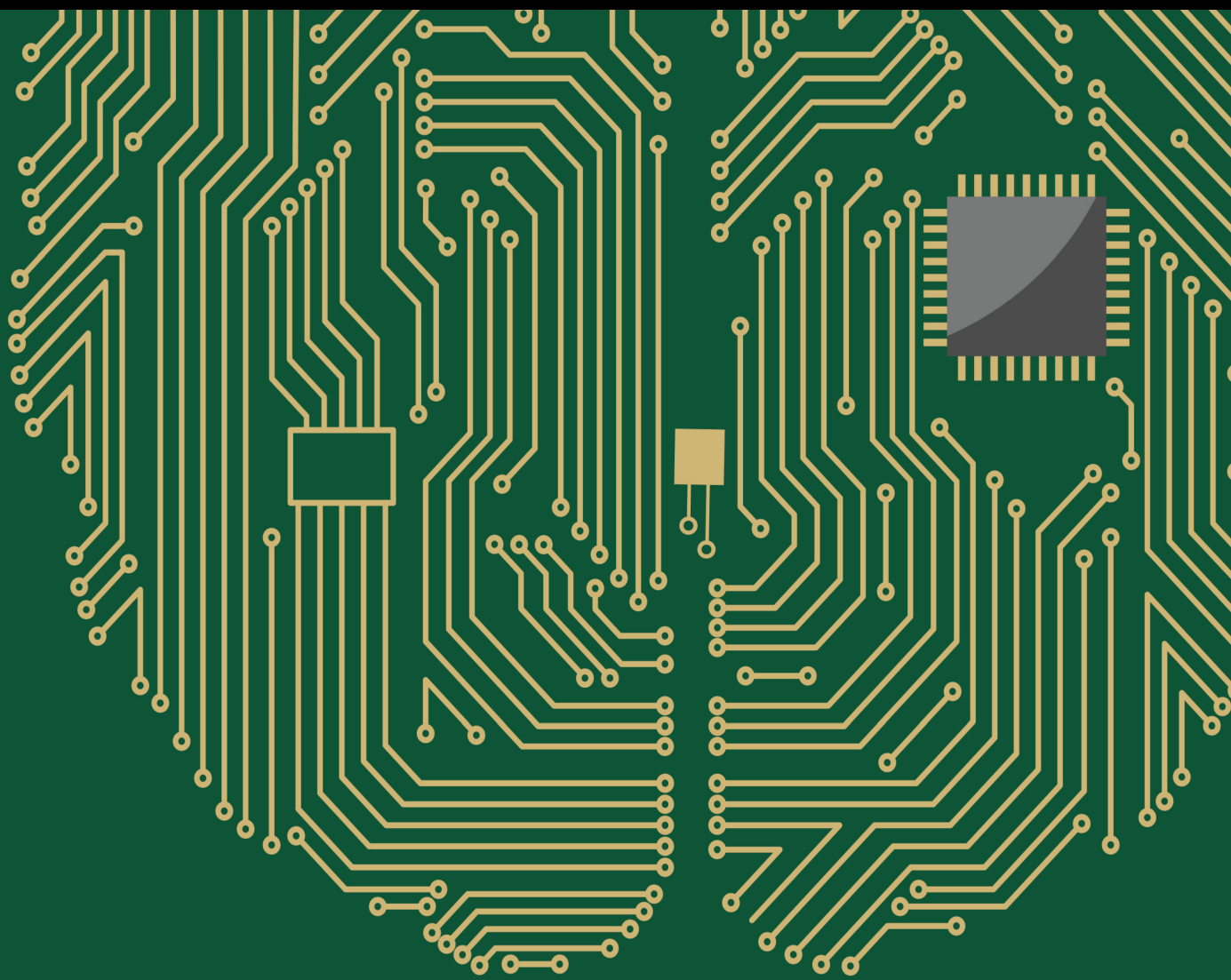


Advanced Artificial Intelligence for Computational Healthcare and Bioinformatics

Lead Guest Editor: Ahmedin M. Ahmed

Guest Editors: Muhammad Tariq, Qiao Xiang, and Xiaolong Zhou





Advanced Artificial Intelligence for Computational Healthcare and Bioinformatics

Computational Intelligence and Neuroscience

**Advanced Artificial Intelligence
for Computational Healthcare and
Bioinformatics**

Lead Guest Editor: Ahmedin M. Ahmed

Guest Editors: Muhammad Tariq, Qiao Xiang, and
Xiaolong Zhou



Copyright © 2023 Hindawi Limited. All rights reserved.

This is a special issue published in "Computational Intelligence and Neuroscience." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chief Editor

Andrzej Cichocki, Poland

Associate Editors

Arnaud Delorme, France
Cheng-Jian Lin , Taiwan
Saeid Sanei, United Kingdom

Academic Editors

Mohamed Abd Elaziz , Egypt
Tariq Ahanger , Saudi Arabia
Muhammad Ahmad, Pakistan
Ricardo Aler , Spain
Nouman Ali, Pakistan
Pietro Aricò , Italy
Lerina Aversano , Italy
Ümit Ağbulut , Turkey
Najib Ben Aoun , Saudi Arabia
Surbhi Bhatia , Saudi Arabia
Daniele Bibbo , Italy
Vince D. Calhoun , USA
Francesco Camastra, Italy
Zhicheng Cao, China
Hubert Cecotti , USA
Jyotir Moy Chatterjee , Nepal
Rupesh Chikara, USA
Marta Cimitile, Italy
Silvia Conforto , Italy
Paolo Crippa , Italy
Christian W. Dawson, United Kingdom
Carmen De Maio , Italy
Thomas DeMarse , USA
Maria Jose Del Jesus, Spain
Arnaud Delorme , France
Anastasios D. Doulamis, Greece
António Dourado , Portugal
Sheng Du , China
Said El Kafhali , Morocco
Mohammad Reza Feizi Derakhshi , Iran
Quanxi Feng, China
Zhong-kai Feng, China
Steven L. Fernandes, USA
Agostino Forestiero , Italy
Piotr Franaszczuk , USA
Thippa Reddy Gadekallu , India
Paolo Gastaldo , Italy
Samanwoy Ghosh-Dastidar, USA

Manuel Graña , Spain
Alberto Guillén , Spain
Gaurav Gupta, India
Rodolfo E. Haber , Spain
Usman Habib , Pakistan
Anandakumar Haldorai , India
José Alfredo Hernández-Pérez , Mexico
Luis Javier Herrera , Spain
Alexander Hošovský , Slovakia
Etienne Hugues, USA
Nadeem Iqbal , Pakistan
Sajad Jafari, Iran
Abdul Rehman Javed , Pakistan
Jing Jin , China
Li Jin, United Kingdom
Kanak Kalita, India
Ryotaro Kamimura , Japan
Pasi A. Karjalainen , Finland
Anitha Karthikeyan, Saint Vincent and the
Grenadines
Elpida Keravnou , Cyprus
Asif Irshad Khan , Saudi Arabia
Muhammad Adnan Khan , Republic of
Korea
Abbas Khosravi, Australia
Tai-hoon Kim, Republic of Korea
Li-Wei Ko , Taiwan
Raşit Köker , Turkey
Deepika Koundal , India
Sunil Kumar , India
Fabio La Foresta, Italy
Kuruva Lakshmana , India
Maciej Lawrynczuk , Poland
Jianli Liu , China
Giosuè Lo Bosco , Italy
Andrea Loddo , Italy
Kezhi Mao, Singapore
Paolo Massobrio , Italy
Gerard McKee, Nigeria
Mohit Mittal , France
Paulo Moura Oliveira , Portugal
Debajyoti Mukhopadhyay , India
Xin Ning , China
Nasimul Noman , Australia
Fivos Panetsos , Spain

Evgeniya Pankratova , Russia
Rocío Pérez de Prado , Spain
Francesco Pistolesi , Italy
Alessandro Sebastian Podda , Italy
David M Powers, Australia
Radu-Emil Precup, Romania
Lorenzo Putzu, Italy
S P Raja, India
Dr.Anand Singh Rajawat , India
Simone Ranaldi , Italy
Upaka Rathnayake, Sri Lanka
Navid Razmjooy, Iran
Carlo Ricciardi, Italy
Jatinderkumar R. Saini , India
Sandhya Samarasinghe , New Zealand
Friedhelm Schwenker, Germany
Mijanur Rahaman Seikh, India
Tapan Senapati , China
Mohammed Shuaib , Malaysia
Kamran Siddique , USA
Gaurav Singal, India
Akansha Singh , India
Chiranjibi Sitaula , Australia
Neelakandan Subramani, India
Le Sun, China
Rawia Tahrir , Iraq
Binhua Tang , China
Carlos M. Travieso-González , Spain
Vinh Truong Hoang , Vietnam
Fath U Min Ullah , Republic of Korea
Pablo Varona , Spain
Roberto A. Vazquez , Mexico
Mario Versaci, Italy
Gennaro Vessio , Italy
Ivan Volosyak , Germany
Leyi Wei , China
Jianghui Wen, China
Lingwei Xu , China
Cornelio Yáñez-Márquez, Mexico
Zaher Mundher Yaseen, Iraq
Yugen Yi , China
Qiangqiang Yuan , China
Miaolei Zhou , China
Michal Zochowski, USA
Rodolfo Zunino, Italy

Contents

Retracted: Home Care System for Mobility Disabilities Based on Intelligent Perception

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9856364, Volume 2023 (2023)

Retracted: Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9891083, Volume 2023 (2023)

Retracted: Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9868231, Volume 2023 (2023)

Retracted: An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9853814, Volume 2023 (2023)

Retracted: On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9816078, Volume 2023 (2023)

Retracted: Analysis of Geocological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9793050, Volume 2023 (2023)

Retracted: Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9760925, Volume 2023 (2023)

BUPNN: Manifold Learning Regularizer-Based Blood Usage Prediction Neural Network for Blood Centers

Lingling Pan, Zelin Zang , Siqi Ma, Wei Hu , and Zhechang Hu


Research Article (13 pages), Article ID 1003310, Volume 2023 (2023)

IGPred-HDnet: Prediction of Immunoglobulin Proteins Using Graphical Features and the Hierarchical Deep Learning-Based Approach

Zakir Ali , Fahad Alturise , Tamim Alkhalifah , and Yaser Daanial Khan 

Research Article (13 pages), Article ID 2465414, Volume 2023 (2023)

[Retracted] Analysis of Geoecological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information

Yuanwen Song , Lei Gao, Haipin He, and Juan Lu



Research Article (13 pages), Article ID 6555005, Volume 2022 (2022)

[Retracted] An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring

Liheng He, Tingru Zhu, and Meng Lv 



Research Article (12 pages), Article ID 4250462, Volume 2022 (2022)

Interpersonal Relationship, Knowledge Characteristic, and Knowledge Sharing Behavior of Online Community Members: A TAM Perspective

Wu Jiarui, Zhang Xiaoli , and Su Jiafu 



Research Article (11 pages), Article ID 4188480, Volume 2022 (2022)

The Chain Mediating Effect of Network Behavior and Decision Self-Efficacy between Work Skills and Perceived Employability Based on Social Cognitive Theory

Liping Yang  and Hong Zhang 


Research Article (12 pages), Article ID 5240947, Volume 2022 (2022)

The Regulatory Effect of Firm Size on Digital Transformation: An Empirical Study of Pharmaceutical Companies in China

Xiaowen Luo  and Shun-Chi Yu 


Research Article (12 pages), Article ID 7731174, Volume 2022 (2022)

[Retracted] Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment

Shaojie Liu 


Research Article (7 pages), Article ID 3415999, Volume 2022 (2022)

[Retracted] Home Care System for Mobility Disabilities Based on Intelligent Perception

Yanyan Qiu and Chunbo Qiu 


Research Article (9 pages), Article ID 9528046, Volume 2022 (2022)

[Retracted] On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data

Xiaoqiong Ding 

Research Article (11 pages), Article ID 7183032, Volume 2022 (2022)


[Retracted] Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture

Xiaoxian Ju 

Research Article (9 pages), Article ID 5222760, Volume 2022 (2022)

Contents

[Retracted] Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model

Xinshu Li 

Research Article (9 pages), Article ID 3590224, Volume 2022 (2022)

DBP-iDWT: Improving DNA-Binding Proteins Prediction Using Multi-Perspective Evolutionary Profile and Discrete Wavelet Transform

Farman Ali , Omar Barukab , Ajay B Gadicha, Shruti Patil, Omar Alghushairy, and Akram Y. Sarhan


Research Article (8 pages), Article ID 2987407, Volume 2022 (2022)

A Lane-Changing Decision-Making Model of Bus Entering considering Bus Priority Based on GRU Neural Network

Wanjun Lv , Yongbo Lv , Jianwei Guo , and Jihui Ma

Research Article (10 pages), Article ID 4558946, Volume 2022 (2022)

An Optimized LIME Scheme for Medical Low Light Level Image Enhancement

Yue Kun, Gong Chunqing, and Gao Yuehui 



Research Article (8 pages), Article ID 9613936, Volume 2022 (2022)

Seismic Safety Design and Analysis of Hydraulic Sluice Chamber Structure Based on Finite Element Method

Xin Cai, Zhenming Cui , Xingwen Guo, Fan Li, and Yanan Zhang


Research Article (10 pages), Article ID 6183588, Volume 2022 (2022)

Differential Privacy via Haar Wavelet Transform and Gaussian Mechanism for Range Query

Dong Chen , Yanjuan Li, Jiaquan Chen, Hongbo Bi, and Xiajun Ding 


Research Article (17 pages), Article ID 8139813, Volume 2022 (2022)

Application and Analysis of Improved Fuzzy Comprehensive Evaluation Method in Goodwill Evaluation and Intangible Asset Management

Mengyue Xu 

Research Article (11 pages), Article ID 2235542, Volume 2022 (2022)

Research on the Coordinated Development of New Rural Production and Living Civilization Construction from the Perspective of Green Transformation and Development

Xiaoyan Li 

Research Article (12 pages), Article ID 1750256, Volume 2022 (2022)

Retraction

Retracted: Home Care System for Mobility Disabilities Based on Intelligent Perception

Computational Intelligence and Neuroscience

Received 26 September 2023; Accepted 26 September 2023; Published 27 September 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] Y. Qiu and C. Qiu, "Home Care System for Mobility Disabilities Based on Intelligent Perception," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9528046, 9 pages, 2022.

Retraction

Retracted: Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Li, "Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 3590224, 9 pages, 2022.

Retraction

Retracted: Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] S. Liu, "Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 3415999, 7 pages, 2022.

Retraction

Retracted: An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] L. He, T. Zhu, and M. Lv, "An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4250462, 12 pages, 2022.

Retraction

Retracted: On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Ding, "On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7183032, 11 pages, 2022.

Retraction

Retracted: Analysis of Geocological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] Y. Song, L. Gao, H. He, and J. Lu, "Analysis of Geocological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 6555005, 13 pages, 2022.

Retraction

Retracted: Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Ju, "Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 5222760, 9 pages, 2022.

Research Article

BUPNN: Manifold Learning Regularizer-Based Blood Usage Prediction Neural Network for Blood Centers

Lingling Pan,^{1,2} Zelin Zang ,³ Siqi Ma,³ Wei Hu ,^{1,2} and Zhechang Hu¹

¹Blood Center of Zhejiang Province, 789 Jianye Road, Binjiang District, Hangzhou 310052, Zhejiang, China

²Key Laboratory of Blood Safety Research, Hangzhou 310052, Zhejiang, China

³AI Lab, School of Engineering, Westlake University, Hangzhou 310052, Zhejiang, China

Correspondence should be addressed to Wei Hu; huwei0507@126.com

Received 21 June 2022; Revised 31 August 2022; Accepted 10 September 2022; Published 31 January 2023

Academic Editor: Ahmedin M. Ahmed

Copyright © 2023 Lingling Pan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Blood centers are an essential component of the healthcare system, as timely blood collection, processing, and efficient blood dispatch are critical to the treatment of patients and the performance of the entire healthcare system. At the same time, an efficient blood dispatching system through the high-precision predictive capability of artificial intelligence is crucial for the efficiency improvement of the blood centers. However, the current artificial intelligence (AI) models for predicting blood usage do not meet the needs of blood centers. The challenges of AI models mainly include lower generalization ability in different hospitals, limited stability under missing values, and low interpretability. An artificial neural network-based model named the blood usage prediction neural network (BUPNN) has been developed to address these challenges. BUPNN includes a novel similarity-based manifold regularizer that aims to enhance network mapping consistency and, thus, overcome the domain bias of different hospitals. Moreover, BUPNN diminishes the performance degradation caused by missing values through data enhancement. Experimental results on a large amount of accurate data demonstrate that BUPNN outperforms the baseline method in classification and regression tasks and excels in generalization and consistency. Moreover, BUPNN has solid potential to be interpreted. Therefore, the decision-making process of BUPNN is explored to the extent that it acts as an aid to the experts in the blood center.

1. Introduction

Blood products are an essential part of the treatment of bleeding, cancer, AIDS, hepatitis, and other diseases [1]. Blood is also an indispensable resource in treating injured patients; whether promptly transfusing the blood is critical for rehabilitating injured patients. At the same time, early surgical interventions and rapid blood transfusions are the primary measures to reduce mortality. Unfortunately, these measures require large amounts of blood to support them.

Blood, however, is significantly different from other medical products. Currently, blood cannot be manufactured or synthesized artificially and can only be donated by others. In addition, blood has a short shelf life, making emergency blood more specific and irreplaceable than other medical products. For example, patients in mass casualty events

caused by earthquakes suffer from fractures, fractures accompanied by multiple organ injuries, and crush injuries. Therefore, the peak period of blood consumption occurs 96 hours after the earthquake [2]. In contrast, patients mainly suffer from burns in mass casualties caused by bombings or fires. The peak blood consumption occurs 24 hours after the event. In addition, ongoing blood transfusions may last for several months [3]. Therefore, modeling the prediction of blood use in patients is a meaningful and challenging topic.

Current blood usage prediction models are not well developed. For example, Wang et al. [4] developed an early transfusion scoring system to predict blood requirements in severe trauma patients in the prehospital or initial phase of emergency resuscitation. However, the abovementioned design is more suitable for triage in a

single hospital than in blood centers since the system does not consider how to avoid performance degradation due to domain deviations between different hospitals. Rebecca et al. [5, 6] summarized the currently available studies exploring how to predict the need for massive transfusion in patients with traumatic injuries, listing the blood consumption scoring system (ABC) and the shock index scoring system (SI). These systems use classical or ML-based methods to predict blood consumption during the treatment of patients. Unfortunately, although the progress is remarkable, but it is unsatisfactory in terms of accuracy and generalization, and thus, the blood demand prediction model cannot be widely used. We summarize the problems into the following three points:

- (i) Data quality is limited. The partnership between blood centers and hospitals makes it very difficult to establish a rigorous feedback system for patient information. In addition, biases and missing data due to differences in hospital equipment create challenges for a constant blood use prediction model.
- (ii) Model generalizability is unsatisfactory. Blood usage prediction models are often built for specific hospitals or cities without considering extending to a wider range of applications and thus have poor generalization performance.
- (iii) Model interpretability is inadequate. Most models can only output a category or blood usage forecast but cannot demonstrate the model's decision process. An interpretable model can better work with experts to help blood centers with blood schedules.

This paper proposes a blood usage prediction neural network, BUPNN, to solve the above problems. First, actual patient clinic data and treatment procedures in 12 hospitals are used as training data. Extensively collected data with biases from various hospitals will provide sufficient information to train a high-performance model. We used multiple data complementation schemes to restore the real problem and overcome missing values. In addition, the BUPNN model MDC is augmented with online data by linear interpolation, which increases the diversity of training data and thus improves the stability of the model under training with missing data. Second, to further improve the generalization performance of BUPNN, a similarity-based loss function is introduced to map data with biases to a stable semantic space by aligning samples from different hospitals in the latent space. Third, we analyze the model based on the deep learning interpretation method to enhance interpretability. The proposed analysis is accompanied by the prediction output of the model in real-time to assist one in understanding the process of the model's decision-making. The interpretable study of BUPNN provides the conditions for computers and experts to help each other in blood consumption prediction.

The main contributions of this study are as follows:

- (i) Representative samples. A large amount of data from twelve hospitals is collected for this study to investigate the implied relationship between patients' chain indicators and blood consumption.
- (ii) Generalizable model. This study designs MDC for online missing data completion, and thus, data augmentation to enhance the model's generalization ability. In addition, the various similarity loss function is designed to improve the model's predictive power across domains.
- (iii) Excellent Performance. Experiments on six different settings demonstrate that our method outperforms all baseline methods.

The rest of this paper is organized as follows. In Section 2, we provide a literature review of demand forecasting methods for blood products. Section 3 provides an initial exploration of the data. We provide the data description, model background, model development, and evaluation of four different models for blood demand forecasting in Section. In Section 4, a comparison of the models is provided, and finally, in Section 5, concluding remarks are provided, including a discussion of ongoing work for this problem.

2. Related Work

2.1. ML Techniques for Medical Problems. The integration of the medical field with ML technology has received much attention in recent years. Two areas that may benefit from the application of ML technology in the medical field are diagnosis and outcome prediction. It includes the possibility of identifying high-risk medical emergencies, such as recurrence or transition to another disease state. Recently, ML algorithms have been successfully used to classify thyroid cancer [7] and predict the progression of COVID-19 [8, 9]. On the other hand, ML-based visualization and dimensionality reduction techniques have the potential to help professionals analyze biological or medical data, guiding them to better understand the data [10, 11]. Furthermore, ML-based feature selection techniques [12, 13] have strong interpretability and the potential to find highly relevant biomarkers for output in a wide range of medical data, leading to new biological or medical discoveries.

2.2. Blood Demand Forecasting. There is limited literature on blood demand forecasting; most investigate univariate time series methods. In these studies, forecasts are based solely on previous demand values without considering other factors affecting the demand. Frankfurter et al. [14] developed transfusion forecasting models using exponential smoothing (ES) methods for a blood collection and distribution center in New York. Critchfield et al. [15] developed models for forecasting blood usage in a blood center using several time series methods, including moving average (MA), winter's approach, and ES. Filho et al. [16] developed a Box-Jenkins seasonal

autoregressive integrated moving average (BJ-SARIMA) model to forecast weekly demand for hospital blood components. Their proposed method, SARIMA, is based on a Box-Jenkins approach that considers seasonal and nonseasonal characteristics of time series data. Later, Filho et al. [17] extended their model by developing an automatic procedure for demand forecasting and changing the model level from hospital level to regional blood center to help managers use the model directly. Kumari and Wijayanayake [18] proposed a blood inventory management model for daily blood supply, focusing on reducing blood shortages. Three time series methods, namely, MA, weighted moving average (WMA), and ES, are used to forecast blood usage and are evaluated based on needs. Fortsch and Khapalova [19] tested various blood demand prediction approaches, such as naive, moving average, exponential smoothing, and multiplicative time series decomposition (TSD). The results show that the Box-Jenkins (ARMA) approach, which uses an autoregressive moving average model, results in the highest prediction accuracy. Lestari et al. [20] applied four models to predict blood component demand, including moving average, weighted moving average, exponential smoothing, exponential smoothing with the trend, and select the best method for their data based on the minimum error between forecasts and the actual values. Volken et al. [21] used generalized additive regression and time-series models with exponential smoothing to predict future whole blood donation and RBC transfusion trends.

Several recent studies consider clinically related indicators. For example, Drackley et al. [22] estimated a long-term blood demand for Ontario, Canada, based on previous transfusions' age and sex-specific patterns. They forecast blood supply and demand for Ontario by considering demand and supply patterns and demographic forecasts, assuming fixed patterns and rates over time. Khaldi et al. [23] applied artificial neural networks (ANNs) to forecast the monthly demand for three blood components: red blood cells (RBCs), blood, and plasma, for a case study in Morocco. Guan et al. [24] proposed an optimisation ordering strategy in which they forecast the blood demand for several days into the future and build an optimal ordering policy based on the predicted direction, concentrating on minimising the wastage. Their primary focus is on an optimal ordering policy. They integrate their demand model into the inventory management problem, meaning they do not precisely try to forecast blood demand. Li et al. [25] developed a hybrid model consisting of seasonal and trend decomposition using Loess (STL) time series and eXtreme Gradient Boosting (XGBoost) for RBC demand forecasting and incorporated it into an inventory management problem.

Recently, Motamedi et al. [26] presented an efficient forecasting model for platelet demand at Canadian Blood Services (CBS). In addition, C. Twumasi and J. Twumasi [27] compared k-nearest neighbour regression (KNN), multi-layer perceptron (MLP), and support vector machine (SVM) via a rolling-origin strategy for forecasting and backcasting blood demand data with missing values and outliers from a government hospital in Ghana. Abolghasemi et al. [28] treat the blood supply problem as an optimisation problem [29] and find that LightGBM provides promising solutions and outperforms other machine learning models.

3. Blood Centers and Datasets

This section describes how blood center works with an example in Zhejiang Province, China. It includes the responsibilities of the blood center and the partnership between the blood center and the hospital, and shows in detail how the proposed model will assist the blood center in accomplishing its mission better.

3.1. Blood Center. Figure 1 shows that, this study considers a provincial centralized blood supply system, including blood centers, blood stations, and hospitals. The entire centralized blood supply system completes the collection, management, and delivery of blood products. The blood center is responsible for collecting blood products, testing for viruses and bacteria, and supplying some hospitals. At the same time, blood centers assume the management, coordination, and operational guidance of blood collection and supply institutions. Blood stations are responsible for collecting, storing, and transporting blood to local hospitals. While receiving blood, hospitals must collect clinical information on current patients for analysis and decision-making at the blood center to make the blood supply more efficient. Our proposed AI blood consumption prediction system (BUPNN) receives clinical information from each hospital and uses it to predict the future blood consumption of each patient. The proposed system helps blood center specialists perform blood collection and transportation better.

3.2. Data Details and Challenges. We collected data in their actual state to build a practically usable model. The data in this study are constructed by processing BS shipping data and the TRUST (Transfusion Research for Utilization, Surveillance, and Tracking) database from Zhejiang Province Blood Center and 12 hospitals in Zhejiang Province. The data include data from 2025 patients, including 1970 data from emergency trauma patients in 10 hospitals in Zhejiang Province from 2018 to 2020 and another 55 from patients in two hospitals in Wenling's explosion in 2020.

Each dataset mainly included the following parts. (1) General patient information, including case number, consultation time, pretest classification, injury time, gender, age, weight, diagnosis, penetrating injury, heart rate, diastolic blood pressure, systolic blood pressure, body temperature, shock index, and Glasgow coma index. (2) Injury status, including pleural effusion, abdominal effusion, extremity injury status, thoracic and abdominal injury status, and pelvic injury status. (3) Laboratory tests, including hemoglobin, erythrocyte pressure product, albumin, hemoglobin value 24 hours after transfusion therapy, base residual, PH (acidity), base deficiency, oxygen saturation, PT (prothrombin time), and APTT (partial thromboplastin time). (4) Burn situation, including burn area and burn depth. (5) Patient regression, including whether blood use: whether to transfuse suspended red blood cells, the first day of hospital transfusion, and the amount of blood transfusion.

For a more straightforward presentation of the data distribution, preliminary dataset statistics are shown in

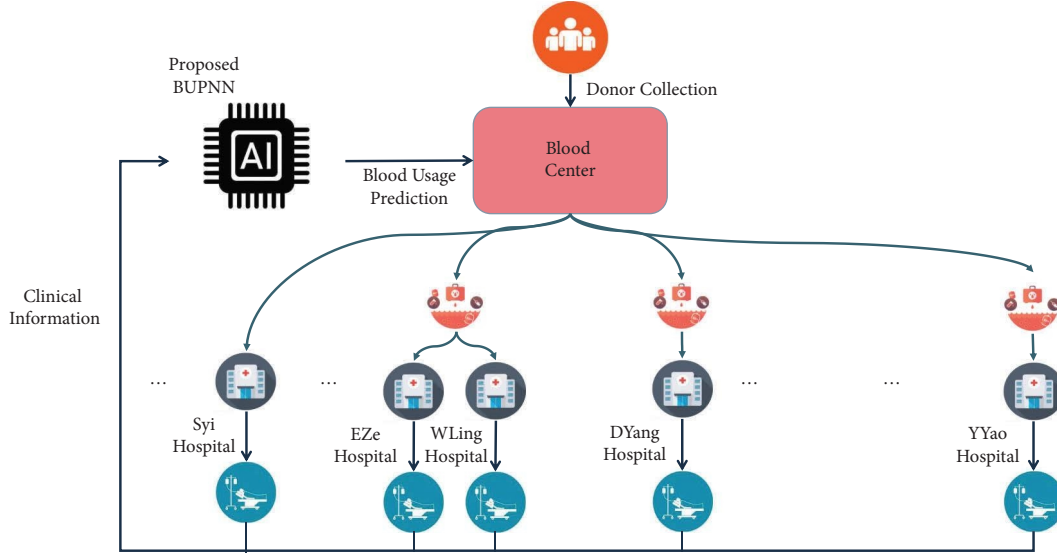


FIGURE 1: BS blood supply chain with one regional blood center and multiple hospitals.

Table 1, and box plots by age, blood consumption, and missing rate are shown in Figure 2.

After a detailed definition of the problem and a description of the dataset, we summarize the main problems faced in this paper.

- (i) Data holds missing values. as shown in Table 1 and Figure 2(c), some hospitals have more than 10% missing values. The missing values cause perturbation to the data distribution, severely affecting the model's training and performance.
- (ii) Data has domain bias. from Table 1 and Figure 2(c), the missing values rate and blood consumption have obvious differences in different hospitals. Such a data distribution impacts the cross-hospital generalization of the model. In addition, the collection of clinical information from different hospitals may also be biased due to differences in testing devices and habits.

4. Methodology

4.1. Problem Definitions. The following definition is made in this paper to discuss the role of predictors.

Definition 1. (Blood Consumption Prediction Problem, BCPP). Let $(\mathbf{X}^s, \mathbf{y}^s)$ be a training dataset where clinical information \mathbf{X}^s is implicitly linked to the blood usage \mathbf{y}^s . The BCPP train a model $F(X|\theta)$ with $(\mathbf{X}^s, \mathbf{y}^s)$, and use the model $F(X|\theta)$ predict the blood usage of new collected data of clinical information \mathbf{X}^t , where θ is the model parameter.

The BCPP includes a classification subproblem and regression subproblem. For classification subproblem \mathbf{y}^s and \mathbf{y}^t are one-hot or category values. For regression subproblem \mathbf{y}^s and \mathbf{y}^t are continuous values.

4.2. Blood Data Complementation. Data complementation is the process of replacing missing data with substituted values. When covering for a data point, it is known as “unit complementation;” when substituting for a component of a data point, it is known as “item complementation.” Missing values introduce substantial noise, making data analysis more complex and less efficient. When one or more patient values are missing, most methods discard data with missing values by default, but data complementation attempts to fill in those values. However, missing data are also a reality, and the model should not require that all data be captured well. Therefore, data complementation is introduced to avoid performance degradation with missing values and improve the actual testing data's performance.

In this study, a single data $x_i \in \mathbf{X}^s$ is complemented by

$$\mathbf{x}_i^c = \mathbb{C}(\mathbf{x}_i) = \{\mathbb{C}(\mathbf{x}_{i,1}), \dots, \mathbb{C}(\mathbf{x}_{i,f}), \dots, \mathbb{C}(\mathbf{x}_{i,n})\},$$

$$\mathbb{C}(\mathbf{x}_{i,f}) = \begin{cases} \mathbf{x}_{i,f} & \mathbf{x}_{i,f} \text{ is not missing,} \\ I_{i,f} & \mathbf{x}_{i,f} \text{ is missing,} \end{cases} \quad (1)$$

where $\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,n}$ are n data components of single data \mathbf{x}_i . If any component is missing, the imputed value I_f is used to fill this missing value. The I_f comes from mean value complementation, median value complementation, and KNN complementation:

$$I_{i,f}^{\text{Mean}} = \text{Mean}(\mathbf{X}_f),$$

$$I_{i,f}^{\text{Median}} = \text{Median}(\mathbf{X}_f), \quad (2)$$

$$I_{i,f}^{\text{KNN}} = \text{KNN} \left(\frac{1}{K} \sum_{j \in N_i^{K-NN}} \mathbf{x}_{j,f} \right),$$

TABLE 1: Statistics of datasets.

Hospital name	Hospital abbreviation	Sample size	Average blood usage	Average missing rate	Male/female ratio	Average age
Dongyang Hospital	DYang	95	8.7	0.03	0.34	49.19
Enze Hospital	EZe	13	0.77	0.18	0.86	50.38
Haining Hospital	HNing	57	15.96	0.17	0.68	55.18
Shiyi Hospital	SYi	72	4.65	0.08	0.24	56.40
Shaoyifu Hospital	SYiFu	191	1.46	0.05	0.41	52.50
Shangyu Hospital	SYu	135	7.92	0.07	0.42	51.08
Wenling Hospital	WLing	42	11.62	0.09	0.45	58.98
Xinchang Hospital	XChang	55	11.09	0.03	0.49	57.89
Xiaoshan Hospital	XShan	62	0	0.08	0.44	50.82
Yongkang Hospital	YKang	194	2.36	0.06	0.62	64.43
Yuyao Hospital	YYao	65	9.85	0.03	0.35	54.68
Zheer Hospital	ZEr	1044	1.44	0.03	0.36	53.93

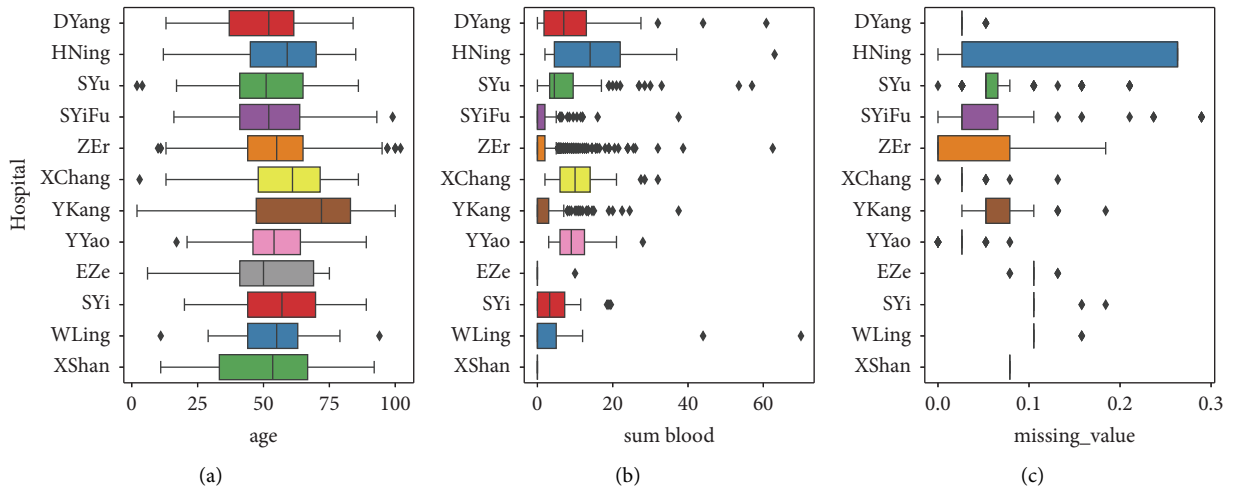


FIGURE 2: Boxplots for the relationship between hospital and age/sum blood/missing value.

where $\text{Mean}(X_f^s)$ and $\text{Median}(X_f^s)$ are the mean value or median value of training datasets on components f . N_i^{K-NN} is the KNN neighborhood of data i in the sense of European distance, $K = 5$ in this paper.

4.3. Cross Hospitals Data Augmentation. Data augmentation is a well-known neural network (NN) training strategy for image classification and signal processing [30]. Data augmentation improves the performance of the methods by precisely fitting the data distribution. First, data augmentation enhances the diversity of data, thereby overcoming overfitting. Second, data augmentation essentially reinforces the fundamental assumption of DR, i.e., the local connectivity of neighborhoods. Finally, it learns refined data distribution by generating more intra-manifold data based on sampled points.

Cross-hospital data augmentation is introduced in a unified framework to generate new data $\mathbf{x}' = \mathbf{T}(\mathbf{x})$:

$$\begin{aligned} \mathbf{x}^a &= \mathbf{T}(\mathbf{x}^C) = \{\tau(\mathbf{x}_{i,1}^C), \dots, \tau(\mathbf{x}_{i,f}^C), \dots, \tau(\mathbf{x}_{i,n}^C)\}, \\ \tau(\mathbf{x}_f^a) &= (1 - r_u) \cdot \mathbf{x}_{i,f}^C + r_u \cdot \bar{\mathbf{x}}_{i,f}^C, \bar{\mathbf{x}} \sim \{N_i^h\}_{h \in H/h_i}, \end{aligned} \quad (3)$$

where the new augmented data \mathbf{x}^a is the combination of each feature $\tau(\mathbf{x}_{i,f}^C)$. $\tau(\mathbf{x}_{i,f}^C)$ is calculated from the linear interpolation of original feature $\mathbf{x}_{i,f}^C$ and augmented feature $\bar{\mathbf{x}}_{i,f}^C$. In addition, the augmented feature $\bar{\mathbf{x}}$ is sampled from the neighborhood $\mathbf{N}(\mathbf{x})$ of \mathbf{x} . N_i^h is the k-NN neighborhood of the data i on the data of hospitals h . H is the set of the hospitals. H/h_i means remove the neighborhood of data i 's hospitals. The combination parameter $r_u \sim U(0, p_U)$ is sampled from the uniform distribution $U(0, p_U)$, and p_U is the hyperparameter.

The cross-hospital augmentation generates new data by combining data i with its neighborhoods in different hospitals. It reduces the terrible influence of the missing data and improves the training data's divergence. Thus, the model learns a precise distribution to enhance the performance of our method. In addition, it works together with the loss function to align data from different hospitals, thus overcoming domain bias.

4.4. Architecture of BUPNN. The proposed BUPNN does not require a unique backbone neural network. The multilayer perceptron network (MLP) is the backbone neural network.

In addition to this, a new network architecture is proposed to enhance generalizability. The proposed neural network architecture is shown in Figure 3.

A typical neural network model uses the network output directly to compute a supervised loss function. It may introduce undesirable phenomena such as overfitting. In this paper, similar to the paper [31], a manifold learning regularizer is proposed to suppress problems such as overfitting using the information in the latent space of the network. As shown in Figure 3, a complete neural network $F(\cdot, w_i, w_j)$ is divided into a latent network $f_i(\cdot, w_i)$ and an output network $f_j(\cdot, w_j)$. The latent network is a preprocessing network that resists the adverse effects of noise and missing value completion. The output network is a dimensional reduction network that maps the data in high-dimensional latent space to the data in low-dimensional latent space. The latent network $f_i(\cdot, w_i)$ maps the data x' input a network latent space and an output network $f_j(\cdot, w_j)$ further map it into the output space:

$$\begin{aligned} y_i &= f_i(x'_i, w_i), \\ y_j &= f_i(x'_j, w_i), \\ z_i &= f_j(y_i, w_j), \\ z_j &= f_j(y_j, w_j), \end{aligned} \quad (4)$$

where x'_i and x'_j are the complementation and augmentation results of origin data x .

Neural networks have powerful fitting performance, but at the same time, there is a risk of overfitting. The typical L2 regularization method can reduce overfitting, but it only limits the complexity of the network without constraining the network in terms of the manifold structure of the data. For example, there is no way to guarantee the distance-preserving and consistency of the network mapping. For this reason, we design a manifold regularizer based on manifold learning to solve this problem, and thus improve the generalization performance of the model and its usability for actual data.

4.5. Loss Function of BUPNN. The loss function of BUPNN consists of two parts; one is the cross-entropy loss which uses label information, and the other is the manifold regularizer loss which uses hospital information and latent space information.

4.5.1. Manifold Regularizer Loss. Manifold regularizer loss handles the domain bias in different hospitals during the training phase and provides a manifold constraint to prevent overfitting. Inconsistent medical equipment and subjective physician diagnoses in different hospitals cause domain bias in data between hospitals. The manifold regularizer loss guides the mapping of the neural networks to be insensitive to hospitals, thus overcoming domain bias (shown in Figure 4). Therefore, the pairwise similarity between nodes is defined first. Considering the dimensional inconsistency of the latent space, we use the t -distribution with variable degrees of freedom as a kernel function to measure the point-pair similarity of the data:

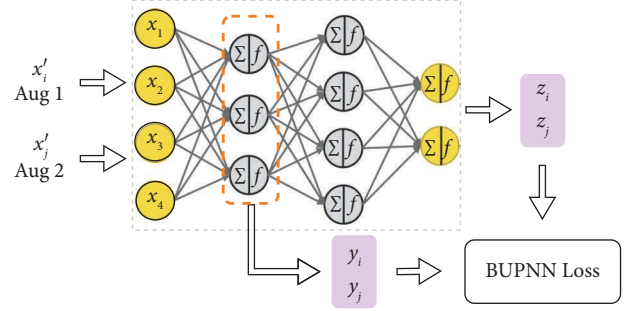


FIGURE 3: Architecture of BUPNN.

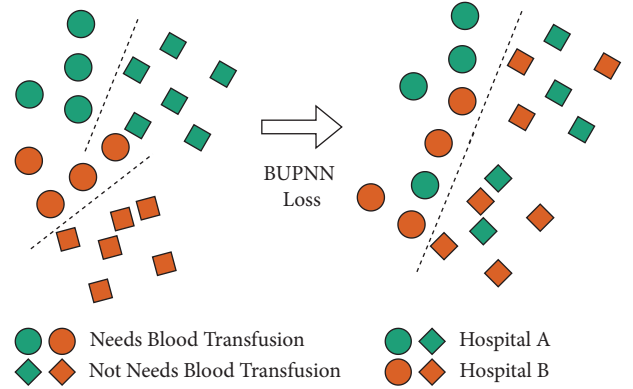


FIGURE 4: How manifold regularizer loss works. Data from the same hospitals are clustered near each other in latent space due to the significant domain bias possessed by the current data. The manifold regularizer loss guides the neural network model to reduce the domain bias by pulling in neighboring nodes across hospitals to mix data from different hospitals.

$$\kappa(d, \nu) = \frac{\text{Gam}(\nu + 1/2)}{\sqrt{\nu\pi}\text{Gam}(\nu/2)} \left(1 + \frac{d^2}{\nu}\right)^{-\nu+1/2}, \quad (5)$$

where $\text{Gam}(\cdot)$ is the Gamma function, and the degree of freedom ν controls the shape of the kernel function. d is the Euclidean pairwise distance of node pairs.

Based on the defined pairwise similarity in a single space, we minimize the similarity difference between two spaces by fuzzy set cross-entropy loss (two-way divergence) [32] $\mathbf{D}(p, q)$:

$$\mathbf{D}(p, q) = p \log q + (1 - p) \log (1 - q), \quad (6)$$

where $p \in [0, 1]$. Notice that $\mathbf{D}(p, q)$ is a continuous version of the cross-entropy loss. In BUPNN, equation (6) is used to guide the pairwise similarity of two latent spaces to fit each other.

Therefore, the loss function of the manifold regularizer is defined as follows:

$$\begin{aligned} L_D &= \sum_{i,j}^B L(x_i, x_j) \\ L(x_i, x_j) &= \begin{cases} \mathbf{D}(1, \kappa(d_{ij}^z, \nu_z)) & \text{if } \mathbf{x}_j = \tau(\mathbf{x}_i), \\ \mathbf{D}(\kappa(d_{ij}^y, \nu_y), \kappa(d_{ij}^z, \nu_z)) & \text{otherwise,} \end{cases} \end{aligned} \quad (7)$$

where B is the batch size and $\mathbf{x}_j = \tau(\mathbf{x}_i)$ indicates whether \mathbf{x}_j is the augmented data of \mathbf{x}_i . If \mathbf{x}_j is the augmented data of \mathbf{x}_i , the loss pulls them together in the latent space; otherwise, the loss keeps a gap between them to preserve the manifold structure. The different degrees of freedom ν_y and ν_z in t -distribution are basic settings according to the dimensions of the space. Equation (7) describes the behavior of a manifold alignment that pairs data collected by different hospitals together in latent space to avoid the detrimental effects caused by domain bias:

$$\begin{aligned} d_{ij}^y &= d(y_i, y_j), \\ d_{ij}^z &= d(z_i, z_j), \end{aligned} \quad (8)$$

where d_{ij}^y and d_{ij}^z are the distance between data node i and node j in spaces \mathbb{R}^{d_y} and \mathbb{R}^{d_z} .

4.5.2. Cross-Entropy Loss. The loss function of the manifold regularizer is essentially an unsupervised term that also requires the use of label information while training the network model. The cross-entropy loss function is introduced simultaneously:

$$L_{CE} = - \sum_{i=1}^N [l_i \ln(\sigma(z_i)) + (1 - l_i) \ln(1 - \sigma(z_i))], \quad (9)$$

where l_i is the label of data node i , $\sigma(z_i)$ is the output of the network model, and $\ln(\cdot)$ is the natural logarithm. When solving the classification subtask, l_i is the category label, and when solving the regression subtask, l_i is the probability label.

The loss function of BUPNN is

$$L = L_D + \beta L_{CE}, \quad (10)$$

where β is a hyperparameter to balance L_D and L_{CE} .

4.6. Pseudocode and Complexity. BUPNN's pseudocode is shown in Algorithm 1. The BUPNN includes the initialization and training phases. In the initialization phase, the kNN neighborhood of every single data is discovered. The time complexity of initialization phases is $O(n^{1.14})$ [33], where n is the number of data. In the training phase, the complexity of the training phases is the same as that of artificial neural networks (ANN). BUPNN calculates the pairwise distance in a batch, so the complexity of training each batch is $O(|B|^2)$, where $|B|$ is the batch size. GPU can well accelerate the pairwise distance, so the training time consumption is the same as that of a typical ANN.

5. Experiments

5.1. Baseline Methods. In this Section, two subtasks, classification subtask and regression subtask, are defined. Seven state-of-the-art baseline classification and regression methods are chosen to discuss the relative advantages of BUPNN. The baseline approach is as follows.

5.1.1. K-Nearest Neighbor Classification/Regression Method (KNN) [37]. The K-nearest neighbor classification/regression method is a nonparametric statistical method. The KNN classification method outputs the prediction by the "majority vote" of its neighbors. The KNN regression method outputs the prediction by the average of its neighbors.

5.1.2. Decision Tree Classification/Regression Method (DT) [35]. A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while, at the same time, an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches representing values for the attribute tested. The leaf node represents a decision on the numerical target. The root node is the topmost decision node in a tree that corresponds to the best predictor. Decision trees can handle both categorical and numerical data.

5.1.3. Random Forest Classification/Regression Method (RF) [34]. Random forests or random decision forests are ensemble learning methods for classification, regression, and other tasks that operate by constructing many decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct decision trees' habit of overfitting their training set.

5.1.4. Extremely Randomized Trees Classification/Regression Method (ET) [38]. Extremely randomized trees add a further step of randomization to random forest, while similar to ordinary random forests in that they are an ensemble of individual trees, there are two main differences: first, each tree is trained using the whole learning sample (rather than a bootstrap sample), and second, the top-down splitting in the tree learner is randomized. Furthermore, instead of computing the locally optimal cut-point for each feature under consideration (based on, e.g., information gain or the Gini impurity), a random cut-point is selected.

5.1.5. Support Vector Machine Classification/Regression Method (SVM) [39]. Support vector machine is a supervised learning model with associated learning algorithms that analyze data for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new standards to one category or the other, making it a nonprobabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). SVM maps training examples to points in space to maximize the width of the gap between the two categories. New models are then mapped into that space and predicted to belong to a class based on which side of the hole they fall.

Input: data: $\mathcal{X} = \{x_i\}_{i=1}^{|\mathcal{X}|}$, learning rate: η , epochs: E , batch size: B, β, γ^z , network: f_θ, g_ϕ ,
Output: graph embedding: $\{e_i\}_{i=1}^{|\mathcal{X}|}$.

- (1) **while** $i = 0; i < E; i ++$ **do**
- (2) $\mathbf{x}^1 = \mathbf{I}(\mathbf{x}) \triangleright$ # blood data complementation in equation (1)
- (3) **while** $b = 0; b < \lceil |\mathcal{X}|/B \rceil; b ++$ **do**
- (4) $\mathbf{x}^{a1}, \mathbf{x}^{a2} = \mathbf{T}(\mathbf{x}^1), \mathbf{T}(\mathbf{x}^1) \triangleright$ # blood data augmentation in equation (3)
- (5) $\mathbf{y}^{a1}, \mathbf{y}^{a2} \leftarrow f_\theta(\mathbf{x}^{a1}), f_\theta(\mathbf{x}^{a2}); \triangleright$ # map input data into \mathbb{R}^{d_y} space in equation (4)
- (6) $\mathbf{z}^{a1}, \mathbf{z}^{a2} \leftarrow f_\theta(\mathbf{y}^{a1}), f_\theta(\mathbf{y}^{a2}); \triangleright$ # map input data into \mathbb{R}^{d_z} space in equation (4)
- (7) $d_{ij}^y \leftarrow d(y^{a1}, y^{a2}); d_{ij}^z \leftarrow d(z^{a1}, z^{a2}); \triangleright$ #calculate distance in \mathbb{R}^{d_y} and \mathbb{R}^{d_z}
- (8) $S^y \leftarrow \kappa(R(B_{ij})d_{ij}^y, \nu_y); S^z \leftarrow \kappa(d_{ij}^z, \nu_z); \triangleright$ #calculate similarity in \mathbb{R}^{d_y} and \mathbb{R}^{d_z} in equation (5)
- (9) $\mathcal{L}_D \leftarrow D(S^y, S^z); \triangleright$ # calculate loss function in equation (10).
- (10) $\theta \leftarrow \theta - \eta \partial \mathcal{L}_D / \partial \theta, \phi \leftarrow \phi - \eta \partial \mathcal{L}_D / \partial \phi; \triangleright$ # update parameters.
- (11) **end while**
- (12) **end while**
- (13) $z_i \leftarrow f_\phi(g_\phi(x_i)); \triangleright$ # calculate the embedding result.

ALGORITHM 1: BUPNN algorithm.

5.1.6. Gradient Boost Classification/Regression Method (GB) [40]. Gradient boosting is a machine learning technique used in regression and classification tasks. It gives a prediction model in an ensemble of weak prediction models, typically decision trees. The resulting algorithm is called gradient-boosted trees. When a decision tree is a weak learner, it usually outperforms a random forest.

5.1.7. Adaptive Boost Classification/Regression Method (ADB) [36]. The adaptive boost algorithm obtains a strong learner by combining a series of weak learners and integrating these vulnerable learners' learning capabilities. Adaptive boost changes the weights of the samples based on the previous learners, increasing the importance of those previously misclassified samples and decreasing the weight of correctly classified samples so that the subsequent learners will focus on those misclassified samples. Finally, these learners are combined into strong learners by weighting.

5.1.8. Light Gradient Boosting Machine Classification/Regression Method (LightGBM) [41, 42]. LightGBM is a distributed gradient boosting framework based on the decision tree algorithm. LightGBM is designed with two main ideas in mind: (1) to reduce the use of data in memory to ensure that a single machine can use as much data as possible without sacrificing speed and (2) to reduce the cost of communication to improve the efficiency when multiple machines are in parallel, and achieve linear acceleration in computation. It can be seen that LightGBM was originally designed to provide a fast and efficient data science tool with a low memory footprint, high accuracy, and support for parallel and large-scale data processing.

5.2. Dataset Partitioning and Grid Search. Table 1 and Figure 2 provide basic information about the data. Five data partitioning schemes are provided in this study to provide a detailed comparison of the performance differences between the different schemes.

Three schemes are with data-complement, including COM-Mea, COM-Med, and COM-KNN as defined in equation (2). First, for COM-Mea, COM-Med, and COM-KNN, the missing values are complemented with a specific method. Following that, the training and testing sets are divided by 90%/10%. Two noncomplement schemes (NC-A and NC-B) are introduced to compare with the complement schemes. NC-A deletes all the data items with missing values, following the typical machine learning scheme. Following the data cleaning, NC-A divides the training and testing sets by 90%/10%. NC-B keeps the same training and testing set division as the data complement schemes and only removes all missing data from the training set and obtains a cleaner training set.

The models are evaluated with 10-fold cross-validation for all the training sets and determine the optimal parameters by grid search. For a fair comparison, we control the search space of all baseline methods to be approximately equal. The search space of the compared process is in Table 2.

5.3. Experimental Setup. We initialize the other NN with the Kaiming initializer. We adopt the AdamW optimizer [43] with learning rate of 0.02 and weight decay of 0.5. All experiments use a fixed MLP network structure, $f_{\theta,w}: (-1, 500, 300, 80)$, $g_\phi: (80, 500, 2)$, where -1 is the features number of the dataset. The number of epochs is 400. The batch size is 2048. $\gamma^y = 100$.

5.4. Comparison on Classification Subtask. Although the blood data for each patient is collected for accurate regression subtasks, it is equally important to predict whether a patient needs a blood transfusion. Especially in emergencies, indicating whether a patient needs a blood transfusion in the short term is more important than estimating the amount of blood used throughout the treatment cycle. Therefore, we first evaluate the performance of the proposed BUPNN on the classification subtask. Then, two evaluation metrics, classification accuracy (ACC) and area under the ROC curve (AUC), are used to compare the classifier's

TABLE 2: Details of grid search.

Methods	Abbreviation	Search space	Note
K-nearest neighbor	KNN	$Nei \in [1, 3, 5, 10, 15, 20]$, $L \in [10, 20, 30, 50, 70, 100]$	$Nei \rightarrow$ neighbor.size, $L \rightarrow$ leaf.size
Random forest	RF	$NE \in [80, 90, 100, 110, 120, 130]$, $MSS \in [2, 3, 4, 5, 6, 7]$	$NE \rightarrow$ boosted.trees.size, $MSS \rightarrow$ samples.split.size
Decision tree	DT	$MSL \in [1, 2, 3, 5, 7, 10]$, $MSS \in [2, 3, 5, 7, 10, 15]$	$MSL \rightarrow$ sample.size.inaleaf, $MSS \rightarrow$ samples.split.size
Extra tree	ET	$MSL \in [1, 2, 3, 5, 7, 10]$, $MSS \in [2, 3, 5, 7, 10, 15]$	$MSL \rightarrow$ sample.size.inaleaf, $MSS \rightarrow$ samples.split.size
Support vector machine	SVM	$M \in [10, 50, 100, 300, 500]$, $T \in [1e^{-4}, 5e^{-4}, 1e^{-3}, 2e^{-3}, 3e^{-3}, 5e^{-3}]$	$NE \rightarrow$ max iterations, $T \rightarrow$ tolerance.for.stopping.criteria
Gradient boosting	GB	$NE \in [80, 90, 100, 110, 120, 130]$, $MSS \in [2, 3, 4, 5, 6, 7]$	$NE \rightarrow$ boosted.trees.size, $MSS \rightarrow$ samples.split.size
Multilayer perceptron	MLP	$L \in [2, 3, 4, 5, 6]$, $WD \in [0.1, 0.2, 0.3, 0.4]$	$L \rightarrow$ number.of.layer $WD \rightarrow$ weight.decay
Adaptive boost	ADB	$NE \in [40, 50, 60, 70, 80, 90]$, $LR \in [0.8, 0.9, 1, 2, 5,]$	$NE \rightarrow$ boosted.trees.size, $LR \rightarrow$ learning.rate\end
Light gradient boosting machine	LGBM	$NE \in [21, 26, 31, 26, 41]$, $L \in [80, 90, 100, 110, 120, 130]$	$NE \rightarrow$ boosted.trees.size, $L \rightarrow$ leaf.size\end
Blood usage prediction neural network	BUPNN (ours)	$\gamma^z \in [0.01, 0.03, 0.05, 0.07]$ $\beta \in [0.1, 0.2], K \in [100, 200, 300]$	$NE \rightarrow$ boosted.trees.size, $L \rightarrow$ leaf.size\end

performance from multiple perspectives. The performance comparison of BUPNN and eight baseline methods is shown in Tables 3 and 4. In addition, the ROC curves of all the compared methods on different schemes are shown in Figure 5. The scatter plot the prediction of BUPNN for COM-Mea schemes are shown in Figure 6.

5.4.1. Data Complementation Delivers Performance Improvements. From the performance results of NC-B, COM-Mid, COM-Mea, and COM-KNN in Tables 3 and 4, we observe that the schemes with data complementation yield better performance. We attribute the reasons to the more abundant training data provided by data complementation. Although the complemented data are imperfect, artificial intelligence models can still learn more helpful information.

Although the model trained with only clean data (NC-A) has the highest performance score, scheme NC-A only includes the clean data and cannot be conveniently generated to the actual data with missing values. However, missing values are unavoidable in real-world application scenarios, so data complementation is a better choice to improve the model’s performance.

5.4.2. BUPNN Outperforms Almost All Baseline Methods. From Tables 3 and 4, we observe that the proposed BUPNN has advantages over all the baseline methods. For the AUC metric, BUPNN has the advantage in all five schemes. However, the COM-MID scheme has the most significant benefits, outperforming the second-best method (LGBM) by 3.71%. For the ACC metric, BUPNN has the advantage in all five schemes. However, the COM-KNN scheme has the most significant benefits, outperforming the second-best method (LGBM) by 2.95%. BUPNN has a 1% average advantage over the second-best

method in both metrics. The reason is attributed to the fact that BUPNN is a neural network-based model, which performs better with richer data. In addition, the proposed manifold loss function can improve the model’s generalization performance and thus enhance the performance of the testing set.

5.4.3. BUPNN Is Better at Handling Complemented Datasets. From four schemes with the same testing set (NC-B, COM-Mid, COM-Mea, and COM-KNN) in Tables 3–5, we observe that BUPNN performs better than all the baseline methods when handling complemented datasets. We attribute this to the data augmentation of the proposed BUPNN model. Data augmentation generates new training data from the completed data and attenuates the effect of missing values, thus guiding BUPNN to learn a smooth manifold.

5.4.4. BUPNN Has Better Generalizability between Different Hospitals. To evaluate the generalization performance of our proposed BUPNN and baseline methods, we tested the ACC with the data from various hospitals (Figure 5). The proposed BUPNN has a dominant performance of five hospitals out of a total of 6 hospitals and has a clear advantage over Shaoyifu Hospital. From Table 1 and Figure 1, we observe that the Shaoyifu Hospital has a relatively obvious domain bias. It has a minimum male-to-female rate, the top three average missing rates, and many outliers in its missing values. We argue that the domain bias influences the performance of the baseline methods, and the proposed BUPNN outperforms the baseline methods due to BUPNN overcoming the domain bias in it. Furthermore, the manifold regularizer loss function provides a good manifold constraint to improve the model’s generalizability between different hospitals (as shown in Figure 4).

TABLE 3: Classification AUC comparison with the baseline methods, the best result is shown in bold. The second result is italicized. The brackets at the right end show how much BUPNN exceeds the optimal metrics in the other methods.

	KNN	RF	MLP	ET	SVM	GB	AdaB	LGBM	BUPNN
NC-A	0.8072	<i>0.9166</i>	0.8436	0.8443	0.8736	0.8949	0.9072	0.8881	0.9229 ($\uparrow 0.0063$)
NC-B	0.7302	<i>0.8109</i>	0.7421	0.8178	0.7116	0.7214	0.8324	0.8119	0.8349 ($\uparrow 0.0240$)
COM-mid	0.8009	<i>0.8526</i>	0.7764	0.8437	0.7435	0.8442	0.842	0.8508	0.8843 ($\uparrow 0.0317$)
COM-mea	0.8054	0.8591	0.8399	0.8252	0.7553	0.8470	0.8562	<i>0.8630</i>	0.8797 ($\uparrow 0.0167$)
COM-KNN	0.8033	0.8575	0.7912	0.8321	0.7739	0.8446	0.8526	<i>0.8620</i>	0.8761 ($\uparrow 0.0141$)
Average	0.7894	<i>0.8593</i>	0.7992	0.8326	0.7716	0.8304	0.8581	0.8552	0.8796 ($\uparrow 0.0202$)

TABLE 4: Classification ACC comparison with the baseline methods, the best result is shown in bold. The second result is italicised. The brackets at the right end show how much BUPNN exceeds the optimal metrics in the other methods.

	KNN	RF	MLP	ET	SVM	GB	AdaB	LGBM	BUPNN
NC-A	0.7113	0.8351	0.7367	0.8144	0.7732	0.8454	0.8557	0.8454	<i>0.8454</i> ($\downarrow 0.0103$)
NC-B	0.6539	0.7531	0.7512	0.7428	0.6235	0.7409	0.7557	0.7445	0.7643 ($\uparrow 0.0086$)
COM-mid	0.7340	<i>0.8030</i>	0.7589	0.7734	0.5567	0.7931	0.7537	0.7783	0.8177 ($\uparrow 0.0147$)
COM-mea	0.7340	0.7931	0.7546	0.7635	0.6305	0.798	0.7931	<i>0.8030</i>	0.8177 ($\uparrow 0.0147$)
COM-KNN	0.7241	<i>0.7931</i>	0.7768	0.7685	0.6059	0.7734	0.7783	0.7734	0.8226 ($\uparrow 0.0295$)
Average	0.7115	<i>0.7955</i>	0.7560	0.7725	0.6380	0.7902	0.7873	0.7890	0.8135 ($\uparrow 0.0180$)

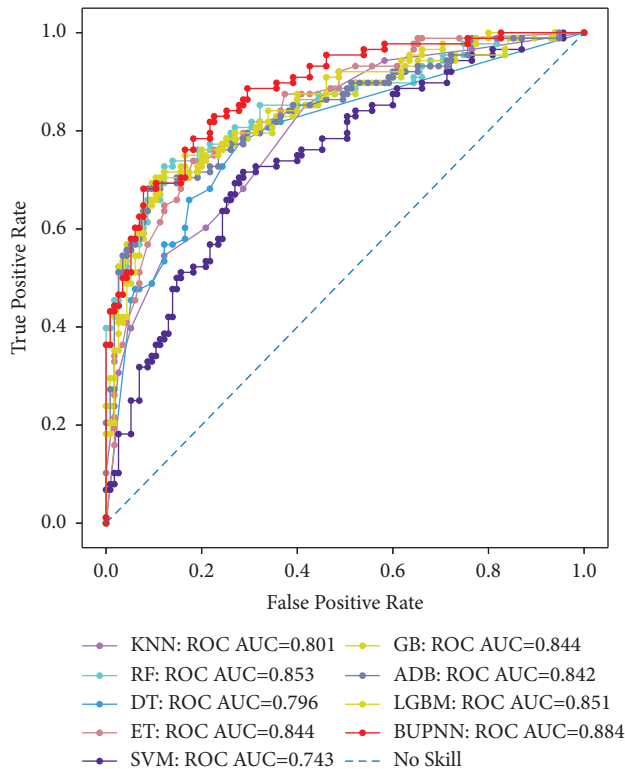


FIGURE 5: The ROC curve comparison for COM-Mea schemes, the missing values are filled with the median value. The closer the curve is to the upper left corner, the better the model's performance. The symmetry of the curve along the line from (0, 1) to (1, 0) indicates the balanced performance of the model.

5.5. Comparison on Regression Subtask. Then, we discuss the performance of BUPNN on the regression subtask. Finally, the regression model is asked to predict the total blood usage in the patient's treatment for advance scheduling at

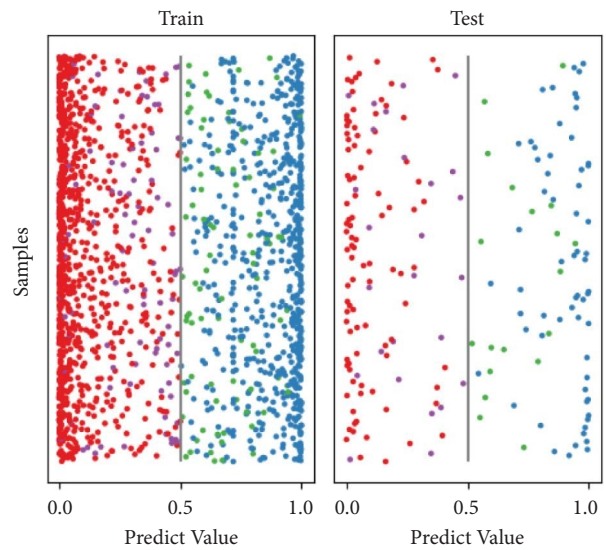


FIGURE 6: The scatter plot the prediction of BUPNN for COM-Mea schemes, the missing values are filled with the median value. The vertical coordinate is used to distinguish different samples, and the horizontal coordinate indicates the predicted value of the BUPNN model for the samples. A predicted value of less than 0.5 indicates that no blood is needed, and a value greater than 0.5 indicates that blood is needed. The color of the scatter indicates whether the prediction is correct or not.

the blood center. The experimental setting of the regression subtask is the same as that of the classification subtask. The data are collected from the same patients, but the target variable is the total blood usage in the patient's treatment. The mean square error (MSE) metric is calculated to measure the performance of the seven methods in the five schemes. The performance comparison is shown in Table 6.

TABLE 5: Median classification ACC comparison with the baseline methods for different hospitals’ all data, the best result are shown in bold. The second result is italicised. The brackets at the right end show how much BUPNN exceeds the optimal metrics in the other methods.

	KNN	RF	ET	MLP	GB	LGBM	SVMVC	ADB	BUPNN
DYang	1.000	1.000	0.571	0.723	0.786	0.714	0.857	0.929	1.000 ($\uparrow 0.000$)
SYi	1.000	0.667	1.000	1.000	1.000	0.667	1.000	1.000	1.000 ($\uparrow 0.000$)
SYiFu	0.796	0.778	0.815	0.735	0.556	0.593	0.630	0.778	0.926 ($\uparrow 0.111$)
WLing	0.500	1.000	0.750	0.250	1.000	1.000	1.000	1.000	1.000 ($\uparrow 0.000$)
ZEr	0.748	0.800	0.743	0.786	0.707	0.833	0.828	0.801	0.834 ($\uparrow 0.001$)
Average	0.765	0.834	0.764	0.698	0.775	0.750	0.833	0.854	0.909 ($\uparrow 0.055$)

TABLE 6: MSE comparison for all data, best results are shown in bold; results with clear advantage are shown in underline. The second result is italicised. The brackets at the right end show how much BUPNN exceeds the optimal metrics in the other methods.

	LR	SVM	MLP	ET	GB	RF	LGBM	Adab	BUPNN
NC-A	0.0095	0.0112	0.0121	0.0194	0.0093	<i>0.0090</i>	0.0090	0.0161	0.0079 ($\uparrow 0.0012$)
COM-mid	<i>0.0042</i>	0.0075	0.0167	0.0103	0.0040	0.0043	0.0042	0.0175	0.0038 ($\uparrow 0.0004$)
COM-mea	0.0042	0.0076	0.0076	0.0074	<i>0.0039</i>	0.0041	0.0044	0.0151	0.0037 ($\uparrow 0.0002$)
COM-KNN	0.0041	0.0065	0.0082	0.0142	0.0040	<i>0.0039</i>	0.0043	0.0109	0.0036 ($\uparrow 0.0003$)
Average	0.0055	0.0082	0.0111	0.0128	<i>0.0053</i>	0.0053	0.0054	0.0149	0.0048 ($\uparrow 0.0005$)

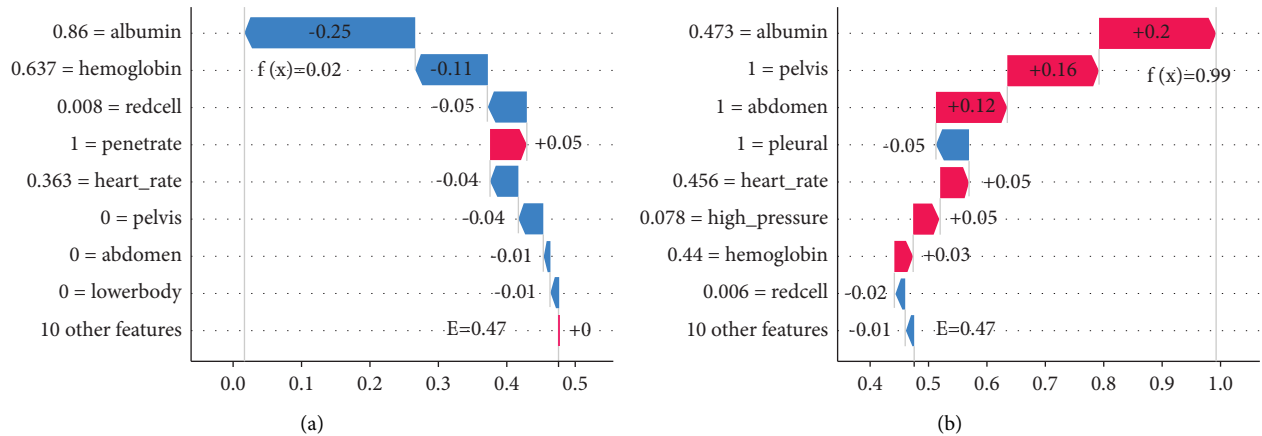


FIGURE 7: Explanation of the decision process of BUPNN for a single sample.

5.5.1. *BUPNN Shows a Consistency Advantage on Regression Subtasks.* From Table 6, we observe that the proposed BUPNN outperforms all the baseline methods on all four schemes. BUPNN has the most significant benefits for the COM-mid scheme, outperforming the second-best method (LGBM) by 0.004 (10.5%). The percentage of improvement shows that the advantage of BUPNN is more evident in the regression subtask. Furthermore, it indicates that the proposed BUPNN is more suitable for handling more difficult regression tasks and that BUPNNs have strong application potential when collecting richer data.

5.6. *The Explanatory Analysis of BUPNN.* Artificial neural network-based models are considered to have a strong performance. Still, their black-box nature leads to difficulty in interpretability, thus making it difficult for the model to be trusted by domain experts. On the other hand, the proposed BUPNN has stronger interpretability because the smoothness of its mapping leads to models that profound model

interpreters can easily interpret. An easy-to-use neural network interpretive tool is introduced to explain the decision process of BUPNN. The visual presentation of the interpretation results is shown in Figure 7.

In Figure 7, the decision processes of BUPNN for three samples are explained by “shap” tool. The vertical coordinates represent the clinical indicators, and the horizontal coordinates represent the predicted values of BUPNN. The images show the decision process of the model from bottom to top. Specifically, the model predicts $E = 0.47$ for an average patient when no information about the patient is observed, i.e., no transfusion is needed. We believe this is reasonable because only patients who are sufficiently harmed need blood transfusions to be treated. Next, using Figure 7(a) as an example, the BUPNN model observed additional patient information, such as “no injury to the lower body,” “no injury to the abdomen,” “no injury to the pelvis,” and the “patient’s heartbeat is not accelerated.” This evidence made the BUPNN firm the original judgment that blood transfusion is unnecessary. Although the penetration

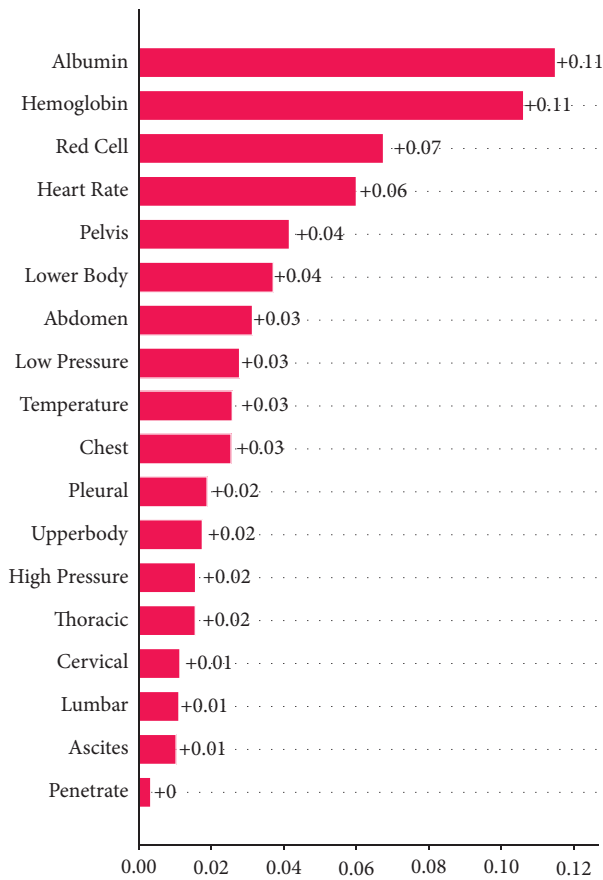


FIGURE 8: Important indicators about blood transfusion obtained by interpretable analysis of BUPNN.

injury raised the probability of transfusion, several essential features (e.g., albumin and hemoglobin) ultimately guided the model to predict $f(x) = 0.02$.

Figure 7(b) provides more examples of how to interpret the decision process of the BUPNN. These two examples show what kind of information needs to be observed by BUPNN that would predict a sample as needing a blood transfusion. Among them, we found some necessary signatures of the need for transfusion, such as the abdomen, pelvis, and pleural.

Next, we calculated the importance of all features in determining whether a blood transfusion is needed or not and displayed them in Figure 8 in the form of a bar chart.

6. Conclusions

In this paper, a neural network model-based blood usage predictor, called deep blood usage neural network (BUPNN), is proposed to serve the scheduling of blood supply in provincial blood centers. The proposed BUPNN receives clinical information from hospital patients and predicts whether a patient needs blood and the amount of blood used. The proposed BUPNN receives clinical data from hospital patients and indicates whether a patient needs blood and the blood consumption amount. The proposed BUPNN mainly solves the problem of predicting blood

usage with high availability and generalization in real-life situations. It can accomplish the prediction task well in different hospitals' missing values and domain bias. BUPNN proposes a manifold learning-based regularizer for the blood prediction problem to improve the model's generalization performance on data from different hospitals and enhance the model using data augmentation and data complementation.

Data Availability

The blood data used to support the findings this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Science Research Foundation of Zhejiang Province (LGF20G030002) "Research on Assisted Decision-making Model of Blood Supply in Public Emergencies".

References

- [1] A. B. Cardillo, J. M. Heal, K. F. Henrichs et al., "Reducing the need for hla-matched platelet transfusion," *New England Journal of Medicine*, vol. 384, no. 25, pp. 2451-2452, 2021.
- [2] H. Hu, X. Lai, C. Tan, N. Yao, and L. Yan, "Factors associated with in-patient mortality in the rapid assessment of adult earthquake trauma patients," *Prehospital and Disaster Medicine*, vol. 37, no. 3, pp. 299-305, 2022.
- [3] O. P. S. P. Sickle Cell Anemia (Stop 2) Trial Investigators, "Discontinuing prophylactic transfusions used to prevent stroke in sickle cell disease," *New England Journal of Medicine*, vol. 353, no. 26, pp. 2769-2778, 2005.
- [4] H. Wang, J. Umejiego, R. D. Robinson et al., "A derivation and validation study of an early blood transfusion needs score for severe trauma patients," *Journal of Clinical Medicine and Research*, vol. 8, no. 8, p. 591, 2016.
- [5] T. C. Nunez, I. V. Voskresensky, L. A. Dossett, R. Shinall, W. D. Dutton, and B. A. Cotton, "Early prediction of massive transfusion in trauma: simple as abc (assessment of blood consumption)?" *Journal of Trauma and Acute Care Surgery*, vol. 66, no. 2, pp. 346-352, 2009.
- [6] R. Schroll, D. Swift, D. Tatum et al., "Accuracy of shock index versus abc score to predict need for massive transfusion in trauma patients," *Injury*, vol. 49, no. 1, pp. 15-19, 2018.
- [7] Y. Sun, S. Selvarajan, Z. Zang et al., "Protein classifier for thyroid nodules learned from rapidly acquired proteotypes," *medRxiv*, 2020.
- [8] F. Saberi-Movahed, M. Mohammadifard, A. Mehrpooya et al., "Decoding clinical biomarker space of covid-19: exploring matrix factorization-based feature selection methods," *Computers in Biology and Medicine*, vol. 146, Article ID 105426, 2022.
- [9] K. Zhou, Y. Sun, L. Li et al., "Eleven routine clinical features predict covid-19 severity uncovered by machine learning of longitudinal measurements," *Computational and Structural Biotechnology Journal*, vol. 19, no. 2001-0370, pp. 3640-3649,

- 2021, <https://www.sciencedirect.com/science/article/pii/S2001037021002609>.
- [10] Z. Zhang, S. Cheng, L. Lu et al., "EVNet: an explainable deep network for dimension reduction," in *IEEE Transactions on Visualization and Computer Graphics*, 2022.
 - [11] A. Mehrpooya, F. Saberi-Movahed, A. N. M. Rezaei-Ravari, M. Eftekhari, and I. Tavassoly, "High dimensionality reduction by matrix factorization for systems pharmacology," *bioRxiv*, vol. 23, no. 1, 2021.
 - [12] F. Saberi-Movahed, M. Eftekhari, and M. Mohtashami, "Supervised feature selection by constituting a basis for the original space of features and matrix factorization," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 1405–1421, 2020.
 - [13] Z. Zang, Y. Xu, Y. Geng, S. Li, and S. Z. Li, "Udrn: unified dimensional reduction neural network for feature selection and feature projection," 2022, <http://arxiv.org/abs/2207.03809>.
 - [14] G. M. Frankfurter, K. E. Kendall, and C. C. Pegels, "Management control of blood through a short-term supply-demand forecast system," *Management Science*, vol. 21, no. 4, pp. 444–452, 1974.
 - [15] G. C. Critchfield, D. P. Connelly, M. S. Ziehwein, L. S. Olesen, C. E. Nelson, and E. P. Scott, "Automatic prediction of platelet utilization by time series analysis in a large tertiary care hospital," *American Journal of Clinical Pathology*, vol. 84, no. 5, pp. 627–631, 1985.
 - [16] O. S. S. Filho, W. Cezarino, and G. R. Salviano, "A decision-making tool for demand forecasting of blood components," *IFAC Proceedings Volumes*, vol. 45, no. 6, pp. 1499–1504, 2012.
 - [17] O. S. S. Filho, M. A. Carvalho, W. Cezarino, R. Silva, and G. Salviano, "Demand forecasting for blood components distribution of a blood supply chain," *IFAC Proceedings Volumes*, vol. 46, no. 24, pp. 565–571, 2013.
 - [18] D. Kumari and A. Wijayanayake, "An efficient inventory model to reduce the wastage of blood in the national blood transfusion service," in *Proceedings of the 2016 Manufacturing & Industrial Engineering Symposium (MIES)*, pp. 1–4, IEEE, Colombo, Sri Lanka, October 2016.
 - [19] S. M. Fortsch and E. A. Khapalova, "Reducing uncertainty in demand for blood," *Operations Research for Health Care*, vol. 9, pp. 16–28, 2016.
 - [20] F. Lestari, U. Anwar, N. Nugraha, and B. Azwar, "Forecasting demand in blood supply chain (case study on blood transfusion unit)," *Proceedings of the World Congress on Engineering*, vol. 2, pp. 16–28, 2017.
 - [21] T. Volken, A. Buser, D. Castelli et al. "Red blood cell use in Switzerland: trends and demographic challenges," *Blood transfusion*, vol. 16, no. 1, pp. 73–82, 2018.
 - [22] A. Drackley, K. B. Newbold, A. Paez, and N. Heddle, "Forecasting ontario's blood supply and demand," *Transfusion (Philadelphia)*, vol. 52, no. 2, pp. 366–374, 2012.
 - [23] R. Khaldi, A. El Afia, R. Chiheb, and R. Faizi, "Artificial neural network based approach for blood demand forecasting: fez transfusion blood center case study," in *Proceedings of the 2nd International Conference on Big Data*, pp. 1–6, Cloud and Applications, Association for Computing Machinery, Tetouan, Morocco, March 2017.
 - [24] L. Guan, X. Tian, S. Gombard et al., "Big data modeling to predict platelet usage and minimize wastage in a tertiary care system," *Proceedings of the National Academy of Sciences*, vol. 114, no. 43, pp. 11368–11373, 2017.
 - [25] N. Li, F. Chiang, D. G. Down, and N. M. Heddle, "A Decision Integration Strategy for Short-Term Demand Forecasting and Ordering for Red Blood Cell Components," *Operations Research for Health Care*, vol. 29, 2020.
 - [26] M. Motamedi, N. Li, D. G. Down, and N. M. Heddle, "Demand forecasting for platelet usage: from univariate time series to multivariate models," arXiv: Learning, 2021.
 - [27] C. Twumasi and J. Twumasi, "Machine learning algorithms for forecasting and backcasting blood demand data with missing values and outliers: a study of tema general hospital of Ghana," *International Journal of Forecasting*, vol. 38, no. 3, pp. 1258–1277, 2022.
 - [28] M. Abolghasemi, B. Abbasi, T. Babaei, and S. Z. Hosseinifard, "How to Effectively Use Machine Learning Models to Predict the Solutions for Optimization Problems: Lessons from Loss Function," 2021, <https://arxiv.org/abs/2105.06618>.
 - [29] Z. Zang, W. Wang, Y. Song et al., "Hybrid deep neural network scheduler for job-shop problem based on convolution two-dimensional transformation," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.
 - [30] D. A. Van Dyk and X. L. Meng, "The art of data augmentation," *Journal of Computational & Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.
 - [31] Z. Zang, S. Li, D. Wu et al., "Dlme: Deep Local-Flatness Manifold Embedding," in *Computer Vision-ECCV 2022: 17th European Conference*, pp. 576–592, Springer, Berlin, Germany, 2022.
 - [32] S. Z. Li, Z. Zang, and L. Wu, "Deep Manifold Transformation for Dimension Reduction," 2020, <https://arxiv.org/abs/2010.14831>.
 - [33] L. McInnes and J. Healy, "Umap: uniform manifold approximation and projection for dimension reduction," 2018, <https://arxiv.org/abs/1802.03426>.
 - [34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
 - [35] L. I. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees. wadsworth," *Biometrics*, vol. 40, p. 358, 1984.
 - [36] C. C. Chang and C. J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.
 - [37] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, *Knn model-based approach in classification*, pp. 986–996, Lecture Notes in Computer Science, Springer, Berlin, Germany, 2003.
 - [38] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
 - [39] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
 - [40] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting algorithms as gradient descent," *Advances in Neural Information Processing Systems*, vol. 12, 1999.
 - [41] Q. Meng, G. Ke, T. Wang et al., "A communication-efficient parallel algorithm for decision tree," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
 - [42] G. Ke, Q. Meng, T. Finley et al., "Lightgbm: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
 - [43] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, <https://arxiv.org/abs/1711.05101>, 2017.

Research Article

IGPred-HDnet: Prediction of Immunoglobulin Proteins Using Graphical Features and the Hierarchical Deep Learning-Based Approach

Zakir Ali ¹, Fahad Alturise ², Tamim Alkhalifah ² and Yaser Daanial Khan ¹

¹Department of Computer Science, School of Science and Technology, University of Management and Technology, Lahore, Pakistan

²Department of Computer, College of Science and Arts in Ar Rass, Qassim University, Ar Rass, Qassim, Saudi Arabia

Correspondence should be addressed to Tamim Alkhalifah; tkhliefh@qu.edu.sa

Received 8 July 2022; Revised 16 September 2022; Accepted 12 October 2022; Published 25 January 2023

Academic Editor: Xiaolong Zhou

Copyright © 2023 Zakir Ali et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Motivation. Immunoglobulin proteins (IGP) (also called antibodies) are glycoproteins that act as B-cell receptors against external or internal antigens like viruses and bacteria. IGPs play a significant role in diverse cellular processes ranging from adhesion to cell recognition. IGP identifications via the in-silico approach are faster and more cost-effective than wet-lab technological methods. **Methods.** In this study, we developed an intelligent theoretical deep learning framework, “IGPred-HDnet” for the discrimination of IGPs and non-IGPs. Three types of promising descriptors are feature extraction based on graphical and statistical features (FEFS), amphiphilic pseudo-amino acid composition (Amp-PseAAC), and dipeptide composition (DPC) to extract the graphical, physicochemical, and sequential features. Next, the extracted attributes are evaluated through machine learning, i.e., decision tree (DT), support vector machine (SVM), k-nearest neighbour (KNN), and hierarchical deep network (HDnet) classifiers. The proposed predictor IGPred-HDnet was trained and tested using a 10-fold cross-validation and independent test. **Results and Conclusion.** The success rates in terms of accuracy (ACC) and Matthew’s correlation coefficient (MCC) of IGPred-HDnet on training and independent dataset (D_{train} D_{test}) are ACC = 98.00%, 99.10%, and MCC = 0.958, and 0.980 points, respectively. The empirical outcomes demonstrate that the IGPred-HDnet model efficacy on both datasets using the novel FEFS feature and HDnet algorithm achieved superior predictions to other existing computational models. We hope this research will provide great insights into the large-scale identification of IGPs and pharmaceutical companies in new drug design.

1. Introduction

Immunoglobulins are serum proteins in the human body. These proteins act as an antibody involved in the various cellular processes such as a decision, binding, or recognition of the cell. Immunoglobulin significantly boosts the immune system by discovering the dangerous macromolecules that entered the body [1]. When unfamiliar elements inject into the body, the immune system has a unique skill to detect the attacker and then activates B lymphocytes to hide the immunoglobulin from invader antigens. For instance, immunoglobulins will deactivate the toxin by altering its chemical

structure when averting its appearance. To provide a shield against bacterial infection, stabilin-2 can attach to both Gram-positive and Gram-negative bacterial contagions.

Immunoglobulins are linked/related to various disease treatments [2], such as autoimmune, inflammation in the skin, and Bechet’s diseases [3, 4]. In other words, intravenous immunoglobulin provides a fighting strength to cure such kinds of diseases for people who have suffered from muscle problems and systemic swelling in skin infections. The use of immunoglobulin for lupus erythematosus dermatosis in association with the treatment of Bechet’s infection has a great potential without any harmful impact

[3, 4]. In Ref. [5], it is shown that immunoglobulins have a better understanding of immunological processes, permitting the development of an enhanced version of drugs to cure the infection. Considering the medical application of immunoglobulin proteins, in-depth knowledge of their functional level is still under development.

Over the past years, immunoglobulin protein classification and characterization have become a hot topic in bioinformatics and computational biology. Wet-lab approaches such as X-ray crystallography and mass spectrometry are used to discover immunoglobulin proteins. However, such laboratory-based approaches are unfavourable due to their high cost and time consumption. In this regard, researchers have designed various machine learning-based methods to identify immunoglobulin protein sequence analysis. Efficient machine learning-based methods can quickly and accurately predict unannotated proteins from large databases. Machine learning techniques are applied in numerous areas of medicine like diagnostics. Clonal dynamics and relative frequencies are utilized to develop an antibody clonal examining framework to explore certain antigenic human monoclonal antibodies [5–7]. In the various field of the healthcare system, immunological and biological usage, including infection control, immunization diagnostics, and B-cell detection, is of key significance [8, 9]. The research community has reported numerous studies related to antigen range that can be selected by specific antibodies or by a group of antibodies, e.g., antibody stock provided by applying a Rep-Seq in many areas [10]. The said key observation headed to another and well-defined technique for tackling the B-cell epitope detection in which the intellectual purpose of a specific antibody is detected [11, 12]. This study incorporates optical, electrochemical, and piezoelectric biosensors to predict complete immunoglobulin degrees, in which electrochemical is most generally employed. Several immunoglobulin optical biosensors depend on surface plasmon resonance (SPR) prediction present in buffer solutions. For an immunoglobulin study, these available state-of-the-art technologies are useful; however, conducting the biochemical study is very expensive in terms of money and time. For accurate and speedy execution of a huge amount of protein data, it is a need of time to develop a computational framework for immunoglobulins. For example, the first phase declares the purpose of immunoglobulins proteins which design a useful and inexpensive framework to predict them efficiently. The research community has designed various frameworks based on machine learning procedures for protein sequence analysis and classification in the last decades [13–17]. In bioinformatics, predicting immunoglobulins transforms protein sequences into feature metrics to uncover the core formation of proteins. The essential characteristics of protein prediction are itemized as follows: feature representation and key feature selection based on their importance and classification. Amino acid composition (AAC), dipeptides (Dip), and tripeptides are feature extraction techniques to extract n -gram features representation, where the occurrence of n -length peptides are utilized as feature matrices [18–20].

Furthermore, another feature extraction method pseudo-amino acid composition (PseAAC), is commonly implemented, considering physicochemical properties among residues [15, 17, 21–23]. The pseudotype protein structure led to a protein density drop in dyscalculia; for this purpose, the notion of pseudo- K -tuple is combined with the idea of PseAAC [24, 25] to design a framework of AAC minimized with pseudo- K -tuples amino acid composition (PseKRAAC) [26]. They developed a classifier IGPred by considering nine (9) physicochemical properties of amino acid-generated proteins with replica ACC [27, 28]. In Ref. [29], a predictor was developed via a support vector machine (SVM) to predict immunoglobulins and non-immunoglobulins. They used PseAAC with nine physical and chemical characteristics of amino acids; A cross-validation technique was used to train a model, and they got 96.3% accuracy. However, the performance is good but still needs an efficient bioinformatics tool to predict immunoglobulin with a less error rate.

Various feature representations and multifaced prediction methods may produce unnecessary knowledge representation [30, 31]. However, to deal with this problem, many studies suggested feature selection algorithms for eliminating unnecessary information to enhance the performance of the prediction methods. The first one is PCC, which stands for Pearson's correlation coefficient, used to measure the significance of feature representation in a subgroup. In contrast, the second part is related to computing the repetition among features representation by using Euclidean distance (ED), cosine distance (CD), and Tanimoto (TO). Maximum-Relevance-Maximum-Distance in [32, 33] and Analysis of Variance (ANOVA) in [34] are typical feature selection approaches. For optimum feature representation, [35–37] used the principal component analysis (PCA) and misclassification error (MCE) to extract optimal feature representation for pentatricopeptide-repeat proteins prediction and got 97.9% accuracy. Li et al. in [33] used the above method to design a model for the prediction of anticancer peptide sequences with 19-dimensional attributes.

Although significant contribution has been devoted to the prediction of IGPs, some shortcomings should be acknowledged in terms of feature-encoding schemes and learning models. One major limitation of the existing methods is the lack of feature learning algorithms to extract the structured pattern information from protein sequences properly. Secondly, only machine learning classifiers are not accurate enough to discriminate IGPs from non-IGPs. Thirdly, the developed immunoglobulin predictors only showed the training dataset results using a cross-validation test while ignoring the external/independent test results. Independent test results are significant as they show the trained model's generalization power.

To our best knowledge, IGPred-HDnet is the first deep learning-based predictor for identifying IGPs. IGPred-HDnet extracts the nominal feature vectors using novel feature descriptors such as FEGS (extract the graphical features), AAPse (extracting physicochemical features), and DPC (sequential features) from the given protein sequence and fed to the hierarchical deep net model (HDnet) as the

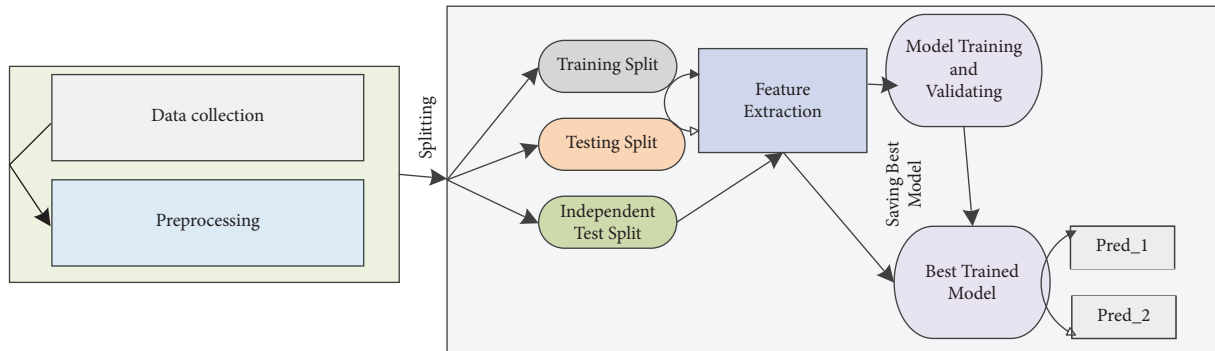


FIGURE 1: Stepwise approaches for classification of IGPs.

base classifier for constructing the model. The model opts for deep representations instead of manually extracted hand-crafted features and aims to perform the classification of IGPs. We have validated the model through exhaustive methods which shows that the overall prediction on both training and testing datasets outperformed the existing state-of-the-art methods. The study provides great insights into the large-scale identification of IGPs which pharmaceutical companies can opt for in novel drug design.

2. Materials and Methods

In the subsequent subsections, we will describe the stepwise approach to the classification of IGPs. Figure 1 shows these stepwise approaches. Firstly, the dataset collection and preprocessing method will be discussed. The feature representation method will be presented in the next section; the classification framework and model evaluation will be discussed in the third stage of the methodology.

2.1. Dataset Construction and Preprocessing. This portion will discuss dataset collection for experimenting, i.e., training and evaluating the designed framework. The dataset contained the immunoglobulins sequences downloaded from the UniProt database present in or outside the cell membrane. There are some standard techniques to assure the quality of the baseline dataset; in the first stage, we eliminated the ambiguous residues, i.e., “B,” “J,” “O,” “X,” “U,” and “Z” from the protein sequences to obtain typical amino acid sequences [38]. We also eliminate the sequence if it is the portion of other proteins. We picked the protein sequences from the human, mouse, and rat categories in the second stage. We used CD-HIT software to diminish hugely indistinguishable bias in the last stage, which caused over-fitting predicted results, and the cutoff value is set at 60%:

$$D = D^+ \cup D^- . \quad (1)$$

Our dataset D consists of 302 samples, with 110 positive D^+ and 192 negative D^- samples of immunoglobulins for training the model:

$$indD = indD^+ \cup indD^- . \quad (2)$$

Our independent dataset $indD$ contains 112 samples to evaluate our trained model, of which 40 are positive $indD^+$ and 72 are negative $indD^-$ samples. Overall, 150 positive and 264 negative samples are provided in Supplementary File S1 and Supplementary File S2, respectively.

2.2. Existing Feature Extraction Schemes. In designing a computerized framework, a series of steps are carried out to predict immunoglobulins. Among them, the feature extraction scheme is a challenging and essential step in formulating a biological sequence into some numerical values [39]. Conventional classification learning models, including K-nearest neighbour (KNN), random forest (RF) [40, 41], and support vector machine (SVM) [42], are based on fixed-length statistical values and are unable to handle the variable-length protein sequence; hence, the features representation algorithm can tackle this problem by extracting the fixed-length feature vector form the variable-length sequences [43–45]. Several researchers have used different feature encoding schemes [46] as shown in Figure 2; however, none of them used the proposed method for extracting vital pattern information from the immunoglobulins. A detailed description is given in Section 2.3.

2.3. Feature Extraction Based on Graphical and Statistical Features (FEFS). Herein, we have opted for a novel feature representation method named Feature Extraction based on Graphical and Statistical features (FEFS) [47] for immunoglobulins sequences, as shown in Figure 3. The proposed deep neural network is not novel; however, the extraction of features through this method is novel. Extracting the hidden pattern information through graphs is different from other sequence-based feature descriptors. The main shortcoming of traditional methods is the loss of sequence order information. For example, amino acid composition and reduced amino acid alphabet cannot retain the protein’s global correlated properties. Furthermore, the manual extraction of features requires extensive approaches which can be somehow not sufficient. These handcrafted features are not that much powerful to discriminate biological sequences as compared to the deep representations, as shown in [15]. The FEFS algorithm was proposed to tackle this issue by

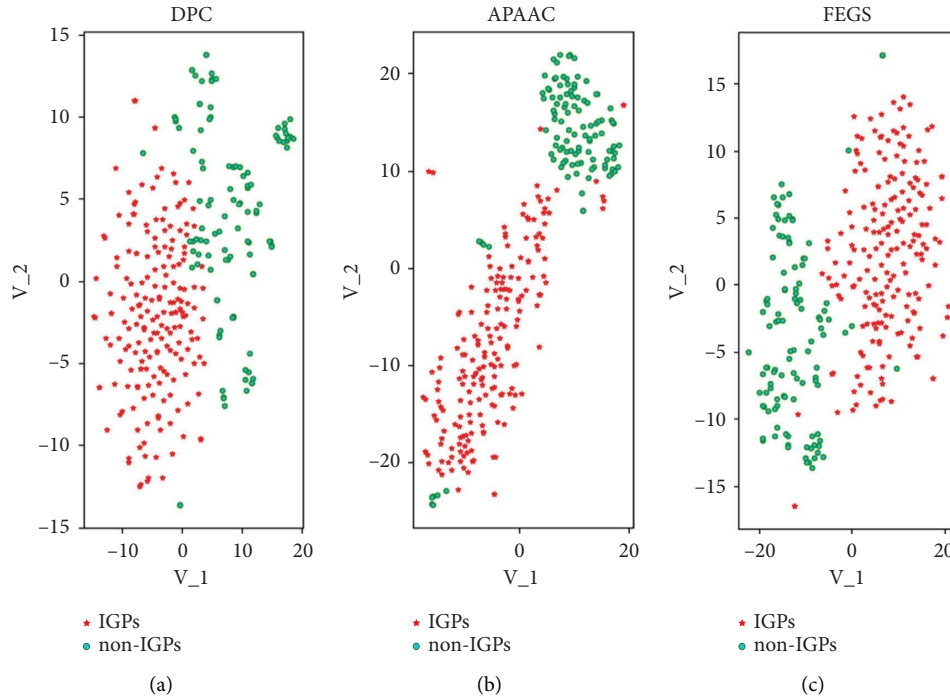


FIGURE 2: Scatter plot of DPC, APAAC, and FECS feature extraction methods.

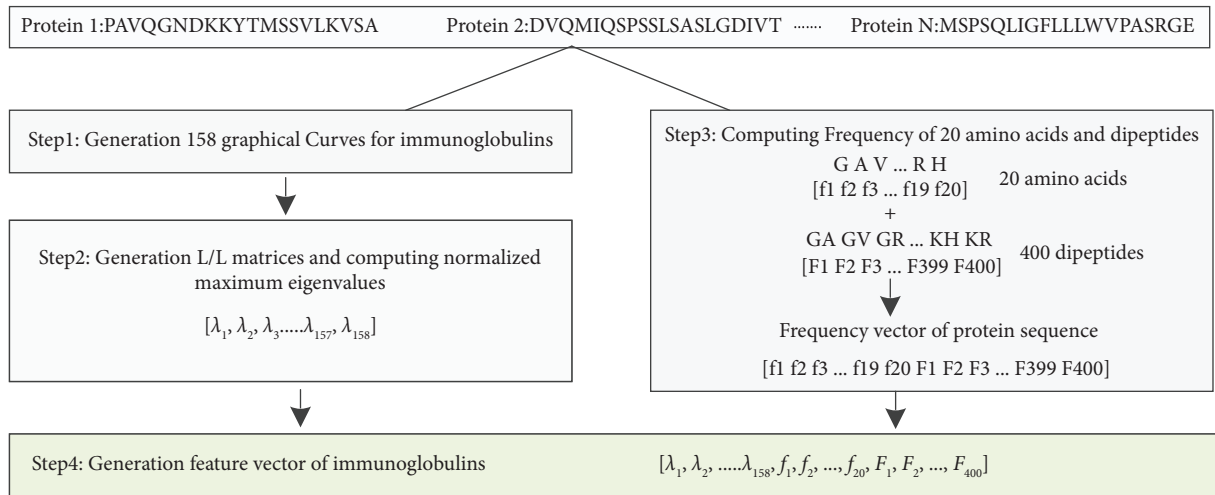


FIGURE 3: A brief explanation of the proposed FECS.

formulating the biological proteins using a three-dimensional curve. The working principle of the FECS algorithm is that initially, FECS employs the graphical depiction of primary proteins using circular cones in 3D space by extending the notion of 3D protein paths. Secondly, using the physico-chemical properties of amino acids that efficiently extract the statistical attributes of protein pairs, FECS seeks to form many circular cones in 3D space. Finally, the 578-dimensional vector is generated by combining mono-amino acid and dipeptide compositions for each protein sequence.

Initially, the protein sequences are provided in the FASTA format as input, and then FECS starts eliminating unnecessary indices with identical values and generating 158 space curves for the subsequent protein sequence.

2.3.1. Generation of 3D Graphical Curves for Immunoglobulins Sequences. In this method, the protein sequences are provided in the FASTA format as input; then, according to their physicochemical indices, 20 amino acids are first linked with 20 points in the 3D area. In the second step, the

graphical curve of an immunoglobulins sequence can be generated by enlarging a 3D protein track centred on a right circular cone.

(1) *Preparation of the 20 Amino Acids and the 400 Amino Acid Sets.* Physicochemical properties (PCP) of amino acids (AAs) play a vital role in analyzing and characterizing protein function. We arranged the 20 AAs with respect to their PCP from lower to higher order. Then, we organized them on the circumference of the bottommost of a right circular cone with a height of 1 by the following formula:

$$\Phi(A_i) = \left(\cos \frac{2\Pi^i}{20}, \sin \frac{2\Pi^i}{20}, 1 \right), i = 1, 2, \dots, 20. \quad (3)$$

The above equation A_i denoted 20 amino acids, whereas all 400 amino acid pairs are linked to the bottom of the right circular cone via the formula below:

$$\Phi(A_i A_j) = \Phi(A_i) + \frac{1}{4}(\Phi(A_j)), i, j = 1, 2, \dots, 20. \quad (4)$$

$A_i A_j$ represents each of the 400 amino acid pairs.

(2) *Building 3D Graphical Curves for Protein Sequences.* Consider that we have a protein sequence S having N AA residues $S = s_1 s_2 \dots s_N$. Constructing the 3D graph for the protein sequence is quite challenging. The 3D graphical curve is generated by enlarging a 3D protein track centred on a right circular cone as follows. Initiating from the origin point $p_0 = (0, 0, 0)$ broadens it to the subsequent point $p_1(x_1, y_1, z_1)$ in the 3D area, conforming to the first AA s_1 and the second point $p_2(x_2, y_2, z_2)$ related to the second AA s_2 and so on till the 3D track is accomplished at the last AA s_N , and via this process, the P path is obtained, coordinating with a 3D graphical curve of the immunoglobulins sequence S , whereas $P_i(x_i, y_i, z_i)$ is the i^{th} amino acid S_i , and the point coordinates x_i, y_i and z_i are described in the following formulas:

$$\Psi(S_i) = \Psi(S_{i-1}) + \sum_{A_1, A_2 \in \{A, C, D, \dots, Y\}} f_{A_1 A_2} \Phi(A_1 A_2). \quad (5)$$

In the above equation, $\Psi(S_0) = (0, 0, 0)$, and $f_{A_1 A_2}$ is the number of amino acid sets determined. The selected 158 physicochemical properties are linked with the exclusive right circular cone; in this way, we got 158 various 3-dimensional graphical curves for every immunoglobulin sequence related to the 158 physicochemical properties of amino acids.

2.3.2. *Numerical Features of Protein Sequences.* Another challenging job is to transform the generated graphical curves into numerical feature vectors for the similarity analysis of immunoglobulins samples. Here, for each curve, the L/L matrix denoted by M is calculated, and off-diagonal values $M_{i,j} (i \neq j)$ are well-defined as a measure of the Euclidean distance and the sum of geometric lengths of boundaries between P_i and P_j of the curve. At the same time, on-diagonal elements are equal to zero. Subsequently, all 158

curves are converted into 158-dimensional feature representation matrices as a graphical features representation described below:

$$V_g = [\lambda_1, \lambda_2, \dots, \lambda_{158}]. \quad (6)$$

There are many other feature extraction techniques in which AAC and DPC are commonly utilized in protein sequence analyses. To count the frequency of AA in a given sequence, normalized by sequence length, AAC is widely used for this process to extract 20 fixed-length features as formulated below:

$$V_a = [f_1, f_2, \dots, f_{20}]. \quad (7)$$

The above equation f represents the number of AA occurrences in the protein sequence. DPC also counts the number of occurrences of the 400 AA sets of the given protein sequence; and it extracts 400 fixed-length features below:

$$V_d = [f_1, f_2, \dots, f_{400}], \quad (8)$$

where f represents the number of occurrences of j^{th} AA sets, i.e., $\{AA, AC, AD, AE, .YY\}$ in the protein sequence. The statistical features, i.e., AA V_a and DPC V_d are merged with graphical features represented V_g to get a 578-dimensional feature vector for the protein sequence S . In general, a dataset that contains N number of immunoglobulins sequences is given to FEGS, then we can get the $N \times 578$ feature representation matrix, in which every row represents a feature representation vector of immunoglobulins sequences.

3. The Proposed Model Workflow

We developed a robust immunoglobulins predictor called Immunoglobulin Proteins Prediction Hierarchical Deep net (IGPred-HDnet). Figure 4 illustrates the flow of the proposed framework, in which the main stages of the IGPred-HDnet framework are shown such as data collection, data distribution, feature representation computation through FEGS, and classification through HDNet and evaluation. In feature representation, a novel feature encoding method is proposed to extract valuable feature representation from immunoglobulin sequences.

3.1. *Hierarchical Deep Net Model (HDnet).* The hierarchical deep net (HDnet) model is an ensemble-based model inspired by [48], which is a substitute for a deep neural network (DNN) to learn hyperlevel feature representation using various resources and efforts. In contrast, DNN used complex architecture, i.e., forward and backward propagation algorithms, to learn hidden information. In developing an HDnet classifier, it is crucial to determine the learning algorithms employed in each layer. In our proposed model, we set the combination of Extreme Gradient Boost (XGBoost) [49, 50], random forest (RF) [51–53], and extremely randomized trees (ERT) [54, 55] classifiers which achieved outstanding performance and feed it with the

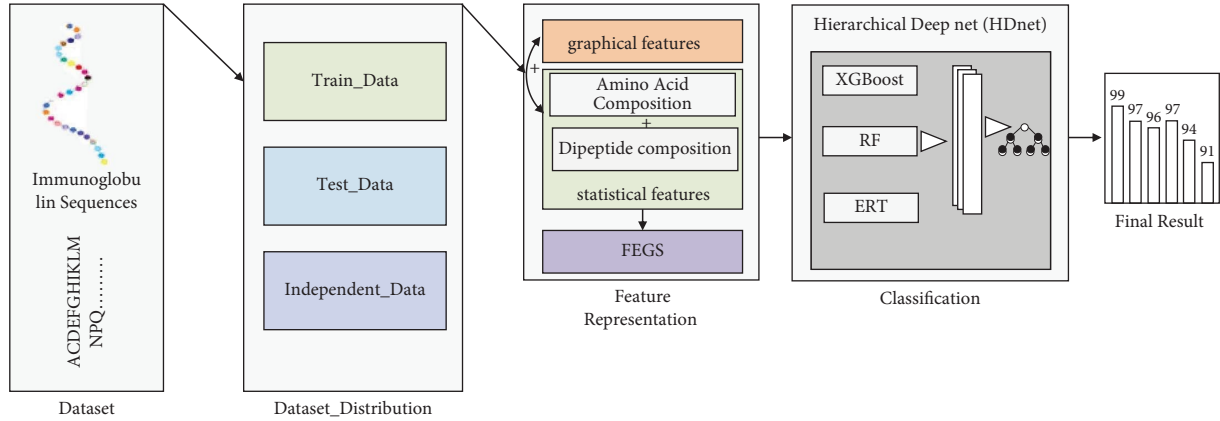


FIGURE 4: Graphical representation of the IGPred-HDnet model.

previously computed 578-dimensional vector. HDnet is based on the deep ensemble method that cascades conventional classifiers, for example, RF, ERT, and XGBoost. Compared to DNN, HDnet uses decision trees instead of various neural network (NN) models for feature representation learning in each layer. Figure 5 shows the generic representation of HDNet, elaborating that if there are multiple feature vectors from multiple encoding schemes, they are concatenated at the level- N . These feature vectors are actually deep representations learnt at different layers, similar to other deep neural networks. Due to the hierarchical type nature, the HDnet model allows the training process to be more robust, and it will be more appropriate for training a limited amount of protein samples. DNN involves various parameters that need tunes during training a model, while our proposed model easily tunes the hyperparameter.

We set the boosting parameter value $k = 20$ for the XGBoost classifier. For RF and ERT, the number of decision trees is also set at 20, and the node values are picked by randomly picking features. In our model, every layer is an ensemble of diverse learners (e.g., six XGBoost, six RF, and six ERT) who accept the feature representation processed by previous layer classification models. The outcome of the previous layer is the input for the subsequent layer for processing. To produce the enhanced feature representation related to the multivariate class vectors, we have integrated, stacked, and summed output as a supreme probability score. The process of training is terminated if enhancement is not observed in performance. Figure 5 reveals the layer-by-layer framework of the HDnet.

4. Performance Evaluation

In this research, we utilized four performance evaluation measures, e.g., accuracy (ACC), specificity (SP), sensitivity (SN), and Matthew correlation coefficient (MCC), to figure out the achievement rate of our proposed prediction models described as

$$ACC = \frac{TP + TN}{TP + FP + TN + FN},$$

$$SP = \frac{TN}{TN + FP},$$

$$SN = \frac{TP}{TP + FN},$$

$$MCC = \frac{TP * TN - FT * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}. \quad (9)$$

In the above equations, TP represents True-IGPs, which are correctly predicted as positive instances, whereas TN corresponds to true non-IGPs, which are correctly classified as negative samples. FN indicates non-IGPs, which the model incorrectly predicts as immunoglobulins.

The performance above measures containing the MCC is dependent on the threshold, which delivers the comprehensive evaluation for the binary class classification. Furthermore, to describe the model performance on a large scale, we utilized the Area Under the ROC (Receiver Operating Characteristic) Curve (AUC), which is in the shape of an independent threshold analysis like a further essential assessment of the model.

5. Proposed Framework Evaluation

In machine learning (ML), the model performance is naturally assessed via cross-validation (CV). There are three tests in the research community to determine the discriminatory power of the designed framework: K-fold also called subsampling, Jackknife, i.e., leave-one-out and independent tests [56, 57]. The Jackknife test provides exceptional and encouraging results to train a model [58]; however, the main cons are computational cast due to a large number of calculations [59]. To overcome the weakness of

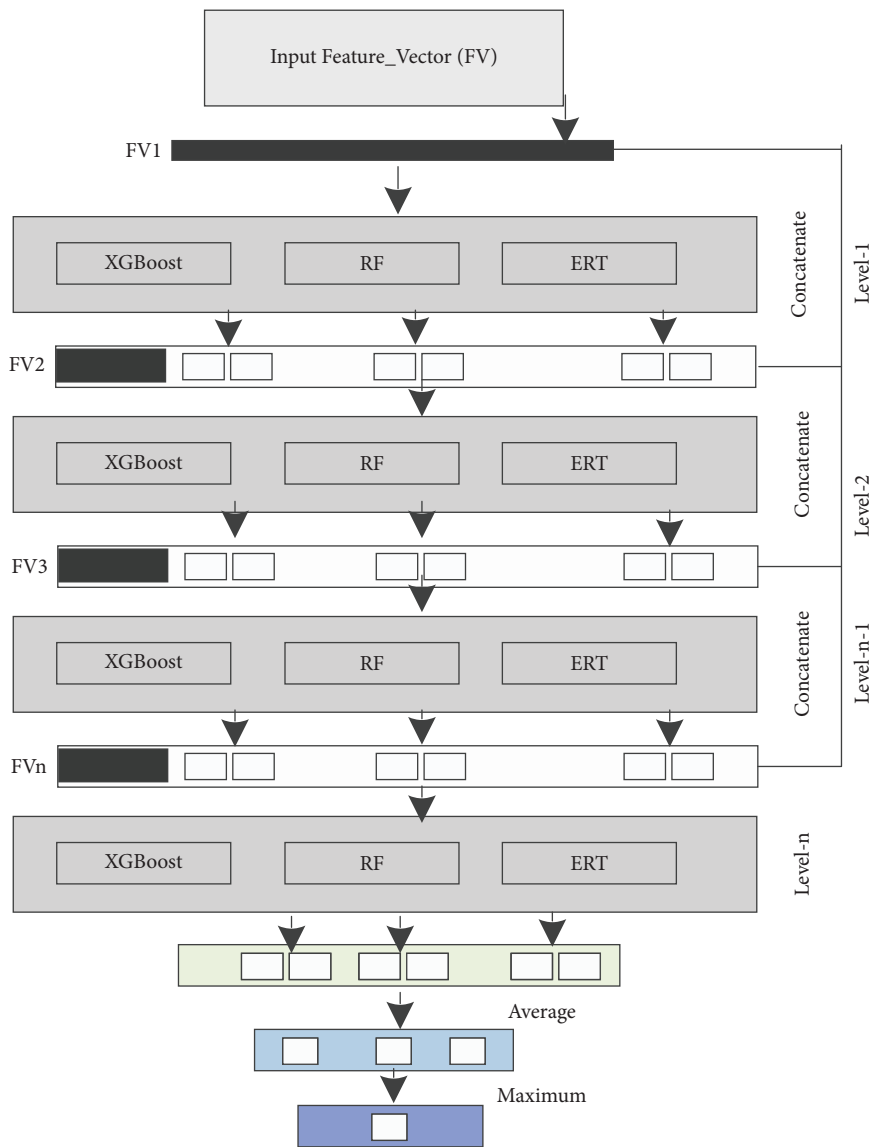


FIGURE 5: The layer-by-layer framework of the HDnet.

the Jackknife and improve the simplification power, we implemented the K -Fold CV test to train our model and test the performance [60]. In this method, we randomly divided the train data into K -folds (subsets), in which $K - 1$ is utilized to train the proposed model, and the leftover is utilized to test the model [61]. Subsequently, for the particular approximation, the obtained results are averaged. We set the value of K to 10 after conducting various experiments.

