC algorithm has better performance, and it can eliminate the interference caused by phase rotation and can suppress the interference introduced by noise. Compared with the SIC algorithm based on the MMSE estimation criterion, the frequency offset preelimination based on the adversarial network algorithm improves the frequency domain frequency offset cancellation performance, especially when the frequency offset is large, the performance of frequency offset cancellation will be significantly improved.

4.2. Constellation Analysis. The downlink multisatellite detection based on the generative adversarial network can be obtained through the simulation of QPSK constellation, which is still selected as the signal mapping. Assuming SNR = 15 dB, the downlink received signal is defined in an urban environment. Figure 15 shows the signal constellation after the MMSE algorithm in [9]. Figure 17 shows the signal constellation after the ML-SIC algorithm in [16]. Figure 18 shows the signal constellation after the OSIC algorithm in [18]. Figure 19 shows the signal constellation after the GAN algorithm proposed in this paper. Compared with the constellation simulation used in Figures 17–19, the adversarial network can better suppress the interference introduced by the larger carrier frequency offset and improve the accuracy of downlink signal cancellation.

4.3. Integrity Analysis. In order to verify the integrity of the three networks more effectively and reflect the differentiation of simulation, we use the simulation of 16QAM constellation for the downlink multisatellite detection. Assuming SNR = 10 dB, the downlink received signal is defined in an urban environment.

Figure 20 shows the signal constellation with no detection. Figure 21 shows the signal constellation after detection with no generator network. Figure 22 shows the signal constellation after detection with no error update network. Figure 23 shows the proposed GAN network. Compared with the constellation simulation used, the adversarial network with three networks, including generator network, discriminator network, and error update network, can better suppress the interference and improve the accuracy of downlink signal cancellation.

4.4. Convergence Analysis. Figure 24 gives the analysis of stability based on the proposed algorithm. The stability of the proposed algorithm is obtained through the convergence learning curve. We research whether the MSE convergence performance of the system combined with the error update network. By defining SNR = 5 dB and SNR = 2 dB, respectively, we get the simulation performance. After 5 iterations, the MSE has reached $10e-3$. Under the same conditions, the convergence speed of no error update is slower, and this is because the error update could improve the convergence learning performance.

5. Conclusion

In this paper, we have proposed the method for satellite downlink signal detection based on the GAN network. We establish the generator network and adversarial network, respectively. The generator network is established with the local generator of virtual satellite signals, and the adversarial network is established for high-precision signal detection. And we also have established the error network with the error signal from satellite downlink. Then we form an adaptive matrix weight adjustment. Compared with traditional shallow networks, such as MMSE, ML-SIC

algorithms, and iterative algorithms, under the same signal-to-noise ratio, the performance is improved by 5 dB.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Qingyang Guan and Shuang Wu contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61501306), Scientific Research Initiation Funds for the Doctoral Program of Xi'an International University (Grant nos. XAIU2019002 and XAIU2018070102), General Project of Science and Technology Department of Shaanxi Province (Grant no. 2020JM-638), and the Natural Science Foundation of Liaoning Province of China (no. 2015020026).

References

[1] S. N. Tran and A. S. d'Avila Garcez, "Deep logic networks: inserting and extracting knowledge from deep belief networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 246–258, 2018.

[2] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] N. C. Luong, D. Thai Hoang, and S. Gong, "Applications of deep reinforcement learning in communications and networking: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[4] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[5] E. Candes and T. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[6] M. M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $l_{2,0}$ norm approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4646–4655, 2010.

[7] R. Prasad, C. R. Murthy, and B. D. Rao, "Joint approximately sparse channel estimation and data detection in OFDM systems using sparse Bayesian learning," *IEEE Transactions on Signal Processing*, vol. 62, no. 14, pp. 3591–3603, 2014.

[8] Z. Chen, L.-Y. Duan, S. Wang et al., "Toward knowledge as a service over networks: a deep learning model communication paradigm," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1349–1363, 2019.

[9] S. Gaur and M. Ingram, "Simple MMSE interference suppression for real and rate-1/2 complex orthogonal space-time block codes," *IEEE Transactions on Wireless Communications*, vol. 7, no. 8, pp. 2901–2904, 2008.

[10] H. Sampath and A. Paulraj, "Linear precoding for space-time coded systems with known fading correlations," *IEEE Communications Letters*, vol. 6, no. 6, pp. 239–241, 2002.

[11] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1051–1064, 2002.

[12] H. Giannakis, P. Stoica, and A. Paulraj, "Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion," *IEEE Transactions on Communications*, vol. 49, no. 12, pp. 2198–2206, 2001.