5.1. Predictive Performance of Hypothesis Learners Using Various Feature Encoding Schemes on Training Dataset D_{train} .

In this section, we experimentally determine the prediction performance of various classifiers, i.e., KNN [62], DT [63], SVM [46, 64], and HDnet using various descriptors, i.e., APAAC (physicochemical features), DPC (sequential features), and FECS (graphical features), as shown in Figure 6. Each learning engine is computed by conducting a ten-fold

CV test on the training dataset D_{train} with four evaluation measures ACC, SN, SP, and MCC. In the case of APAAC feature vectors, the SVM classifier secured the worst AAC = 89.72% and MCC = 0.786, while HDnet achieved a higher ACC of 95.69% and MCC of 0.909 points. Similarly, in the case of the DPC method, again the HDnet classifier produced 0.33% high ACC and 0.007 points MCC, respectively. Furthermore, in the case of the FECS feature method, the highest performance is obtained by the HDnet classifier, which is ACC = 98.00%, SN = 94.55%, SP = 100%, and MCC = 0.958. The second-best predictor is KNN which achieved 90.41% ACC and 0.809 points MCC, while SVM comparatively produce good predictions on all feature methods.

Several judgements are made on the reported results of all classifiers in Table 1. First, the HDnet model consistently produced the best outcomes among the classification algorithms compared to other machine-learning classifiers for

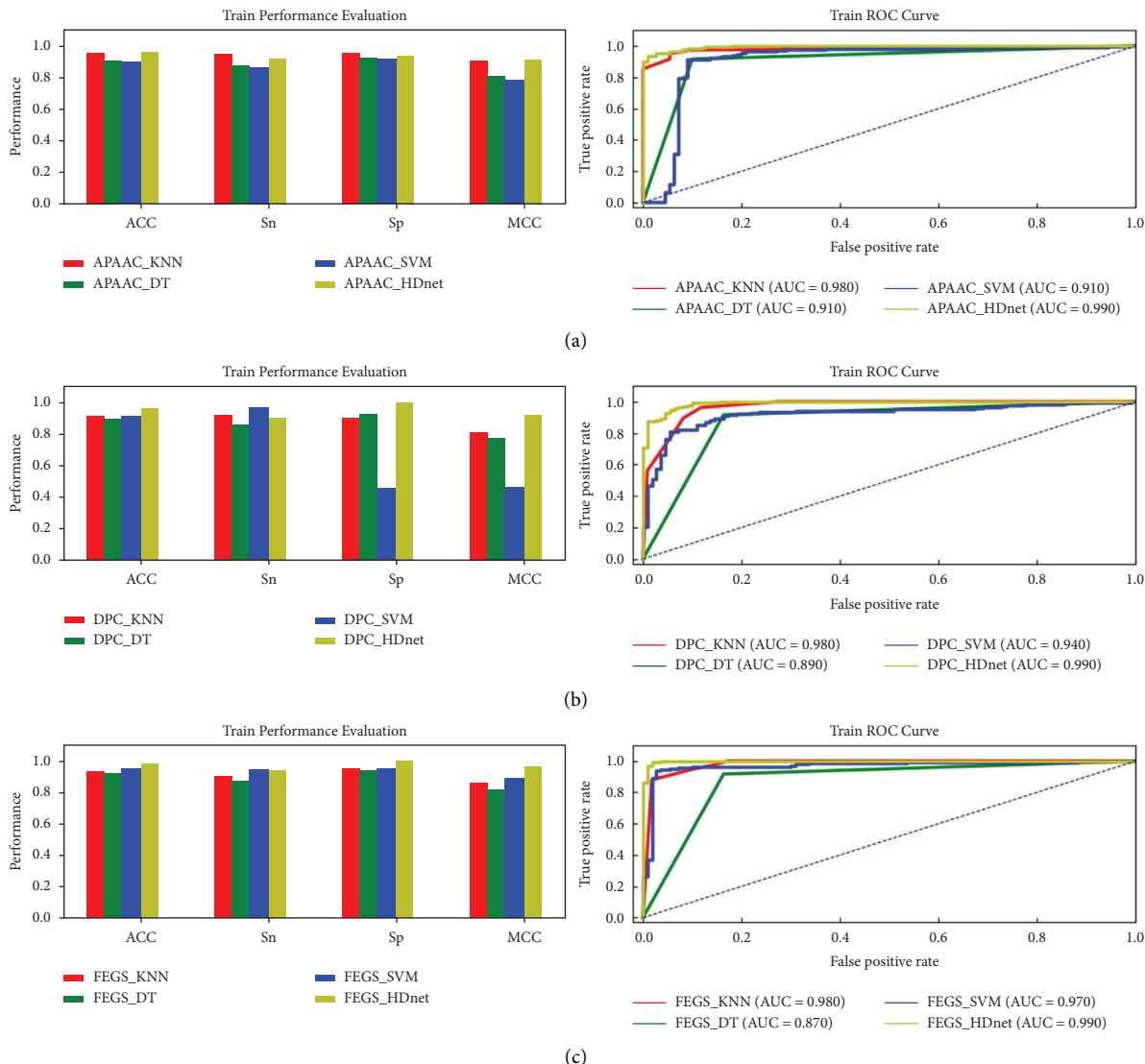


FIGURE 6: Training parameter metric performances evaluation and AUC-ROCs of KNN, DT, SVM, and the proposed HDnet via APAAC, DPC, and the proposed FECS feature extraction methods.

all feature encoding schemes. The main reason is due to the high learning potential of a deep neural network as compared to the conventional classifiers. The internal structure of the HDnet classifier is based on decision trees that enable the model to predict the extracted features better [65]. Further, it is evident in the literature that deeper networks have more learning potential as compared to conventional neural networks [15, 66, 67].

Secondly, among the feature representation approaches, FECS (graphical features) produced the best results for overall hypothesis learners (classifiers) than other feature vectors such as DPC and APAAC. The underlying reason for the high prediction rate of FECS methods is that FECS extracts the conserved local and global graphical, physicochemical and statistical attributes from a protein sequence. As in Figure 1, the visualization influence of the extracted features through t-distributed stochastic neighbour embedding (t-SNE) can be seen. The red colour

represents the IGP class, and the green colour represents the non-IGP class. The features with a high correlation, like DPC and APAAC, cannot incorporate the correct predictions of immunoglobulins. In contrast, the novel features of FECS are less correlated enabling the classifiers to produce high performance.

5.2. Predictive Performance of Hypothesis Learners Using Various Feature Encoding Schemes on the Testing Dataset D_{test} . In this subsection, we examine the success rates of our model via an independent test to show its generalization power. It was ensured that the samples in the independent test D_{test} were unseen, and none of the immunoglobulin samples was used in training the model. Table 1 depicts the prediction outcomes of all classifiers using the APAAC, DPC, and FECS feature methods. Comparative analysis reveals that our proposed learning model HDnet using novel feature FECS

TABLE 1: Analysis of various classifiers using feature encoding schemes on training and testing datasets D_{train} and D_{test} .

Feature-encoding methods	Classifiers	Benchmark dataset					Independent dataset				
		ACC (%)	SN (%)	SP (%)	MCC	F-measure (%)	ACC (%)	SN (%)	SP (%)	MCC	F-measure (%)
APAAC	KNN	95.01	94.55	95.26	0.898	92.95	88.93	97.50	83.33	0.777	85.71
	DT	90.70	87.27	92.66	0.802	86.92	91.96	92.50	91.66	0.829	89.15
	SVM	89.72	86.36	91.66	0.786	85.61	81.25	82.50	80.55	0.612	75.86
	HDnet	95.69	91.82	93.75	0.909	93.83	90.17	87.50	91.66	0.787	86.41
DPC	KNN	90.41	91.82	89.63	0.809	86.77	92.85	97.50	90.27	0.854	90.69
	DT	89.39	85.45	91.66	0.771	85.11	93.75	92.05	94.44	0.864	91.35
	SVM	90.23	96.36	45.58	0.467	69.67	83.03	100	73.61	0.706	80.80
	HDnet	96.02	90.00	99.47	0.916	94.18	91.96	77.50	100	0.829	87.32
FEGS	KNN	93.03	90.00	94.78	0.856	89.71	94.64	100	91.60	0.890	93.00
	DT	91.93	87.27	93.74	0.817	87.95	93.75	97.50	91.60	0.871	91.76
	SVM	94.72	94.55	94.85	0.889	92.96	93.75	100	90.27	0.876	91.95
	HDnet	98.00	94.55	100	0.958	96.94	99.10	97.50	100	0.980	98.73

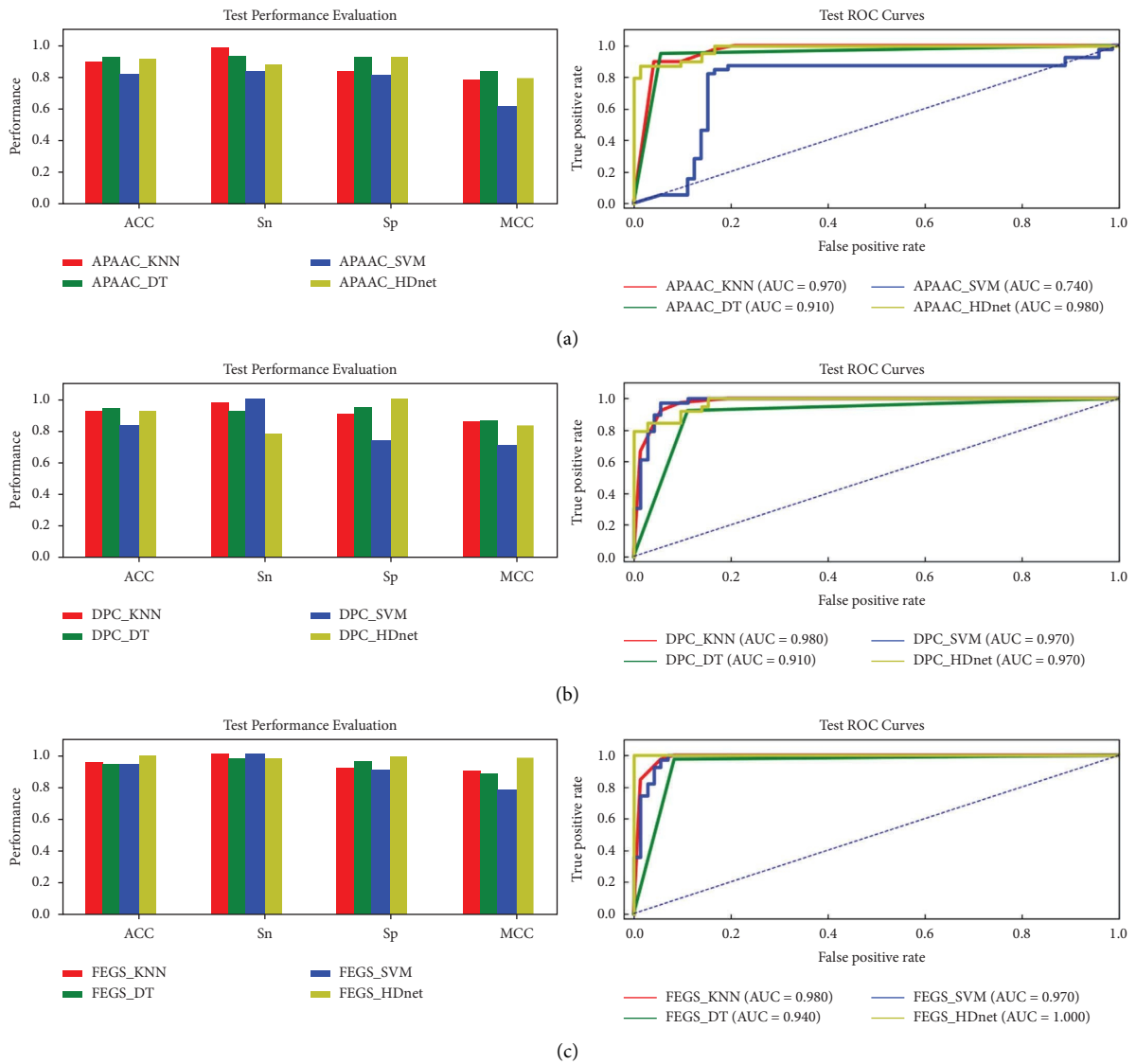


FIGURE 7: Training parameter metric performances evaluation and AUC-ROCs of KNN, DT, SVM, and the proposed HDnet via APAAC, DPC, and the proposed FEGS feature extraction methods.

TABLE 2: Analysis of various classifiers using feature encoding schemes on training and testing datasets D_{train} and D_{test} .

Dataset	Predictor	ACC	SN	SP	MCC	Pre	NPV	F1	AUC
Training	CC_PSSM	0.960	—	—	0.921	0.961	—	—	0.994
	IGPred	0.969	0.963	0.975	—	—	—	—	0.994
	XGBoost	0.972	0.945	0.985	0.950	0.980	—	—	0.970
	IGPred-HDnet	0.980	0.945	1.000	0.958	1.000	1.000	0.971	0.998
Testing	CC_PSSM	0.883	—	—	0.847	0.884	—	—	0.914
	IGPred	0.891	0.886	0.897	—	—	—	—	0.914
	XGBoost	0.894	0.869	0.906	0.874	0.902	—	—	0.892
	IGPred-HDnet	0.991	1.000	0.986	0.980	0.9750	1.000	0.987	1.000

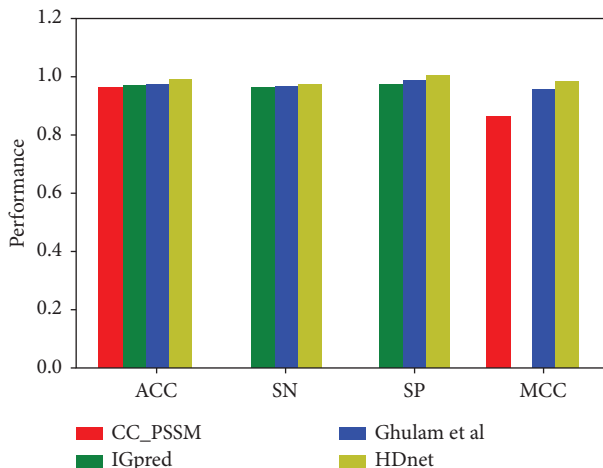


FIGURE 8: Comparison of the proposed method with the existing methods.

achieved outstanding results in terms of all performance metrics, likewise ACC = 99.10%, SN = 97.50%, SP = 100%, and MCC = 0.980 points, respectively. In contrast, the same learning engine using the APAAC feature produced the worst results as shown in Figure 7.

5.3. Predictive Performance of the Proposed Predictor with Existing Methods on Training and Testing Datasets. In this section, we theoretically compare the efficacy of our proposed model with the three developed approaches such as CC-PSSM [39], IGPred [19], and Ghulam et al.’s approach [68] on training and testing datasets. The results in Table 2 are extracted from the previous literature [69]. It is worth noting that none of the existing predictors generated the prediction outcomes on independent tests to show the generalization power of their model. Driven by the novel feature descriptor FECS with the intelligent deep learning-based algorithm HDnet, IGPred-HDnet outperformed the existing methods for IGS identification in terms of all performance metrics, i.e., ACC, SN, SP, MCC, and AUC. On the training benchmark dataset, our method notable increased ACC by 1.9%, SN by 1%, SP by 1.5%, and MCC by 0.026 points over the second-best performer XGBoost. An independent test was performed to investigate further the IGPred-HDnet model’s predictive capability on unseen data. Both the ACC and AUC results are 0.99 and 1.00, as shown in Table 2 and Figure 8.

The underlying reason for achieving high predictions is to extract the graphical-based, physicochemical-based, and sequence-based attributes. Also, the hierarchical type structure of the HDnet classifier enables a better forecast of the IGS samples from the extracted attributes [65].

6. Conclusion and Future Work

IGPs are a crucial constituent of the immune system. Understanding deep insight IGSs can provide useful hints in drug discovery for disease treatment. Thus, the objective of this research was to construct a novel sequence-based computational method for predicting and analyzing IGSs. The proposed theoretical model “IGPred-HDnet” is superior to other advanced immunoglobulin-based predictors due to several reasons. Firstly, we designed an innovative graphical algorithm FECS to capture structured information buried in the protein sample. The structure features produced better results than the other feature schemes. Secondly, we implemented a deep learning model called HDnet for the first time as a learning model for recognizing IGSs.

Despite enhancing the model’s overall performance, further gaps still exist for future, such as several previous publications like Tang et al. [27] established public web-servers that can enrich the applicability of the anticipated model. Also, using novel feature selection algorithms is vital to avoid overfitting and improve the generalization power of the trained model. We hope that the proposed IGPred-HDnet will become a potential tool for large-scale IGSs characterization in particular and other protein problems in general.

Data Availability

The dataset analyzed in this study can be found in the supplementary files.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

Supplementary Materials

Supplementary File S1 contains the positive samples (immunoglobulins sequences). Supplementary File S2 contains the negative samples (nonimmunoglobulins sequences). (*Supplementary Materials*)

References

- [1] A. N. Barclay, “Membrane proteins with immunoglobulin-like domains—a master superfamily of interaction molecules,” in *Seminars in Immunology* Elsevier, Amsterdam The Netherlands, 2003.
- [2] A. H. A. Latiff and M. A. Kerr, “The clinical significance of immunoglobulin A deficiency,” *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, vol. 44, no. 2, pp. 131–139, 2007.
- [3] T. Génereau, O. Chosidow, C. Danel, P. Chérin, and S. Herson, “High-dose intravenous immunoglobulin in cutaneous lupus erythematosus,” *Archives of Dermatology*, vol. 135, no. 9, pp. 1124–1125, 1999.
- [4] J. K. Presto, E. Z. Hejazi, and V. P. Werth, “Biological therapies in the treatment of cutaneous lupus erythematosus,” *Lupus*, vol. 26, no. 2, pp. 115–118, 2017.
- [5] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, and A. Zhavoronkov, “druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico,” *Molecular Pharmaceutics*, vol. 14, no. 9, pp. 3098–3104, 2017.
- [6] M. Ali, S. A. Khan, K. Wennerberg, and T. Aittokallio, “Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach,” *Bioinformatics*, vol. 34, no. 8, pp. 1353–1362, 2018.
- [7] A. A. Shah, H. A. M. Malik, A. Mohammad, Y. D. Khan, and A. Alourani, “Machine learning techniques for identification of carcinogenic mutations, which cause breast adenocarcinoma,” *Scientific Reports*, vol. 12, no. 1, pp. 11738–11815, 2022.
- [8] H. Shirai, C. Prades, R. Vita et al., “Antibody informatics for drug discovery,” *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, vol. 1844, no. 11, pp. 2002–2015, 2014.
- [9] W. Alghamdi, M. Attique, E. Alzahrani, M. Z. Ullah, and Y. D. Khan, “LBCEPred: a machine learning model to predict linear B-cell epitopes,” *Briefings in Bioinformatics*, vol. 23, no. 3, Article ID bbac035, 2022.
- [10] J. Benichou, R. Ben-Hamo, Y. Louzoun, and S. Efroni, “RepSeq: uncovering the immunological repertoire through next-generation sequencing,” *Immunology*, vol. 135, no. 3, pp. 183–191, 2012.
- [11] I. Sela-Culang, S. Ashkenazi, B. Peters, and Y. Ofra, “PEASE: predicting B-cell epitopes utilizing antibody sequence,” *Bioinformatics*, vol. 31, no. 8, pp. 1313–1315, 2015.
- [12] C. K. Hua, A. T. Gacerez, C. L. Sentman, M. E. Ackerman, Y. Choi, and C. Bailey-Kellogg, “Computationally-driven identification of antibody epitopes,” *Elife*, vol. 6, Article ID e29023, 2017.
- [13] Y. Zhai, Y. Chen, Z. Teng, and Y. Zhao, “Identifying antioxidant proteins by using amino acid composition and protein-protein interactions,” *Frontiers in Cell and Developmental Biology*, vol. 8, Article ID 591487, 2020.
- [14] J. Zhang, Z. Zhang, L. Pu, J. Tang, and F. Guo, “AIEpred: an ensemble predictive model of classifier chain to identify anti-inflammatory peptides,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 5, pp. 1831–1840, 2021.
- [15] S. Naseer, W. Hussain, Y. D. Khan, and N. Rasool, “Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations,” *Analytical Biochemistry*, vol. 615, Article ID 114069, 2021.
- [16] A. H. Butt, S. A. Khan, H. Jamil, N. Rasool, and Y. D. Khan, “A prediction model for membrane proteins using moments based features,” *BioMed Research International*, vol. 2016, pp. 1–7, 2016.
- [17] Y. D. Khan, E. Alzahrani, W. Alghamdi, and M. Zaka Ullah, “Sequence-based identification of allergen proteins developed by integration of PseAAC and statistical moments via 5-step rule,” *Current Bioinformatics*, vol. 15, no. 9, pp. 1046–1055, 2020.
- [18] C. Yang, Y. Ding, Q. Meng, J. Tang, and F. Guo, “Granular multiple kernel learning for identifying RNA-binding protein residues via integrating sequence and structure information,” *Neural Computing & Applications*, vol. 33, no. 17, pp. 11387–11399, 2021.
- [19] J. Shao and B. Liu, “ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm,” *Briefings in Bioinformatics*, vol. 22, no. 3, Article ID bbaa192, 2021.
- [20] X. Zhao, H. Wang, H. Li, Y. Wu, and G. Wang, “Identifying plant pentatricopeptide repeat proteins using a variable selection method,” *Frontiers of Plant Science*, vol. 12, Article ID 506681, 2021.
- [21] M. Awais, W. Hussain, N. Rasool, and Y. Daanial Khan, “ITSP-PseAAC: identifying tumor suppressor proteins by using fully connected neural network and PseAAC,” *Current Bioinformatics*, vol. 16, no. 5, pp. 700–709, 2021.
- [22] A. Ehsan, M. K. Mahmood, Y. D. Khan, O. M. Barukab, S. A. Khan, and K. C. Chou, “iHyd-PseAAC (EPSV): identifying hydroxylation sites in proteins by extracting enhanced position and sequence variant feature via Chou’s 5-step rule and general pseudo amino acid composition,” *Current Genomics*, vol. 20, no. 2, pp. 124–133, 2019.
- [23] S. Ilyas, W. Hussain, A. Ashraf, Y. D. Khan, S. A. Khan, and K. C. Chou, “iMethylK-PseAAC: improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general PseAAC via Chou’s 5-steps rule,” *Current Genomics*, vol. 20, no. 4, pp. 275–292, 2019.
- [24] W. Hussain, N. Rasool, and Y. D. Khan, “A sequence-based predictor of Zika virus proteins developed by integration of PseAAC and statistical moments,” *Combinatorial Chemistry & High Throughput Screening*, vol. 23, no. 8, pp. 797–804, 2020.
- [25] S. Naseer, W. Hussain, Y. D. Khan, and N. Rasool, “IPhosS (Deep)-PseAAC: identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-Steps rule,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, p. 1, 2020.
- [26] L.-M. Liu, Y. Xu, and K.-C. Chou, “IPGK-PseAAC: identify lysine phosphoglycylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC,” *Medicinal Chemistry*, vol. 13, no. 6, pp. 552–559, 2017.
- [27] H. Tang, W. Chen, and H. Lin, “Identification of immunoglobulins using Chou’s pseudo amino acid composition with feature selection technique,” *Molecular BioSystems*, vol. 12, no. 4, pp. 1269–1275, 2016.

- [28] S. Naseer, H. Waqar, D. Khan Yaser, and R. Nouman, "NPalmitylDeep-PseAAC: a predictor of N-palmitoylation sites in proteins using deep representations of proteins and PseAAC via modified 5-steps rule," *Current Bioinformatics*, vol. 16, no. 2, pp. 294–305, 2021.
- [29] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Information Sciences*, vol. 418–419, pp. 546–560, 2017.
- [30] B. Song, F. Li, Y. Liu, and X. Zeng, "Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison," *Briefings in Bioinformatics*, vol. 22, no. 6, Article ID bbab282, 2021.
- [31] S. Ahmed, M. Kabir, M. Arif, Z. Ali, and Z. N. Khan Swati, "Prediction of human phosphorylated proteins by extracting multi-perspective discriminative features from the evolutionary profile and physicochemical properties through LFDA," *Chemometrics and Intelligent Laboratory Systems*, vol. 203, Article ID 104066, 2020.
- [32] C. Meng, J. Wu, F. Guo, B. Dong, and L. Xu, "CWLy-pred: a novel cell wall lytic enzyme identifier based on an improved MRMD feature selection method," *Genomics*, vol. 112, no. 6, pp. 4715–4721, 2020.
- [33] Q. Li, W. Zhou, D. Wang, S. Wang, and Q. Li, "Prediction of anticancer peptides using a low-dimensional feature model," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 892, 2020.
- [34] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 215, 2019.
- [35] C. Feng, Q. Zou, and D. Wang, "Using a low correlation high orthogonality feature set and machine learning methods to identify plant pentatricopeptide repeat coding gene/protein," *Neurocomputing*, vol. 424, pp. 246–254, 2021.
- [36] Y. D. Khan, N. S. Khan, S. Farooq et al., "An efficient algorithm for recognition of human actions," *The Scientific World Journal*, vol. 2014, pp. 1–11, 2014.
- [37] S. Saeed, M. K. Mahmood, and Y. D. Khan, "An exposition of facial expression recognition techniques," *Neural Computing & Applications*, vol. 29, no. 9, pp. 425–443, 2018.
- [38] A. A. Shah and Y. D. Khan, "Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification," *Scientific Reports*, vol. 10, no. 1, pp. 16913–17010, 2020.
- [39] M. Arif, M. Hayat, and Z. Jan, "IMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 442, pp. 11–21, 2018.
- [40] E. Alzahrani, W. Alghamdi, M. Z. Ullah, and Y. D. Khan, "Identification of stress response proteins through fusion of machine learning models and statistical paradigms," *Scientific Reports*, vol. 11, no. 1, pp. 21767–21815, 2021.
- [41] S. J. Malebary, R. Khan, and Y. D. Khan, "ProtoPred: advancing oncological research through identification of proto-oncogene proteins," *IEEE Access*, vol. 9, pp. 68788–68797, 2021.
- [42] M. K. Mahmood, A. Ehsan, Y. D. Khan, and K. C. Chou, "IHyd-LysSite (EPSV): identifying hydroxylysine sites in protein using statistical formulation by extracting enhanced position and sequence variant feature technique," *Current Genomics*, vol. 21, no. 7, pp. 536–545, 2020.
- [43] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chemistry*, vol. 11, no. 3, pp. 218–234, 2015.
- [44] F. Ge, Y. Zhang, J. Xu, A. Muhammad, J. Song, and D. J. Yu, "Prediction of disease-associated nsSNPs by integrating multi-scale ResNet models with deep feature fusion," *Briefings in Bioinformatics*, vol. 23, no. 1, Article ID bbab530, 2022.
- [45] F. Ge, A. Muhammad, and D.-J. Yu, "DeepnsSNPs: accurate prediction of non-synonymous single-nucleotide polymorphisms by combining multi-scale convolutional neural network and residue environment information," *Chemometrics and Intelligent Laboratory Systems*, vol. 215, Article ID 104326, 2021.
- [46] W. Hussain, N. Rasool, and Y. D. Khan, "Insights into machine learning-based approaches for virtual screening in drug discovery: existing strategies and streamlining through FP-cadd," *Current Drug Discovery Technologies*, vol. 18, no. 4, pp. 463–472, 2021.
- [47] Z. Mu, T. Yu, X. Liu, H. Zheng, L. Wei, and J. Liu, "FEGS: a novel feature extraction model for protein sequences and its applications," *BMC Bioinformatics*, vol. 22, no. 1, pp. 297–315, 2021.
- [48] Y. Guo, S. Liu, Z. Li, and X. Shang, "BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data," *BMC Bioinformatics*, vol. 19, no. S5, pp. 118–213, 2018.
- [49] 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*, San Francisco, CA, USA, August 2016.
- [50] F. Ge, Y. H. Zhu, J. Xu, A. Muhammad, J. Song, and D. J. Yu, "MutTMPredictor: robust and accurate cascade XGBoost classifier for prediction of mutations in transmembrane proteins," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 6400–6416, 2021.
- [51] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [52] A. H. Butt and Y. D. Khan, "CanLect-Pred: a cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences," *IEEE Access*, vol. 8, pp. 9520–9531, 2020.
- [53] O. Barukab, Y. D. Khan, S. A. Khan, and K. C. Chou, "ISulfoTyr-PseAAC: identify tyrosine sulfation sites by incorporating statistical moments via Chou's 5-steps rule and pseudo components," *Current Genomics*, vol. 20, no. 4, pp. 306–320, 2019.
- [54] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [55] F. Ge, J. Hu, Y. H. Zhu, M. Arif, and D. J. Yu, "TargetMM: accurate missense mutation prediction by utilizing local and global sequence information with classifier ensemble," *Combinatorial Chemistry & High Throughput Screening*, vol. 25, no. 1, pp. 38–52, 2021.
- [56] W. Lee and K. Han, "Constructive prediction of potential RNA aptamers for a protein target," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 5, pp. 1476–1482, 2020.
- [57] M. Shahid, M. Ilyas, W. Hussain, and Y. D. Khan, "ORI-Deep: improving the accuracy for predicting origin of replication sites by using a blend of features and long short-term memory network," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbac001, 2022.
- [58] Y.-C. Wang, X. B. Wang, Z. X. Yang, and N. Y. Deng, "Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature," *Protein and Peptide Letters*, vol. 17, no. 11, pp. 1441–1449, 2010.

- [59] X.-X. Chen, H. Tang, W. C. Li et al., "Identification of bacterial cell wall lyases via pseudo amino acid composition," *BioMed Research International*, vol. 2016, Article ID 1654623, 8 pages, 2016.
- [60] S. J. Malebary and Y. D. Khan, "Evaluating machine learning methodologies for identification of cancer driver genes," *Scientific Reports*, vol. 11, no. 1, pp. 12281–12313, 2021.
- [61] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *Journal of Parallel and Distributed Computing*, vol. 117, pp. 212–217, 2018.
- [62] Z. Ali, A. W. Abbas, T. M. Thasleema, B. Uddin, T. Raaz, and S. A. R. Abid, "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 271–275, 2015.
- [63] X. Chen, M. Wang, and H. Zhang, "The use of classification trees for bioinformatics," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55–63, 2011.
- [64] Y. Gong, B. Liao, D. Peng, and Q. Zou, "Accurate prediction and key feature recognition of immunoglobulin," *Applied Sciences*, vol. 11, no. 15, p. 6894, 2021.
- [65] Z.-H. Zhou and J. Feng, "Deep forest," 2017, <https://arxiv.org/abs/1702.08835>.
- [66] W. Hussain, "SAMP-PFPDeep: improving accuracy of short antimicrobial peptides prediction using three different sequence encodings and deep neural networks," *Briefings in Bioinformatics*, vol. 23, no. 1, Article ID bbab487, 2022.
- [67] C.-S. Wang, P. J. Lin, C. L. Cheng, S. H. Tai, Y. H. Kao Yang, and J. H. Chiang, "Detecting potential adverse drug reactions using a deep neural network model," *Journal of Medical Internet Research*, vol. 21, no. 2, Article ID e11016, 2019.
- [68] A. Ghulam, R. Sikander, F. Ali, Z. N. K. Swati, A. Unar, and D. B. Talpur, "Accurate prediction of immunoglobulin proteins using machine learning model," *Informatics in Medicine Unlocked*, vol. 29, p. 100885, 2022.
- [69] A. Ghulam, R. Sikander, F. Ali, Z. N. Khan Swati, A. Unar, and D. B. Talpur, "Accurate prediction of immunoglobulin proteins using machine learning model," *Informatics in Medicine Unlocked*, vol. 29, Article ID 100885, 2022.

Retraction

Retracted: Analysis of Geocological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] Y. Song, L. Gao, H. He, and J. Lu, "Analysis of Geocological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 6555005, 13 pages, 2022.

Research Article

Analysis of Geocological Restoration in Mountainous Cities Affected by Geological Hazards with Interval Intuitive Fuzzy Information

Yuanwen Song , Lei Gao, Haipin He, and Juan Lu

Department of Geology and Jewelry, Lanzhou Resources and Environment Voc-Tech University, Lanzhou 730021, China

Correspondence should be addressed to Yuanwen Song; songyw@lzre.edu.cn

Received 2 July 2022; Revised 22 July 2022; Accepted 30 August 2022; Published 15 October 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Yuanwen Song et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the progress of the industrial revolution and the development of modern science and technology, China's urbanization process has been promoted. Urban and rural economic and social construction has greatly improved the local appearance and social structure. Human activities and natural ecology have affected the whole geological-ecological process, further aggravated the geological-ecological damage, and caused more serious geological disasters, especially in some places (especially in mountainous areas). In recent years, strong geological disasters have occurred in Wenchuan, Yushu, and Lushan regions of China, which not only seriously endanger the life safety and social life of the affected people, but also damage the geological-ecological structure and social functions of the region, especially in the geographically sensitive Alpine urban areas. It also produced many secondary disasters, such as landslides and land collapses. Mountainous cities and towns have special requirements for construction land, which is difficult to construct. Industrial land resources are in short supply, urban and rural comprehensive construction land is not active, and cultivated land area resources are tight. Compared with plain towns with superior geological conditions, mountain towns are more vulnerable to adverse geological environment such as geological ecology, landform, ecological vegetation, and hydrology. The geographical natural environment, as an organic whole that combines and interacts with the geomorphic natural environment, the biological-ecological environment, and the human social management environment, is the main reason that affects the development of mountain towns. Once the mountain geological ecology is destroyed, a series of geological disasters will often be induced, which will seriously restrict the healthy development of mountain towns. Scientific management of the geological environment plays an important role in the assessment of the geological environment restoration of mountain towns after disasters. Therefore, taking the most beautiful counties in China, Baoxing City, and Tianquan County as examples, on the basis of studying the complex geological-ecological theory of geological disasters, this paper further improves the traditional ecological footprint model in China, and using the interval direct fuzzy information constructs the metric index of ecological restoration scheme of mountain towns, and determines the evaluation index and optimal scheme of ecological restoration. From the aspects of landscape layout construction, disaster prevention and mitigation planning and improvement, and environmental restoration project, the future geocological restoration and response strategy of Lushan County are pointed out, which provide guidance for the postdisaster geocological safety layout construction.

1. Introduction

Geological disaster refers to the disastrous geological event that brings destruction to people's survival, resources, and living conditions under the action of various natural movements and artificial behaviors in geological conditions [1]. China has a vast territory and a large latitude and

longitude span. It is located at the regional boundary, and the tectonic movement is extremely intense [2]. Limited by the monsoon weather, the rainfall is concentrated and large, which is easy to cause various geographical natural disasters [3]. In the mountainous areas of southwest China, the population distribution is obviously affected by the terrain, and it is basically distributed along the relatively flat areas

such as valleys [4]. These conditions have caused many natural geological disasters in China, and the degree of life damage caused by geological disasters is also very high [5]. The occurrence of geological disasters not only does serious harm to the life safety and social life of the people in the affected areas, but also damages the geological-ecological structure and function of the area, especially in the mountainous urban areas sensitive to geological conditions. It also produced many secondary disasters, such as landslides and land collapses. Such natural disasters pose a serious threat to the original landform and ecology. If prevention and control are not carried out, the geological ecology will be in danger of collapse [6]. With the deepening of the country's western development and the construction of an all-round well-off society, the economic development of the western region is rapid and will continue for a long time [7]. However, because the western region straddles two major terrain steps in China, the terrain is undulating, the geological structure is complex, there are many fault zones, and the geological forces are very active [8]. The precipitation in the western region is concentrated, and the human agricultural production activities are strong, causing large-scale geological disasters [9]. Therefore, with the vigorous development of China's market economy, the economic losses caused by geological disasters will be even greater, seriously threatening the life and property safety of the Chinese people, and the development of China's economic and social construction will also be affected and constrained [10].

Since the 1980s, human beings have gradually realized the importance of the Earth's ecological environment for human survival and the development and the urgency of maintaining the Earth's natural ecological system. It has become a key scientific research topic in the world [11]. Ecological restoration is based on biological restoration. According to the actual situation, physical, biochemical restoration, and technical measures are optimized and combined to implement comprehensive restoration of the restoration object. Therefore, ecological restoration is also a disciplinary process [12]. By comprehensively repairing and restoring the degraded natural ecosystem, it will be able to repair the stability and plasticity of the natural ecosystem and its interdisciplinary ecological theory [13]. Geomorphic ecosystem is a complex system affected by many reasons. Because of its vulnerability and sensitivity, a slight change in the geomorphic ecology of the mountain area will involve changes for many reasons, such as space, environment, location, climate, and human factors [14]. Since ancient times, the relationship between people and geological environment has started from ignoring the geological environment, developed to damaging the geological environment, and finally transformed into the ability to restore and develop the geological environment [15]. Today's one-size-fits-all urbanization development mode, extensive land development and preemptive land excavation, and development and utilization have aggravated the geological and ecological crisis. The impact of geological disasters on the urban geological environment is also increasingly significant [16]. This paper summarizes the current urban

postdisaster restoration problems and geocological response strategies. Postdisaster geocological restoration is one of the key issues in the comprehensive response of mountain towns in China. In the postdisaster geological and ecological restoration research, most of the studies focus on the study of artificial environment such as mines [17].

In this context, when studying the evolution of geological natural disasters on the geological natural environment of mountain towns, the level and status of restoration of geological natural disasters on the geological environment of mountain towns are studied in depth, and the future trend of geological environment restoration is predicted according to this model.

2. State of the Art

2.1. Geological Disasters and Their Restoration. From the perspective of disaster types, geological disasters can be divided into major disasters and secondary disasters caused by major disasters. Large-scale natural disasters refer to large-scale natural disasters (large earthquakes, mudflows, collapses, etc.) caused by changes in geological conditions due to natural environment or human factors [18]. Natural disasters are also the largest floods encountered by contemporary cities and towns. Strong floods can also cause many secondary disasters. Therefore, strong natural disasters not only destroy a large number of buildings and hinder traffic, but also may cause many secondary disasters, such as tsunamis, mudstones, and town fires. There are also natural disasters, landslides, and mudstones that have a great impact on cities and towns. Some researchers have discussed the impact of different types of cities and towns on the occurrence of geomorphic natural disasters in urbanization. It mainly studies the relationship between the urbanization development process and the occurrence of geomorphic natural disasters [19].

However, there are few studies in the field of restoration of postdisaster mountain ecosystems, and the existing research is only a relatively general method, which lacks an assessment of the degree of restoration of postdisaster geological ecosystems. In addition, according to the needs of sustainable development policies, in the establishment and development of flexible cities (Figure 1), attention should be paid to the original landform environment to ensure that the use of natural resources cannot damage the original fragile landform ecological environment.

The definition of mountain town belongs to mountain city. It was originally called hillside city in Europe and America. In 1980, with the gradual development of scientific research on hillside cities in China, a corresponding mountain scientific research organization was internationally established. Its main research fields mainly focused on physical geography and biology, and its focus was mainly on the prevention of geological disasters, the conditions of the mountain environment, and the improvement of poor people's lives. There is little information on the concept of mountain cities. The concept of local towns is as follows: villages and towns under complex natural environment conditions such as complex landforms and high slopes.

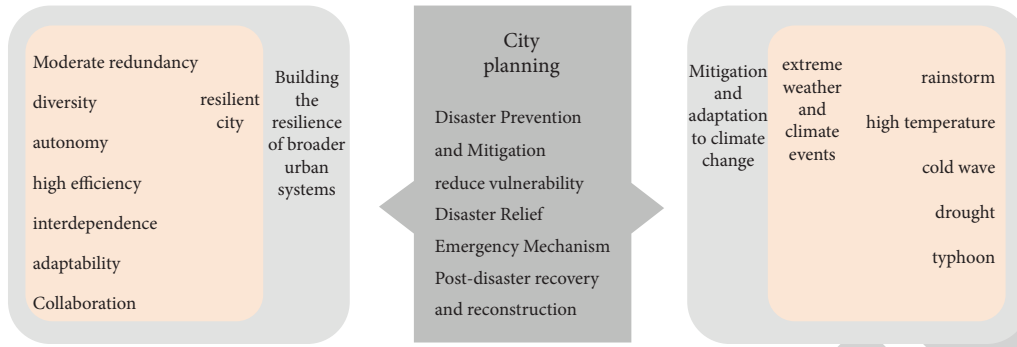


FIGURE 1: Urban planning and construction under resilient cities.

Small towns in mountain areas are the main part of the family environment of mountain residents. Their economy, land-use planning, road traffic, land-use development, and other fields have particularity. In addition, with the regional differences between mountainous areas and Pingchuan areas and the imbalance of local economic and social development, the development of small towns in mountainous areas is also affected by various factors such as fragile ecosystems, mountain landforms, and geological environment. It is urgent to support new science and technology. Typical characteristics of mountain cities are as follows: (1) the following characteristics must be possessed by the geographical location: the city is usually located in the interior of the mountain, or at the junction of the hills and the peaceful land. (2) Unique social traditional culture is as follows: urban and mountainous areas have formed an organic social whole in the process of historical growth. (3) Unique vertical spatial differentiation in space is as follows: combined with different mountain landforms, special residential areas and vertically differentiated living spaces are produced. If the above three points are achieved at the same time, it can be called a mountain city. Due to the research on mountain urbanization, the concept of mountain urbanization has gradually changed from a qualitative concept to a vector concept [20].

The theory of environmental restoration is a branch of ecology, which was first discussed in Keynes's book the restoration process of damaged ecosystems in 1980. Internationally, the main theories related to environmental remediation include environmental remediation, ecological security theory, and environmental remediation. The above concepts all contain the meaning of "restoration and development," that is, the process of restoring the damaged ecosystem to its previous or sustainable development. The difference is that ecological restoration emphasizes natural restoration, while ecological restoration focuses on the mutual promotion of artificial restoration and natural restoration. At present, the main concepts used in academic research are "ecological restoration" and "ecological restoration" [21]. This paper defines ecological restoration as the process of returning a damaged ecosystem from a direction away from its initial state to the initial state before the destruction. Ecological restoration further reduces damage to the environment by actively restoring ecosystems and actively and artificially transforming and constructing the current state. As

shown in Figure 2, the composition of ecological elements in the geological restoration of mountain towns not only reflects the composition of the urban ecosystem but also reflects the particularity of the geological environment.

2.2. Research on Geocological Restoration at Home and Abroad. From 1939, since K Troll introduced the theory of "Earth ecology," and the concept and theory of global ecology have gradually expanded. However, with the expansion of the global ecological theory, the understanding of this aspect is still at an early stage. At this stage, with the in-depth research of various scholars, the current research on geological-ecological restoration focuses on the geological-ecological theory research, geological environment research, geological environment informatization, and geological environment restoration and management (Table 1).

At this stage, foreign geological-ecological restoration mainly focuses on the damage caused by artificial environment and urbanization to the geological ecology. At the 32nd global geological conference held in August 2004, the research scope of the Earth's ecology was extended to the whole Earth circle, and relevant issues on the Earth's ecological restoration were considered. Subsequently, it is considered that the excessive use of land resources is one of the important factors that really affect the healthy development of urban geological ecology when they deeply studied the environmental degradation in the Dhaka metropolitan area of Bangladesh. This not only reduces urban vegetation but also destroys urban ventilation corridors. The main goal of Matthews is the urban geographical diversity under the destruction of mankind. In this paper, various aspects of maintaining natural diversity by using geocological technology are given, and then, the basic methods and steps of natural ecosystem management are defined.

Most of the research studies on the restoration of urban geological students' state in China focus on the evaluation of environmental benefits and urban pollution, and most of them are applied to the research methods of geological students' state. Professor Lin Jingxing, a scholar, pointed out that as an important part of the city, the study of geological student behavior must pay attention to the impact of human activities on the urban geological

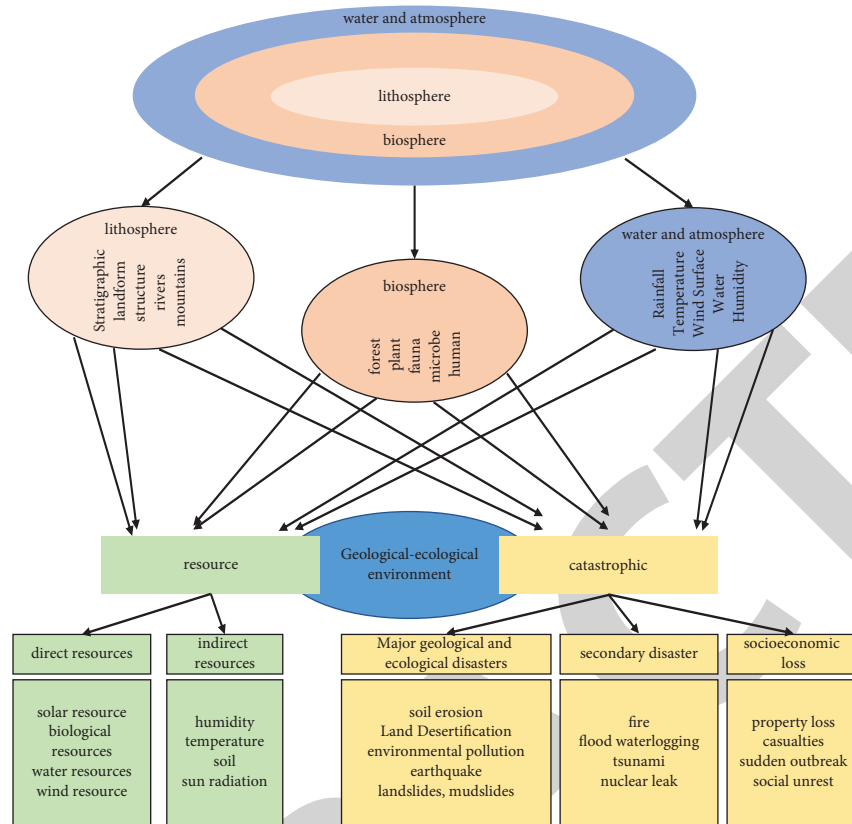


FIGURE 2: Main ecological elements of geological restoration in mountain towns.

TABLE 1: Main research objects, contents, and technical means of geological ecology.

Object	Discussing the content	Main means
Geological structure	Thickness of crust, changes in properties of constituent materials, and plate motion	3S technology, stratigraphy, and geodynamics
Mineral resources	Characteristics of magmatic activity, geophysical characteristics, and geochemical field	Mineralogy, structural geology, and petrology
Geological disaster	Characteristics of geological activities, geological stress field, and Earth plate movement	3S technology, geophysical exploration, and hydrogeology
Climate change	Ocean current energy, greenhouse effect, plant damage, and soil erosion	Structural geology, climatology, and meteorology
The combined effect of human activities	Focus on the comprehensive impact of human activities such as large-scale construction projects, unreasonable construction, and man-made destruction of the environment on the world's natural resources and environment	Geology, physical chemistry, and biology

environment, especially in the prevention, control, and restoration of urban pollution. Lu analyzed the problem of water cut off in the economic and social development of the Yangtze River Basin in China from the perspective of basin geology and urban land quality, and put forward the urban sustainable development strategy based on comprehensive land consolidation, geology, and ecological construction factors. Cui started with the relationship between natural resources and urban geological-ecological environment protection. Taking the Tarim River as an example, he advocated improving the soil according to the use of reasonable groundwater level, restoring the urban geological ecology, and ensuring the reasonable migration and balance of soil water and salt. Jiao Pengcheng took the bad

geological ecology of Lop Nur as the research object, and analyzed the appropriate landform and ecological restoration methods, such as improving the water storage and water transmission projects to save natural resources. Sheng Dongjin analyzed the impact of excessive development of hydropower resources on geological ecology and ecological disasters, and proposed scientific, rational, and orderly utilization of hydropower resources in rivers. Zhao Kunyu first studied the geological and environmental hazard factors of the mine, and on the basis of studying the innate and postnatal influence factors of environmental risks, he has put forward comprehensive defense countermeasures to prevent geological and environmental risks of the tailings dam.

In general, the research of domestic geological environment restoration mainly focuses on the geological environment damage caused by human action, including urban pollution and unreasonable development of resources, promoting the joint development of natural resources and avoiding geological risks. The damage to natural resources and environment caused by natural disasters has rarely been paid attention to.

3. Methodology

3.1. Interval Intuitive Fuzzy Set Theory. Because of the complexity of the emergency action conditions of the ecological environment restoration project and the uncertainty of the policy content, and the limitation of the administrative personnel's mental ability and cognitive ability, the decision-makers often cannot correctly convey the policy content, but the intuitive understanding can be very good, provide ambiguous data, introduce intuitionistic fuzzy sets into time intuitionistic fuzzy sets, and thus gave the concept of interval intuitionistic fuzzy sets.

Let the universe of discourse U be a nonempty finite set and call $\tilde{a} = \langle X, \mu_a(X), \nu_a(X) | X \in U \rangle$ an interval

intuitionistic fuzzy set. The interval intuitionistic fuzzy set is an ordered interval pair composed of a membership interval and a nonmembership interval where the element X in the universe U belongs to \tilde{a} , generally denoted as

$$\begin{aligned} \tilde{a} &= ([\mu_a^-, \mu_a^+], [\nu_a^-, \nu_a^+]) \text{ or,} \\ \tilde{a} &= [a, b], [c, d]. \end{aligned} \quad (1)$$

Among them, μ_a^-, ν_a^- represent the lower bounds of the membership and nonmembership intervals, μ_a^+, ν_a^+ represent the upper bounds of the membership and nonmembership intervals, and Π_a^-, Π_a^+ represent the upper and lower limits of the hesitation degree interval, respectively, and satisfy

$$[\mu_a^-, \mu_a^+] \in [0, 1], [\nu_a^-, \nu_a^+] \in [0, 1], \mu_a^+ + \nu_a^+ \leq 1. \quad (2)$$

Definition 1. Let $\alpha_j = ([a_j, b_j], [c_j, d_j]) (j = 1, 2, \dots, n)$ be a set of interval intuitionistic fuzzy numbers, and IIFWA $Q^n \rightarrow Q$, if

$$\text{IIFWA}_\omega(\alpha_1, \alpha_2, \dots, \alpha_n) = \sum_{j=1}^n \omega_j \alpha_j = \left(\left[1 - \prod_{j=1}^n (1 - a_j)^{\omega_j}, 1 - \prod_{j=1}^n (1 - b_j)^{\omega_j} \right], \left[\prod_{j=1}^n c_j^{\omega_j}, \prod_{j=1}^n d_j^{\omega_j} \right] \right). \quad (3)$$

Among them, $\omega = (\omega_1, \omega_2, \dots, \omega_n)^T$ is the attribute weight of $\alpha_j (j = 1, 2, \dots, n)$, which satisfies $\omega_j \in (0, 1)$ and $\sum_{j=1}^n \omega_j = 1$; then, the function IIFWA is called an n-dimensional interval intuitionistic fuzzy weighted average (IIFWA) operator. In particular, when the attribute weight is $\omega = (1/n, 1/n, \dots, 1/n)^T$, IIFWA degenerates into the interval intuitionistic fuzzy arithmetic mean IIFAA operator.

From the obtained attribute weights, the group decision matrix is assembled using the interval intuition fuzzy weighted average operator to obtain the comprehensive evaluation value of each scheme, the positive ideal scheme is obtained according to the definition, and the projection of each alternative scheme on the positive-ideal scheme vector is calculated, that is,

$$\text{Prj}_{x^+}(X_i) = \frac{C_{\text{IIVIFS}}(X_i, X^+)}{\sqrt{E_{\text{IIVIFS}}(X^+)}}. \quad (4)$$

The larger the P , the closer the alternative x is to the positive-ideal solution x^+ , and the better the alternative X_i . Therefore, the closest degree formula for each method is defined as follows:

$$C(X_i) = \frac{\text{Prj}_{x^+}(X_i)}{\sum_{i=1}^n \text{Prj}_{x^+}(X_i)}. \quad (5)$$

3.2. Improvement of Indicators of Geological-Ecological Footprint Model

3.2.1. Consumption Item Verification. When carrying out the ecological footprint model of geography, the consumer species were verified for the first time. Ecological footprint

refers to the comprehensive transformation of product ecological commodities of people's consumption behavior. Consumption items will affect the correctness of the statistical composition of ecological footprint. Generally, the more detailed the classification is, the more accurate the quantity of ecological products representing people's consumption behavior is, so the calculation result of ecological footprint is more reliable. However, because it is difficult for people to grasp the specific quantity of various consumption patterns in each region in the actual data collection process, investigators in different regions use different consumption types when calculating the ecological footprint of each region, resulting in differences in the calculation of the ecological footprint.

3.2.2. Average Output Check. The average production of ecological products at the national level refers to the production of ecological products based on the biological production area of the country. According to the statistical data of China in recent years and the current actual output, we calculate the annual average production capacity of major crops in the country, so as to enhance the timeliness, accuracy, and practicality of geoenvironmental footprint statistics.

According to the analysis of consumption species in the previous section, according to the geographical environment of resources, the products are divided into four types, namely, farmland, forest, grassland, and water, and the global output is calculated, respectively. In the past five years, the average output per unit of various planting

products was regarded as the national per capita output. The calculation formula is as follows:

$$\overline{EP}_{1k} = \sum_{i=1}^5 \frac{PK_i}{Ak_i}. \quad (6)$$

Among them, EP_{1k} refers to the national average production (kg/hm^2) of the K -th category of products; i refers to the year; in the actual calculation, the average output of the last 5 years is averaged; and PK_i refers to the i -th category of ecological products of the K -th category. Ak_i refers to the total sown/produced area of the country in the i -th year of the k -th ecological product.

(1) *Farmland Products*. The main statistical sources are China Statistical Yearbook and National Agricultural Economic Yearbook. If there is a difference between the same figures in the data center in one year, the same figures given in the high-level yearbook (National Statistical Yearbook) and the New Year yearbook (2018) shall prevail.

(2) *Woodland Products*. Forest products are mainly divided into economic commodities such as trees, chestnuts, and walnuts. The occupied ecological production land is divided into forest land. Because most forest products are hybrid varieties, it is difficult to grasp an accurate yield range. The calculation method is to take the total area of national forest land over the years as the planting area of trees' various forest products. The formula is as follows:

$$\overline{EP}_{2k} = \sum_{i=1}^5 \frac{P_{ki}}{A_{fw}}. \quad (7)$$

Among them, EP_{2k} refers to the average output of forest products (m^3/hm^2), i refers to the year, the average value of the average production in the past 5 years is taken in the calculation (m^3), and A_{fw} refers to the national forest area (hm^2).

(3) *Grass Products*. Wheat belongs to farmland, while dry grass/grass belongs to pasture, and wheat bran belongs to the epidermis of grain or barley after threshing. It is a concomitant product of wheat in the growth process and does not account for the environmental footprint. However, in the processing competitiveness of soybean meal, sunflower seed meal, wheat germ meal, and other foods, the by-products left after the initial processing or extraction are classified as waste and do not occupy the environmental footprint. This is true of grass flour, peanut shell starch, and rice bran. Therefore, the statistics of the geological environmental footprint of Chinese rabbit meat mainly consist of the following two aspects, and the formula is as follows:

$$A_m = A_{m1} + A_{m2} = \sum \frac{P_m \cdot P_{rec} \cdot E_c \cdot eC_i}{eP_i} + \frac{P_m \cdot P_{reg}}{EP_1}, \quad (8)$$

where agile modeling is the geographical ecological footprint of rabbit meat, $AM1$ is the natural ecological footprint of agricultural rabbit meat, AMS is the

ecological footprint of grassland rabbit meat, PM is the production of rabbit meat, and $prec$ is the production of rabbit meat. The proportion of self-cultivated land here is 86%. EC refers to the average grain demand per unit of rabbit meat in the United States. ECI refers to the proportion of the i -th raw material in the average grain demand in the United States. EPI refers to the average output of the i -th raw material in the United States. $Preg$ refers to the proportion of rabbit meat produced from m ranch in the United States, and the average meat production of EP II ranch.

3.2.3. *The Improved Mountain Geological-Ecological Footprint Model*. Based on the above improvements to the traditional ecological footprint model, according to the characteristics of mountain towns and their own conditions, a localized study of the geological-ecological footprint factor is carried out, taking into account the impact of pollutants and water resources on the geological ecology. In order to further improve the theory of geological-ecological footprint and accurately evaluate the sustainable development of Lushan County, the modified ecological footprint model is as follows:

$$\begin{aligned} EF_G &= EF + EFn + EFW = \sum_{i=1}^n \frac{Ci}{EPI} EQi + \frac{QC}{AC} EQi \\ &\quad + \frac{W}{P} \times \gamma w, \\ EC_G &= EC + ECW = \sum_{i=1}^n (AiEQiYFi) \cdot (1 - 12\%) \\ &\quad + 0.4 \times \psi \times \gamma \times \frac{Q}{P}, \\ \frac{ED'}{ER'} &= EC' - EF'. \end{aligned} \quad (9)$$

In the formula, EFG is the environmental footprint of improved landforms, EF is the traditional agricultural ecological footprint, EFN is the environmental footprint of air pollutants, EFW is the environmental footprint of water resource utilization, ECG is the improved environmental carrying capacity, and EC is the environmental footprint. In the traditional natural bearing capacity, ECW refers to the natural carrying capacity of water resources, ED' refers to the improved ecological deficit, and ER' refers to the improved natural surplus.

3.3. *Ecological Restoration Evaluation Index*. The energy consumed in the ecological footprint will form a large amount of carbon dioxide and greenhouse gases in the process of utilization, which will affect the global geology and environment. Therefore, the ratio of ecological footprint to natural environment intensity can be used as the calculation result formula of the ecological pressure index. It expresses the environmental geomorphic characteristics and

natural environment pressure degree of a certain area. The calculation result formula is expressed as follows:

$$EPI = \frac{EF}{EC}, \quad (10)$$

where EPI is the average ecological pressure index of renewable resources, EF is the average environmental footprint of the whole renewable resource utilization system, and EC is the environmental carrying capacity. In order to supplement the lack of the concept of ecological sustainability in the ecological index, the concept of ESI (ecological sustainability index) is introduced in this paper to further explain the problem of ecological sustainability that is different from the single evaluation of ecological footprint. meets human ecological needs. The ecological sustainability index is an integrated indicator system that tracks 21 elements of environmental sustainability, including natural resources, past and present pollution levels, environmental management efforts, environmental contribution to international public affairs, and social capacity to improve environmental performance over the years. Its formula is as follows:

$$ESI = \frac{EC}{EF + EC}. \quad (11)$$

In the formula, ESI is the ecological sustainability index; EF is the average ecological footprint of renewable resources in the region, while EC is the ecological carrying capacity. According to the calculation results of the ecological recovery index and economic recovery index in the previous section, a comprehensive assessment model of postdisaster mountain geological ecology is established, and the comprehensive assessment level of geological-ecological restoration in this field is measured. Finally, EPI, EOI, and EECI are selected to establish the geoeological restoration index, and EPI is the negative index, EOI and EECI are the positive indexes. These three indexes are standardized, and the calculation formula of geoeological restoration index is obtained as follows:

$$GRI = \left(\frac{EPI_{\max} - EPI}{EPI_{\max} - EPI_{\min}} \right) + \left(\frac{EOI_{\max} - EOI}{EOI_{\max} - EOI_{\min}} \right) + \left(\frac{EECI_{\max} - EECI}{EECI_{\max} - EECI_{\min}} \right). \quad (12)$$

In the formula, EPI_{\max} , EPI_{\min} , EOI_{\max} , EOI_{\min} , $EECI_{\max}$, and $EECI_{\min}$ represent the maximum and minimum values of the ecological pressure index, ecological occupation index, and ecological economic coordination index, respectively.

4. Result Analysis and Discussion

Geological and ecological environment protection refers to the organic process of geological natural environment, biological-ecological environment, and human living

environment. It is generally composed of the following three elements: geological natural environment, biological-ecological environment, and human living environment (Figure 3). The geological-ecological environment is the main carrier of geological environment protection. It not only controls the progress of human social activities but also promotes the change of human social environment to geographical environment, making geological-ecological environment and human social environment closely related. Therefore, geology has a great impact on the geographical and ecological environment of urban and rural areas in mountainous areas. Historical experience and recent scientific research have proved that one of the most direct and main reasons affecting the geological ecology of mountain towns is geological disasters, and hydrological disasters have seriously damaged mountain towns. Generally speaking, there are many reasons for geological disasters affecting the geological ecology of mountain towns, but they can be summarized into two aspects, namely, positive factors and negative effects. The two interact and promote each other. This section takes the hard-hit area of 4.20 Lushan earthquake as an example to evaluate the geological-ecological restoration of mountain towns.

4.1. Basic Situation of Research Object Geology and Disasters.

Lushan County is located in Ya'an, Sichuan, China. It is located in the west of Sichuan Basin, northeast of Ya'an, 14 kilometers away from the city center. There are many rivers such as Qingyi River and Lushan River in the county. It is adjacent to Wenchuan City in the north, Chongzhou City in the northeast, Yucheng District in Ya'an City in the southeast, Tianquan City in the southwest, and Baoxing City in the northwest. At 08:02 on April 20, 2013, an earthquake of magnitude 7.0 was triggered in Lushan County, Ya'an, Sichuan (30.3 degrees north latitude and 103.0 degrees east longitude), with a focal depth of 13 kilometers. Strong earthquakes were felt in Chengdu, Chongqing, and Shaanxi [16]. The epicenter occurred in the south section of Longmenshan fault fracture zone (Figure 4). The total fracture diameter of the earthquake damage was 35 minus 40 kilometers. Within three weeks after the epicenter, there were a total of 8014 aftershocks, and 121 aftershocks with a magnitude of 3.0 or above.

Sichuan Bureau of Surveying and Mapping and Geographic Information has inspected the aerial photography of the area with the strongest disaster in the 1300 square kilometer area of the country, obtained the low-altitude resolution pictures, compared and interpreted the high-resolution pictures before the Lushan earthquake, and re-evaluated the mountain disaster as a strong earthquake and secondary disaster. The survey shows that the Lushan earthquake has caused nearly 1400 secondary disasters, most of which are collapses and slides. The earthquake has not only brought a large number of casualties and losses, but also brought great damage to the geological ecology of the scenic county. It is urgent to restore the geological ecology after the disaster.

Three earthquake-stricken areas near Lushan Mountain with a magnitude of 4.20 and a magnitude of 9 were selected

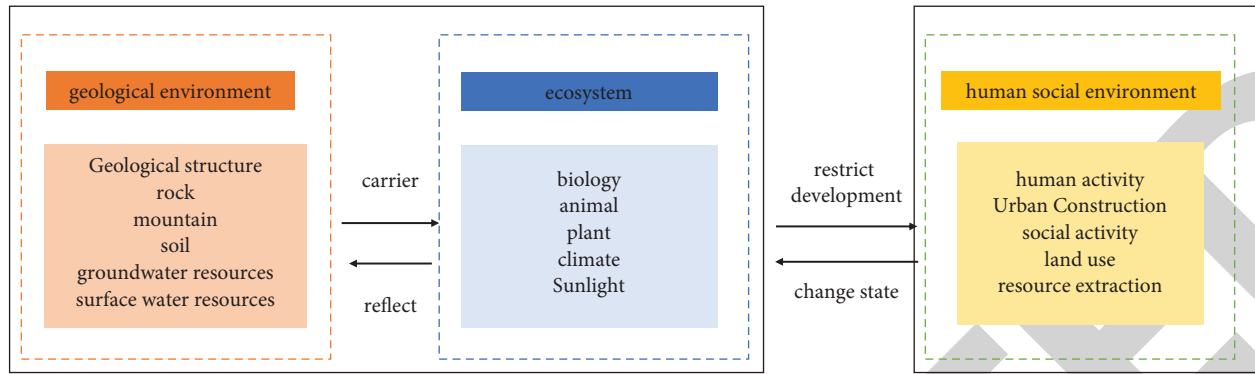


FIGURE 3: Interrelationships between geological and ecological factors.

as cases to carry out a postdisaster assessment of the geological environment restoration in mountainous areas and towns. It is high in the northwest and low in the southeast. The highest altitude is 5360 m. The landform of this area is roughly divided into four categories, mainly the middle mountain belt and the mountain belt, with complex landform (as shown in Figure 5).

4.2. Geological-Ecological Surplus/Deficit over the Years. According to the above calculation method, the ecological profit and loss of Lushan County in 2010–2017 can be calculated, as shown in Figure 4. It can be seen from the map that from 2010 to 2017, the geological environment of Lushan County had an ecological deficit, and the burden on the economic and social development of Lushan County was greatly greater than the bearing capacity of the geographical environment. In 2010, the Earth's ecological construction deficit was $-1^2/\text{people}$. In 2017, the deficit of the Earth's ecological construction was $-0^2/\text{people}$. From 2010 to 2017, the geological environment deficit of Lushan County has decreased by more than 43%. This also shows that the scenery is so beautiful that a lot of efforts have been made in natural resources, ecological construction, and environmental protection. In terms of some measures, although the Lushan earthquake event in 2013 has had a significant negative impact on it, it has generally maintained a deficit situation in geographical and ecological construction, and the negative impact on it is gradually decreasing. However, in 2017, it still maintained a geographical and ecological deficit.

Except for forest land and water resource land, which are in a state of geological and ecological surplus, other types of lands are in a state of geological and ecological deficit, and most of the ecologically productive land in Lushan is in a state of unsustainable development. Even though the geological-ecological deficit of Lushan County has been gradually reduced over the years, the ecological carrying capacity has continued to decline, reflecting that Lushan is currently in a state of geological-ecological imbalance, the sustainable capacity of geological-ecological development is poor, and the geological-ecological security is not optimistic.

According to the calculation method of the geological and ecological footprint of Lushan County in 2017, the geological and ecological footprint of Tianquan County over

the years and the ecological surplus/deficit of Tianquan County over the years were calculated, respectively, and the drawing results are shown in Figure 6.

It can be seen from Figure 6 that the geological-ecological footprint of Tianquan County has gradually decreased from 2010 to 2017, from 1.94732 hm^2 in 2010 to 1.25414 hm^2 in 2017, with a decrease of 35%. In the year when the earthquake was induced, the average geoecological footprint of farmland and grassland also showed a significant reduction trend, but on the whole, the average geoecological footprint of farmland and water area has been significantly reduced. This shows that the county has also strengthened the ecological protection of farmland and water areas. On the one hand, it also shows that the demand for major commodities of farmland and water surface gradually decreases with the increase in time. On the other hand, it also shows that the county has strengthened the ecological protection of farmland and water surface.

Judging from the change degree of the earthquake on the geological environment of Tianquan County, the number of land ecological footprints significantly decreased in 2013, from 1.05727 hm^2 in 2012 to 0.66710 hm^2 in 2013, which shows that the Lushan earthquake in 2013 has indeed changed the local geological environment. The ecological environment has brought about corresponding negative impacts, including the decline of productive demand, the change in labor force composition, and the change in social balance factors (for example, the production capacity of live pigs in China reached 14186 tons in 2012, but only 10234 tons in 2013, with a decrease of nearly 30%). Subsequently, the landform and ecological footprint gradually decreased, indicating that the county pays attention to the geological environment, curbs environmental pollution, and pays attention to the rational use of natural resources.

From 2010 to 2017, in terms of the overall share of natural resources and environmental footprint of various types of soil in Tianquan (Figure 7), the geological-ecological footprint of farmland and fossil energy soil is the highest in Tianquan's geological environmental footprint over the years, which is 406.86% and 44.18%, respectively, followed by water and urban and rural construction land and forest. The share of water and grassland is basically the same. However, the share of water resources in the land is the least. From the change

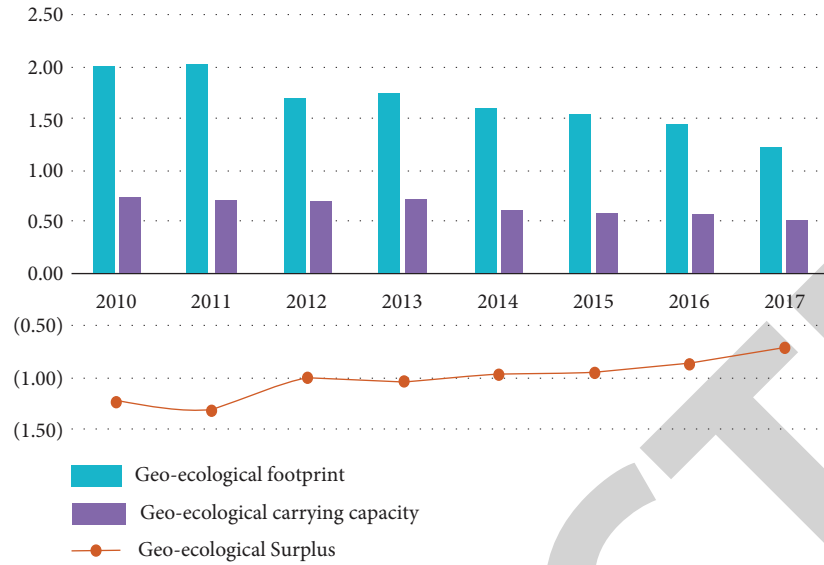


FIGURE 4: The per capita geo-ecological deficit of Lushan County from 2010 to 2017 (hm²/person).

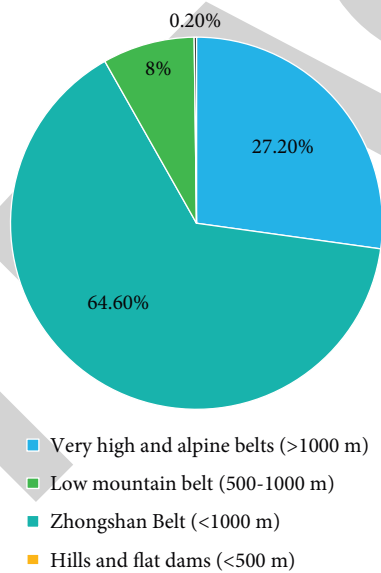


FIGURE 5: The proportion of terrain classification in the study area (changed to a pie chart).

trend of the proportion, the proportion of farmland has greatly changed, from 55.56% in 2010 to 39.62% in 2017. The share of farmland in the geological environment footprint of Tianquan County is gradually decreasing. However, the share of another series of land use represented by fossil energy soil is gradually increasing. Taking the consumption of fossil resources as an example, its utilization rate has significantly increased from 36.70% in 2010 to 50.98% in 2017, indicating that fossil energy consumption is gradually replacing arable land and other crops. The energy cost has also become the main reason restricting the geological ecology of Tianquan County. The utilization scale of fossil resources in Tianquan County is still large. The environmental losses caused by the utilization of natural resources must be paid attention

to the rational use of resources. Generally speaking, the proportion of per capita pollution geo-ecological footprint in Tianquan County is gradually decreasing (except for water areas), which shows that Tianquan County has made certain improvements in environmental pollution, and the pollution geo-ecological footprint is gradually decreasing.

Similarly, the geological and ecological footprint of Baoxing County over the years can be calculated and drawn as shown in Figure 8.

It can be seen from Figure 8 that the geographical and ecological footprint of Baoxing County has gradually decreased from 2010 to 2017, from 1.76533 hm² in 2010 to 1.29829 hm² in 2017, with a decrease of 26%. The pace of geological ecology has also shown a decreasing trend, which



FIGURE 6: Tianquan's total geological and ecological footprint from 2010 to 2017 (hm²/person).

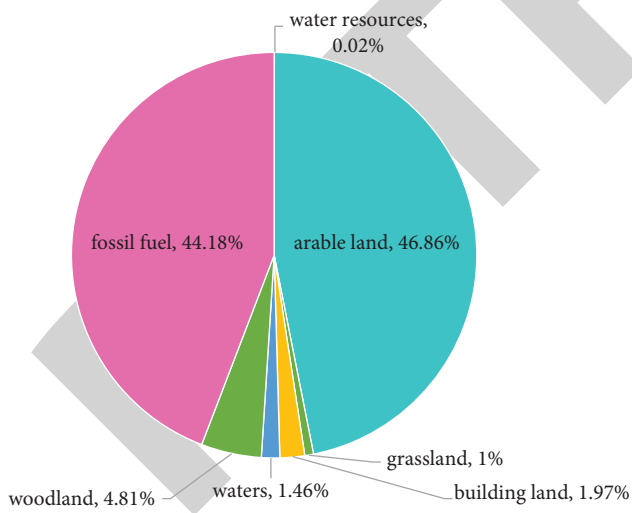


FIGURE 7: Average ratio of geological and ecological footprint of various types of land.

may be due to the gradual decline of the average consumption of productive goods with the increase in time and the change in the number of people. However, in 2013, although the geological-ecological footsteps of farmland and grassland decreased, the geological-ecological footprints of other farmland increased, indicating that the geological-

ecological footsteps of farmland and grassland significantly decreased in that year due to the earthquake (the decrease in the year of the earthquake was much greater than the normal decrease in previous years). It can be inferred that the Lushan earthquake in 2013 did have some negative effects on the geographical ecology of Baoxing, such as the decline of productive agriculture, the change in population composition, and the changes in equilibrium factors. We take corn for example. In 2012, the corn output of Baoxing County reached 9983 tons, while in 2013, it reached 9168 tons, a decrease of 800 tons. Subsequently, the gradual disappearance of geological and environmental footprint indicates that Baoxing County pays attention to postdisaster geological and environmental management and tries its best to reduce the harm of natural disasters.

The change in natural environmental pressure has a positive correlation with environmental chemistry and ecological security. The smaller the ecological pressure, the lower the natural ecological pressure on the region. The environment is relatively safe, and the natural ecological balance is better. On the contrary, it also means that the natural ecological environment in this area is relatively unstable and prone to natural environment disorder. Figure 9 shows the ecological pressure evaluation indicators before and after ecological restoration in the three counties.

We can see that among the three postdisaster demonstration counties, the scenic county has the highest

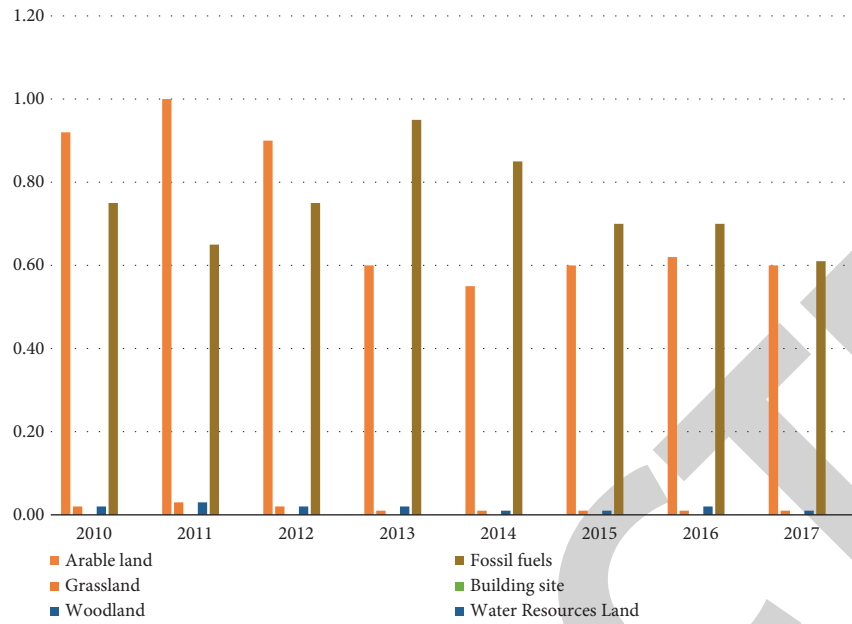


FIGURE 8: Baoxing's total geological and ecological footprint (hm²/person) from 2010 to 2017.

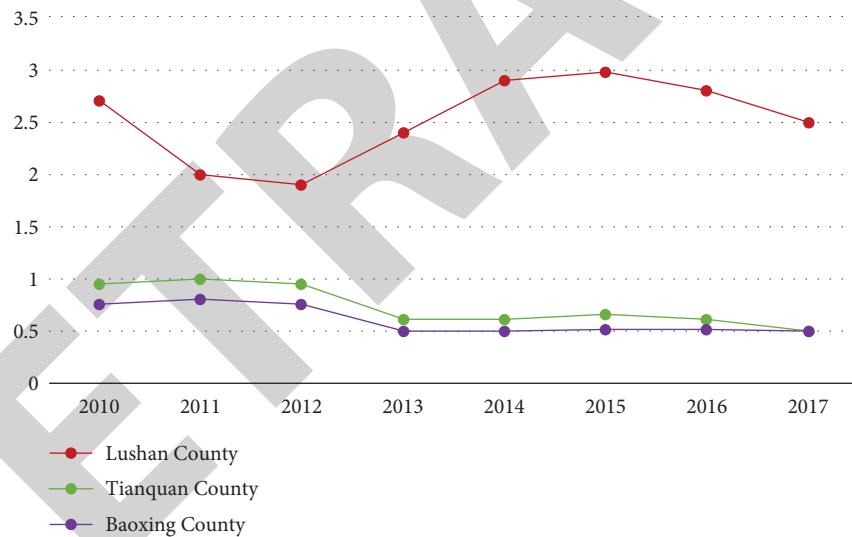


FIGURE 9: Levels of ecological pressure evaluation indicators in Lushan County, Tianquan County, and Baoxing County.

ecological pressure index, reaching an extremely unsafe and unstable area. The reason is that the scenic county not only has low environmental intensity and gradually decreases, but also has a large and gradually expanding environmental footprint. Lushan County has extremely high environmental pressure, which has greatly broken the safety boundary of EPI = 1. Before 2010, the environmental resources consumed in a beautiful year will be supplemented by more than two scenic production resources. The geological environment safety of Lushan County is not optimistic, and corresponding geological environment restoration measures are urgently needed to change the situation. For Baoxing City and Tiancheng City, the ecological pressure values are relatively small, and most of them are in a relatively safe range.

From the perspective of future geographical and ecological security, it is necessary to protect the existing geographical and environmental conditions under the current favorable geographical and environmental conditions. The restoration of geology and environment after the disaster has been gradually promoted.

5. Conclusions

As the restoration and improvement of damaged geological ecology in urban and rural areas not only depend on urban planning and ecological restoration planning, but also have a very complex connotation, level, and personnel composition, they need multidisciplinary coordination,

multidepartments, and multiparticipation in geological-ecological restoration and reconstruction. In the formulation of systems and specific technologies, the urban planning management organs represented by the Urban Planning Commission must make overall planning, command the overall situation, and link up and coordinate with other government departments. According to relevant laws and regulations, specific postdisaster urban geoeological restoration policies and measures will be issued to ensure the effective implementation and implementation of post-disaster urban geoeological restoration policies and measures.

Based on the study of geological disasters and the role of mountain geological environment, this paper combines the related concepts and technologies of postdisaster geomorphic environment restoration assessment abroad. The geoeological footprint model is used to evaluate the level of geoeological restoration in the earthquake stricken areas. First of all, on the basis of elaborating the theory and methods of the conventional ecological footprint model, it further solved the shortcomings of its research, further expanded its research scope, expanded the pollutant accounts related to the mountain geological environment, further established the geoeological footprint model, and then established the assessment model of the postdisaster geoeological restoration level. The ecological restoration level of the postdisaster geography was assessed using the combination of natural restoration indicators and ecological restoration indicators.

In the survey of the per capita deficit of geological and ecological construction in Lushan County from 2010 to 2017, it was found that although the deficit of geological and ecological construction in Lushan County has been gradually decreasing in recent years, the intensity of ecological construction has been decreasing. Through the analysis of the average proportion of geological-ecological footprint of various land users, it is concluded that the proportion of pollutant geological-ecological footprint per capita is gradually decreasing (except for water areas), and the geological-ecological footprint of pollution is gradually declining. In the ecological pressure assessment index grades of Lushan City, Tianquan City and Baoxing County of Lushan County, Tianquan County, and Baoxing County, its geological and ecological safety is not optimistic, and the corresponding geological and ecological restoration means are urgently needed to improve the current situation. Baoxing County and Tianquan County have low ecological pressure values, and most of them are located in relatively safe areas. In the future national geoeological security, it is necessary to gradually promote the restoration of post-disaster geoeological construction according to the actual problems of current geoeological construction under the condition of ensuring the normal geoeological conditions.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the Industrial Support and Guidance Project of Colleges and Universities in Gansu Province in 2019: Research and Key Technology of Coal Mine Reclamation in Gansu Province (no. 2019C-16).

References

- [1] S. Shin, H. Cho, H. C. Kang et al., "Assessment of the value and development of geological heritages in Gangwon Province, Korea," *Journal of the Geological Society of Korea*, vol. 56, no. 6, pp. 683–702, 2020.
- [2] D. Chen, Y. Zhao, and J. Liu, "Characterization and evaluation of rotation accuracy of hydrostatic spindle under the influence of unbalance," *Shock and Vibration*, vol. 2020, no. 2, 16 pages, Article ID 5181453, 2020.
- [3] J. Batista dos Santos Espinelli Junior, G. von Brixen Montzel Duarte da Silva, R. Branco Bastos, E. Badiale Furlong, and R. Carapelli, "Evaluation of the influence of cultivation on the total magnesium concentration and infusion extractability in commercial arabica coffee," *Food Chemistry*, vol. 327, no. 15, Article ID 127012, 2020.
- [4] A. V. Ivanov and A. V. Strizhenok, "Evaluation of the effectiveness of dust screens and the possibilities of taking into account their influence in software models," *Journal of Physics: Conference Series*, vol. 1728, no. 1, Article ID 012008, 2021.
- [5] M. Simarro, J. J. Castillo, J. A. Cabrera, and S. Postigo, "Evaluation of the influence of the speed, preload and span length on the contact forces in the interaction between the pantograph and the overhead conductor rail," *Engineering Structures*, vol. 243, no. 4, Article ID 112678, 2021.
- [6] T. Zeng, K. Yin, H. Jiang, X. Liu, Z. Guo, and D. Peduto, "Groundwater level prediction based on a combined intelligence method for the Sifangbei landslide in the Three Gorges Reservoir Area," *Scientific Reports*, vol. 12, no. 1, pp. 11108–11172, 2022.
- [7] H. Li, M. G. Lotter, K. Kuman, L. Lei, and W. Wang, "Population dynamics during the acheulean at ~ 0.8 ma in east and southeast asia: considering the influence of two geological cataclysms," *Palaeogeography, Palaeoclimatology, Palaeoecology*, vol. 562, no. 1, Article ID 109927, 2021.
- [8] A. Almajed, A. Aldhowian, and K. Abdelrahman, "Geological and geotechnical evaluation of limestone rocks along the riyadh metro project (riyadh city, Saudi arabia)," *Arabian Journal of Geosciences*, vol. 14, no. 2, pp. 89–15, 2021.
- [9] V. Deshko, I. Bilous, I. Sukhodub, and O. Yatsenko, "Evaluation of energy use for heating in residential building under the influence of air exchange modes," *Journal of Building Engineering*, vol. 42, no. 3, Article ID 103020, 2021.
- [10] B. Pawar, S. Park, P. Hu, and Q. Wang, "Applications of resilience engineering principles in different fields with a focus on industrial systems: a literature review," *Journal of Loss Prevention in the Process Industries*, vol. 69, no. 2, Article ID 104366, 2021.
- [11] G. Tian, Z. Guo, and S. Li, "Optimization of tawa landslide treatment scheme based on the AHP-fuzzy comprehensive evaluation method," *IOP Conference Series: Earth and Environmental Science*, vol. 598, no. 1, Article ID 012032, 2020.

Retraction

Retracted: An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] L. He, T. Zhu, and M. Lv, "An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4250462, 12 pages, 2022.

Research Article

An Early Warning Intelligent Algorithm System for Forest Resource Management and Monitoring

Liheng He, Tingru Zhu, and Meng Lv 

College of Civil Engineering, Nanjing Forestry University, Nanjing 210037, China

Correspondence should be addressed to Meng Lv; lvmeng@njfu.edu.cn

Received 26 July 2022; Revised 10 September 2022; Accepted 23 September 2022; Published 11 October 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Liheng He et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The development of remote sensing technology has passed an effective means for forest resource management and monitoring, but remote sensing technology is limited by sensor hardware equipment, and the quality of remote sensing image data is low, which is difficult to meet the needs of forest resource change monitoring. This paper presents a remote sensing image classification method based on the combination of the SSIF algorithm and wavelet denoising. Forest information is extracted from PALSAR/PALSAR-2 radar remote sensing data. The forest distribution map is generated by pixel level fusion algorithm, and the accuracy of the forest distribution map is evaluated by a confusion matrix. The remote sensing image is spatio-temporal fused by the SSIF algorithm to capture more details of forest distribution. The simulation analysis shows that the overall accuracy of the forest classification results obtained by the fusion algorithm is $96\% \pm 1$, and the kappa coefficient is 0.66. The accuracy of forest recognition meets the requirements.

1. Introduction

Forest resources are natural resources with the largest land area, the widest distribution, the most complex components, and the most perfect functions. They are also the main body of the terrestrial ecosystem. They play an irreplaceable role in maintaining species diversity, storing carbon, storing water, providing biological resources, and maintaining ecological balance. Forest resources are affected by natural or non-natural factors, and their area, quality, and function are constantly changing. It is a dynamic renewable resource. Accurately grasping the growth and decline of forest resources is the premise of strengthening the construction of a national ecological civilization, and it is also a necessary measure to maintain the sustainable development of the social economy. The traditional forest resources monitoring method based on the artificial ground investigation has the problems of a heavy workload, high cost, long time-consuming, and low efficiency, which is difficult to meet the current needs of forest resources change monitoring. Remote sensing technology has the advantages of detecting a wide range of the ground, obtaining ground information

quickly, and being less restricted by the ground. It can quickly reflect the real terrain and landform, and can effectively save human and material resources. It has become one of the important means to quickly obtain information in the current forest resource change monitoring [1, 2]. Nowadays, with the rapid development of remote sensing technology, high spatial resolution remote sensing images can better capture the subtle changes of forest resources in the monitoring of forest resources change, which is conducive to the accurate monitoring of forest resources. It is generally considered to be the most important data source for forest resources change monitoring.

2. Literature Review

With the continuous development and innovation of information technology architecture, forest resource dynamic monitoring technology has also gradually developed. With the support of various technologies and years of development, forest resource monitoring and management technology has continuously achieved remarkable results in forestry. Forest resources are constantly changing with the

passage of time and have timeliness and variability. The traditional forest resources investigation and monitoring methods can not fully and effectively monitor forest resources. In terms of technology operation and management, it is difficult to meet the current needs of forestry information monitoring and management. With the continuous innovation and integration of forest resource information management technology and forest resource monitoring technology, relevant forestry departments and forestry research institutes have used computer technology to develop and implement forest resource management system and monitoring system, providing support for forest resource management. At present, the information technology of forest resources has been continuously developed. GIS, RS, and GPS technologies are used to comprehensively reform and adjust the forest resources monitoring system, realize the timely, accurate, and efficient monitoring of forest resources, update the forest resources database, and realize the dynamic adjustment and update of forest resources monitoring [3].

In the forest resources monitoring system, the relevant satellite data are used to conduct a comprehensive remote sensing monitoring experiment on the resources of forestry areas. The space and aviation remote sensing monitoring platform is used to monitor and investigate the ground in combination with GIS technology and GPS, and the distribution type, area, and land type of forest resources are checked. In the middle and late stage, the use of 3S technology in forest resources research institutes, forestry experimental bases, and other places across the country has achieved certain results in the monitoring and management of forest resources, but further development is needed. For the application of 3S technology in forest resources monitoring, relevant scholars and researchers proposed that the actual application of monitoring and the experimental methods of operation should be continuously improved to achieve real-time, efficient, and accurate monitoring of forest resources and provide practical methods for forestry management and relevant departments [4]. With the continuous development of science and technology, remote sensing technology has played an important role in the monitoring of forest resource changes. Remote sensing technologies, such as aerospace and aviation, have shown the advantages of traditional monitoring methods in the monitoring of forest resource changes from the accuracy, timeliness, and coverage of operations. The characteristics of its space technology, such as multiplatform, multiangle, multisensor, high resolution, and hyperspectral, make the remote sensing image technology shorten the monitoring cycle time than the traditional monitoring technology in the monitoring of forest resource changes, and solve the shortcomings of forest resource data update difficulties, forest resource monitoring data lack of spatial distribution information and so on. The scope of forest resources monitoring has been expanded, the processing efficiency of the work has been improved, and the current situation and past changes of forest resources have been grasped in time. As a powerful tool for forest resource management and data processing, GIS makes up for many deficiencies of

traditional forest resource management and monitoring in data updating, data transmission, and data processing. For example, the huge amount of redundant data, the timeliness of forest resource data, and the integration and updating of forest resource data and remote sensing graphic data. However, due to the rapid development of GIS in the later stage, it has been widely used in forestry, agriculture, forest resources monitoring and investigation, urban planning, military deployment, and other fields. In the application of forest resources management, there are mainly forest resources management information system, forest resources auxiliary decision-making system based on GIS, forest fire prevention system, and so on [5, 6].

3. Forest Information Extraction Based on PALSAR Image

3.1. GEE Remote Sensing Cloud Computing Platform. Use JavaScript API to write functions to calculate and process Landsat and PALSAR/PALSAR-2 remote sensing big data stored in GEE on a high-performance cluster server. The data in PALSAR and PALSAR-2 active microwave sensors are used to extract forest information. Its L-band frequency not only has stronger penetration but also is not affected by the alternation of day and night, cloud shadow, and bad weather. It can realize the Earth observation day and night throughout the day, so as to obtain the information of forest vertical structure. Therefore, it is widely used in forest information extraction and other fields. As shown in Table 1, PALSAR and PALSAR-2 sensors have three different observation modes, so they can observe a wider ground width than ordinary radar sensors [7].

3.2. Index Parameter Processing Method. In this study, the annual mosaic data of 2007–2010 and 2015–2021 with a spatial resolution of 25 m from PALSAR-2 sensors covering all the three northern regions are selected. Among them, the annual mosaic data from 2007 to 2010 is from the PALSAR sensor, and the annual mosaic data from 2015 to 2021 is from the PALSAR-2 sensor. Polarization signals based on radar remote sensing image data can be transmitted and received in horizontal and vertical dimensions through L-band, including the HH polarization band in which microwave energy is transmitted and received by radar antenna in the horizontal dimension, and the HV polarization band in which microwave energy is transmitted by radar antenna in the horizontal dimension and received in vertical dimension [8, 9]. In order to reduce the influence of terrain on Earth observation and the distortion of geometric angle, this study uses the 90 m Space Shuttle Radar terrain mission (SRTM) digital elevation model (DEM) to carry out slope correction and orthogonal correction on the backscattering coefficients of HH and HV polarization bands, with a geometric accuracy of about 12 M. The resulting PALSAR/PALSAR-2 mosaic data set includes HH and HV polarized backscatter data, local incident angle and mask information. The size of a pixel is about 25 m × 25 m.

TABLE 1: Data specification of PALSAR/PALSAR-2 sensor.

Observation mode	High resolution		Scanning synthetic aperture	Polarization
Center frequency			1270 MHz (L-band)	
Chirp bandwidth	28 MHz	14 MHz	14 MHz, 28 MHz	14 MHz
Polarization mode	HH/VV	HH + HV/VV + VH	HH/VV	HH + HV + VH + VV
Incident angle	8 to 60°	8 to 60°	18 to 43°	8 to 30°
Spatial resolution	7–44 m	14–88 m	100 m (Repeat scan)	24–89 m
Detection width	40–70 km	40–70 km	250–350 km	20–65 km
Bit length	5 bits	5 bits	5 bits	3/5 bits
Data rate	240 Mps	240 Mps	120 Mps, 240 Mps	240 Mps
Working mode	Side view 34.3°		Side view 34.1°	Side angle of view 21.5°

As shown in formula (1), based on the GEE cloud computing platform, for the corrected mosaic data of pallar/pallar-2 in the three north project area, the HH and HV polarization band amplitude data stored in 16 bit digital (DN) form are converted into backscattering coefficients in decibels (DB) pixel by pixel using the calibration coefficient from JAXA:

$$\gamma^0 (dB) = 10 \log_{10} DN^2 + CF. \quad (1)$$

Here, γ^0 (DB) is the converted backscattering coefficient in dB, DN is the amplitude data of HH or HV polarization bands stored in 16 bit digital form, and CF is the absolute calibration coefficient from JAXA, with a value of -83 . In addition, because the ratio value and difference value of HH and HV polarization bands are of great significance for extracting land use information, this study calculates the difference value difference (HH-HV) and ratio CHH/HV of the converted HH and HV band backscattering coefficients pixel by pixel through GEE cloud computing platform [10].

$$\text{Difference} = HH - HV, \quad (2)$$

$$\text{Ratio} = \frac{HH}{HV}. \quad (3)$$

In formulas (2) and (3), HH represents the backscattering coefficient of the transmission and reception of the radar antenna in the horizontal dimension, and HV represents the backscattering coefficient of the transmission and reception of the radar antenna in the horizontal dimension and the vertical dimension. After pixel by pixel calculation and processing of the PAL/pal-2 mosaic data in the three northern regions on the GEE cloud computing platform, we obtained the PAL/pal-2 data with four bands of HH, HV, difference value, and ratio [11, 12].

3.3. Decision Tree Classification Algorithm to Extract Forest Information. By calculating the frequency distributions of HH, HV, difference value (HH-HV) and ratio value (HH/HV) of four typical land cover types (forest, water, farmland, and urban land) in the three northern regions, the threshold value of extracting forest distribution information using PALSAR/PALSAR-2 radar remote sensing data is determined. First of all, we collected and visually interpreted 1722 sample points in the three north region using Google Earth's ultrahigh resolution images as a training sample set, called regions of interest (ROI),

including 469 forest samples, 432 farmland samples, 429 urban land samples, and 392 water samples. Then, the obtained samples are imported into assets of the GEE cloud computing platform in the form of ShapeFile for analysis, and the annual mosaic data set of PALSAR/PALSAR-2 is imported by using the data resource library in Gee, and the backscatter coefficient is converted pixel by pixel. Then, the HH, HV, HH-HV difference values, and HH/HV ratio values are calculated and used as four bands. Next, the values of the four bands calculated are assigned to the verification sample points one by one, and the results return 1722 sets of characteristic values of samples of interest. Finally, the frequency histograms of HH, HV, difference value (HH-HV), and ratio value (HH/HV) of four typical land cover types are generated, as shown in Figures 1(a)–1(d).

By comparing HH, HV, difference value (HH-HV), and ratio value (HH/HV) of common land cover types, the threshold value of extracting forest information based on PALSAR/PALSAR-2 data in this study is determined [13, 14]. In order to make the classification threshold more reliable, we exclude 2.5% pixels with maximum and minimum backscattering coefficients in HH, HV, HH-HV, and HH/HV and determine the decision tree classification threshold for extracting forest information according to the 95% confidence interval.

As shown in Figure 1, from the frequency histogram of HH and HV backscattering coefficients, it can be seen that the HH and HV values of water bodies are significantly lower than those of forests, farmland, and urban land, and there is almost no overlap with forests. This is because the reflecting surface of the water body is relatively smooth, and most of the backscattering can be reflected through specular reflection. Therefore, the water body can be easily distinguished from forests, urban land, and most farmland land cover types by HH or HV backscattering coefficient. Because the L-band of the PALSAR/PALSAR-2 sensor has strong penetration into the forest, the energy emitted when it is incident will interact with the trunk and branches of trees, resulting in a large amount of volume scattering. Therefore, the HH and HV backscattering coefficients of the forest are high, the difference value (HH-HV) is low, and the HV of the farmland is low. Therefore, the forest can be distinguished from most of the farmland. Due to the complex spatial structure of urban land, the backscatter coefficients of HH and HV are high, and HH-HV and HH/HV also overlap with forests to a certain extent. Therefore, only HH, HV, difference value (HH-HV) and ratio value (HH/HV) are not

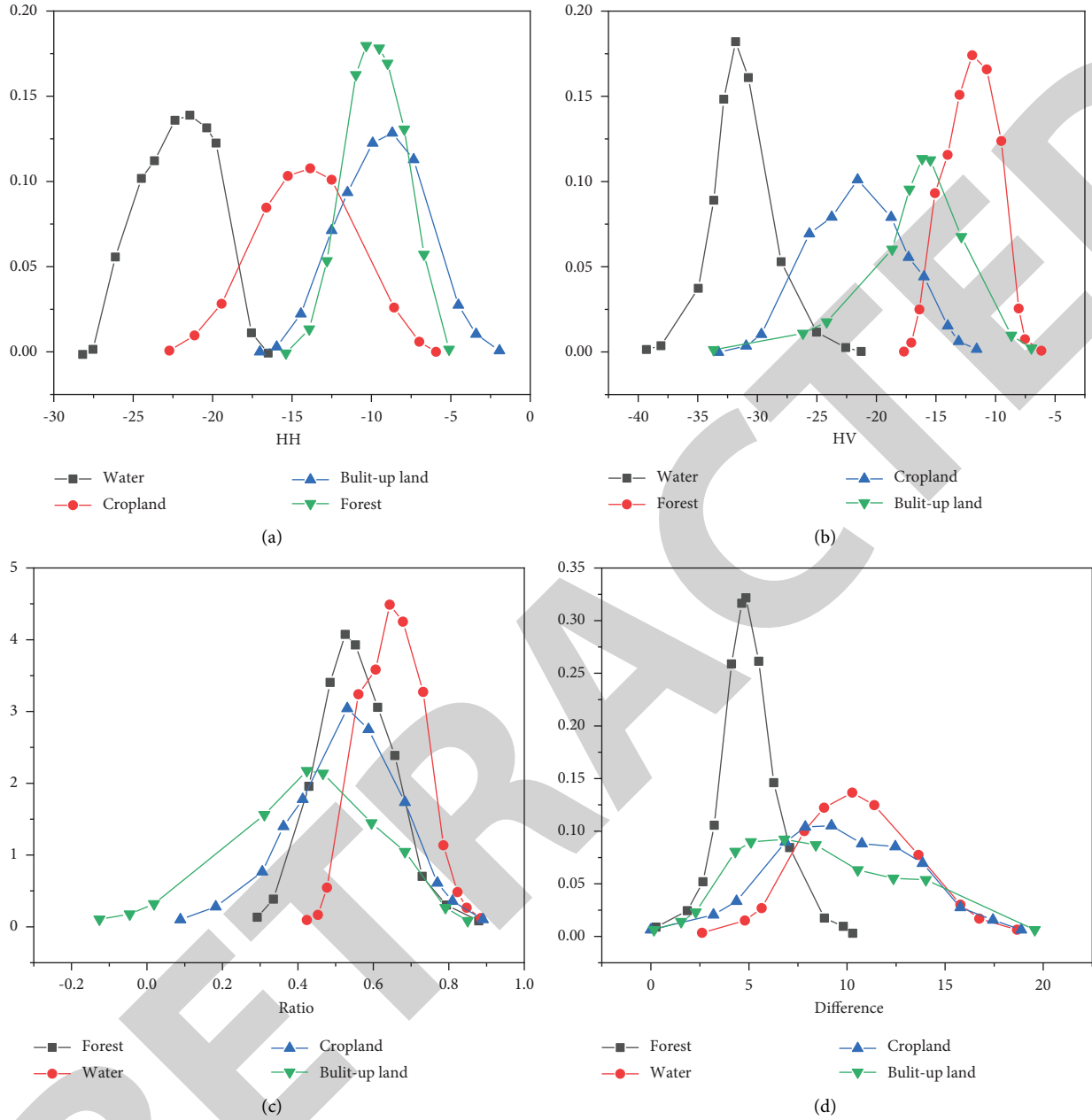


FIGURE 1: Frequency histogram of backscattering coefficients of four typical land cover types.

enough to completely distinguish the types of forest and urban land cover. Finally, forests can be distinguished from water bodies, most farmland, and some urban land cover types. Therefore, as shown in formula (4) and Figure 2, this study constructs a decision tree classification threshold for extracting forests:

$$(-16 < HV < -8) \text{ AND } (2 < \text{Difference} < 8) \text{ AND } (0.3 < \text{Ratio} < 0.85). \quad (4)$$

Based on the above-given analysis of HH, HV, difference value (HH-HV), and ratio value (HH/HV) backscatter coefficient thresholds of four common land cover types, it can be seen that there are obvious rules for extracting information of different land cover types based on PALSAR/PALSAR-2 radar

remote sensing data, which lays a good foundation for building decision trees for classification. Since the focus of this study is the distribution of forests, for the convenience of followup research, we combined the classification results into the forest and nonforest, and generated forest/nonforest distribution maps from 2007 to 2010 and 2015 to 2021, with a spatial resolution of 25 m.

4. Spatiotemporal Fusion Algorithm Based on SSIF Learning

4.1. Remote Sensing Image Degradation Model. A remote sensing image is a digital image obtained by means of photography, aerial scanning, or microwave imaging by

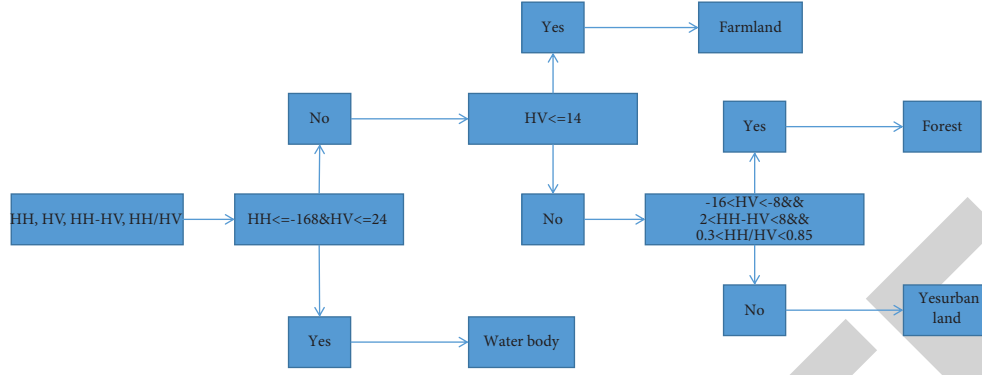


FIGURE 2: Decision tree classification based on PALSAR/PALSAR-2 radar remote sensing data.

sensors mounted on the remote sensing platform. Its basic unit is pixels, and the gray value of surface information is recorded by the DN value [15]. The superresolution reconstruction of remote sensing image is the process of using a certain conversion model to complement the information of low-resolution images, so as to reconstruct high-resolution images. Assuming that the low resolution image l meets a certain degradation relationship with its corresponding high resolution image h , the low resolution image l can be regarded as the result of a series of degradation processes such as down sampling and blurring of the high resolution image h . therefore, the relationship between the two can be established by a remote sensing image degradation model, as shown in relation:

$$L = SBH + n. \quad (5)$$

In (5), L is a low resolution image; H is a high-resolution image; S is the down sampling matrix; B represents fuzzy process; N is the noise generated when H is converted to L .

The super-resolution reconstruction of remote sensing image is the inverse process of the degradation model of remote sensing image, which needs to introduce a priori knowledge to solve. At present, sparse representation theory is one of the most effective methods.

4.1.1. Sparse Representation Principle of Remote Sensing Image. Sparse representation theory originated in the field of signal processing. Its main purpose is to use as few atoms as possible to approximate the original signal from a given super complete dictionary, so as to more refine the information contained in the signal. In recent years, the sparse representation method is mainly used in remote sensing image processing. Relevant research shows that most remote sensing images can be sparse representation because of their relevance and redundancy [16]. The sparse representation model of remote sensing image is shown in the following equation:

$$\min \|a\|_0 \text{ s.t. } X = D \times a. \quad (6)$$

According to the above-given sparse representation theory, the sparse representation of image x can be obtained from a suitable over a complete dictionary, and in

the process of sparse representation, image x should be approximately represented with as few atoms as possible. It can also be understood that the sparse representation coefficient a is mostly equal to or close to 0, that is, assuming that a contains f nonzero terms, which meets the condition $f < n < k$. In order to measure the number of nonzero elements in the sparse coefficient a , l is usually used. Norm is expressed. Theoretically, l is solved in equation. Norm is the simplest sparse measure, but l is used. Norm regularization to achieve sparseness is a NP hard (non determined polynomial hard) problem, so equation (7) is often transformed into the following equation:

$$\min \|a\|_1 \text{ s.t. } \|X - Da\|_2^2 < \varepsilon. \quad (7)$$

By introducing a Lagrange multiplier, (7) can be further optimized into an unconstrained optimization problem:

$$\min \|X - Da\|_2^2 + \gamma \|a\|_1. \quad (8)$$

In (8), γ is a Lagrange multiplier, which can mediate the fidelity of the image and the sparsity of the sparse coefficient. The more 0 elements of a , the better the sparsity.

It can be concluded that there are two optimization variables in sparse representation, dictionary D and sparse coefficient a . The solution to this optimization problem is generally to fix one optimization variable first, then optimize the other variable, and finally proceed alternately. At present, the sparse coefficient solving algorithm can be divided into a greedy algorithm and convex optimization based algorithm. According to the relevant literature, the greedy algorithm has better performance than the convex optimization algorithm, so greedy algorithm is often used for sparse coding [17].

4.1.2. Dictionary Construction. K-SVD (k-singular value decomposition) dictionary learning algorithm is a derivative algorithm of k-means. According to the principle of minimum error, the algorithm decomposes the error term E and through SVD, selects the decomposition term with minimum error to update the dictionary and sparse representation coefficient, and finally obtains the optimal solution through the process of alternating iteration. The specific

steps of the K-SVD dictionary learning algorithm are as follows:

- (1) Randomly select k column vectors from the original sample, initialize dictionary D , normalize all column vectors, and set the number of iterations r
- (2) Fix the dictionary D , and use the OMP algorithm to solve the sparse representation coefficient of each training sample

$$a_i = \underset{D_{r-1,x}}{\operatorname{argmin}} \|X - D_{r-1} \alpha_i\|_F^2, s.t. \|X\| \leq F, i = 1, 2, \dots, k. \quad (9)$$

- (3) Update the dictionary D obtained in step (2) atom by atom and then use formula (9) to perform singular value decomposition (SVD) on the error matrix to extract the most important features in the matrix

$$E_k = U \Delta V^T. \quad (10)$$

- (4) The dictionary D and the sparse representation coefficient are iteratively updated until the final dictionary is obtained

The main task of super-resolution reconstruction based on the principle of sparse representation is to use the method based on sparse representation to obtain the relationship between images with different spatial resolutions, so as to complete the prediction and construction of high-resolution images at another time. Therefore, it is necessary to establish high-resolution and low-resolution dictionaries under the feature blocks of high-resolution and low-resolution images (relatively high and low), and establish the relationship between them. In the process of dictionary construction, if the image blocks corresponding to the high and low resolution images are sampled at the same time, the resulting high and low resolution dictionary also meets the remote sensing image degradation model, and the expression (11) can be obtained:

$$DL = SBD_H + n. \quad (11)$$

By introducing (11) into the sparse representation model of remote sensing images, we can get the following equation:

$$X_L = SBD_H \times \alpha + n_1. \quad (12)$$

(12) is a low resolution image, which can be obtained in the same way:

$$X_H = SBD_H \times \alpha + n_2. \quad (13)$$

According to the observation formulas (12) and (13), images X_c and X_H with different spatial resolutions have the same sparse representation coefficient a in the process of super-resolution reconstruction based on sparse representation, so we can use the same sparse representation coefficient a as a bridge from low resolution remote sensing image to high-resolution remote sensing image reconstruction to realize super-resolution reconstruction [18, 19].

4.2. Sparse Spatio-Temporal Fusion Algorithm Based on Single Pair of Images. The image data used by SSIF algorithm is consistent with that of starm algorithm. Landsat Image at time t_0 , MODIS Image and MODIS Image at time TK are the data sources. Considering the large spatial resolution difference between Landsat Image (high resolution image) and MODIS Image (low resolution image), if the two images are directly fused, it will lead to a huge prediction error. The way to solve this problem is to improve the spatial resolution of the MODIS Image and then integrate the MODIS Image with the original Landsat Image. Through the idea of the remote sensing image degradation model, it is assumed that MODIS Image (corresponding band) is the result of Landsat image degradation through a series of down sampling, blurring, and other processes. Considering the upper limit of super-resolution reconstruction, the spatial resolution of MODIS Image at time t_0 and TK is improved and reconstructed into a transition image with spatial resolution between MODIS and Landsat Image by using the super-resolution reconstruction principle based on sparse representation, and then the information of transition image and Landsat Image is collected by using high pass filtering, so as to generate Landsat Image at prediction time TK [20].

Suppose that Landsat Image and MODIS Image at time t_0 are I_0 and M_0 respectively; The Modsi image at TK time is M_K ; The generated transition images are t_0 and TK respectively. The super-resolution reconstruction of the MODIS image includes two steps, that is, using the known M_0 and I_0 to establish the dictionary training set, and then complete the construction of the transition image. The specific steps are as follows.

4.2.1. High and Low Resolution Dictionary Training.

Obtain the high-resolution features (feature blocks extracted from the difference image of $I_0 - m_0$) and low-resolution features (feature blocks of M_0) of MODIS Image and Landsat Image at time T_0 , list the two feature blocks, input them into the K-SVD dictionary training model, and obtain the low-resolution Dictionary:

$$\{D_L, \Lambda^*\} = \underset{D_L, \Lambda}{\operatorname{argmin}} \left\{ \|x - D_L \Lambda^*\|_F^2 \right\} s.t. \forall_i, \|\alpha_i\|_0 \leq K_0. \quad (14)$$

In equation (14), X is the sampling matrix after the serialization of MODIS gradient image feature blocks, n is the sparse coefficient of each column corresponding to x , where the given sparse coefficient n^* is a nonsingular matrix, and DL is the obtained low resolution dictionary. Since the high-resolution dictionary and the low-resolution dictionary have the same sparse representation coefficient, the corresponding high-resolution dictionary training can be carried out with the following equation:

$$D_G = \underset{D_G}{\operatorname{argmin}} \left\{ \|Y - D_L \Lambda^*\|_F^2 \right\}. \quad (15)$$

Equation (15) needs to be solved by using the generalized inverse matrix of the following equation:

TABLE 2: Forestry land classification standards.

First level	Woodland	Nonforest land
	Forested land	Cultivated land
	Open woodland	Pasture
	Shrub land	Waters
Second level	Immature forest land	Unused land
	Nursery land	Land used for building
	Nonstanding forest land	
	Suitable forest land	
	Land for forestry auxiliary production	

$$D_G = Y(\Lambda^*)^+ = Y\Lambda^{*t}(\Lambda * \Lambda^{*T})^{-1}. \quad (16)$$

4.2.2. Transition Image Construction. After the dictionary training, it is used to construct the transition image. The transition prediction image T_K is generated based on the MODIS Image M_K at T_K time. First, the gradient feature image block of M_K is sampled in the same way as in dictionary training, assuming that the i th column of X_K is X_K ; the OMP algorithm in greedy algorithm can solve the sparse coefficient matrix of its corresponding low resolution dictionary DL, so as to obtain the one-to-one corresponding sparse coefficient matrix A_I in each low resolution image block. Based on the principle of sparse representation, using the sparse coefficient A_K of MODIS Image (M_O) at T_K time and the high-resolution dictionary aligned at t_0 time, the difference image Y_K of high-resolution and low-resolution images at T_K time can be obtained, and finally the transition prediction image T_K can be obtained. Similarly, the transition prediction image t_0 at t_0 time can be obtained.

4.2.3. High Pass Filtering. Since the difference image between Landsat Image and MODIS Image at t_0 time for dictionary training has a high degree of similarity in phenological changes and changes in surface coverage types, assuming that the transition prediction image t_0 and T_K obey the linear change relationship, the following linear transformation model can be obtained:

$$T_k = a \times T_0 + b, \quad (17)$$

$$L_k = a \times L_0 + b. \quad (18)$$

The solution formula of Landsat Image at T_K time can be derived from (17) and (18):

$$L_k = T_k + \frac{T_k}{T_0} \times (L_0 - T_0). \quad (19)$$

4.3. Forest Resource Change Monitoring with High Spatial Resolution Fusion Images

4.3.1. Establishment of Classification System. Before the classification of remote sensing images, it is necessary to establish a classification system according to the information contained in the images of the study area and the purpose of

the study. According to the classification of land use status and the main technical provisions for the planning, design, and investigation of forest resources, the classification standards of forest land are shown in Table 2. The classification system is formulated according to the macro situation of national forest land and forest resources.

Forest resources are the general name of forest land and the forest organisms it grows, and forest land resources are the main body of forest resources. The forest resources referred to in this paper are aimed at forest land resources. Considering that there are still some differences between the high spatial resolution fusion image and the actual GF-2 MSS image, and the insufficient characteristics of the 4 m spatial resolution GF-2 MSS image, the recognition ability of forest land types is limited, so it is difficult to make a fine distinction. Therefore, the shrub land, sparse forest land, and forest land in the study area are collectively referred to as forest land in this paper. Nonforest land is divided into construction land, cultivated land, water body, and unused land, among which residential areas, urban roads, and other construction land are collectively referred to as construction land.

4.3.2. Classification Feature Information Extraction. The purpose of feature information extraction is to represent all the information of the image with the least and most significant features, and improve the discrimination between different ground objects while reducing the dimension, so as to distinguish the ground types to the greatest extent.

(1) *Spectral Characteristics.* In this paper, four bands of red, green, blue, and near infrared of GF-2 MSS image on January 16, 2021, and high spatial resolution fusion image on October 9, 2021, generated by SIF algorithm are selected as spectral characteristic variables to participate in classification.

(2) *Texture Features.* In this paper, the co-occurrence measures texture analysis tool in ENVI 5.3 software is used to calculate the gray level co-occurrence matrix of the first principal component and the second principal component after the principal component transformation. The 7×7 base window is used respectively, and the gray quantization level is set to 64 to extract the texture features. This paper selects three texture features: contrast, variance and mean as auxiliary feature information of image classification.

TABLE 3: Statistics of characteristic parameters.

Characteristic information	Characteristic parameter	Number
Spectral information	B, G, R, NIR, NDVI, RVI, NDWI	7
Texture information	PC1 mean, PC1 variance, PC1 contrast, PC2 mean, PC2 variance PC2 contrast	6

(3) *Vegetation Index*. Vegetation index is an effective index to measure the status of surface vegetation. This kind of index combines images and different bands (often referred to as visible light, red band, and near-infrared band) in a linear or nonlinear way to enhance the vegetation characteristic information. Among them, normalized vegetation index and ratio vegetation index are two commonly used vegetation indexes. Both of them make use of the characteristics of vegetation with strong absorption ability in the red light band and strong reflection ability in the near-infrared band to calculate the band of the near-infrared band and the red light band [21].

Normalized vegetation index (NDVI) is to normalize the red light band and near-infrared band to enhance vegetation information. The index is between $[-1, 1]$. When the value of NDVI is greater than 0, it indicates that there is vegetation cover on the ground; When NDVI value is less than or equal to 0, it means that the surface is not covered by vegetation or the surface is covered by nonforest land such as clouds, water or snow. In short, the NDVI index can enhance the differentiation between forest land and nonforest land.

The ratio vegetation index (RVI) is the division of near-infrared band and visible red band to enhance vegetation information, which is consistent with the concept of normalized vegetation index. Generally speaking, the RVI value of a healthy green vegetation covered area is much greater than 1, and when the RVI value is near 1, it means that there is no vegetation cover or the vegetation coverage is low in this area.

(4) *Normalized Water Index*. There is a large area of water in this study area. In order to increase the contrast between water and forest land and between water and other land types, normalized difference water index (NDWI) is often used. Because the spectral corresponding curve of the water body decreases from the green light band to the infrared band, while the spectral corresponding curve of vegetation is on the contrary, this index can not only highlight the characteristic information of the water body but also weaken the characteristic information of soil and vegetation. The classification feature parameters extracted in this paper are shown in Table 3. The study area is classified according to the following feature parameter set.

4.3.3. *Image Adaptive Appendix Wavelet Denoising*. Signals are usually divided into high-frequency and low-frequency signals, in which low-frequency signals contain useful information in a large range and large scale of the image, while high-frequency signals reflect the edge and detail information of the image, and it is also the location of noise concentration. According to the relevant research, wavelet denoising benefits from its multiresolution, low

entropy, decorrelation, flexibility of base selection, and other characteristics and can effectively remove the noise in remote sensing images. As shown in Figure 3, taking the noisy signal S_n as the input, first complete the wavelet decomposition of the signal, that is, the low-frequency and high-frequency signals in the input signal S_n are separated, and then the decomposed high-frequency coefficients are threshold quantized according to the relevant theory of the selected threshold method. Finally, the processed wavelet coefficients are used for the inverse wavelet transform to complete the signal reconstruction to obtain the denoised signal.

4.4. Comparison and Analysis of Simulation Results

4.4.1. *Accuracy Analysis*. The accuracy evaluation of classification results is a measure of the reliability of classification results, and the confusion matrix is the standard form of accuracy evaluation. The confusion matrix is a comparison matrix obtained by calculating the pixels of the classification data set and the verification data set, which is used to verify whether the classification of the classification data set and the verification data set at the corresponding position are consistent, so as to obtain the classification accuracy of the image. In this paper, the GF-2MSS images on January 16, 2021, and October 9, 2021, are randomly sampled by visual interpretation method. As the real samples, the validation sample data set is constructed. Finally, the classification accuracy of the images is evaluated by the calculation results of the confusion matrix.

Table 4 shows the classification confusion matrix of GF-2 MSS image classification feature set on January 16, 2021, in which the overall classification accuracy of random forest classification is 84.42%, kappa coefficient is 0.81, the overall classification accuracy of adaptive threshold wavelet denoising + random forest classification is 89.92%, kappa coefficient is 0.87, and the overall accuracy of adaptive threshold wavelet denoising + random forest classification method is 5.5% higher than that of random forest classification method. Kappa coefficient increased by 0.06.

According to Table 5, the user accuracy and mapping accuracy of various ground objects in random forest classification are higher than 75%, of which the classification accuracy of water and forest land is higher than 88%, but the classification accuracy of cultivated land and unused land is low, because some cultivated land has low vegetation coverage, and its spectral response curve is similar to that of unused land, which makes it difficult to distinguish them. The areas with high vegetation coverage are close to the reflectance of forest land, which is easy to produce misclassification. In addition, the generation process of remote sensing images will be affected by uncertain factors, resulting

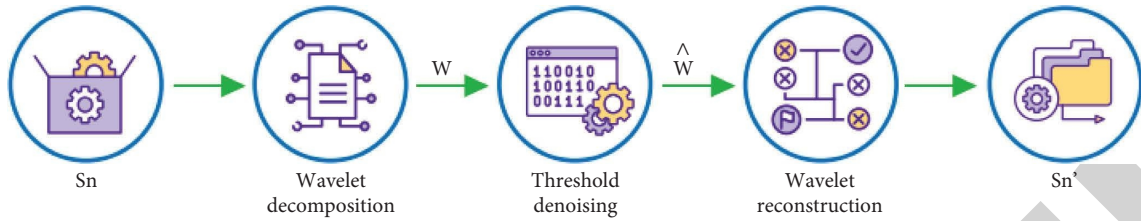


FIGURE 3: Wavelet threshold denoising process.

TABLE 4: Confusion matrix of GF-2MSS image classification feature set on January 16, 2021.

Category	Random forest					Adaptive threshold wavelet denoising + random forest				
	Cultivated land	Land used for building	Woodland	Unused land	Water body	Cultivated land	Land used for building	Woodland	Unused land	Water body
Cultivated land	321	23	30	44	6	350	9	25	32	5
Land used for building	19	340	9	38	5	20	400	7	30	1
Woodland	34	15	400	12	4	6	10	410	8	2
Unused land	23	41	6	322	10	25	20	8	350	4
Water body	3	1	5	6	400	1	2	3	3	420
Overall classification accuracy = 84.42% kappa coefficient = 0.81					Overall classification accuracy = 89.42% kappa coefficient = 0.87					

TABLE 5: Classification accuracy statistics of GF-2 MSS image classification feature set on January 16, 2021.

Category	Random forest				Adaptive threshold wavelet denoising + random forest			
	Cartographic accuracy (%)	User accuracy (%)	Misclassification error (%)	Leakage error (%)	Cartographic accuracy (%)	User accuracy (%)	Misclassification error (%)	Leakage error (%)
Cultivated land	80.25	75.71	24.29	19.75	87.50	83.73	16.27	12.50
Land used for building	80.95	83.33	16.67	19.05	90.48	86.76	13.24	9.52
Woodland	88.89	86.02	13.98	11.11	91.11	94.25	5.75	8.89
Unused land	76.30	80.10	19.90	23.70	82.94	86.42	13.58	17.06
Water body	94.91	96.47	3.53	5.09	97.22	98.13	1.87	2.78

TABLE 6: Confusion matrix of classification feature set of high spatial resolution fusion image on October 9, 2021.

Category	Random forest					Adaptive threshold wavelet denoising + random forest				
	Cultivated land	Land used for building	Woodland	Unused land	Water body	Cultivated land	Land used for building	Woodland	Unused land	Water body
Cultivated land	250	23	45	32	25	250	25	15	29	10
Land used for building	12	240	15	14	0	10	400	21	18	2
Woodland	50	10	400	6	5	40	8	410	6	4
Unused land	24	40	2	322	10	28	36	2	350	4
Water body	4	10	3	5	300	2	1	1	4	300
Overall classification accuracy = 81.4% Kappa coefficient = 0.85					Overall classification accuracy = 84.83% Kappa coefficient = 0.89					

in “abnormal points,” so there is also the possibility of misclassification. The classification method combining adaptive threshold wavelet denoising and random forest has improved the mapping accuracy and user accuracy of cultivated land, unused land, construction land, forest land, and water body. Among them, the mapping accuracy of

cultivated land, construction land, forest land, and unused land has been improved by 7.25%, 9.52%, 2.22%, and 6.64%, respectively, and the user accuracy has been improved by 8.02%, 3.42%, 8.23%, and 6.32%. Due to the large spectral difference between water body and other ground objects and high discrimination, the accuracy improvement is small.

TABLE 7: Classification accuracy statistics of high spatial resolution fusion image classification feature set on October 9, 2021.

Category	Random forest				Adaptive threshold wavelet denoising + random forest			
	Cartographic accuracy (%)	User accuracy (%)	Misclassification error (%)	Leakage error (%)	Cartographic accuracy (%)	User accuracy (%)	Misclassification error (%)	Leakage error (%)
Cultivated land	73.53	64.94	35.06	26.47	76.47	76.47	23.53	23.53
Land used for building	70.00	83.67	16.33	30.00	76.67	81.85	18.15	23.33
Woodland	85.56	84.43	15.57	14.44	91.11	87.61	12.39	8.89
Unused land	82.86	79.23	20.77	17.14	83.71	80.72	19.28	16.29
Water body	87.50	92.72	7.28	12.50	93.75	97.4	2.60	6.25

TABLE 8: Statistical table of forest land changes in the study area from January 2021 to October 2021.

Type of ground feature change	Change area (m ²)	Rate of change (%)
Forest land—unused land	468496	0.97
Woodland cultivated land	1339424	2.78
Woodland water body	19632	0.04
Forest land—construction land	306640	0.64
Unused land forest land	525520	1.09
Cultivated land forest land	860120	1.79
Water forest	66768	0.14
Construction land—forest land	24784	0.05

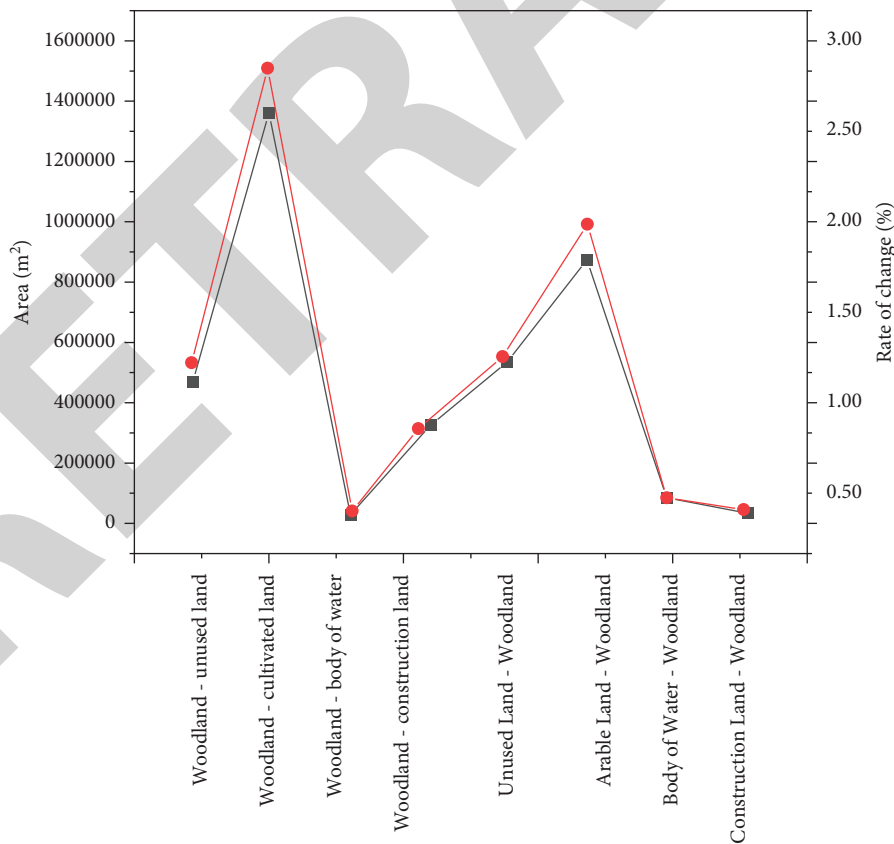


FIGURE 4: Change rate of forest resources.

Table 6 shows the classification confusion matrix of the classification feature set of high spatial resolution fusion images on October 9, 2021, in which the overall accuracy of

random forest classification is 81.4%, and the kappa coefficient is 0.85. The overall classification accuracy of adaptive threshold wavelet denoising + random forest classification is

84.83%, and the kappa coefficient is 0.89. Compared with the random forest classification method, the overall accuracy of the adaptive threshold wavelet denoising + random forest classification method is improved by 3.43%, and the kappa coefficient is improved by 0.04. Combined with Table 7, the adaptive closed-value wavelet denoising + random forest classification method has better improved the mapping accuracy and user accuracy of cultivated land, forest land, unused land, and water, but the user accuracy of construction land has been reduced by 1.81%.

4.4.2. Monitoring Results of Forest Resource Change. In order to obtain the increase and decrease of the number and distribution of forest land after mutual transformation with other land types, and then obtain the change of forest resources in this region, nonforest land is not the research object of this paper, so this paper does not discuss the change between nonforest land. Using the postclassification comparison method to monitor the changes of the two images, the thematic map of forest resource cover change in the study area can be obtained. According to the thematic map of forest resource cover change in the study area, the spatial location information and the direction information of forest land change in the study area can be obtained. The areas of forest land change in the study area are distributed on both sides of the river, and there are many changes with cultivated land and unused land. In order to further obtain the change area of forest land and the change rate between forest land and nonforest land, use ENVI software to calculate the transfer matrix based on the change results of forest resource coverage, and the statistical results are shown in Table 8 and Figure 4.

According to Table 8 and Figure 4, from January to October 2021, the area of forest land turned into unused land was 468496 m², with a change rate of 0.94%, the area of forest land turned into cultivated land was 1339424 m², with a change rate of 2.78%, the area of forest land turned into the water was 19632 m², with a change rate of 0.04%, the area of forest land turned into construction land was 306640 m², with a change rate of 0.64%, and the total reduced area of forest land was 2134192 m². The area from unused land to forest land is 525520 m², with a change rate of 1.09%, the area from cultivated land to forest land is 860120 m², with a change rate of 1.79%, the area from water to forest land is 66768 m², with a change rate of 0.14%, the area from construction land to forest land is 24784 m², with a change rate of 0.050/a, and the total increased area of forest land is 1477192 m². In general, from January to October 2021, the forest area in the study area decreased by 657000 m², that is, the forest coverage decreased by 1.36%, and the forest resources decreased.

5. Conclusion

The process and accuracy evaluation of optical remote sensing data Landsat and radar remote sensing data PAL-SAR/PALSAR-2 fusion algorithm are introduced. Firstly, the nearest neighbor algorithm is used to adjust the spatial resolution of the forest information extracted from the two

remote sensing data sources, and then the appropriate fusion method is used to fuse to generate the final forest distribution map. Next, taking the interannual and intra-annual vegetation index time series generated from Google Earth ultrahigh resolution images and Landsat and MODIS images from collect Earth as reference data, the accuracy of the forest distribution map obtained by the fusion algorithm is evaluated based on the validation sample data set, and compared qualitatively and quantitatively with the four mainstream medium and high spatial resolution forest distribution maps in the market. The forest distribution map obtained in this study has the highest overall accuracy of 96% ± 1% and kappa coefficient of 0.66. In addition, this chapter analyzes the reasons why the classification results of the fusion algorithm are inconsistent with other forest distribution maps from the two aspects of forest definition and data source.

This paper introduces the basic principle of the SSIF algorithm, and then expounds on the principle of remote sensing image degradation model, sparse representation principle, dictionary construction principle, and super-resolution reconstruction principle based on sparse representation theory related to the SSIF algorithm. Finally, the spatio-temporal fusion framework of the Landsat 8 OLI image and GF-2 MSS image of the SSIF algorithm is established respectively, and the spatio-temporal fusion of the Landsat 8 OLI image and GF-2 MSS image is completed. Because continuous time series GF-2mss image data and Landsat 8 OLI image data are difficult to obtain, this paper only considers the spatio-temporal fusion of single temporal medium and high spatial resolution remote sensing images.

The classification method based on adaptive closed value wavelet denoising and random forest can better reduce the “salt and pepper noise” generated by image classification. This is because adaptive threshold wavelet denoising removes some high-frequency components in the image on the basis of retaining the original spectral information of the image, so as to suppress the “noise” in the image, reduce the uncertainty factors in the image classification process, and further improve the classification accuracy of the image. It shows that the classification method based on adaptive closed value wavelet denoising and the random forest is feasible.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the open fund project of Key Laboratory for Carbon Neutrality and Territory Optimization of Ministry of Natural Resources (Grant no. 2021CNT03001).

Research Article

Interpersonal Relationship, Knowledge Characteristic, and Knowledge Sharing Behavior of Online Community Members: A TAM Perspective

Wu Jiarui,¹ Zhang Xiaoli ,² and Su Jiafu ^{3,4,5}

¹Hotel Management Institute, Yunnan Tourism College, Kunming 650221, China

²Anhui Technical College of Mechanical and Electrical Engineering, Wuhu 241000, China

³Chongqing City Vocational College, Chongqing 402160, China

⁴International College, Krirk University, Bangkok 10220, Thailand

⁵National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, China

Correspondence should be addressed to Su Jiafu; jeff.su@cqu.edu.cn

Received 12 July 2022; Accepted 23 September 2022; Published 10 October 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Wu Jiarui et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of information science and technology, online communities are attracting an increasing number of participants, who can share information, create original content, and offer emotional support, thus communicating and spreading knowledge frequently within the community. To develop a model of influencing factors for the knowledge sharing behavior of online community members, this study employs the technology acceptance model (TAM) as a moderator variable based on the social exchange theory. In this study, the influencing factors model for knowledge sharing behavior of online community members was tested using PLS-SEM. The results show that knowledge sharing is motivated by trust and quality of knowledge; the interaction term of perceived usefulness and knowledge quality of the user has a significant negative correlation with the knowledge sharing behavior of online community users; perceived usefulness significantly positively moderates the correlation between knowledge tacitness and knowledge sharing behavior of users; perceived ease of use significantly positively moderates the relationship between knowledge quality and knowledge sharing behavior; perceived ease of use significantly negatively moderates the relationship between knowledge tacit and knowledge sharing behavior. In order to maximize the activity and stickiness of the online community platform, the platform must focus on maintaining and enhancing the platform's credibility and knowledge quality. On the other hand, the online community platform extols its professional utility and ease of operation, which are conducive to the generation of behavior that is conducive to knowledge sharing.

1. Introduction

Following the vigorous spread of Internet technology and Web2.0, the influence of the Internet on the life and work of people is growing with each passing day [1]. The online community based on social media has gradually turned into a new platform for people to share knowledge and publish different ideas, and it also could facilitate business activities. The social network has the potential to change the way people live and facilitate knowledge exchange among individuals. The site contains a number of

groups created by users with similar hobbies and interests who share their views, experiences, and insights on a variety of topics and seek out knowledge and information related to their interests [2]. For instance, the Facebook group is quite different from the traditional social networking for people. The platform facilitates communication between users with common interests, discusses, and shares information. As for another example, Zhihu is a respected and well-known Q&A platform in the Chinese Internet. It is committed to building a knowledge sharing network that will be accessible to everyone, allowing people

to easily share their knowledge, experience, and insights worldwide.

The online community is a product of rapid growth and has a huge influence in the Internet era. Lai and Chen [3] illustrate that an online community consists of a group of people with common interests who interact virtually one-on-one. Additionally, they can assist with the retention of customers, word-of-mouth advertising, product reviews, and customer insight. Online communities are critical for continuing commitment to interact, the relationship of trust, and meaningful interactions with partners, customers, employees, and suppliers [3]. Members could through these communities seek and share knowledge [4]. Xu and Liu [5] demonstrate that due to its characteristics of fairness, openness, cross-time and spatiality, it has the ability to integrate dispersed knowledge groups on a larger scale, even globally. As a result of the growth in online communities over the past few years, online communities have the potential to play a significant role in enhancing the exchange of knowledge.

Recently, scholars [6–9] have conducted some studies on user information sharing behavior in online communities. Gang [6] studied the influence of the motivation of professional virtual community members to participate in knowledge exchange on interactive behavior. Connelly et al. [7] revealed that people may not share knowledge because they are unaware of the needs of knowledge of others. Li et al. [8] detected that online community members' knowledge sharing intentions were significantly affected by altruism, relational trust, personal outcome expectations, and self-efficacy. Sun and Hong [9] showed that human interaction, human interaction with information and expectations of the outcome, expectations of the community, and user knowledge sharing behaviors were significantly positively correlated, and the interaction between people and systems was positively affected by user knowledge sharing behaviors.

The previous research studies bulkily stayed on the single level of human interaction, human-knowledge interaction, or human-system interaction; nevertheless, there is a lack of integrative study. At first, people as the subject of knowledge sharing are the bearers and initiators of knowledge sharing activities. It is likely that the subjective initiative of individuals will affect the judgment of objective phenomena. Such as, emotions have a profound effect on people's willingness to share information. Second, knowledge is composed of explicit and implicit, among which tacit knowledge accounts for a larger proportion than explicit knowledge. This is a state of being that exists in people's minds, which is difficult to describe adequately in language or in written materials, and is not easy to be comprehended by others. Finally, since information network technology is an important tool for knowledge sharing to be carried out smoothly; whereas, in different industries, the use of multiple systems is independent and incompatible with each other, which reduces the effectiveness of knowledge sharing. Therefore, by integrating human interaction, human-knowledge interaction, and human-system interaction, the analysis results are objective and scientific.

This study selects two dimensions of interpersonal relationships (reciprocity and trust) and knowledge characteristics (quality of knowledge and tacit knowledge) as research objects, according to the theory of technology acceptance, the introduction of perceived usefulness and perceived ease of use has a moderating effect on interpersonal relationships and knowledge sharing behaviors, as well as knowledge characteristics and knowledge sharing behaviors. This is of great significance for the online community to remain the willingness of users to participate, maintain the vitality of the community, and promote the active participation of community members in knowledge sharing.

Accordingly, the remainder of the study is organized as follows: Section 2 investigates the influences of interpersonal relationships, knowledge characteristics, and technology acceptance models on knowledge sharing behavior. In Section 3, an assessment of the impact the technology acceptance model has had on the behavior of knowledge sharing is made. Section 4 assesses how knowledge sharing behaviors are influenced by moderator variables. Section 5 concludes the study and discusses and summarizes the contributions and limitations.

2. Literature Review and Hypotheses Development

2.1. Interpersonal Relationships and Knowledge Sharing Behavior. It is imperative that interpersonal communication can be practiced in modern society. For one to truly appreciate others, he or she must understand their needs and desires. Primarily, a mutually beneficial relationship is one in which both parties expect some kind of gain, whether it is a material gain or an emotional one. The effectiveness of interpersonal interactions depends on how beneficial they are for both parties, as a relationship cannot last if one party is always benefited alone. Second, trust is the cornerstone of any relationship between two individuals. A harmonious interpersonal relationship cannot exist without trust as the basis of words and deeds. The term trust refers to the subjective sense of confidence and security that one party has toward the other party, which is often associated with the cooperative behavior of people and the willingness to act [10]. As variables in the social exchange theory, reciprocity and trust can be used to explain models of behavior involving knowledge sharing. Blau[11] elucidated that the social exchange theory has been employed to explain knowledge sharing behavior in terms of people's behavior, outcomes or interests, and the environment. Therefore, this study investigates the effect of reciprocity and trust, two interpersonal factors, on knowledge sharing behavior, based on the social theory as well as social exchange and social cognition theories [12].

2.1.1. Reciprocity and Knowledge Sharing. As an extrinsic motivation, reciprocity suggests that people will exhibit knowledge sharing behaviors with the aim of accumulating rewards. The concept of knowledge sharing in the early years

was defined by Polanyi [12] as a deliberate subjective behavior that can be reused by others through knowledge transfer. As per Okyere-Kwakye [13], individuals may not share initially, since they perceive the activity as a mere cost, unless they intend to share when a positive reward is expected. Using social exchange theory as a basis for reciprocity, it is suggested that individuals only engage in certain behaviors if they expect a positive outcome [11]. It has been found by Kelley and Thibaut [14] that participants in online communities share their individual knowledge when they perceive the behavior of other participants as being similar to their own. Among participants in a community of practice (CoP), Wasko and Faraj [15] demonstrated that reciprocal behaviors enhance knowledge sharing. Thus, the first hypothesis is proposed.

H1: knowledge sharing behavior is positively influenced by reciprocity.

2.1.2. Trust and Knowledge Sharing. It is believed that trust is an intrinsic motivation to act on behalf of people based on an appreciation for the benefits of their actions and is the focal point of all relationships within an organization [10, 16]. As Molm [17] asserted, individuals refrain from engaging in certain activities when they are uncertain of the future rewards associated with them. The result is that people act according to how much trust they have in the system, but they can only build trust in others if they are confident, and there will be no costs associated with transactions [17]. A trusting relationship between two people facilitates easy cooperation. Nevertheless, Molm [17] also noted that trust encourages members of an organization to share knowledge. There will be a trend towards higher cooperation and commitment as long as trust exists among individuals in an organization. Some literatures contend that trust is the most cost-effective technique for enhancing knowledge sharing within an organization [18]. According to this hypothesis, participants are more likely to share knowledge with a recipient if they perceive the recipient to be honest, trustworthy, and reliable. A higher level of trust will prevent the individual from thinking of any future negative things at the event and will encourage a willingness to share knowledge. Thus, the next hypothesis can be formulated.

H2: knowledge sharing behavior is positively influenced by the trust.

2.2. Knowledge Characteristics and Knowledge Sharing. On account of communities are often the basic social units that organize knowledge creation, the study of knowledge sharing in community settings becomes increasingly important [19, 20]. Furthermore, knowledge sharing demands collaboration on the part of both the seekers and contributors [21], as knowledge sharing can be facilitated while knowledge can be exchanged synchronously and asynchronously [22]. Yan and Jian [23] mentioned that as a commodity, knowledge has its natural attributes (use value) and its social attributes (value). The choice of knowledge commodities by consumers is not only solely dependent on their own perception of the quality of content but also upon

the source from which the knowledge was derived. A related study on the knowledge acquisition behavior of users shows that high-quality knowledge promotes users to use knowledge communities.

2.2.1. Knowledge Quality and Knowledge Sharing. Knowledge quality pertains to how well new knowledge satisfies the criteria of study in terms of productivity-enhancing properties, affecting social life, saving trial-and-error costs, and knowledge value-added [24]. The quality of knowledge referred to in this study refers to the higher knowledge quality exchanged and shared in the online community, and the greater the probability of members participating in sharing. The following assumptions are derived from this information:

H3: knowledge sharing behaviors are positively influenced by knowledge quality.

2.2.2. Knowledge Tacitness and Knowledge Sharing. Yan et al. [25] referred that the exchange of information in the online community not only contains a lot of explicit knowledge but also hides a wealth of tacit knowledge. Knowledge sharing is an exchange behavior. Users will be motivated to share knowledge if they perceive the benefits that sharing knowledge can bring to them, whether that is in the form of material reward or spiritual reward [26]. Jeremy [27] explained that it is generally believed that tacit knowledge (as distinct from the more general intangible investment) cannot be codified, imperceptible knowledge of how it is acquired through informal acceptance of learned behaviors and procedures. The knowledge that is tacit is internalized within an organization, exhibits characteristics which cannot be expressed in documents, exhibits a lower degree of fluidity than explicit knowledge, and is very hard to imitate [28]. Therefore, knowledge tacitness possesses the characteristics of a valuable resource.

While knowledge itself is valuable for generating competitive advantage and value within an organization, it is not sufficient to generate those benefits on its own. The value of knowledge cannot be realized until it is shared within the organization in order to generate a competitive advantage. In light of this, it is necessary to integrate and externalize tacitness knowledge in order to gradually convert it into coded and explicit knowledge [29] because the more explicit the knowledge, the easier it will be for organizations to share it. Therefore, the next hypothesis can be made.

H4: knowledge sharing behavior is positively influenced by tacit knowledge.

2.3. Moderating Role of Perceived Usefulness. The technology acceptance model (TAM) proposed by Davis believes that the user's perception of human-computer interaction can be divided into perceived usefulness and perceived ease of use [30]. The term perceived usefulness is defined here as "the degree to which an individual believes that using a specific system will hoist his or her performance on the job." This stems from the definition of the word useful: "capable of

being utilized to one's benefit." In organizational settings, people perform various rewards in order to reinforce their performance. Conversely, a system with a high perceived utility is considered to have an effective use-performance relationship by its users [30].

Recent studies have investigated the regulatory role of TAM factors. It has been demonstrated that a platform that is deemed useful and easy to use will boost influencers, thereby enhancing the satisfaction of members and increasing their behavioral intentions. Considering community participation and contribution in the context of online communities, social media are used by individuals to participate and contribute to online content for two reasons, namely, participation benefits and contribution incentives [31]. As a whole, online communities offer a number of benefits to their members. Participation benefits are structured around four basic needs, namely, needs of functional, needs of social, needs of psychological, and needs of hedonic [32]. First, users can benefit from these platforms on a functional level by joining online communities for specific activities [33]. In parallel, they will also be able to build trust and expand their online social network in the process of information exchange [34]. Thus, the following assumptions are proposed:

H5: perceived usefulness positively modulates reciprocity and knowledge sharing behaviors.

H6: perceived usefulness positively modulates trust and knowledge sharing behaviors.

H7: perceived usefulness positively modulates the quality of knowledge and knowledge sharing behaviors.

H8: perceived usefulness positively modulates tacit knowledge and knowledge sharing behaviors.

2.4. Moderating Role of Perceived Ease of Use. By contrast, in terms of perceived ease of use, it refers to the degree to which a user perceives a particular system as effortless. This stems from the definition of "easy" and "without difficulty or effort." In other words, when all else are equal, users are more likely to accept an app that is considered to be easier to use than other apps [30].

In contrast, contributions to online communities can be attributed to several types of factors. Like those who join online communities for the purpose of participation, members participate in the creation of content as a means of reward. There are five motivations which can be identified with regard to participation in online communities—instrumentality, effectiveness, quality control, acquisition of status, and expectations [31]. These incentives increase not only engagement but also increase contributions, sparking interactive activity on the online platform. Parra-López et al. [35] extended a factor model that assumes that users will bear the cost of effort, spend time learning the system, and suffer privacy losses, all of which may undermine their motivation to contribute in online communities [36–38]. Accordingly, the following assumptions are made:

H9: perceived ease of use positively modulates reciprocity and knowledge sharing behaviors.

H10: perceived ease of use positively modulates trust and knowledge sharing behaviors.

H11: perceived ease of use positively modulates quality of knowledge and knowledge sharing behaviors.

H12: perceived ease of use positively modulates tacit knowledge and knowledge sharing behaviors.

Therefore, this study is based on the social exchange theory, combined with knowledge characteristics, and introduces a technology acceptance model as a moderator variable to develop a model of the factors that influence knowledge sharing behaviors in network communities, as shown in Figure 1.

3. Methodology

3.1. Measurement Development. There were several variables and questionnaires used in this study that were adapted from mature scales used in previous studies, and some of them were designed according to the study object and study requirements set forth in the study. There are three parts to the questionnaire: the first part is the characteristics of the demographic and the duration of using the online community; the second part is the measurement of interpersonal relationships, knowledge characteristics, and the moderator variables of the TAM model; the third part is the measurement of self-construction. All items proposed in this article were designed using a five-level Likert scale (5 means strongly agree and 1 means strongly disagree).

3.2. Data Collection. After completing the questionnaire design, the experts evaluated the questionnaire, reviewed the necessity and usefulness of the questionnaire items for measuring variables, confirmed that all factors were related to the corresponding variables, and assessed the correctness and appropriateness of the word and content of the questionnaire. Some wording and presentation of the questionnaire have been adjusted based on their feedback.

The questionnaire survey was conducted online, processed through the Questionnaire Star platform, and collected data by WeChat forwarding and snowballing. The collection time was from May 4th to 12th, 2022, and a total of 336 valid questionnaires have been recovered.

The features of the demographic profile of the questionnaire sample and the duration of using the online community are shown in Table 1. In terms of gender composition, males and females accounted for 43.75% and 56.25% of the respondents respectively; most of the users were between 18 and 28 years old, and most of them used online communities for 4–6 years, indicating that the respondents have extensive experience in using the online community.

3.3. Data Analysis

3.3.1. Reliability. As a means of analyzing the reliability and validity of the collected effective data, the reliability and validity of the survey samples were evaluated. In reliability testing, the consistency, stability, and dependability of test

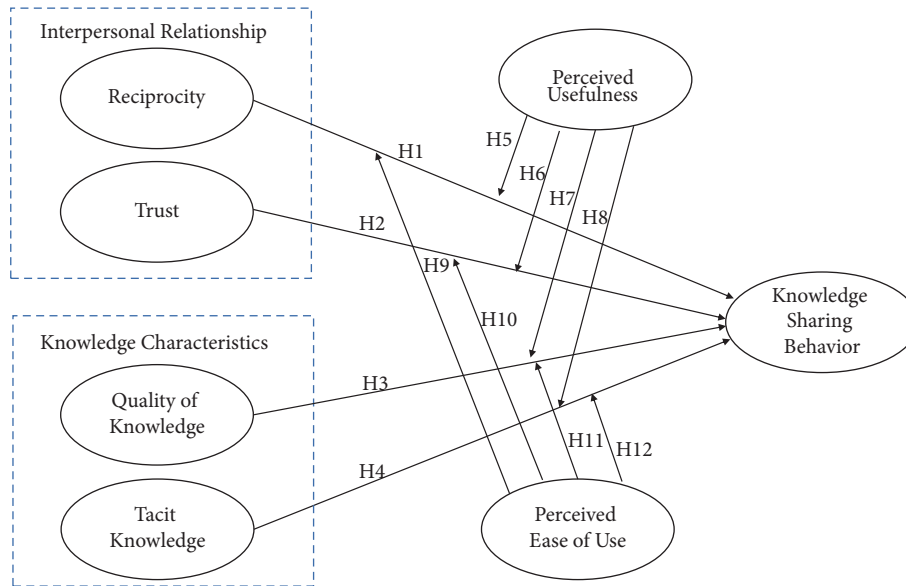


FIGURE 1: Research model.

TABLE 1: Demographic characteristics of the sample and duration of using online.

Variable	Index	WeChat (N = 336)
Gender	Male	43.75%
	Female	56.25%
Age	<18	0.89%
	18–28	67.56%
	29–40	27.38%
	41–48	2.98%
	49–55	0.30%
	55>	0.89%
Occupation	Public officials (public officials of state organs, civil servants, and other state units)	11.31%
	Professional and technical personnel (local professional and technical personnel such as doctors, lawyers, and drivers)	11.01%
	Private enterprises (entrepreneurs, individual, industrial, and commercial households)	11.90%
	Others	65.77%
Duration of using the online community (year)	<1	0.89%
	1–3	16.07%
	4–6	79.46%
	7–10	2.68%
	>10	0.89%

results and data were examined. This study uses SmartPLS3.0 to test the internal consistency coefficient (Cronbach’s α) and composite reliability of each study variable. Table 2 presents the results of the test. The consistency coefficient and composite reliability value of each variable in this study are above the critical value of 0.7 [39], which satisfy the reliability requirements, indicating that the measurement model has high internal consistency and reliability.

3.3.2. *Validity*. Validity includes content validity and construct validity. It is important to note that some of the questions in the questionnaire were adapted from existing

literature, while others were designed in conjunction with the characteristics of the study object, providing validity to their content. There are two primary measures of construct validity, namely, convergent validity and discriminant validity. Convergent validity reflects whether the measurement items of a variable are highly correlated. The main measurement indicators include factor loading and average variance extraction value. It is generally believed that if the factor loading is greater than 0.7 and the AVE is above 0.5, the scale is considered to have high convergent validity. Table 2 elucidates that the standardized loadings of items were mostly higher than 0.7 and the AVE for every construct was greater than 0.5, indicating that the model has good convergent validity.

TABLE 2: Scale properties.

Factor	Item	Standardized loading	Cronbach's alpha	Composite reliability (CR)	Average variance extracted (AVE)
KQ	KQ1	0.818	0.924	0.943	0.768
	KQ2	0.827			
	KQ3	0.929			
	KQ4	0.925			
	KQ5	0.878			
KSB	KSB1	0.836	0.830	0.898	0.747
	KSB2	0.890			
	KSB3	0.865			
PEOU	PEOU1	0.924	0.880	0.926	0.807
	PEOU2	0.925			
	PEOU3	0.844			
PU	PU1	0.930	0.918	0.948	0.86
	PU2	0.926			
	PU3	0.925			
RE	RE1	0.897	0.842	0.905	0.762
	RE2	0.910			
	RE3	0.808			
TK	TK1	0.872	0.772	0.866	0.684
	TK2	0.751			
	TK3	0.853			
TR	TR1	0.868	0.904	0.933	0.776
	TR2	0.885			
	TR3	0.905			
	TR4	0.864			

Discriminant validity reflects whether the correlation between the measurement items of different variables is as small as possible. The standard for evaluating each variable is that the square root of the AVE value must be greater than the correlation coefficient between the variable and the other variables [40]. As illustrated in Table 3, in each case, the square root of the AVE value (the value of the diagonal line) is above than the correlation coefficient between the variable and the other variables, which means that the model has excellent discriminant validity [40].

4. Results Analysis

4.1. PLS-SEM Testing. It is usually possible to use three types of methods when studying the relationship between variables. This study is based on the partial least squares method, which was proposed by World et al. in 1983 [41]. It is a new type of multivariate statistical analysis method, which incorporates both regression analysis and structural equation modeling. It is capable of constructing regression paths for multiple dependent variables and also for testing mediation and moderation effects at one time. It is suitable for small samples, and it is robust to collinearity problems and normally distributed data, and there will be no problems such as the model cannot be identified or the coefficient is greater than 1 which is caused by too many single measurement indicators, which makes up for the shortcomings of the traditional structural equation model.

4.2. Main Effects Test. On the premise of the reliability and validity of the measurement model, the PLS-SEM method was applied with the aid of SmartPLS 3.0 software. The bootstrap

TABLE 3: Correlation coefficient matrix and square roots of AVEs (shown as diagonal elements).

	KQ	KSB	PEOU	PU	RE	TK	TR
KQ	0.877						
KSB	0.813	0.864					
PEOU	0.765	0.793	0.898				
PU	0.685	0.833	0.746	0.927			
RE	0.625	0.628	0.562	0.535	0.873		
TK	0.797	0.743	0.727	0.686	0.547	0.827	
TR	0.829	0.789	0.66	0.659	0.662	0.693	0.881

The bold numbers in Table 3 are the square root of AVE.

repeated sampling method was used to test the significance of the collected data, and the sampling number was 5000.

In this study, the detection model was constructed in three steps for analysis and comparison: a first assessment was made on the effects of the control variable on independent and dependent variables, then the moderator variables were added, and finally, the interaction terms between the moderator variables and their respective variables were added.

In the first test (model 1), the three control variables of gender, age, and occupation have no significant correlation with knowledge sharing behavior. Figure 2 illustrates the effects of each variable on knowledge sharing behavior.

Based on Figure 2, it can be seen that in the first dimension, reciprocity ($\beta=0.110$, $p<0.01$) and trust ($\beta=0.299$, $p<0.001$) have a significant positive impact on knowledge sharing behavior; thus, H1 and H2 were supported; second, quality of knowledge ($\beta=0.325$, $p<0.001$) and tacit knowledge ($\beta=0.214$, $p<0.001$) have a highly significant positive correlation with knowledge sharing behavior; thus, H3 and H4 were supported.

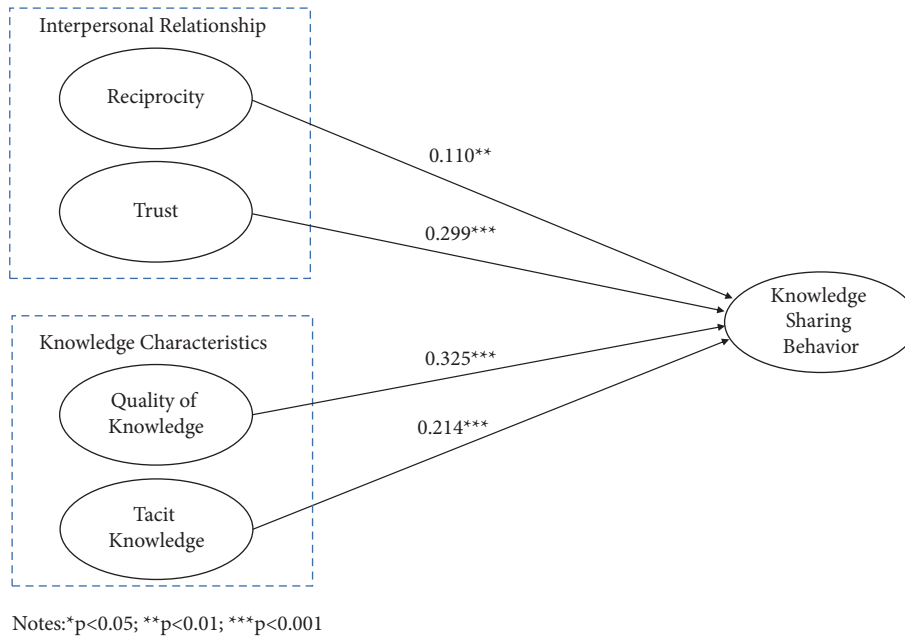


FIGURE 2: Main effects test.

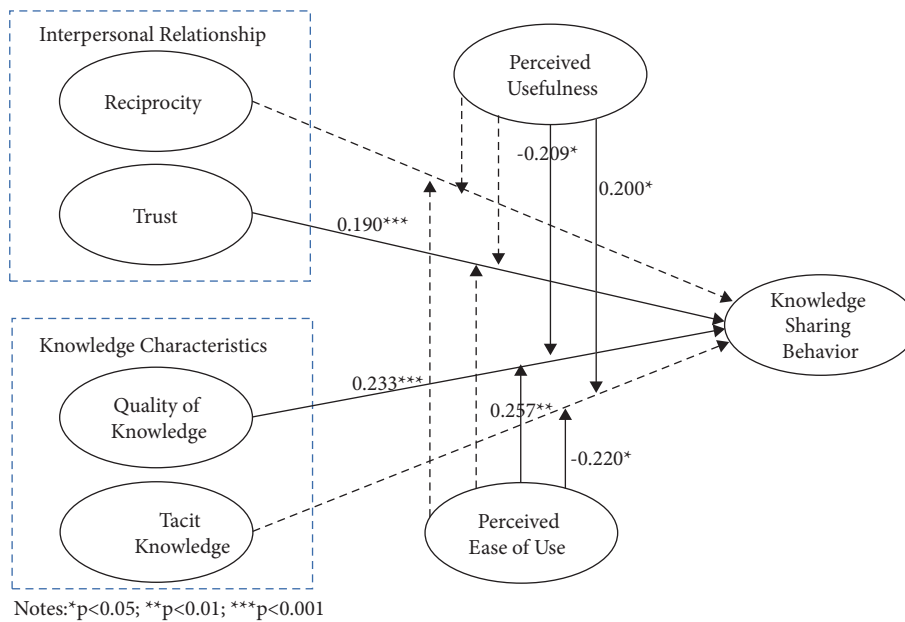


FIGURE 3: Moderated variable interaction test.

4.3. Moderated Variable Interaction Test. The second step of testing (model 2) introduces the moderating variable TAM model, that is, the influence of perceived usefulness and perceived ease of use on knowledge sharing. The test results reveal that perceived usefulness ($\beta = 0.409, p < 0.001$) has a significantly positive impact on knowledge sharing behavior; moreover, perceived ease of use ($\beta = 0.165, p < 0.05$) correlates positively with knowledge sharing behavior. Nevertheless, when the influence of moderator variables on knowledge sharing behavior is added, knowledge sharing behavior is not significantly affected by latent variables such as trust and tacit knowledge.

In the third step of test (model 3), in order to detect the moderating effect, the two TAM model variables are

interacted with four latent variables, respectively, to form eight interaction terms, and the path coefficient and the correlation of the interaction terms to knowledge sharing behavior is detected and shown in Figure 3, and the results of each hypothesis are presented in Table 4.

Figure 3 shows that the relationship between reciprocity and knowledge sharing behaviors is not significantly moderated by perceived usefulness; thus, H5 was not supported. Meanwhile, perceived usefulness also does not have a significantly moderate relationship between trust and knowledge sharing behavior; therefore H6 was not supported. Perceived usefulness significantly negatively moderates the relationship between knowledge quality and knowledge sharing behavior

TABLE 4: Results of a hypothesis test.

Hypothesis	Path	WeChat, N = 336			Supported
		Model 1	Model 2	Model 3	
	(Control variables)				
	Gender	-0.025	-0.053	-0.052	
	Age	0.036	0.050	0.037	
	Occupation	-0.018	0.026	0.033	
	Independent variables				
H1	RE- > KSB	0.110**	0.055	0.050	Partially
H2	TR- > KSB	0.229***	0.212***	0.190***	Yes
H3	KQ- > KSB	0.325***	0.164**	0.233***	Yes
H4	TK- > KSB	0.214***	0.028	-0.001	Partially
	Moderator				
	PU- > KSB		0.409***	0.415***	
	PEOU- > KSB		0.165*	0.143*	
H5	PU × RE- > KSB			-0.068	No
H6	PU × TR- > KSB			0.070	No
H7	PU × KQ- > KSB			-0.209*	No
H8	PU × TK- > KSB			0.200*	Yes
H9	PEOU × RE- > KSB			0.069	No
H10	PEOU × TR- > KSB			-0.099	No
H11	PEOU × KQ- > KSB			0.257**	Yes
H12	PEOU × TK- > KSB			-0.220*	No
R2		0.736	0.837	0.848	

Notes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

($\beta = -0.209$, $p < 0.05$); whereas, in the main effect, knowledge sharing behavior is significantly positively influenced by knowledge quality, indicating that for users with higher perceived usefulness, the quality of their knowledge will reduce their knowledge sharing behaviors. Thereby, H7 was not supported. Knowledge tacitness and knowledge sharing behavior are significantly positively mediated by perceived usefulness ($\beta = -0.200$, $p < 0.05$); thus, H8 was supported.

In the moderating effect test of perceived ease of use, knowledge sharing behavior did not significantly correlate with the interaction items formed by perceived ease of use and reciprocity and perceived ease of use and trust, indicating that neither H9 nor H10 obtained support. Nonetheless, a positive correlation of knowledge sharing behavior exists between perceived ease of use and quality of knowledge as an interaction term ($\beta = -0.257$, $p < 0.01$), and the correlation coefficient is greater than the path coefficient of knowledge quality and knowledge sharing behavior in model 3 ($\beta = 0.233$, $p < 0.001$), that is, the perceived ease of use significantly moderates the relationship between knowledge quality and knowledge sharing behavior; therefore, H11 was supported. The perceived ease of use significantly negatively moderates the relationship between tacitness knowledge and knowledge sharing behavior ($\beta = -0.220$, $p < 0.05$); thus, H12 was not supported, and observations reveal that users with higher perceived ease of use and knowledge tacitness will degrade their behavior towards knowledge sharing.

5. Conclusion

5.1. Discussion. Based on the data study results, this study draws the three following conclusions:

- (1) Trust and quality of knowledge are the main motivations for knowledge sharing behavior, based on the

results of the previous study by Ridings et al. [42] and Zhang et al. [43]. Trust is the premise and foundation of knowledge sharing behavior; trust can be explained as people believe that the recipient is honest, trustworthy, and reliable, and this will motivate them to share their knowledge; hence, they will seek help from other members of the online community when they encounter problems, to ask questions, communicate, and discuss in the community and be willing to accept the experience and knowledge shared by others. Conversely, acquiring high-quality knowledge serves as an internal driving force for fostering knowledge sharing behavior, knowledge quality is the primary driving force for encouraging users to engage in community activities, and high-quality knowledge can satisfy individuals' external knowledge desires. Exchange is conducive to efficient knowledge consolidation and innovation among members and is also avail to the development and growth of online communities.

- (2) Furthermore, reciprocity and knowledge tacitness factors will also possess a positive impact on knowledge sharing. However, there is still considerable room for improvement. Possibly, this is due to the lack of clearly defined reciprocity norms of the online community, the lack of a strong atmosphere of reciprocity, or the degree to which individual differences are felt and the existence of an independent sense of self. Compared with explicit knowledge, tacit knowledge is difficult to describe, cannot be expressed precisely, and shows less mobility. It has strong unique characteristics, and it is difficult for individuals to share and understand. However, there is also a need to take into

consideration the influence of reciprocal and demand for tacit knowledge in knowledge sharing behaviors.

- (3) Perceived usefulness and perceived ease of use have significant moderating effects on knowledge characteristics (quality of knowledge and tacit knowledge) in knowledge sharing behavior. From the perspective of moderating effects, the interaction items of users' perceived usefulness and knowledge quality have a significantly negative correlation with the knowledge sharing behavior of online community users. Users who perceive usefulness generally need community members to help solve problems or provide assistance, and as these problems are usually not complex and difficult to solve, members who are easy to solve and do not require a high-quality knowledge level are more willing to share. Furthermore, high-quality content sharing can increase community stickiness and create a more active community.

Perceived usefulness has a significantly positive moderating impact on the relationship between knowledge tacitness and knowledge sharing behavior. Users in a community are often in a network of relationships formed by holding the same purpose, using social networks to break communication barriers and enhance social links. If individuals in the community exchange and share experiences, everyone could usually perceive each other and actively participate in it, to exchange and share their own experiences.

There is a significant positive moderating effect of perceived ease of use on the relationship between knowledge quality and knowledge sharing behaviors. This could explain that with the growth of information technology, people can easily use technology to spread and share high-quality knowledge. The interaction terms formed by perceived ease of use and knowledge tacitness have a significantly negative correlation with knowledge sharing behavior in online communities. This is because the development of technology has caused the problem of rights protection for the sharing of personal experience.

The ability to easily acquire useful and high-quality knowledge and tacit knowledge for the online community will motivate users to actively participate. This also explains why users pay to participate in some online community knowledge sharing activities related to their profession. Improving the professional and tacit knowledge of the online community can enhance the activity and stickiness of the online community.

It was not found that perceived ease of use and perceived usefulness moderated interpersonal relationships (reciprocity and trust) in knowledge sharing behavior. In terms of perceived ease of use and perceived usefulness, the effects of reciprocity and trust on knowledge sharing behaviors did not change significantly.

5.2. Contribution and Limitation. The sharing of knowledge among members of a network community represents a new

mode of dissemination of knowledge. This study starts from the two dimensions of interpersonal relationships (reciprocity and trust) and knowledge characteristics (knowledge quality and tacit knowledge), meanwhile introducing two dimensions of perceived usefulness and perceived ease of use as moderating variables to construct a new theoretical model. This study found that trust and quality of knowledge were the main motivations for knowledge sharing behavior, and perceived ease of use and perceived usefulness had a significant moderating effect on knowledge characteristics (quality of knowledge and tacit knowledge) in knowledge sharing behavior. In this study, a holistic study is conducted on the interaction of humans, the interaction between humans and knowledge, and the interaction between humans and systems.

Practical guiding significance: (1) the online community platform should focus on maintaining and hoisting the credibility and knowledge quality of the platform, thereby improving the activity and stickiness of the platform and (2) the online community platform exalts the professional pertinence and operation convenience of the platform. In this way, knowledge sharing behaviors are more likely to be generated.

Despite the fact that this study examines knowledge sharing behavior from the perspectives of reciprocity, trust, knowledge quality, and knowledge tacitness, there are still some limitations that remain.

This study does not analyze the differences between different genders and different groups of age levels. For example, there may be differences between males and females or there may be differences in knowledge sharing behaviors among different age groups, and these factors may need to be further studied.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] T. Wang, G. Qi, Y. Zhang, and Y. Liang, "Research on continuous knowledge sharing behavior in open innovation communities," *Information Science*, vol. 36, no. 2, pp. 139–145, 2017.
- [2] X. Zhang and J. Su, "A combined fuzzy DEMATEL and TOPSIS approach for estimating participants in knowledge-intensive crowdsourcing," *Computers & Industrial Engineering*, vol. 137, Article ID 106085, 2019.
- [3] H. M. Lai and T. T. Chen, "Knowledge sharing in interest online communities: a comparison of posters and lurkers," *Computers in Human Behavior*, vol. 35, pp. 295–306, 2014.
- [4] C. W. Phang, A. Kankanhalli, and R. Sabherwal, "Usability and sociability in online communities: a comparative study of knowledge seeking and contribution," *Journal of the Association for Information Systems*, vol. 10, no. 10, pp. 721–747, 2009.

- [5] X. Xu and F. Liu, "Research on the influence of social capital and technical support on knowledge sharing in academic virtual communities: analysis of knowledge sharing elements based on the platform of science net," *Science & Technology Progress and Policy*, vol. 2, pp. 1–10, 2022.
- [6] K. W. Gang and T. Ravichandran, "Exploring the determinants of knowledge exchange in virtual communities," *IEEE Transactions on Engineering Management*, vol. 62, no. 1, pp. 89–99, 2015.
- [7] C. E. Connelly, D. Zweig, J. Webster, and J. P. Trougakos, "Knowledge hiding in organizations," *Journal of Organizational Behavior*, vol. 33, no. 1, pp. 64–88, 2012.
- [8] Z. Li, M. Li, and J. He, "An empirical research influencing factors of knowledge sharing willingness among virtual community members," *Library and Information Service*, vol. 53, no. 53, pp. 53–56, 2009.
- [9] C. Sun and Y. hong, "Research on the influencing factors of knowledge sharing in virtual health community: based on perspective of social cognitive and perceptual interaction," *Information Research*, vol. 264, no. 10, pp. 34–39, 2019.
- [10] D. Gambetta, *Can We Trust? Trust: Making and Breaking Cooperative Relationships*, pp. 213–237, Department of sociology University of Oxford, Oxford, UK, 2000.
- [11] P. Blau, *Exchange and Power in Social Life*, p. 352, John Wiley & Sons, New York, USA, 1964.
- [12] M. Polanyi, *Personal Knowledge: Toward A Post-Critical Philosophy*, p. 373, Harper Torch book, New York, USA, 1962.
- [13] Okyere-Kwakye, "Individual factors and knowledge sharing," *American Journal of Economics and Business Administration*, vol. 3, no. 1, pp. 66–72, 2011.
- [14] M. Kelley and J. W. Thibaut, *Interpersonal Relationship: A Theory of Interdependence*, pp. 1–24, Wiley, New York, USA, 1978.
- [15] M. M. Wasko and S. Faraj, "Why should I share? Examining social capital and knowledge contribution in electronic networks of practice," *MIS Quarterly*, vol. 29, no. 1, pp. 35–57, 2005.
- [16] R. Regilsberger, M. Sasse, and J. McCarthy, "The students dilemma: evaluating trust in computer mediated communication," *International Journal of Human-Computer Studies*, vol. 58, pp. 759–781, 2003.
- [17] L. D. Molm, "Power, Trust and fairness: comparison of negotiated and reciprocal exchange," in *Power and Status (Advances in Group Processes)*, S. R. Thye and E. Lawler, Eds., pp. 31–65, Emerald Group Publishing Limited, Bingley, West Yorkshire, 2003.
- [18] J. H. Dyer and H. Singh, "The relational view: cooperative strategy and sources of interorganizational competitive advantage," *Academy of Management Review*, vol. 23, no. 4, pp. 660–679, 1998.
- [19] S. Jiafu, Y. Yu, and Y. Tao, "Measuring knowledge diffusion efficiency in R&D networks," *Knowledge Management Research and Practice*, vol. 16, no. 2, pp. 208–219, 2018.
- [20] M. Parent, R. B. Gallupe, W. D. Salisbury, and J. M. Handelman, "Knowledge creation in focus groups: can group technologies help?" *Information & Management*, vol. 38, no. 1, pp. 47–58, 2000.
- [21] S. J. Yang and I. Y. Chen, "A social network-based system for supporting interactive collaboration in knowledge sharing over peer-to-peer network," *International Journal of Human-Computer Studies*, vol. 66, no. 1, pp. 36–50, 2008.
- [22] Y. Yang, X. Zhang, and J. Su, "Knowledge transfer efficiency measurement with application for open innovation networks," *International Journal of Technology Management*, vol. 81, no. 1/2, pp. 118–142, 2019.
- [23] B. Yan and L. Jian, "Beyond reciprocity the bystander effect of knowledge response in online knowledge communities," *Computers in Human Behavior*, vol. 76, pp. 18, 2017.
- [24] P. Wang, G. Hou, and L. Yang, "Evolutionary game analysis of knowledge sharing of users in social Q&A community based on knowledge quality & quantity," *Journal of Modern Information*, vol. 38, no. 4, pp. 42–57, 2018.
- [25] Z. Yan, T. Wang, and Y. Chert, "Knowledge sharing in online health communities: A social exchange theory perspective," *Information & Management*, vol. 53, no. 5, pp. 643–653, 2016.
- [26] M. Gagne, "A model of knowledge—sharing motivation," *Human Resource Management*, vol. 48, no. 4, pp. 571–589, 2009.
- [27] J. Howells, "Tacit knowledge," *Technology Analysis & Strategic Management*, vol. 8, no. 2, pp. 91–106, 1996.
- [28] R. M. Grant, "Toward a knowledge-based theory of the firm," *Strategic Management Journal*, vol. 17, no. S2, pp. 109–122, 1996.
- [29] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, New York, USA, 1995.
- [30] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.
- [31] Y. Wang and D. R. Fesenmaier, "Towards understanding members' general participation in and active contribution to an online travel community," *Tourism Management*, vol. 25, no. 6, pp. 709–722, 2004b.
- [32] Y. Wang and D. R. Fesenmaier, "Modeling participation in an online travel community," *Journal of Travel Research*, vol. 42, no. 3, pp. 261–270, 2004a.
- [33] J. Hagel, "Net gain: expanding markets through virtual communities," *Journal of Interactive Marketing*, vol. 13, no. 1, pp. 55–65, 1999.
- [34] J. Preece, *Online Communities: Designing Usability and Supporting Socialbility*, John Wiley & Sons, New York, USA, 2000.
- [35] E. Parra-López, J. Bulchand-Gidumal, D. Gutiérrez-Taño, and R. Díaz-Armas, "Intentions to use social media in organizing and taking vacation trips," *Computers in Human Behavior*, vol. 27, no. 2, pp. 640–654, 2011.
- [36] J. Su, X. Chen, F. Zhang, N. Zhang, and F. Li, "An intelligent method for lead user identification in customer collaborative product innovation," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 5, pp. 1571–1583, 2021.
- [37] J. Su, Y. Yang, and R. Duan, "A CA-based heterogeneous model for knowledge dissemination inside knowledge-based organizations," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 4, pp. 2087–2097, 2018.
- [38] K. Hyan Yoo and U. Gretzel, "The influence of perceived credibility on preferences for recommender systems as sources of advice," *Information Technology & Tourism*, vol. 10, no. 2, pp. 133–146, 2008.
- [39] A. H. Segars, "Assessing the unidimensionality of measurement: a paradigm and illustration within the context of information systems research," *Omega*, vol. 25, no. 1, pp. 107–121, 1997.
- [40] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981.

- [41] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.
- [42] C. M. Ridings, D. Gefen, and B. Arinze, "Some antecedents and effects of trust in virtual communities," *The Journal of Strategic Information Systems*, vol. 11, no. 3-4, pp. 271–295, 2002.
- [43] M. Zhang, guoqing Tang, and yan. Zhang, "Influencing factors analysis of Users' Knowledge sharing behavior in virtual communities based on stimulus-organism-response paradigm," *Information Science*, vol. 35, no. 11, pp. 149–155, 2017.

Research Article

The Chain Mediating Effect of Network Behavior and Decision Self-Efficacy between Work Skills and Perceived Employability Based on Social Cognitive Theory

Liping Yang ¹ and Hong Zhang ²

¹International College, National Institute of Development Administration, Bangkok 10240, Thailand

²Institute of Social Technology, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

Correspondence should be addressed to Hong Zhang; 6422032017@stu.nida.ac.th

Received 9 August 2022; Revised 14 September 2022; Accepted 23 September 2022; Published 4 October 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Liping Yang and Hong Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The purpose of this study is to investigate the chain mediating effects of networking behaviors and decision self-efficacy between work skills development and perceived employability. Structural equations modeling is used to analyze data collected from 813 Chinese students. The results show the following: first, the work skills development is positively correlated with perceived employability. Second, network behavior and decision self-efficacy each have a mediating effect between work skills development and perceived employability. Finally, this study found a chain mediating effect of network behavior and decision self-efficacy between work skills development and perceived employability. Therefore, this research shows that Work-Integrated Learning (WIL) needs to focus not only on skills development and employability outcomes but also on developing a strong network-based platform for stakeholders. In addition, higher education institutions and workplaces should also provide career guidance and counseling centers to help students build confidence in career decision-making and ensure students' mental health care and healthy career development.

1. Introduction

Work skills are essential to people's mental health care. A survey revealed nine work skills in the Indian healthcare industry. They found that with changing supply and demand patterns and customer demand for service excellence, workplaces are increasingly seeking greater proficiency, a serious challenge in today's era. Hence, organizations expect employees to have excellent employability skills. They also found that, in the healthcare industry, employee work skills were positively correlated with patient satisfaction. Therefore, the healthcare industry also needs to train effective work skills to remain competitive [1]. Network behavior described a form of network-based social support, and it has always been closely associated with students' mental health issues. One study used four scales, the Symptom Checklist 90 (SCL-90), the Teacher-Student Relationship Questionnaire

(TSRQ), and the Peer Relationship Scale (PRS), and assessed psychological symptoms, quality of teacher-student relationships, and quality of peer relationships. They found that risk from all types of psychological symptoms was associated with school ties. Furthermore, poor school relationships carry a high risk of mental health problems. So they suggested that school administrators should urgently improve students' school relations [2]. In fact, communication skills in work skills are positively related to self-efficacy. In addition, superior communication skills can aid in treatment and effective care in the healthcare industry, and training courses in communication skills help improve self-efficacy [3]. And then, career decision-making self-efficacy helps to improve the emotional life quality of students. Higher career decision self-efficacy leads to more positive emotions [4]. Furthermore, perceived employability is a crucial psychological protection resource. It reduces the psychological

distress and worry of work seekers due to employment difficulties and also reduces the current fears caused by COVID-19 and it promotes market prosperity. Therefore, colleges and universities should improve the employability of students. For example, career guidance and training to improve students' employability [5]. This study analyzed the relationship between work skills and perceived employability from the perspective of mental health care. Work skills development cultivates the employability of university students, and good employability is an important driving force for students' future career success. Currently, perceived employability places new demands on work skills development. However, most of the current research and discussions in this area focus on the assessment of work skills development [6], students' skills for coping with work readiness [7], the application of work skills development models [8], curriculum mapping [9], and whether work skills development can affect graduate employability [10, 11]. We need to focus on skills development. Because the past study has shown that the ownership of employability skills has the possibility to find out satisfactory careers for students, because they will be even more employable in their working livelihood [12]. Therefore, healthy career success benefits students' mental health care.

Based on social cognitive theory, the implementation of skills must be varied to fit changing environments and serve multiple aims. Cognitive training affects the beginning and middle stages of skill development. The structure of knowledge determines how to select the right skills to achieve specified goals. Continued training makes the skills easy to apply, leading to a certain level of competence [13]. Work skills development can be used as a tool for viewing student progress, and students can use it to assess their own skill levels [6]. Whereas perceived employability involves the person's feeling of his or her probabilities of gaining and keeping employment [14]. Employers value the soft skills of graduates more, and universities can be more inclined to develop soft skills courses, which can improve their employability. As a result, graduates can demonstrate soft skills to employers when they are looking for a job [10]. Consequently, it is in the best benefit of students to acquire new skills and knowledge as this is significant for their employability [15]. This study revealed a new framework for the relationship between work skills development and perceived employability that incorporates network behavior and decision self-efficacy as mediators and uses quantitative methods to verify chain mediation effects, complementing knowledge about the relationship between these variables. Furthermore, the present study has described new knowledge of these cognitive developmental processes and found that work skills development improves perceived employability by enhancing network behaviors to shape confidence in decision self-efficacy. In summary, past research has emphasized the impact of work skills on employability but has not incorporated both network behavior and decision self-efficacy into the research framework. Furthermore, this study first validates the reliability and validity of the network behavior scale and decision self-efficacy scale in a Thai environment, providing a measurement tool for future researchers.

2. Literature Review

Social cognitive theory (SCT) argued that the agents who strive to improve the quality of life and the environment are individuals [16]. Individuals are apt to seek their targets if they think their own capabilities and actions are able of meeting the wished outcomes [17]. Work skills development helps increase their cognition, continuously strengthen their skills and knowledge, and make their network behavior more exploratory, systematic, and meaningful for career development. This research used the SCT to inspect how students in work skills development can improve their capability to decision self-efficacy through network behavior, which in turn influences students' perceived employability. Employability often helps employees be flexible to changes in the work environment to societal and human resource and is explained as features that develop adaptive thinking, actions, and affections that aid personals counter flexibly to alterations in their task circumstance [18]. Therefore, persons with strong employability have a propensity to be buffered against passive influences of unemployment [18, 19]. Furthermore, some respects of employability assist persons with winning work recovery. First, one of the secrets to people's ability to stay employed is to have strong work skills and ongoing training, which enables them to find new jobs [18]. Next, people with sufficient social capital can access more resources in professional networks [20]. Social capital offers job hunters precious chances; social connections can cause job hunters to notice vacancies, bringing to "accidental job chance" [21]. Because knowledge derived from social relationships is positively correlated with people's job fit. When people meet talented insiders, they will gain a more precise view of future work [22]. Therefore, the number of informal career networks remains positively correlated with potential job opportunities and helps people gain greater employment competitiveness [23].

Social cognitive theory (SCT) describes the social transmission of new behavioral patterns [24]. It mainly includes the acquisition of knowledge, the innovation and practice of thinking, and the functional value of these elements. Its function also concerns utilization determinants. In fact, many factors, containing perceived self-efficacy to have a good command of the necessary abilities, ownership of basic resources, and outcome expectations, are related to the benefits and costs of new behavioral patterns, and the key factor that people practice is their perceived barriers and potential opportunities. In addition, social networking is also a major feature [25, 26]. Structural interconnectedness offers latent routes of affect; psychosocial factors greatly decide the destiny of what diffuses by those gregarious networks [13]. Perceived self-efficacy can directly or indirectly influence behavior, so it is critical in SCT [27, 28]. In addition, self-efficacy also positively affects people's motivation for outcome expectations [29–31]. In fact, self-efficacy determines how choices and decisions are made. For example, when people make decisions that do not ensure the success of the predetermined plan and are firmly maintained, especially when people encounter difficulties, it is important for individuals to make decisions with self-

confidence [32]. Therefore, confidence based on support and maintenance efficacy is the key to action psychology, and decision psychology needs its support [33].

According to attribution theory, human motivation can be influenced by the attribution of their performance [34]. Because people often assess whether their expectations are being met, and use the results of those assessments to guide their actions. When people imagine themselves in a situation of success, it means that they have strong efficacy, so the guidance of this efficacy positively promotes their performance and behavior. Conversely, self-doubting people impair their performance because they lack self-efficacy. In addition, when people are doing evaluations, high-performing people are more proactive in pursuing opportunity value [35, 36]. On the other hand, highly productive people have strong strategic sensitivity and high cognitive abilities that help them monitor their living environment more effectively [37]. In addition, people with high self-efficacy are better at asking deep and broad questions, and as a result, they are able to save more time, which is an easy strategy for acquiring knowledge [38].

This study also investigates whether network behavior affects decision self-efficacy. Regarding network behavior, it usually means that a person has a social relationship or a willingness to connect with others. According to social cognitive theory (SCT), SCT emphasizes the concept of collective agency. A central part of the collective agency is that people have confidence in collective strength and the ability to achieve desired outcomes. In other words, collective performance is the result of everyone's efforts [13]. In conclusion, collective self-confidence is positively correlated with people's achievement [39]. High self-efficacy always helps to coordinate and improve collective performance, both at the social and individual levels [13].

2.1. The Influence of Work Skills Development on Perceived Employability. The impact of work skills on employability is very significant because work skills bring competitiveness. For skills development, Chandran [40] described many recommendations like devising a new course, making new teaching outlines, running English word coaching plans, mixing general skills and technological skills, and conforming soft skills into the course to give the power to students with employability. In fact, graduates demand to be competitive to guarantee they can survive in the labor market. To be competitive, well-educated graduates demand to hold themselves with skills. These skills can be a feature to them, and they can decide their marketability [41]. According to Jackson [42], WIL is a tool that enhances graduate work practice and has been shown to improve graduate work skills and employability. Work skills include a range of skills that are used on the job, studies have pointed out that language skills have a positive impact on the employability of international graduates in Norway, and graduate employability is influenced by many work skills such as social skills, communication skills, IQ, and network skills, etc. These work skills all affect the employability of graduates [43]. When it comes to employability, what

matters most is the link between a job seeker's skills and an employer's needs [44]. Business skills in work skills greatly influence employability, especially for business students. Different employers need different job skills. Some employers are very obsessed with the IT skills of job seekers, some employers pursue business-related skills, and some employers need soft skills on the job, such as coordination and communication skills. Some employers require graduates to have office skills such as writing and communication skills, creative and critical thinking skills, and more. Most employers also attach great importance to the actual work experience of graduates, and graduates who have work experience or participated in job placement programs are more concerned employers. In addition, political skills have positive implications for student's entrepreneurial education [45]. Therefore, for work skills development, both hard skills and soft skills are core aspects that reflect the employability of graduates [46]. Therefore, we propose Hypothesis 1.

Hypothesis 1. There will be a positive relationship between work skills development and perceived employability.

2.2. Mediating Role of Networking Behaviors. Batistic and Tymon [47] demonstrated that networks arise from frequent access to resources and information, and it contributes to increased perceived employability. Chen [48] proposed that social networks help graduates improve their employability. University graduates ought to attach importance to the forming of studying conduct based on a sociable network to enhance their employability in China. Craig [49] confirmed that robust ways for improving employability ought to be executed to create even more skilled or equipped employees, such as offering chances for an internship, networking, and short curriculums. In addition to the industry and internship interchange, networking and response for the student are same significant for student's employability [40]. The principle of network behavior is like the knowledge creation process. The knowledge creation process is to share individual ideas, transform scattered tacit knowledge into explicit knowledge shared by organizations, and finally store knowledge in a database to integrate this scattered knowledge [50]. Network behavior is a key career strategy because it means that people have the potential to communicate and connect with potential employers ahead of time for potential employment opportunities. In addition, career outcomes were also associated with online behavior [51]. Networking is explained as a target-guided activity which happens both internally and externally in a team, concentrated on building, developing, and using relationships. Moreover, that is affected by various kinds of person, work, and team level reasons and bring about to advanced reputation and authority, work outcomes, teams gain strategic intelligence and professional success. Therefore, it is held to be of a large career worth for aspirants or organized system [52]. In fact, the moderation of skill development and network behavior was positively associated with perceived employability [53]. Some studies have tested the influence of the superior-subordinate relationship on employees' emotions, and

the research shows that the manager's work autonomy has a positive moderating effect on the relationship [54]. Therefore, maintaining a good interpersonal network has a positive significance for people's lives, and network behavior is conducive to people's good performance. Therefore, we propose Hypothesis 2.

Hypothesis 2. Network behavior plays a mediating role in the associations between work skills development and perceived employability.

2.3. Mediating Role of Decision Self-Efficacy. Makki et al. [55] found that engineering graduates had higher skill levels, had high self-efficacy, and were more eager to explore their career plans. In addition, universities can develop relevant training for them, making them highly employable. Therefore, getting enough work preparedness skills, and cultivate graduates' confidence in their abilities, will guide them toward valid exploration of career selections [56]. Perceived employability (PE) is people's viewpoint of their easy access to employment, and it is positively related to self-efficacy (SE) [57]. PE and SE are distinct but related [58]. Employability is a significant reason that can decide the quality of future graduates, Tentama and Nur [59] explored the role of SE and partner interaction on student employability. They reported SE is positively related to PE [60]. Moreover, Sultana and Malik [61] found that self-efficacy also promotes protean person to develop high perceived internal and external employability. They described the expectation of full mediation of SE on PE. Charoensukmongkol and Pandey [62] pointed that the mediating effect of sales self-efficacy between cultural intelligence and the quality of cross-cultural sales presentations. These positive effects also reflect the objectivity of self-efficacy in improving people's work quality and performance. Therefore, we propose Hypothesis 3.

Hypothesis 3. Decision self-efficacy (DSE) has a mediating role in the associations between work skills development (WSD) and perceived employability (PE).

2.4. The Chain Mediating Effect of Network Behavior and Decision Self-Efficacy. The present study has discussed the relationship between DSE and career exploration. Brown et al. [63] stated that career decision self-efficacy can lead to sustainable careers. Chen et al. [64] concluded that DSE positively predicts sustainable career development. Therefore, self-efficacy and work experience play a key role in students' career development [65]. Career decision self-efficacy predicts the purpose of career exploration [66]. Lack of participation affected career exploration and, furthermore, career self-efficacy had an impact on self-exploration [67]. People need to improve DSE for more work outcomes and sustainable career development [68]. Parents and teachers can actively contribute to DSE [69]. Program participation was positively correlated with DSE, and in addition, career help and support from school staff, as well as career-related connections and activities, supported

participants' DSE [70]. In addition, nontraditional university women with children were more likely to network with shared interests, and these network behaviors were also associated with higher levels of DSE [71]. Lastly, those who felt responsible for others' happiness also highly showed levels of DSE [72]. Therefore, with more beneficial interpersonal relationships they can brook more unpredictability and uncertainty when they make decisions. Evidently, great social bonds and great social functioning can promote their capability of controlling the future [73]. For instance, Pond and Hay [74] reported that the provision of information enhanced self-efficacy. Similarly, it could use by government employees doing organizational restructuring [75]. Network diversity can be in various contexts, such as family and friends, partnerships, someone's advice, and the same purpose of interest [76]. Degree centrality quantifies the relative number of individuals in a team, and it represents how closely an individual is connected to other people in the network [77]. Therefore, degree centrality is positively related to the number of relations in the network. Based on SCT, degree centrality confers information richness and social support [78]. Instead, this should provide a better level of confidence during the transition. The centrality of social networks establishes self-efficacy [79]. Therefore, we propose Hypothesis 4.

Hypothesis 4. Network behavior and decision self-efficacy will have a chain mediating effect between work skill development and perceived employability. In conclusion, this study has four hypotheses, as shown in Figure 1.

3. Research Methods

3.1. Respondents. Our respondents were from the Yunnan University of Business Management in Yunnan Province, China. The location was chosen because Yunnan Province has carried out higher education school-enterprise cooperation "going out" activities [80], so this study could test university students' learning outcomes and have a practical significance in the results. Second, given the good communication between us and the university, the university agreed to participate in the sampling of this study, so there is convenience and transparency in the study. Questionnaires were distributed to 1252 Chinese undergraduate students, and 813 valid questionnaires were returned, with a recovery rate of 65%. Among the interviewees, men accounted for 25.6%, women accounted for 74.4%, and there were more women than men. Young people aged 16 to 19 accounted for 44% and aged 20 to 23 accounted for 56%. Respondents are undergraduate students in various majors; 100% are undergraduates and below, 4.5% are Economics and Business students, followed by Engineering 5.9%, Humanities and Arts 61.4%, and Science 28.2%.

3.2. Procedure. The subjects of this study are Chinese undergraduate students. Respondents participated in work or internship programs to varying degrees. Participants receive support from employers and schools in terms of internships in the workplace, work skills, and other needs. The issues

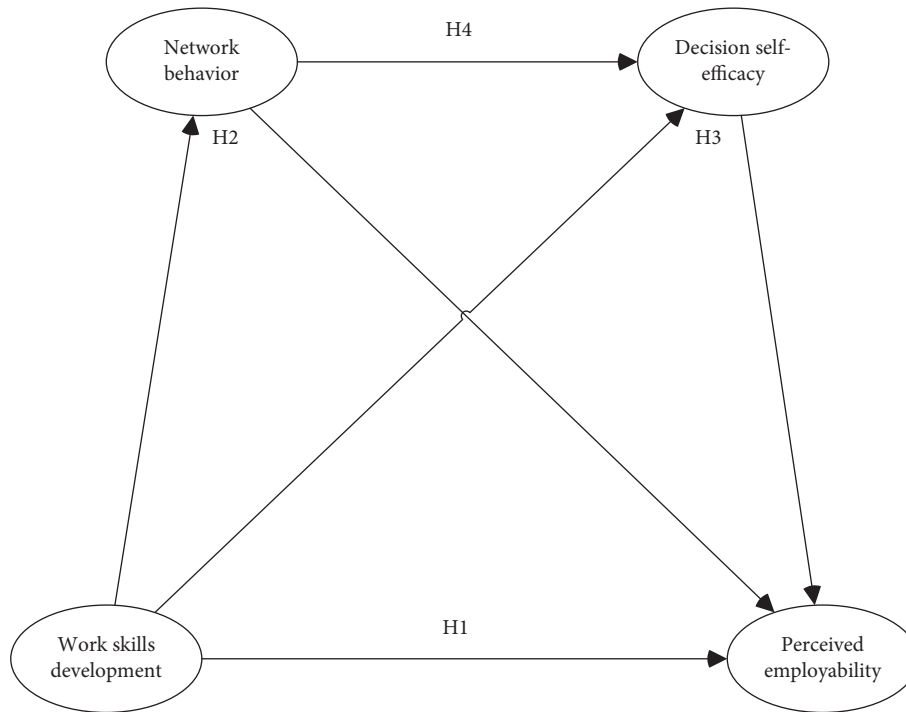


FIGURE 1: Hypothesis model.

involved in this study have been well understood by the respondents and can meet the criteria of empirical analysis. In April 2021, we communicated with the staff members in charge of the Work-Integrated Learning (WIL) internship program at the university and sent an invitation to participate in this research to their students via e-mail. We inform students that these data will be used for research purposes only and that students' personal privacy is kept confidential. In addition, students have the right to choose whether to participate in this research, and they can stop or refuse to participate at any time during the research process. We only provide access to the return form for students who would like to participate in research.

3.3. Measures. To ensure the reliability and validity of the study, the study referred to important relevant literature and selected four authoritative scales. This study summarizes the previous literature, combined with the specific scenarios of WIL, and uses the work skills development scale and the networking behaviors for career development scale, the decision self-efficacy for career exploration scale, and the perceived employability scale, so that the measurement is suitable for students and finally forms a scale. A five-point Likert scale was used in the present study, and undergraduate students evaluate the corresponding items according to their actual conditions.

3.3.1. Work Skills Development. The work skills development measured in this study is mainly based on the students' personal level. We asked students to rate themselves on a 5-point Likert scale. The lowest value is 1 = not developed, 5 =

very well developed. The 21 items are: Communication Skills; Writing Skills; Professionalism in Your Field; Interpersonal Skills; Leadership Skills; Teamwork and Cooperation; Analytical Skills; Initiative; Decision Making Skills; Problem Solving Skills; Flexibility; Self Confidence; Self-Control; Ability To Work Independently; Time Management; Ability and Willingness to Learn; Achievement Orientation; Resilience Skills; Conflict Management; Prioritization, quality, and accuracy of the work; Networking and collaborating in virtual environments [81]. The results of the study showed that the scale's Cronbach's alpha coefficient was 0.968, so it had very high reliability.

3.3.2. Networking Behavior. The scale was used to assess students' career development network behavior. This scale include "I use computers and connections for easy access to employment opportunities; I use many networking tools such as WeChat, Twitter, line, etc. to build connections or get in touch with celebrities in other industries, which facilitates my professional network; Mentors are great and they provide great advice on careers; I have great relationships with people from government agencies who offer good career advice and give me employment opportunities; I talk to family and friends about careers activities to promote the search for more employment opportunities; I am good at using computer networks to contact people of the same major as me; I am good at using social networking tools to promote the realization of my ideal career; I'm good at using the Internet and learning from it; I'm good at using the Internet to advance career skills; I am good at using the Internet to find a job." 10 items. These items are translated into English sentences that are more suitable for Chinese people to understand. The

TABLE 1: Model fit.

Model	χ^2	df	χ^2/df	CFI	TLI	AIC	BIC	RMSEA	SRMR
Four-factor model	4098.889	1169	3.506	0.905	0.900	69487.517	70220.831	0.056	0.033
Three-factor model	6944.082	1172	5.925	0.813	0.804	72326.710	73045.922	0.078	0.077
Two-factor model	9742.295	1174	8.298	0.722	0.710	75120.923	75830.734	0.095	0.087
Single-factor model	10586.684	1175	9.010	0.695	0.682	75963.313	76668.422	0.099	0.087

Four-factor model: work skills development, network behavior, decision self-efficacy, and perceived employability. Three-factor model: work skills development + network behavior, decision self-efficacy, and perceived employability. Two-factor model: work skills development + network behavior + decision self-efficacy and perceived employability. Single-factor model: work skills development + networking behaviors + decision self-efficacy + perceived employability.

TABLE 2: Descriptive statistics and correlation.

Variable	Mean	SD	CR	AVE	1	2	3	4
1. Work skills development	3.307	0.645	0.968	0.591	(0.769)			
2. Networking behaviors	3.378	0.630	0.931	0.574	0.617**	(0.758)		
3. Decision self-efficacy	3.331	0.583	0.942	0.576	0.641**	0.666**	(0.759)	
4. Perceived employability	3.275	0.628	0.888	0.532	0.596**	0.735**	0.740**	(0.729)

Note. * $p < 0.05$; ** $p < 0.01$. Diagonal number is the square root value of AVE.

results of the study showed that the scale's Cronbach's alpha coefficient was 0.930, so it had high reliability.

3.3.3. Decision Self-Efficacy. This scale was made by Lent et al. [82]. The scale is measured by 12 items, A 5-point rating scale will be used to rate each item (1 = no have confidence, 5 = highest confidence). The results of the study showed that the scale's Cronbach's alpha coefficient was 0.942, so it had high reliability.

3.3.4. Perceived Employability. This scale measures perceived internal employability by Rothwell et al. [83]. Rätty et al. [84] used this scale to measure the employability and self-perception of Finnish university students. The scale has been validated in countries such as Turkey [85], Spain [86] and Finland [87]. The developed scale includes seven items. For example: "The labor market has generally a high demand for graduates at the moment," "There are plenty of job vacancies in the geographical area in which I am looking," "I can easily find out about opportunities in my chosen field," "The skills and abilities I possess are what employers are looking for." These seven items measure the students' Self-perceived employability. The results of the study showed that the scale's Cronbach's alpha coefficient was 0.887, so it had high reliability.

3.3.5. Control Variables. Since both age and gender variables are not related to the hypothesized variables, so the study excluded the effect of control variables.

4. Results Analysis

This research uses data analysis tools such as SPSS v23 and Mplus v8.3. The research is mainly analyzed in three stages: first, the model fit, reliability, and validity of the measurement model are analyzed; second, descriptive statistics are used for the research variables; and finally, we tested the chain

TABLE 3: Heterotrait-monotrait ratio statistics (HTMT).

Construct	1	2	3	4
1. Work skills development				
2. Networking behaviors	0.649			
3. Decision self-efficacy	0.671	0.711		
4. Perceived employability	0.643	0.810	0.810	

mediation effect based on the structural equation model of Mplus v8.3 (Model 6) created by Hayes [88].

4.1. Common Method Deviation. Based on Harman's univariate method to test, using the confirmatory factor analysis (CFA), so we ran all measures fixed on one underlying factor and the results showed poor fit ($\chi^2 = 10586.684$, $df = 1175$, $\chi^2/df = 9.010$, $CFI = 0.695$, $TLI = 0.682$, $RMSEA = 0.099$, $SRMR = 0.087$) [89], the results indicated that most of the variation could not be explained by methodological factors [90]. In addition, this study also used exploratory factor analysis (EFA) that is performed without rotation. The results show that the cumulative explained total variance is 45.930%. The first factor explained less than 50% of the variance [91]. Therefore, this study did not find serious common method bias.

4.2. Confirmatory Factor Analysis. We performed a CFA for each variable by Mplus v8.3 to analyze the discriminant validity of all variables. As shown in Table 1, the fit of the four-factor model assumed is the most ideal and all meet the standard, while the fit of other factor models is relatively poor.

4.3. Descriptive Statistics and Correlation. Statistical analysis of the mean, standard deviation, and correlation coefficient of the latent variables was performed using SPSS v23. As shown in Table 2, the study found significant correlations between work skills development, network behavior, decision self-efficacy, and perceived employability.

TABLE 4: Exploratory factor analysis results.

Scales	Items	Factor 1	Factor 2	Factor 3	Factor 4
WSD	WSD1	0.680	0.144	0.122	0.230
	WSD2	0.672	0.082	0.075	0.251
	WSD3	0.683	0.187	0.168	0.135
	WSD4	0.689	0.147	0.095	0.223
	WSD5	0.693	0.175	0.195	0.202
	WSD6	0.710	0.201	0.167	0.150
	WSD7	0.725	0.213	0.153	0.135
	WSD8	0.743	0.159	0.214	0.183
	WSD9	0.730	0.184	0.221	0.194
	WSD10	0.767	0.216	0.163	0.142
	WSD11	0.753	0.232	0.217	0.111
	WSD12	0.726	0.258	0.228	0.110
	WSD13	0.658	0.261	0.163	0.030
	WSD14	0.716	0.281	0.194	0.057
	WSD15	0.710	0.206	0.204	0.094
	WSD16	0.736	0.246	0.211	0.026
	WSD17	0.722	0.214	0.233	0.129
	WSD18	0.756	0.242	0.215	0.085
	WSD19	0.706	0.214	0.227	0.130
	WSD20	0.758	0.243	0.227	0.080
	WSD21	0.701	0.200	0.303	0.122
DSE	DSE1	0.256	0.575	0.292	0.384
	DSE2	0.251	0.612	0.243	0.358
	DSE3	0.267	0.629	0.238	0.345
	DSE4	0.206	0.675	0.201	0.197
	DSE5	0.284	0.725	0.229	0.170
	DSE6	0.289	0.711	0.224	0.142
	DSE7	0.248	0.697	0.237	0.141
	DSE8	0.231	0.732	0.211	0.092
	DSE9	0.247	0.600	0.171	0.264
	DSE10	0.270	0.705	0.168	0.120
	DSE11	0.251	0.695	0.208	0.166
	DSE12	0.273	0.685	0.234	0.129
NB	NB1	0.296	0.220	0.673	0.139
	NB2	0.229	0.215	0.693	0.145
	NB3	0.314	0.258	0.653	0.108
	NB4	0.291	0.254	0.671	0.179
	NB5	0.229	0.294	0.669	0.132
	NB6	0.210	0.258	0.716	0.204
	NB7	0.226	0.209	0.726	0.224
	NB8	0.236	0.165	0.672	0.289
	NB9	0.218	0.197	0.659	0.286
	NB10	0.191	0.214	0.676	0.254
PE	PE1	0.188	0.259	0.302	0.527
	PE2	0.242	0.227	0.355	0.558
	PE3	0.227	0.249	0.309	0.658
	PE4	0.255	0.283	0.321	0.625
	PE5	0.218	0.377	0.320	0.614
	PE6	0.207	0.338	0.310	0.518
	PE7	0.216	0.434	0.260	0.550

Extraction method: principal component analysis. Rotation method: maximum variance rotation method (Varimax). ^aThe rotation has converged after 6 iterations. Note. WSD = work skills development; DSE = decision self-efficacy; NB = networking behaviors; PE = perceived employability.

4.4. Convergent Validity and Discriminant Validity Test. The AVE values corresponding to the four factors are greater than 0.5, and the CR values are higher than 0.7, which means that the data have good convergent validity [92]. As shown in Table 3, using the HTMT value for discriminant validity analysis, all HTMT values are less than 0.85, which means

that the factors have good discriminant validity [93]. The study used SPSS v23 for exploratory factor analysis, fixed the number of four factors, and set the maximum variance rotation method (Varimax) to rotate the data. The KMO value is $0.977 > 0.9$ ($p < 0.001$), Bartlett's Test of Sphericity passed, and the measurement relationship of the factors is

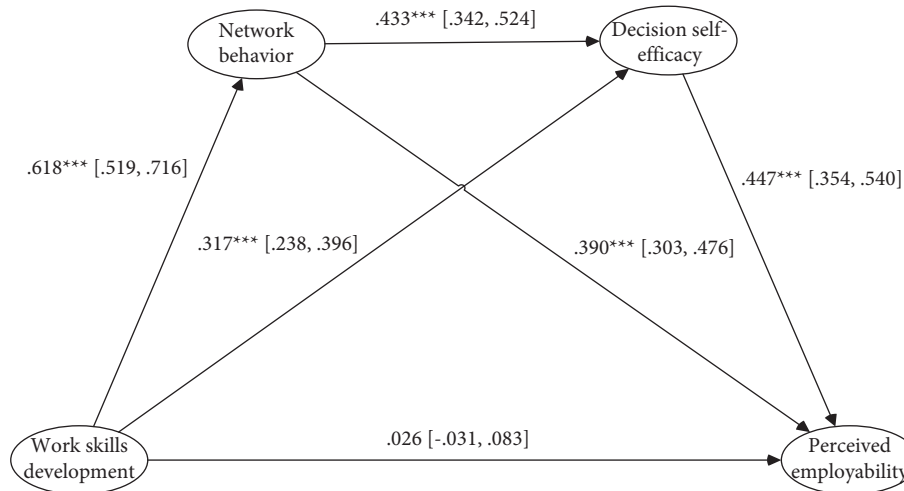


FIGURE 2: The unstandardized path coefficients in model testing. Note: *** $p < 0.001$.

TABLE 5: Chain mediating effect.

Path	Estimation	CI at 95% level		
Total indirect effect	0.502	0.426	0.584	
Total effect	0.528	0.453	0.608	
Direct effect	WSD → PE	0.026	-0.031	0.083
Indirect effect	WSD → NB → PE	0.241	0.194	0.297
	WSD → DSE → PE	0.142	0.106	0.184
	WSD → NB → DSE → PE	0.120	0.092	0.156

Note. CI, confidence Interval; WSD, work skills development; NB, network behavior; DSE, decision self-efficacy; PE, perceived employability.

good, as shown in Table 4. These results explain the relationship between the variables and provide more evidence for more analyses.

4.5. Structural Equation Modeling Analysis. First, construct a structural equation model 1, and the main effect is tested, with work skills development as the independent variable and perceived employability as the dependent variable. The fitting index of model 1 meets the requirements ($\chi^2/df = 4.912$, CFI = 0.919, TLI = 0.912, AIC = 40310.175, BIC = 40709.737, SRMR = 0.0354, and RMSEA = 0.069); therefore, the model is fitted. The main effect test results show that work skills development has a positive effect on perceived employability ($\beta = 0.526$, $p < 0.001$), and supports H1. Second, the establishment of Model 2 and Model 3 regards the network behavior and decision self-efficacy as single mediators. The results show that the model is fitted (model 2: $\chi^2/df = 4.066$, CFI = 0.912, TLI = 0.906, AIC = 2849.672, BIC = 3221.029, SRMR = 0.0353; RMSEA = 0.061; model 3: $\chi^2/df = 3.982$, CFI = 0.910, TLI = 0.905, AIC = 3100.992, BIC = 3491.153, SRMR = 0.0340; RMSEA = 0.061). Through Mplus v8.3, the bootstrap method is used to repeatedly sample 5000 times to test the mediation effect. For model 2, the mediating effect of network behavior was 0.361, with a 95% confidence interval [0.303, 0.429], excluding 0, based on the assumption that H2 was verified. For model 3, the mediating effect of decision self-efficacy is 0.384, with a 95% confidence interval [0.320, 0.455], excluding 0, based on the assumption that H3 is verified.

Finally, the chain mediation effect was tested. A correlation was observed between the two mediator variables in network behavior and decision self-efficacy. The study hypothesized that these two variables have a chain mediating effect between work skills development and perceived employability. According to Mplus v8.3, the 95% confidence interval of the mediating effect was estimated by extracting 5,000 bootstrap samples, and the chain mediation effect of network behavior and decision self-efficacy was tested significantly, as shown in Figure 2. Work skills development → network behavior → perceived employability mediating effect is 0.241, 95% confidence interval is [0.194, 0.297], excluding 0, and mediating effect is significant. Work skills development → decision self-efficacy → perceived employability, the mediating effect is 0.142, the 95% confidence interval is [0.106, 0.184], excluding 0, and the mediating effect is significant. Work skills development → network behavior → decision self-efficacy → perceived employability, the chain mediating effect is 0.120, 95% confidence interval [0.092, 0.156], excluding 0, indicating that networking behavior and decision self-efficacy are between work skills development and perceived employability, and H4 is verified. The results are shown in Table 5.

5. Conclusions

5.1. Research Conclusions. This research explores the influence mechanism of work skills development on perceived employability based on social cognition theory. The

structural equation modeling (SEM) is used to test the individual and continuous mediation effects of networking behavior and decision self-efficacy at the same time, and to verify the ability of students' networking behavior and decision self-efficacy. The multi-chain-based mediation role in the perceived employability relationship provides a new path mechanism for considering the influence of work skills development on perceived employability. Empirical research shows the following results: (1) Main effect test. The results show that there is a positive correlation between work skills development and perceived employability. (2) Mediating effect test. The test results show that networking behavior and decision self-efficacy play a mediating role in work skills development and perceived employability, respectively. Networking behavior enhances the ability of decision self-efficacy for career exploration and plays a continuous mediating role in the impact of work skills development on perceived employability.

5.2. Theoretical Implications. These findings allow us to understand how work skills acquired in Work-Integrated Learning (WIL) correlate with perceived employability and answer the question of whether network behavior, decision self-efficacy, and employability are perceived. Work skills development not only provides social capital such as soft skills but can also improve students' attitudes towards interpersonal networks and confidence in decision-making, thereby enhancing their perception and self-evaluation of employability. First, work skills development and skills that employers perceive as critical to the workplace help reduce the gap in producing high-skilled graduates [94]. And then, students try to connect with others who have the ability to provide work or work assistance [51], for example: through networking (i.e., friendships, parental support, teacher connections) [69], project participation, the ecosystem of career-related activities and other interpersonal activities enhance their decision self-efficacy [70], and the higher the self-efficacy, the naturally improved perceived employability [60]. Therefore, with good interpersonal relationships, students can make decisions more calmly. Clearly, good social bonds can enhance their abilities [73].

We take a unique approach to understand how work skills development affects perceived employability. Structural equation modeling was used to examine the multiple mediating roles of network behaviors and decision self-efficacy (DSE) between work skills development and perceived employability, elucidating the impact of career network behavior and DSE on work skills development, and clarifying the specific path and internal mechanism of career network behavior and DSE in the impact of work skills development on perceived employability. The research results verify that work skills development shapes confidence in DSE by reinforcing network behaviors and that during these cognitive developmental processes, major factors that enhance perceived employability are revealed.

5.3. Managerial Implications. Work-Integrated Learning (WIL) could improve students' employability skills, including real work skills [95], and soft skills [96, 97]. Students, as the main stakeholders of WIL, should consider their own career planning, and students should focus on work skills development, including the training of hard and soft skills for themselves. In addition, students can actively build their own various interpersonal networks during the process of internship [76]. When they are confused about their careers, they could seek career help and support from friends, family members, colleagues, schools, mentors, employers, etc., and build relationships in advance for their career decisions. At the same time, these network connections also help students' judgment ability for career exploration, so students can gain sufficient decision-making confidence and enhance their decision self-efficacy [69]. Students could focus on their work skills and could rely on strong social network relationships and decision self-efficacy to enhance their perception of employability. Therefore, these students could achieve career success in the future. In addition, this study suggested that when conducting WIL, colleges and universities should not only focus on work skills and employability but also actively help students develop interpersonal network relationships, such as the establishment of teacher-student relationships, workplace boss-student relationships, colleague-student relationships, and academic tutor-student relationship. At the same time, colleges and universities should provide more career guidance and advice for students participating in WIL. Taken together, these learning outcomes benefit student mental health care and career development.

5.4. Research Limitations and Future Perspectives. This study just determines the assessment of work skill development from the aspect of the students and fails to collect the related data on the teachers. Moreover, considering perception at distinct periods has distinct effects on people's activity and selection, future studies might use a deep interview from the perspective of students or mixed methods. Furthermore, the impact of work skill development on perceived employability is many sided. Future researchers could add dimensions to study variables in areas such as work skills development. The current study just analyses the mediating factors between work skills development and perceived employability. Therefore, future studies should analyze moderators in the research framework.

Data Availability

Data from this study are available and support the finding.

Conflicts of Interest

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

References



- [1] S. Sisodia and N. Agrawal, "Examining employability skills for healthcare services in India: a descriptive literature review," *International Journal of Service Science, Management, Engineering, and Technology*, vol. 10, no. 3, pp. 63–79, 2019.
- [2] J. Li, J. Li, R. Jia, Y. Wang, S. Qian, and Y. Xu, "Mental health problems and associated school interpersonal relationships among adolescents in China: a cross-sectional study," *Child and Adolescent Psychiatry and Mental Health*, vol. 14, no. 1, p. 12, 2020.
- [3] B. Nørgaard, J. Ammentorp, K. O. Kyvik, and P. E. Kofoed, "Communication skills training increases self-efficacy of health care professionals," *Journal of Continuing Education in the Health Professions*, vol. 32, no. 2, pp. 90–97, 2012.
- [4] E. Işik, "The relationship of career decision self-efficacy, trait anxiety, and affectivity among undergraduate students," *Psychological Reports*, vol. 111, no. 3, pp. 805–813, 2012.
- [5] Petruzzello, G., R. Chiesa, and M.G. Mariani, "The storm doesn't touch me!—the role of perceived employability of students and graduates in the pandemic era," *Sustainability*, vol. 14, no. 7, p. 4303, 2022.
- [6] S. Bandaranaike and J. Willison, "Work skill development framework: an innovative assessment for work integrated learning," in *Proceedings of the 2010 Acen National Conference*, pp. 1–19, Perth, WA, Australia, 27 September 2010.
- [7] D. Fatimah, T. Trisnarningsih, and P. Pujiati, "Soft skills of SMK IT baitunnur students in dealing with work readiness," *International Journal of Multicultural and Multireligious Understanding*, vol. 9, no. 3, pp. 335–343, 2022.
- [8] S. Bandaranaike, "The Work Skill Development [WSD] framework: work-ready competencies for today and tomorrow," in *Proceedings of the A peer reviewed short paper for a presentation at the International conference on Models of Engaged Learning and Teaching (I-MELT)*, vol. 11-13, Adelaide, South Australia, December 2017.
- [9] E. Klein and J. Lewandowski-Cox, "Music technology and Future Work Skills 2020: an employability mapping of Australian undergraduate music technology curriculum," *International Journal of Music Education*, vol. 37, no. 4, pp. 636–653, 2019.
- [10] D. J. Finch, L. K. Hamilton, R. Baldwin, and M. Zehner, "An exploratory study of factors affecting undergraduate employability," *Education + Training*, vol. 55, no. 7, pp. 681–704, 2013.
- [11] C. W. Yao and M. D. Tuliao, "Soft skill development for employability," *Higher Education, Skills and Work-based Learning*, vol. 9, no. 3, pp. 250–263, 2019.
- [12] A. Forrier and L. Sels, "The concept employability: a complex mosaic," *International Journal of Human Resources Development and Management*, vol. 3, no. 2, pp. 102–124, 2003.
- [13] A. Bandura, "Social cognitive theory: an agentic perspective," *Asian Journal of Social Psychology*, vol. 2, no. 1, pp. 21–41, 1999.
- [14] E. Berntson and S. Marklund, "The relationship between perceived employability and subsequent health," *Work & Stress*, vol. 21, no. 3, pp. 279–292, 2007.
- [15] K. Van Dam, "Antecedents and consequences of employability orientation," *European Journal of Work & Organizational Psychology*, vol. 13, no. 1, pp. 29–51, 2004.
- [16] A. Bandura, "Toward a psychology of human agency: pathways and reflections," *Perspectives on Psychological Science*, vol. 13, no. 2, pp. 130–136, 2018.
- [17] A. Bandura, G. V. Caprara, C. Barbaranelli, M. Gerbino, and C. Pastorelli, "Role of affective self regulatory efficacy on diverse spheres of Psychosocial functioning," *Child Development*, vol. 74, pp. 1–14, 2012.
- [18] M. Fugate, A. J. Kinicki, and B. E. Ashforth, "Employability: a psycho-social construct, its dimensions, and applications," *Journal of Vocational Behavior*, vol. 65, no. 1, pp. 14–38, 2004.
- [19] M. A. Gowan and R. Karren, "Employability, well-being and job satisfaction following a job loss," *Journal of Managerial Psychology*, vol. 27, no. 8, pp. 780–798, 2012.
- [20] G. Hoye, E. A. J. Hooft, and F. Lievens, "Networking as a job search behaviour: a social network perspective," *Journal of Occupational and Organizational Psychology*, vol. 82, no. 3, pp. 661–682, 2009.
- [21] S. McDonald, "Right place, right time: serendipity and informal job matching," *Socio-Economic Review*, pp. 307–331, 2009.
- [22] T. Mouw, "Social capital and finding a job: do contacts matter?" *American Sociological Review*, vol. 68, no. 6, pp. 868–898, 2003.
- [23] R. Garg and R. Telang, "To Be or not to Be linked on LinkedIn: job search using online social networks," *SSRN Electronic Journal*, 2011.
- [24] A. Bandura, "Social cognitive theory, some section from Social foundations of thought and action: a social cognitive theory," *Annals of Child Development*, vol. 6, p. 936, 1994.
- [25] M. Granovetter and R. Soong, "Threshold models of diffusion and collective behavior," *Journal of Mathematical Sociology*, vol. 9, no. 3, pp. 165–179, 2010.
- [26] E. M. Rogers and D. L. Kincaid, *Communication Networks: Toward a New Paradigm for Research* New York, US, Free Press, 1981.
- [27] A. Bandura, "Social cognitive theory of self-regulation," *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 248–287, 1991.
- [28] E. A. Locke and G. P. Latham, *A Theory of Goal Setting & Task Performance* Hoboken, New Jersey, Prentice-Hall, 1990.
- [29] I. Azjen, "Attitudes, personality and behavior," *The Journal of Genetic Psychology*, vol. 138, pp. 155–156, 1981.
- [30] J. Rotter, "Social learning theory," in *Expectations and Actions: Expectancy-Value Models in Psychology*, Em N. T. Feather, Ed., pp. 241–260, Erlbaum, Hillsdale, NJ, 1982.
- [31] V. H. Vroom, "Work and motivation," vol. 38, John Willey & Sons., New York, VroomWork and Motivation, 1964.
- [32] A. Bandura, *Self-efficacy: The Exercise of Control*, Freedom and Company, New York, 1997.
- [33] J. Deigh, "Personal being: a Theory for individual psychology. Rom harré," *Ethics*, vol. 95, no. 4, pp. 947–949, 1985.
- [34] B. Weiner, *An Attributional Theory of Motivation and Emotion*, Springer Science & Business Media, 1986.
- [35] N. F. Krueger Jr. and P. R. Dickson, "Perceived self-efficacy and perceptions of opportunity and threat," *Psychological Reports*, vol. 72, no. 3_suppl, pp. 1235–1240, 1993.
- [36] N. Krueger and P. R. Dickson, "How believing in ourselves increases risk taking: perceived self-efficacy and opportunity recognition," *Decision Sciences*, vol. 25, no. 3, pp. 385–400, 1994.
- [37] R. Wood and A. Bandura, "Social cognitive theory of organizational management," *Academy of Management Review*, vol. 14, no. 3, pp. 361–384, 1989.
- [38] S. Debowski, R. E. Wood, and A. Bandura, "Impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic

- search,” *Journal of Applied Psychology*, vol. 86, no. 6, pp. 1129–1141, 2001.
- [39] R. J. Sampson, S. W. Raudenbush, and F. Earls, “Neighborhoods and violent crime: a multilevel study of collective efficacy,” *Science*, vol. 277, no. 5328, pp. 918–924, 1997.
- [40] A. V. Chandran, “Employability skills of management students in Bangalore,” *SECOM Management & Technology Review*, vol. 12, 2019.
- [41] M. Rahmat, K. Ahmad, S. Idris, and N. F. A. Zainal, “Relationship between employability and graduates’ skill,” *Procedia - Social and Behavioral Sciences*, vol. 59, pp. 591–597, 2012.
- [42] D. Jackson, “The contribution of work-integrated learning to undergraduate employability skill outcomes,” 2013, <https://files.eric.ed.gov/fulltext/EJ1113705.pdf>.
- [43] S. Tamrakar and R. Das, “Effects of local language skills on the employability of International graduates in Norway,” 2016, <https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2461878/Tamrakar,%20S.%20og%20Das,%20R.%202016.pdf?sequence=1>.
- [44] N. T. Širca, B. Nastav, D. Lesjak, and V. Sulčić, “The labour market, graduate competences and study programme development: a case study,” *Higher Education in Europe*, vol. 31, no. 1, pp. 53–64, 2006.
- [45] X. Wei, X. Liu, and J. Sha, “How does the entrepreneurship education influence the students’ innovation? Testing on the multiple mediation model,” *Frontiers in Psychology*, vol. 10, p. 1557, 2019.
- [46] J. Andrews and H. Higson, “Graduate employability, “soft skills” versus “hard” business knowledge: a European study,” *Higher Education in Europe*, vol. 33, no. 4, pp. 411–422, 2008.
- [47] S. Batistic and A. Tymon, “Networking behaviour, graduate employability: a social capital perspective,” *Education + Training*, vol. 59, no. 4, pp. 374–388, 2017.
- [48] Y. Chen, “Graduate employability: the perspective of social network learning,” *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 6, pp. 2567–2580, 2017.
- [49] K. Craig, “Perceptions of employability within an undergraduate science department: a case study to define current strategies and recommend improvements,” *Journal of Academic Research and Essays*, 2021.
- [50] L. Yang and V. Aumeboonsuke, “The impact of entrepreneurial orientation on firm performance: the multiple mediating roles of competitive strategy and knowledge creation process,” *Mobile Information Systems*, vol. 2022, pp. 1–10, 2022.
- [51] M. L. Forret and T. W. Dougherty, “Networking behaviors and career outcomes: differences for men and women?” *Journal of Organizational Behavior*, vol. 25, no. 3, pp. 419–437, 2004.
- [52] C. Gibson, J. H. Hardy III, and M. Ronald Buckley, “Understanding the role of networking in organizations,” *Career Development International*, vol. 19, no. 2, pp. 146–161, 2014.
- [53] A. J. Clements and C. Kamau, “Understanding students’ motivation towards proactive career behaviours through goal-setting theory and the job demands–resources model,” *Studies in Higher Education*, vol. 43, no. 12, pp. 2279–2293, 2017.
- [54] P. Charoensukmongkol, “Supervisor-subordinate guanxi and emotional exhaustion: the moderating effect of supervisor job autonomy and workload levels in organizations,” *Asia Pacific Management Review*, vol. 27, no. 1, pp. 40–49, 2022.
- [55] B. I. Makki, M. U. Javaid, and S. Bano, “Level of work readiness skills, career self-efficacy and career exploration of engineering students,” *NFC-IEFR Journal of Engineering and Scientific Research*, vol. 4, no. 1, pp. 91–96, 2016.
- [56] B. I. Makki, R. Salleh, and H. Harun, “Work readiness, career self-efficacy and career exploration: a correlation analysis,” in *Proceedings of the 2015 International Symposium on Technology Management and Emerging Technologies (ISTMET)*, IEEE, Langkawai Island, August 2015.
- [57] H. Ahmed, S. Nawaz, and M. Imran Rasheed, “Self-efficacy, self-esteem, and career success: the role of perceived employability,” *Journal of Management Sciences*, vol. 6, no. 2, pp. 18–32, 2019.
- [58] E. Berntson, K. Näswall, and M. Sverke, “Investigating the relationship between employability and self-efficacy: a cross-lagged analysis,” *European Journal of Work & Organizational Psychology*, vol. 17, no. 4, pp. 413–425, 2008.
- [59] F. Tentama and M. Z. Nur, “The correlation between self-efficacy and peer interaction towards students’ employability in vocational high school,” *International Journal of Evaluation and Research in Education*, vol. 10, no. 1, pp. 8–15, 2021.
- [60] H.-y. Ngo, H. Liu, and F. Cheung, “Perceived employability of Hong Kong employees: its antecedents, moderator and outcomes,” *Personnel Review*, vol. 46, no. 1, pp. 17–35, 2017.
- [61] R. Sultana and O. F. Malik, “Protean career attitude, perceived internal employability and perceived external employability; does self-efficacy make a difference,” *Middle East J. of Management*, vol. 7, no. 4, pp. 343–364, 2020.
- [62] P. Charoensukmongkol and A. Pandey, “The influence of cultural intelligence on sales self-efficacy and cross-cultural sales presentations: does it matter for highly challenge-oriented salespeople?” *Management Research Review*, vol. 43, no. 12, 2020.
- [63] C. Brown, E. E. Darden, M. L. Shelton, and M. C. Dipoto, “Career exploration and self-efficacy of high school students: are there urban/suburban differences?” *Journal of Career Assessment*, vol. 7, no. 3, pp. 227–237, 2016.
- [64] S. Chen, H. Chen, H. Ling, and X. Gu, “How do students become good workers? Investigating the impact of gender and school on the relationship between career decision-making self-efficacy and career exploration,” *Sustainability*, vol. 13, no. 14, p. 7876, 2021.
- [65] P. A. Creed, W. Patton, and L. A. Prideaux, “Predicting change over time in career planning and career exploration for high school students,” *Journal of Adolescence*, vol. 30, no. 3, pp. 377–392, 2007.
- [66] L. A. Ochs and R. T. Roessler, “Predictors of career exploration intentions,” *Rehabilitation Counseling Bulletin*, vol. 47, no. 4, pp. 224–233, 2016.
- [67] S. Kanten, P. Kanten, and M. Yeşiltaş, “The role of career self-efficacy on the effect of parental career behaviors on career exploration: a study on school of tourism and hotel management’ students,” *European Journal of Multidisciplinary Studies*, vol. 3, no. 1, pp. 143–155, 2016.
- [68] L. Xin, F. Tang, M. Li, and W. Zhou, “From school to work: improving graduates’ career decision-making self-efficacy,” *Sustainability*, vol. 12, no. 3, p. 804, 2020.
- [69] G. V. Gushue and M. L. Whitson, “The relationship among support, ethnic identity, career decision self-efficacy, and outcome expectations in african American high school students,” *Journal of Career Development*, vol. 33, no. 2, pp. 112–124, 2016.
- [70] A. Kezar, L. Hypolite, and J. A. Kitchen, “Career self-efficacy: a mixed-methods study of an underexplored research area for first-generation, low-income, and underrepresented college students in a comprehensive college transition program,” *American Behavioral Scientist*, vol. 64, no. 3, pp. 298–324, 2019.