[13] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, 2011.

[14] M. Cirkic and E. Larsson, "SUMIS: near-optimal soft-in soft-out MIMO detection with low and fixed complexity," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3084–3097, 2014.

[15] S. K. Mohammed and E. G. Larsson, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 3, pp. 1059–1071, 2013.

[16] M. B. Çelebi and H. Arslan, "Theoretical analysis of the coexistence of LTE-A signals and design of an ML-SIC receiver," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4626–4639, 2015.

[17] F. Rusek, D. Persson, B. K. Lau, E. Larsson, and T. Marzetta, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[18] J. Ahn, B. Shim, and K. B. Lee, "Sparsity-aware ordered successive interference cancellation for massive machine-type communications," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 134–137, 2018.

[19] P. Ghamisi, M. Dalla Mura, and J. A. Benediktsson, "A survey on spectral-spatial classification techniques based on attribute profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2335–2353, 2015.

[20] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: a review," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6547–6565, 2017.

[21] Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 2966–2977, 2016.

[22] Q. Wang, Z. Meng, and X. Li, "Locality adaptive discriminant analysis for spectral-spatial classification of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2077–2081, 2017.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[24] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[25] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: a technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[26] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "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.

[27] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 1–12, 2015.

[28] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3516–3530, 2017.

[29] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyper spectral image classification," *Journal of Sensors*, vol. 2015, Article ID 258619, 2015.

[30] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 844–853, 2017.

[31] W. Shao and S. Du, "Spectral-spatial feature extraction for hyper-spectral image classification: a dimension reduction and deep learning approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4544–4554, 2016.

[32] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral-spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sensing Letters*, vol. 6, no. 6, pp. 468–477, 2015.

[33] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[34] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sensing*, vol. 9, no. 1, p. 67, 2017.

[35] H. Liang and Q. Li, "Hyperspectral imagery classification using sparse representations of convolutional neural network features," *Remote Sensing*, vol. 8, no. 2, p. 99, 2016.

[36] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, "Hyperspectral images classification with gabor filtering and convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2355–2359, 2017.

[37] C. Zhang, X. Yang, Y. Tang, and W. Zhang, "Learning to generate radar image sequences using two-stage generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 401–405, 2020.

[38] C. Wang, C. Xu, X. Yao, and D. Tao, "Evolutionary generative adversarial networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 6, pp. 921–934, 2019.

[39] Y. Pang, J. Xie, and X. Li, "Visual haze removal by a unified generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3211–3221, 2019.

[40] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, "Gated-GAN: adversarial gated networks for multi-collection style transfer," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 546–560, 2019.

[41] A. Creswell, T. White, V. Dumou, and K. Lin, "Generative adversarial networks: an overview," *IEEE Signal Processing Society*, vol. 35, no. 1, pp. 53–65, 2017.

[42] Y. Sun, J. Tang, X. Shu, Z. Sun, and M. Tistarelli, "Facial age synthesis with label distribution-guided generative adversarial network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2679–2691, 2020.

[43] H. Yu, Z. Li, G. Zhang, P. Liu, and J. Wang, "Extracting and predicting taxi hotspots in spatiotemporal dimensions using conditional generative adversarial neural networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 3680–3692, 2020.

[44] Z. Yuan, H. Li, J. Liu, and J. Luo, "Multiview scene image inpainting based on conditional generative adversarial networks," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 314–323, 2020.

[45] C. Jia, X. Zhang, S. Wang, S. Wang, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 177–189, 2019.

[46] M. Mardani, E. Gong, J. Y. Cheng et al., "Deep generative adversarial neural networks for compressive sensing MRI," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 167–179, 2019.

[47] Y. Lyu, Z. Han, J. Zhong, C. Li, and Z. Liu, "A generic anomaly detection of catenary support components based on generative adversarial networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2439–2448, 2020.

[48] H. Li, S. Zhou, W. Yuan, J. Li, and H. Leung, "Adversarial-example attacks toward android malware detection system," *IEEE Systems Journal*, vol. 14, no. 1, pp. 653–656, 2020.

[49] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1359–1371, 2019.

[50] M. Zhang, M. Gong, Y. Mao, J. Li, and Y. Wu, "Unsupervised feature extraction in hyperspectral images based on was serstein generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2669–2688, 2019.

[51] Q. Xuan, Z. Chen, Y. Liu, H. Huang, G. Bao, and D. Zhang, "Multiview generative adversarial network and its application in pearl classification," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 10, pp. 8244–8252, 2019.