- [71] M. A. Padula, "Reentry women: a literature review with recommendations for counseling and research," *Journal of Counseling and Development*, vol. 73, no. 1, pp. 10–16, 1994.
- [72] J. L. Quimby and K. M. O'Brien, "Predictors of student and career decision-making self-efficacy among nontraditional college women," *The Career Development Quarterly*, vol. 52, no. 4, pp. 323–339, 2004.
- [73] Y. Zhang, L. Cui, G. Zhang, S. Sarasvathy, and R. Anusha, "An exploratory study of antecedents of entrepreneurial decision-making logics: the role of self-efficacy, optimism, and perspective taking," *Emerging Markets Finance and Trade*, vol. 55, no. 4, pp. 781–794, 2018.
- [74] S. B. Pond and M. S. Hay, "The impact of task preview information as a function of recipient self-efficacy," *Journal of Vocational Behavior*, vol. 35, no. 1, pp. 17–29, 1989.
- [75] N. L. Jimmieson, D. J. Terry, and V. J. Callan, "A longitudinal study of employee adaptation to organizational change: the role of change-related information and change-related self-efficacy," *Journal of Occupational Health Psychology*, vol. 9, no. 1, pp. 11–27, 2004.
- [76] R. S. Burt, "A note on social capital and network content," *Social Networks*, vol. 19, no. 4, pp. 355–373, 1997.
- [77] R. T. Sparrowe, R. C. Liden, S. J. Wayne, and M. L. Kraimer, "Social networks and the performance of individuals and groups," *Academy of Management Journal*, vol. 44, no. 2, pp. 316–325, 2001.
- [78] P. V. Marsden, "Homogeneity in confiding relations," *Social Networks*, vol. 10, no. 1, pp. 57–76, 1988.
- [79] J. M. Vardaman, J. M. Amis, B. P. Dyson, P. M. Wright, and R. Van de Graaff Randolph, "Interpreting change as controllable: the role of network centrality and self-efficacy," *Human Relations*, vol. 65, no. 7, pp. 835–859, 2012.
- [80] J. Yun and W. Long, "Thoughts on "going out" of school-enterprise cooperation of higher education in yunnan," in *Proceedings of the 5th International Conference on Arts, Design and Contemporary Education (ICADCE 2019)*.
- [81] C. Pop and B. Khampirat, "Self-assessment instrument to measure the competencies of Namibian graduates: testing of validity and reliability," *Studies In Educational Evaluation*, vol. 60, pp. 130–139, 2019.
- [82] R. W. Lent, I. Ezeofor, M. A. Morrison, L. T. Penn, and G. W. Ireland, "Applying the social cognitive model of career self-management to career exploration and decision-making," *Journal of Vocational Behavior*, vol. 93, pp. 47–57, 2016.
- [83] A. Rothwell, I. Herbert, and F. Rothwell, "Self-perceived employability: construction and initial validation of a scale for university students," *Journal of Vocational Behavior*, vol. 73, no. 1, pp. 1–12, 2008.
- [84] H. Rätty, U. Hytti, K. Kasanen, K. Komulainen, P. Siivonen, and I. Kozlinska, "Perceived employability and ability self among Finnish university students," *European Journal of Psychology of Education*, vol. 35, no. 4, pp. 975–993, 2019.
- [85] U. Karli, "Adaptation and validation of self-perceived employability scale: an analysis of sports department students and graduates," *Educational Research and Reviews*, vol. 11, no. 8, pp. 848–859, 2016.
- [86] R. Vargas, M. I. Sánchez-Queija, A. Rothwell, and Á. Parra, *Self-perceived Employability in Spain*, Education+ Training, 2019.
- [87] H. Rätty, K. Komulainen, C. Harvorsén, A. Nieminen, and M. Korhonen, "University students' perceptions of their "ability selves" and employability: a pilot study," *Nordic Journal of Studies in Educational Policy*, vol. 4, no. 2, pp. 107–115, 2018.
- [88] A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, Vol. 120, Guilford Publications, New York, edn edition, 2013.
- [89] L. t Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no. 1, pp. 1–55, 1999.
- [90] H. Zhou and L. Long, "Statistical remedies for common method biases," *Advances in Psychological Science*, vol. 12, no. 06, p. 942, 2004.
- [91] J. Hair Jr, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis*, 7th Edition, 2010.
- [92] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981.
- [93] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015.
- [94] A. T. St Louis, P. Thompson, T. N. Sulak, M. L. Harvill, and M. E. Moore, "Infusing 21st century skill development into the undergraduate curriculum: the formation of the iBEARS network," *Journal of Microbiology & Biology Education*, vol. 22, no. 2, pp. 001800-21-e221, 2021.
- [95] G. Crebert, M. Bates, B. Bell, C. J. Patrick, and V. Cragolini, "Developing generic skills at university, during work placement and in employment: graduates' perceptions," *Higher Education Research and Development*, vol. 23, no. 2, pp. 147–165, 2004.
- [96] E. Nicholls, M. Walsh, and M. Walsh, "University of Wolverhampton case study," *Education + Training*, vol. 49, no. 3, pp. 201–209, 2007.
- [97] C. Gribble and N. McRae, "Creating a climate for global WIL: barriers to participation and strategies for enhancing international students' involvement in WIL in Canada and Australia," in *Professional Learning in the Work Place for International Students*, pp. 35–55, Springer, 2017.

Research Article

The Regulatory Effect of Firm Size on Digital Transformation: An Empirical Study of Pharmaceutical Companies in China

Xiaowen Luo ^{1,2} and Shun-Chi Yu ¹

¹International College, National Institute of Development Administration, Bangkok 10240, Thailand

²College of Applied Technology, Yunnan Minzu University, Kunming 650504, China

Correspondence should be addressed to Shun-Chi Yu; gavinee188@gmail.com

Received 3 August 2022; Revised 23 August 2022; Accepted 13 September 2022; Published 3 October 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Xiaowen Luo and Shun-Chi Yu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Digital transformation (DT) has been a key way for pharmaceutical companies to enhance innovation and R&D capabilities, improve product quality, reduce costs, and create competitive advantages. The external environment factors and the internal conditions' factors are the main factors affecting the DT of pharmaceutical companies. This research aimed to probe the effects of the external environment factors, the internal conditions' factors, firm size, and control variables on the DT of pharmaceutical companies based on synergetics. Purposive sampling and snowball sampling were used in this research. In addition, this research collected 395 valid data from Chinese pharmaceutical companies through online questionnaires. This research used quantitative analysis, and SPSS and Amos software were applied to data processing analysis. The results of structural equation modelling (SEM) and regression analysis showed that the external environment factors and the internal conditions' factors had a significantly positive correlation with the DT of pharmaceutical companies, and the effects of the internal conditions on the DT of pharmaceutical companies were greater than that of the external environment. In addition, firm size positively moderated the relationship between the external environment, internal conditions, and the DT of pharmaceutical companies. The results of this research not only can provide theoretical reference for scholars but also put forward implementation suggestions of DT for Chinese pharmaceutical company managers.

1. Introduction

China's pharmaceutical industry has developed rapidly, and the scale of China's pharmaceutical market has grown from RMB 955.5 billion in 2012 to RMB 2.15 trillion in 2020 [1]. However, problems such as weak innovation ability and weak competitiveness of China's pharmaceutical industry are prominent. In addition, the intensification of market competition has put forward higher requirements and challenges for operation management and cost control of pharmaceutical companies. Therefore, promoting the DT of pharmaceutical companies is an effective means to promote the transformation and upgrading of pharmaceutical companies to innovative technology and enhance their competitiveness. Macroeconomic factors leading to the DT of the pharmaceutical industry include technological innovation, new regulations, increased drug production costs,

and new demands from users [2]. There are many factors that affect the DT of enterprises, which come from the external environment of the companies, including political, economic, technological, customer needs, and other factors [3]. There are factors from within the company's organization, including strategy, corporate culture, leadership, dynamic capabilities, organizational characteristics, innovation, and other factors [4]. If the main influencing factors of DT can be found and effective measures can be adopted to promote the DT of pharmaceutical companies, it is necessary to study the influencing factors of DT. Haken et al. [5] thought that greater effects can be generated through interaction between systems, often from external and internal collaboration. The application of digital technology can help companies to improve their internal cooperation ability [6]. However, DT is systematic and very complex, which requires multiparty cooperation, such as cooperation between

government and companies, companies and companies, and companies and universities [7].

The aim of this research is to discover the main factors affecting the DT of pharmaceutical companies by reviewing the literature on the external environment, the internal conditions, firm size, control variables, and DT of pharmaceutical companies and to build the model of the factors affecting DT. Quantitative analysis was utilized to test the relationship between variables and DT through SEM and regression analysis. Finally, this research can integrate the factors that affect the DT of pharmaceutical companies and put forward the implementation suggestions of DT for pharmaceutical companies.

2. Literature Review and Research Hypotheses

2.1. The Factors of External Environment. Zhu et al. [8] considered that technological capabilities and competitive pressures can influence digital transformation. Tarutè et al. [3] thought that the digital transformation of small and medium-sized enterprises can be affected by external factors. Wilaisakoolyong [9] considered that digital technologies, costs reduction, productivity improvement, government policies, the upgrading of consumer behaviour, and pressure of market competition are the main factors affecting the digital transformation of companies. Hadia and Hmoodb [10] thought that the requirements of organization expansion, the pursuit of market share growth, and market competition are major reasons for the company to implement digital transformation. Tsenzharik et al. [11] thought that government policies can promote the digital transformation of companies. The relevant government departments should lay down financial support policies and special plans to support companies in implementing digital transformation projects, provide subsidies for those companies that apply digital technologies, and then promote the digitalization of companies [12]. It is necessary for the government to formulate relevant policies to regulate the digital transformation of enterprises, and the ICT infrastructures are the foundation of the digital transformation of enterprises [13]. Digital technology plays a role in ensuring digital transformation of enterprises [14]. Therefore, the above findings lead to the following assumptions.

H1: The external environment can have a significant influence on the DT of pharmaceutical companies

H1a: Customer needs can have a significant influence on the DT of pharmaceutical companies

H1b: Market competition can have a significant influence on the DT of pharmaceutical companies

H1c: Government policy can have a significant influence on the DT of pharmaceutical companies

H1d: Digital technology can have a significant influence on the DT of pharmaceutical companies

2.2. The Factors of Internal Conditions. Factors within the organization's internal conditions can facilitate the digital transformation of companies [15, 16]. Kane et al. [17]

considered that a digital strategy is the key to the success of business transformation. Agile and learning organizations support the digital transformation of companies [18]. Leadership's digital awareness can affect an organization's digital transformation [19]. Senior leadership's support for digital transformation can determine the success or failure of a company's digital transformation [20]. Therefore, the above findings lead to the following assumptions.

H2: The internal conditions can have a significant influence on the DT of pharmaceutical companies

H2a: Digital strategy can have a significant influence on the DT of pharmaceutical companies

H2b: Organization capability can have a significant influence on the DT of pharmaceutical companies

H2c: Leadership can have a significant influence on the DT of pharmaceutical companies

2.3. Firm Size. The firms can be divided into different scales according to the number of employees, income, and total assets. The division standard of firm size in different industries is different. The number of employees of the company can affect whether the company carries out the DT projects. Firm size and its capabilities can affect the success of digital transformation [21, 22]. Firm size can influence the urgency of digital transformation [23]. Firm size has a regulatory effect in artificial intelligence applications on the performance of manufacturing firms [24]. Larger companies are more experienced in management and have more capital, human resources, digital skills, and relatively strong capabilities to advance digitalization. Firm size plays a role in the DT of companies. This study mainly divides firm size based on the number of employees. Therefore, the above findings lead to the following assumptions.

H3: Firm size can have a significantly regulatory influence on the relationship between the external environment and the DT of pharmaceutical companies

H4: Firm size can have a significantly regulatory influence on the relationship between the internal conditions and the DT of pharmaceutical companies

2.4. The Control Variables. The ownership types of Chinese enterprises include state-owned enterprises and private enterprises [25]. Eller et al. [26] considered that the type of ownership and the age of the company have positive effects on the DT of SMEs. The region, industry, and the age and business volume growth of the company can affect the transformation and upgrading of manufacturing companies [27, 28]. Senior leaders play a decisive role in the digital transformation of companies [29]. The current DT situation of companies can affect the implementation of DT in the next step [30]. Therefore, this research selected five control variables, including staff position, the region which the company belongs to, the age of the company, the type of company ownership, and the DT situation of the company.

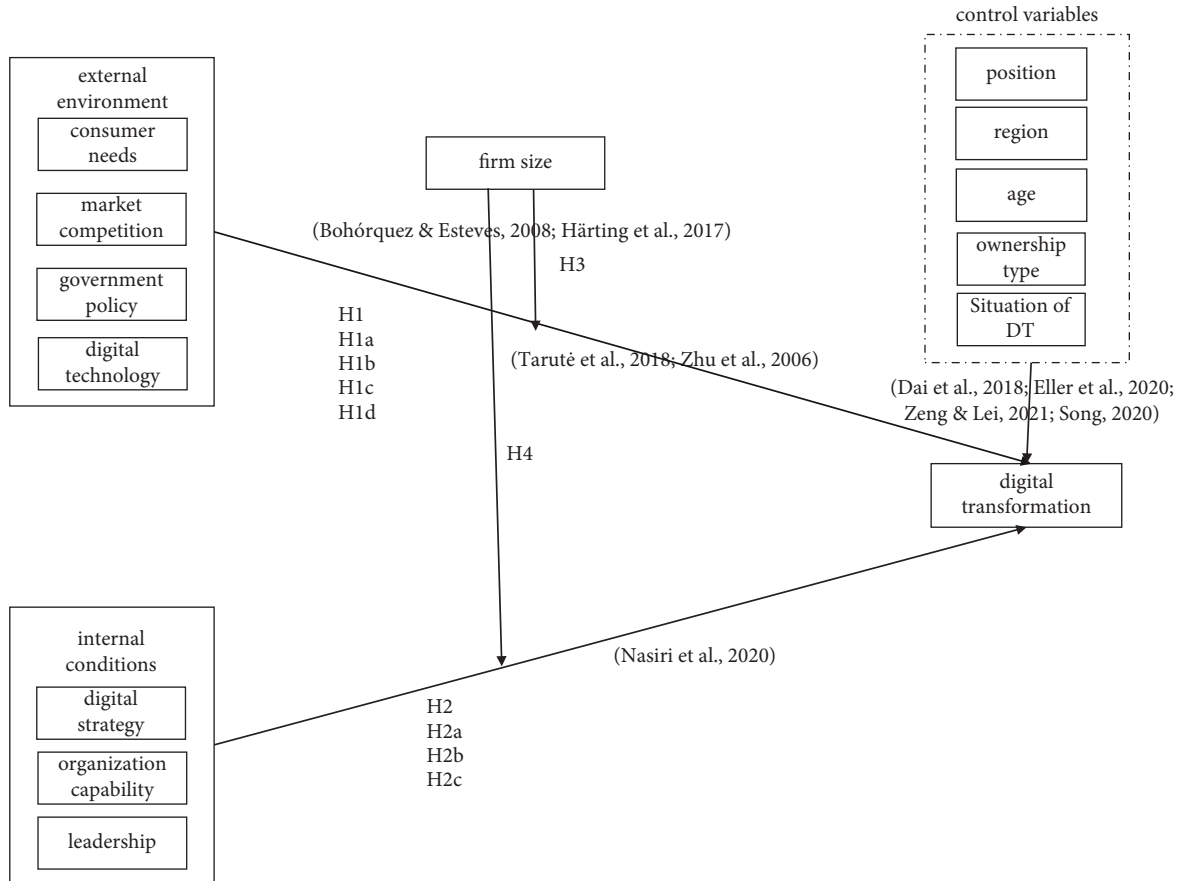


FIGURE 1: The theoretical framework of the DT of pharmaceutical companies.

2.5. *Research Framework.* From the research results of previous scholars, it was concluded that the external environmental factors and the internal conditions' factors were the main factors affecting the DT of pharmaceutical companies [31], and firm size played a regulatory role in the DT of pharmaceutical companies. According to literature, this research formed the related theoretical framework, as shown in Figure 1.

3. Methods

3.1. *Sample and Data Collection.* Since the implementation of pharmaceutical companies' DT required a high degree of professionalism, the sample objects of this research were mainly general staff, the first-line managers, and middle managers and senior managers of pharmaceutical companies.

Purposive sampling focused on candidates with similar characteristics or specific characteristics related to the subject being studied [32]. Snowball sampling pointed at the selection and investigation of several people with the characteristics required for research purposes, relying on them to select people who met the needs of the research [33]. In order to ensure the representativeness of the sample, purposive sampling and snowball sampling were applied to collect data through online questionnaires. From April 10 to May 15, 2022, a number of pharmaceutical companies in

various provinces in China sent out questionnaires, and a total of 443 questionnaires were recovered, and 395 questionnaires were valid, and the validity rate reached 89.16%.

The descriptive statistics for the questionnaire were shown in Table 1. The distribution of the region which the sample belongs to was discrete, and the distribution of the sample objects was a normal distribution, which had good external validity. According to the type of company ownership, a number of state-owned enterprises were 149, accounting for 37.72%. A number of private enterprises were 246, accounting for 62.28%. According to the situation of pharmaceutical companies' DT, pharmaceutical companies that have not undergone DT and have no willingness and plan for DT accounted for 6.33%; pharmaceutical companies that have not undergone DT and have DT willingness and plans accounted for 16.71%; pharmaceutical companies that have undergone DT and whose DT projects are in the early stages of construction accounted for 30.63%; and pharmaceutical companies that have undergone DT and have achieved certain results in DT projects accounted for 46.33%. It can be seen that DT is the general development trend for pharmaceutical companies of China.

3.2. *Measures.* This questionnaire was compiled on the basis of the measurement scale used in previous studies, and the scale had been validated by other scholars. The Likert scale

TABLE 1: Demographic statistics.

Item	Category (N = 395)	Frequency	Percentage (%)
Position	General staff	183	46.33
	First-line manager	77	19.49
	Middle manager	45	11.39
	Senior manager	90	22.78
Region	Northeast	128	32.41
	East	98	24.81
	Central	103	26.08
	West	66	16.71
Age	Within 3 years	75	18.99
	Within 3–5 years	89	22.53
	Within 5–10 years	117	29.62
	More than 10 years	114	28.86
Size	Less than 100 people	47	11.90
	100–300 people	79	20.00
	300–2000 people	145	36.71
	More than 2000 people	124	31.39
Ownership type	State-owned enterprise	149	37.72
	Private enterprise	246	62.28
Situation of DT	No and there is no intention and plan for DT	25	6.33
	No but there is a willingness and plan for DT	66	16.71
	Yes, the DT project is in the early stage of construction	121	30.63
	Yes, the DT project has achieved certain results	183	46.33

and five subjective measures were applied in this research [34]. The external environment scale was mainly composed of customer needs, market competition, government policy, and digital technology [31]. The scale of the customer needs consisted of five measurement items, such as “Do you think the improvement of customer spending power has an important influence on DT?” The scale of market competition was composed of five measurement items, such as “Do you think market competition pressure in the same industry has an important influence on DT?” The scale of government policy was composed of four measurement items, such as “Do you think government financial support and incentives have an important influence on DT?” The scale of digital technology was composed of four measurement items, such as “Do you think the support of information infrastructure has an important influence on DT?” The internal condition scale mainly consisted of digital strategy, organization capability, and leadership [31]. The scale of digital strategy was composed of four measurement items, such as “Do you think developing a digital strategy has an important influence on DT?” The scale of organization capability was composed of six measurement items, such as “Do you think organization agility has an important influence on DT?” The leadership scale was composed of four measurement items, such as “Do you think leadership awareness has an important influence on DT?” The scale of company DT was composed of four measurement items, such as “Do you think company DT will drive organization transformation?”

4. Results Analysis

4.1. Reliability Analysis. SPSS 25.0 was used for reliability analysis in this research. Tavakol and Dennick [35] considered that the corrected-item total correlation (CITC)

value of each dimension exceeds 0.5, and the Cronbach’s alpha value exceeds 0.7, indicating that the scale has high reliability. Cronbach [36] deemed that when the CITC value of a scale item is less than 0.5, and the item should be deleted. According to the reliability analysis results, it can be seen that the external environmental variables were measured through customer needs, market competition, government policy, and digital technology [31]. The CITC value of CN1 in the customer needs dimension was less than 0.5, and the overall reliability of customer needs after CN1 was deleted had increased from 0.695 to 0.822; the CITC value of MC4 in the market competition dimension was less than 0.5, and the overall reliability of market competition after when MC4 was deleted had increased from 0.887 to 0.916; the CITC value of GP3 in the government policy dimension was less than 0.5, and the overall reliability of government policies after GP3 was deleted had increased from 0.731 to 0.913; and the CITC value of TT2 in the digital technology dimension was less than 0.5, and the overall reliability of digital technology after TT2 was deleted had increased from 0.674 to 0.873. Therefore, after removing CN1, MC4, GP3, and TT2, the overall reliability of the external environment had increased from 0.880 to 0.916.

The internal condition variables were measured through digital strategy, organization capability, and leadership. The CITC values of OC2 and OC3 in the organization capability dimension were less than 0.5, and the overall reliability of organization capability after OC2 and OC3 which was deleted had increased from 0.723 to 0.881; the CITC value of LS2 in the leadership dimension was less than 0.5, and the overall reliability of leadership after LS2 was deleted had increased from 0.733 to 0.883. Therefore, after removing OC2, OC3, and LS2, the overall reliability of the internal condition had increased from 0.837 to 0.889. Therefore, after

TABLE 2: The validity of variables.

Variable	CR	AVE	Correlation of variables							
			CN	MC	GP	TT	DS	OC	LS	DT
CN	0.837	0.563	(0.750)							
MC	0.893	0.680	0.428**	(0.824)						
GP	0.839	0.634	0.541**	0.547**	(0.796)					
TT	0.856	0.664	0.429**	0.524**	0.545**	(0.815)				
DS	0.889	0.728	0.220**	0.199**	0.233**	0.169**	(0.853)			
OC	0.861	0.608	0.376**	0.315**	0.342**	0.259**	0.468**	(0.780)		
LS	0.877	0.705	0.274**	0.225**	0.280**	0.180**	0.407**	0.481**	(0.839)	
DT	0.854	0.594	0.418**	0.432**	0.449**	0.396**	0.382**	0.552**	0.428**	(0.771)

Notes: ** $P < 0.01$. The square root of AVE is presented in parentheses.

TABLE 3: The fitting indexes of the models.

Fitting index	χ^2/df	RMR	RMSEA	GFI	AGFI	NFI	IFI	CFI	TLI
Reference value	<5	<0.1	<0.08	>0.8	>0.8	>0.8	>0.8	>0.8	>0.8
The overall CFA model	1.675	0.058	0.041	0.911	0.888	0.932	0.972	0.971	0.966
The second-order factor model of EE	4.16	0.088	0.09	0.899	0.857	0.927	0.944	0.943	0.929
The second-order factor model of IC	1.151	0.029	0.02	0.982	0.969	0.985	0.998	0.998	0.997
The SEM model of DT	1.858	0.163	0.047	0.896	0.876	0.92	0.962	0.961	0.957

removing unreasonable items, the Cronbach’s alpha of each dimension exceeded 0.7. It followed that the questionnaire scale used in this research had good reliability.

4.2. *Validity Analysis.* Confirmatory factor analysis is utilized to verify the validity of scales from three aspects: convergent validity, model fit, and discriminant validity. Factors such as factor loading, composite reliability (CR), and average variance extracted (AVE) are used to measure convergent validity. Factor loading reflects the degree to which measurement items can be applied to reflect latent variables, generally exceeding 0.5 [37]. CR exceeds 0.6, indicating that the internal quality of the model is ideal [38]. AVE reflects the convergence degree of a potential variable, which is required to exceed 0.5 [39]. The standardized factor loadings of each measurement item exceeded 0.5. As shown in Table 2, the CR of each variable exceeded 0.6, indicating that the intrinsic quality of each measurement scale was good. The AVE of each variable all exceeded 0.5. It can be concluded that the scale had good convergent validity.

Fornell and Larcker [40] deemed that the square root of each variable’s AVE exceeds the correlation coefficient value of its column and row, indicating that the discriminant validity satisfies the analysis requirements. According to the results of Table 2, the measurement scale had good discriminant validity.

The operation of AMOS 23.0 had achieved the confirmatory factor model of the overall variables, the second-order factor model of the external environment factors, the second-order factor model of the internal conditions’ factors, and the SEM model of DT. According to the results of Table 3, the fitting indices of each model met the general standard, indicating that the four measurement models had good fitting effects.

4.3. *Correlation Analysis.* Ong and Puteh [39] considered that the correlation test is applied to probe whether there is a certain

correlation between variables. This research made use of Pearson’s correlation to analyse the relationship between variables. The correlation analysis results are shown in Table 4.

On account of control variables, there was a positive correlation between staff position and DT (correlation coefficient (CC) = 0.169, $P < 0.01$); there was a negative correlation between the region of the company and DT (CC = -0.157, $P < 0.05$); the age of the company was positively related to DT (CC = 0.199, $P < 0.01$); the ownership type of the company was negatively related to DT (CC = -0.132, $P < 0.01$); and there was a positive correlation between the DT situation of the company and DT (CC = 0.138, $P < 0.01$). On account of moderating variables, there was a positive correlation between firm size and DT (CC = 0.216, $P < 0.01$). The expansion of firm size had a positive effect on DT. On account of independent variables, the external environment was positively related to DT (CC = 0.535, $P < 0.01$), and customer needs, market competition, government policy, and digital technology variables in the external environment were also positively related to DT. There was a positive correlation between the internal conditions and DT (CC = 0.573, $P < 0.01$), and digital strategy, organization capability, and leadership variables in the internal conditions were also positively related to DT.

In this research, AMOS 23.0 was used to verify the SEM model of the influencing factors of pharmaceutical companies’ DT, and this model was used for parameter estimation and path analysis. The results are shown in Table 5 and Figure 2. The standardized path coefficient of the external environment on DT was 0.43 ($t = 7.094$, $P < 0.001$), indicating that the external environment had a significant influence on DT, so H1 is valid. The standardized path coefficient of the internal conditions on DT was 0.59 ($t = 8.081$, $P < 0.001$), indicating that the internal conditions have a significant influence on DT, so H2 is valid. In addition, since the standardized path coefficient of [IC-DT] was larger than that of [EE-

TABLE 4: The correlation analysis of variables.

	Position	Region	Size	Age	Type	Action	CN	MC	GP	TT	DS	OC	LS	EE	IC	DT
Position	1															
Region	-0.097	1														
Size	0.045	-0.099	1													
Age	0.104*	-0.091	-0.006	1												
Type	0.021	0.035	-0.024	-0.531**	1											
Action	0.079	-0.169**	0.009	0.668**	-0.496**	1										
CN	0.131**	-0.053	0.057	0.051	-0.074	0.039	1									
MC	0.086	-0.075	0.151**	0.172**	-0.146**	0.151**	0.428**	1								
GP	0.121*	-0.125*	0.075	0.098	-0.128*	0.067	0.541**	0.547**	1							
TT	0.035	-0.075	0.111*	0.06	-0.042	0.053	0.429**	0.524**	0.545**	1						
DS	0.211**	-0.093	0.052	0.119*	-0.08	0.089	0.220**	0.199**	0.233**	0.169**	1					
OC	0.200**	-0.064	0.041	0.109*	-0.145**	0.066	0.376**	0.315**	0.342**	0.259**	0.468**	1				
LS	0.151**	-0.091	0.097	0.169**	-0.148**	0.088	0.274**	0.225**	0.280**	0.180**	0.407**	0.481**	1			
EE	0.117*	-0.105*	0.125*	0.121*	-0.125*	0.099	0.745**	0.792**	0.840**	0.789**	0.259**	0.407**	0.303**	1		
IC	0.235**	-0.103*	0.079	0.165**	-0.157**	0.101*	0.367**	0.311**	0.360**	0.256**	0.776**	0.828**	0.786**	0.408**	1	
DT	0.169**	-0.157**	0.216**	0.199**	-0.132**	0.138**	0.418**	0.432**	0.449**	0.396**	0.382**	0.552**	0.428**	0.535**	0.573**	1

Notes: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

TABLE 5: Parameter estimates for the model variables.

Hypothesis	Hypothesis path	Estimate	S.E.	C.R.	P	Standardized estimate
H1	DT <--- EE	0.63	0.088	7.094	***	0.43
H2	DT <--- IC	0.95	0.117	8.081	***	0.59

Notes: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

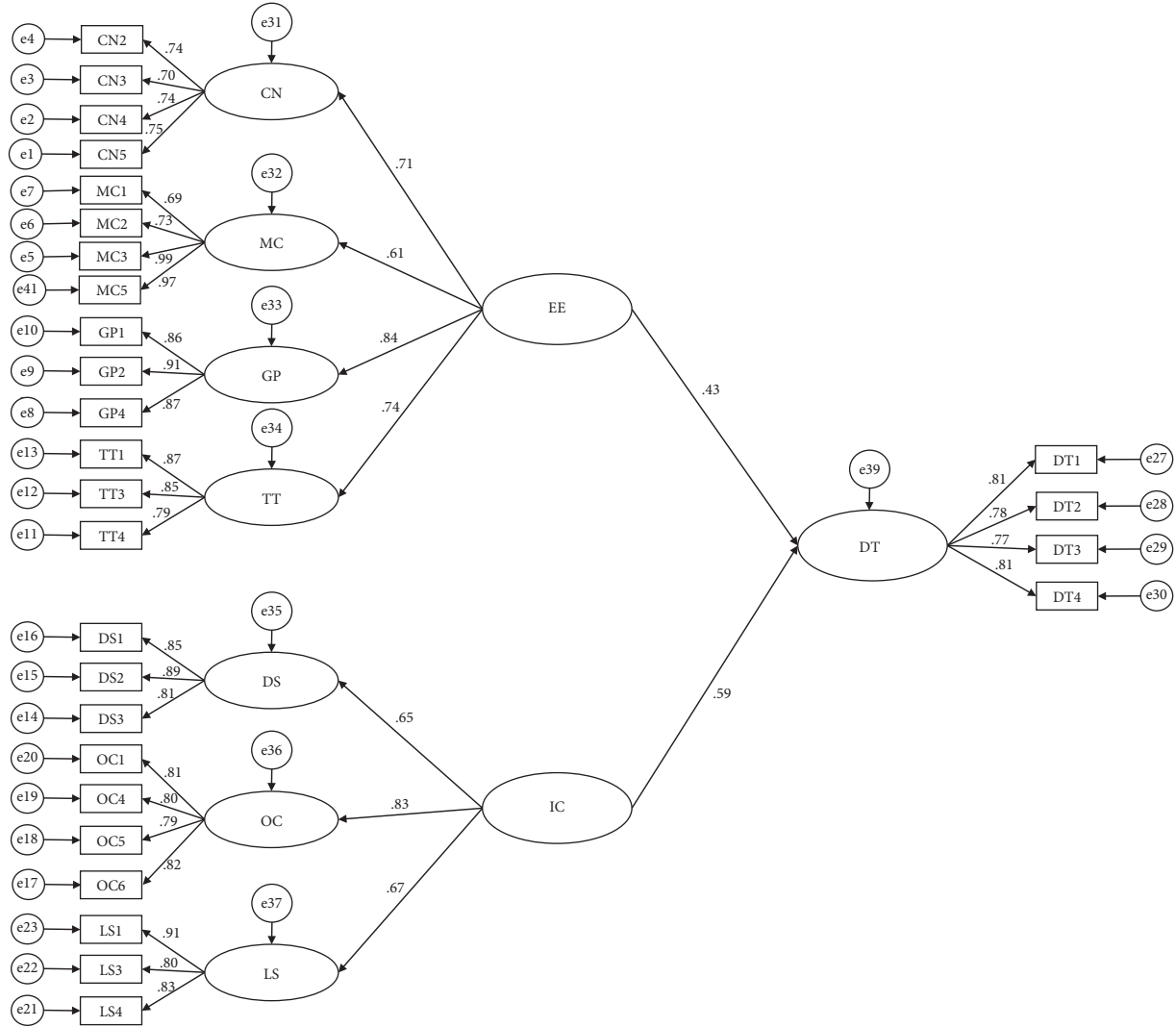


FIGURE 2: The SEM model of influencing factors of pharmaceutical companies' DT.

DT], it showed that the influence of the internal conditions on the DT of pharmaceutical companies exceeded that of the external environment.

4.4. *The Common Method Biases Test.* The Harman single factor method was used in this research. As shown in Table 6, the nonrotating factor analysis showed that there were eight factors with characteristic roots exceeding 1, of which the factor with the largest characteristic root can explain 35.753% of the overall variation. There were no serious common method biases in the data of research [41].

4.5. *Multicollinearity Test.* When the tolerance exceeds 0.1 and the variance inflation factor (VIF) is less than 10, it indicates

that there is no multicollinearity between variables [42]. The test of collinearity in this research found that the tolerance of each variable exceeded 0.1, and the VIF was less than 10, indicating that there was basically no multicollinearity between variables.

4.6. *Regression Analysis.* According to the results of Table 7, Model 1 used the control variables (position, region, age, type, and action) to perform regression analysis on the DT of the company. P3 ($\beta = 0.303, P < 0.05$) and P4 ($\beta = 0.383, P < 0.001$) indicated that the influence level of senior managers on the DT of the company was higher than that of general staff and middle managers; R2 ($\beta = 0.251, P < 0.05$) and R4 ($\beta = -0.463, P < 0.001$) indicated that regions with a high degree of economic development had a greater impact

TABLE 6: The results of common method biases test.

Component	Initial eigenvalue			Extraction sums of squared loadings		
	Total	Variance (%)	Cumulative (%)	Total	Variance (%)	Cumulative (%)
1	10.011	35.753	35.753	10.011	35.753	35.753
2	3.462	12.363	48.116	3.462	12.363	48.116
3	1.703	6.083	54.199	1.703	6.083	54.199
4	1.531	5.469	59.668	1.531	5.469	59.668
5	1.453	5.19	64.858	1.453	5.19	64.858
6	1.36	4.856	69.714	1.36	4.856	69.714
7	1.205	4.305	74.019	1.205	4.305	74.019
8	1.059	3.781	77.8	1.059	3.781	77.8
...						

TABLE 7: Results from the multiple regression analysis of EE on DT.

Independent variables		Dependent variable: DT						
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Position	P2	0.161	0.081	0.098	0.071	0.105	0.043	0.053
	P3	0.303*	0.103	0.301*	0.17	0.228	0.132	0.155
	P4	0.383***	0.254*	0.282**	0.247*	0.33**	0.227*	0.213*
Region	R2	0.251*	0.115	0.163	0.18	0.173	0.1	0.114
	R3	0.067	-0.144	-0.02	-0.009	-0.053	-0.118	-0.127
	R4	-0.463***	-0.24	-0.313*	-0.208	-0.256*	-0.164	-0.202
Age	A2	0.293	0.317*	0.214	0.14	0.166	0.15	0.131
	A3	0.503**	0.536***	0.469**	0.442**	0.408*	0.4**	0.381*
	A4	0.46*	0.47**	0.345*	0.382*	0.382*	0.341*	0.335*
Type	T2	-0.056	-0.017	-0.01	0.037	-0.048	0.013	0.022
Action	Action2	-0.628**	-0.497**	-0.567**	-0.411*	-0.54**	-0.451*	-0.433*
	Action3	-0.449*	-0.401*	-0.409*	-0.244	-0.355	-0.291	-0.29
	Action4	-0.532*	-0.425*	-0.506*	-0.311	-0.424*	-0.351	-0.338
Size			0.163***	0.125**	0.155***	0.141***	0.135***	0.153***
CN			0.389***					
MC				0.352***				
GP					0.348***			
TT						0.331***		
EE							0.589***	0.568***
Zscore(Size) *								0.104**
Zscore(EE)								
F		5.325***	10.420***	10.523***	10.968***	9.662***	14.438***	14.374***
R ²		0.154	0.292	0.294	0.303	0.277	0.364	0.378
Adjusted R ²		0.125	0.264	0.266	0.275	0.248	0.338	0.352
D-W		0.277	0.559	0.552	0.545	0.514	0.688	0.694
Maximum VIF		6.878	6.903	6.882	6.982	6.907	6.924	6.928

Notes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. EE = external environment.

on the DT of the company; A3 ($\beta = 0.503$, $P < 0.01$) and A4 ($\beta = 0.46$, $P < 0.05$) indicated that the longer the company was established, the greater the impact on the DT of the company was; Action 2 ($\beta = -0.628$, $P < 0.01$), Action 3 ($\beta = -0.449$, $P < 0.05$), and Action 4 ($\beta = -0.532$, $P < 0.05$) indicated that the situation of “the company has not carried out DT, and has no willingness and plan for DT” had a significantly negative impact on the DT of the company. Therefore, the staff position, the region which the company belongs to, the age of the company, and the DT situation of the company had varying degrees of impact on the DT of the company. Model 2 reported the results of H1a, which supported a significantly positive relationship between

customer needs and DT ($\beta = 0.389$, $P < 0.001$). Model 3 reported the results of H1b, which supported a significantly positive relationship between market competition and DT ($\beta = 0.352$, $P < 0.001$). Model 4 reported the results of H1c, which supported a significantly positive relationship between government policy and DT ($\beta = 0.348$, $P < 0.001$). Model 5 reported the results of H1d, which supported a significantly positive relationship between digital technology and DT ($\beta = 0.331$, $P < 0.001$). Model 7 reported the results of H3 (the β value of Zscore(Size) * Zscore(EE) was 0.104, $P < 0.01$), and R^2 had increased from 0.364 of Model 6 to 0.378, which confirmed that firm size positively moderated the relationship between the external environment and DT.

TABLE 8: Results from the multiple regression analysis of IC on DT.

Independent variables		Dependent variable: DT					
		Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Position	P2	0.161	0.109	0.03	0.048	-0.001	0.012
	P3	0.303*	0.177	0.058	0.21	0.042	0.047
	P4	0.383***	0.224*	0.125	0.234*	0.083	0.09
Region	R2	0.251*	0.166	0.025	0.128	0.017	-0.001
	R3	0.067	0.026	-0.058	-0.021	-0.052	-0.06
	R4	-0.463***	-0.326*	-0.283*	-0.289*	-0.235*	-0.263*
Age	A2	0.293	0.236	0.245	0.259	0.213	0.201
	A3	0.503**	0.44**	0.458**	0.401*	0.338*	0.323*
	A4	0.46*	0.408*	0.395*	0.333*	0.308*	0.286
Type	T2	-0.056	-0.024	0.084	-0.009	0.049	0.056
Action	Action_2	-0.628**	-0.479*	-0.482**	-0.486*	-0.404*	-0.397*
	Action_3	-0.449*	-0.376	-0.314	-0.293	-0.248	-0.239
	Action_4	-0.532*	-0.418	-0.333	-0.347	-0.263	-0.243
Size			0.156***	0.161***	0.142**	0.145***	0.149***
DS			0.297***				
OC				0.461***			
LS					0.337***		
IC						0.622***	0.639***
Zscore(Size) * Zscore(IC)							0.092**
F		5.325***	8.773***	16.834***	9.842***	16.381***	16.113***
R ²		0.154	0.258	0.4	0.28	0.393	0.405
Adjusted R ²		0.125	0.228	0.376	0.252	0.369	0.38
D-W		0.277	0.435	0.771	0.492	0.686	0.67
Maximum VIF		6.878	6.917	6.926	6.965	6.969	6.979

Notes: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. IC = internal conditions.

According to the results of Table 8, Model 9 reported the results of H2a, which supported a significantly positive relationship between digital strategy and DT ($\beta = 0.297$, $P < 0.001$). Model 10 reported the results of H2b, which supported a significantly positive relationship between organization capability and DT ($\beta = 0.461$, $P < 0.001$). Model 11 reported the results of H2c, which supported a significantly positive relationship between leadership and DT ($\beta = 0.337$, $P < 0.001$). Model 13 reported the results of H4 (the β value of $Zscore(Size) * Zscore(IC)$ was 0.092, $P < 0.01$), and R^2 had increased from 0.369 of Model 12 to 0.38, which confirmed that firm size can positively moderate the relationship between the internal conditions and DT.

In conclusion, the hypotheses of this research were verified to be true according to the previous empirical evidence, as shown in Table 9.

5. Discussion

This research explored the relationship and effects of the external environment factors, the internal conditions' factors, firm size, and other control variables with the DT of pharmaceutical companies. The results of SEM and regression analysis can support all the proposed hypotheses. Firstly, the analysis of the results of the SEM model showed that the external environment factors can significantly affect the DT of pharmaceutical companies, which was in accordance with the previous research results [3, 8]. The internal conditions' factors

TABLE 9: Summary of research hypotheses.

Hypothesis	Validation results
H1: The external environment can have a significant influence on the DT of pharmaceutical companies	Supported
H1a: Customer needs can have a significant influence on the DT of pharmaceutical companies	Supported
H1b: Market competition can have a significant influence on the DT of pharmaceutical companies	Supported
H1c: Government policy can have a significant influence on the DT of pharmaceutical companies	Supported
H1d: Digital technology can have a significant influence on the DT of pharmaceutical companies	Supported
H2: The internal conditions can have a significant influence on the DT of pharmaceutical companies	Supported
H2a: Digital strategy can have a significant influence on the DT of pharmaceutical companies	Supported
H2b: Organization capability can have a significant influence on the DT of pharmaceutical companies	Supported
H2c: Leadership can have a significant influence on the DT of pharmaceutical companies	Supported
H3: Firm size can have a significantly regulatory influence on the relationship between the external environment and the DT of pharmaceutical companies	Supported
H4: Firm size can have a significantly regulatory influence on the relationship between the internal conditions and the DT of pharmaceutical companies	Supported

can significantly affect the DT of pharmaceutical companies, which was in accordance with the previous research results [11, 15]. Particularly, this research found that the internal conditions' factors had a greater influence on pharmaceutical companies' DT than the external environment factors and provided additional support to previous research. Secondly, according to the regression analysis results, it was shown that firm size can positively moderate the relationship between the external environment factors, the internal conditions' factors, and the DT of pharmaceutical companies, which was in accordance with the previous research results [21]. In addition, senior managers can play a decisive role in the formulation and implementation of DT strategies, which was consistent with the previous research results [29].

6. Conclusion

In conclusion, the research provided empirical evidence for the influence of the external environment factors, the internal conditions' factors, and firm size on the DT of Chinese pharmaceutical companies. The external environment factors and the internal conditions' factors were significant factors affecting the DT of pharmaceutical companies, but the internal conditions' factors had a greater influence on DT. In addition, this research found that firm size can positively moderate the relationship between the external environment, the internal conditions, and pharmaceutical companies' DT. The larger the firm size was, the more urgent and influential the need for DT had. Therefore, pharmaceutical companies implemented DT projects according to their firm size and internal conditions. This research can fill the research gaps of pharmaceutical companies' DT by exploring the relationship and effects between the external environment, internal conditions, and firm size on pharmaceutical companies' DT and also put forward suggestions for promoting the DT implementation of pharmaceutical companies. Finally, pharmaceutical companies increasingly have realized that DT is a fast solution for companies to develop, and it is essential to explore the implementation paths of DT according to their own characteristics.

6.1. Theoretical and Practical Implications. The theoretical significance of this research is to build the model of the influencing factors of pharmaceutical companies' DT and confirm the influence of the external environment, internal conditions, and firm size on the DT of pharmaceutical companies. It can enrich the research on the DT of pharmaceutical companies in China.

The practical significance of this research is to provide suggestions for pharmaceutical companies to implement DT. The content include those as follows: (1) To realize strategic digitalization, senior managers of pharmaceutical companies should regard DT as a systematic project, carry out the overall planning, top-level design, and systematic promotion, establish a DT implementation team with the responsibility system of senior managers, and build an organization that supports the mechanisms and incentives of digital operations to ensure the implementation of DT; (2) to realize the digitalization of infrastructure, pharmaceutical companies should make full

use of the new generation of information technology to promote the standardization of systems, interfaces, and network connection protocols of the hardware facilities, so as to enable the facilities of companies to have interconnection and security protection capabilities and form a foundation to support DT; (3) to realize the digitization of resources, pharmaceutical companies can build a digital platform and build a huge resource pool to support the company's efficient aggregation and dynamic allocation of various resources; (4) to realize the digitization of elements, pharmaceutical companies can use digital platforms to plan, store, and manage data resources in a unified manner to support data application innovation; (5) to realize business digitization, pharmaceutical companies can promote to reform and innovate the data-driven business processes such as drugs, R&D, clinical trials, production, operation management, marketing, and patient services and to form the new digital businesses and value-added space; and (6) to realize the benefits of digitalization, pharmaceutical companies can realize the value-added benefits and ecological construction in innovation and economic and social benefits.

6.2. Research Limitations. Due to the limitations of subjective and objective reasons, there are still some limitations in this research.

First of all, although this research initially has built the model of influencing factors of DT of pharmaceutical companies, the comprehensiveness of the model still has certain limitations. Secondly, the cross-sectional data were utilized to test the research hypotheses, but its disadvantage was that it cannot reflect the dynamic process of influencing factors on DT. Thirdly, there are limitations in data acquisition. This research adopted the method of purposive sampling and snowball sampling. Therefore, it is impossible to control the regions and objects involved in the data, which can weaken the persuasiveness of the research results to a certain extent.

Data Availability

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

Conflicts of Interest

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

References

- [1] E R I, "Pharmaceutical Digital Transformation Framework," 2020, <https://www.dx2025.com/archives/90860.html>.
- [2] R. M. L. Dierks, O. Bruyère, J. Y. Reginster, and F. F. Richy, "Macro-economic factors influencing the architectural business model shift in the pharmaceutical industry," *Expert Review of Pharmacoeconomics & Outcomes Research*, vol. 16, no. 5, pp. 571–578, 2016.
- [3] A. Tarutė, J. Duobienė, L. Klovienė, E. Vitkauskaitė, and V. Varaniūtė, "Identifying factors affecting digital transformation of SMEs," in *Proceedings of the 18th International*

- Conference on Electronic Business*, pp. 373–381, Guilin, China, December 2018.
- [4] M. K. Peter, C. Kraft, and J. Lindeque, “Strategic action fields of digital transformation an exploration of the strategic action fields of Swiss SMEs and large enterprises,” *Journal of Strategy and Management*, vol. 13, no. 1, pp. 160–180, 2020.
 - [5] H. Haken, A. Wunderlin, and S. Yigitbasi, “An introduction to synergetics,” *Open Systems & Information Dynamics*, vol. 3, no. 1, pp. 97–130, 1995.
 - [6] X. Zhang and J. Su, “A combined fuzzy DEMATEL and TOPSIS approach for estimating participants in knowledge-intensive crowdsourcing,” *Computers & Industrial Engineering*, vol. 137, Article ID 106085, 2019.
 - [7] S. Jiafu, Y. Yu, and Y. Tao, “Measuring knowledge diffusion efficiency in R&D networks,” *Knowledge Management Research and Practice*, vol. 16, no. 2, pp. 208–219, 2018.
 - [8] K. Zhu, S. Dong, S. X. Xu, and K. L. Kraemer, “Innovation diffusion in global contexts: determinants of post-adoption digital transformation of European companies,” *European Journal of Information Systems*, vol. 15, no. 6, pp. 601–616, 2006.
 - [9] N. Wilaisakoolyong, “The essential factors that improve organization to be digital enterprise in Thailand 4.0,” in *Proceedings of the 2018 22nd International Computer Science and Engineering Conference (ICSEC)*, pp. 1–4, IEEE, Chiang Mai, Thailand, November 2018.
 - [10] A. M. Hadia and S. J. Hmoodb, “Analysis of the role of digital transformation strategies in achieving the edge of financial competition,” *International Journal of Innovation, Creativity, and Change*, vol. 10, no. 11, pp. 19–40, 2020.
 - [11] M. K. Tsenzharik, Y. V. Krylova, and V. I. Steshenko, “Digital transformation in companies: strategic analysis, drivers and models,” *St Petersburg University Journal of Economic Studies*, vol. 36, no. 3, pp. 390–420, 2020.
 - [12] C. L. Chen, Y. C. Lin, W. H. Chen, C. F. Chao, and H. Pandia, “Role of government to enhance digital transformation in small service business,” *Sustainability*, vol. 13, no. 3, pp. 1028–1124, 2021.
 - [13] A. Majchrzak, M. L. Markus, and J. Wareham, “Designing for digital transformation: lessons for information systems research from the study of ICT and societal challenges,” *MIS Quarterly*, vol. 40, no. 2, pp. 267–277, 2016.
 - [14] C. Châlons and N. Dufft, “The Role of IT as an Enabler of Digital Transformation,” *The Drivers of Digital Transformation*, pp. 13–22, Springer, Singapore, 2017.
 - [15] M. Wolf, A. Semm, and C. Erfurth, “Digital Transformation in Companies—Challenges and success Factors,” in *Proceedings of the International Conference on Innovations for Community Services*, pp. 178–193, Springer, Žilina, Slovakia, June 2018.
 - [16] M. Nasiri, M. Saunila, J. Ukko, T. Rantala, and H. Rantanen, “Shaping Digital Innovation via Digital-Related Capabilities,” *Information Systems Frontiers*, pp. 1–18, 2020.
 - [17] G. C. Kane, D. Palmer, A. N. Phillips, D. Kiron, and N. Buckley, “Strategy, not technology, drives digital transformation,” *MIT Sloan Management Review and Deloitte University Press*, vol. 14, pp. 1–25, 2015.
 - [18] H. J. Gergs, “Agility and Organizational Development- Pretty Best Friends? How organizational development can support the digital transformation of companies and how it has to change itself,” *GIO-Grup.-Interakt.-Organ.-Z. Angew. Organ*, vol. 50, no. 2, pp. 101–110, 2019.
 - [19] M. Sow and S. Aborbie, “Impact of leadership on digital transformation,” *Business and Economic Research*, vol. 8, no. 3, pp. 139–148, 2018.
 - [20] A. Singh and T. Hess, “How Chief Digital Officers Promote the Digital Transformation of Their Companies,” *Strategic Information Management*, pp. 202–220, Routledge, London, UK, 2020.
 - [21] R. C. Härting, C. Reichstein, and P. Jozinovic, “The potential value of digitization for business,” *Informatik*, pp. 1647–1657, 2017.
 - [22] V. Bohórquez and J. Esteves, “Analyzing SMEs size as a moderator of ERP impact in SMEs productivity,” *Communications of the IIMA*, vol. 8, no. 3, pp. 67–80, 2008.
 - [23] B. Prasad and P. Junni, “A contingency model of CEO characteristics and firm innovativeness: the moderating role of organizational size,” *Management Decision*, vol. 55, no. 1, pp. 156–177, 2017.
 - [24] J. Chen and Y. Meng, “The Impact of Artificial Intelligence Application on the Performance of Manufacturing Enterprises: The Mediator of Labor Structure and the Moderator of enterprise Scale,” *China Labor*, vol. 4, pp. 28–46, 2021.
 - [25] G. Z. Wu and D. M. You, “Will enterprise Digital Transformation Affect Diversification Strategy?” pp. 1–34, 2021, <https://arxiv.org/abs/2112.06605>.
 - [26] R. Eller, P. Alford, A. Kallmünzer, and M. Peters, “Antecedents, consequences, and challenges of small and medium-sized enterprise digitalization,” *Journal of Business Research*, vol. 112, pp. 119–127, 2020.
 - [27] Y. Dai, N. Li, R. Gu, and X. Zhu, “Can China’s carbon emissions trading rights mechanism transform its manufacturing industry? Based on the perspective of enterprise behavior,” *Sustainability*, vol. 10, no. 7, pp. 2421–2516, 2018.
 - [28] G. Zeng and L. Lei, “Digital transformation and corporate total factor productivity: empirical evidence based on listed enterprises,” *Discrete Dynamics in Nature and Society*, vol. 2021, Article ID 9155861, 6 pages, 2021.
 - [29] J. Zhang, J. Long, and A. M. E. Von Schaeuwen, “How does digital transformation improve organizational resilience?” *Sustainability*, vol. 13, no. 20, Article ID 11487, 2021.
 - [30] Y. Song, *Research on Influencing Factors of Intelligent Manufacturing Transformation in Equipment Manufacturing Enterprises*, pp. 51–52, Changchun University of Technology, Changchun University of Technology, Changchun, China, 2020.
 - [31] X. Luo and S. C. Yu, “Relationship between external environment, internal conditions, and digital transformation from the perspective of synergetics,” *Discrete Dynamics in Nature and Society*, vol. 2022, Article ID 6756548, 12 pages, 2022.
 - [32] I. Etikan, S. A. Musa, and R. S. Alkassim, “Comparison of convenience sampling and purposive sampling,” *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 1–4, 2016.
 - [33] I. Etikan and K. Bala, “Sampling and sampling methods,” *Biometrics & Biostatistics International Journal*, vol. 5, no. 6, pp. 1–3, 2017.
 - [34] G. Norman, “Likert scales, levels of measurement and the “laws” of statistics,” *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 625–632, 2010.
 - [35] M. Tavakol and R. Dennick, “Making sense of Cronbach’s alpha,” *International Journal of Medical Education*, vol. 2, pp. 53–55, 2011.
 - [36] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.
 - [37] C. F. Chen and D. Tsai, “How destination image and evaluative factors affect behavioral intentions?” *Tourism Management*, vol. 28, no. 4, pp. 1115–1122, 2007.

- [38] M. S. Velmurugan and M. S. Velmurugan, "Information technology adoption on 3G mobile phones in India: the empirical analyses with SPSS 20, SmartPLS2. 0M3 and LIS-REL8. 80-part 2," *International Journal of Business Innovation and Research*, vol. 13, no. 2, pp. 147–166, 2017.
- [39] M. H. A. Ong and F. Puteh, "Quantitative data analysis: choosing between SPSS, PLS, and AMOS in social science research," *International Interdisciplinary Journal of Scientific Research*, vol. 3, no. 1, pp. 14–25, 2017.
- [40] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981.
- [41] H. H. Harman, "Modern Factor Analysis," *Book Modern Factor Analysis*, University of Chicago press, Chicago, IL, USA, 1976.
- [42] M. S. Mulyadi and Y. Anwar, "Impact of corporate social responsibility toward firm value and profitability," *The Business Review*, vol. 19, no. 2, pp. 316–322, 2012.

Retraction

Retracted: Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] S. Liu, "Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 3415999, 7 pages, 2022.

Research Article

Construction and Computation of the College English Teaching Path in the Artificial Intelligence Teaching Environment

Shaojie Liu 

College of Foreign Languages, Shaanxi Xueqian Normal University, Xian, Shaanxi 710100, China

Correspondence should be addressed to Shaojie Liu; 23053@snsy.edu.cn

Received 25 July 2022; Revised 29 August 2022; Accepted 6 September 2022; Published 30 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Shaojie Liu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Today, English is the world's main international language and is widely spoken. In this context, the learning of English has long been valued by all countries. In addition, English plays an essential role in the process of economic globalization, specifically in the majority of countries where English is not the native language. At the same time, as China's integration into the world grows, the importance of English has become more pronounced, and the demand for English language skills has increased. At present, China's ever-expanding opening to the outside world, its developing science and technology, and its rising international status have created an urgent need for English-speaking professionals. This will lay the foundation for China's economic and social development and enable China to play a greater and more active role in international affairs, which has great practical and far-reaching historical significance. At present, the traditional English teaching mode has been criticized due to a great number of issues. At the same time, the new English language standards and the change of evaluation methods have created new demands on English teaching methods. As a result, with the development of artificial intelligence technology, the application of artificial intelligence technology in the field of education has brought new directions for the optimization of English classroom. In addition, diversified network technologies also provide new support for the realization of intelligent and collaborative English classroom teaching. To be specific, the maturity of English intelligent speech synthesis and intelligent speech evaluation technologies has enabled intelligent technologies to reach or even surpass the level of human teachers in certain areas. However, due to the limitations of the market environment and evaluation methods, the products that can focus on English classroom teaching are not highly intelligent and less collaborative. What is more, the process of English classroom intelligence has stagnated compared to other subjects. Obviously, from the perspective of integrating artificial intelligence technology with English classroom, it is quite necessary and feasible to design and develop an intelligent teaching assistant that can be applied to college English teaching. As information technology continues to evolve, both national and teacher levels are beginning to focus on how it can be better used in the teaching of subjects. Therefore, the use of online information technology in teaching and learning presents a great opportunity for research and practice in English language learning at the university level.

1. Introduction

English is one of the major international languages spoken in the world today. What is more, English is the most widely spoken language in the world [1]. In other words, English occupies an important place in the process of economic globalization. In fact, the majority of countries place a high priority on the teaching of English [2, 3]. As China's integration into the world grows, its importance becomes more pronounced and the demand for English language skills becomes higher and higher. At the same time, China's

expanding openness to the outside world, its advancing science and technology, and its growing international status have created an urgent need for English-speaking professionals [4]. This will lay the foundation for China's economic and social development and enable China to play a greater and more active role in international affairs [5]. As a result, the teaching of English becomes an important part of the basic education in China, and it is also one of the measures of the quality of a country's nationals. For today's college students, growing up in an open and tolerant era, learning English well means understanding the rich culture behind

the language [6]. In addition, learning English gives college students a new way of seeing the world. At the same time, language is a bridge between us and another country. For example, while we are admiring the exoticism of another country, we can also use English to introduce the world to the glorious history and culture of China. This allows the world to gain a better understanding of China, and it enhances friendship between China and the rest of the world [7]. More importantly, learning English will help college students learn advanced science and technology from abroad, which will help China move forward more quickly. In order to achieve this goal, it is quite necessary to cultivate the comprehensive English application ability of college students [8]. First of all, it is necessary to improve students' English listening and speaking skills. After all, great English listening and speaking skills are positive for improving the overall quality of students in all aspects and laying a solid foundation for their future work as well as social interaction.

College English is a very crucial public course in higher education. To be specific, this kind of curriculum adopts English as the medium as well as language of instruction in order to promote general education [9]. As the most essential content of undergraduate foreign language education, college English has different characteristics from other courses [10]. It is only when its objectives, content, teaching activities, and assessment system fully reflect its unique requirements that it can truly achieve the ultimate goal of teaching this subject [11]. With the advent of the information age, global economic integration and international exchange are deepening in various industries and fields [12, 13]. The requirements for foreign language skills of professionals, especially English as an international language, are becoming more and more common, and the requirements for deeper English skills are becoming more prominent [14]. In international communication, the English intercultural communication ability of key personnel directly will also affect the process of international communication, science and technology exchange, and the smooth transition of cultural conflicts in various fields and even determines the ultimate success of project cooperation [15, 16]. The effective use of English and cross-cultural communication skills are very important.

With the rapid development of information technology, people's learning environment and learning methods are constantly being enriched and improved [17]. In this context, various educational concepts and perspectives based on the information technology environment are also emerging and new. As a result, the combination of traditional teaching design and information technology is inevitable [18]. Therefore, in recent years, the reform of English teaching in universities has been transforming into network technology-based teaching. As shown in Figure 1, multimedia, tablets, cell phones, computers, as well as other information tools are flooding into college English classrooms as effective tools to support teachers' teaching and enrich college English teaching mode [19, 20]. In this context, with the support and assistance of such information technology environment, the shortcomings of the traditional college English teaching mode have been improved to a certain extent [21]. In other

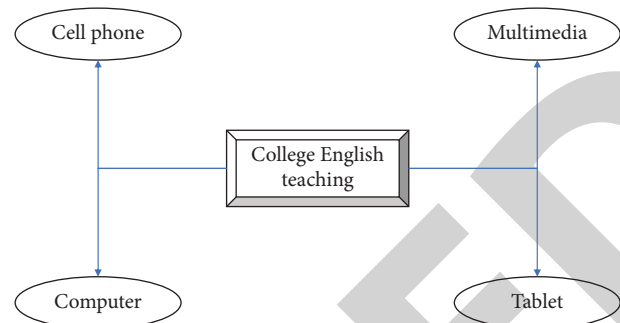


FIGURE 1: Teaching tools of college English teaching.

words, the original single plane teaching method has become more vivid and three-dimensional.

As shown in Figure 2, there are three main modes of teaching English in the university based on information technology environment. The first one is the group learning mode. This mode refers to the teacher's one-to-many teaching of students with the aid of IT equipment. This is also the most common mode of teaching English in college. To be specific, the teacher needs to teach the whole class through PPT and projection equipment [22]. The second mode is the individual learning mode, in which students choose and learn the course content independently through learning systems. The third is the blended learning model. This mode means that teachers use mobile devices as a medium to combine various learning modes such as group learning, independent learning, and social interactive learning in a hierarchical and organized manner through information technology to provide students with more diverse and rich learning options and learning styles [23]. The first and second modes of learning are the most commonly used in today's university English teaching. Their advantages are obvious, but their disadvantages cannot be ignored. In this context, a blended learning model that integrates mobile devices is particularly important [24]. Therefore, it is an important proposition to study the reform of college English teaching mode and promote the effective development of college English teaching by combining various teaching aspects such as group learning, independent learning, social sharing, and multiple evaluation in a hierarchical and organized way, and giving students a more free and efficient learning environment.

Fragmented mobile learning has been a hot topic in recent years in teaching and learning research. The original idea of this concept is to enable learners to selectively separate learning content effectively through the use of mobile tools that transcend the constraints of time and space [25]. This allows learners to freely choose and control the content and timing of their learning and to extract meaningful pieces of knowledge from the vast amount of information available. Fragmented learning is an adaptation of learning styles to the fast-paced nature of modern life. It gives great freedom and independence to learning and deeply meets the modern people's need for free, efficient, and fast learning [26]. For this reason, in recent years, mobile tools such as microblogs, WeChat, and mobile APPs have

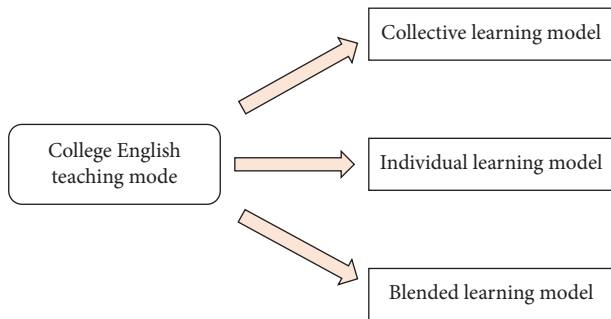


FIGURE 2: Mode of college English teaching.

been adapted to each other and jointly promoted fragmented learning. This can make fragmented mobile learning, i. e., learning in pieces through mobile tools, more popular [27]. The design of a blended university English teaching model cannot be achieved without the support of mobile tools. It allows learning to occur in fragments that are not limited to the classroom, thus giving students more freedom and choice. As a result, diversified network teaching environment has become a hot topic of research in college English classroom teaching.

Big data is ushering in a major transformation. Faced with the huge amount of information resources, people need to develop online learning behavior and habits with a new way of thinking [28]. The widespread use of mobile tools is profoundly changing the way people live and learn. The mobile learning and fragmented learning brought about by mobile tools make learning happen beyond the limitations of time and space, and moreover, it adapts to the learners' needs for learning autonomy and independence [29]. The immediacy and portability of mobile tools also provide new forms and carriers for the presentation and development of university English teaching resources. In other words, more teaching activities can be carried out and run with mobile tools as the scaffolding [30]. Therefore, combining mobile learning with university English courses is a choice that is in line with the trend of technological development and the development of the times. The application of mobile tools should be the key to the effective implementation of blended teaching and learning design and should be the focus of research on English classroom teaching.

What is more, the use of a series of intelligent technology products in the English classroom has enriched the content as well as the mode of English classroom teaching [31]. To a certain extent, this mode has also improved the efficiency of English classroom teaching. However, the use of technology also limits teachers' own development. As a result, how to balance the roles of intelligent technology and teachers in the classroom and how to realize the human-computer collaborative teaching mode are important issues faced by the technology-enhanced classroom teaching. In order to address these issues, this study designs and develops an intelligent English teaching assistant from the perspective of human-computer collaborative teaching to solve the problems of speech input and language environment creation in English teaching [32]. The assistant has the main functions of text-to-speech, automatic speech evaluation, and listening

materials creation. In addition, this study uses the intelligent teaching assistant to assist English teachers to carry out classroom teaching so as to solve a great number of problems in English teaching such as nonstandard speech input, promote the modernization and intelligence of English teaching, and realize the integration of artificial intelligence technology and English subject teaching.

2. Analysis of Learning Situation of College English Teaching

The design and implementation of the teaching process contains a great number of basic aspects and the synergy and influence of many factors need to be considered. As shown in Figure 3, in the process of teaching design and implementation, learners, teachers, teaching objectives, teaching contents, teaching media as well as some other factors are integrated in the whole teaching process. To be specific, they together form the basic structure of the teaching model. At the same time, the synergy of these factors ensures the proper implementation of teaching and learning. As a result, the discussion and analysis of these basic factors can facilitate the design and implementation of teaching and learning. In other words, it is only through a comprehensive and specific understanding of these basic factors that we can make proper judgment and design. In addition, the whole process of teaching and learning can be adjusted and modified according to what may happen.

2.1. Learner Analysis. The learner is the main subject of instructional design and the center of the entire teaching and learning process. In addition, the learner is the starting and ending point for all instructional activities and arrangements. The characteristics of learners influence and guide the development of instructional a result, the analysis of learners should be placed before teaching. In fact, the majority of the learners for college English teaching are young students. These students are influenced by social, media, and family factors, and they have outstanding characteristics in behavior, emotion, thinking, and psychological cognition. At the same time, the rapid development of social economy and science and technology has provided them with a huge amount of information and the means to obtain it. This information can cultivate their active thinking, open personality, and broad vision. As a result, they form their own opinions and values on many issues early and even form their own cultural habits.

Cognitive development is the process by which an individual's ways of acquiring knowledge and skills and their ability to solve related problems change in stages over time. The most important factors that are emphasized are time and mechanism. It can be said that cognitive development represents, in part, the way learners understand the outside world, their thinking patterns, and their ability to learn. These competencies make it possible to design more contextualized instructional activities. After all, learners have a clear understanding of their place in the learning situation and their responsibilities. In addition, they are able to adapt

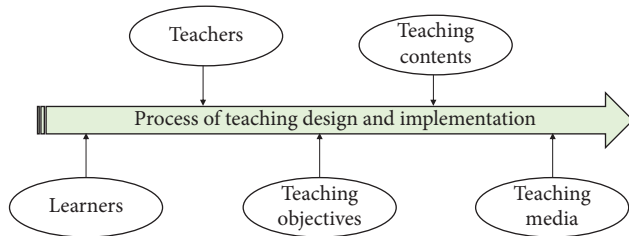


FIGURE 3: Factors in the process of teaching design and implementation.

more independently to changing learning processes free from the constraints of concrete experience. At the same time, they are able to find ways to adjust their own direction of inquiry as they work through problems. As a result, more cooperative and research-based learning can be applied in this process.

2.2. Teaching Objective Analysis. Teaching objectives are specific, explicit statements of the visible behaviors that learners are expected to exhibit as a result of instruction. The original purpose of instruction is to enable learners to change in the direction specified by the goal. The teaching objectives set the direction of the instructional activity and guide the entire instructional process. During the teaching process, the teacher can control and regulate each step in the process. At the same time, the accomplishment of the teaching objectives is a test of the design and practice of the teaching activities. To be specific, the teacher can identify the issues and make corrections by comparing the final teaching results with the objectives set. Figure 4 illustrates the categories of English training objectives in universities.

The content of a university English course is aggregated in the form of units, each of which is a specific topic. The various forms of content are organized under these themes and presented in detail under specific themes. Themes allow for the internal organization and connection of multiple forms and aspects of content. At the same time, these resources facilitate the learning of learners in various ways, such as visual, auditory, verbal, and practice, thus enriching the levels and forms of learning. Figure 5 shows the content structure of university English course units. Course materials are the basic resources for teaching English at the university level, and units are the basic organizational units of university English course materials. The learning content is organized and aggregated in the form of thematic units. As chapter-based and coherent teaching materials, the units of a university English course textbook have a similar organizational structure. However, the textbook is adapted and modified in specific chapters so that the teaching becomes unified as well as clear.

2.3. English Specific Skill. The starting point and ultimate goal of learning a language and culture is to be able to recognize and use it well in daily study and work, and this is also true for learning the English language. The accumulation of a large number of words, the practice of listening

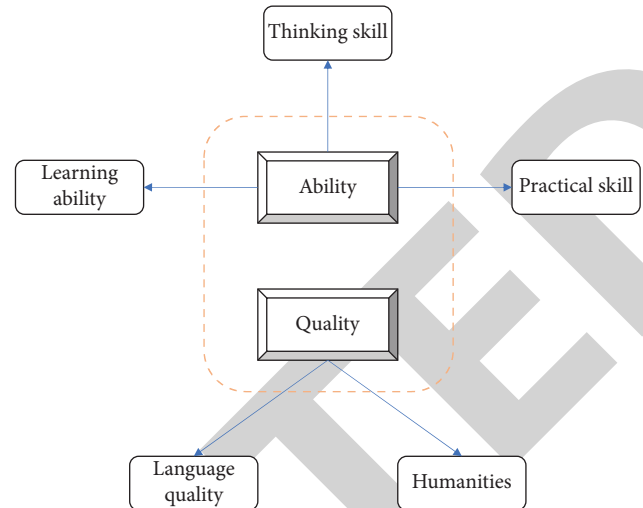


FIGURE 4: Categories of English training objectives in universities.

and speaking, and the memorization of relevant grammatical knowledge are all aimed at the smooth and efficient use of English for one's own benefit. As a result, it is essential to acquire certain English skills for daily life, study, and work, as well as for answering exams. To be specific, it enables people to achieve their goals in a faster and more efficient way. In addition, it can lead to a more skillful learning process in which people develop their own learning habits and find the most appropriate learning path for themselves. Figure 6 illustrates the specific structure of college English skills.

The composition of English language is organized and presented in the form of words, sentences, paragraphs, and chapters, and in this way, it is increased and expanded in number and length. English language learning skills are empirical rules that learners discover over time. Specifically, it is a broad and individualized set of skills that exist in all aspects of English language learning. The learning of English language, like the learning of other languages and cultures, can be developed and refined in multiple dimensions: listening, speaking, reading, writing, and translating. Therefore, the role of English language learning skills is to facilitate learners' effective and practical mastery and understanding of the formal content of words, sentences, paragraphs, texts, and chapters in the dimensions of listening, speaking, reading, writing, and translating.

3. Application of Diversified Network Teaching

The teaching design behavior mainly includes the design of teaching objectives, teaching methods, the development of curriculum resources, and collective lesson planning. Quality teaching resources are the basic guarantee of teaching effectiveness, and the design of English classroom requires the support of various media materials such as sound, text, and images. At present, English teachers mostly use PPT classroom materials to design teaching activities. Therefore, the production of PPT materials has become an inevitable part of every English teacher's lesson preparation,

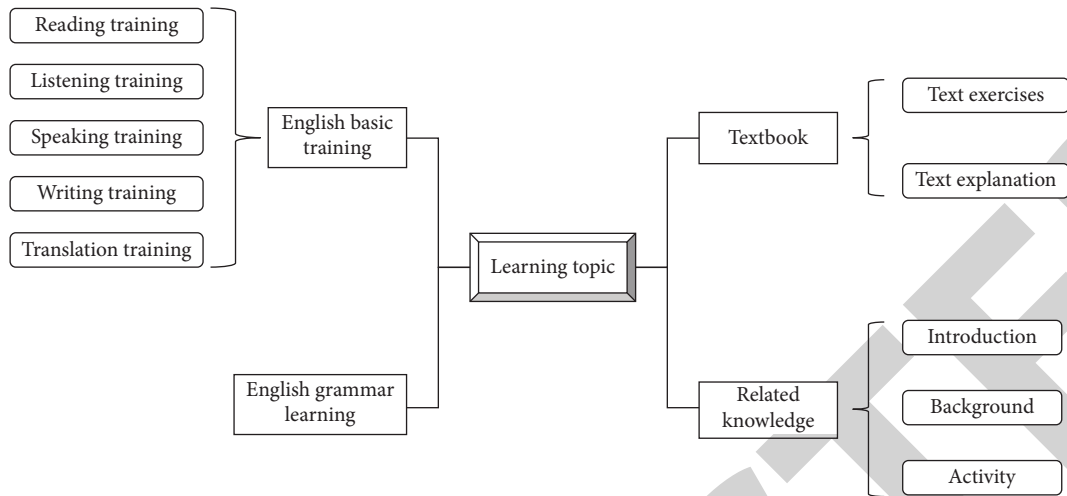


FIGURE 5: Content structure of university English course units.

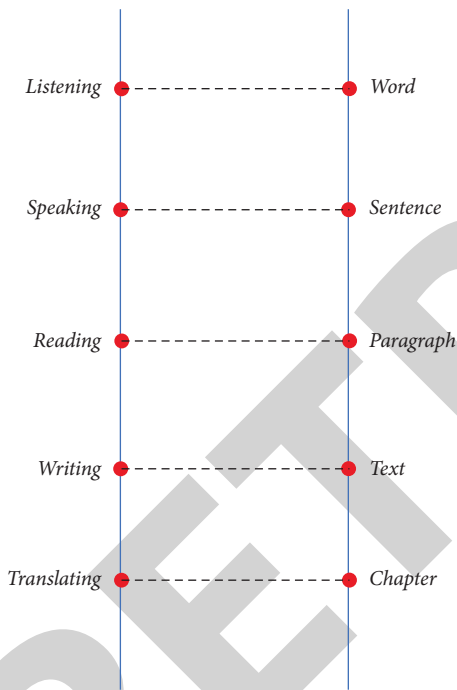


FIGURE 6: Specific structure of college English skills.

which undoubtedly brings great teaching pressure to English teachers. In traditional teaching, there are two ways for teachers to obtain audio materials. The first is the audio materials accompanying the textbook, and the second is the audio materials recorded by teachers themselves. The use of audio materials can ensure the quality of audio, but it lacks flexibility.

3.1. Network Structure Model Design. The C/S model assigns the system tasks to the client and server according to the rules and uses the cooperation between the client and server to accomplish the system tasks. The C/S model has the advantages of fast response time and strong security, but the development and maintenance costs are high, and it does not

have cross-platform capability. To be specific, it mainly applies the browser to act as the client and puts the main task processing into the server. The B/S model has the advantages of cross-platform, no client maintenance, and easy development. However, this model also has the disadvantages of slow response time and lack of personalization. From the practical point of view of practicality, intelligence, and convenience, the English subject intelligent teaching assistant is implemented in the B/S mode. The user can access the server through a common browser, and the client can realize the relevant functions. This model reduces the requirements for the client, and the core functions depend mainly on the server to run, thus reducing the need for hardware conditions. As a result, access to the assistant can be achieved by installing a browser on the running computer. This simplifies the development, use, and maintenance of the assistant and also greatly improves system compatibility. The overall structure design and architecture logic of the intelligent teaching assistant is shown in Figure 7.

The intelligent teaching assistant adopts B/S architecture. Therefore, teachers and administrators can access the system functions of Intelligent Teaching Assistant conveniently and quickly at any time and any place through the campus network. While completing intelligent teaching and management, it can also effectively ensure user information security. Users access the server through the browser, and the data exchange is done directly with the server. After that, the data in the database can be accessed through the server. This can effectively avoid the user's direct manipulation of the data in the database, thus making the user request relatively independent from the database. As a result, data security, reliability, and accuracy are guaranteed.

3.2. Module Design and Analysis. According to the concept of multimedia teaching, not only text but also graphics, illustrations, icons, video animation, audio, and other information media are needed in English teaching. In addition, it is necessary to use scientific technology to combine these media in a unified way. The design and creation of



FIGURE 7: Structure design and architecture logic of the intelligent teaching assistant.

courseware provides teachers with the ability to quickly design activities and present content visually. To be specific, this function can effectively combine multiple teaching media to achieve a uniform and fast presentation of teaching content. The intelligent courseware editing and creation function should reflect the teacher's dominant position. As an intelligent tool for collaborating with teachers to design teaching activities, the assistant can be used to create intelligent courseware easily and quickly by intelligent means. On the basis of this, it can have the intelligent voice integration function, which is not available in traditional slides so that teachers can free themselves from the heavy workload of creating courseware and focus on classroom organization and management, activity design, as well as teaching reflection.

The smart courseware editing feature is used by teachers to create multimedia web pages for classroom use. With the built-in visual web authoring tool, teachers can store the created pages in the cloud database. At the same time, they can use the browser to render the visual content of the web page when they use it. In this way, intelligent web pages can be created that provide a multidimensional display of English text in a randomized point-and-click manner. The intelligent voice courseware contains all the content that can be included in a web page. As a result, text, images, audio, video, tables, and other teaching materials can be inserted. In addition, it also has the functions of setting font and font size, setting image size, setting background color, and hyperlinks, etc.[33]

4. Conclusion

Unlike the traditional university English resource model, the university English teaching resource model in the context of multiple networks considers the relevant functions and characteristics of mobile tools. Specifically, it enables students to participate in the accumulation and improvement of teaching resources without the use of functions. As a result, teachers are no longer the only source of resource materials, and students are able to participate and interact more through the provision and sharing of resources. The flow of teaching resources is no longer a one-way process from teachers to students, but a mobile tool that allows more people to participate in the process of resource generation. The resource material is expanded and fed back in the cycle, and the release of relevant content can be the source of new feedback. As a result, resources are added in this circular pattern. At the same time, the resource structure is no longer only a medium for providing relevant materials but also a platform for mutual communication.

However, because there are few public resources, teachers still need to design and create their own intelligent

voice courseware although the Intelligent Teaching Assistant provides a public resource library. As a result, there are fewer resources that can be directly used by teachers. As AI technology advances and intelligent teaching continues to evolve, the impact on the English classroom will deepen. The researchers and developers are still facing serious challenges, and the optimization and improvement of the shortcomings in this study will be part of the author's future research work.

Data Availability

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

Conflicts of Interest

All authors declare that there are no conflicts of interest regarding this study.

References

- [1] W. Bing, "The college English teaching reform based on MOOC," *English Language Teaching*, vol. 10, no. 2, pp. 19–22, 2017.
- [2] C. Meng-yue, L. Dan, and W. Jun, "A study of college English culture intelligence-aided teaching system and teaching pattern," *English Language Teaching*, vol. 13, no. 3, pp. 77–83, 2020.
- [3] G. Wang, "On the application of cooperative learning in college English teaching," *International Education Studies*, vol. 13, no. 6, pp. 62–66, 2020.
- [4] Y. Shu, "Experimental data analysis of college English teaching based on computer multimedia technology," *Computer-Aided Design and Applications*, vol. 17, no. 2, pp. 46–56, 2020.
- [5] Y. Gao, "Computer-aided instruction in college English teaching under the network environment," *Computer-Aided Design and Applications*, vol. 18, no. 4, pp. 141–151, 2021.
- [6] S. Cui, "Research on across-cultural communication in college English teaching based on cloud platform," *Journal of Computational and Theoretical Nanoscience*, vol. 14, no. 1, pp. 89–93, 2017.
- [7] S. Kong, "Practice of college English teaching reform based on online open course," *English Language Teaching*, vol. 12, no. 5, pp. 156–160, 2019.
- [8] H. Fu, J. Niu, Z. Wu et al., *Influencing Factors of Stereotypes on Wastewater Treatment Plants-Case Study of 9 Wastewater Treatment Plants in Xi'an, China*, pp. 1–10, Environmental Management, Houston, Texas, 2022.
- [9] Y. Zhang and L. Zuo, "College English teaching status and individualized teaching design in the context of mobile learning," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 12, p. 85, 2019.

Retraction

Retracted: Home Care System for Mobility Disabilities Based on Intelligent Perception

Computational Intelligence and Neuroscience

Received 26 September 2023; Accepted 26 September 2023; Published 27 September 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] Y. Qiu and C. Qiu, "Home Care System for Mobility Disabilities Based on Intelligent Perception," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9528046, 9 pages, 2022.

Research Article

Home Care System for Mobility Disabilities Based on Intelligent Perception

Yanyan Qiu and Chunbo Qiu 

The Affiliated Hospital of Medical School, Ningbo University, Ningbo 315000, China

Correspondence should be addressed to Chunbo Qiu; qiuchunbo0311@163.com

Received 18 July 2022; Revised 29 August 2022; Accepted 6 September 2022; Published 30 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Yanyan Qiu and Chunbo Qiu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to explore how to realize home care for the elderly with mobility difficulties, this paper proposes a home care system for the elderly with mobility difficulties based on intelligent perception. This method explores the research of home care for mobility disabilities by recommending key technical problems and solutions based on information represented by intelligent perception. The research shows that the home care system based on intelligent perception can effectively solve the nursing problems of the elderly, which is about 60% more efficient than the traditional methods. The combination of intelligent perception and reasonable home care mode will improve the social and economic benefits of health services and promote the balance between supply and demand of the whole health services.

1. Introduction

The 21st century is an age of aging population [1]. Population aging has brought a profound impact on the global political, economic, social, and cultural development, and the pressure of pension, medical care, social services, and other aspects of a huge elderly group is increasingly prominent. In the face of limited medical resources and rapidly increasing medical costs, the problem of healthy aging of the elderly has become an urgent topic for the elderly society to study and solve [2]. Since Switzerland put forward the concept of local aging in the 1960s, the strategies of developed countries to cope with aging no longer focus on institutional care but have launched community-based and home-based services, which is also the way of elderly services emphasized and supported in the United Nations principles for the elderly and the United Nations Declaration on aging. With the arrival of the “silver wave,” China has also begun to pay attention to the special service field of home care, and the development of home care has increasingly important practical significance [3].

Home care is an important support for the home-based elderly care system [4]. China’s population aging shows

three characteristics: the largest absolute number of elderly people, the fastest aging rate, and “getting old before getting rich” [5]. At present, China’s economic conditions are not rich, and the social security system is not perfect. The state cannot invest a lot of money to set up a large number of nursing homes for the elderly, nursing homes for the elderly, rehabilitation centers, and other medical and healthcare institutions to centrally manage such a large-scale elderly population. The vast majority of the elderly will still stay in communities and families. Due to the rapid development of population aging and the dual impact of family planning policy, the trend of “four two one” family and family size miniaturization is irreversible. The proportion of “empty nest elderly families” has gradually increased, the pace of work of the children of the elderly has accelerated, the distance of living apart has widened, and the family care function has gradually weakened. The traditional family pension is not enough to bear all the pension responsibilities. “Home-based elderly care” is an important way to solve the problem of elderly care in the new era, and it is valued by governments at all levels and supported by the elderly [6].

Carrying out family nursing is an inevitable requirement to adapt to medical reform [7]. Limited medical resources

and soaring medical costs are common problems faced by today's aging society. The world is actively pursuing medical system reform to seek effective ways to meet the health needs of the elderly and reduce medical costs. It has become an important measure for the reform of the medical system to shorten the average length of stay, improve the turnover rate of hospital beds, and reasonably apply the limited manpower, material resources, beds, and medical equipment to the key treatment of critical, urgent, and severe patients. A large number of stable patients and chronic patients treated by the hospital will receive follow-up treatment and rehabilitation care at home and in the community, which requires the community to provide nursing services that are seamlessly integrated with the treatment and rehabilitation of diseases outside the hospital, so as to ensure the smooth progress of the reform of the medical system. Carrying out community family nursing service is an inevitable requirement for deepening medical reform and developing community health services. Therefore, carrying out home care for the elderly has become an inevitable choice for China to promote the benign operation of the home-based elderly care system, meet the health needs of the elderly, comply with the reform of the medical system and the development needs of nursing itself, and alleviate the pressure of population aging. Intelligent perceptual home care system is a very good choice and the future development trend [8]. The specific process is shown in Figure 1.

2. Literature Review

Zhang and others said that in 1980, the Joint Committee for the evaluation of health institutions in the United States clearly formulated the third standard of nursing services: the nursing department should clearly formulate a method to determine the nursing needs of patients in order to provide appropriate nursing measures [9]. Therefore, many hospitals began to adopt patient classification system to improve the quality of nursing, and the management method of patient classification has developed rapidly. Legido Quigley et al. define patient classification as classifying patients according to the care they need during a specific period of time [10]. Szatmári and Hoffman emphasize that the whole process of patient classification is to predict what kind of individual care patients should receive in a specific time [11]. Patient classification is mainly realized through a patient classification evaluation system, that is, a set of evaluation tools. The patient classification and evaluation system include a broader and more complex meaning than patient classification. It is a framework that organizes patients by evaluating the relevant indicators of patients' nursing needs and carries out classification management according to the level of nursing needs.

Miao et al. believe that "patient classification evaluation" is the most sensitive method to evaluate the type, degree, and prognosis of patients' home care needs [12]. It can systematically evaluate the patient's condition, so as to correctly monitor and determine the patient's individualized needs. In terms of nursing management, "patient classification assessment" helps to establish the manpower allocation

standard for caring for patients, plan and predict the required nursing manpower, and calculate the complexity and time of nursing measures.

Carrying out family nursing is the need of nursing self-development. The core of the development of nursing is to meet the people's growing demand for health services, take health as the center and demand as the guide, constantly innovate the mode of nursing services, expand the connotation of work, and vigorously develop elderly care, chronic disease care, hospice care, and other nursing services based on communities and families, so as to meet the needs of social development. Carrying out family nursing is one of the most effective ways to promote nursing work out of the hospital, into the community and into the family. It is also the need of the development of nursing in the aging society.

Ji et al. put forward and gave the concept of artificial neural network and the mathematical model of artificial neuron, thus creating an era of ANN research [13]. The early ANN did not have the ability to extract complex features, and the parameters needed to be manually adjusted, which violated the requirements of "intelligence." Therefore, in the process of continuous evolution and improvement, the MLP model with multiple hidden layers was proposed. At the same time, researchers used the back-propagation algorithm to optimize network parameters and basically formed the prototype of deep learning. MLP is composed of input layer, hidden layer, and output layer. MLP does not specify the number of hidden layers, so the appropriate number of hidden layers can be selected according to each demand. The combination of artificial neural network and home care can greatly improve the happiness index and nursing problems of the elderly.

Home care is an evolving concept. Wu et al. put forward the concept of home care. At that time, he hired nurses to take care of his sick wife at home, which made him realize the importance and necessity of the nursing staff's family visit service [14]. With the help of Nightingale, he founded a district visiting nursing institution, began to train full-time visiting nurses, established a family visiting nursing system, and initiated family nursing activities to help the poor improve their health conditions, laying the foundation of modern family nursing. Xiao and others said that visiting nursing activities began to appear in the United States, Canada, the Netherlands, Australia, Germany, and other countries [15]. In 1885, the first visiting nurse association was established in the United States. By the 1960s, the visiting nurse association had developed in the whole country. In 1965, the United States formulated and passed family nursing regulations. After hundreds of years of development, developed countries in Europe and the United States have formed a relatively complete family nursing service system with their own characteristics.

3. Method

3.1. Long-Term and Short-Term Memory Units. As for the structure of long-term and short-term memory (LSTM) units, "e" represents the addition of vector corresponding

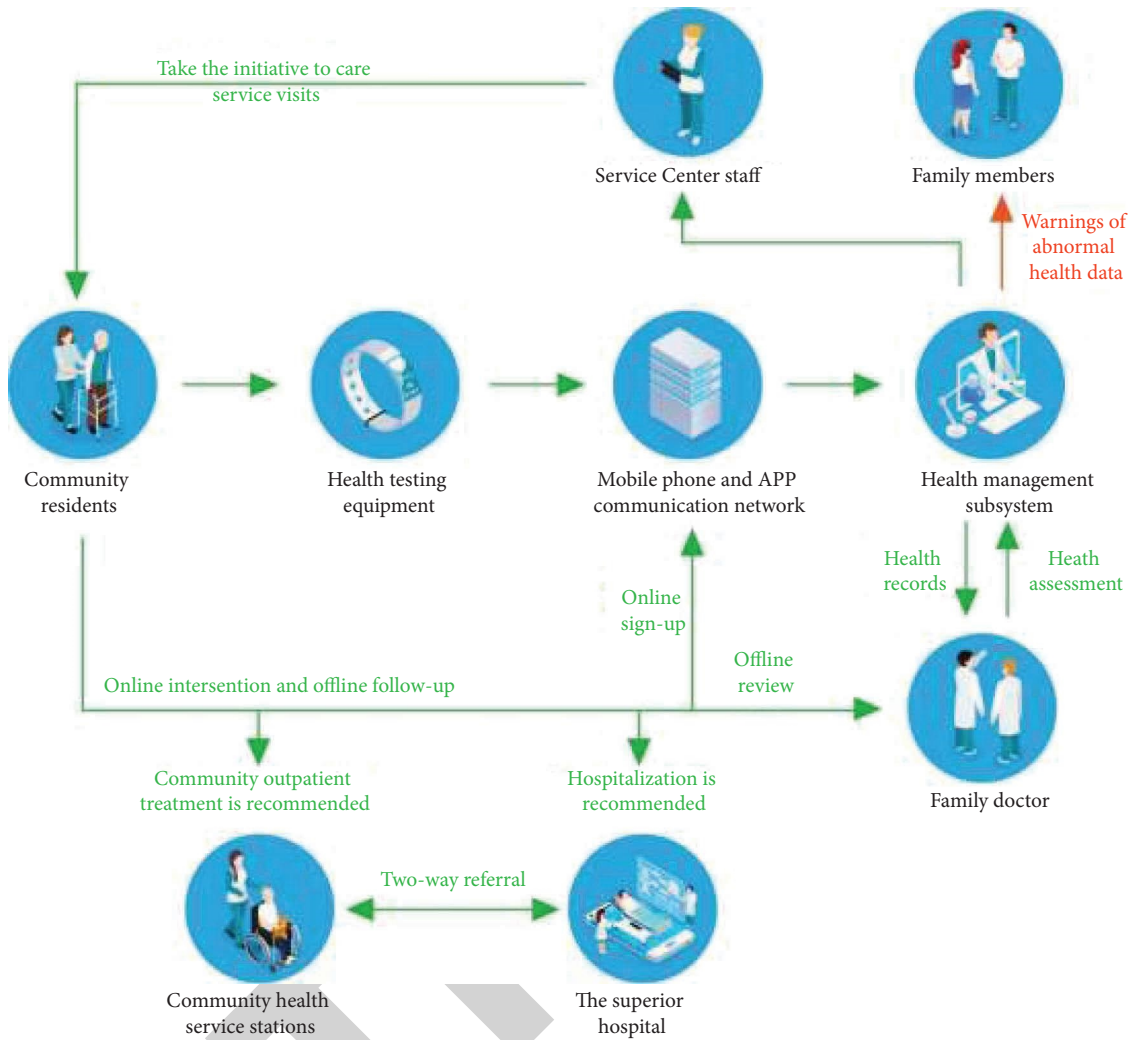


FIGURE 1: Intelligent induction home care system.

elements [16]. The LSTM cell contains a memory cell module and three gate structures that run through the whole cell. The memory cell module provides a long-term path of gradient propagation for the LSTM cyclic neural network, thus alleviating the problem of gradient disappearance in RNN [17]. The gate structure is responsible for protecting and controlling the state of memory cells. Each gate structure learns the mode of information screening from the input of the current time through its own weight matrix and sigmoid activation function. Specifically, the LSTM unit includes a forgetting gate, an input gate, and an output gate. The forgetting gate determines how much information in the previous memory cell state is discarded, the input gate controls how much input information at the current time can be added to the memory cell state, and the output gate determines which information in the updated memory cell state can be used as a hidden state. It should be noted that the input information controlled by the input gate comes from the external input at the current time and the hidden state at the previous time. The information to be output determined by the output gate is the result of processing the state of memory cells [18].

Assuming that the current time is t , f_t , i_t , and o_t , respectively, represent the status of the forgetting gate, input gate, and output gate at the current time, their calculation process is shown in the following formulas:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (1)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad (2)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o). \quad (3)$$

After defining the function of the three gate structure, we update the state of memory cells. First, we use the forgetting gate to process the memory cell state at the previous moment and then add the obtained cell state to the input information controlled by the input gate, so as to obtain the updated memory cell state. The mathematical expression is shown in the following formula:

$$c_t = f_t \odot c_{t-1} + i_t \odot z_{t1}, \quad (4)$$

where z_t is the input information, and the definition is shown in the following formula:

$$zt = \phi(Wc[h_{t-1}, xt] + bc). \quad (5)$$

Finally, the updated state of the output gate and the memory cell is calculated by Hadamard product to obtain the hidden state of the LSTM at the current time, as shown in the following formula:

$$h_t = o_t \odot \phi(c_t). \quad (6)$$

Assuming that the distribution of the real data is P_{data} (z), the generator maps the noise Z to the generated distribution P_G . Its purpose is to hope that the mapped generated distribution is as close as possible to the real distribution. Therefore, we can get the optimization problem of the generator, as shown in the following formula:

$$G^* = \operatorname{argmin} \operatorname{Div}(P_G, P_{\text{data}}). \quad (7)$$

Although we do not know the calculation formula of P_G and P_{data} , we can take some samples from these two distributions and use these samples to describe the distribution. For P_{data} , we can sample from a given data set. For P_G , we get the output samples of the generator by inputting randomly sampled noise into the generator. With these samples, the distance between P_G and P_{data} can be calculated through the discriminator. The optimization of the discriminator is shown in the following formula:

$$D^* = \operatorname{argmin} V(D, G), \quad (8)$$

where, as shown in the following formula,

$$v(G, D) = E_{r \sim P_{\text{data}}} [\log D(x)] + E_{r \sim P_G} [\log(1 - D(x))]. \quad (9)$$

Therefore, the optimization goal of the discriminator is to maximize $V(G, D)$ for a given generator. Let $D(z)$ be the optimization variable and directly derive $V(G, D)$ to obtain d^* as shown in the following formula:

$$D^* = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)}. \quad (10)$$

When the state action transition probability is known, the optimal strategy can be obtained by dynamic programming according to the Behrman equation as shown in the following formulas:

$$v^\pi(s) = \sum_{a \in A} \pi(a|s), \quad (11)$$

$$Q^\pi(s) = r(s, a). \quad (12)$$

However, in most cases, the state transition probability is unknown, so the time difference method can be used to estimate the value function. Here, we need to introduce the Behrman optimal equation first. For all $s \in S$, the corresponding action value function Q^* of the optimal strategy satisfies the Behrman optimal equation, as shown in the following formula:

$$Q^*(s, a) = E[r(s, a)]. \quad (13)$$

3.2. Meaning of Home Care Needs of the Elderly. The home nursing needs of the elderly are a kind of health needs, and it is the starting point of carrying out home nursing work. Health demand refers to the amount of health services that people are willing and able to purchase at a certain price level within a certain period of time [19]. Requirements can be divided into two categories: one is requirements transformed from needs. Only when people's health needs are transformed into needs, can they have the behavior of seeking medical and health-care services, and it is possible to use health resources. Health needs are caused by the gap between the actual health level and the "ideal health level," including the health needs recognized by individuals and the health needs determined by medical experts. The two are sometimes consistent, as shown in Table 1.

Grading of the Japanese nursing service system is as follows: the elderly who need nursing services are divided into six grades according to their daily living ability, and different grades represent different levels of nursing needs of the elderly. Nursing institutions provide a corresponding amount of nursing services and funds according to the level of the elderly as shown in Table 2.

Behavioral ability (including activity ability, clinical performance, and working ability) is an important index used to evaluate the patient's health level, the impact of disease on the patient's life, and guide treatment and rehabilitation [20]. The most commonly used scale for assessing behavioral competence is Karnofsky performance status (KPS) as shown in Table 3.

The evaluation of behavior ability is mainly based on the doctor's observation and understanding of the patient, rather than the specific investigation of the patient's health status. The evaluation process and rating are simple and fast, without lengthy evaluation items. Behavior ability evaluation is particularly suitable for guiding the treatment needs of cancer patients and predicting prognosis. Classification of Taiwan's home care system: patients are classified into five grades according to their self-care ability, waking time, and range of activities. Grading is mainly used to evaluate the qualifications and service needs of home care applicants and set the entry and exit criteria. Applicants with grade 3 and above are the objects of home care [21] as shown in Table 4.

Some experts divided the comprehensive health function of the elderly into six grades from the five aspects of social resources, economic resources, physical health, daily living ability, and mental health and reflected the health status of the elderly with the grade level of different dimensions, as shown in Table 5.

It is not difficult to see that physical health is an important aspect of reflecting the health level and an important condition factor affecting the self-care strength of the elderly. It directly determines the self-care needs of the elderly when their health is poor and is of great significance for evaluating the needs of professional family nursing services. However, for the elderly, a special group, their self-care ability is no longer determined by a single physical health, and it is often difficult to identify their self-care ability only by disease-related physical health assessment [22]. Who proposed that the ability to live independently should be the main index to evaluate the health

TABLE 1: Determination of health needs of individuals and medical experts.

Medical experts	Individual	
	Have health needs	No health needs
Have health needs	A	B
Medical experts	C	D

TABLE 2: Health classification of patients in Japanese nursing care service.

Grade	Health
Need help	Basically have the ability of daily living, but some help is needed for bathing, etc
Level 1 care	Unstable standing or walking, excretion, bathing, etc. need some help
Level 2 care	Difficulty in standing, excretion, etc. need help
Level 3 care	Unable to stand, turning over, excreting, bathing, dressing, etc. need help completely
Level 4 care	Rely on others for excretion, bathing, dressing, etc
Level 5 care	Rely on others in daily life

TABLE 3: Karnofsky performance status,KPS.

Behavioral capacity status	Level
Normal, without symptoms and signs	100
Able to perform normal activities, with mild symptoms and signs	90
Barely able to perform normal activities, with some symptoms or signs	80
You can take care of yourself, but you cannot maintain normal life and work	70

TABLE 4: Coriolis scale.

Behavioral capacity status	Level
Fully mobile without any restrictions	0
Able to walk and maintain light work, such as simple housework, but limited to physical exertion activities	1
Able to walk and maintain self-care, but unable to work or housework	2
More than 50% of the time when you are awake, you can get up and do not have to limit your activities to bed or chair	3
Can only maintain limited self-care, more than 50% of the waking time, activities are limited to bed or chair 3, completely unable to move, cannot take any self-care, and completely limited to bed or chair	4

level: the physical function status can better reflect the health status, health services, and social service needs of the elderly. Therefore, to evaluate the self-care strength of the elderly, identify the degree of their self-care defects, and explore their home care needs, we cannot emphasize physical health or physical function in isolation. The two are equally important and complementary, and neither is indispensable.

It is impossible for any expert to be authoritative on every problem in the prediction, and the authority or not has a considerable impact on the reliability of the evaluation [23]. Therefore, when dealing with the evaluation results, we must consider the authority of experts on a certain issue. The authority of experts is generally determined by two factors: one is the basis for experts to judge the scheme, which can be expressed by the judgment coefficient C_i . The second is the familiarity of experts with the problem, which is expressed by the familiarity coefficient C_s . These two indicators deserve to be obtained, mainly through expert self-evaluation. The degree of authority is the arithmetic mean of the judgment coefficient and the familiarity coefficient as shown in the following formula:

$$C_a = \frac{(C_i + C_s)}{2}. \tag{14}$$

In the judgment basis, we provide four questions. Please make a self-evaluation by the expert. The results are shown in Table 6.

According to the self-evaluation standard coefficient based on expert judgment, the coefficient and $C_i = 1.0$ have the greatest impact on expert opinions, $c_i = 0.8$ has the middle impact, and $C_i = 0.6$ has the least impact, calculate the arithmetic mean of the sum of all experts' self-evaluation, as shown in Tables 7 and 8.

C_i is calculated as shown in the following formula:

$$C = \sum M_j W_j I_j M = (0.7 \times 1 + 0.8 \times 6 + 0.9 \times 7 + 1.0 \times 2) \div 16 = 0.8625. \tag{15}$$

The self-assessment of experts' familiarity mainly involves two aspects, as shown in Table 9.

TABLE 5: Physical health level.

Excellent: often or occasionally engage in intense sports
Good: no obvious disease or disability, only receive routine medical care, such as physical examination once a year
Mild disorder: only mild disease or disability, which can be improved by treatment or corrective measures
Moderate disorder: one or more disorders, if not very painful, require continuous medical care
Severe disorder: there are one or more kinds of disabilities. The injury is either very painful and life-threatening, or it needs extensive treatment
Completely disabled: lying in bed and requiring 24-hour medical assistance or nursing care to maintain life

TABLE 6: Judgment basis and quantification of influence degree.

Judgment basis	Impact on expert judgment		
	Large	Middle	Small
Understanding of the current situation of community home nursing in China	0.3	0.2	0.1
Understanding of the current situation of community home care in the United States	0.1	0.1	0.1

TABLE 7: Frequency of self-evaluation of expert judgment basis.

Judgment basis	Large	Middle	Small
	Frequency (%)	Frequency (%)	Frequency (%)
Understanding of the current situation of community home nursing in China	9 (56.25)	7 (43.75)	0 (0)
Understanding of the current situation of community home care in the United States	2 (12.50)	10 (62.50)	4 (25.00)

TABLE 8: Distribution of the number of experts based on self-evaluation.

$C_i = 1.0$	2
$C_i = 0.9$	7
$C_i = 0.8$	6
$C_i = 0.7$	1

The experts' familiarity with the assessment issues is shown in Table 10.

C_s , C_a are calculated as shown in the following formulas:

$$C_s = \sum M_j W_{j1} M = (0.9 \times 12 + 0.7 \times 4 + 0.5 \times 0 + 0.3 \times 0) / 16 = 0.85, \quad (16)$$

$$C_a = (C_i + C_s) / 2 = (0.85 + 0.8625) / 2 = 0.86. \quad (17)$$

General experts believe that $C_a \geq 0.70$ is an acceptable reliability. Therefore, from the results of the expert authority evaluation of this consultation, the 16 experts have a high degree of authority on the content of this evaluation, and the results are credible [24].

The enthusiasm coefficient of experts represents the degree of concern and cooperation of experts in this study, which is the key issue of expert consultation.

The degree of coordination of expert opinions refers to whether there are major differences in the evaluation opinions given by all experts on all indicators. It is expressed by the coordination coefficient W , and the value of W is 0~1. The greater the W , the better the degree of coordination. The coordination coefficient of the two rounds of consultation is shown in Table 12. After inspection, the coordination coefficient is significant, indicating that the coordination of

expert opinions is good, and the evaluation results are acceptable.

4. Results and Analysis

The prevalence rate of chronic diseases is the ratio of the number of chronic diseases surveyed in the first half of the survey to the total number of people surveyed, mainly reflecting the health status of the elderly [25, 26]. The diagnosis of chronic diseases in this study is based on the outpatient medical records or discharge diagnosis sheets of the elderly in addition to the doctors' inquiry of the respondents. The survey shows that the prevalence of chronic diseases in the elderly is high, among which the top five chronic diseases are hypertension (58.2%), coronary heart disease (31.0%), cataract (26.8%), stroke (21.8%), and diabetes (19.0%), while the prevalence of hyperlipidemia (18.6%), lung disease (11.8%), arthritis (10.6%), cervical spondylosis (10.0%), osteoporosis (9.6%), and Alzheimer's disease (5.0%) is also relatively high as shown in Table 13.

According to the distribution of the number of elderly patients at all ages, the trend of elderly patients at all ages was analyzed by SAS software (Cochran-Mantel-Haenszel Statistics), $X^2_{cmh} = 23.26$, $P < 0.001$. It can be seen that the difference in the number of elderly patients at all ages is significant. With the increase of age, the proportion of elderly patients suffering from multiple diseases at the same time gradually increases.

According to the extraction results of principal component analysis, six factors/principal components with eigenvalues greater than 1.0 are extracted, and their cumulative contribution to the total variance is 65.711%. Some experts pointed out that only when the cumulative variance contribution rate of public factors reaches at least 40%, it can be considered that the scale has a good structural validity of 16%.

TABLE 9: Frequency of self-evaluation of experts' familiarity.

Index	Familiarity frequency (%)	Familiar frequency (%)	General frequency (%)	Less familiar frequency (%)	Frequency of not knowing (%)
Nursing operation technology	14 (87.5)	2 (12.5)	0	0	0
Health status of the elderly	10 (62.5)	6 (37.5)	0	0	0

TABLE 10: Expert familiarity self-assessment results.

Score	0.9	0.7	0.5	0.3	0.1
Number of experts	12	4	0	0	0

TABLE 11: Responses to two rounds of expert consultation.

	Total number of experts	Number of responding experts	Response rate RR (%)
First round	16	16	100
Second round	16	16	100

TABLE 12: Coordination degree of expert consultation of family nursing service project.

Index	Number of indicators	Kenda11's W	X^2	P value
First round	96	0.644	979.323	<0.00
Second round	114	0.866	1566.126	<0.00

TABLE 13: Distribution of the number of elderly patients.

Number of sickness	Number of people	%	Cumulative percentage
0	15	3.0	3.0
1	91	18.2	21.2
2	174	34.8	56.0
3	100	20.0	76.0
4	72	14.4	90.4
5+	48	9.6	100.0
Total	500	100	

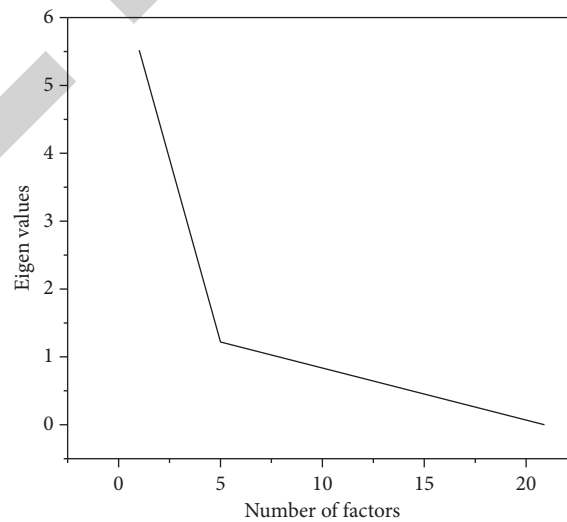


FIGURE 2: Factor number and eigenvalue steep slope diagram.

Therefore, it can be considered that the structural validity of the elderly health function assessment scale is good. It is also determined by the decreasing variation of factors in the steep slope map, which is also 6 factors, as shown in Figure 2.

Finally, the component matrix analysis after the rotation axis is carried out to investigate the level items contained in the common factors. After average orthogonal rotation, as a result, only factor coefficients greater than 0.4 are listed.

5. Conclusion

With the aging of the population, China's health supply pressure is increasing. In order to further rationally allocate health resources for the elderly, the government has placed community health services in an important strategic position. Family nursing, as an important form of community health services, plays a vital role in meeting the growing needs of elderly health services. A reasonable family nursing model will improve the social and economic benefits of health services and promote the balance between supply and demand of the whole health services.

At present, the national policy background of vigorously developing community health services and the gradual promotion and improvement of community health service mode have built a good platform for the research and development of family nursing services. We will take this opportunity to focus on the research of the family nursing service model. The research on family nursing service mode is a service management system project with the characteristics of the complex system, difficult implementation, and periodic application. The research on family nursing needs assessment and service content is one of the important parts of the research on the family nursing service model. The combination of intelligent perception technology and home care service mode can make the elderly life more convenient and simple in their later years, which is the development trend of home care in the future.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

- [1] A. Collins, J. E. H. Brown, J. Mills, and J. Philip, "The impact of public health palliative care interventions on health system outcomes: a systematic review," *Palliative Medicine*, vol. 35, no. 3, pp. 473–485, 2021.
- [2] S. Sarma, D. H. S. Lamkuche, and D. E. C. Blessie, "Securing communication in the iot- based power constrained devices in health care system," *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, no. 7, pp. 115–121, 2021.
- [3] C. Yi and X. Feng, "Home interactive elderly care two-way video healthcare system design," *Journal of Healthcare Engineering*, vol. 2021, no. 1, 11 pages, Article ID 6693617, 2021.
- [4] M. Sango and Y. Nakamura, "Role of medical institutions responsible for home care in the medical care provision system for intractable neurological diseases," *Brain and nerve = Shinkei kenkyū no shinpo*, vol. 72, no. 8, pp. 893–899, 2020.
- [5] M. Jelonek, T. Herrmann, M. Ksoll, and N. Altmann, "Ethnographically derived socio-technical analysis for information system support in intensive home care," *Complex Systems Informatics and Modeling Quarterly*, vol. 38, no. 22, pp. 1–20, 2020.
- [6] M. Gülşen and S. Arslan, "Home care in gastrointestinal system-oriented cancer surgery," *Türkiye Klinikleri Journal of Internal Medicine*, vol. 6, no. 1, pp. 44–48, 2021.
- [7] M. Lussier, M. Couture, M. Moreau et al., "Integrating an ambient assisted living monitoring system into clinical decision-making in home care: an embedded case study," *Gerontechnology*, vol. 19, no. 1, pp. 77–92, 2020.
- [8] S. Kim and P. H. Jun, "Comparison of children's subjective well-being by types of out-of-home care system: focusing on foster care and residential care," *Journal of Life-span Studies*, vol. 10, no. 1, pp. 1–18, 2020.
- [9] G. Zhang, Z. Mei, B. Lo et al., "A non-invasive blood glucose monitoring system based on smartphone ppg signal processing and machine learning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 7209–7218, 2020.
- [10] H. Legido-Quigley, L. Otero, D. L. Parra, C. Alvarez-Dardet, and J. M. Martin-Moreno, "Re: will austerity cuts dismantle the Spanish healthcare system?" *BMJ*, vol. 346, no. 24, pp. 1322–1324, 2020.
- [11] A. Szatmári and I. Hoffman, "The transformation of the municipal social care system in Hungary -in the light of the provision of home care services," *Lex localis - Journal of Local Self-Government*, vol. 18, no. 4, pp. 691–712, 2020.
- [12] Y. L. Miao, W. F. Cheng, Y. C. Ji, S. Zhang, and Y. L. Kong, "Aspect-based sentiment analysis in Chinese based on mobile reviews for bilstm-crf," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 5, pp. 8697–8707, 2021.
- [13] W. Ji, C. Chen, G. Xie, L. Zhu, and X. Hei, "An intelligent fault diagnosis method based on curve segmentation and svm for rail transit turnout," *Journal of Intelligent and Fuzzy Systems*, vol. 41, no. 1, pp. 1–11, 2021.
- [14] Q. Wu, Q. Feng, Y. Ren, Q. Xia, Z. Wang, and B. Cai, "An intelligent preventive maintenance method based on reinforcement learning for battery energy storage systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8254–8264, 2021.
- [15] Z. Xiao, X. Fu, L. Zhang, and R. S. M. Goh, "Traffic pattern mining and forecasting technologies in maritime traffic service networks: a comprehensive survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1796–1825, 2020.
- [16] M. Wu, W. Wang, and R. Han, "Dynamic monitoring system of high slope of rock and soil based on internet of things perception," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 5953–5962, 2021.
- [17] P. S. Kuzmin, "Smart metering systems: An empirical analysis of technology perception factors," *Strategic Decisions and Risk Management*, vol. 12, no. 1, pp. 8–23, 2021.
- [18] H. Wang and A. Li, "A systematic approach for English education model based on the neural network algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 1, pp. 1–12, 2020.
- [19] H. Lian, X. Pei, and X. Guo, "A local environment model based on multi-sensor perception for intelligent vehicles," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 15427–15436, 2021.

Retraction

Retracted: On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Ding, "On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7183032, 11 pages, 2022.

Research Article

On the Intelligent Computing Model of Diagnosis Teaching in Preschool Education in Colleges and Universities under the Background of Big Data

Xiaoqiong Ding 

Yunnan College of Business Management, Kunming City, Yunnan Province 650032, China

Correspondence should be addressed to Xiaoqiong Ding; xiaoxiangyu@hbut.edu.cn

Received 29 July 2022; Revised 31 August 2022; Accepted 6 September 2022; Published 30 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Xiaoqiong Ding. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to infer the cognitive state of students and provide teachers with the potential learning state of students, a diagnostic teaching model for preschool education in colleges and universities under the background of big data is proposed. By increasing students' programming ability and modeling students' theoretical and practical abilities at the same time, the cognitive diagnosis is introduced into the field of computer teaching, so as to make it applicable to computer classrooms and provide students' cognitive information needed for teaching. The experimental results show that the advantages of the CDF-CSE approach gradually emerge as the training data become sparse (the proportion of training data decreases from 80% to 20%). In the combined questions of the three datasets, when the training set is 20% and MAE is used as the criterion, the CDF-CSE model improves by 47.8%, 65.8%, and 49.8%, respectively, compared with the other methods that perform best on the training set. When the number of questions is small, the CDF-CSE model improves by 37.8%, 42.5%, and 27.7% on RMSE on three datasets, respectively, compared with the best-performing other methods. When there are more questions, it has 32.3%, 36.5%, and 45.6% improvement, respectively. It is concluded that this model can accurately predict students' performance in computer courses and provide detailed and rich cognitive reports.

1. Introduction

With the comprehensive popularization of Internet technology, the 21st century has entered an era of information explosion. Information technology has fully entered human social life, and informatization has become the general trend of the development of the current era [1, 2]. In this context, all walks of life have accumulated a large amount of available data, and it has become an inevitable trend to improve and upgrade related industries through the mining and application of these data. With the gradual advancement of informatization in the field of education, various intelligent teaching assistance platforms have emerged one after another. Due to their massive resources and learning characteristics independent of time and space, these teaching assistance platforms have developed rapidly and attracted a large number of users. At present, the intelligent teaching

assistance system is gradually becoming a mainstream education method in the knowledge era, which has gradually become an ideal learning environment for learners to carry out collaborative learning, knowledge construction, and wisdom development. As a new type of modern education platform, the intelligent teaching assistance system has completed the transformation from "teacher-centered" to "student-centered." On the one hand, the intelligent teaching assistance system is free from the constraints of time and space and has the advantages of autonomy and openness. On the other hand, the current teaching assistance system is still based on the knowledge-infusion teaching method, similar to traditional education, which is composed of learning materials.

At present, the research on educational assistive technology for educational examination data has attracted the gradual attention of researchers in the fields of education

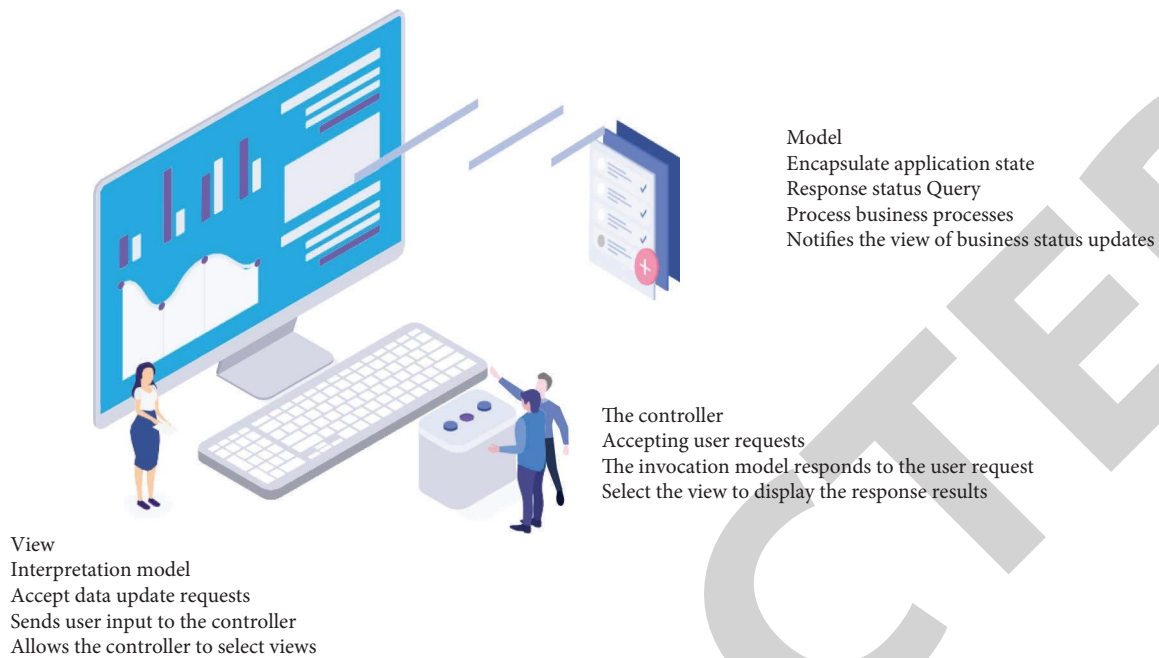


FIGURE 1: Auxiliary education model.

and data mining, but the current educational assistive technology still faces great challenges [3]. On the one hand, effective educational assistance is based on adequate and accurate understanding of students, that is, a comprehensive analysis of students' learning ability and cognitive level is required. Therefore, student modeling has become an important direction and research basis of educational auxiliary algorithms. On the other hand, on the basis of student modeling, intelligent auxiliary technologies for the educational process should also be studied and improved, such as the automatic construction of collaborative learning groups and the prediction of students' performance trends, as shown in Figure 1.

2. Literature Review

Miulescu et al. developed a cognitive diagnosis model of disciplinary knowledge ability based on the Bayesian model test method. The purpose of using the cognitive diagnosis model was to achieve different results according to the DINA model data for different structures. Differentiation between items could also be analyzed using this method. A likelihood function was established between item level and correlation [4]. Qin et al. pointed out that a diagnostic assessment of a person's cognitive processes, processing skills, and knowledge structure was often referred to as cognitive diagnosis or cognitive diagnostic assessment. Cognitive diagnostic assessments were used to measure individual-specific knowledge structures (knowledge structures) and processing skills (processing skills) [5]. Qiu et al. believed that cognitive diagnostic tests should test at least three users' cognitive characteristics [6]. The first was that the user had significant knowledge and skills in a specific cognitive domain. However, this knowledge and capabilities could be achieved at a

higher level. And the second was the basic structure of knowledge, which could indicate that the user could use some skills to carry out the knowledge construction. The third was the cognitive process of users. In summary, the cognitive diagnosis was based on traditional measurement theory, but it emphasized that test measurement must thoroughly examine the subject's internal mental processes [7]. Xu et al. believed that the cognitive diagnosis model could be simply divided into two parts, namely, generalized cognitive diagnosis and narrow cognitive diagnosis [8]. The generalized diagnosis mainly indicated the observation of user scores and the internal characteristics of users, which could not only be used to construct psychological theories but also be used in education and teaching. Cognitive diagnosis in a narrow sense refers to classifying subjects according to whether they have mastered the skills or characteristics tested in the fields of education and teaching, mainly to achieve better communication between teachers and students. Daghestani et al. used a visual method to display the results of cognitive diagnosis to more intuitively display the student's diagnosis report [9]. Prada et al. used a cognitive diagnostic model to evaluate students' ability to master knowledge points and make personalized test item recommendations [10]. Bhat et al. found that there was a large amount of data on MOOCs related to knowledge points. Therefore, they tried to extract features from various types of materials such as texts and teaching videos on MOOCs and proposed a method based on representation learning for the first time (representation learning-based method) to automatically infer the learning order between knowledge points in MOOCs [11]. In addition to machine learning methods, Chang et al. turned their attention to information theory to formulate an information-theoretic view of knowledge point prerequisite relationships and

automatically construct relationships between concepts from text corpora, so as to generate reading lists for students to help them learn the relevant material in the best way [12].

In the research study, the students' programming abilities were modeled based on a probabilistic graphical model, and cognitive diagnosis was introduced into the field of programming education. At the same time, the students' theoretical level and practical ability were modeled. By using the same knowledge points that appear in theoretical questions and experiments, the two capabilities were linked for a more comprehensive diagnostic report.

3. Research Methods

3.1. Problem Definition. Now, defining our problem, we assumed that there are a total of M students, K knowledge points, N_i questions, and N_e experiments. R is a score matrix of M rows and N_i columns. And R_{ji} represents the j student's scores on the i th question, where $j \in \{1, \dots, M\}, i \in \{1, \dots, N_i\}, R_{ji} \in [0, 1]$. For programming experiments in the course, R_{je}' indicates the j student's scores on the experiment e , where $e \in \{1, \dots, N_e\}, R_{je}' \in [0, 1]$. The matrix Q indicates the knowledge points contained in each topic, with a total of N_i rows and K columns. q_{ik} indicates whether the topic contains k knowledge points, where $k \in \{1, \dots, K\}$. The matrix indicating the knowledge points contained in each experiment is denoted as Q' , with a total of N_e rows and K columns. q_{ek} indicates whether the experiment e contains k knowledge points. Under normal circumstances, $q_{ik} = 1$ means that the topic i requires k knowledge points, and zero is the opposite. The meaning of q_{ik} is similar. In the model, the matrix Q is normalized, such that $\sum_{k=1}^K q_{ik} = 1, \sum_{k=1}^K q_{ke}' = 1$.

R, R', Q, Q' are known. Our goal is to give students' cognitive assessment in computer teaching, which is divided into the following three aspects [13].

- (1) Diagnose j student's programming ability c_j
- (2) Give j student the theoretical mastery α_{jk} of k knowledge point and the corresponding practical application ability β_{jk}
- (3) Predict the j student's answers η_{ji} to new theoretical question i and the practical status η_{je}' of experiment e , where $\alpha_{jk}, \beta_{jk} \in (0, 1)$, 0 means no mastery and 1 means complete mastery

The specific application process of cognitive diagnosis in computer course teaching: First, students pass exams or homework, answer theoretical questions, or write codes. Teachers or teaching assistants give a specific score according to the students' answers. The cognitive diagnostic model uses these scores, along with other teaching information, to make inferences about students' abilities and generate cognitive reports that are returned to students. Students can carry out the targeted training based on cognitive reports to check and fill gaps [14].

3.2. Cognitive Diagnosis Model

3.2.1. Cognitive Diagnosis Model

(1) *Code Capability.* To assess students' coding ability, we refer to the research findings in educational psychology. Everyone has a high-order latent characteristic that indicates a general ability to learn a subject. In the model, in order to represent the practice level of students, the high-level latent feature is visualized as a parameter that each student i has to represent the student's programming learning ability c_j . This parameter is not for a certain knowledge point, but for the embodiment of his own programming ability [15]. In Layman's terms, each student j has an independent parameter c_j that represents the student's ability to write programs. The programming ability c_j and the student's theoretical knowledge level together determine the corresponding level of the code written by the student.

(2) *The Mastery and Application of Knowledge Points.* α_{jk} is recorded as the j student's ability to use knowledge point k to solve theoretical problems, and at the same time, β_{jk} is recorded as the ability of j student to use knowledge point k to write code. In the model, it is reasonably assumed that there is no direct correlation between the various abilities, that is, the abilities of a student in different skills do not affect each other, and the abilities of different students are independent of each other. According to the experience gained in the past teaching process, the two abilities of the same skill are related. Students can apply it only after they have mastered the skill theoretically. Therefore, a hypothesis is proposed.

The j student's ability to program using knowledge point k is proportional to the student's theoretical mastery of the knowledge point and basic coding ability, i.e., $\beta_{jk} = c_j \alpha_{jk}$.

That is to say, it is believed that a person's ability to use knowledge point k to program depends on his theoretical level of the corresponding knowledge points, which is limited by his basic programming ability.

(3) *The Degree of Mastery of the Topic.* η_{ji} is used to indicate the student's mastery of a theoretical question, and η_{je}' is used to indicate the student's mastery of an experiment. The traditional cognitive diagnosis model considers that a student's mastery of a topic is related to the knowledge points that the student has learned and the knowledge points needed to answer the question. In practice, each topic contains one or more knowledge points [16]. If the student answers the question completely or gives a partial answer, it means that the student has used the specific knowledge points required by the question, and it is recognized that the student has mastered or partially mastered the corresponding knowledge points. Based on the above analysis, we define students' mastery of the problem as the following assumptions.

The degree of master of a student j on topic i is related to the level of the student's mastery of the knowledge points

examined by the topic, that is, the degree of master of the student on the topic $\eta_{ji} = \sum_{k=1}^K \alpha_{jk} q_{ki}$ and the degree of mastery of the student in the e experiment $\eta'_{je} = \sum_{k=1}^K \beta_{jk} q'_{ke}$.

In short, a student's score on a certain question is proportional to the student's mastery of the knowledge point examined in the question. The more proficient the student at playing the knowledge point, the better his performance in the question that includes the knowledge point.

(4) *Actual Score*. In actual situations, due to mistakes or guesswork, students may give some correct answers without mastering the knowledge points or write wrong answers due to subjective factors such as being careless for a while, resulting in a slight deviation between the actual score and the actual mastery degree [17]. So, it is necessary to take into account the impact of these factors on students' final scores in the model. Therefore, refer to the method of probability matrix decomposition in the recommendation system, that is, there is a certain gap between the score predicted by the model and the user's actual score on the item, and this gap conforms to the Gaussian distribution. Therefore, the actual score is simulated according to the students' mastery of the problem as given in the following formulae:

$$R_{ji} \sim \mathcal{N}(\eta_{ji}, \sigma_R^{-1} \mathbf{I}), \quad (1)$$

$$R'_{je} \sim \mathcal{N}(\eta'_{je}, \sigma_{R'}^{-1} \mathbf{I}). \quad (2)$$

σ_R and $\sigma_{R'}$ are the hyperparameters and \mathbf{I} is the identity matrix. In our model, the actual score follows a Gaussian distribution with the mean value of the mastery level, that is, the actual score should be related to the students' mastery of the problem. However, some external factors will lead to the probability of deviation.

(5) *Summary*. In order to understand the abovementioned methods better, the proposed model is summarized as the probability graph, in which the gray circles are the observed values and the white circles are the unknown quantities. It can be seen from the figure that the observable value is the M students' performance R on the N_i questions and R' on the N_e experiments, as well as the matrix Q and Q' of the corresponding relationship between each question and K knowledge points. The mastery of

degree $\{\alpha_{jk}\}_{k=1}^K$ of each student j in each knowledge point k corresponds to a programming level $\{\beta_{jk}\}_{k=1}^K \cdot \alpha_{jk}$, and c_j determine β_{jk} . The student's j master degree η_{ji} on question i is determined by the student's master's degree $\{\alpha_{jk}\}_{k=1}^K$ and the knowledge point $\{q_{ki}\}_{k=1}^K$ investigated by the topic, and student's j code master's degree η'_{je} on experiment e is determined by the knowledge point coding ability $\{\beta_{jk}\}_{k=1}^K$ and the knowledge point $\{q'_{ke}\}_{k=1}^K$ investigated by the experiment question. Finally, the actual scores of student j on question i and experiment e , R_{ji} and R'_{je} , are affected by the degree of master's $\{\alpha_{jk}\}_{k=1}^K$ and $\{\beta_{jk}\}_{k=1}^K$, respectively [18]. Following the setting of the HO-DINA model, our parameters obey the prior distribution of the following formulae:

$$c_j \sim \mathcal{N}(\mu_c, \sigma_c^{-1} \mathbf{I}), \quad (3)$$

$$\alpha_{jk} \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^{-1} \mathbf{I}). \quad (4)$$

In formulae (3) and (4), σ_c and σ_α are the hyperparameters.

3.2.2. *Parameter Optimization*. According to the above probability graph and the assumed prior of parameters, we can obtain the posterior distribution of c and α , as shown in the following formulae.

$$p(\mathbf{c}, \boldsymbol{\alpha} | R, R') \propto p(R | \boldsymbol{\alpha}) p(R' | \mathbf{c}, \boldsymbol{\alpha}) p(\mathbf{c}) p(\boldsymbol{\alpha}). \quad (5)$$

In formula (5),

$$p(R_{ji} | \alpha_j) = \mathcal{N}\left(\sum_{k=1}^K \alpha_{jk} q_{ki}, \sigma_R^{-1} \mathbf{I}\right), \quad (6)$$

$$p(R'_{je} | c_j, \alpha_j) = \mathcal{N}\left(\sum_{k=1}^K c_j \alpha_{jk} q'_{ke}, \sigma_{R'}^{-1} \mathbf{I}\right), \quad (7)$$

$$p(c_j) = \mathcal{N}(\mu_c, \sigma_c^{-1} \mathbf{I}), \quad (8)$$

$$p(\alpha_{jk}) = \mathcal{N}(\mu_\alpha, \sigma_\alpha^{-1} \mathbf{I}). \quad (9)$$

Taking the negative logarithm of the posterior distribution, the following formula can be obtained.

$$F(c, \alpha) = \frac{\sigma_R}{2} \sum_{j=1}^N \sum_{i=1}^{N_i} (R_{ji} - \eta_{ji})^2 + \frac{\sigma_{R'}}{2} \sum_{j=1}^M \sum_{e=1}^{N_e} (R'_{je} - \eta'_{je})^2 + \frac{\sigma_\alpha}{2} \sum_{j=1}^M \sum_{k=1}^K (\alpha_{jk} - \mu_\alpha)^2 + \frac{\sigma_c}{2} \sum_{j=1}^M (c_j - \mu_c)^2 + C. \quad (10)$$

In formula (10), C is a constant, f is our loss function, and our goal is to minimize the error function $F(c, \alpha)$.

Next, the way to optimize our parameters is introduced. The optimization process is mainly divided into two steps, and the detailed process is as follows:

(1) *Step 1 Optimize α* . It is fixed and each student's ability to answer each question is independent of each other, that is, α_{jk} are independent of each other, which means that the whole parameter optimization can be decomposed into multiple independent parts, and optimize each part of α_{jk} in parallel to improve the efficiency of the algorithm [19]. In the research, the

gradient descent method was used to optimize a single part, and the following formula is obtained:

$$\alpha^{new} = \alpha^{old} - r_1 \frac{\partial}{\partial \alpha} F(c, \alpha). \quad (11)$$

In formula (11), α^{old} is the parameter before the update, α^{new} is the parameter after the update, r_1 is the learning rate, and $\partial/\partial \alpha F(c, \alpha)$ is the partial derivative of the function $F(c, \alpha)$ on α .

$$\begin{aligned} \frac{\partial}{\partial \alpha} F(c, \alpha) = & -\sigma_R \sum_{i=1}^{N_i} q_{ki} (R_{ji} - \eta_{ji}) \\ & - \sigma_{R'} \sum_{e=1}^{N_e} c_j q_{ke}' (R_{je}' - \eta_{je}') + \sigma_\alpha (\alpha_{jk} - \mu_\alpha). \end{aligned} \quad (12)$$

(2) *Step 2 Optimize. c.* Similarly, α is fixed and each student j 's parameters c_j are independent of each other, so the gradient descent method is used to optimize each c_j simultaneously, as shown in the following formula.

$$c^{new} = c^{old} - r_2 \frac{\partial}{\partial c} F(c, \alpha). \quad (13)$$

In formula (13), c^{old} is the parameter before the update, c^{new} is the parameter after the update, r_2 is the learning rate, and $\partial/\partial c F(c, \alpha)$ is the partial derivative of the function $F(c, \alpha)$ on c .

$$\frac{\partial}{\partial c} F(c, \alpha) = -\sigma_{R'} \sum_{e=1}^{N_e} (R_{je}' - \eta_{je}') \left(\sum_{k=1}^K \alpha_{jk} q_{ke}' \right) + \sigma_c (c_j - \mu_c). \quad (14)$$

According to formulae (11) and (13), the value of the sum is iteratively updated, and the abovementioned process is only stopped when the model converges or the number of iterations reaches a certain number of times [20]. Finally, the algorithm outputs the students' cognitive parameters and the students' basic coding ability, and we can make predictions about the students' grades based on these two parameters.

3.3. Experiment

3.3.1. Experimental Data. To validate the proposed model, the preschool data of students in a computer course at a university were collected to validate our model. The collected real-world datasets were organized, cleaned, and formatted, excluding some special data, such as the data of students who barely handed in homework. Our experiments were conducted on two real datasets (data from the computer courses "Data Structure" and "Network Security") and a simulated dataset, all three datasets containing students' theoretical R and experimental grades R' , as well as questions and the knowledge examined in the experiment Q and Q' . In the real dataset, the students' grades and the knowledge points required by the topic were provided by teachers or teaching assistants [21]. The basic situation of the dataset is given in Table 1.

(1) *Course "Data Structures" Dataset.* The first dataset is the student practice data collected in the course "Data

Structure," containing 96 students, 58 theoretical questions, and 10 experimental questions, a total of 6528 pieces of data. The course contains a total of 19 knowledge points, including common concepts of data structures such as "find," "graph," and "binary tree." Theoretical questions are mainly theoretical investigations of this knowledge. While, experiments include classical implementations of various algorithms as well as simulated real-world scenarios, such as "bank queues" and "flight reservations" [22].

(2) *Course "Network Security" Dataset.* The second real dataset is a dataset collected from the course "Network Security," which contains 10 theoretical questions answered by 194 students and 8 experimental answers, a total of 3492 pieces of data. The course contains a total of 7 knowledge points, involving "encryption/decryption," "buffering," and so on. The content of the experiment includes "shell-code" and "vulnerability" and other related content.

(3) *Simulation Dataset.* Finally, a simulated dataset is constructed. In the simulated dataset, we set up 1000 students, 200 theoretical questions, 50 experiments, and a total of 250,000 pieces of data. The dataset contains a total of 20 knowledge points.

First, the theoretical questions are taken as an example to introduce how to simulate students' answering situations [23]. In order to simulate the different skill levels of different students, each student is assigned a personalized personal model. Assuming that the students' mastery of knowledge points obeys the Gaussian distribution, the students' ability determines the mean and variance of the personal model. For students with a higher degree of mastery of knowledge points, the mean of the personal model is higher; while for students with a low degree of mastery of knowledge points, the mean of the personal model is correspondingly lower. At the same time, the variance of the model corresponding to the careful student is smaller, that is, the student's score can better reflect his true level. Relatively, the variance of the personal model corresponding to the careless student is larger because his answering situation may not be stable. In addition, matrix Q is generated to link questions and knowledge points together. The students' answers to each question are generated based on the student's personal model and the knowledge points contained in the questions indicated by the matrix Q . To reflect the instability of the students' on-the-spot performance, a random error in the range $[-0.1, 0.1]$ is added to the simulated scores. If the error value is negative, it means that the student does not play at their true level; if it is positive, it means that the student has played well or guessed part of the answer. Finally, our simulated scores range between $[0, 1]$, and even after adding the error value, it will not exceed this range.

A similar method is used to simulate the experimental performance of students. A student's practical ability on a certain knowledge point is related to the student's theoretical mastery. Therefore, to simulate the students' practical ability, a random number in the range $[0, 0.1]$ is subtracted from the students' theoretical ability. This ensures that the students' practical ability is linked to their theoretical ability. But in

TABLE 1: Dataset overview.

Dataset	Number of students	Knowledge points	Number of theoretical questions	Number of programming questions	Data volume
“Data structure”	96	19	58	10	6528
“Network security”	194	7	10	8	3492
“Simulation dataset”	1000	20	200	50	250000

TABLE 2: Experimental results for predicting student performance.

Model	“Data structure”		“Network security”		“Simulation dataset”	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
IRT	0.358 ± 0.02	0.415 ± 0.02	0.203 ± 0.02	0.247 ± 0.02	0.227 ± 0.02	0.327 ± 0.02
PMF-5D	0.316 ± 0.02	0.380 ± 0.02	0.195 ± 0.02	0.246 ± 0.02	0.212 ± 0.02	0.265 ± 0.02
PMF-10D	0.319 ± 0.02	0.383 ± 0.02	0.196 ± 0.02	0.248 ± 0.02	0.219 ± 0.02	0.265 ± 0.02
PMF-KD	0.319 ± 0.02	0.383 ± 0.02	0.196 ± 0.02	0.248 ± 0.02	0.213 ± 0.02	0.265 ± 0.02
FuzzyCDF	0.283 ± 0.01	0.376 ± 0.02	0.210 ± 0.01	0.230 ± 0.02	0.220 ± 0.01	0.314 ± 0.02
NeuralCD	0.464 ± 0.03	0.481 ± 0.02	0.429 ± 0.02	0.498 ± 0.02	0.120 ± 0.02	0.235 ± 0.02
CDF-CSE	0.147 ± 0.01	0.231 ± 0.01	0.066 ± 0.01	0.146 ± 0.01	0.106 ± 0.01	0.132 ± 0.01

order to conform to the actual situation, the students’ practical ability will not exceed their theoretical mastery of knowledge points. Finally, students’ scores on experimental questions are simulated by using the method described above.

3.3.2. *Experimental Settings.* Compared with the four methods, the validity of the proposed cognitive diagnosis model is proved.

Item Response Theory (IRT). IRT is a classical cognitive diagnostic method, which models students’ potential abilities and characteristics of questions.

Probability Matrix Factorization (PMF). PMF is a latent factor model that projects students and questions into a low-dimensional space.

FuzzyCDF. This method introduces the concept of fuzzy sets, regards the student’s performance as a continuous value, and combines a variety of classical cognitive diagnosis methods to improve itself.

NeuralCD. This approach combines deep learning and remains interpretable.

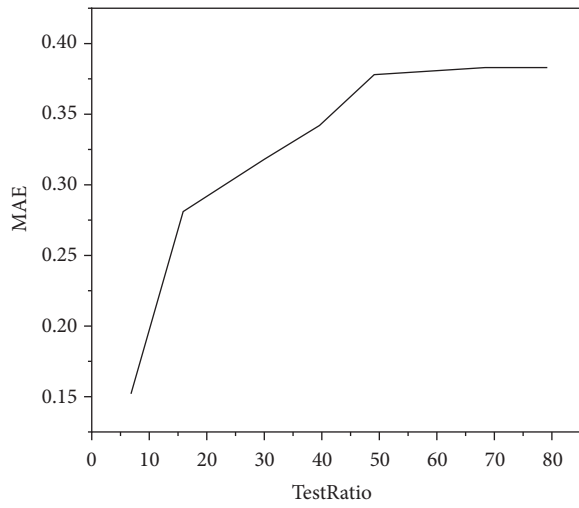
Mean absolute error (MAE) and root mean square error (RMSE) are used to evaluate the performance of each model. During the training of CDF-CSE, when the parameters do not change (the difference between the results of two consecutive iterations is less than 10^{-10}), the training is stopped and the results are obtained. For each experiment, it runs 100 times, and the results are averaged. It is worth mentioning that the IRT model and FuzzyCDF are implemented using the MCMC method. In the experiments of these two algorithms, the number of iterations is set as 10000, and sample parameters are obtained based on the results of the last 2500 iterations. To make a fair comparison, the parameters are adjusted to record the best performance of each algorithm. Finally, all algorithms are implemented in Python and run on Windows 10 machines with 8 GB of memory and an i5 3.2 GHz CPU.

4. Result Analysis

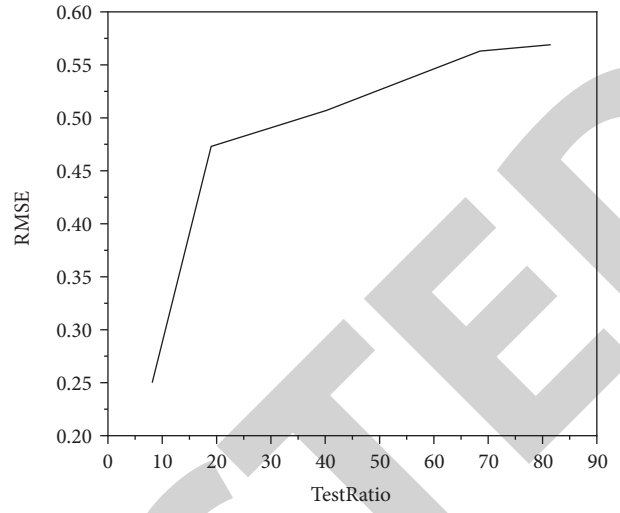
4.1. *Score Prediction.* First, the accuracy of each model in predicting students’ scores is compared to judge whether the cognitive results given by the model are reliable. Parameters α and β are the theoretical and practical abilities of students given by the model. These two parameters are used to predict students’ scores, and the cognitive results obtained by each method are tested by the error between predicted scores and real scores. In the experiment, different implementations of matrix decomposition methods, namely, PMF-5D, PMF-10D, and PMD-KD were used as PMFS with 5, 10, and K (knowledge point number) potential factors, respectively.

First, the fixed training set accounted for 80%, and the test set accounted for 20%. In the dataset containing both theoretical and experimental data, the performance of each model is compared. Table 2 shows the experimental results for predicting student performance. Overall, CDF-CSE performance is the best among the three datasets because it can establish a correlation between theory and practice [24]. It is worth mentioning that NeuralCD uses a deep learning method, which requires a large number of instances for training, so it performs poorly on the two smaller datasets and performs well on the larger simulated datasets, but it is still not as good as CDF-CSE.

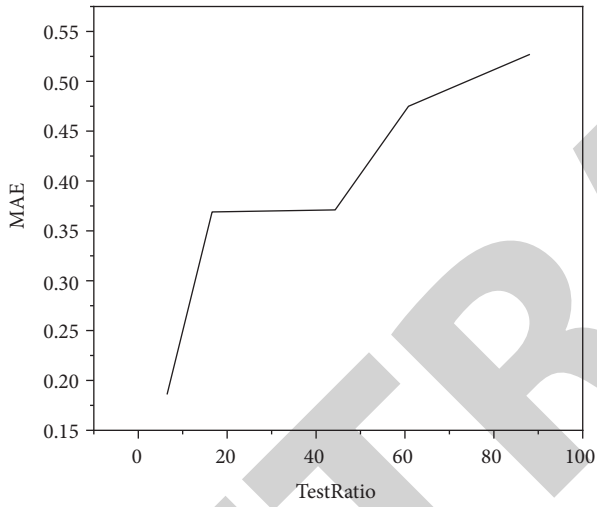
In order to observe the performance of different datasets in different levels of sparse levels, a training set of different sizes is constructed, with 10%–80% of each candidate’s scoring data, and the rest are for the tests. Since the NeuralCD is poor in small datasets, it is not compared in this experiment. In addition to the model of the NeuralCD, the models will be compared to the dataset that contains the theoretical problem, the dataset that contains only the experimental problem, and the performance of the dataset that contains two types of data. From the different angles mentioned above, the overall comparison is different. Figures 2(a)–2(f), 3(a)–3(f), and 4(a)–4(f) show the results of the CDF-CSE and other methods in three different



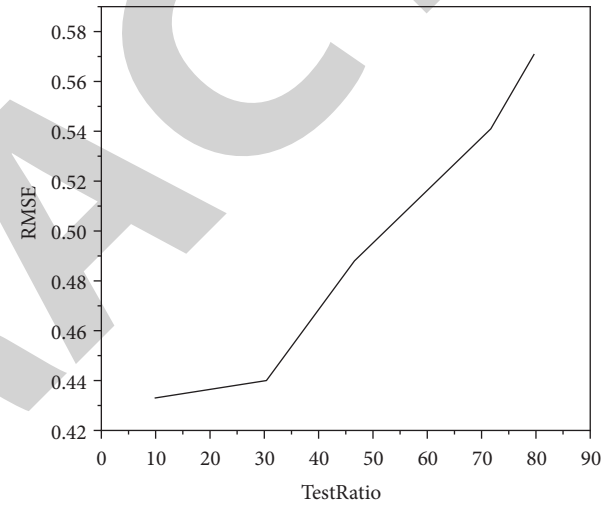
(a)



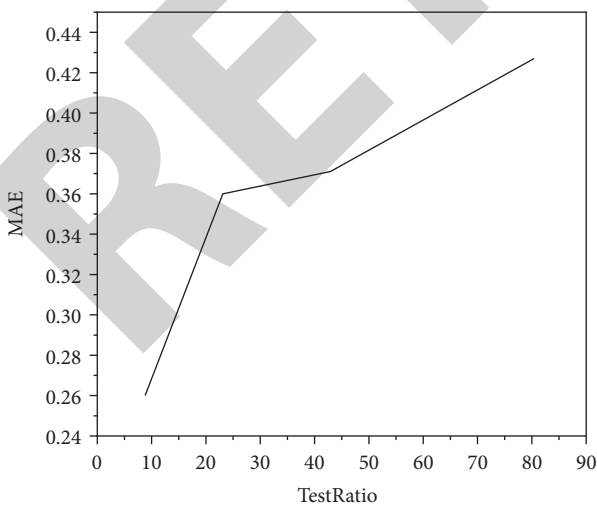
(b)



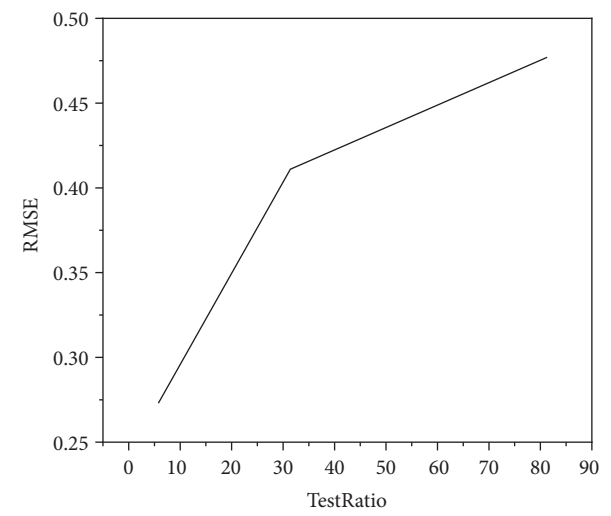
(c)



(d)



(e)



(f)

FIGURE 2: The performance of each model on the “data structure” course datasets. (a) Data structure-theoretical-MAE. (b) Data structure-theoretical-RMSE. (c) Network security-experimental-MAE. (d) Network security-experimental-RMSE. (e) Data structure-mix-MAE. (f) Data structure-mix-RMSE.

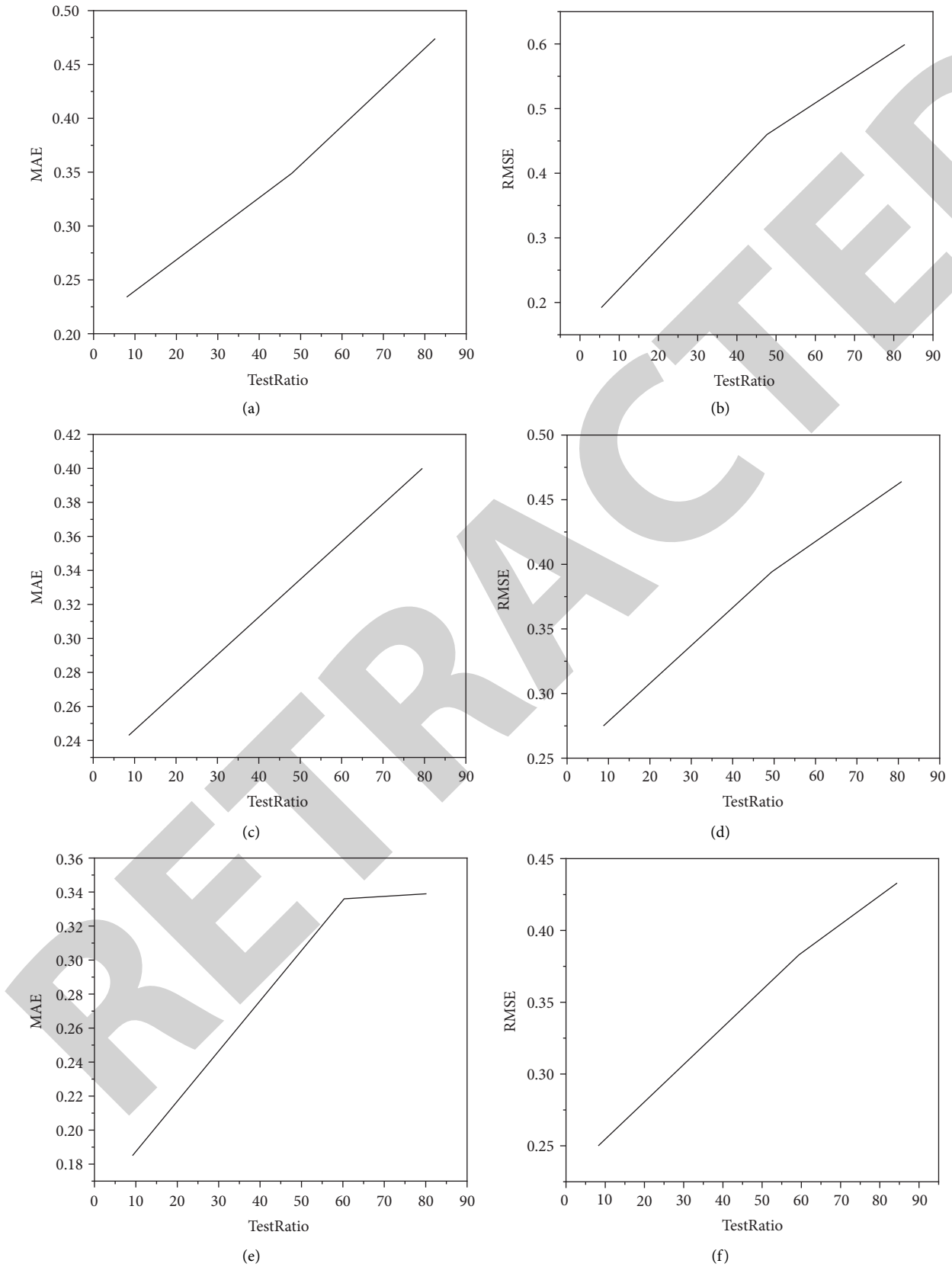
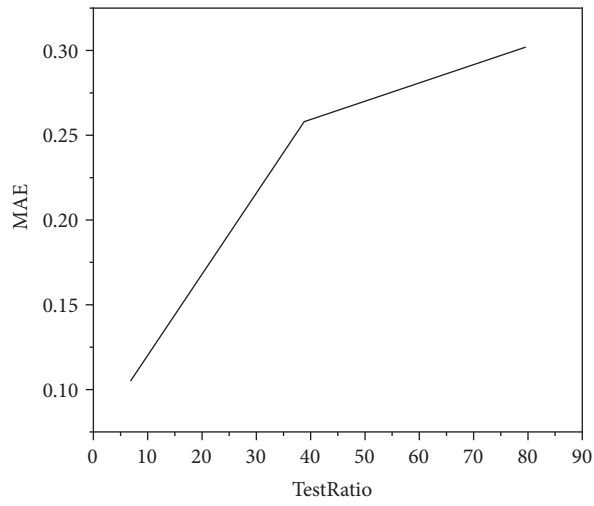
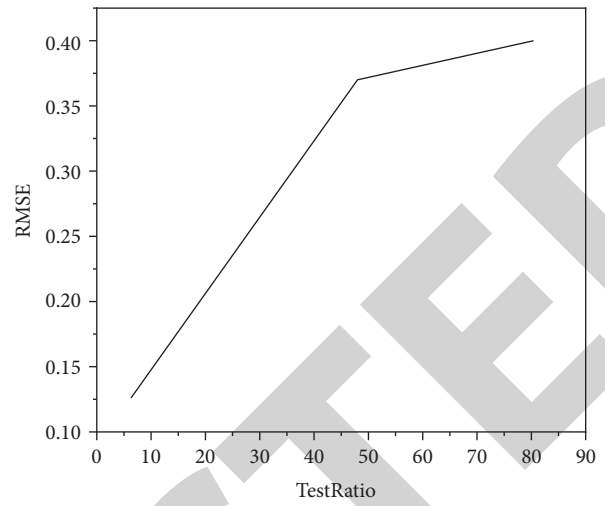


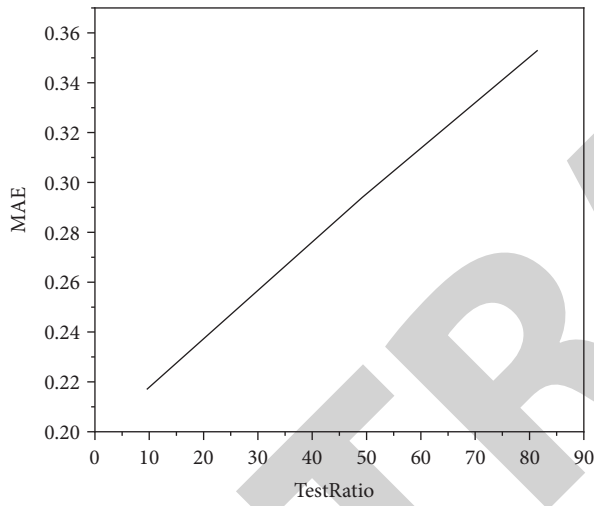
FIGURE 3: Performance of each model on the “network security” course datasets. (a) Network security-theoretical-MAE. (b) Network security-theoretical-RMSE. (c) Network security-experimental-MAE. (d) Network security-experimental-RMSE. (e) Network security-mix-MAE. (f) Network security-mix-RMSE.



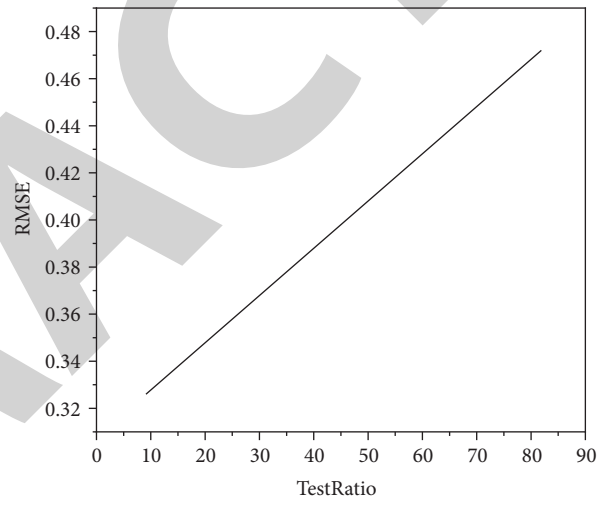
(a)



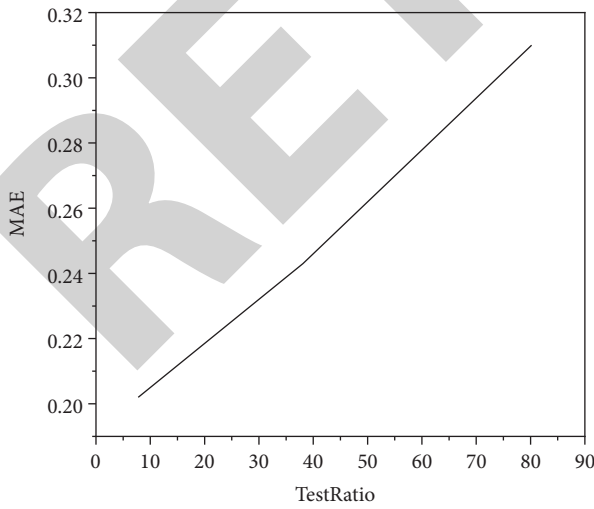
(b)



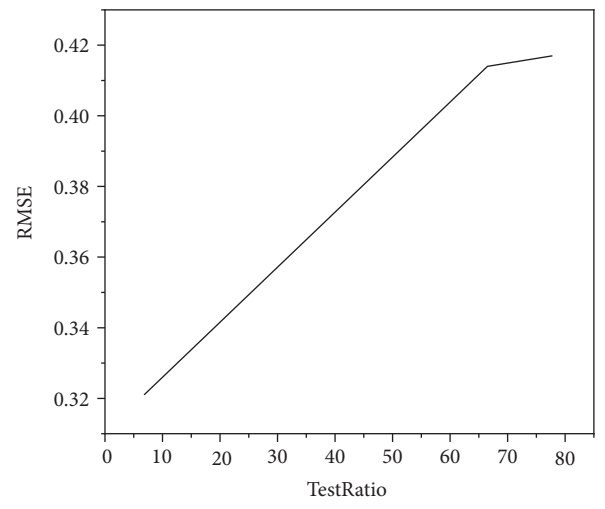
(c)



(d)



(e)



(f)

FIGURE 4: Performance of individual models on simulation datasets. (a) Synthetic-theoretical-MAE. (b) Synthetic-theoretical-RMSE. (c) Synthetic-experimental-MAE. (d) Synthetic-experimental-RMSE. (e) Synthetic-mix-MAE. (f) Synthetic-mix-RMSE.

datasets. In the results of the empirical research, it is observed that the CDF-CSE performs best in all datasets.

Specifically, it outperforms PMF in terms of combining teaching assumptions, outperforms IRT in terms of quantitatively analyzing students, and outperforms all other methods in terms of combining theory and experimentation. In addition, the parameters obtained by CDF-CSE can directly represent the cognitive state of students inferred by the model, which is interpretable. However, the parameters obtained by IRT and PMF cannot give students' abilities in various knowledge points. Such a model, even if it can accurately predict students' scores, has little effect in diagnosing students' cognitive status. More importantly, as the training data become sparser (the proportion of training data drops from 80% to 20%), the advantages of the CDF-CSE method gradually emerge. For example, in the comprehensive question of three datasets, when the training set is 20% and MAE is used as the judging criterion, CDF-CSE is improved by 47.8%, 65.8%, and 49.8% compared with other methods that perform best on the training set.

Obviously, CDF-CSE is more accurate than other methods because the CDF-CSE method can be trained on datasets with two different problems, theoretical and experimental. That is, compared with other models that only consider one kind of problem, CDF-CSE can obtain more information during the training process. For the dataset containing two types of problems, the model will provide different probability assumptions for the two kinds of problems. This matches the actual situation. Even in the special case where the probability distribution of students' scores on both questions is the same, CDF-CSE can be trained normally. However, other models can only consider one probability distribution, which inevitably produces errors. In other words, on different types of questions that examine the same knowledge points, students' performance should be the same, and CDF-CSE makes full use of this feature. When the model observes that students have a good grasp of a knowledge point in theory, the model reasonably predicts that the students can use the knowledge point well in practice. At the same time, we can also infer the theoretical level of students according to the situation of students in the experiment. This approach is in line with the actual teaching experience, and the experimental results also prove the feasibility of the method. In contrast, other models do not do this, so they occasionally perform poorly on some datasets. And compared with other models, the assumptions of CDF-CSE are also more applicable to subjective questions and computer teaching, which also leads us to achieve good performance on all three datasets.

To sum up, CDF-CSE can capture the characteristics of students more accurately, and it is also more suitable for practical teaching scenarios with sparse data.

4.2. Course Process Follow-Up. Furthermore, we hope that cognitive diagnostic models can not only evaluate students after the course is completed but also provide feedback to students as the course progresses. In this way, the cognitive diagnosis model can help students

understand their own cognitive structure in the learning process and timely check and fill in gaps to improve learning efficiency. Therefore, the process of taking a class is simulated, and experiments are carried out in the process. In the experiment, a dataset containing two kinds of questions is used, and the data volume of the training set is fixed at 80%, and the rest is used as the test set. At the same time, according to the chronological order, only a small number of questions are used for training at the beginning, and the number of questions is gradually increased after getting the results.

CDF-CSE still performs the best on all datasets. It performs well in the early stages (when there are few knowledge concepts and questions). Its advantages gradually become apparent as the amount of data increases. For example, when the number of items is small, the CDF-CSE model achieves 37.8%, 42.5%, and 27.7% improvement in the RMSE metric on the three datasets, respectively, compared with the other best-performing methods. Improvement. It has 32.3%, 36.5%, and 45.6% improvement when the number of questions is large.

The result proves that the method of combining theoretical and experimental performance to analyze the cognitive state of students is still feasible in the case of little data. Even in the early stages of teaching, CDF-CSE can use more information than other models to better analyze the characteristics of students. As the course develops, the analysis results are gradually more accurate. To sum up, the proposed CDF-CSE can well follow up the whole process of computer course teaching.

5. Conclusions

According to the experimental results, the CDF-CSE method is superior to other methods in predicting student performance. This is because CDF-CSE can extract common features from two types of questions and at the same time can distinguish the differences between them, so as to diagnose students' theoretical cognition and practical ability. Compared with other models, CDF-CSE is more suitable for teaching computer courses, and the results given are more abundant and accurate. The experimental results also confirm that our model can be applied to different situations. Therefore, more comprehensive cognitive information can be analyzed using different data from students. From the experimental results, it is concluded that CDF-CSE can solve the problem of inaccurate feedback of traditional cognitive diagnostic models in computer course teaching and provide more detailed and interpretable cognitive analysis results. In computer science teaching courses, such results help teachers understand the teaching situation and can also help students learn in a targeted manner.

Data Availability

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

Retraction

Retracted: Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Ju, "Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 5222760, 9 pages, 2022.

Research Article

Application of Big Data Technology to Promote Agricultural Structure Adjustment and High-Quality Development of Modern Agriculture

Xiaoxian Ju 

Henan University of Animal Husbandry and Economy, Zhengzhou 450046, China

Correspondence should be addressed to Xiaoxian Ju; 80702@hnuah.edu.cn

Received 10 July 2022; Revised 3 August 2022; Accepted 15 September 2022; Published 29 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Xiaoxian Ju. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The implementation of the strategy of rural revitalization is a major ministerial work made by the Nineteenth National Congress of the Communist Party of China and is the general grasp of contemporary agriculture, peasants, and rural work. In recent years, with the rapid development of remote sensing technology and deep learning technology, the demand for the technology for the classification of crops on satellite remote sensing images based on deep learning technology has increased in agricultural insurance and land survey. Therefore, this paper trains one, which is 85.9%–92.8%, the accuracy of corn classification is 77%–93%, and the accuracy of forest classification is 77%–87.6%. Subsequently, the overall accuracy of classifying all directories through the multi-temporal validation data set between May 2017 and October 2017 reached 92.6%. Such a multi-time combination method can be used for monthly, timely, and efficient iteration of agricultural insurance and crop yield estimation, which will be more accurate each time. These methods can also be further applied to the growth and change monitoring of large agricultural planting areas, adding bricks and tiles to China's agricultural remote sensing. If the countryside is to be revitalized, agriculture must develop rapidly at the same time, industries must flourish, ecology must be livable, rural customs must be civilized, and life must be prosperous. Modern agriculture is a comprehensive circulation system with high yield, high quality, low consumption, ecology, environmental protection, and high efficiency. The development of modern agriculture is inseparable from the industrialization of agriculture, the globalization of agriculture, the digitization of agriculture, the integration of agriculture, the adjustment of agricultural structure, and agricultural innovation. Only the continuous development of modern agriculture can make rural revitalization enter a new journey.

1. Introduction

China is now a powerful agricultural country with 1.4 billion people. Agricultural products determine the social value of grain and directly affect the living level of people [1]. Grain production is related to national transportation and people's livelihood, which is the basic consensus of this paper. Food production needs to ensure safety [2] All these are important sources of our sense of security. The mountains are green, the forest is stone, and the world is safe. Since 2004, China's grain production has continued for 16 years, with a bumper harvest every year [3]. At present, the state vigorously promotes the strategy of rural revitalization, and leading cadres at all levels are actively committed to optimizing rural

formats, especially taking the reorganization of various resources as the starting point, actively using the advantages of land, space, labor, and other advantages to vigorously develop modern agriculture and help rural revitalization.

The latest opinions on food security issues were put forward at the second National Congress in 2020 [4]. According to the principle of "ensuring the basic self-sufficiency of grain and the absolute safety of rations," the government has formulated the agricultural policy of putting people first, maintaining the balance of domestic and foreign demand, insisting on expanding production, appropriately liberalizing the customs of some agricultural products, and importing domestic crops of grain [5]. Scientific and technological development is an important strategy for

China's food security. We will always adhere to the path of food security with Chinese characteristics. Under the epidemic situation in 2020, the country and society mentioned in the text will still maintain stable social order, but the safe supply of agricultural products and food will still be improved [6]. Under the premise of changes in the external environment, the country's economic development will remain relatively stable. Through their own agricultural development, the Chinese people have not only maintained the stability of the people's production but also maintained the foundation for national economic recovery. Therefore, the stability of agricultural structure and the high-quality development of the modern agricultural system are indispensable [7].

Annual crop conditions play a significant role in harvest. Determining the maturity of grain crop coverage plays an important role in promoting national food security and the sustainable development of the rural economy. Through macro-control, China can understand the situation of crops and select various types of cultivation methods. Adapt to the growing needs of mankind and the ever-changing production requirements [8]. The countryside must be revitalized, not without agriculture. To develop agriculture, is impossible to do without the progress of agricultural science and technology, without the training of professional farmers, and even more without the innovation of agricultural production and management.

In the practice of agricultural land use, the situation of crop planting land in China is very complicated. The landforms such as flat land, mountains, and hills are variable and undulating with a large span. In the plain area, some land is suitable for large-scale or intensive agricultural production; in the hilly area, it is suitable for fruit cultivation or returning farmland to forest; in the mountainous area, it is suitable for developing green planting. However, only by rationally developing the planting industry according to different terrain can we promote the healthy and sustainable development of rural areas [9]. Under the interference of various reasons, China's agriculture is facing great historical challenges. Under this premise, only by finding the right way can China's agricultural products greatly improve their output efficiency [10].

If statistical analysis is carried out only by the manual measurement method, its cost is quite expensive and cannot be popularized, because it requires a large number of locally collected survey data and a large number of manual operations. Land statistics based on big data analysis is a modern scientific research method [11]. For example, the complex model represented by in-depth learning can automatically obtain agricultural environmental information and automatically adjust the industrial structure through learning a large number of remote sensing satellite image data, which can not only save labor costs and time costs but also maximize benefits [12]. Therefore, agriculture should focus on diversified development, and high-quality links should be carried out with new village planning, new industrial development, cultural inheritance, etc., so that modern agriculture can show a thriving development outlook. China is a large country with a vast source and very rich agricultural

resources and a large agricultural population is large, so to rejuvenate the Chinese nation, we cannot but attach importance to agriculture, and only with the modernization and development of agriculture can we have the smooth realization of rural revitalization.

2. State of the Art

2.1. Modern Agricultural Digitalization. Modern agricultural digitalization includes high-tech electronic intelligent control equipment such as seedling breeding, cultivation management, soil, and environmental management, so that agricultural production methods can achieve intelligent management, remote diagnosis, remote control, and disaster early warning.

The application of deep learning technology in agricultural remote sensing analysis is , and crop classification methods will be increasingly diverse. Historically, different forms of deep learning have had a great impact on crop coverage surveys. More satellite data are now easier to obtain than in the past [13]. The use and construction of large digital changes for the countryside. Zero people think that deep learning technology is very important for remote sensing data processing, especially in the era of remote sensing big data processing [14]. Deep learning technology will become the most advanced biological remote sensing technology. With the help of remote sensing big data analysis, zero can provide services for fine agriculture through deep cognition technology. Specifically, by applying the semantic segmentation mode, crop information can be obtained more efficiently in remote sensing video. Based on the difference of crop species, the semantic segmentation mode can also get the marks of each pixel in the remote sensing image. Remote sensing semantic division is often used in agricultural image classification [15].

Based on the richness of remote sensing data, the main task is to identify nonRGB images because such images not only contain RGB bands but also other bands in the image data. It is extremely difficult and flexible to make all the effective bands in remote sensing images [16]. Li. Extracted rice fields using a multi band medium high resolution image spectrometer (MODIS). In addition, Huang. Also used the ones. These two studies use more RGB signals than traditional remote sensing image processing. With the continuous movement of the moon in the sky, it can continuously obtain a large number of multi-temporal images. In this paper, multi-temporal data can also be used to observe the change of landforms or crop growth [17]. For example, it can monitor the ecological diversity of grassland vegetation and even eliminate the impact of clouds in remote sensing images. All these functions can be realized with the help of more time phase diagrams. This facilitates the practical application of multispectral remote sensing images in this paper [18].

At present, China's agriculture has entered a new stage of high-quality development, in accordance with the new requirements of high-quality development, it is necessary to accelerate the pace of modern industrialized agricultural development and further promote the integrated

development of primary, secondary, and tertiary industries. First, it is necessary to expand in terms of resources, markets, business formats, and products; second, it is necessary to do a good job in combining agriculture and leisure tourism, agriculture and circulation culture, agriculture and healthy old-age care, and agriculture and local education; the third is to realize the integration of various business forms and the complementary advantages of various business forms.

However, there are still a number of problems that must be overcome. Therefore, if people want to use the full band signals and multi-phase signals of satellite images at the same time, it is necessary to investigate how computers can carry such large-scale data processing in the process of deep learning training. Although single-phase remote sensing images are easy to obtain, the image sequences for geographical coordinates. An effective multi-temporal semantic segmentation model requires a large amount of temporal information training. So, of course, another way to add training data is also. Therefore, we must find a way to study how to train a good multi-time-related semantic segmentation mode without abundant temporal information [19].

2.2. Multi Temporal Remote Sensing Satellite Image. With the development of remote sensing technology, the spatial clarity of satellite remote sensing images has been gradually improved, so that people can more accurately obtain the actual data in remote sensing images. World rural development is the basic outlet and inevitable development of rural areas. Its basic purpose is to make full use of the advantages of all kinds of modern agricultural products in the world, deeply participate in global rural cooperation, optimize the allocation of agricultural product resources, expand the supply of modern agricultural products, and further improve the income of farmers.

At present, there are mainly three kinds of analysis technologies used in multispectral remote sensing, namely, semisupervised classification technology, supervised analysis technology and unsupervised analysis technology. Semisupervised classification technology mainly includes self-training, joint training, semisupervised vector method and traditional naive Bayesian classification. 2. Unsupervised methods are mainly divided into some clustering methods and principal component methods, such as k-means clustering method and the PCA linear dimension reduction method. The supervised method is the most common classification method, and the neural network method is the most common. In addition, there are the Mahalanobis distance judgment method, Fisher: linear judgment method, second normal form distance judgment method, and so on [20].

2.3. Semantic Segmentation Neural Network. The convolution layer: since people can obtain information about images by convolution operation, the

Some features of the original signal are enhanced, and noise is reduced. As shown in Figure 1, by mapping the input to the input, taking filter: as the convolution kernel and (1) as the convolution operation, we can obtain the results of all the

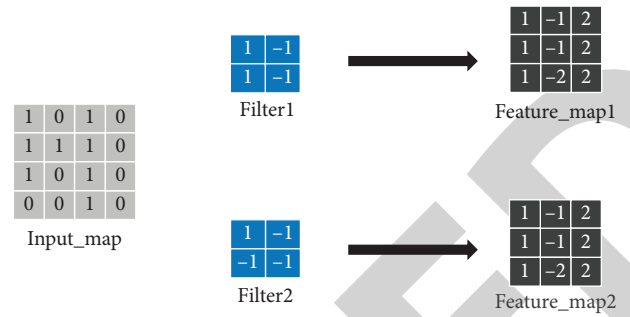


FIGURE 1: Principle of convolution operation.

feature maps in the figure. The results of feature mapping can be used as feature maps in neural networks.

$$(f * g)(n) = \sum_0^{\infty} (f(x) * g(n-x)). \quad (1)$$

Full connection layer: many-to-many mapping is adopted for each input and output, and multi-dimensional signals are mapped to the entire output layer through SoftMax parameters. SoftMax can be used to obtain the corresponding confidence of all input and output data in this article. The standard expression of the SoftMax parameter is equation (2). Where X represents the output framework of the upper layer of the model, Y represents the final division conclusion of SoftMax, and w_i is the weight value of this layer.

$$p(y | x) = \frac{\exp(W_y \cdot x)}{\sum_{C=1}^C \exp(W_c \cdot x)}. \quad (2)$$

2.4. Loss Function. The loss function is used to measure the difference between the simulated predicted value and the actual label value (ideal output value). Generally speaking, the loss function directly affects the quality of simulation. The loss function generally includes two components: experience risk loss function and structural risk loss function. The structural risk loss function refers to the empirical risk loss function plus the regular term. Realize the further exchange of agricultural products and labor to form an interconnected global whole; the second is to actively do a good job in international agricultural technology exchanges and cooperation, and take the initiative to exchange and dock global high-tech achievements; the third is to reform the traditional farming mode and develop organic agriculture and ecological agriculture; the fourth is to increase cooperation in the global market to realize the operation of agricultural commodities in the world.

Common loss functions are 0-1 loss functions, that is, the correct value is 1 and the error value is 0. The 0-1 loss function is usually used to judge the number of errors and is not suitable for fitting the real difference results. As formula :

$$L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases}. \quad (3)$$

The absolute value loss function can describe the deviation of the calculation results in the one-dimensional straight line, as shown in formula :

$$L(Y, f(x)) = |Y - f(x)|. \quad (4)$$

Log loss function is also one of the most commonly used loss functions. This logarithmic function can also well describe the difference between numerical distribution results. It is a loss function used when analyzing scenarios and is used to describe various types of confidence quantities. However, it has high robustness but poor robustness to noise. and the robustness against noise is poor. Formula (2) five

$$L(Y, P(Y | X)) = -\log P(Y | X). \quad (5)$$

The square loss function is often used in regression and pixel-level prediction. It is easy to calculate, but the weight update is slow. It can better express the direct difference between the output value and the predicted value:

$$L(Y | f(X)) = \exp [-\gamma f(x)]. \quad (6)$$

Cross entropy loss function, such as 2.7. It is mathematically known as the likelihood function. When the calculation of the square difference function is slow, this paper usually uses the cross line to calculate the function. It is often used in multi-classification tasks. In this paper, softmax is used as the activation parameter of the output method, and the cross entropy loss method is generally used. It can overcome the phenomenon that the weight adjustment speed of the square loss function is slow, and it can also automatically adapt to the update according to the error degree:

$$C = -\frac{1}{n} \sum_x [y \ln a + (1 - y) \ln (1 - a)]. \quad (7)$$

2.5. U-NET. The U-net network is a classical semantic segmentation neural network, and its two sides are symmetrical. It is a standard neural network with an encoder decoder architecture. It can complete end-to-end image mapping, such as the basic structure of u-net shown in Figure 2. As an extension of the FCN network, the u-net network can also be applied to the case where the amount of marked data is relatively small. In general, the number of labeled data for medical data is relatively small, so u-net can better meet the training of small sample number. The u-net network integrates the feature structure information in the down sampling process into the up sampling process through the integration method, so that the coordinate signals in the original image features can be more accurately retained.

3. Methodology

The Deeplab network is a series of network structures proposed by Hekaiming and others. It is also a common and practical network structure in recent years. From the deeplab V1 network to today's deeplab v3+ network, the accuracy of

the deeplab network in semantic segmentation has been greatly improved.

The overall structure of the deeplabv3+ network is shown in Figure 3 It is also a network model of the encoder-decoder structure. The encoder part of the network uses the entire deeplabv3 network model and the DCNN (deep convolutional neural network) Generally, only the convolutional neural network of the RESNET system can be used. The architecture adopts the ASPP architecture and introduces multi-scale image signals. In the decoder part, the deep labv3 + network combines the low-level characteristic signal of the deep labv3 network with the high-level characteristic signal output by the deep labv2. Then, the conclusion is obtained by convolution calculation. To promote a new round of high-quality industrial revitalization in Dongtai rural areas, it is necessary to further emancipate the mind, based on the further implementation of the national agricultural modernization revitalization policy and the demand for agricultural and rural high-quality industries, and adhering to the government's purpose of serving agriculture and the people's livelihood, conscientiously shoulder the responsibility of industrial revitalization, actively seek new problems, pay close attention to policy implementation, implement the concept of integrated development of three industries and the idea of full value chain development. Great progress has been made in the fields of quality brand building and green industry development. We will actively integrate the primary, secondary, and tertiary industries in rural areas, reform and open up new development, and vigorously promote agriculture and rural high-quality industries.

A satellite image is a very convenient tool for agricultural investigation. With the progress of deep learning, this paper can use the method of deep learning to learn crops. In the business of agricultural insurance, in-depth learning technology and skills are used to effectively extract the planting areas and types of crops, which can quickly produce results and save human costs. Policy innovation is the power source of rural industrial revitalization, and policy guarantee is the driving force of rural industrial revitalization. First, comprehensively promote agricultural reform. We should grasp the relationship between top-level design and basic exploration, overall integration and key breakthroughs, solidly promote the reform of the rural "Three Rights Separation," the collective property rights system and the rural financial system, accelerate the structural transformation of the rural supply side, and further improve the input system for agricultural economic development.

The main data collection area of this paper is Jilin, which is located in the center of Northeast China, adjacent to Dalian Province, Inner Mongolia, Heilongjiang Province, the Russian Federation and South Korea. It is also the main geographical center of Northeast China studied in this paper, and the regional shape changes significantly. From the perspective of landform, the landform is inclined from southeast to northwest of Jianguo, showing the characteristics of high in the southeast and east and low in the northwest of Jianguo. The whole province can be divided into two types of terrain: the eastern hills and the central and

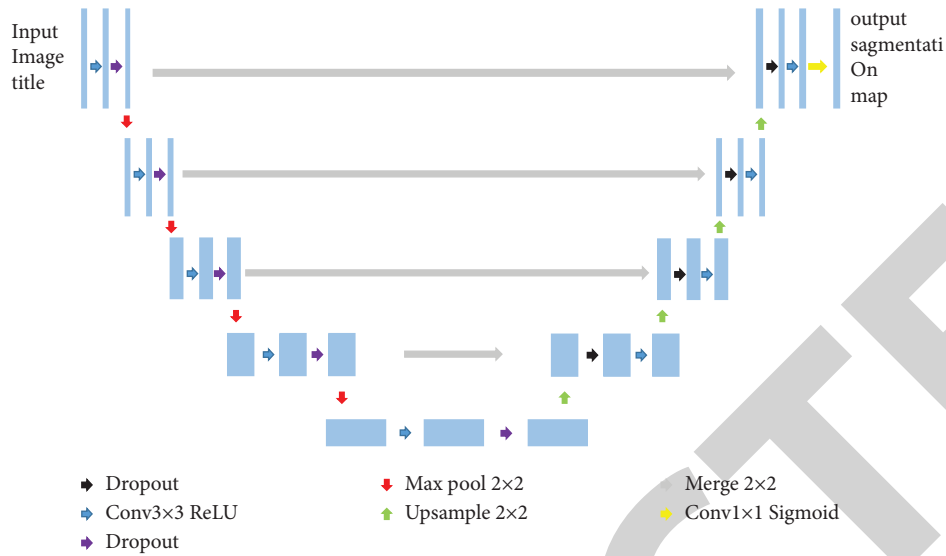


FIGURE 2: The basic structure of a U-net network.

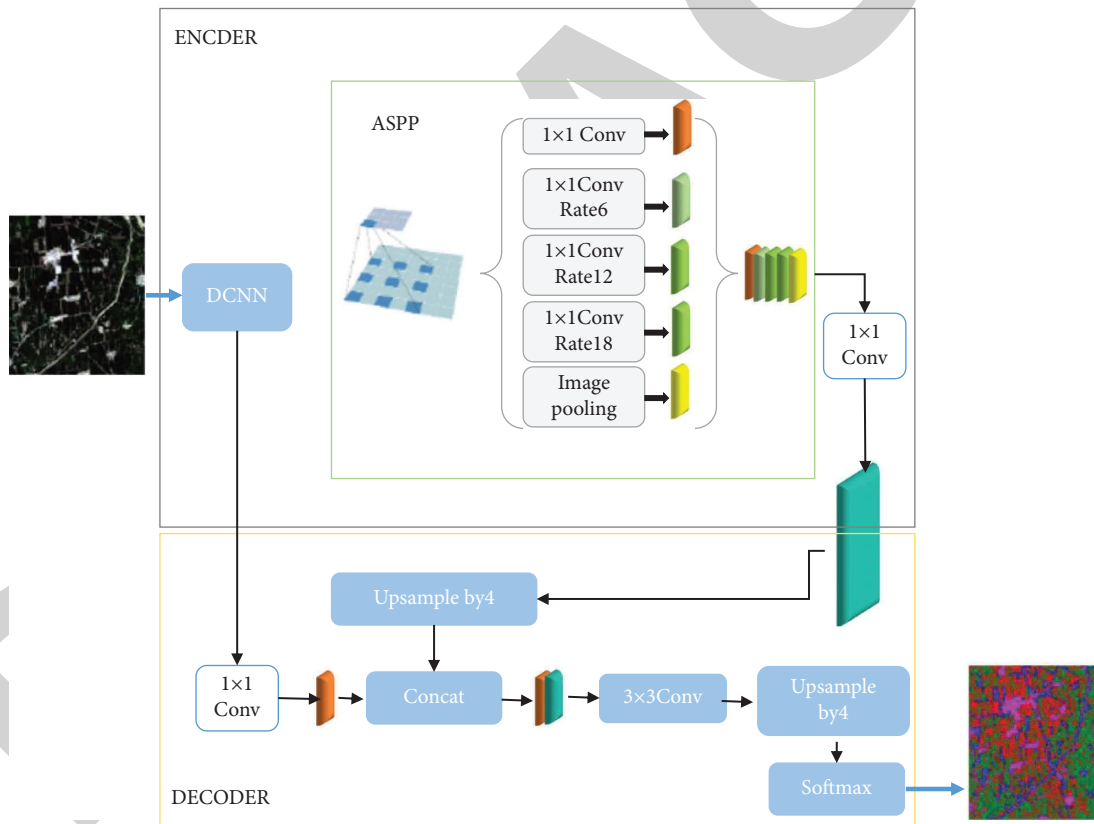


FIGURE 3: Structure diagram of the DeepLabV3+ network.

western alluvial plains. Jilin is the main grain and oil production province in China. By 2019, the sown area of grain and oil will be 5.645 million hectares, including 840 hectares of rice, 400 hectares of wheat, 421.5 hectares of corn, and 1 million hectares of wheat 1996. The cumulative total output of grain and oil was 38.78 million tons.

In such an important area, in order to ensure that the growth of crops can be monitored and farmers can plant

crops at ease, the state has implemented an agricultural insurance policy for major crops. The main purpose of this paper is to monitor and extract these planting areas.

The data in this paper is from the remote sensing satellite Landsat 8, launched by the United States in 2013. It is equipped with an oli land imager to detect the soil, and a TIRS thermal infrared sensor to detect the infrared spectrum information of the sun. This paper mainly uses its

TABLE 1: Introduction to oli parameters of Landsat 8 satellite.

Band	Pixel resolution (m)	Wavelength range (μm)	Main applications
Band 1 coastal	30	0.433–0.453	Shoreline observation
Band 2 blue	30	0.450–0.515	Water penetration
Band 3 green	30	0.525–0.600	Basic discrimination of vegetation
Band 4 red	30	0.630–0.680	Chlorophyll absorption degree
Band 5 NIR	30	0.845–0.885	Distinguish the wetness of the land
Band 6 SWIR 1	30	1.560–1.660	Distinguish bare ground
Band 7 SWIR 2	30	2.100–2.300	Discrimination of rocks and minerals
Band 8 pan	15	0.500–0.680	Enhanced resolution
Band 9 cirrus	30	1.360–1.390	Cloud detection usage

old method. Oli is divided into nine beams, as shown in Table 1.

4. Result Analysis and Discussion

The neural network architecture used in this experiment is very complex, and it must rely on more powerful technical computing power than the GPU. In this paper, the PyTorch framework based on CUDA will be used to recode the program. The machine configuration used in this article is shown in Table 2:

This paper uses the oli part of landsat8, which is a part of multispectral data and consists of 7 channels.

We can train the data set in this paper through deeplabv3+, deeplabv3 and the u-net network. In the preset weight part of the training system, we can initialize the DCNN part of the system with the magenet pretraining weight. One of the learning rates is the default starting point of the system, which is 50 rounds, that is, the learning rate is 0.1, 0.01, 0.001, 0.0001 in descending order. The optimizers of the three networks are designed as SGD (gradient descent algorithms), and the loss function uses the cross-entropy loss function. According to the new needs of economic construction, we will intensify the development of modern agricultural products and continue to promote the integration of primary, secondary, and tertiary industries.

4.1. Evaluation of Experimental Results for Three Different Models. This paper can clearly observe that the convergence speed and accuracy of the deeplabv3+ network are higher than those of the other two networks. Compared with the original deeplabv3 network, the deeplabv3+ network can more effectively combine the low-level features and high-level features. In Figures 4 and 5 It can be seen from Figure 4 that the deep labv3+ network has obvious convergence speed and accuracy.

Higher than the other two networks.

In this paper, the data set from May to the end of September 2017 is divided into four stages according to the growth and development stages, mainly because the crop characteristics of the four stages are quite different. There is a very simple neural network modeling for each data set. The DCNN module of Deep Labv3+ is used to replace the network model in this paper. First, taking the above SE model as an example, the attention model is directly referred

TABLE 2: Equipment configuration.

Project	Content
Central processing unit	Intel(R) core(TM) i7-7700K CPU @ 4.20 GHz
Memory	128 G
Graphics card	2 * NVIDIA GeForce GTX 1080Ti
Operating system	Ubuntu16.04
CUDA	Cuda10.0 with cudnn
Data processing	Python3.7, pytorch

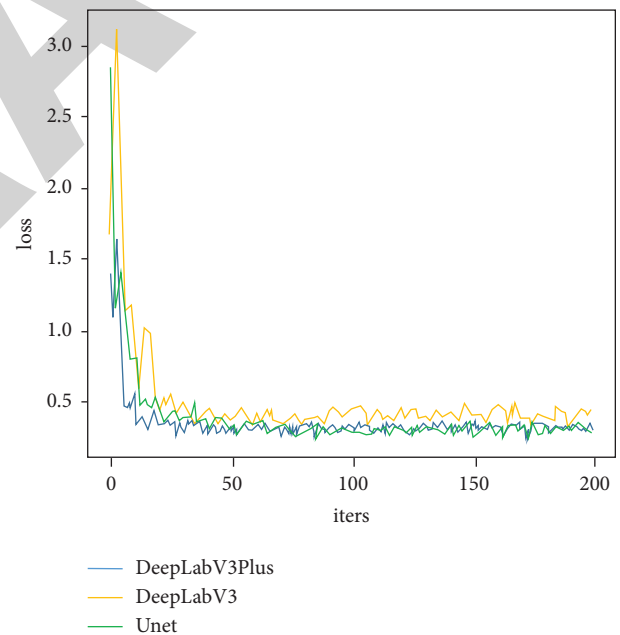


FIGURE 4: Loss graph.

to the network structure of ResNet. This chapter uses ResNet to compare with the three-depth ResNet system. The experimental results are shown in the following Table 3:

In the above table, it is obvious that the performance of Se ResNet network is slightly higher than that of other networks.

4.2. Crop Classification Using Multi Temporal Remote Sensing Data Based on Posterior Probability. The model trained based on a single time period is limited in its actual process.

TABLE 3: Comparison of experimental data collected under different network structures in different stages.

Name	SE-ResNext101	ResNext101	ResNet50	ResNet101	ResNet152
Data part 1	90	89.3	87.7	88	84
Data part 2	85.7	84.1	84.9	83.4	84.7
Data part 3	88.1	82.5	84.5	87.8	87.6
Data part 4	88.7	85.7	86.45	86.3	86.6

We always look forward to finding a way to break from the traditional method of determining crop types based on one month's survey. Our hope is that we can judge the types of crops within a specific range through several months of investigation and research, or prepare remote sensing maps of a region through many remote sensing satellites many times, so as to make the conclusion more credible.

The training idea of multi temporal image data is usually to align the remote sensing images taken by satellites at different time points for a single geographical coordinate region. That is to say, the training images must be in the same time dimension, and ensure that the same crops are logically consistent in the image pixel coordinates under the same geographical coordinates. Logically, such clipping is described to assume that there are clouds in the picture. Therefore, according to the logic of the neural network, the clipping here belongs to an unknown type, which will lead to mismatching with the image coordinates of the target position. Of course, a few errors can be tolerated, but a large number of errors always occur in actual operation.

4.3. Single Phase Division of Multi-Phase Data. Due to the limitation of Superposition Technology in integrated training, and considering the difficulty of training large time phase data sets, this chapter presents a single time training method and a step-by-step method of final synthesis. According to the growth process of these crops, the training of large temporal data sets includes four disjoint stages. In this way, we can use these data separately in the way in Chapter 3. Then, this paper trains these models with four sub-methods and summarizes the results of simulation training with some of these methods.

The first is the division of data sets. In this paper, the data set is divided according to four growth time stages, and the results in the following table are obtained. Each image is divided into $512 \times 512 \times 7$ sizes. Then, the four basic models are trained according to the data in the following table. Similar to the third chapter, the paper adopts three semantic segmentation modes, namely u-net, deeplav3, and deeplav3+ network. The posterior probability of each model prediction, the article also saves as 20% of the experimental data as the test set, and the test set does not participate in the model exercise. Table 4 shows the total amount of data collected for training.

After basic data separation, this paper can get four groups of separated models. How to merge the models according to stacking's integrated learning method in the prediction stage?

It notes the corresponding simple equilibrium and weighted average, meta learning method. Although the

TABLE 4: Data volume in the data set of each time period.

Dateset sub-part	Amount
Part 1	848
Part 2	992
Part 3	1072
Part 4	1328

TABLE 5: Confusion matrix obtained on the test set by Deepl avb3+ model of each training set.

	Ture Label	Prediction			
		Rice	Com	Forest	Other
Part 1	Rice	0.883	0.017	0.038	0.062
	Corn	0.096	0.776	0.093	0.034
	Forest	0.098	0.027	0.777	0.097
	Other	0.043	0.004	0.033	0.92
Part 2	Rice	0.928	0.037	0.024	0.011
	Corn	0.076	0.803	0.069	0.052
	Forest	0.117	0.123	0.741	0.019
	Other	0.015	0.013	0.007	0.964
Part 3	Rice	0.859	0.007	0.103	0.03
	Corn	0.117	0.771	0.1	0.011
	Forest	0.217	0.026	0.744	0.011
	Other	0.047	0.002	0.007	0.945
Part 4	Rice	0.86	0.042	0.083	0.015
	Corn	0.062	0.934	0.003	0.001
	Forest	0.116	0.001	0.876	0.006
	Other	0.121	0.009	0.033	0.836

above method is simpler and more convenient than the method based on its average value, the overall effect will be low. Because it is easy to receive the effect of the shortcomings of a single model by using this model, the models have their own advantages in compounding. This chapter gives a model compounding method based on a posteriori probability:

$$H(x) = \underset{j}{\text{Argmax}} \sum_{i=1}^4 g(y_j, h_i(x)), \quad (8)$$

$$g(y_j, h_i(x)) = P(y_j | h_i(x)),$$

$$y = f(W \cdot d(x)).$$

The each mode operation, indicates the conditional, that is, when the predicted value is $h(x)$, the actual real number is Y_j . By adding the accumulated pixel values of the target area through multiple modules according to the weight, the predicted value of the target area can be obtained.

In this paper, the four models are trained in data sets, and it is found that the deeplavb3+ network.

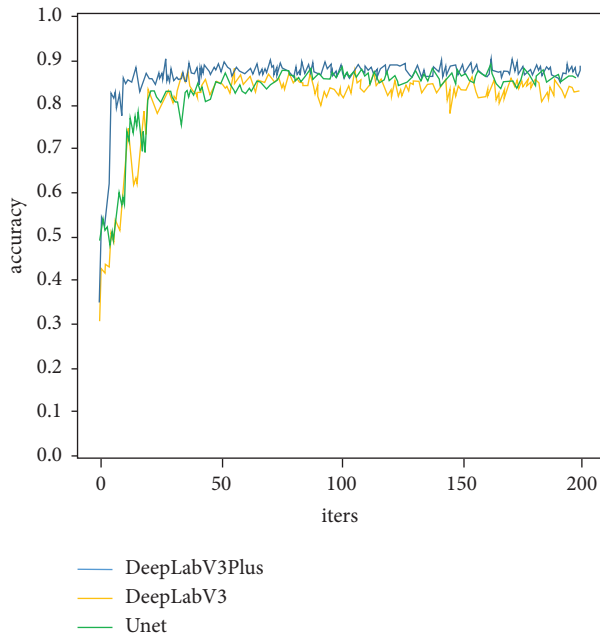


FIGURE 5: Accuracy graph.

Maintain the best performance in the. Based on the deeplabv3+ model, this paper tests the accuracy of the test set, as shown in Table 5.

This table is its corresponding test set to indicate the posttest probability value of each model under its own classification. From the perspective of posttest probability, the method in this paper will help to highlight the highest probability results in the classification results. Therefore, it can be assumed that the classification result of a model is rice, but it may also be wheat or forest land, although the probability is not very high. However, this paper has obtained the accuracy of the prediction results of various crops in each model from the actual test set and used it as the posttest probability. Among the classification results of the specified patterns, the posttest probability results are also very effective.

5. Conclusion

In order to overcome the limitation of the single period model in the practical production process, this paper also proposes a multi-temporal remote sensing data division based on a posteriori probability. The development of agricultural modernization and the implementation of the strategy of comprehensive rural revitalization have enabled modern agriculture and rural revitalization to enter a new stage of development in an all-round way. Only when we firmly follow the road of development with Chinese characteristics in the new era, proceed from practical problems, follow the path of socialist construction of new rural areas, new agriculture, and new rural areas, and realize the transformation and upgrading of modern agriculture from traditional Chinese rural areas, can we certainly achieve the modernization of Chinese rural areas, and achieve comprehensive rural revitalization.

In addition, because the probability analysis is discrete, the information in the control process is iterated once a year. In this way, as the crop growth cycle continues to change, the model will help to integrate the information of each growth cycle in the growth cycle. In addition, the Superposition Technology of the model can also tolerate the information loss in a specific period, such as the above effects brought by cloud coverage and long satellite return periods.

Data Availability

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

Conflicts of Interest

The author declares that they have no conflicts of interest.

Acknowledgments

The research is supported by: General project of Humanities and Social Sciences of Henan Provincial Department of Education in 2020 (2020-ZZJH-185).

References

- [1] C. Desconsi, "The production of development expectations and the pace of agricultural expansion: analysis of land-use practices by small farmers in northern Mato Grosso State – Brazil," *Caderno de Geografia*, vol. 30, no. 3, pp. 418–420, 2020.
- [2] A. Gomez, M. Narayan, and L. Zhao, "Effects of nano-enabled agricultural strategies on food quality: current knowledge and future research needs," *Journal of Hazardous Materials*, vol. 40, no. 1, Article ID 123385, 2020.
- [3] C. Nye, "Agriculture's "other" contingent labour source. Agricultural contractors and relationships of interdependence at the farmer-contractor interface," *Journal of Rural Studies*, vol. 78, no. 22, pp. 223–233, 2020.
- [4] S. Kamari and F. Ghorbani, "Extraction of highly pure silica from rice husk as an agricultural by-product and its application in the production of magnetic mesoporous silica MCM-41," *Biomass Conversion and Biorefinery*, vol. 11, no. 6, pp. 3001–3009, 2021.
- [5] I. Grigorescu, C. Dumitric, and M. Dumitracu, "Urban development and the (Re)use of the communist-built industrial and agricultural sites after 1990 the showcase of buchar-est-ilfov development region," *Land*, vol. 10, no. 12, pp. 227–236, 2021.
- [6] I. Oblova, I. Gerasimova, and J. Sishchuk, "Case-study based development of professional communicative competence of agricultural and environmental engineering students," *E3S Web of Conferences*, vol. 175, no. 1, Article ID 15035, 2020.
- [7] L. Teplova, V. Chuikova, and L. Afanasyeva, "Agricultural consumer credit co-operations one of the most important resources for the development of small and medium business," *Russian Journal of Management*, vol. 7, no. 4, pp. 51–55, 2020.
- [8] T. George, "A new look at agricultural development and the non-agriculture economy in low-income countries," *Global Food Security*, vol. 26, no. 13, Article ID 100449, 2020.
- [9] K. Mazur, "Innovative development of the agricultural market: research of modern tendencies and strategies,"

Retraction

Retracted: Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

- [1] X. Li, "Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 3590224, 9 pages, 2022.

Research Article

Prediction and Analysis of Housing Price Based on the Generalized Linear Regression Model

Xinshu Li 

Faculty of Science, University of Melbourne, Melbourne, Victoria 3010, Australia

Correspondence should be addressed to Xinshu Li; lxs_elle@163.com

Received 14 July 2022; Revised 22 August 2022; Accepted 30 August 2022; Published 29 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Xinshu Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to improve the reliability of housing price prediction and analysis, this article combines the generalized linear regression model to build a real estate price prediction model and analyzes the basic knowledge of data mining. On the basis of this prior knowledge, this article investigates the cluster analysis algorithm and selects the generalized linear regression model as the research focus based on its definition and the characteristics of stock data. Moreover, this article analyzes the estimation methods of the generalized linear regression model and the nonparametric regression model, and then gives the estimation method of a partial linear model. In addition, this article verifies the validity of the model proposed in this article by means of simulation research. Through the simulation and comparison experiments, it can be seen that the housing price prediction system based on the generalized regression model proposed in this article has a high housing price prediction accuracy.

1. Introduction

Nowadays, many people invest in real estate, because sometimes, it can bring a lot of capital income, which has developed very violently in our country. Investment in real estate is also a reflection of a local real estate development situation to a certain extent. For example, if investment is hot, it means that the development is unbalanced, and the supply is in short supply. If investment is cold, it means that the recent real estate market is relatively stable. Moreover, it is also a reflection of the development of a city [1]. However, it does not mean that the more the investment, the better the real estate will develop, the more benefits the investors will get, and there may be negative situations, and some places have suffered from this situation. Moreover, real estate investment in many places has gone wrong, exceeding demand. This is a waste of resources and irresponsible for people's lives, and there will also be situations where workers are not paid, causing social chaos [2]. Therefore, in order to prevent these situations from happening, it is very important to control the investment in real estate. In the process of urbanization, the government should actively formulate reasonable measures to control the situation and ensure that

the proportion of investment in fixed assets remains around 25%. Moreover, this limit is not fixed. It also depends on the development of the city. After the initial stage of urban development, this ratio can be appropriately reduced, because there is no need for so many houses at this time [3]. What we need to know very clearly is that we cannot make reasonable improvement measures in a timely manner. The reason for this is that we always know the problem after the situation arises, and it has a certain lag effect. Therefore, when formulating, it is necessary to fully and comprehensively consider the development situation and changes in the relationship between supply and demand, strive to achieve standards that can meet the long-term development situation, and try to avoid the rise in housing prices due to incorrect measures. This is not only a guarantee for the stable development of society, but also a guarantee for people's lives. Only when people are stable can a country develop well [4].

At present, the methods of housing price forecasting can be divided into two categories. One is a multifactor analysis model based on the analysis of the influencing factors of housing prices, and the other is a single-factor analysis based on time series. In the multivariate analysis models, most of

them only consider the parameters that affect housing prices, such as multiple regression models, but do not consider the nonparametric factors. The absence of some nonparametric influencing factors is likely to lead to a decrease in the accuracy of the prediction model [5]. In the process of reviewing the literature, only one paper was found that used a partial linear model to predict the average sales price of commercial housing across the country, and the results of the paper showed that the partial linear model was better than the linear regression model in predicting housing prices. This is because the partial linear model considers both linear and nonlinear factors affecting housing prices [6]. However, considering that there are many factors affecting housing prices, there will be a curse of dimensionality when using a partial linear model. The additive model can eliminate the disaster of dimensionality, so it is of great practical significance to build a housing price prediction model based on the additive model. In addition, it also has important theoretical significance to establish a housing price prediction model on the basis of the additive model [7]. In different places, housing prices are affected by local policies and special events, and the fluctuation laws of housing prices are also different.

The definition of real estate in economics is mainly divided into narrow sense and broad sense. Real estate in the broad sense is understood as the sum of real estate commodity relations generated in the exchange process [8]; real estate in the narrow sense is understood as a place used for real estate rental, sale, mortgage, and other commodity transactions [9]. It is worth mentioning that the real estate price in this article is both an equilibrium price and a market price [10].

In the real estate market, consumers refer to buyers, suppliers refer to real estate developers, and the equilibrium price refers to the equilibrium price between the quantity of a certain type of real estate provided by real estate developers. Real estate usually cannot play its role as a commodity independently, and its main value is reflected in its use, which is a kind of induced demand [11].

As a commodity, real estate conforms to the relationship between supply and price in economic theory, but the real estate itself has a large investment scale, a long construction period, and the visibility of recoverable interest rates is longer than that of general commodity cycles. Generally speaking, when the supply of real estate increases, housing prices will rise, which is a positive relationship. However, the entire process of real estate supply is relatively long, and it will have an informatization impact on the market during the construction period, such as investment in real estate. Elements such as quantity, real estate development, and related infrastructure construction will enter the market in the form of information, thereby affecting the judgment and analysis of various market players, and then affecting the changes in real estate prices [12]. However, the effect of this influence path is relatively slow, and the price changes are relatively lag, and the lag time is positively correlated with the risk it brings. The lag time here refers to the time that

changes in the supply of real estate act on prices. For developers, the greater the risk, the less incentive they have to develop real estate, the smaller the quantity of real estate provided, the insufficient supply to meet the demand, and the price rise; on the contrary, when the information is more comprehensive and the uncertainty is less, the lower the risk, the greater the motivation for real estate developers to develop, the supply exceeds the demand, and the price decreases [13].

Whether it is a house buyer or a real estate developer, they will make a psychological assessment of the future economic situation when making decisions. This is mainly because people are uncertain about the risks they will face in the future, and an early warning mechanism will be generated in advance. In economics, anticipation is defined as a psychological effect, which refers to the expected effect that people will collect, analyze, and judge before making economic decisions [14]. The real estate market is of great significance in my country and has always been the focus of all sectors of society. However, the information people have is limited, and it is impossible to effectively predict the future real estate market trends and avoid the impact of the economic environment on real estate. Therefore, expectations are correct. Both the buyer and the real estate developer are very important. For home buyers, they will have a rough expectation of future house prices based on the current and future real estate market conditions, as well as the trend of real estate prices in the past period of time. If the real estate price is on an upward trend in the future, then the demand for housing will also increase whether it is for consumption demand or investment demand [15]; if the real estate price is on a downward trend in the future, then demand will also drop. In addition, the surrounding supporting equipment of real estate also affects the psychological expectations of home buyers, including traffic conditions, infrastructure construction, and green area, especially schools, hospitals, and other factors [16].

This article uses the generalized linear regression model to construct the real estate price prediction model, verifies the validity of the model in this article by comparing the actual data and simulation research, and promotes the accuracy of subsequent real estate price prediction.

2. Generalized Linear Regression Prediction

2.1. Cluster Analysis. Data similarity refers to the calculation method of similarity between data objects. There are generally two methods to describe the similarity between data objects: one is the distance, and the other is the similarity coefficient. The so-called distance refers to depicting the distance between objects according to the relationship between close and distant, that is, putting the closest ones together and combining them into one class. The similarity coefficient is a numerical value between 0 and 1, indicating the similarity between the two. When the similarity coefficient is closer to 1, the similarity between the two is greater; when the similarity coefficient is closer to 0, the similarity

between the two is smaller. Cluster analysis is often divided into R-type clustering and Q-type clustering. Among them, the R-type clustering often uses the correlation coefficient to describe the similarity, and the Q-type clustering often uses the distance measurement to describe the similarity.

2.1.1. Two Categorical Variables. For binary variables, the similarity matrix is used to describe the similarity between objects, and it considers observations (x_i, x_j) , where $x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$, $x_j^T = (x_{j1}, x_{j2}, \dots, x_{jp})$ and $x_{ik}, x_{jk} \in \{0, 1\}$.

Among them, there exists $a_t, t = 1, 2, 3, 4$, and its value depends on the pair (x_i, x_j) . Therefore, the similarity characterization formula used in daily life can be given as

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}, \quad (1)$$

where δ, λ is the weight coefficient.

2.1.2. Continuous Variables

(1) Numerical Continuous Variables. The following formula is defined as

$$d_{ij} = \|x_i - x_j\| = \left(\sum_{k=1}^n |x_{ik} - x_{jk}|^r \right)^{1/r}, \quad (2)$$

where x_{ik} represents the k th attribute value of the i th object, $d_{ii} = 0, \forall i = 1, 2, \dots, n$.

When $r = 1$, the above formula is the absolute distance; when $r = 2$, the above formula is the Euclidean distance; when $r = \infty$, the above formula is the Chebyshev distance.

(2) Vector-Type Continuous Variable. If we encounter numerical variables that are not on the same metric, we first need to standardize them, so we introduce a more general metric—Mahalanobis distance, where the data object is in the form of a vector:

$$d = (x - y)^{-1} \sum (x - y)^T. \quad (3)$$

The similarity coefficient is a numerical value between 0 and 1, indicating the similarity between the two. The similarity between the two is smaller. In the following, a few commonly used formulas are introduced:

(1) Exponential similarity coefficient formula

$$r_{ij} = \frac{1}{m} \sum_{k=1}^m \exp \left[-\frac{3}{4} \frac{(x_{ij} - x_{jk})^2}{s_k^2} \right] (i, j = 1, 2, \dots, n), \quad (4)$$

$$s_k = \left[\frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2 \right]^{(1/2)} \quad k = (1, 2, \dots, m), \quad (5)$$

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik} \quad k = (1, 2, \dots, m).$$

(2) Cosine formula of the included angle

$$r_{ij} = \frac{|\sum_{k=1}^m x_{ik} x_{jk}|}{\left[\sum_{k=1}^m x_{ik}^2 \sum_{k=1}^m x_{jk}^2 \right]^{(1/2)}}, (i, j = 1, 2, \dots, n). \quad (6)$$

(3) Correlation coefficient formula

$$r_{ij} = \frac{|\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)|}{\left[\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^m (x_{jk} - \bar{x}_j)^2 \right]^{(1/2)}}, (i, j = 1, 2, \dots, n), \quad (7)$$

where $\bar{x}_i = (1/m) \sum_{k=1}^m x_{ik}$; $\bar{x}_j = (1/m) \sum_{k=1}^m x_{jk}$.

The algorithm flow is as follows: first, the algorithm needs to determine the number of cluster categories k and perform initial clustering on the dataset.

The algorithm steps are as follows:

(1) The algorithm selects k initial clustering centers: $z_1^{(1)}, z_2^{(1)}, \dots, z_k^{(1)}$, where the superscript indicates the number of iterative operations in the clustering process.

(2) When the r th iteration has been performed, if for a certain sample x , there is

$$d(x, z_j^{(r)}) = \min \{d(x, z_i^{(r)}), i = 1, 2, \dots, k\}, \quad (8)$$

then, $x \in S_j^{(r)}$. $S_j^{(r)}$ is a subset of samples with $z_j^{(r)}$ as the cluster center. In this way, that is, the principle of minimum distance, all samples are assigned to k cluster centers.

(3) The algorithm calculates the reclassified cluster centers:

$$z_j^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{x \in S_j^{(r)}} x, (j = 1, 2, \dots, K), \quad (9)$$

where $n_j^{(r)}$ is the number of samples included in $S_j^{(r)}$.

(4) If $z_i^{(r+1)} = z_j^{(r)}, j = 1, 2, \dots, K$, the algorithm ends; otherwise, the algorithm goes to (2).

2.2. Random Process. Such a random process is called a Markov chain, if it takes only a finite or listable number of values, and for any Q and any state W, we have

$$\begin{aligned} P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} \\ = P\{X_{n+1} = j | X_n = i\} \\ = P_{ij}, \end{aligned} \quad (10)$$

where $X_n = i$ indicates that the process is in state i at time n , and $\{0, 1, 2, \dots\}$ is called the state space of the process, denoted as S . The above formula describes the characteristics of the Markov chain, which is called the Markov property. When given the past states X_0, X_1, \dots, X_{n-1} and the present state X_n , the conditional distribution of the future state X_{n+1} is independent of the past state and only depends on the present state. That is to say, the Markov chain has no

aftereffect. We also conduct research on housing prices based on the ineffectiveness of Markov chains.

The conditional probability $P_{ij} = P\{X_{n+1} = j | X_n = i\}$ is the one-step transition probability of the Markov chain $\{X_n, n = 0, 1, 2, \dots\}$, referred to as the transition probability, denoted as P_{ij} . It represents the probability of being in state i and moving to state j next. When the transition probability $Q = 1$ of the Markov chain is only related to the states i, j , and it has nothing to do with n , it is called a time-aligned Markov chain. Otherwise, it is called a non-time-aligned chain. The transition probability matrix is given as

$$P = (p_{ij}) = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0,N-1} \\ p_{10} & p_{11} & \cdots & p_{1,N-1} \\ \vdots & \vdots & \vdots & \vdots \\ p_{N-1,0} & p_{N-1,1} & \cdots & p_{N-1,N-1} \end{bmatrix}, \quad (11)$$

where P is called the transition probability matrix, generally referred to as the transition matrix. Since the probability is non-negative and the process must transition to some state, it is easy to see that $p_{ij} (i, j \in S)$ has the following properties:

- (1) $p_{ij} \geq 0, i, j \in S$;
- (2) $\sum p_{ij} = 1, \forall i \in S$.

In practical applications, the one-step transition probability is generally difficult to obtain directly, and the method of using frequency instead of probability is often considered to count the number of transitions m_{ij} from a fixed state i to other states j , and count the total number of times in state i . Therefore, we get $p_{ij} = (m_{ij}/m_i)$.

The n -step transition probability is called the conditional probability:

$$p_{ij}^{(n)} = P(X_{m+n} = j | X_m = i), i, j \in S; m \geq 0; n \geq 1. \quad (12)$$

It is the n -step transition probability of the Markov chain, and correspondingly, $P^{(n)} = (p_{ij}^{(n)})$ is called the n -step transition probability matrix. When $n = 1$, $p_{ij}^{(1)} = p_{ij}$, $P^{(1)} = P$. In addition, it stipulates

$$p_{ij}^{(0)} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \quad (13)$$

It has nothing to do with the state that the intermediate $n-1$ steps transition through.

Classification and nature of states:

- (1) Irreducible Markov chain

State i is said to be reachable to state $j (i, j \in S)$, and if there exists $n \geq 0$ such that $p_{ij}^n > 0$.

We classify any two intercommunication states into a class, the states in the same class should all be intercommunicated, and any state cannot belong to two different classes at the same time.

From this, we can get the definition of irreducible Markov chain.

Moreover, it is specially stipulated that when the abovementioned set is an empty set, the period of i is said to be infinite.

- (2) Always return state.

For any state i, j , $f_{ij}^{(n)}$ is the probability of reaching j for the first time from i after n steps, and obviously, $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$. If $f_{ij} = 1$, state j is said to be a constant return state. If $f_{ij} < 1$, the state j is said to be a nonrecurrent state or an instantaneous state.

For the constant return state i , u_i represents the average number of steps (time) required to start from i and then return to i , as shown in the following formula:

$$u_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}. \quad (14)$$

Among them, if $u_i < +\infty$, then i is called the normal return state. If $u_i = +\infty$, then i is called the zero return state. If i is a normal return state and is aperiodic, it is called an ergodic state. If i is an ergodic state and $f_{ii}^{(1)} = 1$, i is called an absorbing state, and obviously $u_i = 1$.

For an irreducible aperiodic Markov chain, if it is ergodic, then $\pi_j = \lim_{n \rightarrow \infty} p_{ij}^{(n)} > 0 (j \in S)$ is a stationary distribution and the only stationary distribution. A stationary distribution does not exist if the states are all instantaneous or all zeros are recurring. The stationary distribution satisfies the following formula:

$$\begin{cases} \sum_{j \in S} \pi_j = 1 \\ \sum_{i \in S} \pi_i p_{ij}^{(n)} = \pi_j. \end{cases} \quad (15)$$

For the traversed Markov chain, if all states are connected and are normal return states with period 1, the limit is given as

$$\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j, j \in S. \quad (16)$$

It is called the limit distribution of the Markov chain, that is, $\pi_j = (1/u_j)$. The limiting distribution is the stationary distribution and the only stationary distribution.

2.2.1. Calculation of the One-Step Transition Probability Matrix.

In practical applications, it is generally difficult to directly obtain the one-step transition probability, so the method of using frequency instead of probability is often considered.

First, the frequency transition matrix M is obtained, that is, to count the number m_{ij} of transitions from a fixed state i to other states j :

$$M = \begin{bmatrix} m_{00} & m_{01} & \cdots & m_{0,N-1} \\ m_{10} & m_{11} & \cdots & m_{1,N-1} \\ \vdots & \vdots & \vdots & \vdots \\ m_{N-1,0} & m_{N-1,1} & \cdots & m_{N-1,N-1} \end{bmatrix}. \quad (17)$$

Then, the total number of times in state i is counted, which is calculated according to the following formula:

$$P_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{j \in S} m_{ij}} & \sum_{j \in S} m_{ij} > 0, \\ 0 & \sum_{j \in S} m_{ij} = 0, \end{cases} \quad (18)$$

where $m_i = \sum_{j=0}^N m_{ij}$

Finally, the transition probability matrix P is calculated, that is, the probability is replaced by frequency:

$$P = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0,N-1} \\ P_{10} & P_{11} & \cdots & P_{1,N-1} \\ \vdots & \vdots & \vdots & \vdots \\ P_{N-1,0} & P_{N-1,1} & \cdots & P_{N-1,N-1} \end{bmatrix}. \quad (19)$$

2.2.2. Markov Test. Before using the Markov chain to build a prediction model, the Markov property of the sequence $\{X_t, t \in T\}$ must be checked. The test is performed using the χ^2 statistic.

Test statistics:

$$\tilde{\chi}^2 = 2 \sum_{i=1}^m \sum_{j=1}^m n_{ij} \left| \log \left(\frac{P_{ij}}{\hat{P}_{.j}} \right) \right|, \quad (20)$$

where $\hat{P}_{.j} = \sum_{i=1}^m n_{ij} / \sum_{i=1}^m \sum_{k=1}^m n_{ik}$ and $P_{ij} = n_{ij} / \sum_{k=1}^m n_{ik}$. When m is larger, the above obeys the chi-square distribution, and the degree of freedom is $(m-1)^2$.

The confidence level α is chosen. If the statistic is $\tilde{\chi}^2 > \chi_{\alpha}^2 (m-1)^2$, then the null hypothesis is rejected, and the sequence $\{X_t, t \in T\}$ is considered to be Markov's, and the model can be used to make predictions after passing the Markov test, and vice versa.

2.2.3. Stable Distribution. From the transition probability matrix, a stationary distribution π_j is derived.

$$\begin{cases} \sum_{j \in S} \pi_j = 1, \\ \sum_{i \in S} \pi_i P_{ij}^{(n)} = \pi_j. \end{cases} \quad (21)$$

2.2.4. Making Predictions Based on the Initial State. It is known that the initial state is x_0 , and if it is in state i , then $x_0 = (0, \dots, 0, 1, 0, \dots, 0)$, that is, the probability of being in state i at this time is 1, and the rest are 0.

$$x_1 = x_0 * p. \quad (22)$$

The state of the system at time $t+1$ can be obtained.

We set $x_1 = (x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(n)})$. According to the principle of maximization, we can get

$$\max \{(x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(n)})\} = x_1^{(j)}. \quad (23)$$

Therefore, the next moment is most likely to be in state j .

2.3. Time Series Analysis

2.3.1. Stationary Time Series. Stationary time series mean and variance do not change systematically and do not change periodically. Each observation value in this type of series basically fluctuates at a fixed level. Although the degree of fluctuation is different in different time periods, there is no certain rule, and its fluctuation can be regarded as random. Stationary distribution includes general autoregressive model $AR(p)$, moving average model $MA(q)$, and autoregressive moving average model $ARMA(p, q)$.

An autoregressive model is a process of using itself as a regression variable, that is, a linear regression model that uses the linear combination of random variables at several previous moments to describe random variables at a certain time in the future. It is a common form in time series.

Consider a time series y_1, y_2, \dots, y_n , p -order autoregressive model (abbreviated $AR(p)$) indicating that Z in the series is a linear combination of the first p series and a function of the error term; the general form of the mathematical model is

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \omega_t, \quad (24)$$

where p is called the order of the autoregressive model, denoted as $AR(p)$. $\alpha_1, \dots, \alpha_p$ is the model parameter, ω_t is white noise with mean 0 and variance σ^2 .

There are similarities between moving average $MA(q)$ models and autoregressive models.

If a univariate time series data is $\{y_t; t = 1, 2, \dots\}$,

$$y_t = \omega_t + \beta_1 \omega_{t-1} + \cdots + \beta_p \omega_{t-p}. \quad (25)$$

AR models are attempts to capture and explain the momentum and mean reversal effects of financial trading markets. The MA model is an attempt to capture and explain the observed oscillatory effects in the white noise term, which can be understood as the effects of unintended events that affect the observed process.

The $ARMA$ model is a combination of the two. Its main disadvantage is that it ignores the fluctuation clustering phenomenon often seen in financial market time series data. The model formula is as follows:

$$\begin{aligned} y_t &= \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \omega_t + \beta_1 \omega_{t-1} + \cdots + \beta_q \omega_{t-q} \\ &= \sum_{i=1}^p \alpha_i y_{t-i} + \omega_t + \sum_{i=1}^q \beta_i \omega_{t-i}. \end{aligned} \quad (26)$$

2.3.2. Nonstationary Time Series. Nonstationary distributions include (p, d, q) .

Finally, the $ARIMA$ model is used as a comparative model to highlight the accuracy and practicability of the model built in this subject. Figure 1 shows the specific modeling steps.

If the sequence is nonstationary, it can be made stationary with the help of difference operation. Nonstationary series can be written as

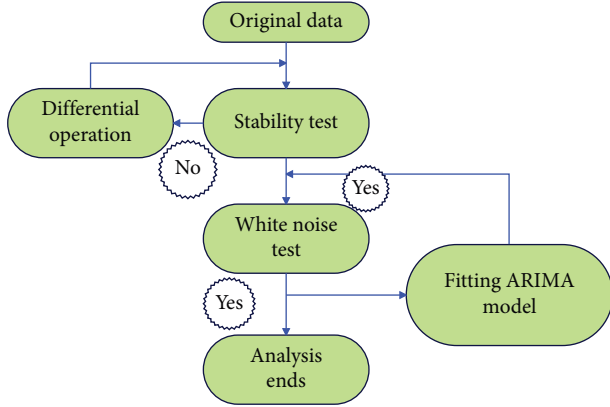


FIGURE 1: Flowchart of ARIMA modeling steps.

TABLE 1: Determination of p , q -order.

Model	ACF	PACF
AR(p)	Trailing	p -order truncation
MA(q)	q -order truncation	Trailing
ARMA(p, q)	Trailing	Trailing

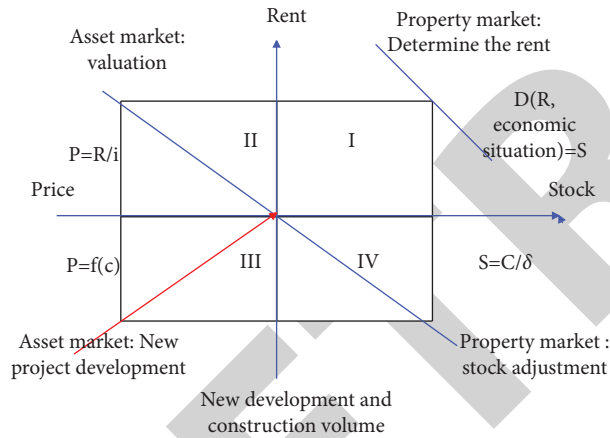


FIGURE 2: Four-quadrant equilibrium diagram of the real estate market.

$$X_t = \mu_t + \xi_t = \sum_{j=0}^d \beta_j t^j + \Theta(B)\varepsilon_t. \quad (27)$$

Among them, $\{\varepsilon_t\}$ is a white noise sequence with zero mean.

For example, the first difference is

$$\nabla X_t = X_t - X_{t-1}. \quad (28)$$

The formula for calculation is as follows:

$$\text{ACF}(k) = \rho_k = \frac{\text{cov}(y_t, y_{t-k})}{\text{Var}(y_t)}, \quad (29)$$

where k represents the number of lag periods.

The determination of the p , q -order of the ARIMA(p, d, q) model is determined by ACF and PACF, as shown in Table 1.

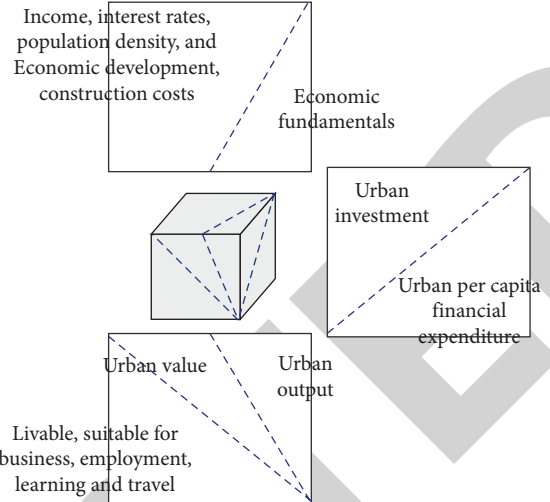


FIGURE 3: Cube of influencing factors of urban real estate prices.

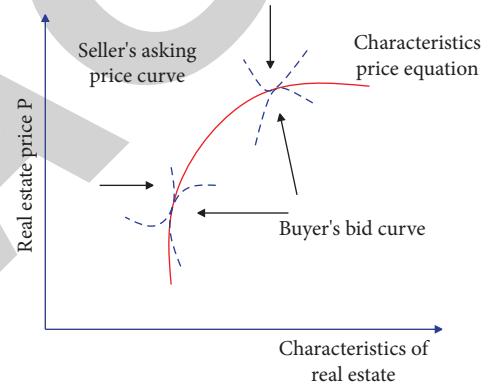


FIGURE 4: Equilibrium process of house price in the hedonic price model.

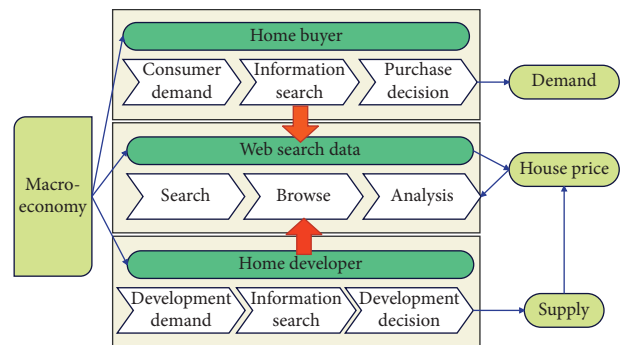


FIGURE 5: Conceptual framework.

Null hypothesis: the residual sequence is a white noise sequence.

$$H_0: \rho_1 = \rho_2 = \dots = \rho_m = 0, \forall m \geq 1. \quad (30)$$

Alternative hypothesis: the residual sequence is a non-white noise sequence.

$$H_1: \text{there is at least one } \rho_k \neq 0, \forall m \geq 1, k \leq m$$

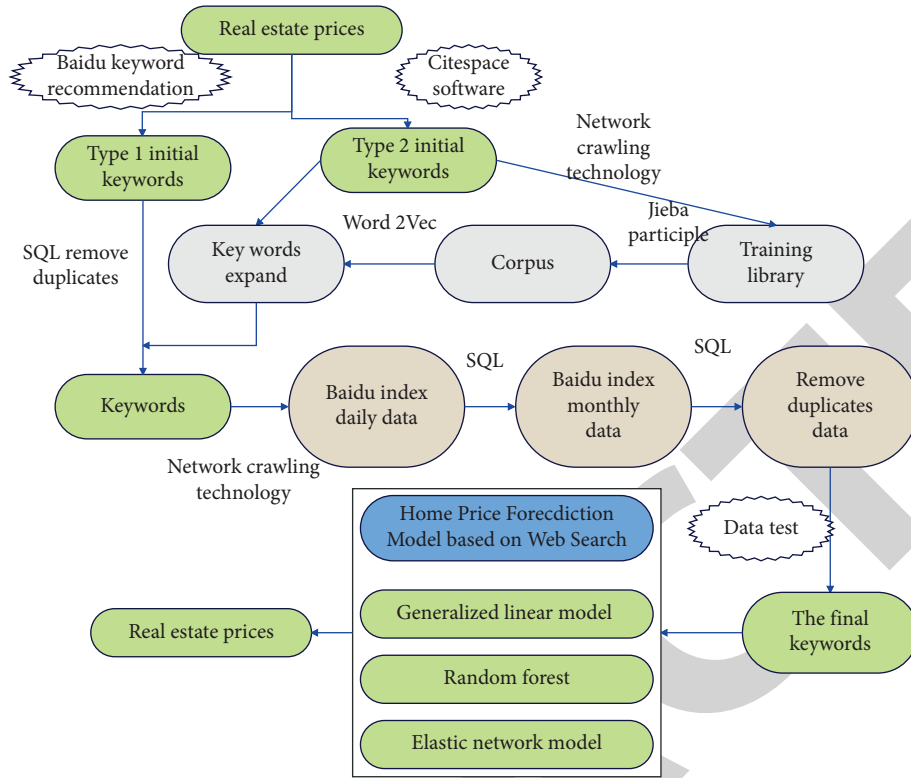


FIGURE 6: Technical framework.

Test statistics:

$$LB = N(N + 2) \sum_{k=1}^m \frac{\hat{P}_k^2}{N - K} \sim \chi^2(m). \quad (31)$$

If the test result is to reject the null hypothesis, it means that the residual sequence is a nonwhite noise sequence, and the useful information in the residual sequence has not been fully extracted. It further shows that the fitted model is not significant. If the residual sequence is a white noise sequence, the null hypothesis is not rejected, indicating that the fitted model is significantly effective.

3. Prediction and Analysis of Housing Price Based on Generalized Linear Regression Model

Due to the duality of real estate, it can not only provide services for households as consumption, but also provide assets for households as investment. Therefore, the real estate market can be regarded as consisting of three sub-markets that interact with each other: the real estate use market, the real estate asset holding market, and the real estate production market. According to their interrelationships, a four-quadrant model is constructed using the rectangular coordinate quadrants, as shown in Figure 2.

Based on the idea of The Economic Cycle Cube, a cube of influencing factors of urban real estate prices is constructed, as shown in Figure 3.

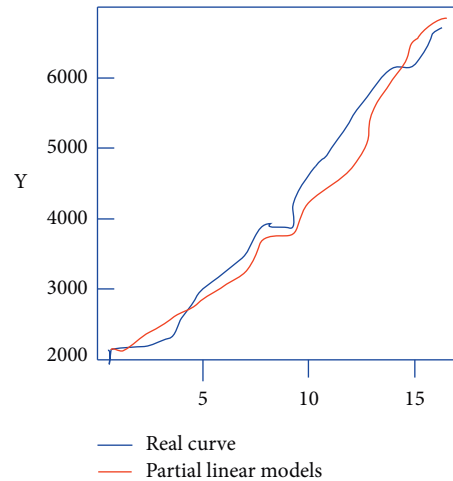


FIGURE 7: The real curve of the national average sales price of commercial housing and the fitted curve simulated by a partial linear model.

A heterogeneous commodity has different characteristics that meet the needs of consumers, and the implicit price of these heterogeneous characteristics can be calculated by regression, as shown in Figure 4.

Based on the theory of supply and demand, this article divides the participants in the real estate market into real estate developers and home buyers, and divides their behavior into three stages, as shown in Figure 5.

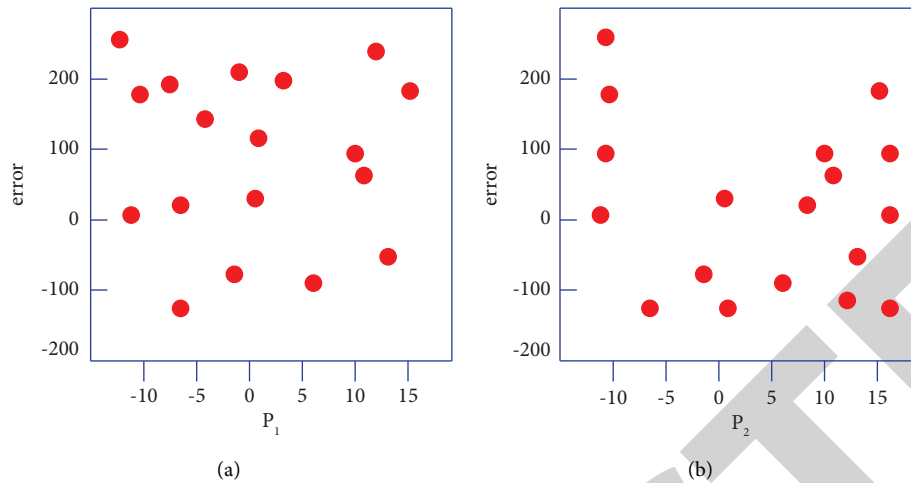


FIGURE 8: Scatter plot of error terms and principal components. (a) Scatter plot of the error term and the first principal component and (b) scatter plot of the error term and the second principal component.

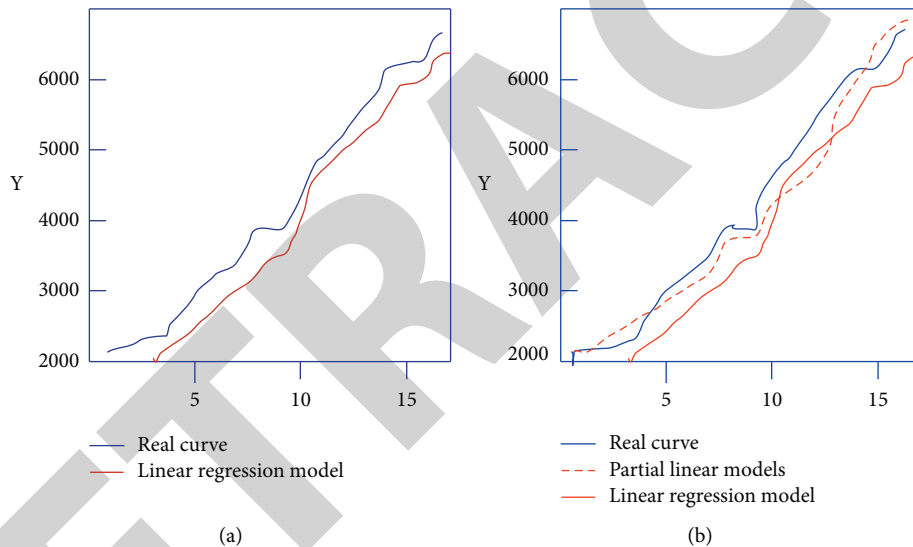


FIGURE 9: Fitting curve. (a) The real curve of the national average selling price of commercial housing and the fitted curve simulated by a linear regression model and (b) the true curve of the national average sales price of commercial housing and the fitted curve of the two models.

In terms of keyword selection and network information crawling processing, this article mainly adopts the technical means of machine learning, as shown in Figure 6.

According to the established partial linear model, the fitting result of the model is calculated and compared with the real value of the average sales price of commercial housing, and the two curves are drawn, as shown in Figure 7.

It can be seen that the fitted value is relatively close to the real value, and the fitting curve and the real curve have the same trend of change, and the simulation effect is good. The scatter plots of the error term and the two principal components (GDP (100 million yuan) and urban population (10,000 people)) are made, respectively, as shown in Figures 8(a) and 8(b).

According to the established linear regression model, the fitting result of the model is calculated and compared with the real value of the average sales price of commercial

housing, and the two curves are drawn, as shown in Figure 9(a). In order to compare the fitting results of the partial linear model and the linear regression model more clearly, we draw the real curve of the national average sales price of commercial housing, the fitted curve of the partial linear model, and the fitted curve of the linear regression model in one graph, as shown in Figure 9(b).

It can be seen from the above research that the housing price prediction system based on the generalized regression model proposed in this article has a high housing price prediction accuracy.

4. Conclusion

The demand for commercial housing is generally divided into self-occupied demand, investment demand, and speculative demand. Self-occupation demand is self-occupation;

Research Article

DBP-iDWT: Improving DNA-Binding Proteins Prediction Using Multi-Perspective Evolutionary Profile and Discrete Wavelet Transform

Farman Ali ¹, Omar Barukab ², Ajay B Gadicha,³ Shruti Patil,⁴ Omar Alghushairy,⁵ and Akram Y. Sarhan⁶

¹Department of Elementary and Secondary Education, Peshawar, Khyber Pakhtunkhwa, Pakistan

²Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Jeddah 21911, Saudi Arabia

³Department of Computer Science and Engineering, P.R. Pote, Collage of Engineering and Management, Amravati, India

⁴Symbiosis Institute of Technology, Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International University, Pune, India

⁵Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

⁶Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia

Correspondence should be addressed to Farman Ali; farman335@yahoo.com

Received 29 June 2022; Revised 19 August 2022; Accepted 9 September 2022; Published 28 September 2022

Academic Editor: Xiaolong Zhou

Copyright © 2022 Farman Ali et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DNA-binding proteins (DBPs) have crucial biotic activities including DNA replication, recombination, and transcription. DBPs are highly concerned with chronic diseases and are used in the manufacturing of antibiotics and steroids. A series of predictors were established to identify DBPs. However, researchers are still working to further enhance the identification of DBPs. This research designed a novel predictor to identify DBPs more accurately. The features from the sequences are transformed by F-PSSM (Filtered position-specific scoring matrix), PSSM-DPC (Position specific scoring matrix-dipeptide composition), and R-PSSM (Reduced position-specific scoring matrix). To eliminate the noisy attributes, we extended DWT (discrete wavelet transform) to F-PSSM, PSSM-DPC, and R-PSSM and introduced three novel descriptors, namely, F-PSSM-DWT, PSSM-DPC-DWT, and R-PSSM-DWT. Onward, the training of the four models were performed using LiXGB (Light eXtreme gradient boosting), XGB (eXtreme gradient boosting, ERT (extremely randomized trees), and Adaboost. LiXGB with R-PSSM-DWT has attained 6.55% higher accuracy on training and 5.93% on testing dataset than the best existing predictors. The results reveal the excellent performance of our novel predictor over the past studies. DBP-iDWT would be fruitful for establishing more operative therapeutic strategies for fatal disease treatment.

1. Introduction

DNA-binding proteins perform many crucial activities like DNA translation, repair, translation, and damage [1]. DBPs are directly encoded into the genome of about 2–5% of the prokaryotic and 6–7% of eukaryotic [2]. Several DBPs are responsible for gene transcription and replication, and some DBPs shape the DNA into a specific structure, called chromatin [3]. The research on DBPs is significant in diverse fatal

disease treatment and production of drugs. For instance, nuclear receptors are the key components of tamoxifen and bicalutamide medicines which are used in cancer treatment. Similarly, glucocorticoid receptors participate in the production of dexamethasone, which is utilized in autoimmune diseases and anti-inflammatory, allergies, and asthma treatment [4–6]. Onward, Inhibitor of DNA-binding (ID) proteins are closely related to tumor-associated processes including chemoresistance, tumorigenesis, and angiogenesis. In

addition, ID proteins are also directly concerned with lung, cervical, and prostate cancers [7].

Protein sequences are rapidly growing in the online database. A series of predictors were developed for diverse biological problems including iRNA-PseTNC [8], iACP-GAEnsC [9], cACP-2LFS [10], DP-BINDER [11], Deep-AntiFP [12], cACP [13], iAtbP-Hyb-EnC [14], iAFPs-EnC-GA [15], and cACP-DeepGram [16]. It is highly demanding to predict DBPs by computational approaches. Several predictors were introduced using the primary sequential information and structural features. Structured-based predictors produce good prediction results, but structural features for all proteins are unavailable. Some of the structure-based protocols are iDBPs [17], DBD-Hunter [18], and Seq(DNA) [19]. Sequence-based systems have been developed using sequential information, more convenient and easy to employ for large datasets. Therefore, many sequence-based systems were adopted for DNA-binding proteins identification. Among these methods: DBP-DeepCNN [20], DNA-Prot [21], iDNA-Prot [22], iDNA-Prot[dis] [23], Kmer1 + ACC [24], Local-DPP [25], DBPPred-PDSD [26], DPP-PseAAC [27], and StackDPPred [28]. Consequently, Li et al. extracted features by a convolutional neural network (CNN) and Bi-LSTM [29]. Onward, Zhao et al. the features of the proteins are analyzed by six methods and classification is performed with XGBoost [30]. Each computational method contributed well to enhancing the prediction of DBPs. However, more efforts are needed to improve prediction of DBPs. Considering this, a new method (DBP-iDWT) is established to identify DBPs accurately. The contribution of our research is as follows:

- (i) Designed three new feature descriptors i.e., F-PSSM-DWT, PSSM-DPC-DWT, and R-PSSM-DWT
- (ii) LiXGB is applied for model training and prediction
- (iii) Constructed a new computational model (DBP-iDWT) for improving DBPs identification

In addition to LiXGB, the features set is fed into three classification algorithms, namely ERT, XGB, and Adaboost. The efficacy of each classifier was assessed with ten-fold test, while the generalization capability was assessed by a testing set. LiXGB using R-PSSM-DWT secured the highest prediction outcomes than past methods. The flowchart of the DBP-iDWT is depicted in Figure 1.

The rest portion of the manuscript comprises three parts. Section 2 comprises details regarding datasets and methodologies; in Section 3, the performance of classifiers has illustrated; and Section 4 summarizes the conclusion.

2. Materials and Methods

2.1. Selection of Datasets. We selected two datasets from the previous work [31]. One dataset (PDB14189) is employed model training and the other dataset is deployed as a testing dataset. PDB14189 was collected from the UniProt database [32]. To design a standard dataset, they removed more than 25% of similar sequences by CD-HIT toolkit. The final

training dataset comprises 7129 DBPs and 7060 non-DBPs. The independent set was retrieved by a procedure explained in reference [33]. The similar sequences with a cutoff value 25% are removed. The final testing dataset contains 1153 DBPs and 1119 non-DBPs.

2.2. Feature Descriptors. In this work, the patterns are discovered with PSSM-DPC-DWT, F-PSSM-DWT, and R-PSSM-DWT. These approaches are elaborated in the following parts.

2.2.1. Position-specific Scoring Matrix (PSSM). Recently, evolutionary features are successfully implemented and improve the prediction results of many predictors [1, 20]. We also implemented PSSM for the formulation of evolutionary patterns. Each sequence is searched against the NCBI database applying the PSI-BLAST program for the alignment of homologous features [34].

The PSSM can be denoted as follows:

$$PSSM = (P_1, P_2, \dots, P_j, \dots, P_{20})^T, \quad (1)$$

$$P_{i,j} = (P_{1,j}, P_{2,j}, \dots, P_{L,j}), \quad (i = 1, 2, \dots, L),$$

where T and $P_{i,j}$ indicate the transpose operator and score of j type of amino acid in the i^{th} position of query sequence.

2.2.2. Filtered Position-specific Scoring Matrix (F-PSSM). PSSM transforms the evolutionary patterns into numerical forms. It may comprise some negative scores which can lead to generating similar feature vectors despite different sequences. To cope with this hurdle, F-PSSM filters all the negative scores in the preprocessing step. The detail of dimension formulation is provided in [35].

2.2.3. Position-specific Scoring Matrix-Dipeptide Composition (PSSM-DPC). The local sequence-order patterns contains informative feature which are explored by incorporating DPC into PSSM. DPC calculates the frequency of continuous amino acids and produces a dimension of 400 [36]. DPC is calculated as follows:

$$PSSM - DPC = (G_{1,1}, \dots, G_{1,20}, G_{2,1}, \dots, G_{2,20}, \dots, G_{20,1}, \dots, G_{20,20})^T, \quad (2)$$

where

$$P_{i,j} = \frac{1}{L} \sum_{k=1}^{L-1} G_{k,i} \times G_{k+1,j} \quad (1 \leq i, j \leq 20). \quad (3)$$

2.2.4. Reduced Position Specific Scoring Matrix (R-PSSM). It is believed that there exist several similarities among 20 unique amino acids. Based on these similarities, researchers categorized these residues into groups. Li et al. [37] suggested that according to some specific residue the following groups can be formed:

$$G(i) = \begin{cases} Y, & if i = F, Y, W; \\ L, & if i = M, L; \\ V, & if i = I, V; \\ S, & if i = A, T, S; \\ N, & if i = N, H; \\ E, & if i = Q, E, D; \\ K, & if i = R, K; \\ i, & otherwise. \end{cases} \quad (4)$$

Using the Li et al. rule, the $L \times 20$ PSSM is converted to $L \times 10$ matrix by the following equations:

$$G_1 = \frac{F + Y + W}{3},$$

$$G_2 = \frac{M + L}{2},$$

$$G_3 = \frac{I + V}{2},$$

$$G_4 = \frac{A + T + S}{3},$$

$$G_5 = \frac{N + H}{2},$$

$$G_6 = \frac{Q + E + D}{3},$$

$$G_7 = \frac{R + K}{2},$$

$$G_8 = C,$$

$$G_9 = G,$$

$$G_{10} = P.$$

If $r_1 r_2 r_3 \dots r_L$ is a given protein sequence, then its reduced PSSM (R-PSSM) is indicated as follows:

$$RP = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ r_1 & R_{1,1} & R_{1,2} & R_{1,3} & R_{1,4} & R_{1,5} & R_{1,6} & R_{1,7} & R_{1,8} & R_{1,9} & R_{1,10} \\ r_2 & R_{2,1} & R_{2,2} & R_{2,3} & R_{2,4} & R_{2,5} & R_{2,6} & R_{2,7} & R_{2,8} & R_{2,9} & R_{2,10} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_L & R_{L,1} & R_{L,2} & R_{L,3} & R_{L,4} & R_{L,5} & R_{L,6} & R_{L,7} & R_{L,8} & R_{L,9} & R_{L,10} \end{bmatrix}. \quad (6)$$

We obtain 110 feature vector from RP.

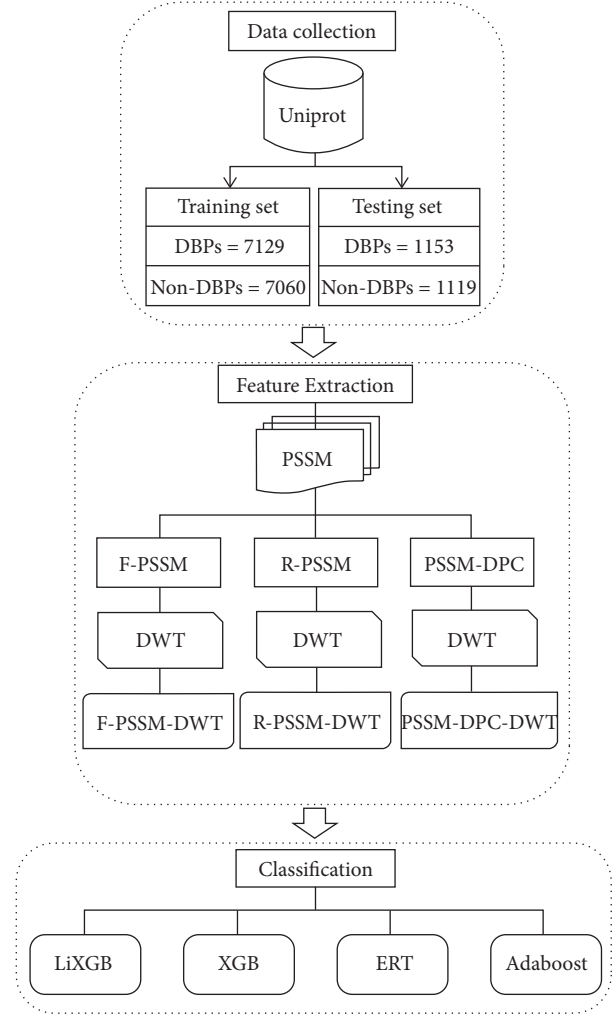


FIGURE 1: Architecture of the proposed model.

2.2.5. Discrete Wavelet Transform. To achieve only salient information, some compression approaches like DWT is applied in research areas. DWT is used for compression of signals and denoising [38, 39]. DWT divides a signal into low-frequency and high-frequency components [40]. Low frequencies are more important than high-frequencies [41]. The Low frequencies are onward split into low and high levels to achieve discriminative patterns. DWT is computed as follows:

$$X(m, n) = \sqrt{\frac{1}{m}} \int_0^y f(y) \Psi\left(\frac{y-n}{m}\right) dy, \quad (7)$$

where m represents the scale variable and n shows the translation variable. $X(m, n)$ is the transform coefficient. The low and high frequencies of a signal $f(t)$ is computed as follows:

$$C_{i,low}[a] = \sum_{k=1}^N s[k]L[2a-k],$$

$$C_{i,high}[a] = \sum_{k=1}^N s[k]H[2a-k],$$
(8)

where $C_{i,high}[a]$ and $C_{i,low}[a]$ are the high and low frequencies of the signal. H , $s[k]$, and L , represent the high pass filter, discrete signal, and low pass filter, respectively.

To obtain only important features and eliminate the less informative and noisy patterns, DWT is extended into F-PSSM, PSSM-DPC, and R-PSSM to split into low and high frequencies up to two levels. Finally, PSSM-DPC-DWT, R-PSSM-DWT, and F-PSSM-DWT novel feature descriptors are constructed. The dimension of each feature set is 512 after applying DWT. Figure 2 depicts the schematic view of Two-level DWT.

2.3. Light eXtreme Gradient Boosting. During the establishment of the predictor, the model training is performed by a classifier. Gradient Boosting Machine (GBM) classifier uses decision trees for the construction of a model. The model performance is improved with loss function [42]. Unlike GBM, eXtreme Gradient Boosting (XGB) employs an objective function. XGB concatenates loss function and regularization for regulating the model complexity. It performs parallel computations to optimize the computational speed. Due to these benefits of XGB, Light eXtreme Gradient Boosting (LiXGB) was proposed [43]. LiXGB possesses many additional features like lower memory, higher efficiency, and fast model training speed that improve the model performance. LiXGB minimizes the model training time of the large datasets. We utilized the hyperparameters like max depth, estimator, eta, lambda, and alpha. The “eta” maintains the learning rate, “estimator” constructs trees, “max depth” is used for controlling the tree depth, “alpha” shrinks the high dimension of the dataset, and “lambda” avoids the overfitting. Other parameters have been kept as default. These hyperparameters are also summarized in Table 1.

2.4. Proposed Model Validation Methodologies. The model performance is examined by different validation approaches. The commonly used validation methods are k-fold and jackknife [44–47]. However, the jackknife is time-consuming and costly [48–50]. During 10-fold cross validation, training set is split into 10-folds. The 9 folds are used for model training and 1 fold is used for model validation. This process is repeated 10 times so that each fold is used for the test exactly once. The final prediction is the average of all tested folds [51–54]. The current work performance is evaluated with 10-fold and five indexes, i.e., specificity (Sp), F-measure, sensitivity (Sn), accuracy (Acc), and Mathew’s correlation coefficient (MCC) for evaluating the model performance [55–58]. These parameters are computed as follows:

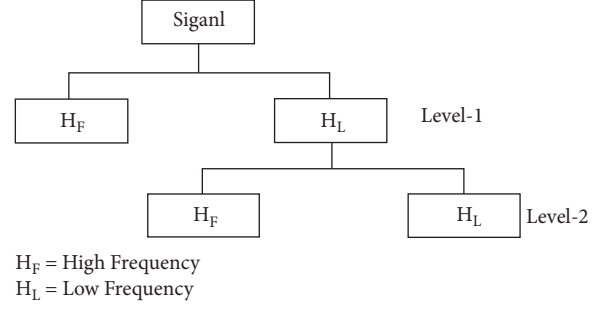


FIGURE 2: 2-level structure of DWT.

TABLE 1: Applied parameters with values.

Parameter	Value
Era	0.1
No. of estimator	500
Alpha	1
Lambda	1
Max depth	8

$$Acc = 1 - \frac{H_-^+ + H_+^-}{H^+ + H^-},$$

$$Sn = 1 - \frac{H_-^+}{H^+},$$

$$Sp = 1 - \frac{H_+^-}{H^-},$$

$$MCC = \frac{1 - (H_-^+ + H_+^-/H^+ + H^-)}{\sqrt{(1 + (H_-^+ + H_+^-/H^+))(1 + (H_-^+ + H_+^-/H^-))}},$$

$$F1\ Score = \frac{(2 * precision * recall)}{(precision + recall)},$$

$$Precision = \frac{H^+}{H_-^+ + H^+},$$

$$Recall = \frac{H^+}{H_+^- + H^+},$$

(9)

where H^+ is used to denote the DBPs, H^- is the non-DBPs, H_+^- shows the prediction of non-DBPs which the model predicted mistakenly as DBPs, and H_-^+ represents the DBPs which are classified by the model as non-DBPs.

3. Results and Discussion

After performing experiments on the models, In this part, we will elaborate the obtained results of the learning algorithms via the extracted feature sets of the training and testing sequences.

TABLE 2: Results of encoders before DWT.

Model	Encoder	Acc (%)	Sn (%)	Sp (%)	MCC (%)
Adaboost	F-PSSM	71.52	80.42	62.54	43.67
	PSSM-DPC	80.05	78.44	81.69	60.15
	R-PSSM	80.07	76.15	84.02	60.35
ERT	F-PSSM	75.18	84.74	58.97	44.56
	PSSM-DPC	79.22	73.18	85.31	58.42
	R-PSSM	79.56	74.99	84.18	59.40
XGB	F-PSSM	74.57	82.17	66.90	49.67
	PSSM-DPC	81.53	76.15	86.97	63.47
	R-PSSM	81.63	76.48	86.84	63.64
LiXGB	F-PSSM	76.60	82.47	66.01	48.75
	PSSM-DPC	83.54	84.61	82.46	67.10
	R-PSSM	83.62	82.30	84.96	67.27

3.1. Results of Feature Encoders before DWT. In this section, we have reported the outcomes of F-PSSM, PSSM-DPC, and R-PSSM in Table 2. The performance of the individual descriptor is analyzed by 10-fold test and assessment indices. On F-PSSM, the accuracies secured by LiXGB, XGB, ERT, and Adaboost are 76.60%, 74.57%, 75.18%, and 71.52%, respectively. Among all classifiers, LiXGB achieved the best accuracy. On PSSM-DPC, all classifiers enhanced the prediction results and generated 83.62%, 81.63%, 79.56%, and 80.07% accuracies by LiXGB, XGB, ERT, and Adaboost, respectively. Similarly, the classifiers also improved the performance on the R-PSSM descriptor using all evaluation parameters. LiXGB attained the highest (83.62%) accuracy. The predictions indicate that LiXGB possesses higher learning power comparatively XGB, ERT, and Adaboost.

3.2. Results of Feature Encoders after DWT. The features extracted by representative methods may contain some noisy, redundant, or less informative features. To avoid such features, DWT is applied to F-PSSM, PSSM-DPC, and R-PSSM. DWT considers the informative patterns and improves the performance of the model. After applying DWT, we achieve F-PSSM-DWT, PSSM-DPC-DWT, and R-PSSM-DWT. Each feature is fed into Adaboost, ERT, XGB, and LiXGB in order to examine the performance over these feature descriptors and results are summarized in Table 3. With 10-fold test, Adaboost, ERT, XGB, and LiXGB produced 73.20%, 77.26%, 75.37%, and 79.40% accuracies which are 1.68%, 2.08%, 0.80%, and 2.80% than F-PSSM, PSSM-DPC, and R-PSSM, respectively. Similarly, the classifiers also boosted the performance on PSSM-DPC-DWT on all evaluation parameters. Furthermore, with R-PSSM-DWT, Adaboost, ERT, XGB, and LiXGB have enhanced the accuracies by 2.16%, 3.49%, 1.98%, and 3.22% than R-PSSM. These results demonstrate that all classifiers show improvement in performance after applying DWT. Among all feature descriptors, the best results are secured by R-PSSM-DWT.

LiXGB has constantly depicted better achievement than other classifiers. LiXGB enhanced the performance and generated 3.23%, 3.79%, and 4.61% higher accuracies than

TABLE 3: Results of feature encoders after DWT.

Model	Encoder	Acc (%)	Sn (%)	Sp (%)	MCC (%)
Adaboost	F-PSSM-DWT	73.20	82.35	56.67	40.21
	PSSM-DPC-DWT	81.81	80.45	83.19	63.66
	R-PSSM-DWT	82.23	77.68	86.81	64.77
ERT	F-PSSM-DWT	77.26	79.91	74.59	54.58
	PSSM-DPC-DWT	81.53	76.15	86.97	63.47
	R-PSSM-DWT	83.05	81.30	84.82	66.15
XGB	F-PSSM-DWT	75.37	83.43	60.81	45.31
	PSSM-DPC-DWT	82.45	83.65	81.25	64.91
	R-PSSM-DWT	83.61	82.66	84.56	67.23
LiXGB	F-PSSM-DWT	79.40	83.11	75.65	58.94
	PSSM-DPC-DWT	84.74	84.30	85.19	69.49
	R-PSSM-DWT	86.84	86.60	87.08	73.69

XGB, ERT, and Adaboost with R-PSSM-DWT. It is concluded that the performance of LiXGB is superior to other classifiers.

3.3. Comparison with Existing Predictors Using Training Set. Several methods have been implemented for the identification of DBPs. The proposed work is compared with past studies including iDNA-Prot [22], iDNA-Prot|dis [23], TargetDBP [59], MsDBP [60], PDBP-CNN [29], and XGBoost [30] and summarized the results in Table 4. Our proposed study improved the accuracy by 4.82%, sensitivity by 10.58%, and MCC by 0.09 than the best predictor (PDBP-CNN). Similarly, The DBP-iDWT enhanced 5.42% Acc, 2.49% Sn, 8.65% Sp, and 0.11 MCC than the second best study (XGBoost). In the same fashion, our predictor performance is superior to past studies using all four assessment parameters. The outcomes verified that DBP-iDWT can discriminate DBPs with high precision.

3.4. Comparison with Past Predictors Using Independent Set. A method is considered effective if it has high generality for the new sequences. We also evaluated the proposed work using a testing dataset. The results compared with past studies like PseDNA-Pro, iDNAPro-PseAAC, iDNAProtES, DPP-PseAAC, TargetDBP, MsDBP, and PDBP-Fusion as noted in Table 5. It is noted that our predictor (DBP-iDWT) raised 5.06% Acc, 17.06% Sn, 8.22% Sp, and 0.10 MCC than PDBP-Fusion. Similarly, DBP-iDWT improved 6.14% Acc, 14.02% Sn, and 0.13 MCC than TargetDBP. Onward, the proposed study also secured higher prediction results than other past methods in Table 5.

These results analysis confirm that the incorporation of DWT into R-PSSM in conjunction with LiXGB can identify DBPs more accurately. Past studies have reported that the selection of the best features can improve the model performance [61–63]. In this study, we also implemented feature selection approach including mRmR and SVM-RFE, however, no improvement in the model performance is observed.

TABLE 4: Comparative analysis with past work on the training set.

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
iDNA-prot	75.40	83.81	64.73	0.50
iDNA-prot dis	77.30	79.40	75.27	0.54
TargetDBP	79.71	79.56	79.85	0.59
MsDBP	80.29	80.87	79.72	0.60
PDBP-CNN	82.02	87.49	76.50	0.64
XGBoost	81.42	84.11	78.43	0.62
DBP-iDWT	86.84	86.60	87.08	0.73

TABLE 5: Comparative analysis with past work using testing dataset.

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
PseDNA-pro	67.23	78.38	56.08	0.35
iDNAPro-PseAAC	66.22	78.37	54.05	0.33
iDNAProt-ES	68.58	95.95	41.22	0.44
DPP-PseAAC	61.15	55.41	66.89	0.22
TargetDBP	76.69	76.35	77.03	0.53
MsDBP	66.99	70.69	63.18	0.33
PDBP-fusion	77.77	73.31	66.85	0.56
DBP-iDWT	82.83	90.37	75.07	0.66

4. Conclusion and Future Vision

DBPs play an active role in many biological functions and drug designing. We have designed a predictor for improving DBPs prediction with high precision. The global information, local features, sequence-order patterns, and correlated factors are explored by PSSM-DPC-DWT, R-PSSM-DWT, and PSSM-DPC-DWT.

The models are trained with LiXGB, XGB, ERT, and Adaboost. It is concluded that R-PSSM-DWT with LiXGB has effectively attained superlative performance than other predictors. The successful outcomes of the proposed study is due to factors like utilization of effective descriptors, application of a compression scheme, and appropriate classifier.

DBP-iDWT will be effective for the identification of DBPs due to its promising prediction power than other predictors and perform an active role in drug development. DBP-iDWT would be fruitful for establishing more operative therapeutic strategies for fatal disease treatment. In addition, we will apply advanced deep learning frameworks [64–67] in our future work to further improve the DBPs prediction.

Data Availability

The data and code are freely available at <https://github.com/Farman335/DBP-DWTPred>.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Farman Ali: Conceptualization, Methodology. Harish Kumar and Shruti Patil: data collection, writing-original draft preparation. Omar Barukab and Ajay B Gadicha: Visualization, performed experiments suggested by reviewers. Omar Alghushairy and Akram Y Sarhan: Code writing, editing, and reviewed the paper.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under Grant no. RGP.2/198/43.

References

- [1] S. Ahmed, M. Kabir, Z. Ali, M. Arif, F. Ali, and D.-J. Yu, "An integrated feature selection algorithm for cancer classification using gene expression data," *Combinatorial Chemistry and High Throughput Screening*, vol. 21, pp. 631–645, 2018.
- [2] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, "An overview of the structures of protein-DNA complexes," *Genome Biology*, vol. 1, 2000.
- [3] K. Sandman, S. L. Pereira, and J. N. Reeve, "Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome," *Cellular and Molecular Life Sciences*, vol. 54, no. 12, pp. 1350–1364, 1998.
- [4] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there," *Nature Reviews Drug Discovery*, vol. 5, no. 12, pp. 993–996, 2006.
- [5] H. Gronemeyer, J. A. Gustafsson, and V. Laudet, "Principles for modulation of the nuclear receptor superfamily," *Nature Reviews Drug Discovery*, vol. 3, no. 11, pp. 950–964, 2004.
- [6] W. H. Hudson, I. M. S. d. Vera, J. C. Nwachukwu et al., "Cryptic glucocorticoid receptor-binding sites pervade genomic NF- κ B response elements," *Nature Communications*, vol. 9, no. 1, p. 1337, 2018.
- [7] H. A. Sikder, M. K. Devlin, S. Dunlap, B. Ryu, and R. M. Alani, "Id proteins in cell growth and tumorigenesis," *Cancer Cell*, vol. 3, no. 6, pp. 525–530, 2003.
- [8] S. Akbar, M. Hayat, M. Iqbal, and M. Tahir, "iRNA-PseTNC: identification of RNA 5-methylcytosine sites using hybrid vector space of pseudo nucleotide composition," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 451–460, 2020.
- [9] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artificial Intelligence in Medicine*, vol. 79, pp. 62–70, 2017.
- [10] S. Akbar, M. Hayat, M. Tahir, and K. T. Chong, "cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach," *IEEE Access*, vol. 8, pp. 131939–131948, 2020.
- [11] F. Ali, S. Ahmed, Z. N. K. Swati, and S. Akbar, "DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information," *Journal of Computer-Aided Molecular Design*, vol. 33, no. 7, pp. 645–658, 2019.
- [12] A. Ahmad, S. Akbar, S. Khan et al., "Deep-AntiFP: prediction of antifungal peptides using distant multi-informative features incorporating with deep neural networks," *Chemo-metrics and Intelligent Laboratory Systems*, vol. 208, Article ID 104214, 2021.

- [13] S. Akbar, A. U. Rahman, M. Hayat, and M. Sohail, "CACP: classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components," *Chemometrics and Intelligent Laboratory Systems*, vol. 196, Article ID 103912, 2020.
- [14] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "iAtbP-hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Computers in Biology and Medicine*, vol. 137, Article ID 104778, 2021.
- [15] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: Identifying Antifungal Peptides Using Sequential and Evolutionary Descriptors Based Multi-Information Fusion and Ensemble Learning Approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 222, Article ID 104516, 2022.
- [16] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artificial Intelligence in Medicine*, vol. 131, Article ID 102349, 2022.
- [17] G. Nimrod, M. Schushan, A. Szilágyi, C. Leslie, and N. Ben-Tal, "iDBPs: a web server for the identification of DNA binding proteins," *Bioinformatics*, vol. 26, no. 5, pp. 692–693, 2010.
- [18] M. Gao and J. Skolnick, "DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions," *Nucleic Acids Research*, vol. 36, no. 12, pp. 3978–3992, 2008.
- [19] H. Zhao, J. Wang, Y. Zhou, and Y. Yang, "Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome," *PLoS One*, vol. 9, no. 5, Article ID e96694, 2014.
- [20] F. Ali, H. Kumar, S. Patil, A. Ahmed, A. Banjar, and A. Daud, "DBP-DeepCNN: Prediction of DNA-Binding Proteins Using Wavelet-Based Denoising and Deep Learning," *Chemometrics and Intelligent Laboratory Systems*, vol. 229, Article ID 104639, 2022.
- [21] K. K. Kumar, G. Pugalenth, and P. N. Suganthan, "DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 6, pp. 679–686, 2009.
- [22] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "IDNA-Prot: identification of DNA binding proteins using random forest with grey model," *PLoS One*, vol. 6, no. 9, Article ID e24756, 2011.
- [23] B. Liu, J. Xu, X. Lan et al., "IDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS One*, vol. 9, Article ID e106691, 2014.
- [24] Q. Dong, S. Wang, K. Wang, X. Liu, and B. Liu, "Identification of DNA-binding proteins by auto-cross covariance transformation," in *Proceedings of the Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 470–475, Washington DC, USA, November 2015.
- [25] L. Wei, J. Tang, and Q. Zou, "Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information," *Information Sciences*, vol. 384, pp. 135–144, 2017.
- [26] F. Ali, M. Kabir, M. Arif et al., "DBPPred-PDSD: machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet Transform and optimized integrated features space," *Chemometrics and Intelligent Laboratory Systems*, vol. 182, pp. 21–30, 2018.
- [27] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 452, pp. 22–34, 2018.
- [28] A. Mishra, P. Pokhrel, and M. T. Hoque, "StackDPPred: a stacking based prediction of DNA-binding protein from sequence," *Bioinformatics*, vol. 35, no. 3, pp. 433–441, 2018.
- [29] G. Li, X. Du, X. Li, L. Zou, G. Zhang, and Z. Wu, "Prediction of DNA binding proteins using local features and long-term dependencies with primary sequences based on deep learning," *PeerJ*, vol. 9, Article ID e11262, 2021.
- [30] Z. Zhao, W. Yang, Y. Zhai, Y. Liang, and Y. Zhao, "Identify DNA-binding proteins through the extreme gradient boosting algorithm," *Frontiers in Genetics*, vol. 12, Article ID 821996, 2021.
- [31] X. Du, Y. Diao, H. Liu, and S. Li, "MsDBP: exploring DNA-binding proteins by integrating multi-scale sequence information via Chou's 5-steps rule," *J Proteome Res*, vol. 18, 2019.
- [32] X. Ma, J. Guo, and X. Sun, "DNABP: identification of DNA-binding proteins based on feature selection using a random forest and predicting binding residues," *PLoS One*, vol. 11, no. 12, Article ID e0167345, 2016.
- [33] C. Zou, J. Gong, and H. Li, "An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis," *BMC Bioinformatics*, vol. 14, no. 1, p. 90, 2013.
- [34] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.
- [35] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, "PPIevo: protein-protein interaction prediction from PSSM based evolutionary information," *Genomics*, vol. 102, no. 4, pp. 237–242, 2013.
- [36] F. Ali and M. Hayat, "Machine learning approaches for discrimination of Extracellular Matrix proteins using hybrid feature space," *Journal of Theoretical Biology*, vol. 403, pp. 30–37, 2016.
- [37] T. Li, K. Fan, J. Wang, and W. Wang, "Reduction of protein sequence complexity by residue grouping," *Protein Engineering Design and Selection*, vol. 16, no. 5, pp. 323–330, 2003.
- [38] R. Moshrefi, M. G. Mahjani, and M. Jafarian, "Application of wavelet entropy in analysis of electrochemical noise for corrosion type identification," *Electrochemistry Communications*, vol. 48, pp. 49–51, 2014.
- [39] X. Wang, J. Wang, C. Fu, and Y. Gao, "Determination of corrosion type by wavelet-based fractal dimension from electrochemical noise," *International Journal of Electrochemical Science*, vol. 8, pp. 7211–7222, 2013.
- [40] B. Yu, S. Li, C. Chen et al., "Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition," *Chemometrics and Intelligent Laboratory Systems*, vol. 167, pp. 102–112, 2017.
- [41] M. Hayat, A. Khan, and M. Yeasin, "Prediction of membrane proteins using split amino acid and ensemble classification," *Amino Acids*, vol. 42, no. 6, pp. 2447–2460, 2012.
- [42] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Computers in Biology and Medicine*, vol. 121, Article ID 103761, 2020.

- [43] X. Wang, Y. Zhang, B. Yu et al., "Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis," *Computers in Biology and Medicine*, vol. 134, Article ID 104516, 2021.
- [44] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, "iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 204, Article ID 104103, 2020.
- [45] F. Ali, S. Akbar, A. Ghulam, Z. A. Maher, A. Unar, and D. B. Talpur, "AFP-CMBPred: computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information," *Computers in Biology and Medicine*, vol. 139, Article ID 105006, 2021.
- [46] I. A. Khan, D. Pi, N. Khan et al., "A Privacy-Conserving Framework Based Intrusion Detection Method for Detecting and Recognizing Malicious Behaviours in Cyber-Physical Power Networks," *Applied Intelligence*, vol. 51, pp. 1–16, 2021.
- [47] P. Chaudhari, H. Agrawal, and K. Kotecha, "Data augmentation using MG-GAN for improved cancer classification on gene expression data," *Soft Computing*, vol. 24, no. 15, pp. 11381–11391, 2020.
- [48] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 384, pp. 78–83, 2015.
- [49] F. Ali, H. Kumar, S. Patil, A. Ahmad, A. Babour, and A. Daud, "Deep-GHBP: improving prediction of Growth Hormone-binding proteins using deep learning model," *Biomedical Signal Processing and Control*, vol. 78, Article ID 103856, 2022.
- [50] F. Ali, H. Kumar, S. Patil, K. Kotecha, A. Banjar, and A. Daud, "Target-DBPPred: an intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting," *Computers in Biology and Medicine*, vol. 145, Article ID 105533, 2022.
- [51] O. Barukab, F. Ali, W. Alghamdi, Y. Bassam, and S. A. Khan, "DBP-CNN: deep learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network," *Expert Systems with Applications*, vol. 197, Article ID 116729, 2022.
- [52] O. Barukab, F. Ali, and S. A. Khan, "DBP-GAPred: an intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning," *Journal of Bioinformatics and Computational Biology*, vol. 19, no. 04, Article ID 2150018, 2021.
- [53] A. Ghulam, F. Ali, R. Sikander, A. Ahmad, A. Ahmed, and S. Patil, "ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network," *Chemometrics and Intelligent Laboratory Systems*, vol. 226, Article ID 104589, 2022.
- [54] A. Ghulam, R. Sikander, F. Ali, Z. N. K. Swati, A. Unar, and D. B. Talpur, "Accurate prediction of immunoglobulin proteins using machine learning model," *Informatics in Medicine Unlocked*, vol. 29, Article ID 100885, 2022.
- [55] Z. U. Khan, F. Ali, I. Ahmad, M. Hayat, and D. Pi, "iPredCNC: computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 195, Article ID 103876, 2019.
- [56] Z. U. Khan, F. Ali, I. A. Khan, Y. Hussain, and D. Pi, "iRSpot-SPI: deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components," *Chemometrics and Intelligent Laboratory Systems*, vol. 189, pp. 169–180, 2019.
- [57] Z. U. Khan, D. Pi, S. Yao, A. Nawaz, F. Ali, and S. Ali, "piEnPred: a bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm," *Frontiers of Computer Science*, vol. 15, no. 6, pp. 156904–156911, 2021.
- [58] M. Ullah, A. Iltaf, Q. Hou, F. Ali, and C. Liu, "A foreground extraction approach using convolutional neural network with graph cut," in *Proceedings of the 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, pp. 40–44, Chongqing, China, June 2018.
- [59] J. Hu, X.-G. Zhou, Y.-H. Zhu, D.-J. Yu, and G.-J. Zhang, "TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 4, pp. 1419–1429, 2020 Jul-Aug.
- [60] X. Du, Y. Diao, H. Liu, and S. Li, "MsDBP: exploring DNA-binding proteins by integrating multiscale sequence information via Chou's five-step rule," *Journal of Proteome Research*, vol. 18, no. 8, pp. 3119–3132, 2019.
- [61] G. Sanghani and K. Kotecha, "Incremental personalized E-mail spam filter using novel TFDCR feature selection with dynamic feature update," *Expert Systems with Applications*, vol. 115, pp. 287–299, 2019.
- [62] A. Ahmad, S. Akbar, M. Hayat, F. Ali, and M. Sohail, "Identification of Antioxidant Proteins Using a Discriminative Intelligent Model of K-Space Amino Acid Pairs Based Descriptors Incorporating with Ensemble Feature Selection," *Biocybernetics and Biomedical Engineering*, vol. 42, 2020.
- [63] S. Akbar, M. Hayat, M. Kabir, and M. Iqbal, "iAFP-gap-SMOTe: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins," *Letters in Organic Chemistry*, vol. 16, no. 4, pp. 294–302, 2019.
- [64] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59800–59821, 2021.
- [65] F. Ali, M. Arif, Z. U. Khan, M. Kabir, S. Ahmed, and D.-J. Yu, "SDBP-Pred: prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM," *Analytical Biochemistry*, vol. 589, Article ID 113494, 2020.
- [66] F. Ali, F. Ali, A. Ghulam et al., "Deep-PCL: a deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features," *Chemometrics and Intelligent Laboratory Systems*, vol. 221, Article ID 104484, 2022.
- [67] R. Sikander, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," *Scientific Reports*, vol. 12, pp. 5505–5509, 2022.

Research Article

A Lane-Changing Decision-Making Model of Bus Entering considering Bus Priority Based on GRU Neural Network

Wanjun Lv , Yongbo Lv , Jianwei Guo , and Jihui Ma

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Yongbo Lv; yblv@bjtu.edu.cn

Received 26 July 2022; Revised 30 August 2022; Accepted 9 September 2022; Published 24 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Wanjun Lv et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A mandatory lane change occurs when buses are ready to enter the station, which will easily cause a reduction of urban road capacity and induce traffic congestion. Using deep learning methods to make lane-changing decisions has become one of the research hotspots in the field of public transportation, especially with the development of the Cooperative Vehicle-Infrastructure System. Aiming at the exploration of the bus lane-changing rules and decisions during entering, we built a GRU neural network model considering bus priority by using the first real-world V2X (vehicle to everything) dataset. Firstly, we illustrated the image and point cloud data processing by coordinate transformation. Secondly, the Kalman filtering algorithm was used to evaluate the vehicle state. Combined with the bus priority rules, we propose a flexible right-of-way lane in front of the bus stop. And then, we obtain the feature variables as inputs to the model. The XGBoost algorithm was chosen to train the GRU model. Results show that the model has higher identification accuracy for lane-changing maneuvers by comparison with other models. It plays a very important role in providing a decision basis for more refined bus operation management.

1. Introduction

Lane-changing behavior plays an important role in traffic flow theory. There are two types of lane-changing operations, namely, mandatory lane-changing (MLC) and discretionary lane-changing (DLC) [1]. Recently, researchers have shown an increased interest in the MLC maneuvers but paid less attention to the DLC scenes [2, 3]. MLC means that drivers are forced to leave the current lane because external factors have disturbed them or they are planning to reach their destination. A bus changing lanes and entering a station is one of the typical scenarios of MLC. Within the vicinity of the station, the bus must complete the lane change and prepare to stop at the station. Therefore, drivers are likely to adopt a radical lane change strategy, causing the surrounding vehicles to slow down, change lanes, and even queue, resulting in traffic congestion. More importantly, due to the large size of the bus itself, it will block the sight of the surrounding car drivers when changing lanes, which will easily lead to traffic accidents. Therefore, it is very necessary

to study lane-changing behavior during the bus entering process.

One of the most significant current discussions in lane-changing is decision-making. Early lane change decision methods were mainly rule-based methods, which determined whether a vehicle met certain lane change rules by building a set of decision rules [4, 5]. Furthermore, models based on discrete choices are also used in lane change decisions [6–8]. This kind of model can take driver characteristics into account and at the same time has a simpler structure. However, the influence of different factors on the outcome of lane change is linear in these models, whereas lane change behavior is a process of coupled influence of multiple factors, and the relationship among the factors is often nonlinear. In recent years, in order to represent the nonlinear relationship between the input and output data more accurately, machine learning algorithms are increasingly being applied to lane-changing decision-making [9], such as support vector machine (SVM) [10], Bayesian filtering (BF) [11], and so on.

For MLC scenarios, many previous studies have focused on lane-changing behaviors in freeway on-ramp merging areas. There are fewer methods that have tried to solve the important issues of the lane-changing maneuvers that occur on buses entering. With the development of autonomous driving (AD) technology and the cooperative vehicle infrastructure system (CVIS), it is easier to collect enough bus-related data for machine learning algorithm training.

The purpose of this research is to model the bus lane-changing decision-making process using the first real-world V2X dataset for CVIS, with attention to the flexible bus priority along the bus entering route. The main contributions of this study are as follows:

- (1) This paper focuses on the process of bus lane-changing decision-making, one of the MLC behaviors, which can seriously affect traffic conditions and is rarely explored by previous research studies.
- (2) Based on real traffic data, a GRU neural network is used to solve the nonlinear bus lane-changing decision-making problem during entering. Combining the V2X technology, we take the bus priority into account, which is helpful to update real-time feature variables.

Section 2 reviews the literature on lane-changing methods. Section 3 is concerned with the methodology used for this study. The training results and discussion are shown in Section 4. Finally, Section 5 gives a brief summary and areas for further research.

2. Related Work

Lane-changing is one of the common driving behaviors in traffic flow. Gipps [4] is the first scholar who built the vehicle lane-changing decision model to test the factors that affect drivers' lane-changing decisions in urban traffic. He proposed a logical frame for the lane-changing decision process. To better understand the mechanisms of the lane-changing model, other authors expanded the investigations by examining the diversified aspects of expressways and urban roads, including the lane-changing probability [12], drivers' aggressive and courtesy behaviors [13], and the triggered negative effects [14]. On the other hand, some authors also tried to develop Gipps's model. Kesting et al. first demonstrated the minimizing overall braking induced by Lane Change (MOBIL) model and considered the vehicle acceleration changes on the basis of the previous model [5]. As connected vehicles appeared, there was a large volume of published studies describing the role of cooperative merging strategies [15] and optimization algorithms to reveal the detailed mandatory lane-changing behaviors on freeways [16].

Besides, a considerable amount of literature uses utility theory to investigate lane-changing behaviors. The study of utility theory was first carried out by Ahmed et al. to examine the process of lane-changing decision-making on freeways. Toledo et al. integrated DLC and MLC behaviors by using the utility model [17]. Sun and Eleftheriadou provided an in-depth analysis of the influences of drivers' characteristics on

lane-changing behaviors by observing drivers' profiles and corresponding vehicle trajectory data [18, 19].

Recently, a number of machine learning-based lane-changing decision models have also emerged. Tang et al. utilized an adaptive learning approach to develop a fuzzy neural network-based prediction model for lane change behavior [20]. In response to the problem that it is difficult to model the multiparameter nonlinear features in the lane change decision process, the authors' team developed a support vector machine-based lane change decision model [21]. A neural network-based lane change decision model was developed in the paper [22]. Multilayer perceptron (MLP) and convolutional neural network (CNN) architectures were combined to demonstrate how the lane change acceptance of a specific driver can be learned from lane change intention and the surrounding environment. The work [23] proposed a deep reinforcement learning (DRL)-based motion planning strategy for autonomous driving tasks in highway scenarios where an autonomous vehicle merges into two-lane road traffic flow and realizes the lane-changing maneuvers.

These existing LC models were mostly developed for freeways and are not likely to be applied to buses that are entering. In the above analysis, we introduce the GRU neural network to improve the MLC decision process during bus entering.

3. Materials and Methods

3.1. Data Processing. The DAIR-V2X dataset is the first large-scale, multimodality, multiview dataset captured from real scenarios for CVIS and contains data captured from the Vehicle-Infrastructure Cooperative view [24]. It contains 71,254 LiDAR frames and 71,254 camera frames captured in intersection scenes where a well-equipped vehicle passes through intersections with infrastructure sensors deployed. 40% of the frames are captured from infrastructure sensors and 60% of the frames are captured from vehicle sensors. All of them are precisely labeled by expert annotators. The dataset covers 10 km of city roads, 10 km of highways, 28 intersections, and 38 km² of driving regions with diverse weather and lighting variations. The subdatasets are presented in Table 1.

In this paper, the DAIR-V2X-C dataset containing 38,845 frames of image data and 38845 frames of point cloud data were selected to train the model. In order to obtain information about the buses and their surrounding vehicles, the 3-dimensional (3D) LiDAR point cloud data and the 2-dimensional (2D) camera image set need to be correlated through spatial coordinate transformation. There are three coordinate systems used in the dataset, including the world coordinate, the camera coordinate, and the LiDAR coordinate. The schematic diagram of coordinate transformation is shown in Figure 1.

The Lidar-to-Camera transformation can be obtained by simply multiplying the Lidar-to-World and World-to-Camera transformations. In order to generate the depth map D_G of the ground plane with the same size as the image,

TABLE 1: The details of DAIR-V2X dataset.

DAIR-V2X dataset	View	Image	Point cloud
DAIR-V2X-C	Vehicle-infrastructure cooperative	38845	38845
DAIR-V2X-V	Single vehicle	22325	22325
DAIR-V2X-I	Infrastructure	10084	10084

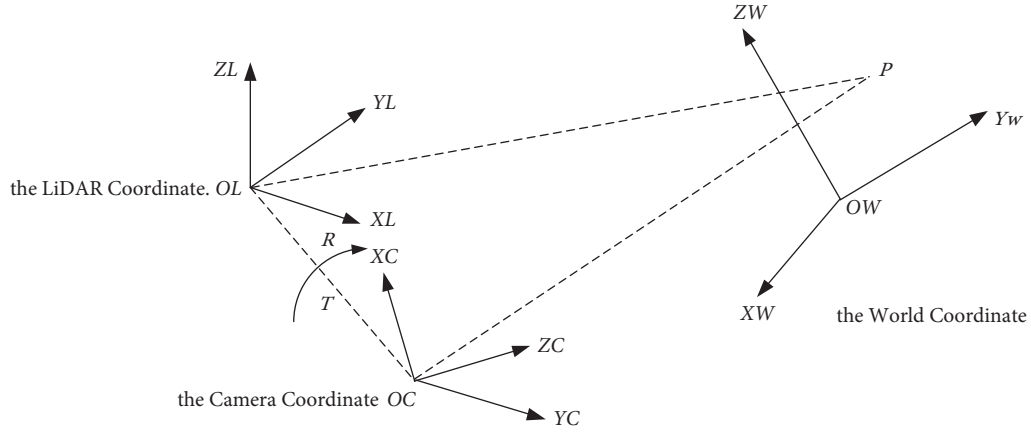


FIGURE 1: Schematic diagram of coordinate transformation.

we use the ground plane equation $G(\alpha, \beta, \gamma, d)$ and camera intrinsic $K^{3 \times 3}$ as shown in the following equation:

$$\begin{cases} Z[x, y, 1]^T = K^{3 \times 3}[X, Y, Z]^T \\ G^{1 \times 4}[X, Y, Z, 1]^T = 0 \end{cases}, \quad (1)$$

where $[x, y]$ is the pixel in the image coordinates and $[X, Y, Z]$ is the corresponding 3D point in the camera coordinate that lies on the ground plane. Thus, the depth Z can be derived with the known 2D image points and the ground plane equation G .

After completing the matching of point cloud information and image information through coordinate conversion, we use the Kalman filtering algorithm to evaluate the vehicle state, including the 2 steps, namely, predicting and updating.

The first step is predicting, as shown in the following equation:

$$\begin{cases} x_{t+1} = A\hat{x}_t + Bu_t + w_t \\ P_{t+1} = A\hat{P}_tA^T + Q \end{cases}, \quad (2)$$

where x_{t+1} is the predicted value of the system state vector at moment $t+1$. \hat{x}_t and u_t are the optimal estimated value and input control vector of the state vector at moment t , respectively. Matrices A and B denote the state transfer matrix and control matrix, respectively. w_t is the noise of the system. P_{t+1} represents the predicted value of the state estimation covariance matrix at moment $t+1$. \hat{P}_t means the optimal estimation of the state estimation covariance matrix at moment t . Q is the process noise covariance matrix.

The second step is updating, as shown in the following equation:

$$\begin{cases} K_{t+1} = P_{t+1}H^T(HP_{t+1}H^T + R)^{-1} \\ \hat{x}_{t+1} = x_{t+1} + K_{t+1}(z_{t+1} - Hx_{t+1}) \\ \hat{P}_{t+1} = (I - K_{t+1}H)P_{t+1} \end{cases}, \quad (3)$$

where K_{t+1} denotes the Kalman gain coefficient. H is the measurement matrix. R is the measurement noise covariance matrix. \hat{x}_{t+1} denotes the optimal estimate of the state vector after the update at moment $t+1$. z_{t+1} is the measurement value at moment $t+1$. \hat{P}_{t+1} is the optimal estimation value of the state evaluation covariance matrix at moment $t+1$.

Because the bus lane-changing process does not involve external inputs to the system, both B and w_t in (2) are not taken into account. In this paper, we assume that the vehicle is treated as a uniform acceleration motion in each sampling period. Suppose the state vector is $[x, v, a]^T$, which means position, speed, and acceleration. A_k is shown in (4). H is the unit matrix.

$$A_k = \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}. \quad (4)$$

The state estimation of the target vehicle by applying the Kalman filtering algorithm jointly to the 3D LiDAR point cloud data and image data can obtain the position, velocity, and acceleration state information of the target vehicle.

3.2. Extraction of Feature Variables. Targeting buses' apparently purposeful and time-sensitive lane-changing

behavior, it is necessary to guarantee the priority of right-of-way during their approach to the station. The process is illustrated in Figure 2.

The green area is the flexible right-of-way lane upstream of the bus stop, the length of which is determined by the traffic and roadway network status. FT is the upstream vehicle in the adjacent lane that needs to slow down or change lanes to avoid collision during the bus switches into this lane. PT_0 is located downstream in the flexible right-of-way lane and will not turn right in the intersection, so it is requested to leave this lane; PT_1 is located downstream in the flexible right-of-way lane and needs to turn right in the intersection, so it is not necessary to leave the current lane; PC is the vehicle located downstream in the lane where the bus is located. It is not allowed to change in the green area. The FC is the vehicle located upstream in the lane where the bus is located. Vehicles in other lanes do not directly affect the bus and are not analyzed.

According to Gipps' safe distance rules [4], D_{PC} and D_{PT} indicate the safe distance between the subject vehicle (the bus) and the preceding vehicle in the current lane and the target lane, respectively.

$$\begin{cases} D_{PC}^i = v_b^i \tau + \frac{(v_b^i)^2}{2d_b} - \frac{(v_{PC}^i)^2}{2d_{PC}} \\ D_{PT}^i = v_b^i \tau + \frac{(v_b^i)^2}{2d_b} - \frac{(v_{PT}^i)^2}{2d_{PT}} \end{cases}, \quad (5)$$

where τ indicates the reaction time of bus driver. v_b^i , v_{PC}^i , and v_{PT}^i are the initial velocity of the bus, PC, and PT in each sampling period, respectively. d_b , d_{PC} , and d_{PT} are the maximum deceleration of the corresponding vehicle.

Similarly, the safe distance between the bus and the following vehicle before and after the bus changes the lane (enters the green area in Figure 2) can be formulated as follows:

$$\begin{cases} D_{FC}^i = v_{FC}^i \tau + \frac{(v_{FC}^i)^2}{2d_{FC}} - \frac{(v_b^i)^2}{2d_b} \\ D_{FT}^i = v_{FT}^i \tau + \frac{(v_{FT}^i)^2}{2d_{FT}} - \frac{(v_b^i)^2}{2d_b} \end{cases}, \quad (6)$$

where v_{FC}^i and v_{FT}^i denote the initial velocity of the FC and FT in each sampling period, respectively. d_{FC} and d_{FT} represent the maximum deceleration of the corresponding vehicle.

In a real-world lane-changing scenario, the bus tends to finish the lane-changing process as soon as possible. Some studies adopt the time ratio t_b^f/t_{\max} to represent this characteristic, where t_b^f denotes the time used to finish the lane-changing process and t_{\max} is the longest allowable time for lane-changing maneuver. The traffic efficiency

description in this paper is developed by the above-mentioned method, and the index can be calculated as follows:

$$\frac{t_b^f}{t_{\max}} = \frac{\sum_i^N x_b^{if} - x_b^{i0}/v_b^i}{t_{\max}}, \quad (7)$$

where x_b^{if} and x_b^{i0} represent the final and initial position of the i th sampling period, respectively. t_b^f can be approximately calculated by $\sum_i^N x_b^{if} - x_b^{i0}/v_b^i$.

After considering the safety and efficiency, we can obtain the feature variables as shown in Table 2.

Thanks to the advantages of CVIS, using V2X communication seems to be the logical solution. Other vehicles on the road can check whether there are approaching buses from behind periodically within a specific sensing distance. These feature parameters can be transmitted by vehicle-to-vehicle technology according to the expected wireless communication range of V2X systems. In case a vehicle detects an approaching bus that is entering the station in its adjacent or current lane, the vehicle will determine whether it affects the ideal bus operation to guarantee its priority. Of course, this paper assumes that all vehicles have real-time communication capabilities and does not reject the bus lane-changing requirement.

3.3. Lane-Changing Decision-Making Model. The gated recurrent unit (GRU) network is one of the variants of long-short-term memory (LSTM) networks in recurrent neural network (RNN), which mainly replaces the two gating units of LSTM (forgetting gate and input gate) with one gating unit (update gate). The GRU model is more streamlined than the standard LSTM model, with only two gating units in the hidden layer neuron structure and fewer network parameters, and its model structure is shown in Figure 3. It has been well developed and applied in the field of time series prediction.

The GRU neural network works similarly to the LSTM in which it has two gates, namely, a reset gate and an update gate. Both the reset gate and the update gate receive the hidden state h_{t-1} at the previous moment and the input information x_t at the current moment and the output gating signals are denoted as r_t and z_t , respectively. The gating unit also consists of a sigmoid function and a dot product operation. The GRU works as follows.

Short time series dependencies are obtained via the reset gate as follows:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r). \quad (8)$$

Calculate update gate as follows:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z). \quad (9)$$

The candidate vector is obtained after updating as follows:

$$\tilde{h}_t = \tanh(W_{xh}x_t + W_{hh}(r_t h_{t-1}) + b_c). \quad (10)$$

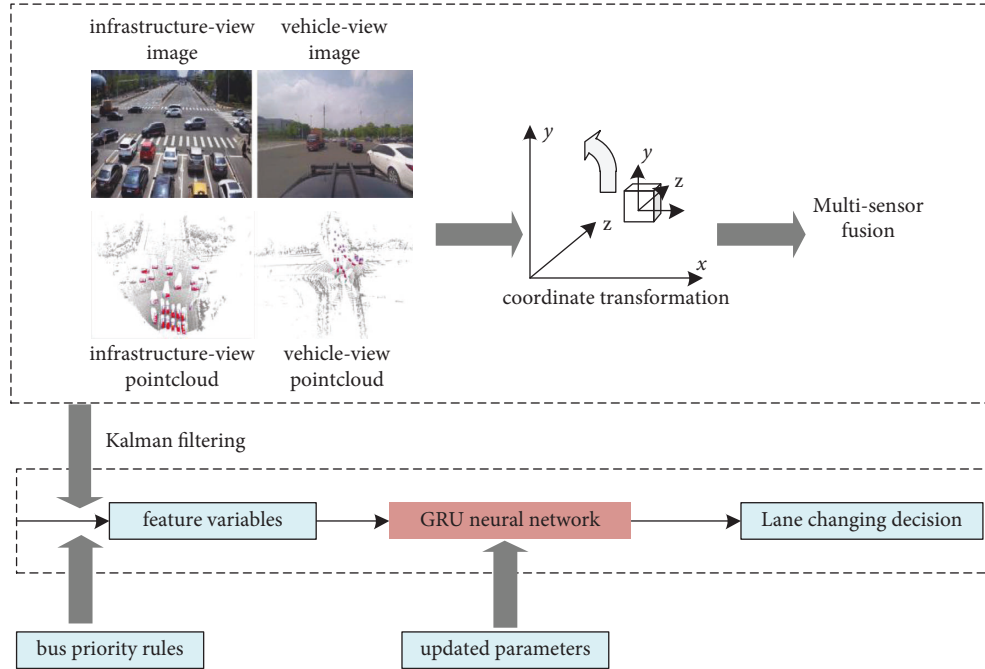


FIGURE 4: The lane-changing decision-making model of bus entering.

4. Results and Discussion

4.1. Testing Data and Evaluation Metrics. We adopted the first large-scale DAIR-V2X dataset, which provides high-quality traffic data for research. The captured images obtained by cameras and the point clouds scanned by the LiDAR are calibrated and aligned between different sensors. The transformations among the world, the camera, and LiDAR are obtained, as well as the ground plane equation. The 2D-3D joint annotation is carried out by projecting the point clouds onto the images and adjusting the 3D bounding boxes manually to fit the 2D instance. As shown in Figure 5, there are a total of 9 kinds of traffic elements in the dataset. The objective of this paper is to show that, the corresponding number of samples is more than 60,000.

We used the scaled exponential linear unit (SeLU) as the activation function in the hidden layers. Compared with the traditional sigmoid and the popular rectified linear unit (ReLU), SeLU can converge towards zero mean y and unit variance automatically. Therefore, it has the same effect as batch normalization, which means that exploding and vanishing gradients can be avoided. The SeLU activation function is shown as follows:

$$\text{SeLU}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha \cdot e^x - \alpha & \text{if } x \leq 0 \end{cases}, \quad (12)$$

where λ and α are two parameters with typical values, namely, 1.0507 and around 1.6733, respectively.

According to our previous research [25], the operating time of the XGBoost algorithm is much faster than that of other programs on a single device. It can scale up to billions

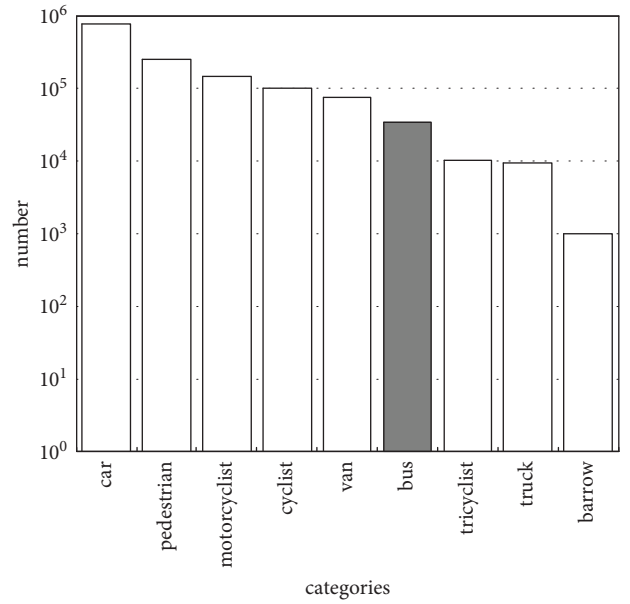


FIGURE 5: The number of different vehicles categories in DAIR-V2X dataset.

of examples in the case of distributed or limited memory. We chose XGBoost algorithm to train the GRU neural network in this paper.

An appropriate evaluation framework can help to estimate performance. Basically, samples in classification problems are divided into four categories, including true positives, true negatives, false positives, and false negatives. We introduce this evaluation framework for the model in Table 3.

TABLE 3: Evaluation framework of bus lane-changing decision model.

	Training condition		
	Bus lane changing	TP	FN
True condition	Bus lane changing	TP	FN
	Bus lane keeping	FP	TN

TABLE 4: GRU neural network tested.

Structure	Hidden layer		
	1	2	3
1	30	0	0
2	50	0	0
3	100	0	0
4	10	10	0
5	30	10	0
6	50	10	0
7	10	10	5
8	10	10	10
9	30	10	10

TP and FN are the probabilities of actual lane-changing events correctly and wrongly classified as lane-changing and lane keeping events, respectively. FP and TN are the numbers of actual lane keeping events wrongly and correctly classified as lane-changing and lane keeping events, respectively. The following evaluation indicators (13) and (14) calculated by these four parameters are chosen in this paper:

$$\text{accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (13)$$

$$\text{true positive rate (TPR)} = \frac{TP}{TP + FN}. \quad (14)$$

Furthermore, the ROC curve can describe the relationship between TP and FP. The ROC score which calculates the area under the ROC curve is also a judging basis for the effect of algorithm.

4.2. The Optimal GRU Structure. In the paper [26], it was stated that overfitting in training neural networks was serious and should be given more attention. When overfitting happens, the error on the testing dataset is still high although the error of the model on the training dataset could be ignored. To avoid this phenomenon, various network structures, from simple to complicated, are tested in this paper. The different neural network structures are illustrated in Table 4 (where 0 means that there is no layer in the corresponding structure).

We use the cross-validation method to avoid overfitting, and 30% of the training samples are chosen as the validation set. Further statistical tests revealed that the model works efficiently with the help of the cross-validation method. We separate all the samples into 3 parts: 30% as the training dataset, 30% as the validation set, and the remaining 40%

as the test dataset. For all the neural network schemes in Table 4, combined with the flexible right-of-way lane in front of the bus stop, generally at least 4,000 to 14,000 samples (10% to 40% of the total sample) are required. The results are shown in Figure 6.

Results show that structure 5 (3 hidden layers that contain 30, 10, and 0 neurons, respectively), trained by 10,000 samples, obtained the best performance. Furthermore, we selected 10,000 training samples to test different lengths of the historical time interval. Results show that considering the 8-second historical input is always the best choice. Figure 7 gives an example of ACC of the model.

To examine the accuracy of the bus lane-changing model in this paper, we compared the performance of the stochastic gradient descent (SGD) algorithm, the Random Forest model, and the SVM model with our model in the dataset. The indexes of different models are presented in Table 5. It can be seen that the GRU model in this paper performs better than other models. The ACC value of our model is up to 92.45%, much higher than that of others. It is worth noting that the TPR can guarantee the safety during bus lane-changing, which is vital for the driver to operate the lane-changing successfully. Hence, it could conceivably be noticed that the higher the value of TPR is, the safer the lane-changing will be. At the same time, the false negative would not have a negative impact on driving safety because it only leads to the loss of an opportunity to change lanes. Therefore, it is necessary to pay more attention to the accuracy of lane-changing maneuvers. The results of the model indicate that it is a safe and effective bus lane-changing decision-making system.

Figure 8 compares ROC curves among SGD, Random Forest, SVM, and GRU. It is apparent that the performance of the proposed GRU neural network is better than that of other binary classifiers, which means that the same conclusion can be drawn as what we get in Table 5. The gated

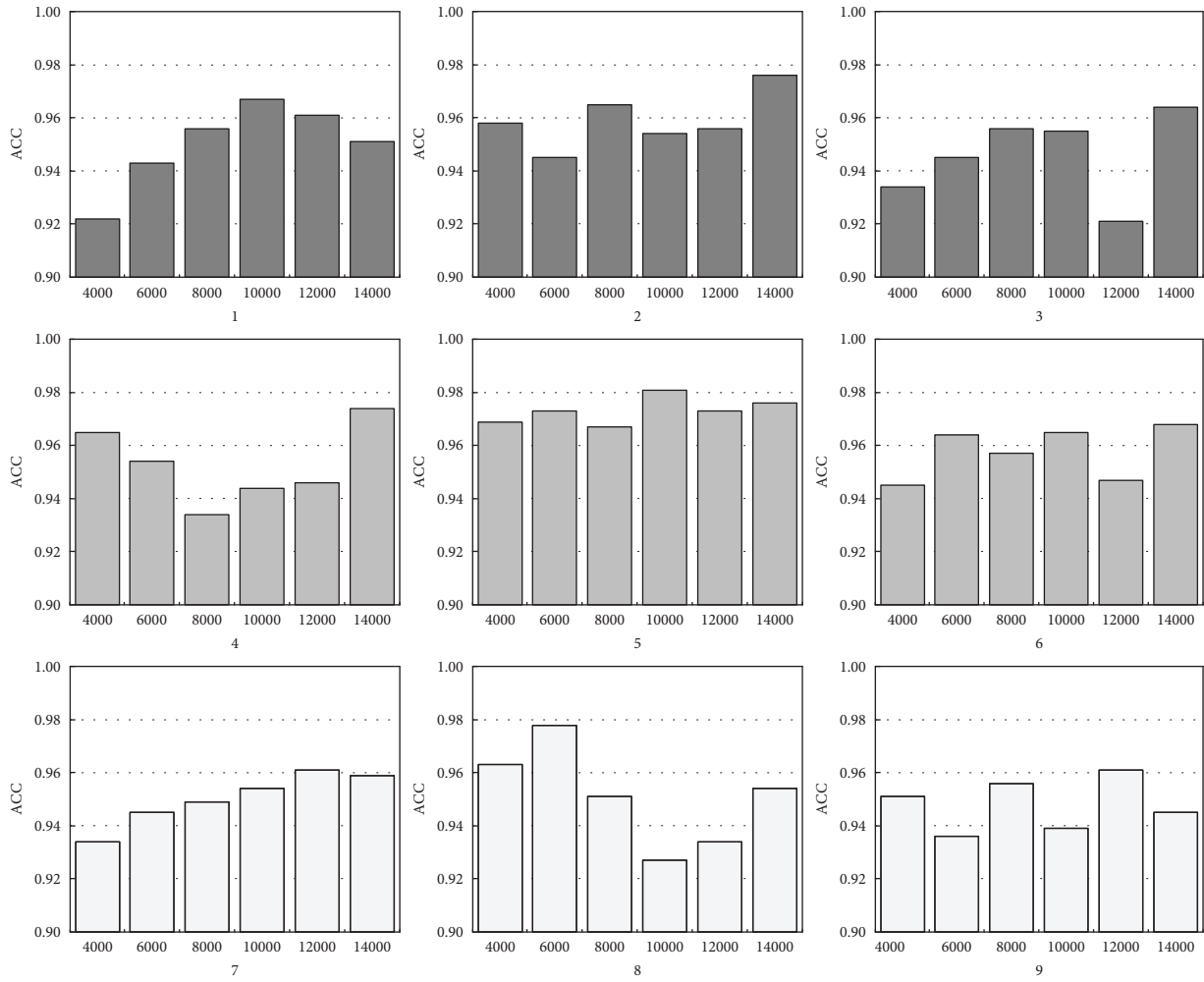


FIGURE 6: ACC values of different GRU network structures.

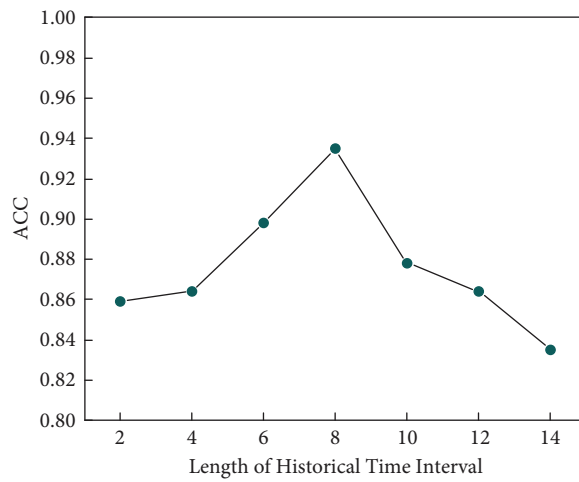


FIGURE 7: ACC values of different length of the historical time interval.

TABLE 5: Evaluation values of different models for lane-changing.

Model	ACC (%)	TPR (%)
SGD	76.34	95.41
Random forest	81.23	93.25
SVM	89.35	90.33
GRU	92.45	96.68

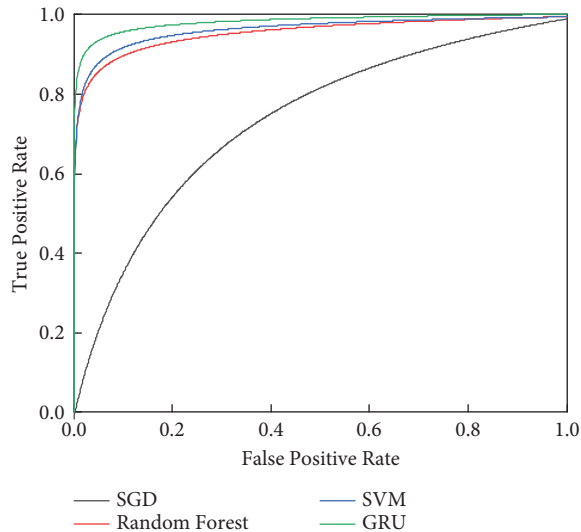


FIGURE 8: Comparison of ROC curves among different models.

branch plays an important role in correlation analysis, which explicitly extracts the characteristics of the feature variables, which is helpful to the lane-changing decision.

5. Conclusion

Based on the DAIR-V2X dataset, this study builds a lane-changing decision-making model of a bus entering based on the GRU neural network. Image and point cloud data are matched by coordinate transformation to obtain the number of bus vehicles. To reduce the noise interference, the Kalman filtering algorithm is used to evaluate the buses' position, speed, and acceleration. Meanwhile, the proposed method takes the bus priority rules into consideration when it enters the station, thus getting novel feature variables as input vectors. After calculating the experimental data, we choose the network structure "3 hidden layers that contain 30, 10, and 0 neurons, respectively," as the optimal structure and 8 seconds as the length of the historical time interval because the outcome performs best. By comparison and verification, it was proved that the proposed model has higher accuracy. The ACC value of our model is up to 92.45%.

Further research regarding the role of lane-changing would be worthwhile. This study considers the rules to ensure the priority of buses entering with the precondition that there is a 100% penetration rate of V2X technology in all vehicles. In a realistic road environment, the existence of disobedience to the lane-changing rules needs to be considered.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (61872036).

References

- [1] A. Talebpour, H. S. Mahmassani, and S. H. Hamdar, "Modeling lane-changing behavior in a connected environment: a game theory approach," *Transportation Research Procedia*, vol. 7, pp. 420–440, 2015.
- [2] S. Hao, L. Yang, and Y. Shi, "Data-driven car-following model based on rough set theory," *IET Intelligent Transport Systems*, vol. 12, no. 1, pp. 49–57, 2018.
- [3] Y. Dou, Y. Fang, C. Hu, R. Zheng, and F. Yan, "Gated branch neural network for mandatory lane changing suggestion at the on-ramps of highway," *IET Intelligent Transport Systems*, vol. 13, no. 1, pp. 48–54, 2019.
- [4] P. G. Gipps, "A model for the structure of lane-changing decisions," *Transportation Research Part B: Methodological*, vol. 20, no. 5, pp. 403–414, 1986.
- [5] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model MOBIL for car-following models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1999, no. 1, pp. 86–94, 2007.
- [6] D. J. Sun and L. Elefteriadou, "A driver behavior-based lane-changing model for urban arterial streets," *Transportation Science*, vol. 48, no. 2, pp. 184–205, 2014.
- [7] C. Ng, S. Susilawati, M. A. S. Kamal, and I. M. L. Chew, "Development of a binary logistic lane change model and its validation using empirical freeway data," *Transportation Business: Transport Dynamics*, vol. 8, no. 1, pp. 49–71, 2020.
- [8] H. van Lint, W. Schakel, G. Tamminga, P. Knoppers, and A. Verbraeck, "Getting the human factor into traffic flow models: new open-source design to simulate next generation of traffic operations," *Transportation Research Record Journal of the Transportation Research Board*, vol. 2561, no. 1, pp. 25–33, 2016.
- [9] D. F. Xie, Z. Z. Fang, B. Jia, and Z. He, "A data-driven lane-changing model based on deep learning," *Transportation Research Part C: Emerging Technologies*, vol. 106, pp. 41–60, 2019.
- [10] Y. D. F. Yan and D. Feng, "Lane changing prediction at highway lane drops using support vector machine and artificial neural network classifiers," in *Proceedings of the 2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 901–906, IEEE, Banff, AB, Canada, July 2016.
- [11] X. Li, W. Wang, and M. Roetting, "Estimating driver's lane-change intent considering driving style and contextual traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3258–3271, 2019.
- [12] Q. I. Yang and H. N. Koutsopoulos, "A Microscopic Traffic Simulator for evaluation of dynamic traffic management

- systems,” *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 3, pp. 113–129, 1996.
- [13] P. Hidas, “Modelling lane changing and merging in microscopic traffic simulation,” *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 5-6, pp. 351–371, 2002.
- [14] P. Hidas, “Modelling vehicle interactions in microscopic simulation of merging and weaving,” *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 1, pp. 37–62, 2005.
- [15] X. Guan, M. Hu, P. Kar, and S. Guan, “Cooperative merging strategy for connected and automated vehicles at highway on-ramps,” *SPIEL*, vol. 145, p. 121740Q, 2022.
- [16] A. Omidvar, L. Elefteriadou, M. Pourmehr, and C. Letter, “Optimizing freeway merge operations under conventional and automated vehicle traffic,” *Journal of Transportation Engineering, Part A: Systems*, vol. 1467 pages, 2020.
- [17] T. TomerN, K. Haris, and E. B. Moshe, “Modeling integrated lane-changing behavior,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1857, no. 1.
- [18] D. J. Sun and L. Elefteriadou, “Lane-changing behavior on urban streets: a focus group-based study,” *Applied Ergonomics*, vol. 42, no. 5, pp. 682–691, 2011.
- [19] D. J. Sun and L. Elefteriadou, “Lane-changing behavior on urban streets: an “in-vehicle” field experiment-based study,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 27, no. 7, pp. 525–542, 2012.
- [20] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, “Lane-changes prediction based on adaptive fuzzy neural network,” *Expert Systems with Applications*, vol. 91, pp. 452–463, 2018.
- [21] Y. Liu, X. Wang, L. Li, S. Cheng, and Z. Chen, “A novel lane change decision-making model of autonomous vehicle based on support vector machine,” *IEEE Access*, vol. 7, pp. 26543–26550, 2019.
- [22] A. Díaz-Álvarez, M. Clavijo, F. Jiménez, E. Talavera, and F. Serradilla, “Modelling the human lane-change execution behaviour through multilayer perceptrons and convolutional neural networks,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 56, pp. 134–148, 2018.
- [23] K. Lv, X. Pei, C. Chen, and J. Xu, “A safe and efficient lane change decision-making strategy of autonomous driving based on deep reinforcement learning,” *Mathematics*, vol. 10, no. 9, p. 1551, 2022.
- [24] H. Yu, “DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection,” 2022, <https://arxiv.org/abs/2204.05575>.
- [25] W. Lv, Y. Lv, Q. Ouyang, and Y. Ren, “A bus passenger flow prediction model fused with point-of-interest data based on extreme gradient boosting,” *Applied Sciences*, vol. 12, no. 3, p. 940, 2022.
- [26] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

Research Article

An Optimized LIME Scheme for Medical Low Light Level Image Enhancement

Yue Kun, Gong Chunqing, and Gao Yuehui 

Tianjin Modern Vocational Technology College, Tianjin, China

Correspondence should be addressed to Gao Yuehui; yuehui_gao@163.com

Received 11 June 2022; Revised 22 July 2022; Accepted 5 August 2022; Published 22 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Yue Kun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The role of medical image technology in the medical field is becoming more and more obvious. Doctors can use medical image technology to more accurately understand the patient's condition and make accurate judgments and diagnosis and treatment in order to make a large number of medical blurred images clear and easy to identify. Inspired by the human vision system (HVS), we propose a simple and effective method of low-light image enhancement. In the proposed method, first a sampler is used to get the optimal exposure ratio for the camera response model. Then, a generator is used to synthesize dual-exposure images that are well exposed in the regions where the original image is underexposed. Next, the enhanced image is processed by using a part of low-light image enhancement via the illumination map estimation (LIME) algorithm, and the weight matrix of the two images will be determined when fusing. After that, the combiner is used to get the synthesized image with all pixels well exposed, and finally, a postprocessing part is added to make the output image perform better. In the postprocessing part, the best gray range of the image is adjusted, and the image is denoised and recomposed by using the block machine 3-dimensional (BM3D) model. Experiment results show that the proposed method can enhance low-light images with less visual information distortions when compared with those of several recent effective methods. When it is applied in the field of medical images, it is convenient for medical workers to accurately grasp the details and characteristics of medical images and help medical workers analyze and judge the images more conveniently.

1. Introduction

Low-light images are a ubiquitous image model. Its main features are low brightness and a darker area at the center of the image. Such images have low visibility, are not easy to observe and analyze, and are not conducive to related applications, which makes digital image processing face major challenges. Figure 1 provides three such examples, such as the painting on the wall in the first example, almost “buried” in the dark. In order to enhance low-light images, people have researched many classic algorithms, but the processing results of these algorithms are more or less problematic. For example, the enhancement algorithm based on wavelet transform [1] can not only describe the outline of the image but also highlight the details of the image due to its multiresolution characteristics, but it has little effect on the change of the image contrast. The histogram-based enhancement algorithm [2] mainly uses histogram

equalization to improve the contrast of the image, but the grayscale of the processed image is reduced, and some details are lost. In [3], the authors try to enhance the contrast while maintaining the naturalness of the lighting. While it prevents the output image from being overenhanced, its performance is not very appealing in terms of efficiency and visual effects. The *illu-adj* algorithm proposed in [4] is able to process the *H* and *I* components separately and achieve image enhancement in the HSI color space. However, the image processed by this scheme is still dark, and the effect is not good. In conclusion, so far there is no perfect method for low-light image enhancement.

Inspired by the human visual system (HVS) [5], this paper proposes a simple and effective innovative method: a low-light image enhancement method based on an optimized LIME scheme. This paper compares this innovative algorithm with four other algorithms: multilevel color consistency theory (MSRCR) [6], simultaneous reflection

and illumination estimation algorithm (SRIE) [7], light map estimation-based method (LIME) [8], and the multibias fusion method (MF) [9]. Experimental data show that our proposed method has unique advantages in luminance order error (LOE), visual information fidelity (VIF), and subjective vision.

Furthermore, the main contributions of this paper can be summarized as follows:

- (1) This paper proposes a low-light image enhancement scheme that optimizes LIME.
- (2) We find the optimal exposure ratio so that the composite image has a good exposure effect in the underexposed area of the original image.
- (3) We design a low-complexity calculation scheme to obtain the weight matrix.
- (4) In order to improve the performance of the output image, we increase the postprocessing part. In the postprocessing part, the optimal grayscale range of the image is adjusted, and the image is denoised and reconstructed using the BM3D model.

2. Optimize LIME Scheme

Our optimized LIME scheme consists of three parts: pre-processing, weight matrix construction by the LIME method, and postprocessing. The overall algorithm framework is shown in Figure 2.

Among them, the first part consists of a double-exposure sampler, a generator, and an evaluator. The sampler can obtain better exposure of the camera response model; the generator can synthesize the image at the optimal exposure rate to make it in the underexposed area of the original image. The inner exposure is good; since the weight matrix of all pixels is nonuniform, i. e., well exposed pixels are given larger weights, and poorly exposed pixels are given smaller weights; therefore, the weights of all images are weighted using the estimator matrix, which is evaluated such that the output matrix is normalized per pixel. Second, we use the LIME algorithm to determine the weight matrix when the two images are fused. By fusing the composite image and the input image according to this weight matrix, an enhanced image with all pixels well exposed can be obtained. Finally, the postprocessing part improves the performance of the output image by adjusting the optimal grayscale of the image (the flowchart of the postprocessing algorithm is shown in Figure 3). In addition, we use the BM3D model to denoise and reconstruct the image to further optimize the performance after image enhancement.

According to our algorithm framework, the enhanced image can be defined as

$$\mathbf{R}^c = \mathbf{W} \circ \mathbf{P}^c + \circ g(\mathbf{P}^c, \hat{k}). \quad (1)$$

Among them, c is the index of the three color channels, R is the result of image enhancement, W represents the weight matrix, g is the brightness transfer function (BTF), which \hat{k} represents the optimal exposure rate, and P is the input original image, \circ which represents the correspondence

between the two matrices. The elements of the position are multiplied. It is obvious from this formula that to get an enhanced image, we can start with 3 parts: double exposure evaluator (evaluate W), double exposure generator (determine g), and double exposure sample (determine \hat{k}).

2.1. Double Exposure Sampler. The double exposure sampler will determine the optimal exposure \hat{k} for the resulting image so that the input image and the resulting image represent as much information as possible, while leaving the composite image well exposed in areas of the original image that were underexposed.

A well-exposed image has better visibility than an underexposed or overexposed image and can give people more information. So, we have to find the best exposure \hat{k} to provide the most information. To find \hat{k} , we define image entropy as

$$\mathcal{H}(B) = - \sum_{i=1}^N p_i \cdot \log_2 p_i, \quad (2)$$

where B represents the luminance component of the image and p_i represents the i th element of the histogram of B .

Since the entropy of a well-exposed image is higher than that of an underexposed or overexposed image, therefore, it is reasonable to use entropy to find the optimal exposure. We compute the optimal exposure \hat{k} by maximizing the image entropy that enhances brightness, which is expressed as

$$\hat{k} = \operatorname{argmax} \mathcal{H}(g(\mathbf{B}, k)). \quad (3)$$

Since the image entropy first increases and then decreases with the increase of the exposure rate \hat{k} , \hat{k} can be solved by the one-dimensional minimization method. To improve computational efficiency, we resize the input image to 50×50 during optimization.

2.2. Double Exposure Generator. In this part, we use the camera response model to implement the double exposure generator, and the camera is constructed by BTF g .

To estimate g , we select two images with different exposures, named P_0 and P_1 , respectively. The BTF model can be described by a two-parameter function as

$$\mathbf{P}_1 = (g(\mathbf{P}_0, k)) = \beta \mathbf{P}_0^\gamma. \quad (4)$$

Among them, k is the exposure rate, and β and γ are the parameters related to the exposure rate k in the BTF model. It is observed that the different color channels have approximately the same model parameters. The fundamental reason is that for a typical camera, the response curves of different color channels are approximately the same.

Since the BTF of most cameras is nonlinear, our BTF g can usually be solved by

$$g(\mathbf{P}, k) = e^{b(1-k^a)} \mathbf{P}(k^a), \quad (5)$$

where β and γ are two model parameters, which can be calculated from camera parameters a , b and exposure k . We assume that no information about the camera is provided,



FIGURE 1: Naturally captured low-light images.

and we need to use a fixed camera parameter ($a = -0.3293$, $b = 1.1258$) that can fit most cameras.

2.3. Double Exposure Evaluator. The design of the weight matrix W is the key to realizing the enhancement algorithm, which can enhance the low contrast in the underexposed areas while preserving the contrast in the well-exposed areas. We need to assign larger weight values to well-exposed pixels and smaller weight values to underexposed pixels. Intuitively, the weight matrix is positively related to the scene illumination. Since areas of high illumination are more likely to be well exposed, they should be assigned large weight values to preserve their contrast. Inspired by the LIME method [10], this paper calculates the weight matrix by the following formula:

$$W_h(x) \leftarrow \sum_{y \in \Omega(x)} \frac{G_\delta(x, y)}{\left| \sum_{y \in \Omega(x)} G_\sigma(x, y) \nabla_h \hat{T}(y) \right| + \epsilon} \quad (6)$$

$$W_v(x) \leftarrow \sum_{y \in \Omega(x)} \frac{G_\delta(x, y)}{\left| \sum_{y \in \Omega(x)} G_\sigma(x, y) \nabla_v \hat{T}(y) \right| + \epsilon}$$

Here, ∇_h represents the gradient in the horizontal direction, ∇_v represents the gradient in the vertical direction, \hat{T} represents the initial estimate of the light map, ϵ is a very small constant to avoid a denominator of 0, $\Omega(x)$ is the area centered on the pixel x , and y is the position index within the area. $G_\delta(x, y)$ can be expressed as

$$G_\sigma(x, y) \propto \exp\left(-\frac{\text{dist}(x, y)}{2\sigma^2}\right), \quad (7)$$

where $\text{dist}(x, y)$ is used to measure the spatial Euclidean distance between the x and y positions.

2.4. Light Map Estimation. As one of the earliest color constancy methods, the Max-RGB [11] method attempts to estimate the illuminance by finding the maximum value of the three color channels (R , G , and B channels), but this estimation can only improve the global illuminance. In this paper, in order to deal with nonuniform lighting, we choose to use the following initial estimates:

$$\hat{\mathbf{T}}(x) \leftarrow \max_{c \in \{R, G, B\}} \mathbf{P}^c(x). \quad (8)$$

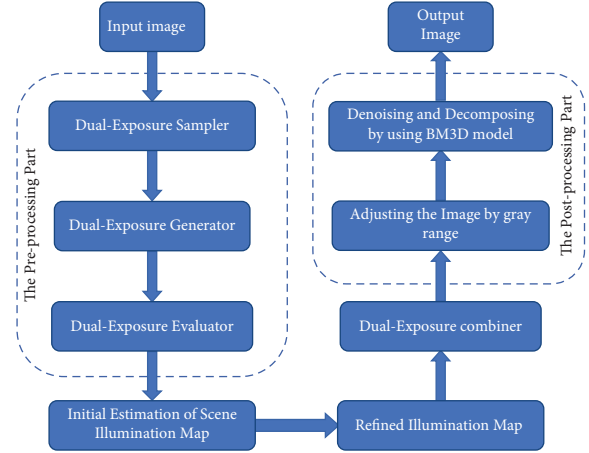


FIGURE 2: Algorithm framework.

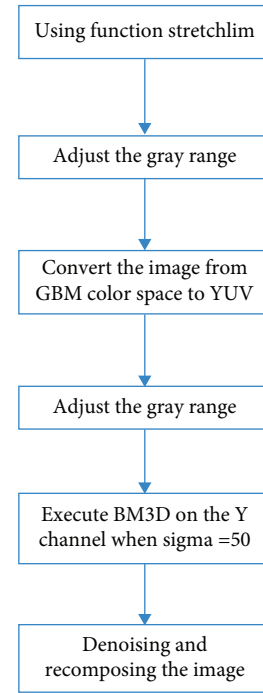


FIGURE 3: Flowchart of the postprocessing algorithm.

For each individual pixel x , the rationale behind the above operation is that the illuminance is at least the maximum value of the three channels at some pixel location.

For the initial estimation of light maps, a good solution should preserve both overall structure and smooth texture details. To solve this problem, we propose to solve the following optimization problems:

$$\min_{\mathbf{T}} \|\hat{\mathbf{T}} - \mathbf{T}\|_F^2 + \alpha \|\mathbf{W} \circ \nabla \mathbf{T}\|_1, \quad (9)$$

where α is a coefficient that balances these two terms, $\|\cdot\|_F$ and $\|\cdot\|_1$ represent the F -norm and 1-norm, respectively, \mathbf{T} is the original light map. $\nabla \mathbf{T}$ is the first derivative filter, and it contains $\nabla_h \mathbf{T}$ (horizontal) and $\nabla_v \mathbf{T}$ (vertical). In Equation

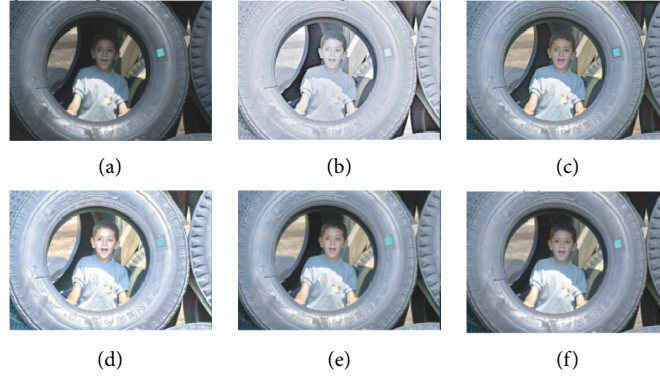


FIGURE 4: BOY1 enhancement results. (a) Original. (b) MSRCR. (c) SRIE. (d) LIME. (e) MF. (f) Proposed.

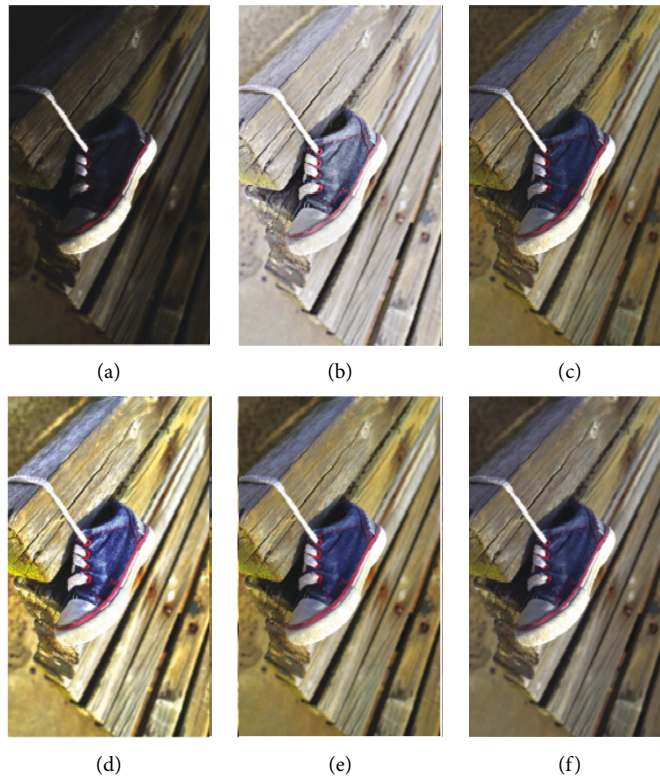


FIGURE 5: SHOE enhancement results. (a) Original. (b) MSRCR. (c) SRIE. (d) LIME. (e) MF. (f) Proposed.

(9), the first term guarantees fidelity between $\hat{\mathbf{T}}$ and \mathbf{T} , while the second term considers (structure-aware) smoothness.

2.5. Low Complexity Solution. On carefully analyzing the problem (9), it is not difficult to find that the origin of the iterative process is the sparse weighted gradient term, that is, the 1-norm term $\|\mathbf{W} \circ \nabla \mathbf{T}\|_1$. The combination of the 1-norm and the gradient operation of T makes the calculation somewhat complicated. Therefore, in order to reduce the computational complexity, formula (10) is obtained:

$$\lim_{\epsilon \rightarrow 0^+} \sum_x \sum_{d \in \{h,v\}} \frac{\mathbf{W}_d(x) (\nabla_d \mathbf{T}(x))^2}{|\nabla_d \mathbf{T}(x) + \epsilon|} = \mathbf{W} \circ \nabla \mathbf{T}_1. \quad (10)$$

That is, $\mathbf{W} \circ \nabla \mathbf{T}_1$ can be approximated by using $\sum_x \sum_{d \in \{h,v\}} \mathbf{W}_d(x) (\nabla_d \mathbf{T}(x))^2 / |\nabla_d \mathbf{T}(x) + \epsilon|$. Therefore, the problem in (9) can be approximately described as

$$\min_{\hat{\mathbf{T}}} \|\hat{\mathbf{T}} - \mathbf{T}\|_F^2 + \alpha \sum_x \sum_{d \in \{h,v\}} \frac{\mathbf{W}_d(x) (\nabla_d \mathbf{T}(x))^2}{|\nabla_d \hat{\mathbf{T}}(x) + \epsilon|. \quad (11)$$

Since equation (11) contains a quadratic term, the problem can be solved directly by solving the following equation:

$$\left(\mathbf{I} + \sum_{d \in \{u,v\}} \mathbf{D}_d^T \text{Diag}(\tilde{\mathbf{W}}_d) (\mathbf{D}_d) = -t \right), \quad (12)$$

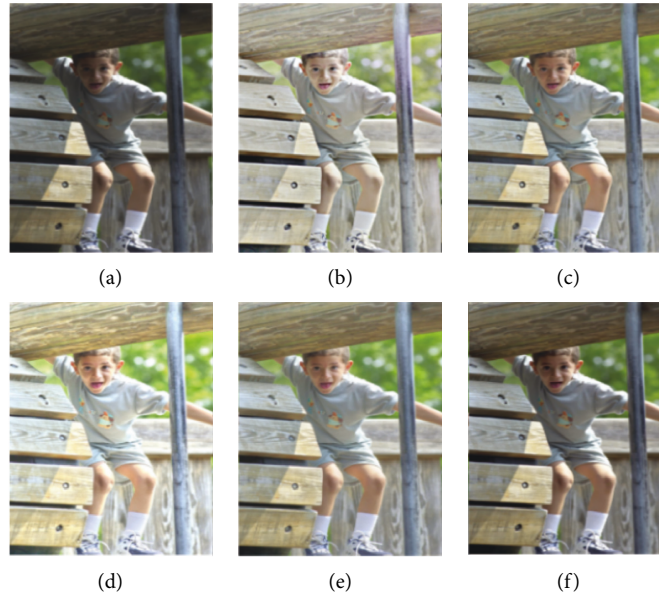


FIGURE 6: BOY2 enhancement results. (a) Original. (b) MSRCR. (c) SRIE. (d) LIME. (e) MF. (f) Proposed.

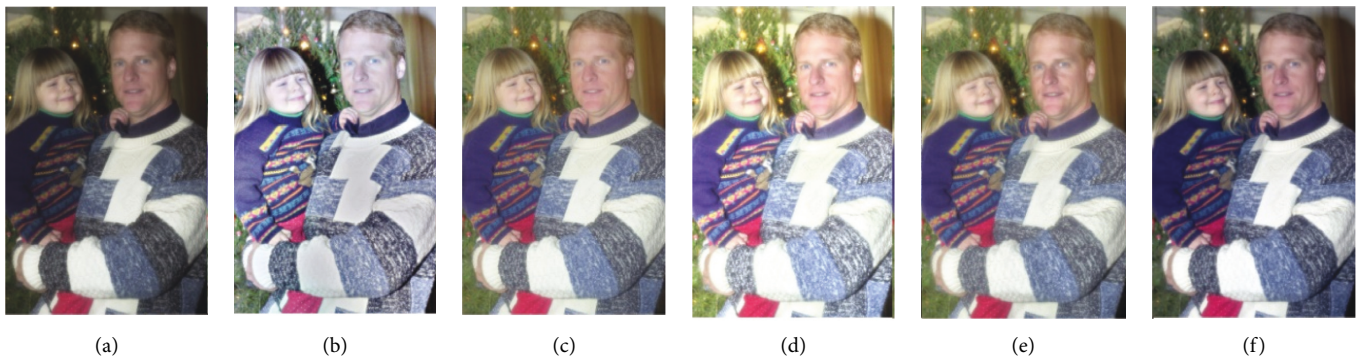


FIGURE 7: FATHER enhancement results. (a) Original. (b) MSRCR. (c) SRIE. (d) LIME. (e) MF. (f) Proposed.

where \tilde{w}_d is the vectorization of \tilde{W}_d , which is expressed as $\tilde{w}_d(x) \leftarrow W_d(x) / |\nabla \hat{T}(x)| + \epsilon$. $\text{Diag}(x)$ is a diagonal matrix constructed with the vector x . L is a symmetric positive definite Laplacian matrix, and there are many practical solutions in [12–16]. After doing all the above steps, we can combine them to get an enhanced image R^c .

2.6. Adjust the Grayscale Range. In this section, we use the `imadjust` function in MATLAB to adjust the grayscale range of the image and set the gamma value in the function to 0.9 to achieve better visual effects.

2.7. Denoising and Restructuring Using BM3D Models. To further improve the visual effect, we also employ a denoising recombination technique. Considering the comprehensive performance, this paper chooses the BM3D [17] model.

BM3D (block matching 3D, a three-dimensional block matching algorithm) can be said to be one of the best denoising algorithms at present. The main idea of this

algorithm is to find similar blocks in the image for filtering processing, which can be divided into two steps: one is basic estimation, and a simple denoising operation is performed. The second is the final estimation, and more detailed denoising is performed to further improve the peak signal-to-noise ratio.

In this paper, to further reduce the amount of computation, we convert the input image P from the RGB color space to the YUV color space and perform BM3D only on the Y channel. In addition, the noise level is different for different input regions due to different magnifications. BM3D models have the same effect on different areas. Therefore, to avoid imbalances in processing, such as some locations (dark areas) being well removed and some locations (bright areas) being oversmoothed, we use the following operations to maintain the balance.

$$\mathbf{R}_f \leftarrow \mathbf{R} \circ \mathbf{T} + \mathbf{R}_d \circ (1 - \mathbf{T}), \quad (13)$$

where \mathbf{R}_d and \mathbf{R}_f are the results after denoising and recombination.

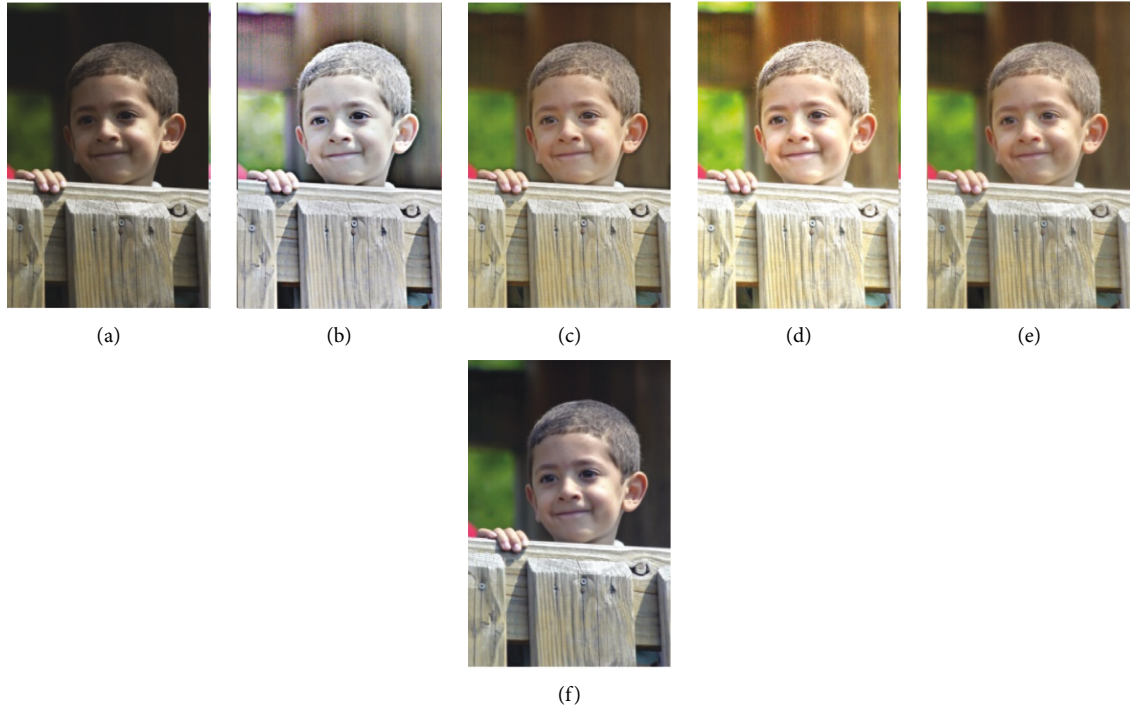


FIGURE 8: BOY3 enhancement results. (a) Original. (b) MSRCR. (c) SRIE. (d) LIME. (e) MF. (f) Proposed.

TABLE 1: Comparison of quantitative performance of different methods in terms of LOE.

Method and image	Comparison of quantitative performance of different methods in terms of LOE				
	BOY	SHOE	BOY2	FATHER	BOY3
MSRCR	1600.15	1673.6	1600.33	1940.24	2160.49
SRIE	599.34	618.83	709.31	564.33	499.77
LIME	1388.30	1288.88	1434.57	1368.46	1329.36
MF	742.75	538.23	807.26	652.99	685.92
Proposed	303.49	446.43	289.62	430.12	301.81

3. Result Analysis

In this section, we compare our scheme with other methods, including MSRCR, SRIE, LIME, and MF methods. The results are shown in Figures 4–8.

To verify the effectiveness of the proposed algorithm, we process some typical low-light images. The names are BOY1, SHOE, BOY2, FATHER, and BOY3. Here are their comparison results:

Furthermore, to demonstrate the effectiveness of the proposed algorithm, we evaluate the enhanced images using the following 3 metrics:

- (1) Lightness order error (LOE)
- (2) Visual information fidelity (VIF)
- (3) Human subjective visual evaluation

3.1. Luminance Sequence Error (LOE). The relative order of brightness represents the light source direction and brightness changes, and the naturalness of the enhanced image is related to the relative order of brightness in different areas. Therefore, we adopt the luminance order error (LOE)

as an objective measure of performance. The definition of LOE is as follows:

$$\text{LOE} = \frac{1}{m} \sum_{x=1}^m \sum_{y=1}^m (U(\mathbf{Q}(x), \mathbf{Q}(y)) \oplus U(\mathbf{Q}_r(x), \mathbf{Q}_r(y))), \quad (14)$$

where m is the number of pixels. The function $U(p, q)$ outputs 1 where $p \geq q$ and 0 otherwise. The symbol “ \oplus ” represents the exclusive or operator. $\mathbf{Q}(x)$ and $\mathbf{Q}_r(x)$ are the maximum values between the R , G , and B channels at the x position of the enhanced image and the reference image, respectively. The lower the LOE, the better the image enhancement and the more natural brightness can be maintained. The results are shown in Table 1.

3.2. Visual Information Fidelity (VIF). VIF models the quality assessment problem as an information fidelity criterion, quantifying the mutual information between the reference image and the distorted image relative to the image information extracted by HVS. The VIF metric [18] can be expressed as

TABLE 2: Quantitative performance results of VIF.

Method and image	Quantitative performance results of VIF				
	BOY	SHOE	BOY2	FATHER	BOY3
MSRCR	0.43525	0.29493	0.66658	0.24967	0.47728
SRIE	0.55014	0.50491	0.77618	0.62694	0.72162
LIME	0.30723	0.20269	0.60919	0.31385	0.51933
MF	0.61596	0.45259	0.98047	0.70466	0.84256
Proposed	0.7585	0.68341	0.89024	0.70322	0.8943

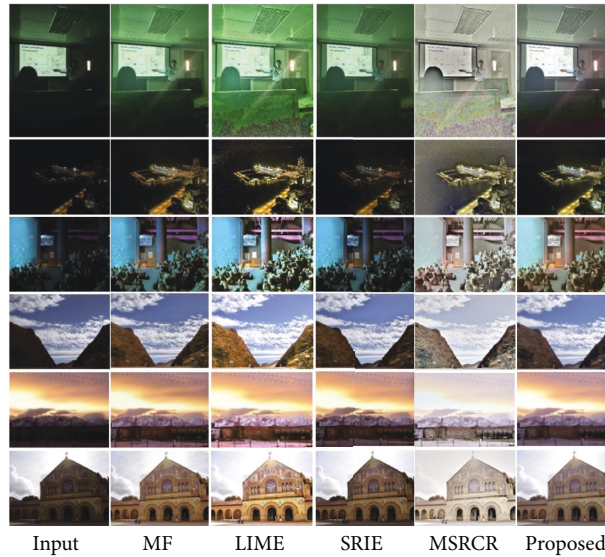


FIGURE 9: Comparison of the effects of different datasets.

TABLE 3: Comparison of quantitative performance of different methods in terms of LOE.

Method	Peng	Dormitory	Screen	Sky	Cloud	Home	LOE(Ave)
MF	275.71	469.04	357.42	642.77	412.88	629.72	464.55
LIME	1382.11	1025.02	1062.62	1175.31	1239.2	1164.51	1174.89
SRIE	283.47	1183.61	412.73	660.49	327.17	701.66	594.88
MSRCR	4442.73	1553.33	1561.13	1797.32	1197.44	1404.82	1992.76
Proposed	525.67	988.44	659.26	405.95	355.22	670.76	600.86

TABLE 4: Comparison of quantization performance of different in terms of VIF.

Method	Peng	Dormitory	Screen	Sky	Cloud	Home	VIF(Ave)
MF	0.62566	0.38453	0.24674	0.62679	0.57326	0.3492	0.4678
LIME	0.19281	0.14978	0.08449	0.33217	0.2367	0.14688	0.19048
SRIE	0.61524	0.54838	0.38424	0.62283	0.55206	0.4555	0.52970
MSRCR	0.16534	0.27782	0.11369	0.45671	0.29517	0.15386	0.24375
Proposed	0.72139	0.63082	0.2164	0.92307	0.87397	0.45882	0.63742

$$VIF = \frac{I(C; F)}{I(C; E)}, \quad (15)$$

where E is the image perceived by the HVS. $I(C; F)$ and $I(C; E)$ represent the information that the brain can extract in reference and test images, respectively. The larger the VIF value, the better the image fidelity. The results are shown in Table 2.

3.3. Subjective Evaluation. Figure 9 is another dataset on which we tested, and we can see the enhancement effect on low-light images under different methods. From the perspective of human vision, MF, SRIE, and our proposed method can effectively enhance images; the LIME algorithm and the MSRCR method have a poor enhancement effect. This is consistent with the data validation of Tables 3 and 4.

The above enhanced images and data can intuitively show that compared with the current mainstream low-light image enhancement methods, the method proposed in this paper can more effectively reduce the distortion of visual information, and the overall visual effect is better.

4. Conclusion

This paper proposes an efficient low-light image enhancement method. The core idea of this method is to synthesize double-exposure images through the camera response model, obtain an optimized light map based on the LIME method, and finally perform postprocessing to further improve the image enhancement effect. The experimental results show that compared with other existing common low-light image enhancement schemes, the method proposed in this paper can achieve better image enhancement effects both subjectively and objectively.

In the future, we should consider environmental factors and avoid over-enhancement due to the unknown environmental conditions. In addition, we can use machine learning-related theories to make further improvements. In our experiments, we used a limited dataset. So, we can try more datasets for more accurate data analysis.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the China University Industry-University-Research Innovation Fund [Grant no. 2021BCE01006].

References

- [1] B. Zhang, Y. Peng, W. Jie, F. Qiming, Z. Yongze, and L. Yu, "An ultraviolet image enhancement algorithm in Curvelet domain based on human visual characteristics," *Telecommunications Science*, vol. 35, no. 3, pp. 99–106, 2019.
- [2] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: a deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.
- [3] M. H. Asmare, V. S. Asirvadam, and A. F. M. Hani, "Image enhancement based on contourlet transform," *Signal Image and Video Processing*, vol. 9, no. 7, pp. 1679–1690, 2014.
- [4] Y. Chen, P. Feng, J. Yang, and Z. Xie, "An adaptive enhancement algorithm of low illumine color video image," *International Journal of Sensing, Computing & Control*, vol. 2, no. 2, pp. 79–86, 2012.
- [5] S. Dhar, H. Roy, M. Majumder, C. Biswas, and A. Sarkar, "A novel method for underwater image segmentation based on M-band wavelet transform and human psychovisual phenomenon (HVS)," in *Proceedings of the 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 21–25, Kolkata, India, November 2017.
- [6] M. Fang, Y. Wang, H. Li, and J. Xu, "Detail Maintained Low-Light Video Image Enhancement Algorithm," in *Proceedings of the 2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1140–1144, Changchun, China, August 2018.
- [7] S. Y. Yu and H. Zhu, "Low-illumination image enhancement algorithm based on a physical lighting model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 28–37, 2019.
- [8] Y. Qiu, Y. Yang, Z. Lin, P. Chen, Y. Luo, and W. Huang, "Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV," *China Communications*, vol. 17, no. 3, pp. 46–57, March 2020.
- [9] Z. Ying, G. Li, and W. Gao, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," *Journal of LaTeX Class Files*, vol. 14, no. 8, pp. 1–10, Nov. 2017.
- [10] X. Guo, Y. Li, and H. Ling, "LIME: LowLight image enhancement via illumination map estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, 2017.
- [11] X. Fu, Y. Sun, M. Liwang, Y. Huang, X. P. Zhang, and X. Ding, "A Novel Retinex Based Approach for Image Enhancement with Illumination Adjustment," in *Proceedings of the 2014 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.
- [12] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Transactions on Graphics*, vol. 31, no. 6, pp. 1–10, 2012.
- [13] E. H. Land, "The Retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, Dec. 1977.
- [14] C. Rang, Y. Jia, Y. Yang, Y. Zhu, and Y. Wang, "Fusion of infrared and visible images based on multi-scale edge-preserving decomposition and sparse representation," in *Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–9, Beijing, China, October 2018.
- [15] D. Krishnan and R. Szeliski, "Multigrid and multilevel preconditioners for computational photography," *ACM Transactions on Graphics*, vol. 30, no. 6, pp. 1–10, 2011.
- [16] T. Sun, C. Jung, P. Ke, H. Song, and J. Hwang, "Readability Enhancement of Low Light Videos Based on Discrete Wavelet Transform," in *Proceedings of the 2017 IEEE International Symposium on Multimedia (ISM)*, pp. 342–345, Taichung, Taiwan, December 2017.
- [17] T. Song, Y. Liu, M. Zhao, L. Cao, and S. Liu, "Optimization BM3D algorithm based on pseudo-3D transform," in *Proceedings of the 2019 18th International Conference on Optical Communications and Networks (ICOON)*, pp. 1–3, Huangshan, China, August 2019.
- [18] T. Kuo, Y. Wei, and K. Wan, "Color image quality assessment based on VIF," in *Proceedings of the 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC)*, pp. 96–100, Singapore, July 2019.

Research Article

Seismic Safety Design and Analysis of Hydraulic Sluice Chamber Structure Based on Finite Element Method

Xin Cai,^{1,2} Zhenming Cui ,¹ Xingwen Guo,¹ Fan Li,^{1,2} and Yanan Zhang¹

¹College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210000, China

²College of Mechanics and Materials, Hohai University, Nanjing 210000, China

Correspondence should be addressed to Zhenming Cui; 140402020004@hhu.edu.cn

Received 21 July 2022; Revised 24 August 2022; Accepted 30 August 2022; Published 16 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Xin Cai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

One of the important links in the safety evaluation of sluices is the aseismic safety examination. In order to ensure the daily safe operation of sluices, it is necessary to conduct a normalized aseismic safety examination of sluices, and it is also necessary to study the aseismic safety examination of return sluices. Based on the application of ADINA finite element analysis software, a three-dimensional finite element model of the gate chamber structure is established, and the seismic response of the gate chamber structure is calculated and analyzed by the mode decomposition response spectrum method. The seismic safety of the gate chamber structure is evaluated comprehensively. The results show that 2.00 MPa of tension stress is generated at the junction of the pier and the gate. According to the structural mechanical method, the maximum tensile stress that can be endured is 4.41 MPa, which meets the safety requirements. There is a large tension stress zone between the elevator floor and some parts of the elevator, which far exceeds the standard tension strength value of the concrete moving shaft. Considering the safety, corresponding aseismic reinforcement measures should be taken. The structure of the gate chamber is non-slip and stable, and the safety factor is larger than the standard value of the Gate Design Specification (SL265-2016), which meets the safety requirements. The aseismic safety of the gate chamber structure meets the requirements of the “Standard for Seismic Design of Water Conservancy Buildings” (GB5127-2018), but it has safety defects and the aseismic grade is *B*.

1. Introduction

According to the national water conservancy safety data in 2009 and the record data of sluice in 2008, there are 97 small (1) type or above directly managed sluice in the Yellow River Basin. The main structural forms are divided into open type and culvert type [1]. The basic stratum types of the gate are generally the fourth system of the altered series, the upper altered series of the fourth system, the middle altered series of the fourth system, the lower altered series of the fourth system, the upper Triassic series, and the upper Cambrian series, which are mainly caused by artificial accumulation, alluvial deposition, and flood. The soil properties are divided into 24 different lithology, such as fine sandy loam, loam, sandy soil, silty soil, sandy gravel, light silty loam, sandy soil, silty sand, and screened fine soil. The second largest earthquake area in China's history is the North China earthquake zone [2].

Due to the change of the seismic design specifications of domestic buildings, the seismic design of locks mainly faces the following three problems: (1) the original engineering design did not fully consider the seismic protection and the seismic protection needs to be fully considered according to the standard. According to the existing national seismic work parameter zoning map (gb18306-2015), the structure of building fortification strength and seismic capacity assessment needs to be redefined; (2) the original design considered earthquake protection, but the building fortification strength is low, such as 6 and 7 degrees. Now, according to the code for seismic design of water conservancy building materials (sl191-2008), the fortification strength has increased, and the seismic capacity needs to be reviewed; (3) earthquake is considered in the original design, and the fortification intensity has not changed [3]. Strengthening the protection measures against earthquakes

leads to insufficient earthquake resistance in the original design, and it is necessary to focus on the review of the protection measures against earthquakes [4]. Today, with the development of calculation methods and the improvement of social and economic technology, the seismic characteristics of building materials will become increasingly important. Compared with the current requirements for seismic fortification, the initial design of gate using quasi-static calculation method is slightly insufficient. Therefore, more advanced calculation methods and theories are needed to analyze the comprehensive seismic capacity of buildings.

There are many natural disasters caused by earthquakes and they occur relatively frequently in China. Seismic safety evaluation is the main part of ship lock safety evaluation project. The throat of ship lock construction is the lock chamber, and the design of the lock chamber plays an important role in the safety of the project. Therefore, it is necessary to discuss the seismic stability of the gate chamber structure. At present, most researchers use quasi-static method to review and analyze the gate structure [5]. In fact, the sluice is a three-dimensional thin-wall structure [6]. Connections between buildings, such as traffic bridges, have large errors in the results [7]. In recent years, with the rapid development of computer technology, the finite element numerical simulation technology has been widely used in the seismic analysis of sluices [8]. Therefore, it is necessary to use the three-dimensional finite element numerical simulation technology to study the seismic resistance of the gate chamber structure [9].

2. State of the Art

2.1. Overview of Sluice Gate. The gate refers to a low head hydraulic structure that can adjust the flow of the water body by opening the gate to adjust the flow or adjust the temperature. As the gate plays a major role in regulating water flow and generating electricity, it is widely used in the field of water conservancy. As early as the spring and autumn period, the Chinese began to build ship locks. It is reported that the earliest ship lock in China was built in today's Anhui Province. There are five ship locks in China. According to statistics, by 2008, China had built more than 50000 flood discharge gates, including more than 480 large-scale flood discharge gates, which improved the local drainage and irrigation capacity and accumulated rich experience in promoting projects.

2.1.1. Classification of Sluice Gates. According to the different responsibilities of sluice structures in water conservancy, sluice gates can be classified as follows:

- (1) *Control the brake.* Generally speaking, the control brake is built on the barrage between the main river and the branch river, and the water level is controlled through the ship lock to meet the requirements of upstream and downstream water diversion and navigation, control the upstream and downstream water flow to ensure the safety of the upstream and

downstream rivers, or control the flow direction according to the downstream water supply requirements [10]. If the control gate and the sluice are built together somewhere, it can not only control the inflow flow but also completely discharge the excess water, thus protecting the safety of hydraulic structures [11].

- (2) *Entry lock (entrance lock).* It is mostly built on the banks of rivers, ponds, or lakes or above the diversion channels of agricultural irrigation areas to regulate the amount of tap water and ensure the demand of agricultural irrigation water supply, industrial water, and domestic water [12].
- (3) *Flood gate.* Generally, mountain torrents, which are mostly located on the side of the river and larger than the safe flow of the downstream river, can be divided into flood storage area, flood detention area, and flood diversion channel [13].

2.1.2. Components of the Sluice. According to the composition of sluice, it can be generally divided into three parts: (1) sluice chamber [14] and (2) upstream connection section [15]. The sluice chamber is the most important part in the structure of the sluice. The sluice chamber consists of the bottom plate, middle pier, side pier, gate, daughter wall, maintenance bridge, work bridge, and traffic bridge. The upstream and downstream water level and flow can be changed by adjusting the opening of the gate. The bottom plate of the gate can reduce the effect of osmotic pressure on the whole gate. All the structures of the room are self-respecting. Upstream connection section is usually composed of the above subgrade, bottom protection, upper inverted chute, wing beam, and embankment on both banks. In particular, the mat ensures the chamber and reduces scour and osmotic pressure. The upstream bottom and the squeezed riverbed greatly reduce the scouring of the water flow. The upstream wing wall is designed to form a favorable water flow situation, let the water flow smoothly into the gate area, and achieve the functions of soil blocking, anti-scouring, and anti-lateral seepage.

2.2. Calculation Theory of Sluice Seismic Design

2.2.1. Influence of Earthquake on Sluice Gate. The gate structure has been damaged by strong earthquakes for many times, which seriously affected the normal operation of the gate. However, according to the statistics of the earthquake disaster data after the previous major earthquakes, the main earthquake disaster causes of the gate in China are the fracture of the protection pool, the bottom plate, the damping pool, and the bottom plate structure. The main factors leading to this fracture phenomenon are as follows: (1) the strong earthquake resistance causes the imbalance of the sluice, causing structural displacement, fracture, and even collapse or uplift; (2) under the influence of earthquake damage force, the strength and safety of the structure are damaged, resulting in fracture, deflection, and even

collapse [2], and due to the influence of earthquake force, the strength and safety of the structure are damaged, resulting in cracks, deflection, and even collapse [16]. Earthquake resistance is a complex and changeable natural event. There are many factors that affect the impact of earthquake on the gate. The main reasons are as follows: (1) seismic speed and time; (2) the shockproof function of the gate itself; (3) the basic geological structure of the gate, the time of its generation, and the function of the building; and (4) static load applied on the sluice structure [17]. According to the Geological Survey Report of Locks in the Middle and Lower Reaches of the Yellow River and the Engineering Geological Research Atlas of the Lower Reaches of the Yellow River, 42 locks in the middle and lower reaches of the Yellow River are investigated. And the histogram of their foundation soil quality is shown in Figure 1.

According to the geological conditions, soil types and soil thickness of 42 sluice gates in the middle and lower reaches of the Yellow River, the average thickness of various soil layers is calculated. The calculation results are as follows: the proportion of fill is 0.5%, the proportion of silt is 15.1%, the proportion of loam is 26.3%, the proportion of fine sand loam is 17.1%, the proportion of screened fine soil is 17.5%, and the proportion of neutral clay is 7.3%. The proportion of coarse sand is 2.2%, the proportion of silt is 2.8%, and the proportion of clay is 10.7%, and Figure 2 shows the composition map of the foundation soil of the sluice chamber section in the middle and lower reaches of the Yellow River.

3. Methodology

3.1. Seismic Calculation Theory of Sluice Gate

3.1.1. Seismic Response of a Single Degree of Freedom System.

The so-called single degree of freedom system means that the structure is only a motion acceleration system [18]. In the seismic analysis, the ground is assumed to be in rigid motion, and the structure is a vibration system with mass M -spring K -damping c located on this rigid plane. According to Newton's second law, the following relationship can be obtained:

$$M\ddot{x}(t) + c\dot{x} + Kx(t) = -M\ddot{x}_g(t). \quad (1)$$

In construction engineering, the fixed vibration period T of the structure is generally used as the fixed vibration frequency ω ($t = 2\pi/a$) to represent the seismic response spectrum [19]. There are the following approximate relationships between relative displacement, relative velocity, and absolute acceleration:

$$S_D(\omega, \xi) = |x|_{\max} = \frac{S_V(\omega, \xi)}{\omega}, \quad (2)$$

$$S_A(\omega, \xi) = \left| \ddot{x} + \ddot{x}_g \right|_{\max} = \omega^2 S_D(\omega, \xi). \quad (3)$$

Because of these relationships, some textbooks often draw three kinds of spectra together, which is also called three coordinate spectrum, or three spectrum diagram [19]. Seismic acceleration response spectrum is very common in building seismic engineering.

3.1.2. Seismic Response of Multi-DOF Systems. Finite element method is the most commonly used method to analyze seismic response of multi-degree-of-freedom systems. The array superposition method is commonly used in seismic response analysis of sluice. The first step of the analysis and research is to deal with the characteristics of the structural system so as to achieve the first few low-order structures with large research on the vibration of the structural system. The second step is to calculate the motion equation of the system through the orthogonality of the stratum. Using the decoupling method, the stratum responses corresponding to different strata are obtained, and then, the responses of each stratum are superposed according to certain principles to obtain the comprehensive seismic response of the structure. The free vibration equation of the undamped multi-degree-of-freedom linear system after finite element discretization is

$$[M]\{\ddot{u}(t)\} + [K]\{u(t)\} = \{0\}, \quad (4)$$

where $[M]$ and $[K]$ are the overall stiffness matrix and mass matrix of the system, respectively; $\{u(t)\}$ is the displacement reflection of the system.

$$\{u(t)\} \text{ can be expressed as: } \{u(t)\} = \{U\} \sin(\omega t + \phi). \quad (5)$$

Since the formation of the system has the characteristic of weighted orthogonality, its dynamic response can be expanded according to its formation, that is,

$$\{u(t)\} = \sum_{i=1}^N q_i(t), \quad (6)$$

where $q_j(t)$ is the generalized coordinate of the amplitude change of the formation, which reflects the contribution of the j th-order formation to the total response of the system at time t .

Modal superposition method can get the response characteristics of the structural system in the whole seismic response, so it is called modal superposition time history analysis method. In the design practice, usually, the maximum seismic response of the system is the most unfavorable and most valued. Therefore, on the basis of modal superposition time history method, a new superposition response spectrum method is formed by combining the theory of seismic response spectrum and modal superposition theory so that the maximum seismic response of the system can be calculated simply. At present, this technology has been applied in the seismic engineering design of building engineering.

The focus of the mode shape superposition response spectrum method is to find the maximum value of the structural response, so it is necessary to first obtain the maximum value of the structural response of each mode shape. It can be seen from the above formula that the

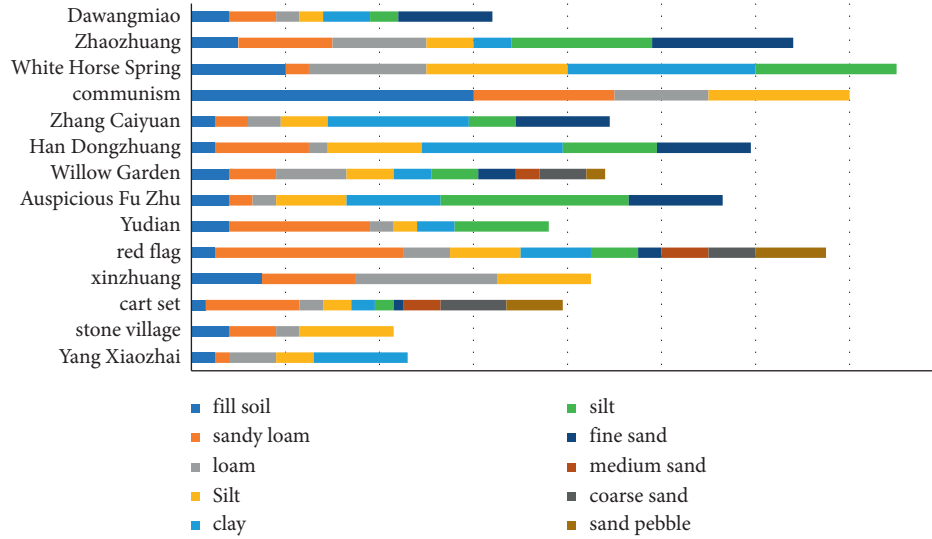


FIGURE 1: Geological column map of the bottom floor of the sluice gates in the middle and lower reaches of the yellow river.

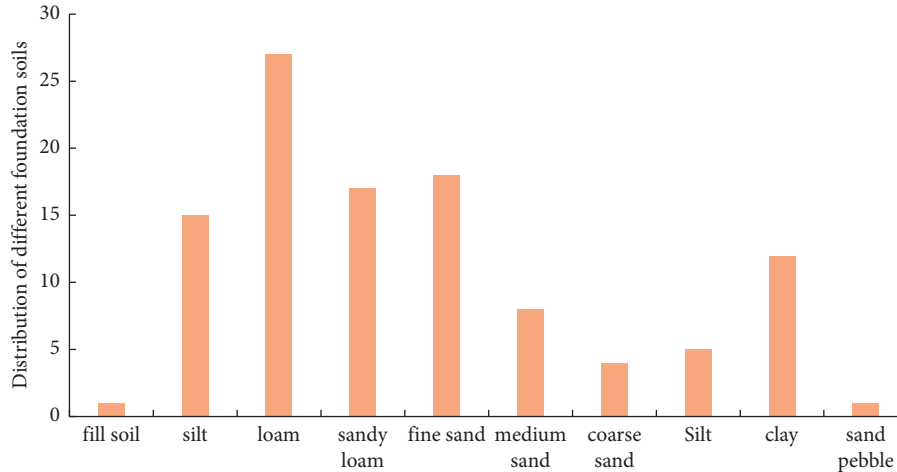


FIGURE 2: The composition of different foundation soils of the bottom plate of the sluice gate in the middle and lower reaches of the yellow river.

maximum value $\{S\}_j$ of the relative displacement response corresponding to the mode shape j can be expressed as

$$\{S\}_j = \{u(t)\}_j^{\max} = \gamma_j \{\Phi\}_j \left| \delta_j(t) \right|_{\max}. \quad (7)$$

3.1.3. Mode Shape Decomposition Response Spectrum Method. This time, the standard design maximum response spectrum specified in the “standard for seismic design of hydraulic structures” (cb51247-2018) (hereinafter referred to as the design code standard) will be used for the seismic evaluation and design of the gate chamber design. As shown in Figure 3, the damping ratio is about 7%, and the representative value of the maximum response spectrum β and Max is 2.25.

According to the provisions of the national standard value of earthquake resistant design strength, when the seismic action effect is estimated by using the mode

decomposition response spectrum method, the seismic action effect of each mode can be combined according to the square root and square root. The specific calculation formula is as follows:

$$S_E = \sqrt{\sum_i^m \sum_j^m \rho_{ij} S_i S_j}, \quad (8)$$

where S_E is the seismic action effect; S_i and S_j are the seismic action effect of the i th and j th order modes, respectively; m is the number of modes used in the calculation.

3.2. Tensile Stress Review. At present, there is very little information about how to evaluate the stability in the range where the tensile stress is greater than the standard deviation of the axial tensile strength of concrete in the finite element calculation. However, according to the conclusion of the

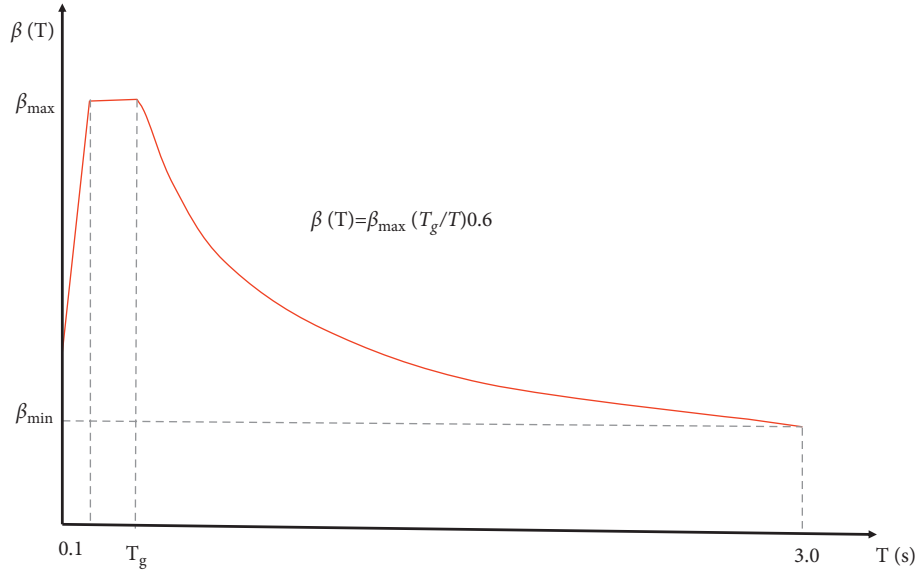


FIGURE 3: Standard design response spectra.

finite element seismic review of the open gate, it is not difficult to see that in the case of a large earthquake, the intersection between the gate pier and the gate bottom plate can often produce a large tensile stress area, and the tensile stress in these areas is often greater than the standard value of the concrete dynamic axial tensile strength. Under the influence of earthquake, the gate pier is an eccentric compression member. Therefore, for the sake of safety and considering the worst case, the gate pier can be considered as a pure bending structure. According to the code for design of hydraulic concrete structures (sl191-2008), considering the bending capacity of the positive diameter and the normal rectangular section or inverted T -shaped section of the flange at the tension side, the bending capacity of the positive diameter of the bending member must meet the following conditions:

$$KM_s \leq f_c b x \left(h_0 - \frac{x}{2} \right) + f_y' A_s' (h_0 - a_s'). \quad (9)$$

In the formula: K is the safety factor of bearing capacity; M_s is the design value of bending moment, N.m; f_c is the design value of concrete axial compressive strength, Pa; A_s is the cross-sectional area of longitudinal tension steel bar, m^2 .

Meanwhile, for pure curved surface components, the normal stress at any point on the section can be measured. The equation is

$$\sigma = \frac{M y}{I_z}, \quad (10)$$

where m is the maximum bending moment on the longitudinal section. To obtain the maximum normal stress acceptable on the longitudinal section, here M can be taken as MS , and I is the maximum moment of inertia on the longitudinal section with respect to the neutral axis Z ; Y is the required maximum internal stress.

By comparing the stress data obtained by the above method with the finite element results, the range where the

tensile stress is greater than the standard value of the dynamic axial tensile strength of cement can be checked.

4. Result Analysis and Discussion

4.1. Finite Element Model Construction and Parameter Design.

There are four holes in a gate project, i.e., two reinforced concrete open sluice gates, with joints between each joint. The total length of the gate chamber is 12.50 m, the total width of the gate chamber is 47.45 m, the thickness of the side pier is 1.40 m, the thickness of the middle pier is 1.40 m, the thickness of the middle joint pier is 1.85 m, and the net width of each hole is 10.00 m. Considering the difference of earth pressure and water pressure at the two ends of the side pier, this calculation will focus on the combination of the side hole and the gate chamber.

According to the specific standards of the gate chamber design, the structure set includes the gate bottom plate, gate pier, steel gate, cross beam, and side hole 3D finite element model of the hoist frame structure. Using Cartesian coordinate system, take x as horizontal azimuth, y as downstream azimuth, and Z as vertical azimuth. In the calculation process, three-dimensional stability constraints are applied to the bottom of the gate pier and gate. However, it must be noted that the construction of the hub has been carried out for more than 50 years. According to the monitoring data, the land subsidence has been basically balanced. The maximum displacement value is less than 0.5 mm. Therefore, for the convenience of analysis and calculation, the influence of foundation is ignored. In addition, since the gate system of the pulling machine room on each pier of Luqiao road exists separately, thin-layer elements are set between the two adjacent road and bridge construction and hoisting engineering rooms for finite element calculation. The relative independence between the two lifting bays and the highway bridge can be realized without participating in the calculation.

In this calculation result, if the reinforcement unit is not considered, any cement unit represents plain cement. In order to reflect the effect of reinforcement diameter on the elastic modulus of cement, the equivalent elastic modulus is used to simulate the elastic modulus of reinforced concrete. These calculation results of the material parameters used are listed in Table 1.

4.2. Experimental Results and Analysis. The structure's self-vibration characteristics are analyzed by means of structure's self-vibration characteristics. Considering the influence of water in front of the gate on the structure of the gate chamber and the self-vibration of the first five frequencies and modes, the characteristic parameters of the gate chamber structure are obtained. Wester-Gard additional mass method is used to simulate structural effects. The first five natural frequencies and periods of the lowering gate chamber structure under normal water level are shown in Table 2. From Table 2, it can be seen that the natural fundamental frequency of gate chamber structure under normal impounding condition is 4.038 Hz, and the second-order natural vibration frequency is similar to the basic frequency. This is mainly due to the independence of each hole in the gate chamber structure and the hoist room on the gate pier. The first and second vibration modes of the gate chamber structure are two independent hoist rooms along the river direction.

Figure 4 shows the schematic diagram of the side pier and reinforcement of the maximum diameter section of each unit of the sluice. Through the tensile stress review method introduced in Chapter 2.2, the maximum pressure on the longitudinal section of the side pier of each unit can be calculated by equations (1) to (3). The bending moment is 2879.6 kn. M. It can be seen from formula (4) that the maximum allowable tensile stress at the intersection of side pier and gate bottom is 4.41 MPa, equal to 2.00 MPa, which is in line with general safety regulations.

In the stage of sluice modal analysis, according to the calculated diversion water level standard, the upstream water level of the sluice house section is 37.40 m, while the downstream water level of the sluice house section is only 37.40 m. The soil boundary is generally fixed, and the following design modes are generally adopted: (1) fixed boundary + no mass foundation model, (2) fixed boundary + massless foundation model + hydrodynamic pressure, (3) stable boundary + mass base model, and (4) stable boundary + mass basic model + dynamic water pressure; the first 10 natural vibration frequencies of the sluice design are shown in Table 3.

According to the above calculation model and parameters, the dynamic response results of the sluice under the two working conditions are calculated. Now only the displacement diagram of the X -direction of the cross section of the sluice chamber structure, frame bridge, and hoisting machine room under the action of the Henghe-direction earthquake is given. Figure 5 is a displacement diagram in the X direction of the cross section ($z = 0.715$ m, $z = -6.64$ m) of the sluice chamber structure under the action of the Henghe earthquake.

The analysis of the displacement results of the sluice structure under the earthquake conditions in the Yokogawa direction shows that the displacement of the bottom plate and the middle part of the top plate of the sluice chamber structure behind the sluice parapet is larger than the displacement on both sides.

Under the two seismic conditions, the maximum displacement of the sluice structure in the X , Y , and Z directions and the deformation of the different structures of the sluice in the X , Y , and Z directions are shown in Table 4.

The analysis of the structural deformation results of the two seismic conditions shows that when the overall model of the sluice including the frame bridge of the sluice and the hoisting machine room is established, the sluice chamber is under the pressure of the side fill in the direction of the river; in the direction of the river, the sluice chamber is constrained horizontally. Under the action of the river-direction earthquake, the sluice structure has a translational displacement in the transverse direction of the river, and its value is 3.61 cm. Under the action of the river-direction earthquake, the sluice chamber structure has no translational displacement.

Table 5 shows the calculation results of stability against sliding of gate chamber structure superimposed by dynamic and static state. It can be seen from Table 3 that fixed boundary is used for soil boundary and mass-free foundation model is used for gate modal analysis. Based on the integration of self-vibration frequency of gate structure and soil, the hydrodynamic pressure of gate chamber also acts on the pier surface through mass unit MASS21. According to the finite element dynamic calculation, the maximum horizontal seismic pressure in the side hole gate house is about 2863.72 kn. Since the earthquake action is random reciprocating, when the ground plane seismic inertial motion faces the downstream, according to the static action calculation, the anti-skid stability safety factor K of the gate house structure is 2.30, which meets the requirements. When the ground inertial motion faces the upstream horizontal earthquake, the anti-skid stability safety factor K of the gate house structure is 4.64 according to the static action calculation, which meets the requirements.

Because of the dynamic displacement response curve in the Hengchuan direction, the dynamic displacement difference of the typical displacement node between the rack bridge and the hoist room with time is relatively small, and only the dynamic displacement strain curve of the typical node of the gate house with time in the Hengchuan direction is given (see Figure 6).

The comparison results of sluice structure displacement of different models show that the dynamic displacement of the sluice chamber structure is larger than that of the sluice frame bridge and the hoisting machine room. Under the same boundary foundation model, the dynamic displacement of the sluice chamber structure can be calculated: (1) the peak dynamic displacement of the model is smaller than the peak value of the model, (2) the dynamic displacement of the model and the sluice chamber structure, (3) the model is small, and (4) the model is dynamic. In addition, under the same boundary foundation model, the

TABLE 1: Concrete material parameters.

Material number	Material name	Density	Elastic mold	Poisson's ratio	Standard value of dynamic axial compressive strength	Standard value of dynamic axial tensile strength
1	Brake floor	2548.00	33.23	0.167	19.66	1.97
2	Pier	2548.00	31.20	0.167	16.22	1.62
3	Highway bridge	2548.00	31.52	0.167	16.72	1.67
4	Open and close the machine room	2548.00	31.52	0.167	16.72	1.67

TABLE 2: Period table of natural frequency of gate chamber structure under different conditions.

Order	Normal water level	
	Frequency	Cycle
1	4.038	0.248
2	4.178	0.239
3	7.708	0.130
4	13.013	0.077
5	13.820	0.072

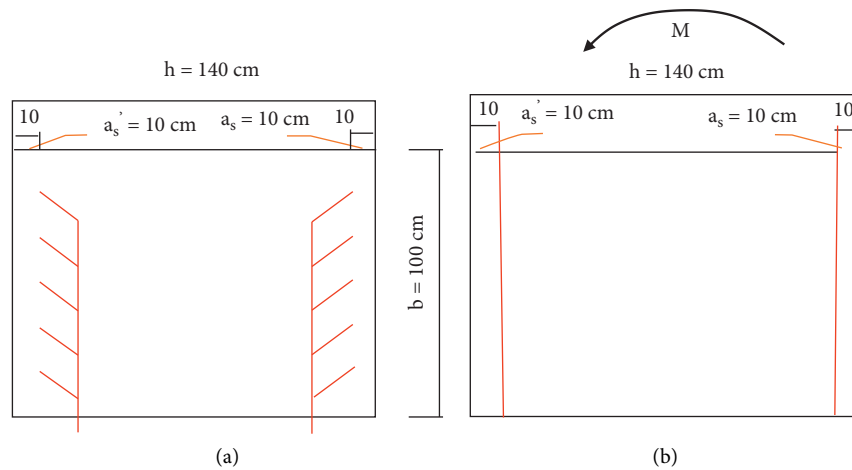


FIGURE 4: Reinforcement diagram of side pier per unit length of sluice. (a) Cross section. (b) Longitudinal section.

TABLE 3: First 10 natural vibration frequencies of four models of the sluice.

Order	1/Hz	2/Hz	3/Hz	4/Hz
1	1.720	0.084	0.942	0.084
2	1.932	0.109	0.992	0.109
3	3.606	0.136	1.081	0.136
4	3.686	1.361	1.220	0.961
5	4.302	3.606	1.569	1.059
6	5.885	5.781	1.569	1.099
7	7.039	7.039	1.751	1.569
8	12.180	7.961	1.828	1.569
9	14.584	11.083	1.881	1.748
10	15.482	11.697	1.956	1.764

dynamic displacement of the sluice bridge and the hoist room structure is the same due to different time, while the dynamic displacement of the sluice chamber structure is

the same at the beginning time. After 1.25 s, the dynamic displacement peak value of the model is higher than that of the model without mass.

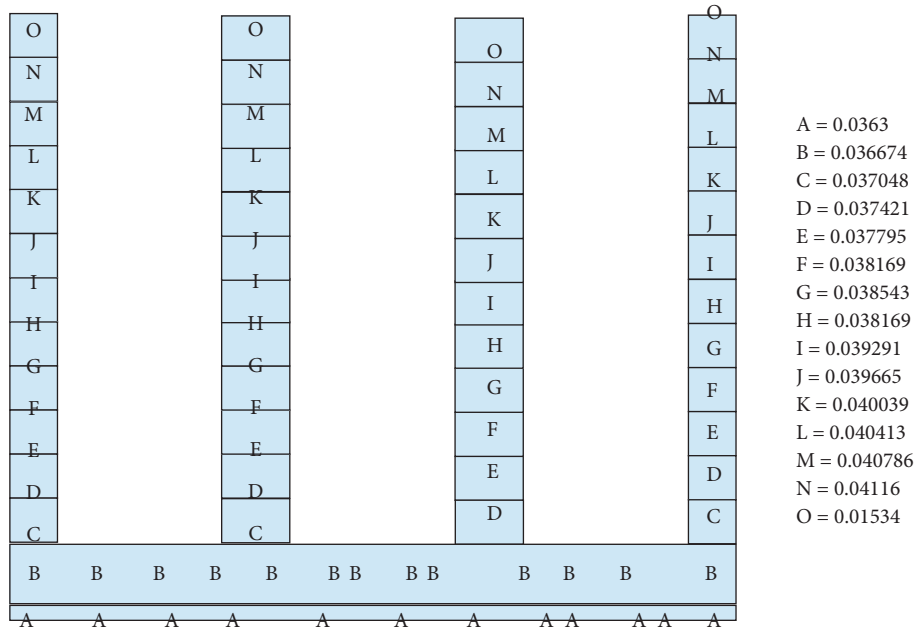


FIGURE 5: Displacement S_x contour map (m) of cross section ($z = 6.64$ m) during the Henghe earthquake.

TABLE 4: Maximum displacement of sluice structures in X, Y, and Z directions under two seismic conditions.

Working condition	S_x	S_y	S_z
Working condition 1	8.87	1.44	0.38
Working condition 2	0.05	0.29	5.24

TABLE 5: Calculation and analysis table of anti-sliding stability of gate chamber structure superimposed by dynamic and static state.

Working condition	Vertical load under static condition	Horizontal load under static condition	Horizontal seismic inertia	Friction coefficient	Anti-skid stability factor	Canonical value
Horizontal seismic inertial force upstream	25197.52	964.93	-2863.72	0.35	4.64	1.10
Horizontal seismic inertial force downstream	25197.52	964.93	2863.72	0.35	2.30	1.10

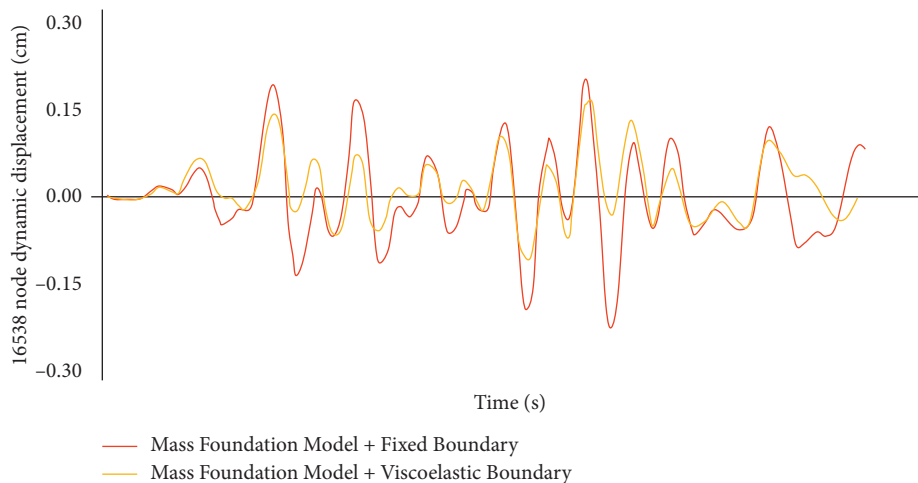


FIGURE 6: Impact curve of lower boundary form on results of quality basis model.

5. Conclusion

This paper summarizes the damage caused by the earthquake to the gate structure and studies the causes of the damage caused by the earthquake to the gate structure. On the basis of this analysis and research, the basic concept of the seismic analysis and research is systematically expounded; the role of the gate, foundation, and bottom plate in the seismic design is comprehensively investigated; and the whole gate design is calculated and analyzed. Through the finite element numerical simulation technology, the vibration resistance test of the gate house structure is carried out, and the following results are obtained according to the calculation: (1) under the normal horizontal condition, the fundamental frequency of the gate house structure is 3.57 Hz, and the first mode is the vibration of the lifting chamber along the river; (2) when the gate pier intersects with the gate bottom plate, the gate mouth has a tensile stress of 2.0 MPa. According to the quantity of reinforcement here, the maximum tensile stress that the place can bear is about 4.41 MPa, which has met the safety requirements. This value far exceeds the standard value of dynamic axial tensile strength of concrete. The reinforced concrete at the bottom of the bridge falls off, the tendon is exposed, and the crack is expanded. (3) Under normal operation, the maximum sliding stability safety factor of the gate chamber structure is 2.30, which is in line with the safety regulations. According to the guidelines for safety evaluation of sluices, the seismic safety of sluices conforms to the relevant provisions of the standard for seismic design of water conservancy buildings and has design defects that do not affect the overall safety, and its seismic grade is *B*.

Data Availability

The labeled data set used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The study was supported by Research on Optimization Design of Sluice Structure on Soft Soil Foundation (51365112), Research on Macroscopic Mechanical Properties of Cementitious Sand-Gravel Dam Based on Micro-Level Damage Evolution Characteristics (519040311), and Optimization Design Theory of Cemented Particle Dam Based on Structural Interface Evolution (518051512).

References

- [1] A. Dost and A. K. Chaudhary, "Seismic resistant design and analysis of (G+15), (G+20) and (G+25) residential building and comparison of the seismic effects on them," *IJIRT*, vol. 15, no. 1, pp. 99–107, 2021.
- [2] M. Abdolmohammadi, J. Ahmadi, and K. Nasserzadeh, "Extending the performance-based design method for seismic response analysis of self-centering structures," *International Journal of Civil Engineering*, vol. 20, no. 6, pp. 721–734, 2022.
- [3] V. H. Sobrado, R. Yaranga, and J. D. Orihuela, "Analysis of seismic bidirectionality on response of reinforced concrete structures with irregularities of l-shaped plan and soft story," *IOP Publishing*, vol. 15, no. 21, pp. 33–41, 2020.
- [4] S. Barbarelli, M. Amelio, and T. Castiglione, "Design and analysis of a new wave energy converter based on a point absorber and a hydraulic system harvesting energy from waves near the shore in calm seas," *International Journal of Energy Research*, vol. 5, no. 104, pp. 12–20, 2020.
- [5] A. Mazaheri, M. C. Seifabad, and S. Mahdavi, "Seismic sensitivity analysis of rigidity and thickness of tunnel lining by using Ground Structure interaction method case study: roudbar lorestan dam," *Geotechnical & Geological Engineering*, vol. 15, no. 9, pp. 1–26, 2020.
- [6] S. K. David, N. Biswal, and K. Muduli, "Design and analysis of an hydraulic trash compactor," *Test Engineering and Management*, vol. 82, no. 20, pp. 8877–8888, 2020.
- [7] Z. Qi, X. Peng, J. Man, C. Lin, and W. Duan, "Design method and finite element analysis of a new prefabricated steel special-shaped lattice column," *E3S Web of Conferences*, vol. 198, no. 16, Article ID 03012, 2020.
- [8] K. Fujita, K. Koyama, K. Minami et al., "High-fidelity nonlinear low-order unstructured implicit finite-element seismic simulation of important structures by accelerated element-by-element method," *Journal of Computational Science*, vol. 49, no. 1, Article ID 101277, 2021.
- [9] M. M. Jegan, B. Pitchia Krishnan, M. K. Shanmugam, P. Infant Kebin Raj, and K. T. Bose, "Design and analysis of hydraulic fixture for hydraulic lift housing," *Journal of Physics: Conference Series*, vol. 1964, no. 7, Article ID 072019, 2021.
- [10] S. Abbasi, R. Daneshfaraz, and M. A. Lotfollahi-Yaghin, "A review on the design and analysis of the hydraulic performance of labyrinth weirs," *Civil Engineering*, vol. 53, no. 11, pp. 1–21, 2021.
- [11] H. Shi, Y. Yue, H. Wang, J. Xu, and X. Mei, "Design and performance analysis of human walking induced energy recovery system by means of hydraulic energy conversion and storage," *Energy Conversion and Management*, vol. 217, no. 6, Article ID 113008, 2020.
- [12] Y. Zhu, Y. He, and J. Ren, "Structure analysis and optimization design of the base of four-pillar hydraulic testing machine," *Journal of Physics: Conference Series*, vol. 2002, no. 1, Article ID 012036, 2021.
- [13] X. Ji, S. Liao, X. Zhang, X. Gao, and R. Ma, "Finite element analysis and optimization design of beam structure of forging press," *Journal of Physics: Conference Series*, vol. 1983, no. 1, Article ID 012018, 2021.

- [14] G. M. Kravchenko, E. V. Trufanova, and K. K. Mazhiev, "The comparison of methods for modeling seismic impact on buildings and structures," *IOP Publishing Ltd*, vol. 7, no. 4, pp. 22–31, 2022.
- [15] N. Manage and N. T. Medagedara, "Design and development of a vision-based hydraulic operated mending machine for repairing 3W forks," *Journal of Instrumentation: English edition*, vol. 9, no. 4, p. 10, 2020.
- [16] A. A. Zamani and S. Etedali, "Seismic response prediction of open- and closed-loop structural control systems using a multi-state-dependent parameter estimation approach," *International Journal of Computational Methods*, vol. 19, no. 05, pp. 105–111, 2022.
- [17] P. Pokharel, "1 Rakesh Ghimire, 2 and pratik lamichhane 1 efficacy and safety of paromomycin for visceral leishmaniasis: a systematic review," *Journal of Tropical Medicine*, vol. 2021, Article ID 8629039, 9 pages, 2021.
- [18] C. Li, L. Su, H. Liao, C. Zhang, and S Xiao, "Modeling of rapid evaluation for seismic stability of soil slope by finite element limit analysis," *Computers and Geotechnics*, vol. 133, no. 9, Article ID 104074, 2021.
- [19] P. Wei and L. Wei, "Calculation and analysis of seismic response dynamics of gravity dam," *IOP Conference Series: Materials Science and Engineering*, vol. 780, no. 6, Article ID 062013, 2020.

Research Article

Differential Privacy via Haar Wavelet Transform and Gaussian Mechanism for Range Query

Dong Chen , **Yanjuan Li**, **Jiaquan Chen**, **Hongbo Bi**, and **Xiajun Ding** 

College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China

Correspondence should be addressed to Xiajun Ding; 37050@qzc.edu.cn

Received 13 May 2022; Revised 3 July 2022; Accepted 22 August 2022; Published 12 September 2022

Academic Editor: Lorenzo Putzu

Copyright © 2022 Dong Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Range query is the hot topic of the privacy-preserving data publishing. To preserve privacy, the large range query means more accumulate noise will be injected into the input data. This study presents a research on differential privacy for range query via Haar wavelet transform and Gaussian mechanism. First, the noise injected into the input data via Laplace mechanism is analyzed, and we conclude that it is difficult to judge the level of privacy protection based on the Haar wavelet transform and Laplace mechanism for range query because the sum of independent random Laplace variables is not a variable of a Laplace distribution. Second, the method of injecting noise into Haar wavelet coefficients via Gaussian mechanism is proposed in this study. Finally, the maximum variance for any range query under the framework of Haar wavelet transform and Gaussian mechanism is given. The analysis shows that using Haar wavelet transform and Gaussian mechanism, we can preserve the differential privacy for each input data and any range query, and the variance of noise is far less than that just using the Gaussian mechanism. In an experimental study on the dataset age extracted from IPUM's census data of the United States, we confirm that the proposed mechanism has much smaller maximum variance of noises than the Gaussian mechanism for range-count queries.

1. Introduction

Over the past ten years, differential privacy has become one of the important methods in the area of privacy-preserving for statistical databases. Differential privacy is a promising scheme for publishing statistical query results of sensitive data, which has a strong privacy guarantee for opponents with arbitrary background knowledge [1–6]. The strong privacy guarantee of differential privacy ensures that any individuals in the data set will not significantly affect the analysis results of the data set. At present, three basic mechanisms are widely used to ensure differential privacy: Laplace mechanism, Gaussian mechanism, and exponential mechanism. Laplacian and Gaussian mechanisms are applicable to numerical queries, and exponential mechanisms are applicable to non-numeric queries [7–9]. Recently, differential privacy is adopted on many research field, such as social network publishing [10–12], crowdsourced data publication [13, 14], and genomic privacy [15–17].

Along with a long-range query scope, the accumulation of noise in the range query answered for privacy preserving

can affect the usability of the released data [18, 19]. To reduce the accumulation of noise, the method of hierarchical decompositions is usually employed [20]. Zhang et al. proposed a differentially private algorithm for hierarchical decompositions and named it as PrivTree. This histogram construction algorithm eliminates the dependency on a predefined limit parameter. The privacy-preserving range query is adopted in the field of Internet of Things (IoT) in recent years [21–23]. Cai et al. studied the transaction approximate range counting problem of large IoT data. They proposed a sampling-based method to generate approximate counting results. For privacy reasons, these results will be further disturbed and then published. It is theoretically proved that this result achieves unbiasedness, bounded variance and enhances privacy guarantee under differential privacy. Mahdikhani et al. proposed a communication efficient privacy protection range query in the fog-enhanced Internet of things. The feature of this scheme is that it adopts the Paillier homomorphic cryptosystem and the ingenious bloom filter data structure to achieve better privacy and higher count aggregation efficiency in the range query

scenario of protecting privacy. Histogram is a representative and popular tool for data publishing and visualization tasks. Nowadays, protecting private data and preventing the leakage of sensitive information have become one of the main challenges faced by histogram [24–26]. Histogram is the result of a group of counting queries. It is the core statistical tool for reporting data distribution. In fact, it is regarded as the basic method of many other statistical analyses, such as range query [27]. The advantage of histogram representation is that it limits the sensitivity to noise. For example, when histograms are used to support range or count queries, adding or deleting a single record will affect at most one box. Therefore, the sensitivity of range or count query on the histogram is equal to 1, and the amount of additional noise per box will be relatively small [28]. For the differential privacy of long-range queries on the histogram, the accumulation of noise is a key issue that needs to be focused.

Discrete wavelet transform (DWT) is an important technology in signal and image processing [29–31]. Lifting scheme, also called second generation wavelet, has many advantages comparing with the first generation wavelet, such as in-place computation, integer-to-integer transforms, and speed [32–34]. Wavelet-based privacy preserving is studied in recent years [35–37]. Xiao et al. propose the differential privacy via Haar wavelet transform. They introduce a data publishing technique named Privelet. Privelet not only ensures ϵ -differential privacy but also provides accurate results for range query by injecting less noise into wavelet coefficients. The mechanism that can be used to build the differential privacy in Privelet is Laplace distribution. The Laplace mechanism, which is used to guarantee differential privacy in Privelet, maybe not a good choice for building the privacy-preserving system based on discrete wavelet. The reason is that the Laplace noise does not have the property of additivity. That is, the sum of two Laplace distributions is not a Laplace distribution. That means we cannot obtain an analyzable noise distribution by wavelet reconstruction where the Laplace noise is injected into the wavelet coefficients.

The Gaussian mechanism for differential privacy is proposed by Dwork [38, 39]. The Gaussian noise can be used in the structuring of hierarchical decompositions, such as wavelet transforms. The property of additivity of Gaussian noise is very important for the reconstruction of noise data. On the one hand, additivity can ensure that the reconstructed noise is still Gaussian noise; on the other hand, some noise can be eliminated during reconstruction.

In view of the above analysis, we will do some research on differential privacy via Gaussian mechanism and lifting scheme of Haar wavelet transform for range query in this study. In summary, the main contributions of this work are as follows:

- (1) Differential privacy using lifting Haar wavelet transform and Laplace mechanism is analyzed in this study. The distribution of noise injected into the input data via wavelet reconstruction is discussed and we conclude that they are not noise of Laplace distribution.
- (2) Differential privacy based on lifting Haar wavelet transform and Gaussian mechanism is constructed in this study. For range query, our analysis shows

that the noise actually added into a certain range of original data is much less than the sum of noise at each data for the proposed mechanism.

- (3) Differential privacy for range query via lifting Haar wavelet and Gaussian mechanism is discussed. We give an algorithm to compute the maximum variance of any range query for any given parameter l (suppose the length of input data is 2^l). Moreover, we give a coarse estimation of the maximum variance of range query using a function expression. Finally, we give an experimental study using the proposed mechanism, and the results show the proposed mechanism has a much smaller maximum variance of noise than the Gaussian mechanism for range query.

The remainder of the study is organized as follows: Section 2 introduces the fundamental definitions and theorems about the differential privacy and its implement mechanism. Section 3 gives the theorems for how to inject Gaussian noise into the Haar wavelet coefficients. Section 4 analysis the noise of range query under the framework of Gaussian mechanism and Haar wavelet. First, the computing method for the variance of range query is given. Second, the algorithm of computing maximum variance for any range query is introduced, and how to get the interval of the range query when obtaining the maximum variance is introduced in detail. Finally, the coarse estimation of the maximum variance of range query is given as a function expression. Section 5 introduces the experimental verification of the computing of maximum variance for the range query based on Gaussian mechanism and lifting Haar wavelet. Conclusions is given in Section 6.

2. Preliminaries

In this section, the fundamental definitions and theorems about the differential privacy and its implement mechanism are introduced first. Furthermore, the method of injecting Laplace noise into the Haar wavelet coefficients is given. They are the basis of the other sections.

2.1. Differential Privacy

Definition 1 ((ϵ, δ) -Differential privacy [38, 39]). A randomized mechanism M with domain $\mathbb{N}^{|x|} \rightarrow \mathbb{R}^d$ is (ϵ, δ) -differential privacy if for all $S \subseteq \text{Range}(M)$ and for all $x, y \in \mathbb{N}^{|x|}$ such that $\|x - y\|_1 \leq 1$.

$$\Pr [M(x) \in S] \leq \exp(\epsilon) \Pr [M(y) \in S] + \delta, \quad (1)$$

where the symbol $\|x\|_1$ denotes the ℓ_1 -norm of a database x , $\|x\|_1 = \sum |x_i|$, and $\|x - y\|_1$ denotes the ℓ_1 -distance between two databases x and y .

2.2. Laplace Mechanism

Definition 2 (ℓ_1 -sensitivity [39]). Let $\mathbb{N}^{|x|} \rightarrow \mathbb{R}^d$ be an arbitrary d -dimensional function, then define the ℓ_1 -sensitivity of function f as follows:

$$\Delta_1 f = \max_{x, y \in \mathbb{N}^{|x|}} \|f(x) - f(y)\|_1, \quad (2)$$

$$\|x - y\|_1 = 1$$

Definition 3. (Laplace distribution, $Lap(\lambda)$). The Laplace distribution with mean zero and scale λ is the distribution with probability density function:

$$Lap(x|\lambda) = \left(\frac{1}{2\lambda}\right) \exp\left(-\frac{|x|}{\lambda}\right). \quad (3)$$

In Definition 3, the variance of this distribution is $\sigma^2 = 2\lambda^2$. We write $Lap(\lambda)$ to denote the Laplace distribution with mean zero and scale λ in this study.

Theorem 1 (Laplace mechanism [39]). *Let f is a function with ℓ_1 -sensitivity, the Laplace mechanism, which adds independently random drawn noise distributed as $Lap(\Delta_1 f/\epsilon)$ into each of the d components of the output, preserves $(\epsilon, 0)$ -differential privacy.*

Remark 1. Throughout the study, we use the term “noise” to refer to a random variable with a zero mean.

2.3. Gaussian Mechanism

Definition 4 (ℓ_2 -sensitivity [39]). Let $f: \mathbb{N}^{|x|} \rightarrow \mathbb{R}^d$ be an arbitrary d -dimensional function, then define the ℓ_2 -sensitivity of function f as follows:

$$\Delta_2 f = \max_{x, y \in \mathbb{N}^{|x|}} \|f(x) - f(y)\|_2, \quad (4)$$

$$\|x - y\|_2 = 1$$

where the symbol $\|x\|_2$ denotes the ℓ_2 -norm of a database x , $\|x\|_2 = \sum x_i^2$, and $\|f(x) - f(y)\|_2$ denotes the ℓ_2 -distance between $f(x)$ and $f(y)$.

Definition 5 (Gaussian distribution, $Gauss(\sigma^2)$). The Gaussian distribution with mean zero and variance σ^2 is the distribution with probability density function:

$$Gauss(x|\sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)\exp(-x^2/(2\sigma^2))}. \quad (5)$$

In Definition 5, the variance of this distribution is σ^2 . We write $Gauss(\sigma^2)$ to denote the Gaussian distribution with mean zero and variance σ^2 .

Theorem 2 (Gaussian mechanism [38, 39]). *Let f be a function with ℓ_2 -sensitivity and let $\epsilon \in (0, 1)$ be arbitrary. For $\sigma \geq \Delta_2 f \cdot \sqrt{2 \ln(1.25/\delta)}/\epsilon$, the Gaussian mechanism, which adds independently drawn random noise distributed as $Gauss(\sigma^2)$ into each of the d components of the output, ensures (ϵ, δ) -differential privacy.*

In Theorem 2, to ensure (ϵ, δ) -differential privacy, we can inject the Gaussian noise with $\sigma^2 = 2 \ln(1.25/\delta) \cdot (\Delta_2 f/\epsilon)^2$ into the input data directly.

2.4. Injecting Noise into the Input Data via Lifting Haar Wavelet

2.4.1. Lifting Scheme of Haar Wavelet. The lifting scheme of Haar wavelet transform is shown in Figure 1. In Figure 1, $x(z)$ is the input data, $x_o(z)$ and $x_e(z)$ denote the odd indexed samples and even indexed samples, respectively. $a(z)$ and $d(z)$ are the approximate coefficients and detail coefficients, respectively. For lifting scheme of Haar wavelet, we have $p(z) = -1$ and $u(z) = 1/2$.

In Figure 1, we have

$$\begin{aligned} x(z) &= \sum_{i=0}^{n-1} x_i \cdot z^{-i} \\ &= \sum x_{2k} \cdot (z^2)^{-k} + z^{-1} \cdot \sum x_{2k+1} \cdot (z^2)^{-k} \\ &= x_e(z^2) + z^{-1} x_o(z^2). \end{aligned} \quad (6)$$

Therefore, the approximate coefficients $a(z)$ and detail coefficients $d(z)$ can be given as follows:

$$a(z) = \frac{1}{2} (x_e(z) + x_o(z)), \quad (7)$$

$$d(z) = \frac{1}{2} (x_e(z) - x_o(z)). \quad (8)$$

In Figure 1, the lifting structure has the reconstruction property, that is

$$x'_o(z) = x_o(z), x'_e(z) = x_e(z), \hat{x}(z) = x(z). \quad (9)$$

Figure 1 shows one-level decomposition and reconstruction via lifting Haar wavelet transform. The wavelet transform usually consists of many decomposition levels. We can apply the same procedure to the approximate coefficients $a(z)$ to get the multilevel Haar wavelet decomposition, as shown in Figure 2.

In Figure 2, the top decomposition level is 3 ($l=3$), $c_{3,0}$ is the approximate coefficient, $c_{k,i}$ ($i \neq 0$) denotes the i th wavelet coefficient in k th decomposition level, and x_m ($m \in [0, 7]$) denotes the input data. In Figure 2, we observe that the number of wavelet coefficients in k th decomposition level is 2^{l-k} .

In Figure 2, given the Haar wavelet coefficients, any entry x_m can be easily reconstructed as follows:

$$x_m = c_{l,0} + \sum_{c_{k,i} \in C \setminus \{c_{l,0}\}} (c_{k,i} \cdot g_{k,i}), \quad (10)$$

where $c_{l,0}$ is the approximate coefficient, $c_{k,i}$ ($i \neq 0$) denotes the i th wavelet coefficient in k th decomposition level, and $g_{k,i}$ equals 1 (-1) if x_m is in the left (right) subtree of $c_{k,i}$, equals 0 if x_m is not in any subtree of $c_{k,i}$. For example,

$$x_1 = 3 = c_{3,0} + c_{3,1} + c_{2,1} - c_{1,1}. \quad (11)$$

In Figure 2, if we inject the noise into the approximate coefficients and detail coefficients, then we can obtain the reconstruction data with noise.

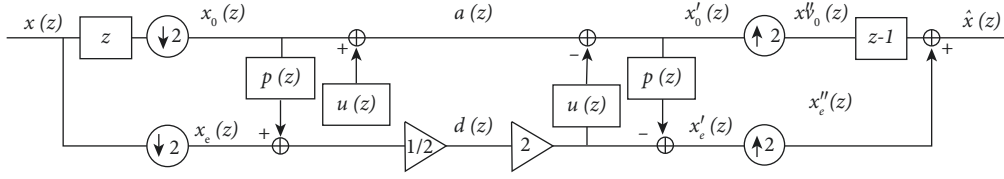


FIGURE 1: Lifting scheme of Haar wavelet transform.

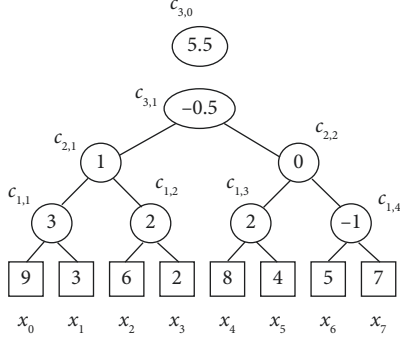


FIGURE 2: Multilevel lifting Haar wavelet transform.

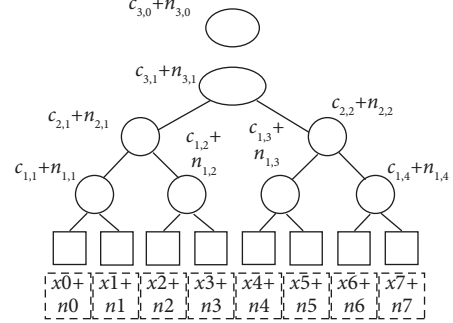


FIGURE 3: Getting input data with noise.

2.4.2. Injecting Noise into Haar Wavelet Coefficients. For the noise injected into Haar wavelet coefficients, the tree structure of noise can be obtained by changing the symbol “ c ” and “ x ” to n in Figure 2 because they use the same decomposition of multilevel lifting Haar wavelet transform.

Referring to equation (10), the noise n_m that injected into data x_m can be given as follows:

$$n_m = n_{l,0} + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} (n_{k,i} \cdot g_{k,i}), \quad (12)$$

where $n_{l,0}$ is the noise injected into approximate coefficient, $n_{k,i}$ ($i \neq 0$) denotes the noise injected into the i th wavelet coefficient in k th decomposition level, and $g_{k,i}$ equals 1 (−1) if n_m is in the left (right) subtree of $n_{k,i}$ and equals 0 if n_m is not in any subtree of $n_{k,i}$.

The range sum of these noise has a special property; that is, some subnoise items can be eliminated when computing some sum of range count. For example, referring to Figure 2 and equation (12), we have

$$\sum_{i=0}^{2^l-1} n_i = \sum_{i=0}^7 n_i = 8 \cdot n_{3,0}. \quad (13)$$

In the above equation, the other subnoise items except $n_{3,0}$ have been eliminated. This gives us the inspiration to apply this property to range query for differential privacy.

2.4.3. Getting Input Data with Noise. Based on the above two sections, we reconstruct the input data with noise by using the multilevel lifting Haar wavelet transform. Considering Figure 2, the input data with noise is shown in Figure 3.

In Figure 3, x_m ($m \in [0, 7]$) denotes the input data reconstructed, n_m is the noise injected into data x_m . $x_m + n_m$ denotes the input data with noise. Referring to equations (10) and (12), we have

$$\begin{aligned} x_m + n_m &= c_{l,0} + \sum_{c_{k,i} \in C \setminus \{c_{l,0}\}} (c_{k,i} \cdot g_{k,i}) + n_{l,0} + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} (n_{k,i} \cdot g_{k,i}) \\ &= (c_{l,0} + n_{l,0}) + \sum_{\substack{c_{k,i} \in C \setminus \{c_{l,0}\} \\ n_{k,i} \in N \setminus \{n_{l,0}\}}} ((c_{k,i} + n_{k,i}) \cdot g_{k,i}), \end{aligned} \quad (14)$$

where the meanings of the symbols $c_{l,0}$, $n_{l,0}$, $c_{k,i}$, $n_{k,i}$, and $g_{k,i}$ are as stated before.

Based on the analysis of above, we conclude that the input data with noised can be obtained by injecting the noise, such as Laplace noise or Gaussian noise, into the approximate and detail coefficients. Moreover, the noise injected into each input data is the sum of the noise injected into approximate and detail coefficients.

2.4.4. Injecting Noise via Haar Wavelet and Laplace Mechanism. In equation (12), we set $n_{k,i}$ as the noise with the Laplace distribution, as given in Definition 3. We have

$$n_{k,i} \sim \text{Lap}\left(\frac{\lambda}{2^k}\right), \quad (15)$$

where λ is the scale parameter of Laplace distribution and k denotes the k th decomposition level of lifting Haar wavelet transform.

According to equations (12) and (15), there is

$$n_m \sim \text{Lap}\left(\frac{\lambda}{2^l}\right) + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} \left(\text{Lap}\left(\frac{\lambda}{2^k}\right) \cdot g_{k,i}\right), \quad (16)$$

where the symbols of n_m , $n_{k,i}$, $n_{l,0}$, and $g_{k,i}$ are the same as those in equation (12).

Using equations (12) and (16), we can describe the Laplace noise injected into Haar wavelet coefficients, as listed in Table 1.

TABLE 1: Laplace noise injected into Haar wavelet coefficients.

$n_{0=}$	$n_{3,0}$	$+n_{3,1}$	$+n_{2,1}$	$+n_{1,1}$
$n_{1=}$	$n_{3,0}$	$+n_{3,1}$	$+n_{2,1}$	$-n_{1,1}$
$n_{2=}$	$n_{3,0}$	$+n_{3,1}$	$-n_{2,1}$	$+n_{1,2}$
$n_{3=}$	$n_{3,0}$	$+n_{3,1}$	$-n_{2,1}$	$-n_{1,2}$
$n_{4=}$	$n_{3,0}$	$-n_{3,1}$	$+n_{2,2}$	$+n_{1,3}$
$n_{5=}$	$n_{3,0}$	$-n_{3,1}$	$+n_{2,2}$	$-n_{1,3}$
$n_{6=}$	$n_{3,0}$	$-n_{3,1}$	$-n_{2,2}$	$+n_{1,4}$
$n_{7=}$	$n_{3,0}$	$-n_{3,1}$	$-n_{2,2}$	$-n_{1,4}$
$n_{k,i} \sim$	Lap ($\lambda/2^3$)	Lap ($\lambda/2^3$)	Lap ($\lambda/2^2$)	Lap ($\lambda/2^1$)

According to Table 1, letting the range of the range query is n_0 to n_7 , we have

$$\sum_{i=0}^{2^l-1} n_i = \sum_{i=0}^7 n_i = 8 \cdot n_{3,0} \sim \text{Lap}\left(8 \cdot \frac{\lambda}{2^3}\right) = \text{Lap}(\lambda). \quad (17)$$

That means the sum of all noise injected into the input data is a noise with Laplace distribution with mean zero and scale λ .

Letting $\lambda = \Delta_1 f / \epsilon$, then $\epsilon = \Delta_1 f / \lambda$, according to Theorem 1, we conclude the $(\epsilon, 0)$ -differential privacy is preserved for the range query from n_0 to n_7 using Laplace mechanism.

According to Table 1, letting the range of the range query is n_1 to n_3 , we have

$$n_1 + n_2 + n_3 = 3 \cdot n_{3,0} + 3 \cdot n_{3,1} - n_{2,1} - n_{1,1}. \quad (18)$$

Therefore,

$$\begin{aligned} n_1 + n_2 + n_3 &\sim \text{Lap}\left(3 \cdot \frac{\lambda}{2^3}\right) + \text{Lap}\left(3 \cdot \frac{\lambda}{2^3}\right) \\ &- \text{Lap}\left(\frac{\lambda}{2^2}\right) - \text{Lap}\left(\frac{\lambda}{2^1}\right). \end{aligned} \quad (19)$$

As we know, the sum of independent random Laplace variables is not a variable of Laplace distribution, so the composite noise of range query of $n_1 + n_2 + n_3$ that injected into input data $x_1 + x_2 + x_3$ is not a noise with Laplace distribution. Therefore, we conclude that it is difficult to judge the level of differential privacy protection based on the Haar wavelet transform and Laplace mechanism.

To solve this problem, we consider adopting the Gaussian mechanism for the differential privacy via Haar wavelet transform in the next section.

3. Injecting Noise into Haar Wavelet Coefficients via Gaussian Mechanism

To inject Gaussian noise into Haar wavelet coefficients in Figure 3, we can set $n_{k,i}$ as the noise with the Gaussian distribution, as given in Definition 5.

Let

$$n_{k,i} \sim \text{Gauss}\left(\frac{3\sigma^2}{4^k}\right), \quad (20)$$

where $3\sigma^2/4^k$ is the variance of Gaussian distribution.

TABLE 2: Gaussian noise injected into wavelet coefficients.

$n_{0=}$	$n_{3,0}$	$+n_{3,1}$	$+n_{2,1}$	$+n_{1,1}$
$n_{1=}$	$n_{3,0}$	$+n_{3,1}$	$+n_{2,1}$	$-n_{1,1}$
$n_{2=}$	$n_{3,0}$	$+n_{3,1}$	$-n_{2,1}$	$+n_{1,2}$
$n_{3=}$	$n_{3,0}$	$+n_{3,1}$	$-n_{2,1}$	$-n_{1,2}$
$n_{4=}$	$n_{3,0}$	$-n_{3,1}$	$+n_{2,2}$	$+n_{1,3}$
$n_{5=}$	$n_{3,0}$	$-n_{3,1}$	$+n_{2,2}$	$-n_{1,3}$
$n_{6=}$	$n_{3,0}$	$-n_{3,1}$	$-n_{2,2}$	$+n_{1,4}$
$n_{7=}$	$n_{3,0}$	$-n_{3,1}$	$-n_{2,2}$	$-n_{1,4}$
$n_{k,i} \sim$	Gauss ($3\sigma^2/4^3$)	Gauss ($3\sigma^2/4^3$)	Gauss ($3\sigma^2/4^2$)	Gauss ($3\sigma^2/4^1$)

According to equations (12) and (20), we have

$$n_m \sim \text{Gauss}\left(\frac{3\sigma^2}{4^l}\right) + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} \left(\text{Gauss}\left(\frac{3\sigma^2}{4^k}\right) \cdot g_{k,i} \right), \quad (21)$$

where the symbols of n_m , $n_{k,i}$, $n_{l,0}$, and $g_{k,i}$ are same as those in equation (12).

Using equations (20) and (21), we can describe the Gaussian noise injected into Haar wavelet coefficients, as listed in Table 2.

Theorem 3. Suppose that X_1 and X_2 are independent random variables, and X_i has Gaussian distribution with mean zero and variance σ_i^2 for $i \in \{1, 2\}$. Then, $X_1 \pm X_2$ is Gaussian distribution with mean zero and variance $\sigma_1^2 + \sigma_2^2$; kX_1 is Gaussian distribution with mean zero and variance $(k\sigma_1)^2$.

The proof of Theorem 3 will not be given because it is a basic property of Gaussian distribution.

According to Table 2 and Theorem 3, there is

$$\sum_{i=0}^{2^l-1} n_i = \sum_{i=0}^7 n_i = 8 \cdot n_{3,0} = \text{Gauss}\left(8^2 \cdot \frac{3\sigma^2}{4^3}\right) = \text{Gauss}(3\sigma^2). \quad (22)$$

That means the sum of all noise injected into the input data is a noise with Gaussian distribution. We analyze the distribution of the noise injected into each input data as follows.

Theorem 4. Injecting Gaussian noise with variance $\sigma_k^2 = 3\sigma^2/4^k$ into the Haar wavelet coefficients in the k th decomposition level (the maximum decomposition level is l , as shown in Figure 3), the noise injected into each input data via Haar wavelet reconstruction is Gaussian noise with variance $(1 + 2/4^l) \sigma^2$.

Proof. According to Definition 5, Theorem 3, Table 2, and equation (21), we have

$$\begin{aligned} n_m &\sim \text{Gauss}\left(\frac{3\sigma^2}{4^l}\right) \pm \sum_{i=1}^l \text{Gauss}\left(\frac{3\sigma^2}{4^i}\right), \\ &\sim \text{Gauss}\left(\left(\frac{1}{4^l} + \sum_{i=1}^l \frac{1}{4^i}\right) \cdot 3\sigma^2\right), \\ &\sim \text{Gauss}\left(\left(1 + \frac{2}{4^l}\right) \cdot \sigma^2\right). \end{aligned} \quad (23)$$

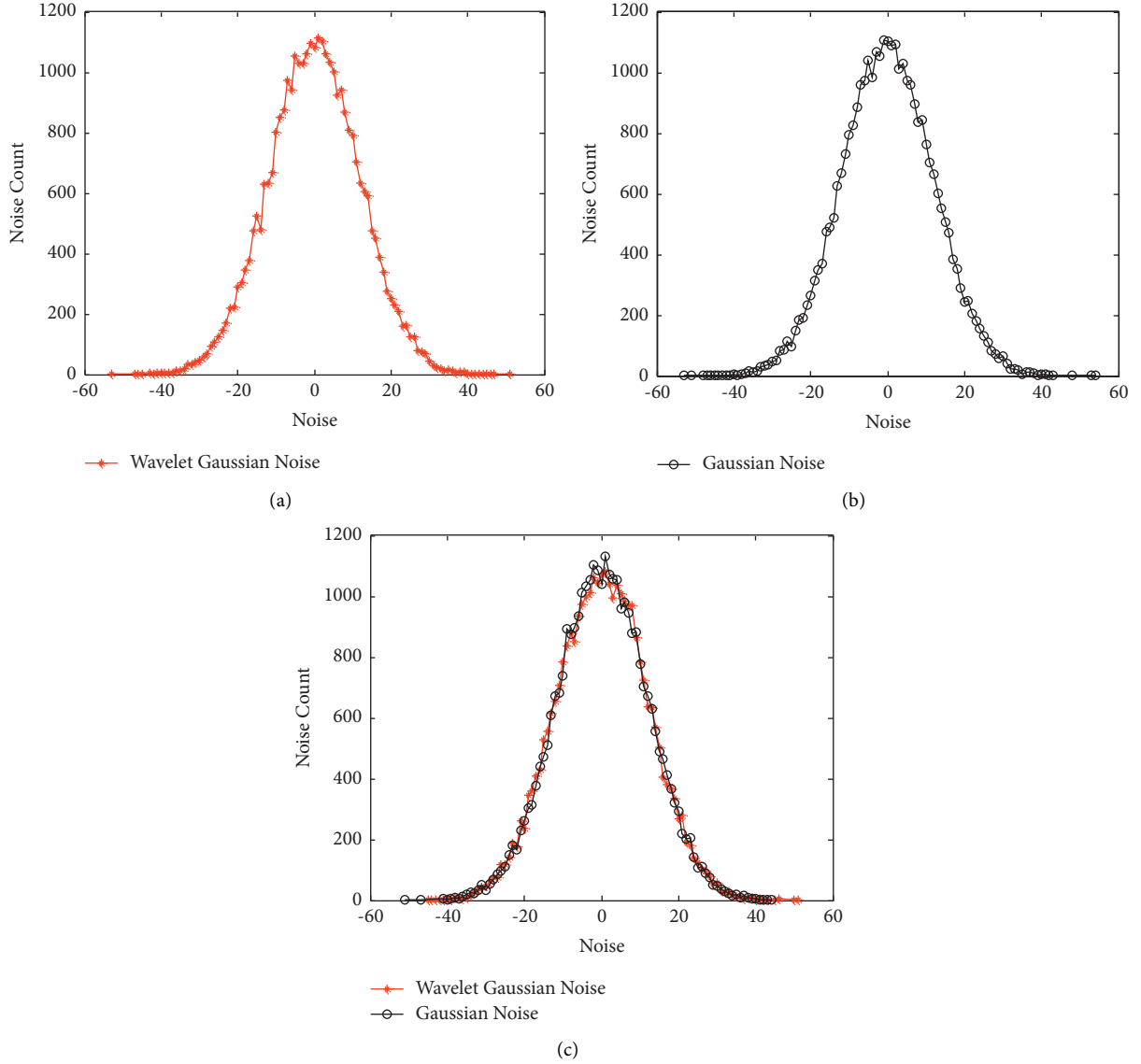


FIGURE 4: Comparison between injecting Gaussian noise into wavelet coefficients using Theorem 4 and injecting Gaussian noise into input data directly.

The proof is completed.

Using Theorems 2 and 4, we can obtain the (ϵ, δ) -differential privacy with variance $(1 + 2/4^l) \sigma^2$ for each input data under the framework of Gaussian mechanism via Haar wavelet transform. \square

Theorem 5 (Differential privacy using Gaussian mechanism and Haar wavelet). *Let f be a function with ℓ_2 -sensitivity, let $\epsilon \in (0, 1)$ be arbitrary, and let $\sigma = \Delta_2 f \cdot \sqrt{2 \ln(1.25/\delta)}/\epsilon$. The mechanism adopting Gaussian and Haar wavelet, which adds Gaussian noise with variance $\sigma_k^2 = 3\sigma^2/4^k$ into the Haar wavelet coefficients in the k th decomposition level (Figure 3), ensures (ϵ, δ) -differential privacy.*

Proof. In Theorem 4, if the Gaussian noise with variance $\sigma_k^2 = 3\sigma^2/4^k$ is injected into the wavelet coefficients, the reconstructed data will be the one with the Gaussian noise

with variance $\sqrt{1 + 2/4^l} \sigma$. According to Theorem 2, the condition of $\sqrt{1 + 2/4^l} \sigma \geq \Delta_2 f \cdot \sqrt{2 \ln(1.25/\delta)}/\epsilon$ should be satisfied. Therefore, letting $\sigma = \Delta_2 f \cdot \sqrt{2 \ln(1.25/\delta)}/\epsilon$, the condition is met and the proof is completed.

We simulate the process of injecting Gaussian noise with $\sigma = 12$ into wavelet coefficients with 15-level decomposition using Theorem 4 and injecting Gaussian noise into input data directly and draw the noise-count figures as follows.

Figure 4(a) shows the count of the noise injected into input data by injecting the Gaussian noise with $\sigma = 12$ into Haar wavelet coefficients with 15-level decomposition using equation (20); Figure 4(b) denotes the noise count by injecting Gaussian noise with mean zero and variance $(1 + 2/4^l) \sigma^2$ into the input data directly. In Figure 4(c), we draw the two curves together and we find that they are almost overlapped. Figure 4(c) shows that the method injecting Gaussian noise into Haar wavelet coefficients has the same

level of differential privacy protection as the method injecting Gaussian noise into the input data directly. In this study, we will focus on the application of range query.

According to Theorems 3–5, we find that the distribution of noise for the range query using Gaussian mechanism and Haar wavelet is a Gaussian distribution. Therefore, we can calculate the variance of noise easily, for example, as listed in Table 2, we have

$$\begin{aligned}
n_1 + n_2 + n_3 &= n_{3,0} + n_{3,1} + n_{2,1} - n_{1,1} \\
&+ n_{3,0} + n_{3,1} - n_{2,1} + n_{1,2} \\
&+ n_{3,0} + n_{3,1} - n_{2,1} - n_{1,2} \\
&= 3n_{3,0} + 3n_{3,1} - n_{2,1} - n_{1,1}, \\
&\sim 3\text{Gauss}\left(\frac{3\sigma^2}{4^3}\right) + 3\text{Gauss}\left(\frac{4\sigma^2}{4^3}\right) \\
&- \text{Gauss}\left(\frac{3\sigma^2}{4^2}\right) - \text{Gauss}\left(\frac{3\sigma^2}{4^1}\right), \\
&\sim \text{Gauss}\left(3^2 \cdot \frac{3\sigma^2}{4^3} + 3^2 \cdot \frac{3\sigma^2}{4^3} + \frac{3\sigma^2}{4^2} + \frac{3\sigma^2}{4^1}\right), \\
&\sim \text{Gauss}\left(\frac{57}{32}\sigma^2\right).
\end{aligned} \tag{24}$$

From this example, we find that some noises (such as $n_{2,1}$ and $-n_{2,1}$, $n_{1,2}$ and $-n_{1,2}$) are eliminated by the operation of addition. According to Theorem 4, the variance of noise injected into each input data should be $(1 + 2/4^3)\sigma^2$ for $l=3$. The total noise variance is $3 * (1 + 2/4^3)\sigma^2 = (99/32) * \sigma^2$. Compared with equation (24), we conclude that, for the range query, the noise actually added into a certain range of the original data is much less than the sum of the noise at each data. Therefore, it is a very important property for Gaussian mechanism to be used on range query. \square

4. Noise of Range Query under the Framework of Haar Wavelet and Gaussian Mechanism

In this section, we discuss how to compute the noise of the range query under the framework of Haar wavelet transform and Gaussian mechanism. First, we give the computing method for the variance of range query in detail. Second, the interval of the range query when obtaining the maximum variance is introduced. Third, to speed up the computing of maximum variance, we observe the results of the range-count interval when getting the maximum variance and give a speed computing method. Finally, we give a coarse estimation of the maximum variance of range query as a function expression.

4.1. Computing Method for the Variance of Range Query. In Figure 3, we choose the Gaussian noise and inject them into the approximate coefficient and each wavelet coefficient.

TABLE 3: Noise variance and decomposition level k .

k	Variance of noise	Number of wavelet coefficients
3	$3\sigma^2/4^3$	2^{3-3}
2	$3\sigma^2/4^2$	2^{3-2}
1	$3\sigma^2/4^1$	2^{3-1}

The variance of Gaussian noise injected into approximate coefficient is $3\sigma^2/4^3$. The variance of Gaussian noise injected into each wavelet coefficient is $3\sigma^2/4^k$ for level k ($k \in [1, 3]$). The relationship between decomposition level and variance of noise is listed in Table 3.

The noise-sum of range query via Haar wavelet transform for interval S can be given by the following equation (Figure 3):

$$n_{sum} = |S| \cdot n_{l,0} + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} (n_{k,i} \cdot (\alpha(n_{k,i}) - \beta(n_{k,i}))), \tag{25}$$

where S is the interval of any range query, $n_{k,i}$ presents the i th noise injected into wavelet coefficient in k th decomposition level, $\alpha(n_{k,i})$ denotes the number of left leaves in the left subtree of $n_{k,i}$ that are contained in S , and $\beta(n_{k,i})$ denotes the number of right leaves in the right subtree of $n_{k,i}$ that are contained in S (Figure 3).

Now we analyze the noise variance of range query. According to equation (20), we know that the noise injected into approximate coefficient is $n_{l,0}$ and its variance is $3\sigma^2/4^l$; the noise injected into wavelet coefficient is $n_{k,i}$ and its variance is $3\sigma^2/4^k$. Therefore, according to Theorem 3, we can compute the noise-variance of range query by replacing $n_{l,0}$ and $n_{k,i}$ with $3\sigma^2/4^l$ and $3\sigma^2/4^k$ in equation (25), respectively.

$$\begin{aligned}
\sigma_{sum}^2 &= |S|^2 \cdot \frac{3\sigma^2}{4^l} + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} \left(\frac{3\sigma^2}{4^k} (\alpha(n_{k,i}) - \beta(n_{k,i}))^2 \right) \\
&= \left(|S|^2 \cdot \frac{1}{4^l} + \sum_{n_{k,i} \in N \setminus \{n_{l,0}\}} \left(\frac{1}{4^k} (\alpha(n_{k,i}) - \beta(n_{k,i}))^2 \right) \right) \cdot 3\sigma^2.
\end{aligned} \tag{26}$$

To compute the value of σ_{sum}^2 , we need to calculate the values of $\alpha(n_{k,i})$ and $\beta(n_{k,i})$ firstly. In Figure 3, the length of interval of leaves in the subtree of $n_{k,i}$ is 2^k . The left point of this interval has the subscript $(i-1) \cdot 2^k$ and the right point of this interval has the subscript $i \cdot 2^k - 1$. Therefore, the subtree of $n_{k,i}$ has the subscript interval of leaves.

$$S_{n_{k,i}} = [(i-1) \cdot 2^k, i \cdot 2^k - 1]. \tag{27}$$

For example, the wavelet coefficient $n_{2,2}$ in Figure 3 has the subscript interval of leaves [4, 7].

According to equation (27), we can obtain the left-half interval $[\alpha_L, \alpha_R]$ and right-half interval $[\beta_L, \beta_R]$ of $S_{n_{k,i}}$:

$$\begin{aligned}
[\alpha_L, \alpha_R] &= [(i-1) \cdot 2^k, i \cdot 2^k - 2^{k-1} - 1], \\
[\beta_L, \beta_R] &= [i \cdot 2^k - 2^{k-1}, i \cdot 2^k - 1].
\end{aligned} \tag{28}$$

```

Input: the maximum decomposition level  $l$ 
Output:  $\sigma^2_{sumMax}$ ,  $S_L$  and  $S_R$  (for  $\sigma^2_{sumMax}$ )
(1)  $\sigma^2_{sumMax} = 0$ 
(2) For  $S_L = 0$  to  $2^l - 1$ 
(3)   For  $S_R = S_L$  to  $2^l - 1$ 
(4)      $sum = 0$ 
(5)     For  $k = 1$  to  $l$ 
(6)       For  $i = 1$  to  $2^{l-k}$ 
(7)         Compute  $\alpha_L, \alpha_R, \beta_L, \beta_R$  of  $n_{k,i}$  using equations (38) and (39)
(8)         If  $S_R < \alpha_L$  or  $S_L > \alpha_R$  then  $\alpha(n_{k,i}) = 0$ 
(9)         Elseif  $S_L \geq \alpha_L$  and  $S_R \leq \alpha_R$  then  $\alpha(n_{k,i}) = S_R - S_L + 1$ 
(10)        Elseif  $S_L < \alpha_L$  and  $S_R > \alpha_R$  then  $\alpha(n_{k,i}) = \alpha_R - \alpha_L + 1$ 
(11)        Elseif  $S_L < \alpha_L$  and  $\alpha_L \leq S_R \leq \alpha_R$  then  $\alpha(n_{k,i}) = S_R - \alpha_L + 1$ 
(12)        Elseif  $S_R > \alpha_R$  and  $\alpha_L \leq S_L \leq \alpha_R$  then  $\alpha(n_{k,i}) = \alpha_R - S_L + 1$ 
(13)        End If
(14)        If  $S_R < \beta_L$  or  $S_L > \beta_R$  then  $\beta(n_{k,i}) = 0$ 
(15)        Elseif  $S_L \geq \beta_L$  and  $S_R \leq \beta_R$  then  $\beta(n_{k,i}) = S_R - S_L + 1$ 
(16)        Elseif  $S_L < \beta_L$  and  $S_R > \beta_R$  then  $\beta(n_{k,i}) = \beta_R - \beta_L + 1$ 
(17)        Elseif  $S_L < \beta_L$  and  $\beta_L \leq S_R \leq \beta_R$  then  $\beta(n_{k,i}) = S_R - \beta_L + 1$ 
(18)        Elseif  $S_R > \beta_R$  and  $\beta_L \leq S_L \leq \beta_R$  then  $\beta(n_{k,i}) = \beta_R - S_L + 1$ 
(19)        End If
(20)        Compute  $sum = sum + (1/4^k) (\alpha(n_{k,i}) - \beta(n_{k,i}))^2$ 
(21)      End For
(22)    End For
(23)    Compute  $\sigma^2_{sum}$  using  $sum$  and (26)
(24)    If  $\sigma^2_{sum} > \sigma^2_{sumMax}$ 
(25)       $\sigma^2_{sumMax} = \sigma^2_{sum}$ 
(26)    End IF
(27)  End For
(28) End For
(29) Import  $\sigma^2_{sumMax}$ ,  $S_L$  and  $S_R$ .

```

ALGORITHM 1: Compute σ^2_{sumMax} for fix l ($l \geq 2$).

Therefore, $\alpha(n_{k,i})$ and $\beta(n_{k,i})$ can be given by computing the number of intersection between S and $[\alpha_L, \alpha_R]$, S and $[\beta_L, \beta_R]$, respectively.

$$\begin{aligned} \alpha(n_{k,i}) &= |S \cap [\alpha_L, \alpha_R]|, \\ \beta(n_{k,i}) &= |S \cap [\beta_L, \beta_R]|. \end{aligned} \quad (29)$$

Let $S = [S_L, S_R]$, where S_L and S_R denote the left and right points of the given range query interval, respectively. Therefore, we have

$$\alpha(n_{k,i}) = |[S_L, S_R] \cap [\alpha_L, \alpha_R]|, \quad (30)$$

$$\beta(n_{k,i}) = |[S_L, S_R] \cap [\beta_L, \beta_R]|, \quad (31)$$

where $k \in [1, l]$ and $i \in [1, 2^{l-k}]$ (Figure 3).

4.2. Maximum Variance of Range Query. The aim of this study is to obtain the maximum value of range query for any fixed maximum decomposition level l (the number of input data is 2^l). According to equation (26), we have

$$\sigma^2_{sumMax} = \max \left(|S|^2 \cdot \frac{1}{4^l} + \sum_{n_{k,i} \in N \setminus \{n_{0,0}\}} \left(\frac{1}{4^k} (\alpha(n_{k,i}) - \beta(n_{k,i}))^2 \right) \right) \cdot 3\sigma^2. \quad (32)$$

In equation (32), the given parameter is l . To compute the value of σ^2_{sumMax} , we need to calculate any range query interval $S = [S_L, S_R]$ in all data and obtain the count $\alpha(c_{k,i})$ and $\beta(c_{k,i})$ using equations (30) and (31). We give the pseudocode of computing σ^2_{sumMax} as follows:

Algorithm 1 illustrates the details of the algorithm of computing σ^2_{sumMax} for fix l ($l \geq 2$). Step 1 is the initialization of σ^2_{sumMax} . Steps 2 to 3 are the range loop of S_L and S_R . Step 5 is the loop of the subscript of decomposition level k . Step 6 is the loop of the subscript of the wavelet coefficient in k th decomposition level. Step 7 is the computation of the $\alpha_L, \alpha_R, \beta_L$, and β_R of $n_{k,i}$. Steps 8 to 13 denote the computation of $\alpha(n_{k,i})$. Steps 14 to 19 denote the computation of $\beta(n_{k,i})$. Step 20 is the computation of right part of σ^2_{sum} using equation (26). Step 23 is the computation of σ^2_{sum} using equation (26). Steps 24 to 26 denote the computation of σ^2_{sumMax} using equation (34). Step 29 denotes the output of Algorithm 1.

According to Algorithm 1, we calculate the values of σ^2_{sumMax} , S_L and S_R , as listed in Table 4.

In Table 4, $S_R - S_L + 1$ denotes the length of interval for the σ^2_{sumMax} . It will take a very long time to compute the σ^2_{sumMax} using Algorithm 1 when $l > 14$, so we need to find some method to speed up Algorithm 1.

4.3. Speeding Algorithm for Computing the Maximum Variance of Range Query. Observing the values of S_L and S_R in

TABLE 4: Max-variance of range query via Haar wavelet and Gaussian mechanism (l from 2 to 14).

l	$\sigma^2_{\text{sumMax}} (* \sigma^2)$	S_L	S_R	$S_R - S_L + 1$
2	3.000000	0	3	4
3	3.562500	1	6	6
4	4.265625	1	14	14
5	4.910156	3	28	26
6	5.586914	5	58	54
7	6.248291	11	116	106
8	6.917542	21	234	214
9	7.582901	43	468	426
10	8.250217	85	938	854
11	8.916558	171	1876	1706
12	9.583388	341	3754	3414
13	10.249973	683	7508	6826
14	10.916680	1365	15018	13654

TABLE 5: Statistic results of S_L and S_R .

l	S_L	S_R
2	0	3
3	1	6
4	$1 \times 2 - 1$	14
5	$1 \times 2 + 1$	28
6	$3 \times 2 - 1$	58
7	$5 \times 2 + 1$	116
8	$11 \times 2 - 1$	234
9	$21 \times 2 + 1$	468
10	$43 \times 2 - 1$	938
11	$85 \times 2 + 1$	1876
12	$171 \times 2 - 1$	3754
13	$341 \times 2 + 1$	7508
14	$683 \times 2 - 1$	15018

Table 4, we give some statistical rules to compute them directly in Table 5.

According to Table 5, we can compute σ^2_{sumMax} using the following algorithm:

Algorithm 2 illustrates the details of the algorithm of computing $\sigma^2_{\text{sumMax},l}$ for any l ($l \geq 2$). Step 1 is the initialization of S_L and S_R . Step 2 is the loop of the maximum decomposition level l . Steps 3 to 7 denote the computation of S_L and S_R according to the maximum decomposition level l . Step 10 is the loop of the subscript of decomposition level k . Step 11 is the loop of the subscript of the wavelet coefficient in k th decomposition level. Step 12 is the computation of the α_L , α_R , β_L , and β_R of $n_{k,i}$. Steps 13 to 18 denote the computation of $\alpha(n_{k,i})$. Steps 19 to 24 denote the computation of $\beta(n_{k,i})$. Step 25 is the computation of right part of σ^2_{sum} using equation (26). Step 28 is the computation of σ^2_{sum} using equation (26). Step 30 denotes the output of Algorithm 2 for any l .

S_L and S_R can also be given directly by simplifying the results in Table 5.

If l is an even number,

$$S_L = \frac{2^{l-2} - 1}{3}, S_R = \frac{11 \times 2^{l-2} - 2}{3}. \quad (33)$$

If l is an odd number,

$$S_L = \frac{2^{l-2} + 1}{3}, S_R = \frac{11 \times 2^{l-2} - 4}{3}. \quad (34)$$

Therefore, we give the values of σ^2_{sumMax} , S_L , S_R , and $S_R - S_L + 1$ for l from 2 to 30 in Table 6.

In Table 6, S_L and S_R denote the left and right points of the range query interval when the σ^2_{sumMax} is met. $S_R - S_L + 1$ denotes the length of interval for the σ^2_{sumMax} . In Table 6, we observe that σ^2_{sumMax} will increase about $(2/3) \sigma^2$ if the parameter l increases 1.

4.4. Coarse Estimation of the Maximum Variance of Range Query. In previous sections, the maximum variance of range queries via Gaussian mechanism and Haar wavelet transform is given for any l . But it is obtained using a computer program, not from a function expression. In this section, the coarse estimation of the maximum variance is given in Theorem 6, and it is a function expression with parameters l and σ^2 .

Theorem 6 (Coarse estimation of the maximum variance). *Let N be a set of independent Gaussian noise $n_{k,i} \in N$ with a variance $3\sigma^2/4^k$, which is injected into one-dimensional Haar wavelet coefficients and approximate coefficient (Figure 3). Suppose $l = \log_2|N|$, that means the number of Gaussian noise injected into Haar wavelet*

Input:
Output: $l, \sigma^2_{\text{sumMax}, b}, S_L$ and S_R (for l and $\sigma^2_{\text{sumMax}, i}$)

- (1) $S_L = 0, S_R = 3$ (for $l = 2$)
- (2) For $l = 3$ to 30
- (3) If l is odd then
- (4) $S_L = S_L \times 2 + 1, S_R = S_R \times 2$
- (5) Elseif l is even then
- (6) $S_L = S_L \times 2 - 1, S_R = S_R \times 2 + 2$
- (7) End If
- (8) $\sigma^2_{\text{sumMax}, l} = 0$
- (9) $sum = 0$
- (10) For $k = 1$ to l
- (11) For $i = 1$ to 2^{l-k}
- (12) Compute $\alpha_L, \alpha_R, \beta_L, \beta_R$ of $n_{k, i}$ using equations (38) and (39)
- (13) If $S_R < \alpha_L$ or $S_L > \alpha_R$ then $\alpha(n_{k, i}) = 0$
- (14) Elseif $S_L \geq \alpha_L$ and $S_R \leq \alpha_R$ then $\alpha(n_{k, i}) = S_R - S_L + 1$
- (15) Elseif $S_L < \alpha_L$ and $S_R > \alpha_R$ then $\alpha(n_{k, i}) = \alpha_R - \alpha_L + 1$
- (16) Elseif $S_L < \alpha_L$ and $\alpha_L \leq S_R \leq \alpha_R$ then $\alpha(n_{k, i}) = S_R - \alpha_L + 1$
- (17) Elseif $S_R > \alpha_R$ and $\alpha_L \leq S_L \leq \alpha_R$ then $\alpha(n_{k, i}) = \alpha_R - S_L + 1$
- (18) End If
- (19) If $S_R < \beta_L$ or $S_L > \beta_R$ then $\beta(n_{k, i}) = 0$
- (20) Elseif $S_L \geq \beta_L$ and $S_R \leq \beta_R$ then $\beta(n_{k, i}) = S_R - S_L + 1$
- (21) Elseif $S_L < \beta_L$ and $S_R > \beta_R$ then $\beta(n_{k, i}) = \beta_R - \beta_L + 1$
- (22) Elseif $S_L < \beta_L$ and $\beta_L \leq S_R \leq \beta_R$ then $\beta(n_{k, i}) = S_R - \beta_L + 1$
- (23) Elseif $S_R > \beta_R$ and $\beta_L \leq S_L \leq \beta_R$ then $\beta(n_{k, i}) = \beta_R - S_L + 1$
- (24) End If
- (25) Compute $sum = sum + (1/4^k) (\alpha(n_{k, i}) - \beta(n_{k, i}))^2$
- (26) End For
- (27) End For
- (28) Compute σ^2_{sum} using sum and (26)
- (29) $\sigma^2_{\text{sumMax}, l} = \sigma^2_{\text{sum}}$
- (30) Import $l, \sigma^2_{\text{sumMax}, b}, S_L$ and S_R
- (31) End For.

ALGORITHM 2: Compute $\sigma^2_{\text{sumMax}, l}$ for any l ($l \geq 2$).

coefficients and approximate coefficient is 2^l (the number of input data is also 2^l). Let M be the noisy data reconstructed from $C + N$ (C is the set of one-dimensional Haar wavelet coefficients of the input data, refer to Figure 2). Then, for any range query answered using M , the variance of noise in the answer is at most $((6l + 9)/4)\sigma^2$.

Proof. Referring to Figure 3 and equation (26), we observe that for any noise $n_{k,i}$, if none of the leaves under $n_{k,i}$ is contained in S , then there is $\alpha(n_{k,i}) = \beta(n_{k,i}) = 0$. On the other hand, if all leaves under $n_{k,i}$ are covered by S , then $\alpha(n_{k,i}) = \beta(n_{k,i}) = 2^{k-1}$. Therefore, $\alpha(n_{k,i}) - \beta(n_{k,i}) \neq 0$, if and only if the left or right subtree of $n_{k,i}$ partially intersects S . At any level of the decomposition tree except for the l th level, there exist at most two such noises. At the level l , at most one such noise that letting the condition $\alpha(n_{k,i}) - \beta(n_{k,i}) \neq 0$ be sufficient.

Considering a noise $n_{k,i}$ at level k ($k \in [1, l]$), such that $\alpha(n_{k,i}) - \beta(n_{k,i}) \neq 0$. Since the left (right) subtree of $n_{k,i}$ contains at most 2^{k-1} leaves, we have $\alpha(n_{k,i}), \beta(n_{k,i}) \in [0, 2^{k-1}]$. So, there is $|\alpha(n_{k,i}) - \beta(n_{k,i})| \leq 2^{k-1}$. Therefore, the variance of the range query about the noise $n_{k,i}$ ($k \in [1, l]$) at most is

$$(\alpha(n_{k,i}) - \beta(n_{k,i}))^2 \cdot \left(\frac{3\sigma^2}{4^k}\right) \leq 4^{k-1} \cdot \left(\frac{3\sigma^2}{4^k}\right) = 3\sigma^2/4. \quad (35)$$

On the other hand, the noise in the approximate coefficient ($n_{l,0}$) has a variance at most:

$$(2^l)^2 \cdot \left(\frac{3\sigma^2}{4^l}\right) = 4^l \cdot \left(\frac{3\sigma^2}{4^l}\right) = 3\sigma^2. \quad (36)$$

Therefore, the total variance injected into wavelet coefficients of 1 to $l-1$ level is $2 \cdot (l-1) \cdot 3\sigma^2/4$, and the variance injected into wavelet coefficients of level l is $1 \cdot 3\sigma^2/4$. According to equation (26), the variance of noise at most is

$$3\sigma^2 + 2 \cdot (l-1) \cdot \frac{3\sigma^2}{4} + 1 \cdot \frac{3\sigma^2}{4} = \left(\frac{6l+9}{4}\right)\sigma^2, \quad (37)$$

which completes the proof.

This conclusion in Theorem 6 can also be obtained by observing Table 2. Now, we give the intuitive explanation of Theorem 6.

According to Table 2, we can give a coarse estimation of the maximum variance of range query. In Table 2, we insert a row at the bottom to calculate the maximum variance sum for each column. The new table is shown as follows.

TABLE 6: Max-variance of range query via Haar wavelet and Gaussian mechanism (any l).

l	$\sigma^2_{\text{sumMax}} (* \sigma^2)$	S_L	S_R	$S_R - S_L + 1$
2	3.000000	0	3	4
3	3.562500	1	6	6
4	4.265625	1	14	14
5	4.910156	3	28	26
6	5.586914	5	58	54
7	6.248291	11	116	106
8	6.917542	21	234	214
9	7.582901	43	468	426
10	8.250217	85	938	854
11	8.916558	171	1876	1706
12	9.583388	341	3754	3414
13	10.249973	683	7508	6826
14	10.916680	1365	15018	13654
15	11.583327	2731	30036	27306
16	12.250003	5461	60074	54614
17	12.916665	10923	120148	109226
18	13.583334	21845	240298	218454
19	14.250000	43691	480596	436906
20	14.916667	87381	961194	873814
21	15.583333	174763	1922388	1747626
22	16.250000	349525	3844778	3495254
23	16.916667	699051	7689556	6990506
24	17.583333	1398101	15379114	13981014
25	18.250000	2796203	30758228	27962026
26	18.916667	5592405	61516458	55924054
27	19.583333	11184811	123032916	111848106
28	20.250000	22369621	246065834	223696214
29	20.916667	44739243	492131668	447392426
30	21.583333	89478485	984263338	894784854

In Table 7, each value of the last row is given by calculating the maximum sum of variance of some continued parts for each column according to Theorem 3. For example, the value of column 2 in the last row, $3\sigma^2$, is calculated by range from n_0 to n_7 :

$$8n_{3,0} \sim 8\text{Gauss}\left(\frac{3\sigma^2}{4^3}\right) = \text{Gauss}\left(8^2 \cdot \frac{3\sigma^2}{4^3}\right) = \text{Gauss}(3\sigma^2). \quad (38)$$

The value of column 3 in the last row, $3\sigma^2/4$, is calculated by range from n_0 to n_3 , or from n_4 to n_7 :

$$\pm 4n_{3,1} \sim 4\text{Gauss}\left(\frac{3\sigma^2}{4^3}\right) = \text{Gauss}\left(4^2 \cdot \frac{3\sigma^2}{4^3}\right) = \text{Gauss}\left(\frac{3\sigma^2}{4}\right). \quad (39)$$

The value of column 4 in the last row, $3\sigma^2/2$, is calculated by range from n_2 to n_5 :

$$\begin{aligned} -2n_{2,1} + 2n_{2,2} &\sim 2\text{Gauss}\left(\frac{3\sigma^2}{4^2}\right) + 2\text{Gauss}\left(\frac{3\sigma^2}{4^2}\right), \\ &\sim \text{Gauss}\left(2^2 \cdot \frac{3\sigma^2}{4^2}\right) + \text{Gauss}\left(2^2 \cdot \frac{3\sigma^2}{4^2}\right), \\ &\sim \text{Gauss}\left(2 \cdot 2^2 \cdot \frac{3\sigma^2}{4^2}\right), \\ &\sim \text{Gauss}\left(\frac{3\sigma^2}{2}\right). \end{aligned} \quad (40)$$

The value of column 5 in the last row, $3\sigma^2/2$, is calculated by range from n_1 to n_2 or from n_3 to n_4 or from n_5 to n_6 :

$$\begin{aligned} -n_{1,1} + n_{1,2} &\sim \text{Gauss}(3\sigma^2/4 + 3\sigma^2/4) = \text{Gauss}(3\sigma^2/2) \\ -n_{1,2} + n_{1,3} &\sim \text{Gauss}(3\sigma^2/4 + 3\sigma^2/4) = \text{Gauss}(3\sigma^2/2). \\ -n_{1,3} + n_{1,4} &\sim \text{Gauss}(3\sigma^2/4 + 3\sigma^2/4) = \text{Gauss}(3\sigma^2/2) \end{aligned} \quad (41)$$

Therefore, we can obtain an estimation of the maximum variance of range query as follows:

$$\sigma^2_{\text{sumMaxEstim}} = 3\sigma^2 + \frac{3}{4}\sigma^2 + \frac{3}{2}\sigma^2 + \frac{3}{2}\sigma^2 = \frac{27}{4}\sigma^2 = \frac{6 \cdot 3 + 9}{4}\sigma^2. \quad (42)$$

In Table 7, the wavelet decomposition level is 3 (the number of input data is 2^3). Supposing that the decomposition level is l , then we can give an estimation of the maximum variance of range query:

$$\sigma^2_{\text{sumMaxEstim}} = \left(\frac{6l + 9}{4}\right)\sigma^2. \quad (43)$$

Note that equation (43) gives the same result with the conclusion in Theorem 6.

Comparing the estimation maximum variances (equation (43)) with the real maximum variances (Table 6), we find that

$$\sigma^2_{\text{sumMax}} \ll \sigma^2_{\text{sumMaxEstim}}. \quad (44)$$

TABLE 7: Coarse estimation of max-variance for range query.

$n_{0=}$	$n_{3,0}$	$+n_{3,1}$	$+n_{2,1}$	$+n_{1,1}$
$n_{1=}$	$n_{3,0}$	$+n_{3,1}$	$+n_{2,1}$	$-n_{1,1}$
$n_{2=}$	$n_{3,0}$	$+n_{3,1}$	$-n_{2,1}$	$+n_{1,2}$
$n_{3=}$	$n_{3,0}$	$+n_{3,1}$	$-n_{2,1}$	$-n_{1,2}$
$n_{4=}$	$n_{3,0}$	$-n_{3,1}$	$+n_{2,2}$	$+n_{1,3}$
$n_{5=}$	$n_{3,0}$	$-n_{3,1}$	$+n_{2,2}$	$-n_{1,3}$
$n_{6=}$	$n_{3,0}$	$-n_{3,1}$	$-n_{2,2}$	$+n_{1,4}$
$n_{7=}$	$n_{3,0}$	$-n_{3,1}$	$-n_{2,2}$	$-n_{1,4}$
$n_{k_i} \sim$	Gauss ($3\sigma^2/4^3$)	Gauss ($3\sigma^2/4^3$)	Gauss ($3\sigma^2/4^2$)	Gauss ($3\sigma^2/4^1$)
$\sigma_{sumMax}^2 =$	$3\sigma^2$	$+3\sigma^2/4$	$+3\sigma^2/2$	$+3\sigma^2/2$

$\sigma_{sumMaxEstim}^2$ provides a function expression using parameters l and σ^2 for the maximum variance of range query via Haar wavelet transform, but it has a very large error comparing the real maximum variance, comparing equation (43) with Table 6. Therefore, for the analysis of the practical applications, we prefer to use the maximum variance σ_{sumMax}^2 as listed in Table 6. \square

5. Experimental Verification

This section introduces the experimental verification of the proposed framework, that is, the computing of maximum variance for range query based on Gaussian mechanism and lifting Haar wavelet. We use the dataset age, which contains census records of individuals from the United States. The age has 107, 974, and 787 records, each of which corresponds to the age of an individual, extracted from the IPUM's census data of the United States [40]. The ages range from 0 to 135 and just have 128 values (ages 121, 123, 127, 128, 131, 132, 133, and 134 are empty). We count the number of each age and give the histogram of age as the input file of our experiments. Given a query length L , we test all the possible range queries with length L and report the maximum variance of the range query for input data.

The noise injected into the input data via Gaussian mechanism is Gaussian noise. The variance of Gaussian noise is the sample variance. Therefore, the sample variance is adopted in this study and is computed by the following equation:

$$Var(n) = \frac{1}{m-1} \sum_{i=1}^m \left(n_i - \frac{1}{m} \sum_{j=1}^m n_j \right)^2, \quad (45)$$

where n_i (or n_j) is the noise injected into the input data and m is the number of the input data.

We research the maximum variance of the range query of noise when ε chosen in the set $\{0.5, 0.75, 1.0, 1.25\}$ and δ chosen in the set $\{0.1, 0.01, 0.001\}$. For each special ε , we draw the maximum variance of the range count of noise using Gaussian mechanism and Gaussian mechanism with Haar wavelet transform when δ is equal to 0.1, 0.01, and 0.001.

For any ε and δ , we can calculate the σ of Gaussian noise using the equation $\sigma = \sqrt{2 \ln(1.25/\delta)}/\varepsilon$ in Theorem 5, and the results are given in Table 8.

To compute the variance, we inject the Gaussian noise into the input data or the Haar wavelet coefficients 10000 times. To compute the maximum variance of the range query

TABLE 8: σ for some ε and δ .

ε	δ	σ
0.5	0.1	4.4951
0.5	0.01	6.2150
0.5	0.001	7.5530
0.75	0.1	2.9967
0.75	0.01	4.1433
0.75	0.001	5.0353
1.0	0.1	2.2475
1.0	0.01	3.1075
1.0	0.001	3.7765
1.25	0.1	1.7980
1.25	0.01	2.4860
1.25	0.001	3.0212

with fixed ε and δ , each variance of the range query for range size k needs to be computed firstly. Then, the maximum value of variance for range size k can be given by comparing all the variance of the k -range queries.

For input data with length 128 (such as 128 histogram), we can draw the maximum variance diagram of the range query using Gaussian mechanism and Gaussian mechanism with Haar wavelet transform for any range sizes, as shown in Figure 5.

In Figure 5, ‘‘Gauss’’ means injecting noise into each histogram data via Gaussian mechanism directly and then gets the noise of range query by the operation of addition. First, ‘‘GaussWave’’ denotes injecting noise into the lifting Haar wavelet coefficients using Theorem 5. Second, the noise injected into each histogram is obtained by the inverse wavelet transform. Finally, the range query for any range size is obtained by injecting the noise together.

In Figure 5, we observe that the maximum variance is increasing linearly with the ‘‘range size’’ for ‘‘Gauss.’’ In Figure 5, for any ε and δ , the maximum variance of the noise using ‘‘GaussWave’’ method is far less than the noise using ‘‘Gauss’’ method.

To observe the variation tendency of ‘‘GaussWave’’ in Figure 5, we just draw the maximum variance diagram of the range query using Gaussian mechanism with Haar wavelet transform, as shown in Figure 6.

In Figure 6, for any ε and δ , the maximum variance of the noise using ‘‘GaussWave’’ method increases with the increase of range size before it gets the maximum value, and it will decrease with the increase of range size after it has gotten the maximum value.

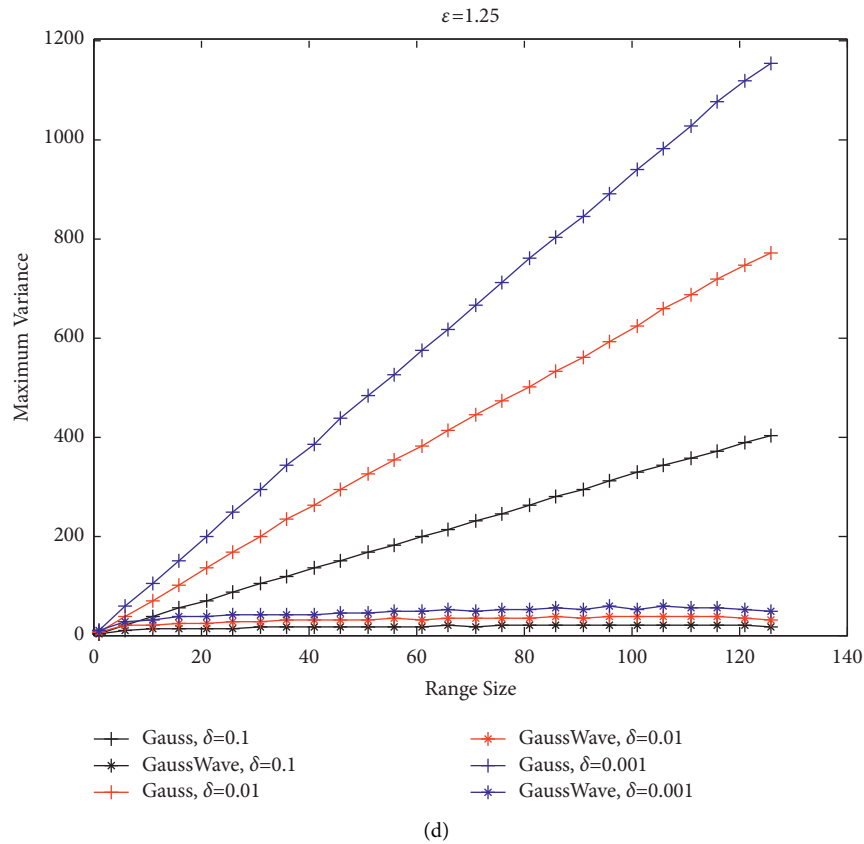
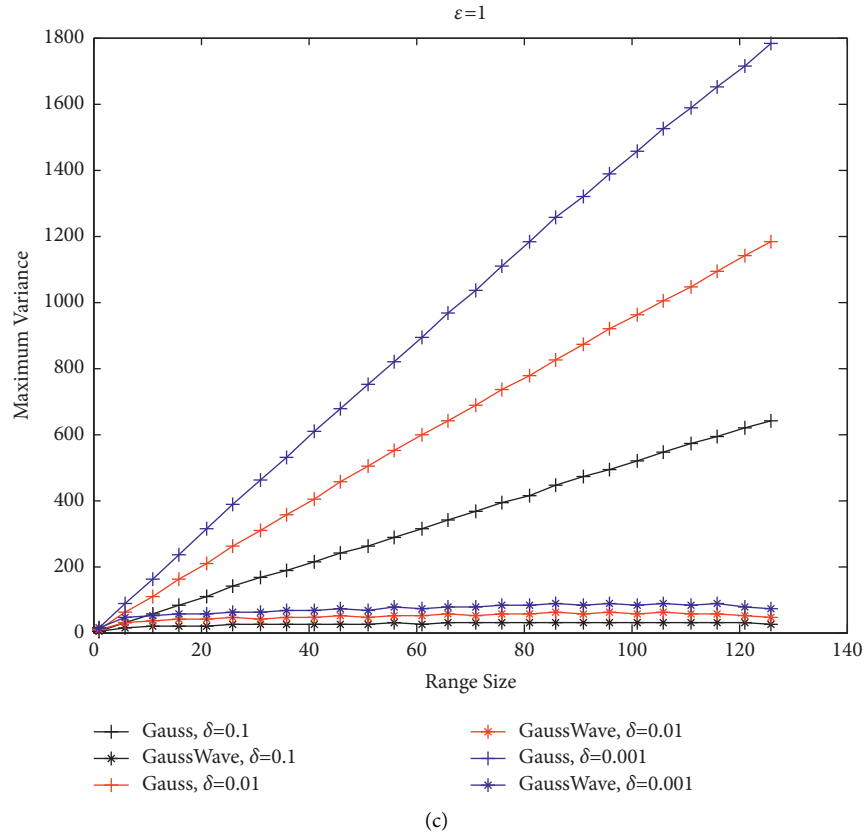


FIGURE 5: Maximum variance of range query using Gaussian mechanism and Gaussian mechanism with Haar wavelet on age (the United States). (a) $\epsilon = 0.5$. (b) $\epsilon = 0.75$. (c) $\epsilon = 1.0$. (d) $\epsilon = 1.25$.

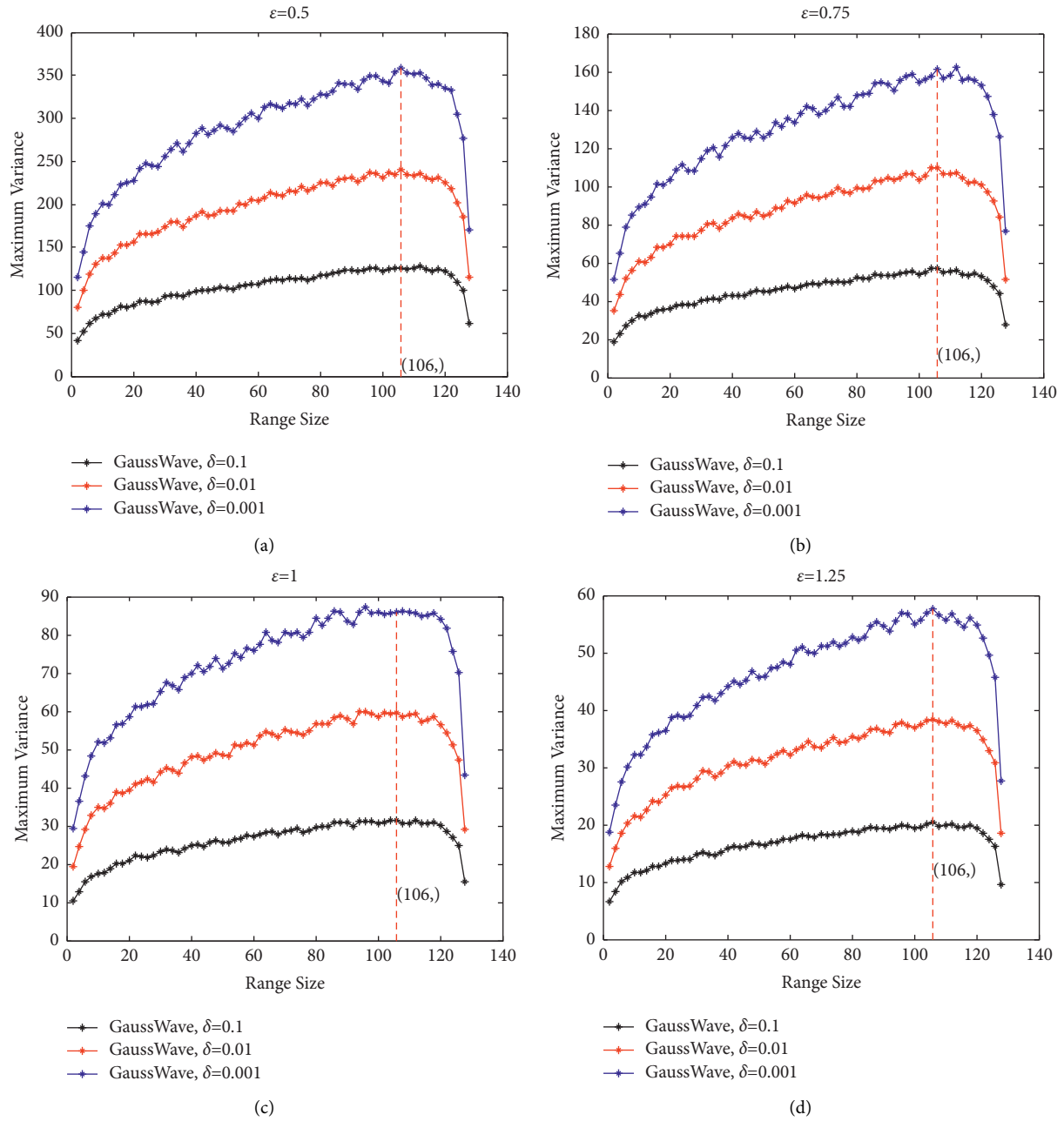


FIGURE 6: Maximum variance of range query using Gaussian mechanism with Haar wavelet on age (the United States). (a) $\epsilon = 0.5$. (b) $\epsilon = 0.75$. (c) $\epsilon = 1.0$. (d) $\epsilon = 1.25$.

TABLE 9: Comparison of maximum value between experimental result and theoretical analysis for range 1 to 128.

ϵ	δ	Range size	Max-value	σ^2_{sumMax}	Difference
0.5	0.1	106	126.093366	126.252493	-0.159127
0.5	0.01	106	240.611130	241.347894	-0.736764
0.5	0.001	106	358.354318	356.451312	1.903006
0.75	0.1	106	56.957740	56.1109710	0.846769
0.75	0.01	106	109.807594	107.264005	2.543589
0.75	0.001	106	161.471355	158.420708	3.050647
1.0	0.1	106	31.521392	31.5617190	-0.040327
1.0	0.01	106	59.637942	60.336974	-0.699032
1.0	0.001	106	86.007548	89.112828	-3.105280
1.25	0.1	106	20.466279	20.199500	0.266779
1.25	0.01	106	38.351021	38.615663	-0.264642
1.25	0.001	106	57.678236	57.032210	0.646026

In Table 6, we observe that the max-variance of range query via Gaussian mechanism with Haar wavelet is $6.248291 * \sigma^2$ when $l=7$ ($2^7=128$) and the length of interval is 106. Considering the value of σ in Table 8, we give the comparison of maximum value between experimental result and theoretical analysis for range query in Table 9.

In Table 9, the column “max-value” presents the experimental result of the maximum value of maximum variance for range query. The column “ σ^2_{sumMax} ” presents the result of theoretical analysis and computed using the formula $\sigma^2_{\text{sumMax}} = 6.248291 * \sigma^2$. The last column “difference” is the difference of columns “max-value” and “ σ^2_{sumMax} .” In Table 9, we find that the results of experiment and theoretical analysis are in substantial agreement.

This section gives the experimental verification of the framework of the maximum variance computing for range query. This framework on privacy preserving is built using Gaussian mechanism and Haar wavelet. In Figure 6, for any ϵ and δ , the maximum variance of the noise increases with the increase of range size before it gets the maximum value of 106, and it will decrease with the increase of range size after it has gotten the maximum value. In Table 9, the experimental value and the theoretical value of maximum variance for range count are compared, and the results show that they are in substantial agreement.

6. Conclusions

In this study, we proposed a new differential privacy framework via Haar wavelet transform and Gaussian mechanism for the range query. The theorems for how to inject Gaussian noise into the Haar wavelet coefficients are given. The noise of range query under the theoretical framework of Haar wavelet and Gaussian mechanism is analyzed. The algorithm to compute the maximum variance of any range query for any given parameter l is introduced. A coarse estimation of the maximum variance of range query using a function expression is given. The experimental results show that the maximum variance of the noise using Gaussian mechanism and Haar wavelet is far less than the noise using Gaussian mechanism. The experimental verification of the computing of maximum variance for range query based on lifting Haar wavelet and Gaussian mechanism is proposed, and the results show the experimental value and the theoretical value of maximum variance for range count are substantial agreement. For future work, we plan to apply our method to the privacy protection of histogram publication. Furthermore, we want to investigate how to assemble our method and machine learning algorithm, such as the decision tree and random forest.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Research Start-Up Funding Project of Qu Zhou University (nos. BSYJ202112 and BSYJ202109) and the Project for Public Interest Research Projects of Science and Technology Program of Zhejiang Province, China (nos. LGF21F010002, LGN21C130001, and LGG21F030002).

References

- [1] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Preliminary of Differential Privacy*, *Differential Privacy and Applications*, pp. 7–16, Springer, Cham, 2017.
- [2] P. Jain, M. Gyanchandani, and N. Khare, “Differential privacy: its technological prescriptive using big data,” *Journal of Big Data*, vol. 5, no. 1, pp. 15–24, 2018.
- [3] H. B. Kartal, X. Liu, and X. B. Li, “Differential privacy for the vast majority,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 10, no. 2, pp. 1–15, 2019.
- [4] J. Lin, J. Niu, X. Liu, and M. Guizani, “Protecting your shopping preference with differential privacy,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 1965–1978, 2021.
- [5] H. Wang and H. Wang, “Correlated tuple data release via differential privacy,” *Information Sciences*, vol. 560, pp. 347–369, Article ID 4267921, 2021.
- [6] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megias, “Individual differential privacy: a utility-preserving formulation of differential privacy guarantees,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1418–1429, 2017.
- [7] H. Wang and Z. Xu, “CTS-DP: publishing correlated time-series data via differential privacy,” *Knowledge-Based Systems*, vol. 122, pp. 167–179, 2017.
- [8] M. U. Hassan, M. H. Rehmani, and J. Chen, “Differential privacy techniques for cyber physical systems: a survey,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 746–789, 2020.
- [9] Z. Huang, R. Hu, Y. Guo, E. Chan-Tin, and Y. Gong, “DP-ADMM: ADMM-based distributed learning with differential privacy,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1002–1012, 2020.
- [10] X. Li, J. Yang, X. Sun, and Z. Jianpei, “Differential Privacy for Edge Weights in Social Networks,” *Security and Communication Networks*, vol. 2017, pp. 1–10, 2017.
- [11] P. Liu, Y. X. Xu, Q. Jiang et al., “Local differential privacy for social network publishing,” *Neurocomputing*, vol. 391, pp. 273–279, 2020.
- [12] H. Huang, D. Zhang, F. Xiao, K. Wang, J. Gu, and R. Wang, “Privacy-preserving approach PBCN in social network with differential privacy,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 931–945, 2020.
- [13] X. Ren, C. M. Yu, W. Yu et al., “ϵ-high-dimensional crowdsourced data publication with local differential privacy,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.
- [14] J. Wei, Y. Lin, X. Yao, and J. Zhang, “Differential privacy-based location protection in spatial crowdsourcing,” *IEEE Transactions on Services Computing*, vol. 15, pp. 1–14, 2019.
- [15] N. Almadhoun, E. Ayday, and Ö. Ulusoy, “Differential privacy under dependent tuples—the case of genomic privacy,” *Bioinformatics*, vol. 36, no. 6, pp. 1696–1703, 2020.

- [16] J. L. Raisaro, G. Choi, S. Pradervand et al., "Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1413–1426, 2018.
- [17] M. Z. Hasan, M. S. R. Mahdi, M. N. Sadat, and N. Mohammed, "Secure count query on encrypted genomic data," *Journal of Biomedical Informatics*, vol. 81, pp. 41–52, 2018.
- [18] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially private data publishing and analysis: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, 2017.
- [19] L. Qian, T. Song, and A. Liang, "The optimization of the range query in differential privacy," in *Proceedings of the International Conference on Computer Science and Network Technology (ICCSNT)*, vol. 1, pp. 618–623, IEEE, July 2015.
- [20] J. Zhang, X. Xiao, and X. Xie, "Privtree: a differentially private algorithm for hierarchical decompositions," in *Proceedings of the 2016 International Conference on Management of Data*, pp. 155–170, San Francisco, CA, USA, January 2016.
- [21] Z. Cai and Z. He, "Trading private range counting over big IoT data," *IEEE*, in *Proceedings of the IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 144–153, Dallas, TX, USA, July 2019.
- [22] H. Mahdikhani, R. Lu, Y. Zheng, and A. Ghorbani, "Achieving efficient and privacy-preserving range query in fog-enhanced IoT with bloom filter," in *Proceedings of the IEEE International Conference on Communications*, pp. 1–6, Dublin, Ireland, July 2020.
- [23] H. Mahdikhani, R. Lu, J. Shao, and A. Ghorbani, "Using reduced paths to achieve efficient privacy-preserving range query in fog-based IoT," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4762–4774, 2021.
- [24] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *The VLDB Journal*, vol. 22, pp. 797–822, 2013.
- [25] X. Meng, H. Li, and J. Cui, "Different strategies for differentially private histogram publication," *Journal of Communications and Information Networks*, vol. 2, no. 3, pp. 68–77, 2017.
- [26] Q. Han, B. Shao, L. Li, Z. Ma, H. Zhang, and X. Du, "Publishing histograms with outliers under data differential privacy," *Security and Communication Networks*, vol. 9, no. 14, pp. 2313–2322, 2016.
- [27] H. Li, J. Cui, X. Meng, and J. Ma, "IHP: improving the utility in differentially private histogram publication," *Distributed and Parallel Databases*, vol. 37, no. 4, pp. 721–750, 2019.
- [28] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Differentially Private Data Publishing: Interactive Setting*, "Differential Privacy and Applications", pp. 23–34, Springer, Cham, 2017.
- [29] F. M. Bayer, A. J. Kozakevicius, and R. J. Cintra, "An iterative wavelet threshold for signal denoising," *Signal Processing*, vol. 162, pp. 10–20, 2019.
- [30] M. Sharma, S. Patel, and U. R. Acharya, "Automated detection of abnormal EEG signals using localized wavelet filter banks," *Pattern Recognition Letters*, vol. 133, pp. 188–194, 2020.
- [31] T. Suzuki, "Wavelet-based spectral-spatial transforms for CFA-sampled raw camera image compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 433–444, 2020.
- [32] Y. Zhang, "The fast image encryption algorithm based on lifting scheme and chaos," *Information Sciences*, vol. 520, pp. 177–194, 2020.
- [33] S. P. Singh and G. Bhatnagar, "A simplified watermarking algorithm based on lifting wavelet transform," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20765–20786, 2019.
- [34] M. B. Nagare, B. D. Patil, and R. S. Holambe, "On the design of biorthogonal halfband filterbanks with almost tight rational coefficients," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 4, pp. 790–794, 2020.
- [35] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1200–1214, 2011.
- [36] P. Derbeko, S. Dolev, and E. Gudes, "Wavelet-based dynamic and privacy-preserving similitude data models for edge computing," *Wireless Networks*, vol. 27, no. 1, pp. 351–366, 2021.
- [37] C. Lin, P. Wang, H. Song, Y. Zhou, Q. Liu, and G. Wu, "A differential privacy protection scheme for sensitive big data in body sensor networks," *Annals of Telecommunications*, vol. 71, no. 9–10, pp. 465–475, 2016.
- [38] C. Dwork, K. Talwar, A. Thakurta, and Li. Zhang, "Analyze gauss: optimal bounds for privacy-preserving principal component analysis," in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 11–20, May 2014.
- [39] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, pp. 211–407, 2013.
- [40] S. Ruggles, K. Genadek, and R. Goeken, *Integrated public use microdata series: Version 6.0 [dataset]*, University of Minnesota, vol. 10, p. 010, Minneapolis, 2017.

Research Article

Application and Analysis of Improved Fuzzy Comprehensive Evaluation Method in Goodwill Evaluation and Intangible Asset Management

Mengyue Xu 

University of Toronto, Ontario, Mississauga L5L1C6, Canada

Correspondence should be addressed to Mengyue Xu; mengyue.xu@mail.utoronto.ca

Received 15 July 2022; Revised 10 August 2022; Accepted 16 August 2022; Published 9 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Mengyue Xu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to improve the effect of goodwill evaluation and intangible asset management, this paper combines the improved fuzzy comprehensive evaluation method to construct an intelligent algorithm. In order to consider the many influencing factors of intangible asset prices and their own influence, this experiment adopts the ARDL model for analysis. Based on the traditional LSTM model, this paper innovatively introduces sparse principal component analysis for feature dimensionality reduction and, on the other hand, adds a convolutional neural network to the model, which is suitable for feature extraction. In addition, this paper constructs a goodwill evaluation and intangible asset management model based on the improved fuzzy comprehensive evaluation method. The research shows that the improved fuzzy comprehensive evaluation method proposed in this paper has a very good application effect in goodwill evaluation and intangible asset management.

1. Introduction

The capital value of intangible assets can be determined through the evaluation of intangible assets. Intangible assets can be valued as human shares in the process of group formation, listing, merger, and joint venture of an enterprise. Moreover, an accurate assessment of the value of intangible assets can provide an important protection for the company's intellectual property rights and proprietary technologies. At the same time, it can protect the company from having a value compensation standard when it is infringed, and based on this, it can ask the infringer for infringement damages, so as to avoid huge losses to the company [1]. The evaluation of the value of intangible assets by business owners can not only determine the value of assets [2]. In addition, information about the enterprise can also be obtained and provided to managers. Although the management of intangible assets of enterprises has gradually taken shape in recent years, the managers of enterprises are not clear about how many intangible assets an enterprise has and the value of the intangible assets, which can easily form a

blind spot in enterprise asset management. After asset evaluation, the true value of assets can be fully revealed, which is more conducive to managers to make appropriate business plans according to the situation of the enterprise, and can also enhance the confidence of investors [3]. Second, intangible assets with clear value can generate economic benefits in economic activities, enabling enterprises to use externally for paid or mortgage loans to obtain funds when there is a problem of capital turnover. Furthermore, the evaluation of intangible assets can also reflect the influence and development of the brand in the industry and region, and understand the value of the corporate brand [4]. The development and management of the brand and the utilization of resources within the enterprise can be further understood by evaluating the value of intangible assets. Therefore, by adjusting the business plan of the enterprise through the choice of consumers among the industries, it is possible to obtain the operation plan and target of the next cycle and strive to maximize the profit of the enterprise [5].

From the perspective of the social function of asset appraisal, the social function of asset appraisal is to provide

value scales for asset business and capital markets. Most of the current asset appraisal projects involve corporate restructuring or property rights changes, such as corporate mergers and acquisitions, listings, and sino-foreign joint ventures. The evaluation conclusion is an important basis for the asset business parties to conclude the transaction. If the capital price is inaccurate, it will not only mislead the allocation of resources, and reduce the efficiency of social resource allocation, but also cause misunderstandings in property rights transactions and various capital businesses, damage the legitimate rights and interests of the parties involved, and affect the security, prosperity, and stability of the capital market [6].

The analytic hierarchy process (AHP) is a method that combines quantitative and qualitative analysis methods for problems that cannot be quantitatively researched by establishing mathematical models and data, and follows the sequence of decomposition first and synthesis later, step by step, to deal with some complex problems, so as to obtain satisfactory decision-making results [7]. Therefore, AHP is also widely used in the evaluation of intangible assets. In the evaluation of machinery and equipment, the analytic hierarchy process is introduced to improve the new rate. Based on the purpose of determining the weight, the analytic hierarchy process is used to construct a judgment matrix, the weight and order of the required indicators are obtained, and the corresponding weight of each indicator is clarified. Combined with specific standards, the assignment of each influencing factor is completed to ensure the accuracy of the index system and avoid the influence of subjective factors [8]. In the process of implementing the construction of the data asset evaluation system, it is first necessary to analyze the relevant factors that affect the value of the data assets and combine the AHP to complete the construction of the final evaluation system. At the same time, it is necessary to ensure that the evaluation system can cover the corresponding costs and applications of the data assets. [9]. We calculate the main weights of different evaluation indicators and the corresponding target layer, design specific operation methods, and verify the operability of the evaluation system with examples [10]. The influencing factors and related indicators of intangible assets are obtained through the expert group, and the relative importance of each factor is obtained by using the analytic hierarchy process [11]. In the research of intangible assets of network database, the value evaluation of database assets is completed by the income method and the analytic hierarchy process, a certain degree of correction is carried out, and then, the weight of each asset is used to complete the evaluation of intangible assets [12]. With the help of the AHP, the evaluation and analysis of the goodwill value of the household appliance industry are carried out, and the factor model that affects the goodwill value is constructed to achieve the effect of revising the evaluation results of the traditional income method and improve the accuracy of the evaluation results [13]. In asset evaluation of all walks of life, especially intangible asset evaluation, the combination of AHP and income method has been widely recognized.

As far as the value transfer of the intangible assets of enterprises is concerned, it is generally expressed as external investment and transfer and sale. This transfer must be premised on the profitability of intangible assets. At this point, the focus of both parties in the transaction is on the function of the intangible asset, and the price of the intangible asset is determined accordingly. The way intangible assets play a role is obviously different from that of tangible assets [14].

Intangible assets do not have material entities and cannot be directly sensed by people. They are invisible assets, but such assets are often attached to certain entities. Physical entities such as equipment and production lines play their role, trademark rights are reflected through product quality, goodwill is embedded in the overall asset portfolio of the enterprise, and the value of intangible assets is materialized in the depth and breadth of tangible assets, which determines the degree of intangible assets. Therefore, the transfer price of intangible assets must fully consider the scope of intangible assets used to “arm” tangible assets [15].

The principles of economics tell us that the organic combination of people, wealth, and things is the source of asset appreciation. In the primitive capital accumulation process of capitalism, if an enterprise achieves assets of several hundred million, it requires the efforts of several generations. In modern economic society, the value added of intangible assets is on the rise. The value rapidly increases [16]. Intangible assets are like catalysts in chemical reactions, and levers in mechanical physics, accelerating and amplifying the effects of tangible assets. The high efficiency of the use value of intangible assets can bring future excess returns to enterprises [17].

The present value of income of intangible assets is to measure the actual value of intangible assets from the perspective of future expected income, which is beneficial to enterprises in the process of transferring intangible assets. It is more objective to observe and confirm the value of intangible assets in a certain period of time in the future, but the present value method of income inevitably has strong subjective judgment factors [18]. At present, this method mainly measures the transfer price of intangible assets, that is, the foreign investment and transfer business of intangible assets. Generally, the transfer price is greater than the book cost. For the original enterprise, if it is a transfer sale, the transfer price is the transfer price, and the original book cost is the transfer cost; if it is an external investment, the transfer price is the investment cost, and the value-added part (transfer price—book cost) is converted into enterprise capital reserve [19].

This paper combines the improved fuzzy comprehensive evaluation method to conduct research on goodwill evaluation and intangible asset management, and to improve the evaluation effect of intangible assets.

2. Intangible Asset Evaluation Algorithm

The full name of the ARDL model is the autoregressive distributed lag model. It includes the dependent variable, the independent variable, and the lag term of the independent

variable. The model does not require that the variables must be of the same order, both I (0) and I (1) can be used, and the model can directly test the short-term and long-term relationships between the dependent variable and the independent variable. As long as the ordinary least square (OLS) assumption is satisfied, the model can be directly estimated. In order to consider the many influencing factors of intangible asset prices and their own influence, this experiment adopts the ARDL model for analysis.

The structure of the ARDL(p, q_1, q_2, \dots, q_k) model is as follows:

$$f(L, P)y_t = \sum_{i=1}^k \beta_i(L, q_i)x_{it} + dw_t + ut. \quad (1)$$

Among them,

$$f(L, P) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p, \quad (2)$$

$$\beta_i(L, q_i) = 1 - \beta_{i1} - \beta_{i2} L^2 - \dots - \beta_{iq_i} L^{q_i}. \quad (3)$$

In formulas (1)–(3), p is used to represent the lag order of y_t , and q_i is the lag order of x_{it} , $i = 1, 2, 3, \dots, k$. The lag operator L is defined as follows: $Ly_t = y_{t-1}$, w_t represents a vector with d rows and 1 column.

Regarding the determination of the lag order of the variable, the modified AIC criterion can be used. The criterion has strong theoretical advantages, and at the same time, the dummy variable with the minimum value is input to solve the problem of outliers. In regression, using these dummy variables can improve the reliability of the model.

To solve the “long dependency problem” of recurrent neural networks, long short-term memory neural networks (LSTMs) were proposed. It is the most commonly used model in recursion and a variant of the recurrent neural network. Compared with the recurrent neural network, the LSTM method increases the memory cells used to form the processing. Based on the “long dependency problem,” coupled with the research and modification of many scholars, the more common LSTM structure now is the introduction of “gate.” The “gate” structure can select the most effective feature among many features for processing, so as to achieve the purpose of controlling the flow of information. A schematic diagram of the basic structure of LSTM is shown in Figure 1.

The structure of the recurrent neural network is a chain connection. The obtained information accepts the input of the previous node, outputs the current node state, and transmits it to the next node for input again. As can be seen from Figure 1, on the whole, the structure of LSTM is the same as that of the recurrent neural network, but the internal input and output have undergone major changes, and the number of transfer states has changed. Each storage unit of LSTM consists of a forget gate, an input gate, and an output gate. The functions of the three gates are to adjust the state of the transmitted information. In the process of transmission, the function of the forget gate can be summarized as “forget the unimportant and leave only the important.” The forget gate controls which information of the previous cell state

C_{t-1} needs to be left and which needs to be forgotten through calculation. The sigmoid function exists as a gated state unit. When the function value is 0, all information is discarded, and when the function value is 1, all information is left. The wf and bf represent the weights and biases of the forget gate.

$$f_t = \sigma(wf \cdot [h_{t-1}, x_t] + bf). \quad (4)$$

Next, the input gate corresponds to the selection memory stage, and the function of this stage is to selectively “remember” the incoming information. Those who feel important information can occupy a larger proportion, and the unimportant information can be appropriately discarded. The gate control signal of the input gate is controlled by the tanh function. The calculation formula of this stage is as follows:

$$\begin{aligned} it &= \sigma(wf \cdot [h_{t-1}, x_t] + b_i) \\ \widehat{Ct} &= \tan h(W_c \cdot [h_{t-1}, x_t] + b_c) \\ Ct &= f_t * Ct - 1 + it * \widehat{Ct}. \end{aligned} \quad (5)$$

The last structure of the memory is the output gate structure, which corresponds to the output stage of the model. This stage mainly controls which information obtained will be regarded as the output of the current state. This process is controlled by the tanh function, which is equivalent to scaling the cell state obtained in the previous stage. The calculation formula of the entire output gate structure is as follows:

$$\begin{aligned} \sigma_t &+ \sigma(w_o \cdot [h_t - 1, x_t] + b_0) \\ h_t &= \sigma_t * \tan h(Ct). \end{aligned} \quad (6)$$

The above is the basic model structure and calculation process of the LSTM model. The model needs to update the parameters in the empirical process, and the update process follows the gradient descent rule and the backpropagation rule.

Based on the traditional LSTM model, this paper innovatively introduces the sparse principal component analysis (SPCA) for feature dimension reduction on the one hand and adds a convolutional neural network to the model on the other hand. The convolutional neural network is suitable for feature extraction, and it is mainly used to deal with image classification problems in the early stage of its development. For example, it inputs a picture into the machine, and the machine tells us the category of this picture (such as flowers and people), and the convolutional neural network classifies by grabbing the distinctive category features in the picture. Nowadays, some studies also use it on one-dimensional time-series forecasting problems. In this paper, the convolutional layer is added to the traditional LSTM model, which can discover the data features more deeply. Using this network structure to train the prediction model can also obtain better prediction results.

2.1. Sparse Principal Component Analysis. Considering that some of the economic characteristic indicators that affect the

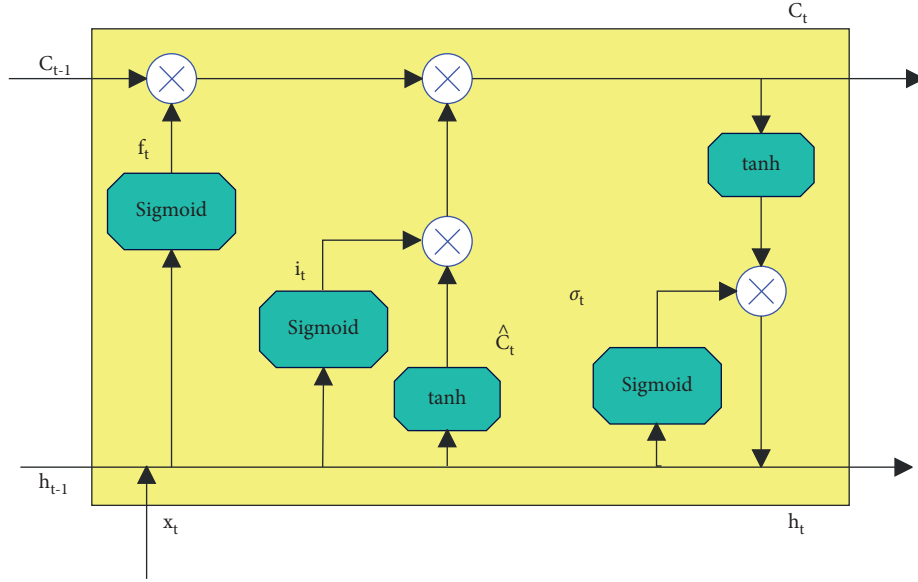


FIGURE 1: Schematic diagram of LSTM memory cell structure.

price level of intangible assets selected in this paper are not completely independent of each other, such as oil and gold prices, there may be a certain correlation, resulting in redundant information between each characteristic. The correlation between variables may lead to problems such as the unstable structure of the model and heavy training burden. In this paper, SPCA is used for feature dimensionality reduction. Relatively speaking, although the principal component analysis can extract features, sometimes the extracted features are difficult to explain. By adding a penalty function, SPCA can sparse the principal component coefficients and highlight important components, which is conducive to interpretation.

The SPCA model is essentially an optimization problem, and the objectives to be optimized are as follows:

$$\begin{aligned} \min_{A,B} & \|X - XBA^T\|_F^2 \lambda \sum_{i=1}^k \|\beta_i\|^2 + \sum_{i=1}^k \lambda_{1,i} \|\beta_i\|_1 \\ \text{s.t.} & A^T A = Ik. \end{aligned} \quad (7)$$

In formula (10), X represents the collected data matrix, A and B are the coordinate blocks obtained by the coordinate descent method, and the initial value is calculated according to the load obtained by the principal component analysis. β_i is the load obtained from the solution, p is the number of variables, and n is the number of samples. Regarding the value of λ , if $n \leq p$ is in the dataset, $\lambda > 0$ is required; otherwise, $\lambda = 0$ is taken. $\lambda_{1,i}$ is a positive parameter, used to control the parameter β_i .

Next, one of the two coordinate blocks is fixed, the other coordinate block is solved, and then, the solution is exchanged. The first k principal components obtained are denoted as $Y_i = Xv_i, i = 1, 2, \dots, k$, and the obtained sparse principal components are used for fitting. The optimization problem (10) is transformed into k independent elastic net problems, and the formula is shown in

$$\min_{\beta_i \in R^2} \frac{1}{2} \|X\beta - y_i\|_2^2 + \lambda \|\beta\|_1, i = 1, 2, \dots, k. \quad (8)$$

In the process of solving the above formula, since each problem is a standard lasso problem, it is possible to separately solve a subproblem first and then solve the remaining subproblems in the same way. After solving, the sparse load can be obtained, that is, β_i in formula (10). Multiplying the sparse loading matrix with the initially collected sample data matrix can get a new sample set after dimensionality reduction.

Regarding the coordinate descent method used to solve the lasso problem, the method is simple and easy to implement. The main principle is to solve only along a certain coordinate direction and fix other directions. The problem transformation of the solution process is as follows:

$$\beta_i^* = \operatorname{argmin}_{\beta} \left\{ \sum_{j=1}^n X_{ji}^2 \beta_i^2 - \left(\sum_{j=1}^n X_{ji} r_j^{(i)} \right) \beta_i + \lambda |\beta_i| \right\}. \quad (9)$$

Among them, $r_j^{(i)} = y_j - \sum_{k \neq i} x_{jk} \beta_k$.

Before solving problem (9), proposition 1 needs to be given first.

Proposition 1. $\forall a \in R$, and $\lambda > 0$, $S\lambda(a)/b$ represents the minimum point of the function $1/2bz^2 - az + \lambda|z|$, where $S\lambda(a)$ represents the soft threshold operator, and the value is as follows:

$$S\lambda(a) = \begin{cases} a - \lambda, & a > \lambda \\ 0, & |a| \leq \lambda \\ a + \lambda, & a < -\lambda. \end{cases} \quad (10)$$

According to proposition 1, we can get the following:

$$\beta_i = \frac{s\lambda \sum_{j=1}^n x_j r_j^{(i)}}{\sum_{j=1}^n x_j^2}. \quad (11)$$

Therefore, the load formula solved by the solution coordinate descent method is as follows:

$$\beta_i = \frac{s\lambda (\langle x_i, y \rangle - \sum_m: |v_m| > 0 \langle x_i, x_m \rangle v_m + v_i)}{x_i^T x_i}. \quad (12)$$

Subsequently, k times are solved to obtain the sparse load matrix $G = (\beta_1, \beta_2, \dots, \beta_k)$. Finally, the k principal components can be obtained by multiplying the obtained sparse loading matrix G with the original data feature matrix Z .

2.2. LSTM Model. During the model training process of LSTM, the activation function will have an impact on the model. If the network layer and the activation function are too deep and the weight is too small, it will easily lead to the disappearance of the gradient, or the phenomenon of gradient explosion will occur. For this, batch normalization (BN) is added to the model. BN uses the minibatch during learning as a unit, and normalizes the data of the node to make the mean value of 0 and the variance of 1. After normalization, scale is added to restore the data to its original state. Through BN, the learning rate of the model can be increased, and the training data can be disrupted to improve accuracy. Since it is normalized, BN will be affected by the size of the number of samples (recorded as Batch_size) selected for one training. Therefore, this experiment needs to adjust the Batch_size parameter.

In addition, the LSTM model also needs to set the number of layers of the LSTM stack, the objective function, etc. The more appropriate the parameter settings, the better the model effect and the stronger the model's predictive ability. The parameters in this experiment are set to two, the first is the number of neurons in the LSTM layer n , and the other is the minibatch size m of BN.

This paper analyzes these two parameters. In theory, the increase in the number of neurons in the hidden layer will increase the number of extracted features. The more features, the smaller the error. However, the number of neurons should not be too many, and too many will lead to overfitting.

Therefore, in this experiment, we first set the minibatch size to 0 and the number of iterations to 100. Under this condition, we sequentially increased the number of neurons one by one to analyze the change in the error value. Overall, when the number of LSTM hidden neurons is 50 and the minibatch size in the Batch_norm parameter is 40, and the RMSE and MAE values are relatively low. Therefore, when only the parameters of the LSTM model are considered, this set of parameters can be selected for model prediction.

2.3. CNN-LSTM Model. In this paper, SPCA is added to the traditional LSTM model for feature extraction, and a convolutional neural network (CNN, convolutional neural network) is added to the SPCA-LSTM model to extract local

features. The convolutional neural network is similar to the fully connected neural network. The special feature of the network structure is that the convolutional layer and the pooling layer are added. The convolutional layer is mainly to enhance the information of the features and extract the original features. The pooling layer reduces the dimension through the operation of the pooling function.

Based on the traditional LSTM model, the CNN-LSTM network structure diagram in this paper is shown in Figure 2, which mainly includes an input layer, a one-dimensional convolutional layer, a pooling layer, an LSTM layer, a fully connected layer, and an output layer.

In the convolutional layer, the traditional convolutional neural network is mainly used for image processing, and the output tensor is generated by sliding the convolution kernel in different dimensions, so as to perform three-dimensional convolutional layer operations to solve image processing problems. However, for time-series forecasting problems, due to its one-dimensional data characteristics, one-dimensional convolutional layers can be used for calculation. A schematic diagram of the operation process is shown in Figure 3.

As can be seen from Figure 3, in the process of convolution operation, the size of the convolution kernel is very important, which is related to the effect of the entire convolution operation. In general, the size setting of the convolution kernel is arbitrary. However, if the convolution kernel is too large, each operation of the convolution process will take a lot of time, resulting in low efficiency and a large amount of operation in the convolution process. If the convolution kernel is too small, only part of the features can be extracted during the feature extraction process. The resulting result is as if the hair feature of only one face image is extracted in the image processing problem, which is not conducive to image classification and may lead to wrong classification results.

In this model, the input size is assumed to be (Height, Weight), abbreviated as (H, W), the size of the convolution kernel is (FH, FW), and the size of the convolution operation is (KH, KW). The padding of the convolution process is set to P and the stride to S. Then, the KH and KW calculation formulas are as follows:

$$\begin{aligned} KH &= \frac{H + 2P - FH}{S} + 1 \\ KW &= \frac{H + 2P - FW}{S} + 1. \end{aligned} \quad (13)$$

It can be seen from the calculation formula of KH, KW that four parameters need to be set during the convolution operation: the number of convolution kernels n , the size of the convolution kernel (FH,FW), the stride S, and the activation function f . At present, the commonly used activation functions are sigmoid, ReLU, etc. The ReLU function is used in this paper. The input of this model is positive, so the problem of gradient disappearance is not easy to occur, but the gradient explosion that is easy to accumulate should also be carefully accumulated during the experiment. The ReLU function expression is as follows:

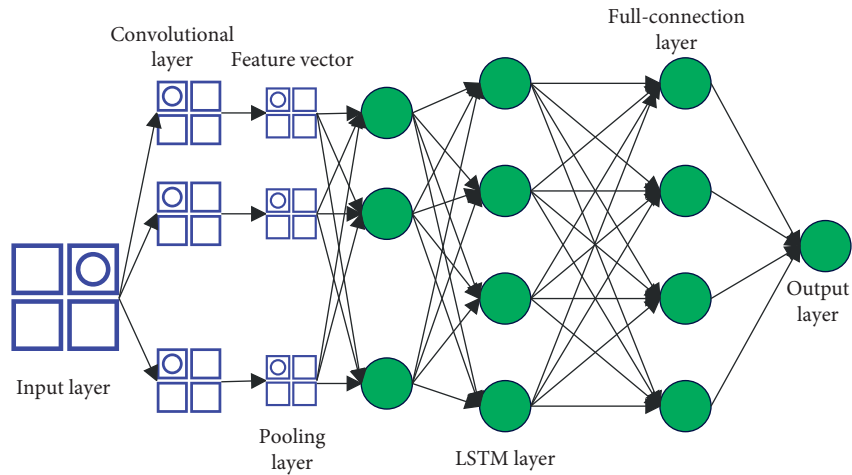


FIGURE 2: Structure diagram of CNN-LSTM network.

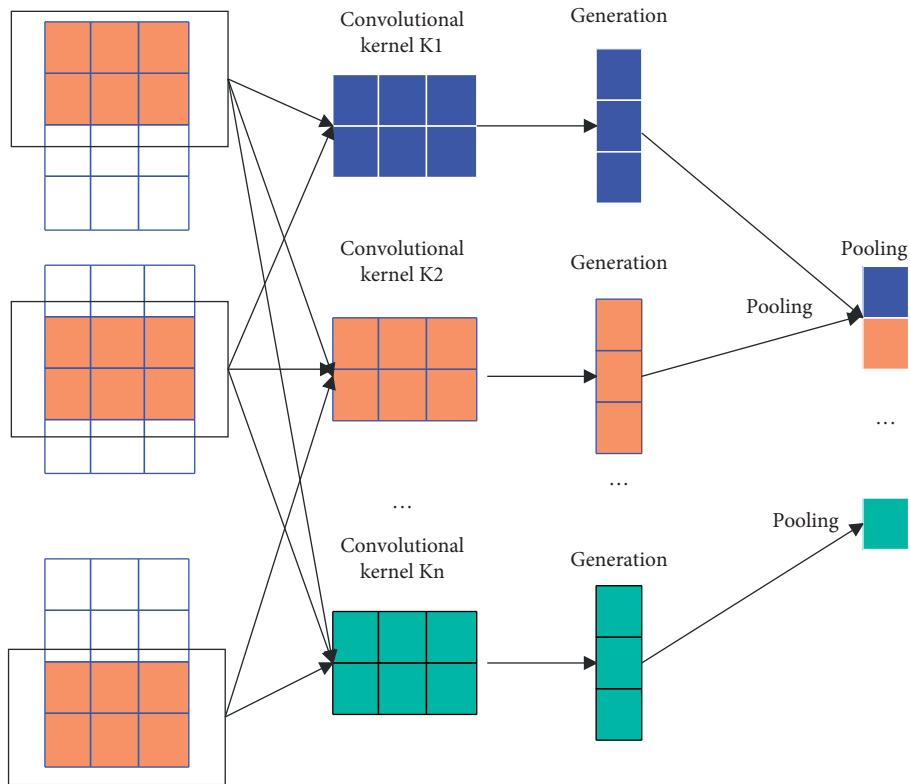


FIGURE 3: Schematic diagram of the convolution operation process.

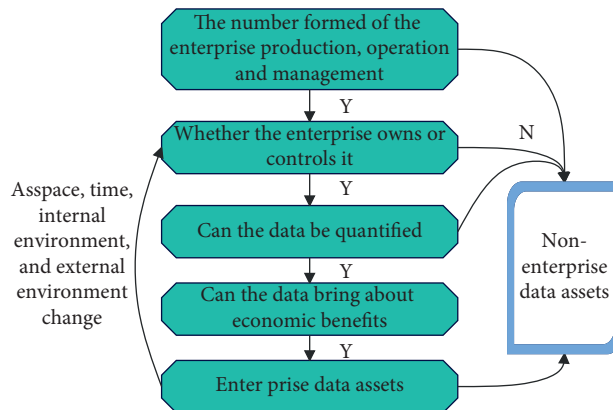


FIGURE 4: Identification process of intangible assets.

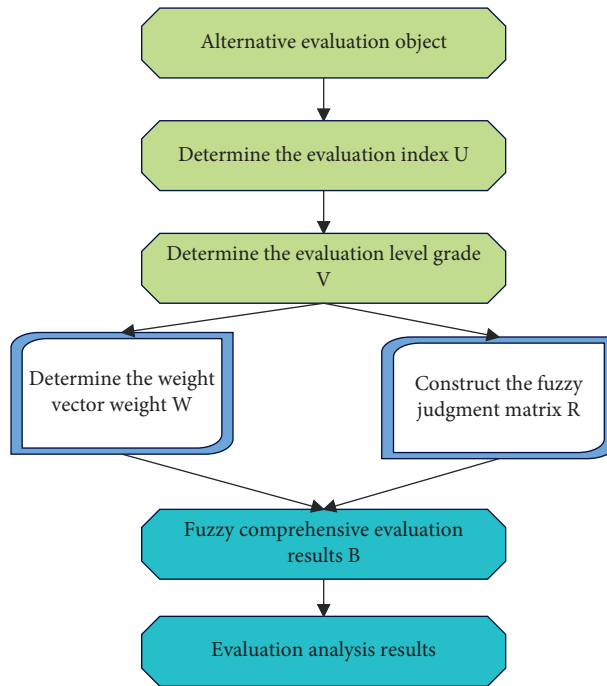


FIGURE 5: Fuzzy comprehensive evaluation method.

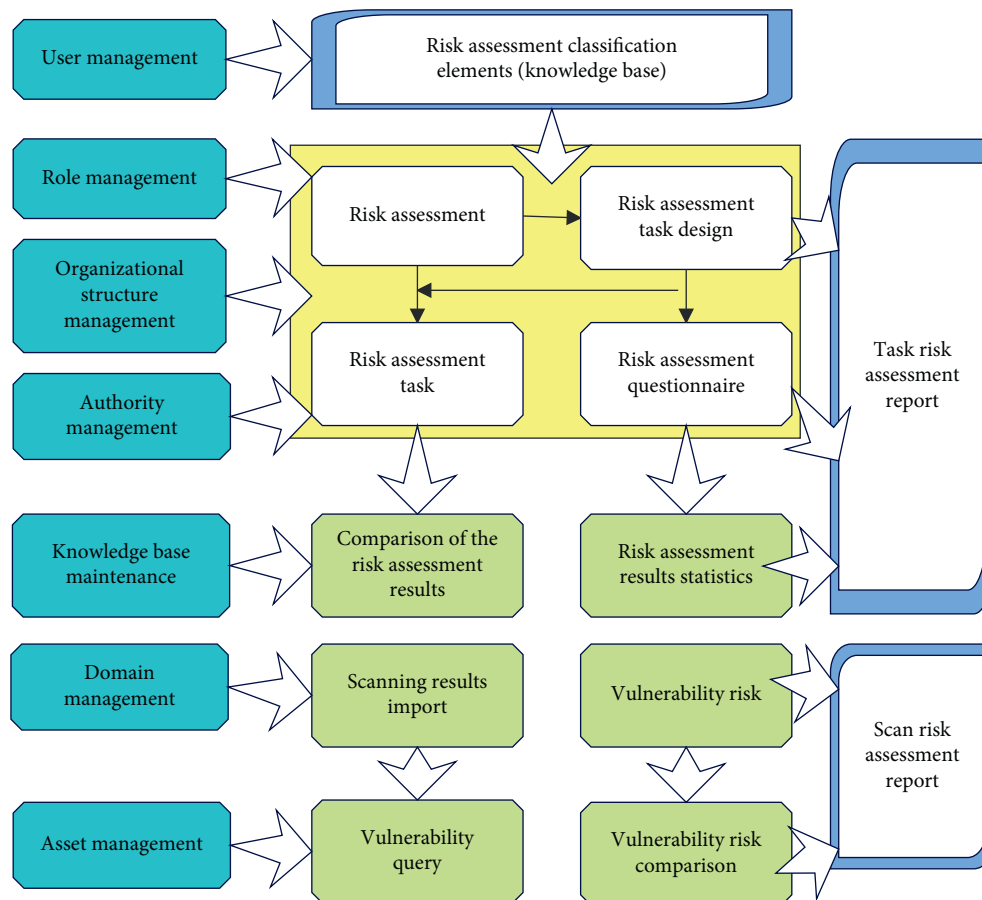


FIGURE 6: Overall design diagram of intangible asset security risk assessment.

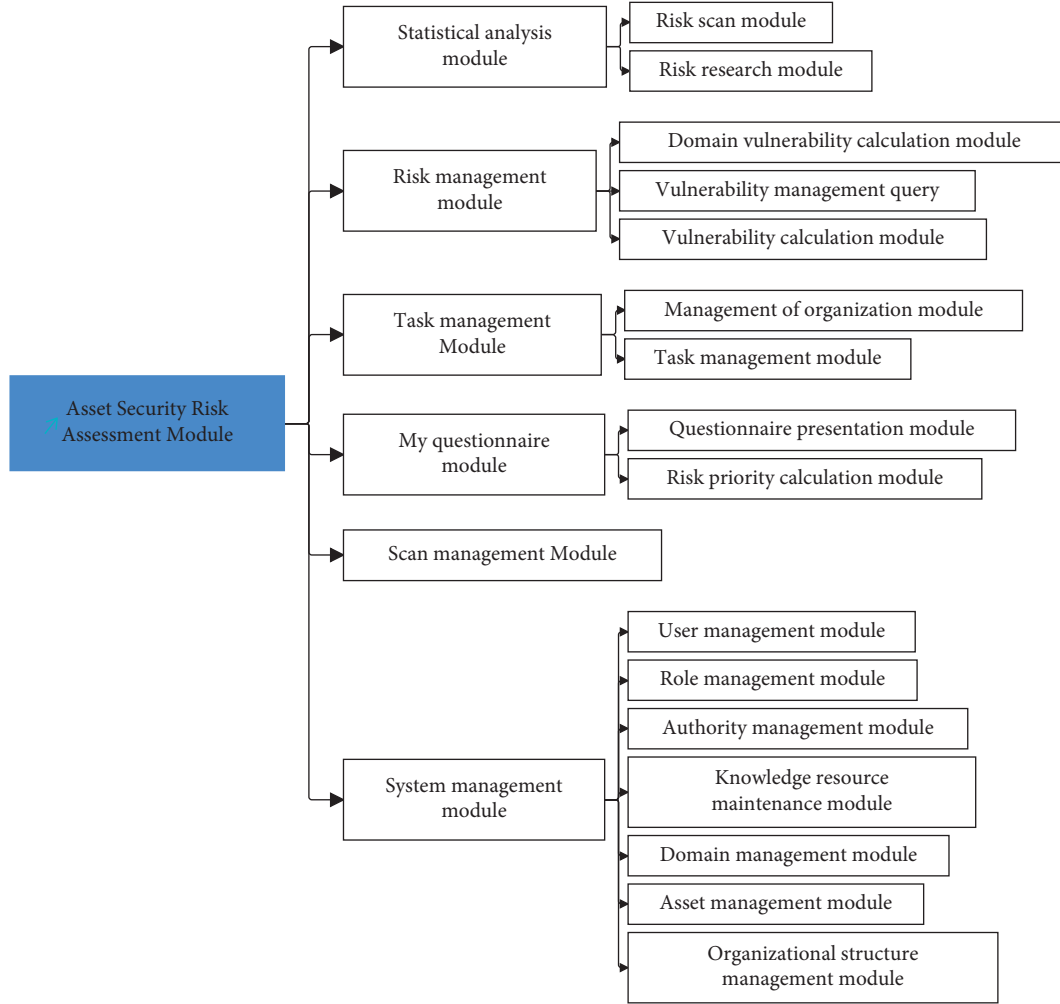


FIGURE 7: The division diagram of the asset security risk assessment module.

$$f(x) = \begin{cases} x & (x > 0) \\ 0 & (x \leq 0). \end{cases} \quad (14)$$

After the convolutional layer operation, the output vector can be expressed in the following form: output=(batch, new-steps, filters), new-steps refers to the new step size of the output, and filters represents the number of feature vectors.

Pooling is a process of abstracting information, and it is a spatial operation that reduces the amount of model parameters and optimization. The common pooling methods include average pooling and max pooling. Average pooling is to select the average value in the area, and the average value can often retain the overall characteristics of the data. Max pooling is to take out the maximum value in the target area, which tends to preserve texture features better. Therefore, average pooling is used in this experiment. The

size w of the pooling window is usually set to the same value as the stride S .

In the LSTM layer and the fully connected layer, random orthogonal matrix initialization is used for weight initialization, and the parameter to be considered in the LSTM layer is the number of neurons u . The fully connected layer uses a linear activation function and is responsible for synthesizing the information. The output layer is used to output prediction results.

The loss function is calculated using mean squared error with the following formula:

$$E = \frac{1}{2} \sum_k (y_k - tk)^2. \quad (15)$$

Among them, tk represents the original input data, y_k represents the output of the neural network, and k represents the dimension of the data. In the training process, batch

TABLE 1: Application effect of the improved fuzzy comprehensive evaluation method in goodwill evaluation.

Num	Goodwill assessment
1	85.41
2	83.67
3	88.69
4	81.76
5	90.81
6	90.51
7	81.32
8	86.24
9	83.36
10	86.63
11	79.40
12	88.70
13	88.27
14	88.50
15	83.07
16	87.61
17	88.03
18	79.84
19	82.16
20	85.67
21	81.15
22	81.29
23	81.07
24	90.24
25	90.42
26	88.61
27	88.65
28	80.51
29	86.53
30	79.63
31	88.26
32	80.17
33	89.56
34	82.80
35	89.70
36	84.42
37	80.03
38	81.32
39	90.04
40	81.76
41	89.96
42	88.88
43	80.42
44	87.68
45	79.22
46	79.12
47	87.09
48	89.72
49	87.89
50	85.32
51	88.03
52	90.44
53	82.75
54	80.03

TABLE 2: Application effect of the improved fuzzy comprehensive evaluation method in intangible asset management.

Num	Intangible asset management
1	85.82
2	90.38
3	84.80
4	90.01
5	87.70
6	89.43
7	92.89
8	85.80
9	90.74
10	84.89
11	91.50
12	88.80
13	84.86
14	86.96
15	91.12
16	91.08
17	90.42
18	84.24
19	90.39
20	89.45
21	85.21
22	89.45
23	89.61
24	91.16
25	92.16
26	90.87
27	92.90
28	92.62
29	89.67
30	92.99
31	90.01
32	89.88
33	92.43
34	85.58
35	90.73
36	91.32
37	92.30
38	90.17
39	92.34
40	88.36
41	89.82
42	92.52
43	89.47
44	86.86
45	92.15
46	92.44
47	84.13
48	87.57
49	88.13
50	92.31
51	84.08
52	85.24
53	86.26
54	90.28

normalization (BN) parameters are also added to each layer, and the size m of Batch_size needs to be determined after experiments.

3. Application of Improved Fuzzy Comprehensive Evaluation Method in Goodwill Evaluation and Intangible Asset Management

The identification process of intangible assets is shown in Figure 4.

The fuzzy comprehensive evaluation method is applied to goodwill evaluation and intangible asset management, as shown in Figure 5.

System design work should be performed top-down. First of all, this paper designs the overall structure and then goes deeper into it layer by layer until it reaches the design of each module. The system should be supported by functions such as user management, role management, organizational structure management, authority management, knowledge source maintenance, domain management, and asset management. Through the comprehensive analysis to find out the relationship between various functions, the overall design of this system is shown in Figure 6.

The basic principle of system module division according to software engineering theory is high cohesion and low coupling. The advantage of this is to facilitate the separation of module code and improve the reusability of module code. The system developed based on this principle has good maintainability and expansibility. This system uses web service technology to support the development and operation of the system under the J2EE-based environment. The system mainly includes six functional modules: system management module, scanning management module, statistical analysis module, my questionnaire module, task management module, and risk management module. The module division diagram is shown in Figure 7.

Based on the above model, the application effect of the improved fuzzy comprehensive evaluation method proposed in this paper in goodwill evaluation and intangible asset management is evaluated, and the results shown in the following Tables 1 and 2 are obtained.

It can be seen from the above research that the improved fuzzy comprehensive evaluation method proposed in this paper has a very good application effect in goodwill evaluation and intangible asset management.

4. Conclusions

Intangible assets are a comprehensive resource of an enterprise. In addition, intangible assets attract foreign investment to attract external investors to invest in the enterprise. The talent team, trademark, corporate culture, corporate customer list, sales channels, etc. in the enterprise are all important resources of the enterprise and must be evaluated for intangible assets. The competition of business is the competition of talents and technology. Therefore, intangible assets occupy an important position in today's business environment, which leads to intangible assets

occupying an important position in today's business environment. The evaluation of intangible assets is a kind of protection and value substantiation for intangible assets. This paper combines the improved fuzzy comprehensive evaluation method to conduct research on goodwill evaluation and intangible asset management. The experimental simulation shows that the improved fuzzy comprehensive evaluation method proposed in this paper has a very good application effect in goodwill evaluation and intangible asset management.

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] I. Podhorska, L. Gajanova, J. Klietkova, and G. H. Popescu, "Analysis of internally generated goodwill indicators: a case study of the Slovak republic," *Organizacija*, vol. 52, no. 4, pp. 271–285, 2019.
- [2] T. Krabec, R. Čížinská, and B. Rýdlová, "Deconstruction of inherent goodwill generated in owner-managed companies in the Czech republic," *International Advances in Economic Research*, vol. 27, no. 3, pp. 253–255, 2021.
- [3] P. M. Johnson, T. J. Lopez, and T. L. Sorensen, "Did SFAS 141/142 improve the market's understanding of net assets, goodwill, or other intangible assets?" *Review of Quantitative Finance and Accounting*, vol. 56, no. 3, pp. 891–915, 2021.
- [4] E. Pechlivanidis, D. Ginoglou, and P. Barmpoutis, "Can intangible assets predict future performance? A deep learning approach," *International Journal of Accounting and Information Management*, vol. 30, no. 1, pp. 61–72, 2021.
- [5] Y. Wang, T. Li, D. Wang, and K. Liu, "Has damage from goodwill impairment grown in China? Analysis and response," *China Journal of Accounting Studies*, vol. 9, no. 2, pp. 168–194, 2021.
- [6] M. Gietzmann and Y. Wang, "Goodwill valuations certified by independent experts: bigger and cleaner impairments?" *Journal of Business Finance & Accounting*, vol. 47, no. 1–2, pp. 27–51, 2020.
- [7] I. Starko, "Normal regulation of the recognition and evaluation of the goodwill," *Review of Accounting Studies*, vol. 17, no. 4, pp. 749–780, 2018.
- [8] M. P. Bauman and K. W. Shaw, "Value relevance of customer-related intangible assets," *Research in Accounting Regulation*, vol. 30, no. 2, pp. 95–102, 2018.
- [9] A. Dugar and J. Pozharny, "Equity investing in the age of intangibles," *Financial Analysts Journal*, vol. 77, no. 2, pp. 21–42, 2021.
- [10] A. A. Ghosh and C. Xing, "Goodwill impairment and audit effort," *Accounting Horizons*, vol. 35, no. 4, pp. 83–103, 2021.
- [11] M. Cordazzo and P. Rossi, "The influence of IFRS mandatory adoption on value relevance of intangible assets in Italy," *Journal of Applied Accounting Research*, vol. 21, no. 3, pp. 415–436, 2020.
- [12] K. V. d. Oliveira, P. R. B. Lustosa, and A. d. O. Gonçalves, "O goodwill na perspectiva da appreciative inquiry (AI): a

- inovação transformando o patrimônio intangível,” *Revista Contemporânea de Contabilidade*, vol. 18, no. 47, pp. 03–17, 2021.
- [13] A. N. Minina and O. Y. Patlasov, “Methodology for brand assessment as an intangible asset in agricultural holdings,” *Science of the Person: Humanitarian Researches*, vol. 15, no. 2, pp. 196–203, 2021.
- [14] E. M. Nichita, “Intangible assets—insights from a literature review,” *Journal of Accounting and Management Information Systems*, vol. 18, no. 2, pp. 224–261, 2019.
- [15] Y. V. Pozdnyakov, S. Zoryana, and G. Tetiana, “Price-forming factors choice grounding at intangible assets with negative depreciation independent valuation/appraising,” *Independent Journal of Management & Production*, vol. 11, no. 6, pp. 2112–2139, 2020.
- [16] M. Kaygusuzoğlu, A. Karahan, and F. Yilmaz, “Implementations on identification and reporting of goodwill impairment (value lowness) in Turkish accounting standarts: a research on the businesses in the istanbul stock exchange,” *Dicle Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, vol. 11, no. 22, pp. 519–542, 2021.
- [17] P. V. C. Okoye, N. Offor, and M. I. Juliana, “Effect of intangible assets on performance of quoted companies in Nigeria,” *International Journal of Innovative Finance and Economics Research*, vol. 7, no. 3, pp. 58–66, 2019.
- [18] L. Karpenko, P. Pashko, P. Voronzhak, H. Kalach, and M. Nazarov, “Formation of the system of fair business practice of the company under conditions of corporate responsibility,” *Academy of Strategic Management Journal*, vol. 18, no. 2, pp. 1–8, 2019.
- [19] J. G. Tchatchou Tchaptchet and O. Colot, “Goodwill’s accounting practices in Belgium and compliance with IAS 36 required disclosures,” *International Business Research*, vol. 12, no. 3, pp. 139–152, 2019.

Research Article

Research on the Coordinated Development of New Rural Production and Living Civilization Construction from the Perspective of Green Transformation and Development

Xiaoyan Li 

Xinyang Vocational and Technical College, Xinyang 464000, China

Correspondence should be addressed to Xiaoyan Li; lixiaoyan@xyvtc.edu.cn

Received 8 July 2022; Revised 19 July 2022; Accepted 25 July 2022; Published 6 September 2022

Academic Editor: Ahmedin M. Ahmed

Copyright © 2022 Xiaoyan Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to improve the collaborative development effect of new rural production and living civilization construction, this article studies the collaborative development of rural production and living civilization construction combined with intelligent data processing algorithms and builds a model from the perspective of green transformation and development. In order to find as many and evenly distributed points as possible on the Pareto interface, a coevolution operator is designed in this article. The Pareto solutions of the two populations exchange information through the action of the co-occurrence operator and the absorption operator, so that the algorithm can find more and better noninferior solutions. The research shows that the collaborative development model of new rural production and living civilization construction has a certain effect, and it has a certain role in promoting the collaborative development of new rural production and living civilization construction.

1. Introduction

System civilization mainly includes economic system civilization and political system civilization. The so-called economic system refers to the economic basis at a certain stage of human history, that is, the sum of production relations [1]. Moreover, it is the foundation of political systems and social ideology. The economic system is the organization and operation mechanism of economic activities carried out by human beings to earn a living, and it is the concrete embodiment of the basic social economic system. Agriculture is the foundation of the development of the national economy, and the establishment of a distinctive socialist market economy system also includes the construction of the rural market economy system [2]. Since 80% of the population is in the countryside, this means that the productive force of the society is the rural productive force. Whether the economy can develop or not depends first on whether the rural areas can develop, and the key to rural development is the awakening of the self-awareness of rural productive forces. That is to say, the value embodiment of

“institutional civilization” construction must first promote the healthy development of rural productive forces [3].

In the research on the value of ecological culture, Chen Shoupeng and Yang Lixin pointed out that the continuous development of society is due to the strong culture as the support and guarantee. As an advanced cultural form, ecological culture provides a strong driving force for the sustainable development of society. It is embodied in the fact that ecological culture provides the society with correct ecological values, ecological world outlook, ecological ethics, ecological science and technology, and ecological aesthetics, which is conducive to improving the civilization of the whole society [4]. Ecological culture is the key to solving the ecological crisis. It can not only provide ideological resources and theoretical references for improving human ecological awareness, but also promote the all-round development of human beings and the transition from modern production and lifestyle to ecological production and lifestyle [5]. While analyzing the value of ecological culture from the perspective of socialist culture construction, Lai Zhangsheng and Hu Xiaoyu believe that ecological culture

should be an important part of characteristic socialist culture, which can enrich the new connotation of characteristic socialist harmonious culture and promote the innovation of characteristic socialist culture [6]. Ecological culture can lay a spiritual foundation for the construction of a socialist harmonious society, provide theoretical guidance for the concept of green development, and lay the groundwork for the construction of a socialist ecological civilization [7].

Rural ecological culture and ecological culture also have a narrow and broad difference. The broad rural ecological culture includes both the rural material production level and the ecological cultural content at the spiritual and institutional level; the narrow rural ecological culture specifically refers to the guidance of ecological values. The formed comprehensive cultural system, which covers rural “good” culture, awe culture, collectivism culture, and other cultural systems, not only pursues a friendly and harmonious system between man and nature, but also pursues the relationship between man and society [8]. From the perspective of the overall rural development, it is believed that rural ecological culture is a cultural form that coordinates various rural development and harmonious coexistence with nature, and the construction of rural ecological culture is one of the important means to ensure the harmonious, healthy, and sustainable development of rural areas in the ideological field [9]. For the problems existing in the development of rural ecological culture, due to the influence of traditional feudal concepts in rural areas, farmers’ ecological awareness is weak, and their ecological cultural concepts are relatively backward. Moreover, the construction of rural ecological culture has just started; various systems and mechanisms are not perfect; and farmers pay too much attention to short-term interests, ignoring the long-term comprehensive social, economic, and ecological benefits of rural construction [10]. The problem of rural ecological culture construction is that the spiritual culture of farmers is insufficient; the brain drain is serious; and the support policies for ecological culture in rural areas are not perfect, lack guarantees, and have weak infrastructure [11]. From the perspective of spiritual, institutional, and material construction, the dilemma of rural ecological culture construction is manifested in the emptiness of farmers’ ecological spiritual culture, the lack of ecological institutional culture, and the lack of ecological material culture [12].

The main way to build rural ecological culture is to change the current rural economic development mode, build an ecological agriculture industry system, continuously improve rural ecological facilities, and spread advanced culture and promote excellent traditional culture [13]. More detailed suggestions are made on the development of rural ecological culture. It is believed that the construction of rural ecological culture should shape and cultivate the traditional rural production and life style ecologically from the material form, and regulate and constrain the rural governments, social organizations, and enterprises at all levels from the institutional form. The majority of farmers spiritually establish the concept of ecological civilization of farmers [14]. The development of ecological culture requires an economic

foundation. The construction of beautiful villages cannot be separated from the emerging economic foundation of rural areas. Second, the government should play a leading role and strengthen the protection of ecosystems [15]. In addition to promoting the transformation of rural economic development and enhancing the quality of farmers, Yang Rongrong believes that it is also necessary to build a good social environment and cultivate a strong social atmosphere. Some scholars have proposed to inherit and protect rural ecological culture and build ecological cultural reserves [16]. Scholars have made suggestions for the government, farmers, carrier construction, etc. The ultimate goal is to better develop rural ecological culture, completely change the unreasonable production and lifestyle in rural areas, and build a pattern of harmonious coexistence between rural people and nature [17].

Eco-Marxism mainly considers the causes of ecological crisis, the solution path, and the idea of ecological socialism. Ecological Marxism believes that the emergence of ecological crisis is due to the concept of human control of nature [18]. Ecological Marxism believes that the origin of ecological crisis lies in the capitalist system. Capitalism is a self-expanding system of economic development. Its purpose is the infinite growth of capital, while nature cannot expand infinitely. The contradiction between nature and capitalism is uncoordinated. Moreover, the ecological environment appears to be the alienation of the relationship between man and nature, but in fact it is the alienation of the practice subject under the capitalist system [19]. The best choice to protect the natural environment is to choose advanced socialism because socialism can overcome the contradiction between the expansion of capital profits and ecology, and realize the renewal of the entire mode of production. The thought of ecological Marxism seeks the insights of Marxism from the perspective of ecology, which has a positive enlightenment significance for the current handling of economic development and environmental protection [20].

This article studies the coordinated development of rural production and living civilization construction in combination with intelligent data processing algorithms, and constructs a model from the perspective of green transformation and development to provide a reference for the subsequent coordinated development of rural production and living civilization construction.

2. Intelligent Agriculture Development Data Analysis Algorithm

2.1. Basic Theory of Evolutionary Algorithms and Basic Concepts of Multiobjective Optimization. The principle of genetic algorithm is essentially based on Darwin’s theory of natural selection. Usually, it is necessary to use a small selection pressure (to explore a certain breadth of the search space) in the early stage of genetic search, and use a large selection pressure (to limit the search space) in the late stage. The direction of the genetic search is directed to some of the more useful regions of the search space. The selection methods usually used are as follows: roulette wheel selection,

$(\mu + \lambda)$ selection, tournament selection, elitist selection, and other forms.

Wheel selection method determines the selection probability of the individual according to the ratio of each chromosome fitness value. The probability of each individual entering the next generation is equal to the ratio of its fitness value to the sum of the fitness values of the individuals in the entire population. The higher the fitness value, the greater the probability of the individual being selected, and the greater the probability of the individual entering the next generation. The specific principle is as follows:

- ① The algorithm calculates the probability $P_i = \text{Fit}(x_i) / \sum_{j=1}^N \text{Fit}(x_j)$, $i = 1, 2, \dots, N$, of the initial selection of each individual in the population, and the sum of the fitness of individuals in the population is $\sum_{j=1}^N \text{Fit}(x_j)$, where $\text{Fit}(x_j)$ is the fitness value of the individual x_j .
- ② The algorithm calculates the cumulative probability $q_i = \sum_{j=1}^i P_j$, $i = 1, 2, \dots, N$ of each individual in the population.
- ③ The algorithm generates a random number $r \in [0, 1]$ according to a uniform distribution.
- ④ If $r \leq q_1$, the algorithm chooses x_1 ; otherwise, if $q_i < r \leq q_{i+1}$, $1 \leq i < N - 1$, the algorithm chooses x_{i+1} . If $k < N$, the algorithm sets $k = k + 1$ and turns to ③; otherwise, the algorithm stops.

As shown in Figure 1, $q_1 = p_1 = 0.45$, $q_2 = P_1 + P_2 = 0.75$, $q_3 = P_1 + P_2 + P_3 = 0.95$, $q_4 = P_1 + P_2 + P_3 + P_4 = 1$, x_1 has the highest probability of being selected into the next generation, and x_4 has the lowest probability of being selected into the next generation.

The algorithm assumes that two individuals x_i^t and x_j^t are arithmetically crossed, and the two new individuals generated after the crossover are as follows:

$$\begin{cases} x_i^{t+1} = \alpha x_j^t + (1 - \alpha)x_i^t, \\ x_j^{t+1} = \alpha x_i^t + (1 - \alpha)x_j^t. \end{cases} \quad (1)$$

Among them, $\alpha \in [0, 1]$ is a random number. In the main operation process of arithmetic crossover, ① the algorithm randomly generates the coefficient α when two individuals are linearly combined; ② the algorithm generates two new individuals according to formula (1).

2.1.1. Mutation Operator. The genetic algorithm solves problems with the help of the principles and ideas of biological evolution. The following is the basic flow chart of genetic algorithm (Figure 2):

The following multiobjective problem is considered.

$$\min_{x \in \Omega} F(x) = (f_1(x), f_2(x), \dots, f_M(x)). \quad (2)$$

Among them, $x = (x_1, x_2, \dots, x_n)$ is an n -dimensional vector, Ω is the feasible solution space, and $f_1(x), f_2(x), \dots, f_M(x)$ is m -objective functions. Generally, for a single-objective optimization problem, the global

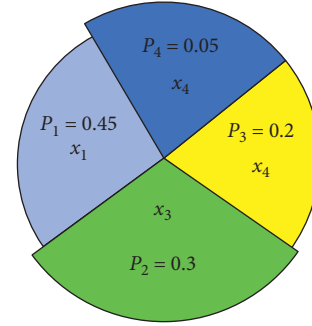


FIGURE 1: Wheel selection method.

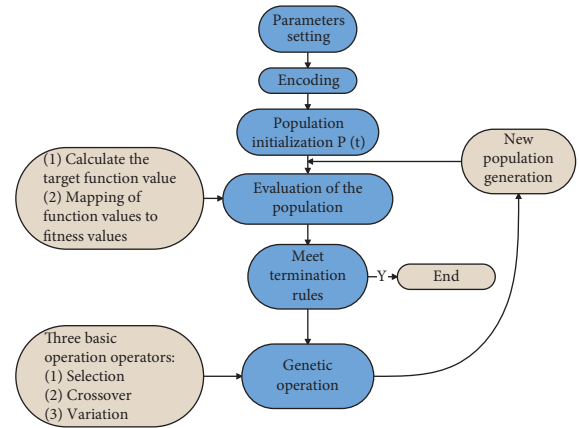


FIGURE 2: Flowchart of genetic algorithm.

optimal solution is the solution that makes the objective function optimal. However, for a multiobjective optimization problem such as formula (2), usually, the functions to be optimized of $f_1(x), f_2(x), \dots, f_M(x)$ are in conflict with each other, and they cannot achieve the desired optimal value at the same time. Scholars thought of using Pareto optimal solution to describe the solution of this optimization problem, in which multiple objective functions have conflicting functions. Pareto optimal solution is a commonly used definition of multiobjective optimization. The following basic definitions are now given:

Definition 1. A vector $u = (u_1, u_2, \dots, u_M)$ is said to be noninferior to another vector $v = (v_1, v_2, \dots, v_M)$, if and only if $\forall i \in \{1, 2, \dots, M\}$ has $u_i \leq v_i$ and $\exists j \in \{1, 2, \dots, M\}$ to makes $u_j < v_j$.

Definition 2. In the Pareto optimal solution, for any two points $x_1, x_2 \in \Omega$, if the following conditions hold:

$$\begin{cases} f_i(x_1) \leq f_i(x_2), \forall i \in \{1, 2, \dots, M\}, \\ f_j(x_1) < f_j(x_2), \exists j \in \{1, 2, \dots, M\}. \end{cases} \quad (3)$$

If vector $(f_1(x_1), f_2(x_1), \dots, f_M(x_1))$ is better than vector $(f_1(x_2), f_2(x_2), \dots, f_M(x_2))$, then x_1 is said to be better than x_2 . It can be seen from the first condition of formula (3) that for M targets, x_1 is no worse than x_2 . It can be seen from the second condition that x_1 is better than x_2 in

at least one objective. Therefore, x_1 should indeed be better than x_2 when formula (3) is satisfied.

Definition 3. If there is no better solution than x_1 , in Ω , then x_1 , is said to be a Pareto optimal solution (or an efficient solution, or a noninferior solution) of formula. (2).

Definition 4. In the Pareto optimal solution set, for formula (2), the set $E(f, \Omega) = \{x \in \Omega \mid x \text{ is called the Pareto optimal solution (or effective solution set or noninferior solution set) of formula (2).}$

Definition 5. In the Pareto frontier, for formula (2), the set $FE(f, \Omega) = \{(f_1(x), f_2(x), \dots, f_M(x)) \mid x \in E(f, \Omega)\}$, (4) is called the Pareto frontier (or effective frontier) of formula (2).

Figure 3 shows the Pareto optimal solution set and the corresponding Pareto front for a problem with two-objective functions as independent variables.

Ω is the decision variable space, where

$$Y = \{y \in R^M \mid y = (f_1(x), f_2(x), \dots, f_M(x)), x \in \Omega\}, \quad (5)$$

is the objective function space.

The optimal solution of a multiobjective optimization problem is always on the boundary line (or surface) of the search area. As shown in Figure 4, the thick line segment represents the optimal interface (Pareto frontier) of the two-objective optimization. The optimal interface of three-objective optimization is a surface, and the optimal interface of three or more objectives constitutes a hypersurface. The solid points A, B, C, D, and E in Figure 4 are all on the optimal boundary, and they are all Pareto solutions, and they are nondominated with each other. However, the hollow points F, G, H, I, and J are not on the optimal boundary within the search area, so they are not optimal solutions, but dominated. Moreover, it is directly or indirectly governed by the optimal solution above the optimal boundary.

The commonly used methods for solving multiobjective optimization problems are as follows: weighted sum method, constraint method, objective programming method, max-min method, and so on.

(1) *Weighted Sum Method.* The weighted sum method can be expressed as follows: for each subobjective function $f_i(x)$, ($i = 1, 2, \dots, M$), a different weight $\omega_i \geq 0$ ($i = 1, 2, \dots, M$) is assigned as the coefficient of each objective function, and $\sum_{i=1}^M \omega_i = 1$. The linear weighted sum of each subobjective function is expressed as follows:

$$\begin{cases} \min & f(x) = \sum_{i=1}^M \omega_i f_i(x), \\ \text{s.t.} & x \in \Omega. \end{cases} \quad (6)$$

Thousands of different Pareto optimal solutions are obtained by choosing thousands of different weight combinations. This method is extremely simple, but it is also a more classical method for solving multiobjective optimization

problems. Its advantage is that it is well understood and easy to calculate; its disadvantage is that the selection of weights is related to the relative importance of each target. In addition, the weighted sum method is sensitive to the shape of the Pareto interface, and the optimal solution for the convex part of the Pareto interface cannot be obtained.

(2) *ε -Constrained Method.* The main idea of this method is to use the important objective function of multiobjective optimization as the optimization objective, while other objective functions are used as constraint functions to solve. The optimization problem transformed by this constraint method is expressed as follows:

$$P_k(\varepsilon_k) \begin{cases} \min & f_k(x), \\ \text{s.t.} & f_j(x) \leq \varepsilon_j^k, x \in \Omega, 1, 1 \leq j \leq k. \end{cases} \quad (7)$$

The advantage of this method is that it is easy to operate, where different values of ε_j are selected during the optimization process, and multiple Pareto optimal solutions can be found, which is very practical in practical problems. However, the disadvantage is that certain prior knowledge is usually required to select a suitable ε_j , and this prior knowledge is usually unknown.

There are many types of multiobjective evolutionary algorithms (MOE), and the methods they use are also very different, which is difficult to describe by a general framework. For the convenience of understanding, we design a flowchart of a multiobjective evolutionary algorithm based on Pareto optimality (Figure 5).

It assumes that there are n factors, each with q levels. When n and q are given, the purpose of uniform design is to find q combinations from all q^n combinations so that the q combinations are uniformly distributed in all possible combination spaces.

$$U(n, q) = [U_{i,j}]_{q \times n}, \quad (8)$$

It means that the selected q combinations are uniformly distributed, where $U_{i,j}$ represents the level of the j -th factor in the i -th combination. When q is prime and $q > n$, it has been shown that $U_{i,j}$ can be calculated as follows:

$$U_{i,j} = (i\sigma^{j-1} \bmod q) + 1. \quad (9)$$

The values of σ are listed in Table 1.

2.2. *A Multiobjective Coevolutionary Algorithm Based on Uniform Design.* In multiobjective optimization, each objective function value has different orders of magnitude. If only the simple weighted summation of the objective function is used as the fitness function, then the value of the fitness function will be dominated by the one with the larger order of magnitude of the objective function value. For example, for two-objective optimization, $f_1(x)$ represents cost, and its value range is $300 \leq f_1(x) \leq 500$, $f_2(x)$ represents reliability, and its value range is $0 \leq f_2(x) \leq 1$. Then, the value of $\omega_1 f_1(x) + \omega_2 f_2(x)$ will be dominated by $f_1(x)$. To avoid similar situations, we normalize the objective function as follows:

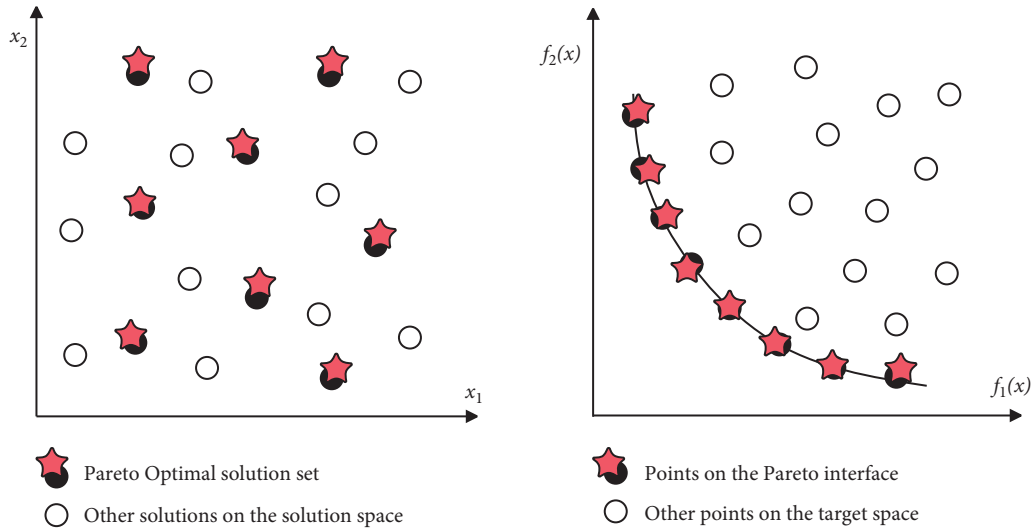


FIGURE 3: Pareto optimal solution in the solution space and Pareto optimal solution in the target space in two-objective planning.

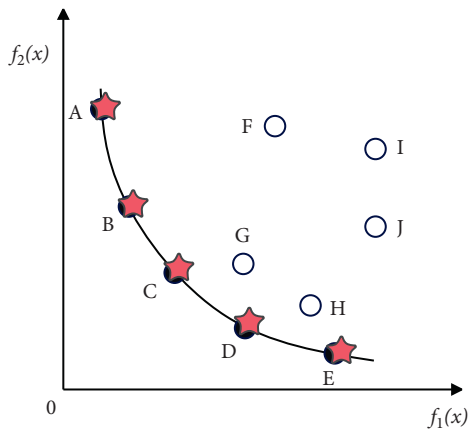


FIGURE 4: Pareto frontier of two-objective optimization.

$$h_i(x) = \frac{f_1(x)}{\max_{y \in \Omega} \{|f_1(y)|\}}. \quad (10)$$

Among them, Ω is the point set of the current population, and $h_i(x)$ is the normalized objective function.

For a given N , the constructed N fitness functions are as follows:

$$\text{fitness}_i = \omega_{i,1}h_1(x) + \omega_{i,2}h_2(x) + \dots + \omega_{i,M}h_M(x), 1 \leq i \leq N. \quad (11)$$

Among them, $\omega_i = (\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,M})$ is the weight vector, and $\omega_{i,j} \geq 0, j = 1, 2, \dots, M, \sum_{j=1}^M \omega_{i,j} = 1$.

In order to make the search approach to the Pareto front evenly, a uniform design method is used to construct the weight vector. In the target space, considering each target as a factor, there are M factors to construct N fitness functions, then N weight vectors are needed, and each factor takes N

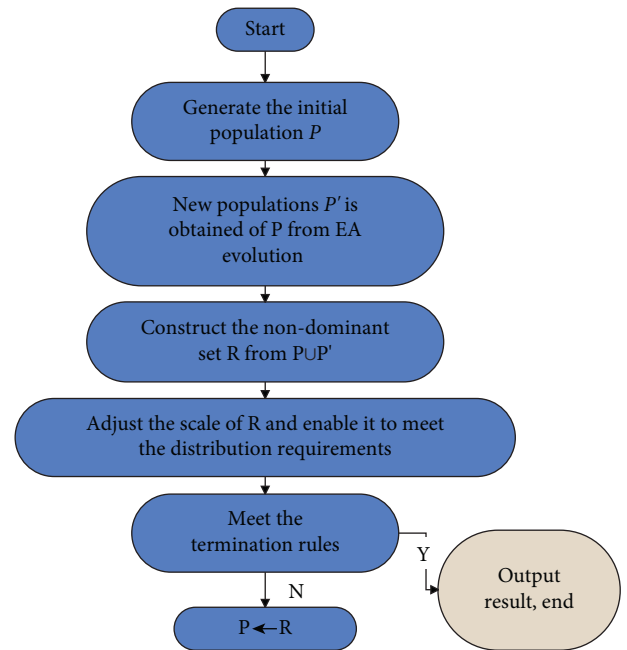


FIGURE 5: Basic flow chart of multiobjective evolutionary algorithm (MOEA).

levels. Then, using the uniform array $U(M, N)$, $\omega_{i,j} (i = 1, 2, \dots, N; j = 1, 2, \dots, M)$ is calculated as follows:

$$\omega_{i,j} = \frac{U_{i,j}}{U_{i,1} + U_{i,2} + \dots + U_{i,M}}. \quad (12)$$

Example 1. The $M = 3, q = 5$ means that there are 3 objective functions, and 5 uniformly distributed weight vectors are to be generated, that is, a uniform array with 3 factors and 5 levels is constructed. It can be seen from Table 1 that $\sigma = 2$, and the uniform array is generated by formula (6)

TABLE 1: The values of σ .

Level of each factor	Value for each factor	σ
5	2.00-4.00	2.00
7	2.00-6.00	3.00
11	2.00-10.00	7.00
13	2.00	5.00
	3.00	4.00
	4.00-12.00	6.00
17	2.00-16.00	10.00
19	2.00-3.00	8.00
	4.00-18.00	14.00
23	2.00,13.00-14.00, 20.00-22.00	7.00
	8.00-12.00	15.00
	3.00-9.00,1 5.00-19.00	17.00
29	2.00	12.00
	3.00	9.00
	4.00-7.00	16.00
	8.00-12.00, 16.00-34.00	8.00
	13.00-15.	14.00
31	25.00-28.00	18.00
	2.0,5.00-12.00, 20.00-30.00	12.00
	3.00-4.0, 13.00-19.00	22.00

Then, according to formula (3), 5 weight vectors are generated, and then formula (3) is used to construct the corresponding fitness function:

$l = (l_1, l_2, \dots, l_N), u = (u_1, u_2, \dots, u_N)$ Among them, $[l, u]$ is the feasible region, if $x = (x_1, x_2, \dots, x_N) \in [l, u]$, x is called a feasible solution of the problem. $[l, u]$ is divided into H different subspaces $[l(1), u(1)], [l(2), u(2)], \dots, [l(H), u(H)]$, and the parameter H can take the value 2, 2^2 , or 2^3 and so on.

The solution interval is divided into two subspaces in the following manner. First, the component with the largest value range is selected, then the solution space is equally divided into two subspaces along the direction of this component, and then the two subspaces are divided into four subspaces according to this method. For any subspace, such as $[l(1), u(1)]$, the coordinate axis with the largest value range is selected, and the two subspaces are divided into four subspaces along this coordinate direction. The above process is repeated until the solution space is divided into H subspaces. The following is the specific algorithm.

Among them, $k = (j_1 - 1)n_2n_3 \dots n_N + (j_2 - 1)n_3n_4 \dots n_N + \dots + (j_N - 1)n_N + j_N$.

After the algorithm divides the solution space into H subspaces, the sample points are selected for each subspace according to the following methods. The algorithm considers an arbitrary subspace, such as the k -th subspace, and denotes it as follows:

$$[l(k), u(k)] = [(l_1(k), l_2(k), \dots, l_N(k)), (u_1(k), u_2(k), \dots, u_N(k))]. \quad (13)$$

In this subspace, the algorithm discretizes the value range of x_i into Q levels $\alpha_{i,1}(k), \alpha_{i,2}(k), \dots, \alpha_{i,Q}(k)$, where Q is a parameter and is a prime number, and $a_{i,j}(k)$ is calculated according to the following formula:

$$a_{i,j}(k) = \begin{cases} l_1(k), & j = 1, \\ l_1(k) + (j-1) \frac{u_i(k) - l_i(k)}{Q-1}, & 2 \leq j \leq Q-1, \\ u_1(k), & j = Q. \end{cases} \quad (14)$$

According to Algorithm 1, it can be known that the difference between any two adjacent levels is the same. It is recorded as $\alpha_i(k) = (\alpha_{i,1}(k), \alpha_{i,2}(k), \dots, \alpha_{i,Q}(k))$, after discretizing the value area of x_i into Q levels, there will be Q^N points in each subspace. Then, a uniform array $U(N, Q)$ is used to select Q sample points in each subspace, denoted as follows:

$$\begin{cases} (\alpha_{1,U_{1,1}}(k), \alpha_{2,U_{1,2}}(k), \dots, \alpha_{N,U_{1,N}}(k)), \\ (\alpha_{1,U_{2,1}}(k), \alpha_{2,U_{2,2}}(k), \dots, \alpha_{N,U_{2,N}}(k)), \\ \dots \\ (\alpha_{1,U_{Q,1}}(k), \alpha_{2,U_{Q,2}}(k), \dots, \alpha_{N,U_{Q,N}}(k)). \end{cases} \quad (15)$$

The algorithm performs the above operations on all H subspaces, with a total of HQ sample points.

Among the HQ sample points, pop points are selected as the initial population. To search in D directions, D fitness functions are constructed, where D is a parameter and a prime number. Through each fitness function, all HQ points are evaluated to select the best $\lfloor \text{pop}/D \rfloor$, and finally a total of pop points are selected to form an initial group. The specific algorithm is as follows.

Example 2. The algorithm considers a two-dimensional solution space: the algorithm sets $1 \leq x_1 \leq 4, -20 \leq x_2 \leq 10$ and its feasible solution area $[l, u]$ is $[(1, -20)(4, 10)]$. The parameters are set as follows: $H = 4, Q = 5, D = 6, \text{pop} = 10$. Algorithm 1 is executed to divide the solution space into the following four subspaces, and Algorithm 2 is executed to generate the initial population:

- (1) According to Algorithm 1, there are $a = (1, -20), z = (4, 10)$.
- (2) Among them, $\Delta_1 = 3, \Delta_2 = 7.5, n_1 = 1, n_2 = 4$, and the following four subspaces are obtained after division:

$$\begin{cases} [l(1), u(1)] = [(1, 20), (4, -12.5)], \\ [l(2), u(2)] = [(1, -12.5), (4, -5)], \\ [l(3), u(3)] = [(1, -5), (4, 23, 5)], \\ [l(4), u(4)] = [(1, 20), (4, -12.5)]. \end{cases} \quad (16)$$

Next, Algorithm 2 is performed to generate the initial population.

- (3) The result obtained by discretizing the first subspace $[l(1), u(1)] = [(1, 20), (4, -12.5)]$ according to formula (11) is as follows:

- (1) The algorithm is set to $a = l, z = u$, and the following process is repeated $\log_2 H$ times: the h -th dimension is selected in the solution algorithm space to satisfy $z_h - a_h = \max_{1 \leq i \leq N} \{z_i, a_i\}$, and $z_h = (a_h + z_h)/2$ is set.
- (2) The algorithm calculates $\Delta_i = z_i - a_i, n_i = ((u_i - l_i)/\Delta_i), i = 1, 2, \dots, N$, and calculates the feasible region $[l(k), u(k)], 1 \leq j_i \leq n_i, 1 \leq i \leq N$ of the new subspace according to the following formula:

$$\begin{cases} l(k) = l + ((j_1 - 1)\Delta_1, (j_2 - 1)\Delta_2, \dots, (j_N - 1)\Delta_N), \\ u(k) = l + (j_1\Delta_1, j_2\Delta_2, \dots, j_N\Delta_N). \end{cases}$$

ALGORITHM 1: The algorithm divides the solution space into H subspaces.

- (1) Algorithm 1 is executed to divide the feasible solution region $[l(1), u(1)], [l(2), u(2)], \dots, [l(H), u(H)]$ into H subspaces, denoted as:
- (2) The algorithm discretizes each subspace according to the formula (11), and then selects Q sample points in each subspace according to the uniform array $U(N, Q)$ according to the formula (3) and generates a total of HQ sample points.
- (3) The algorithm evaluates the HQ sample points generated in step (2) according to each fitness function to select the best $\lfloor \text{pop}/D \rfloor$, and finally selects a total of pop points to form the initial group.

ALGORITHM 2: The algorithm generates the initial population.

$$\begin{cases} \alpha_1(1) = (1, 1.75, 2.5, 3, 25, 4), \\ \alpha_2(1) = (-20, -10.125, -16 - 23, -14, 375, -12.5). \end{cases} \quad (17)$$

Then, according to the uniform array $U = (2, 5) = \begin{bmatrix} 2 & 3 & 4 & 5 & 1 \\ 3 & 5 & 2 & 4 & 1 \end{bmatrix}$ and formula (3), five sample points are selected as follows:

$$\begin{cases} (1.75, -16.5), \\ (2.5, -12.5), \\ (3.25, -18.125), \\ (4, -14.375), \\ (1, -20). \end{cases} \quad (18)$$

The algorithm performs similar operations on the other three subspaces, and a total of 20 sample points are obtained.

- (4) The algorithm evaluates the 20 sample points according to the five fitness functions, each fitness function selects $\lfloor 20/10 \rfloor = 2$ best points, and a total of 10 sample points are selected from the five fitness functions to form the initial group.

2.3. Genetic Operator. Symbiotic operators $P_t^a = (P_1^a, P_2^a, \dots, P_M^a)$ and $P_t^b = (P_1^b, P_2^b, \dots, P_M^b)$ are two parent groups, and the corresponding nondominated solution sets are denoted as O_t^a and O_t^b , respectively. For any $(x_1^a, x_2^a, \dots, x_n^a)$ in the population P_t^a , an individual $(o_1^b, o_2^b, \dots, o_n^b)$ is randomly selected in O_t^b . Afterwards, a new individual is generated according to the following formula, denoted as $(z_1^a, z_2^a, \dots, z_n^a)$, and the next generation group product is generated. Among them, r represents a uniform number randomly distributed in $[-1, 1]$.

$$z_i^a = o_i^b + r \cdot (o_i^b - x_i^a), \quad i = 1, 2, \dots, n. \quad (19)$$

In the same way, for any individual $(x_1^b, x_2^b, \dots, x_n^b)$ in the group P_t^b , the algorithm randomly selects an individual $(o_1^a, o_2^a, \dots, o_n^a)$ in O_t^a and generates a new individual $(z_1^b, z_2^b, \dots, z_n^b)$ according to the following formula, and then, the next generation of group P_{t+1}^a is generated.

$$z_i^b = o_i^a + r \cdot (o_i^a - x_i^b), \quad i = 1, 2, \dots, n. \quad (20)$$

P_t^a and P_t^b are two groups that have evolved separately, that is, search in different spatial regions in the variable space. By the function of formulas (13) and (14), the two groups exchange information with each other.

Attraction operators $P_t^a = (X_1, X_2, \dots, X_M)$ and $P_t^b = (Y_1, Y_2, \dots, Y_M)$ are two parent groups, and the corresponding nondominated solution sets are denoted as O_t^a and O_t^b , respectively. If

$$\forall X \in O_t^a, \exists Y \in O_t^b, Y > X, \quad (21)$$

is established, then the group P_t^b absorbs the group P_t^a to generate a new group $P_{t+1}^b(l_1, l_2, \dots, l_{N+M})$. Among them, $l_i = Y_i, i = 1, 2, \dots, N$, and $l_i, i = N + 1, N + 2, \dots, N + M$ is generated by:

$$l_{i,j} = o_j + r \cdot (o_j - x_{i-N,j}), \quad j = 1, 2, \dots, n. \quad (22)$$

From the above expression, we can know that O_t^b is better than O_t^a , that is, the solution in group P_t^b is better than the solution in group P_t^a . In this case, it is much less likely that the quality of the solution obtained by evolving population P_t^a is better than the quality of the solution obtained by evolving population P_t^b . Therefore, in this competition, group P_t^a will inevitably disappear. In addition, there may be some relatively useful information in the group P_t^a , which can be utilized by (16) to generate better or useful individuals.

Discrete operator: the population $P_t^a = (P_1^a, P_2^a, \dots, P_M^a)$ is randomly discretized into two subpopulations, denoted as $P_t^{ia} = (Z_1^a, Z_2^a, \dots, Z_{M/2}^a)$ and $P_t^{ib} = (Z_1^b, Z_2^b, \dots, Z_{M/2}^b)$, respectively. A population is randomly selected, and the algorithm takes P_t^{ia} and performs the following mutation operations on each individual in P_t^{ib} :

$$z_k^a = \begin{cases} z_k^a + \frac{(u_k^a - l_k^a)}{t} \cdot rr \leq 0.5, \\ z_k^a - \frac{(u_k^a - l_k^a)}{t} \cdot \text{otherwise,} \end{cases} \quad k = 1, 2, \dots, n. \quad (23)$$

Among them, r represents the uniform number randomly distributed in the $[0, 1]$ interval, and t represents the evolutionary algebra. u_k^a and l_k^a are the upper and lower bounds of the k -th coordinate of the point in the search space, respectively. The purpose of doing this is that there may be some more useful information in the group P_t^a . After the group P_t^a is acted on by a discrete operator, a new individual is obtained and the diversity of the group is increased. It is also possible to obtain some solutions that are better than the current population, or to find solutions for nonconvex parts of the Pareto interface.

3. Research on the Coordinated Development of New Rural Production and Living Civilization Construction from the Perspective of Green Transformation and Development

The energy system diagram reflects the basic structure inside each system and the relationship between energy and matter inside and outside the system. The level of the energy conversion rate of each legend and its respective components determine the arrangement order of the legends inside and outside the boundary of the system diagram, and the energy conversion rate decreases from right to left, as shown in Figure 6.

4. Simulation Test Research

This article proposes a method for calculating the degree of distribution, which is used to measure whether the Pareto optimal solution set obtained by the algorithm is uniformly distributed in the target space. The function is defined as follows:

$$S = \sqrt{\frac{1}{|Q|-1} \sum_{i=1}^{|Q|} (d_i - \bar{d})^2}. \quad (24)$$

Among them,

$$d_1 = \min_{j \in Q \wedge j \neq 1} \sum_{k=1}^m |f_k^i - f_k^j|, \bar{d} = \left(\frac{\sum_{i=1}^{|Q|} d_i}{|Q|}, i, j \in Q \right), \quad (25)$$

where m is the number of objective functions, and Q is the set of nondominated solutions obtained by the algorithm. $|Q|$ represents the number of elements in the nondominated

solution set "Q." The value of S is the S metric. When the solutions in the set Q tend to be uniformly distributed, the smaller the value of the corresponding distance evaluation function S is, the more uniformly the noninferior solution set obtained by the algorithm is distributed in the target space.

The measure of broadness (maximum spread) mainly compares the performance of the algorithm by calculating the perimeter of the polyhedron formed by all extreme values in the solution set P , which is defined as follows:

$$M = \sqrt{\sum_{k=1}^m (\max_i^N f_k^i - \min_i^N f_k^i)^2}. \quad (26)$$

Among them, N is the number of nondominated solutions, and m is the dimension of the target space. The larger the value of M , the wider the range of the solution set P , and the value of M is called the M measure.

To verify the effectiveness of the proposed new algorithm, the following test functions are selected.

$$\begin{cases} \text{questionone} \left\{ \begin{array}{l} \min f_1(x) = \frac{1}{x_1^2 + x_2^2 + 1}, \\ \min f_2(x) = x_1^2 + 3x_2^2, \quad + 1 \\ \text{subject to } -3 \leq x_1 \leq 3, \\ -5 \leq x_2 \leq 5, \end{array} \right. \\ \\ \text{questiontwo} \left\{ \begin{array}{l} \min f_1(x) = 2\sqrt{x_1}, \\ \min f_2(x) = x_1(1 - x_2) + 5, \\ \text{subject to } 1 \leq x_1 \leq 4, \\ 1 \leq x_2 \leq 2, \end{array} \right. \\ \\ \text{questionthree} \left\{ \begin{array}{l} \min f_1(x) = x_1^2 + x_2^2, \\ \min f_2(x) = (x_1 - 5)^2 + (x_2 - 5)^2, \\ \text{subject to } -5 \leq x_1 \leq 10, \\ -5 \leq x_2 \leq 10, \end{array} \right. \\ \\ \text{questionfour} \left\{ \begin{array}{l} \min f_1(x) = 1 - \exp(-(x_1 - 1)^2 - (x_2 + 1)^2), \\ \min f_2(x) = 1 - \exp(-(x_1 + 1)^2 - (x_2 - 1)^2), \\ \text{subject to } -10 \leq x_1 \leq 10, \\ -10 \leq x_2 \leq 10. \end{array} \right. \end{cases} \quad (27)$$

Figures 7 to 10 visually compare the nondominated solution sets obtained by the algorithm UCEMOA in this article and the classical algorithm NSGA-II in a certain run.

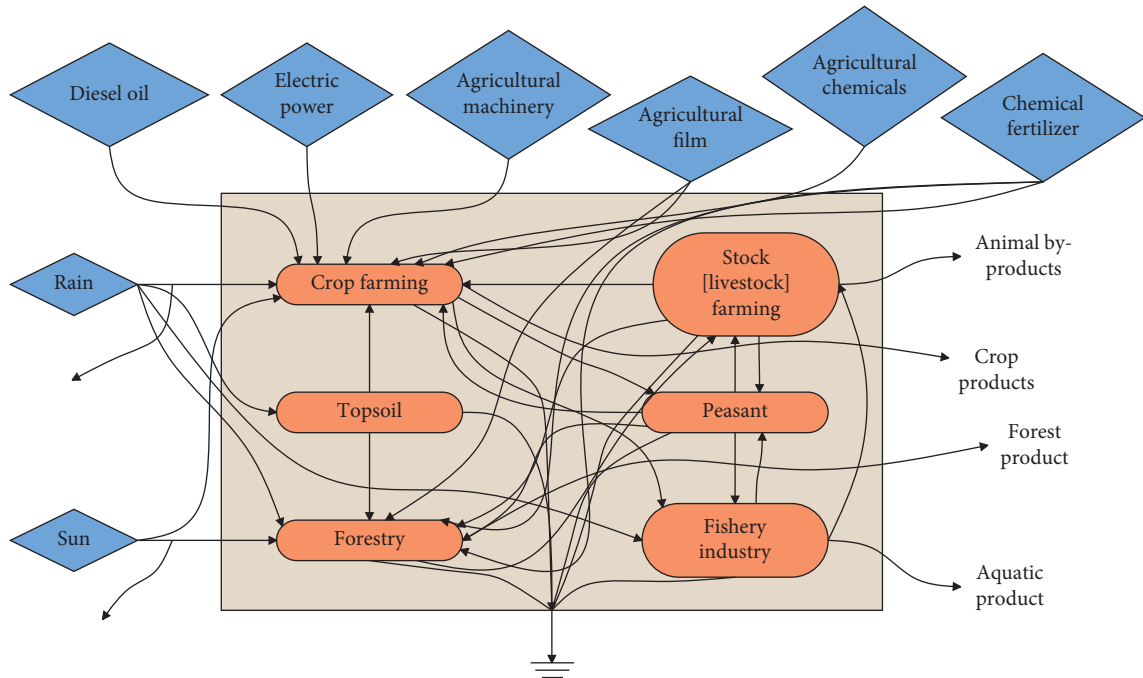


FIGURE 6: Schematic diagram of energy flow in the agro-ecosystem.

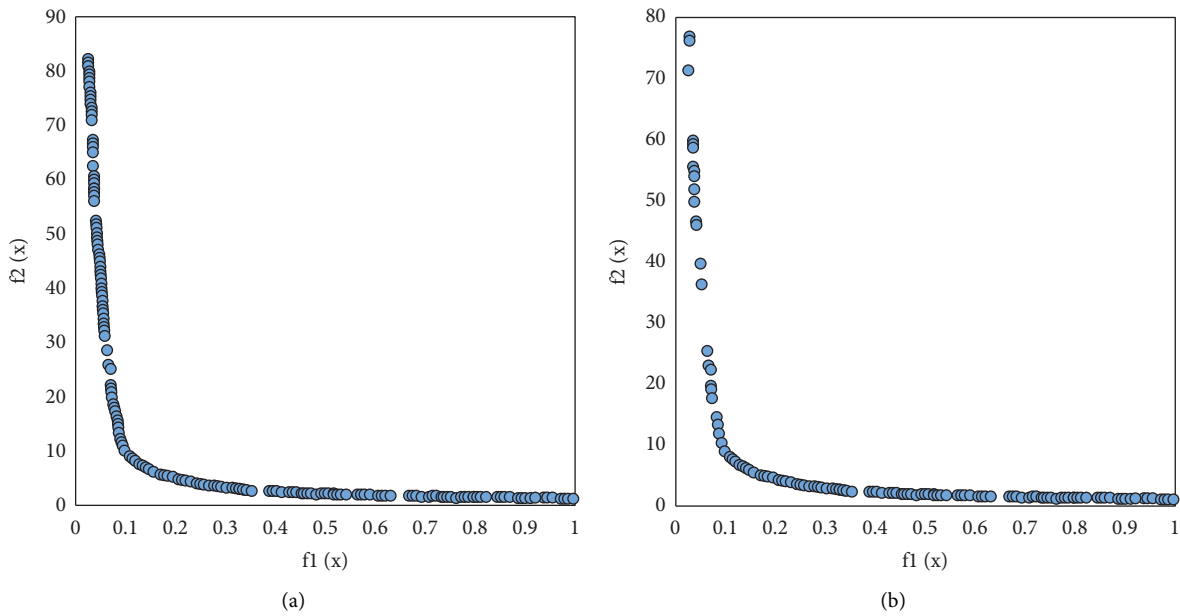


FIGURE 7: Simulation results for problem 1. (a) The running result of UCEMOA on problem 1. (b) The running results of NSGA-II on problem 1.

It can be seen from Figures 6-8 that the algorithm UCEMOA obtains more noninferior solutions than NSGA-II. The following intervals are for the horizontal axis unless otherwise specified. It can be seen from 5 that between the interval [0, 0.1] and [0.6, 0.9], the algorithm UCEMOA in this article obtains more and uniformly distributed

noninferior solutions than NSGA-II. It can be seen from Figure 6 that for the nonconvex Pareto optimal solution interface, the algorithm UCEMOA searches for more uniformly distributed and broad noninferior solutions than NSGA-II. It can be seen from Figure 7 that in the interval [0, 4], the algorithm UCEMOA in this article obtains more

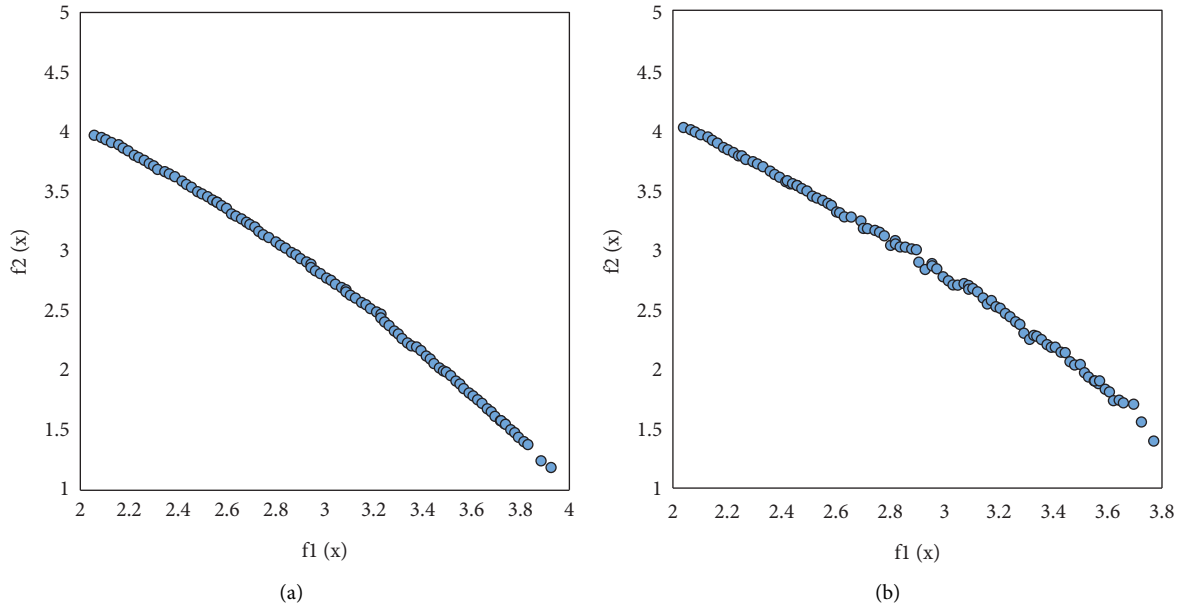


FIGURE 8: Simulation results for problem 2. (a) The running result of UCEMOA on problem 2. (b) The running results of NSGA-II on problem 2.

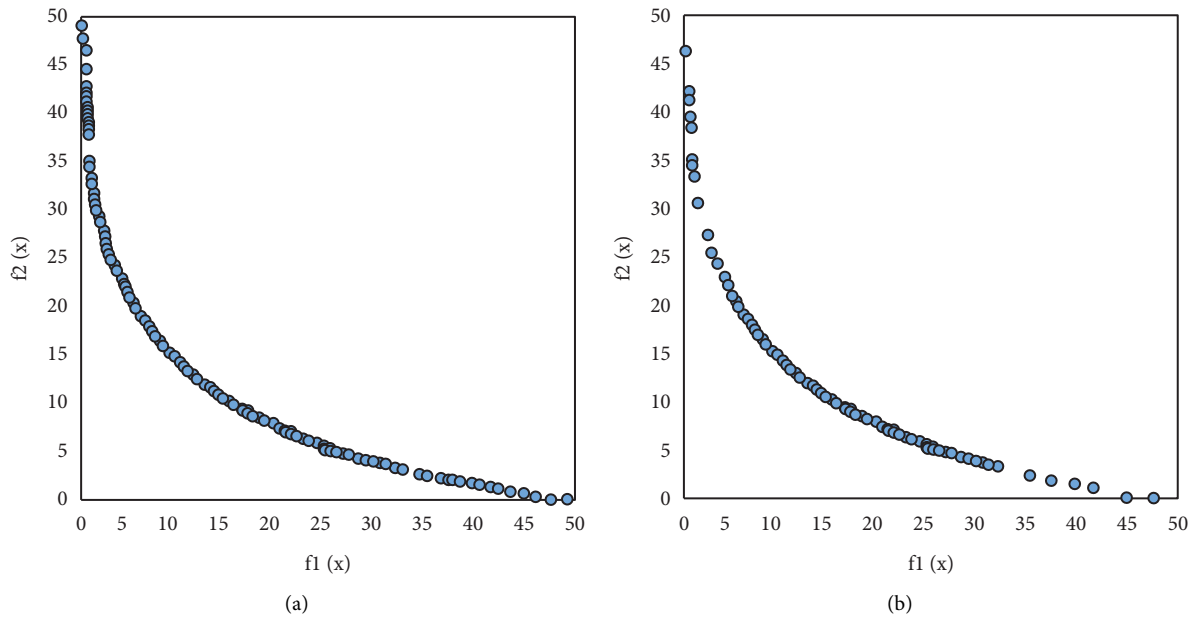


FIGURE 9: Simulation results for problem 3. (a) The running result of UCEMOA on problem 3. (b) The running results of NSGA-II on problem 3.

uniformly distributed and broad noninferior solutions than NSGA-II. In the interval $[30, 50]$, our algorithm UCEMOA searches more noninferior solutions than NSGA-II. It can be seen from Figure 8 that in the interval $[0, 0.5]$, the algorithm UCEMOA in this article obtains more noninferior solutions than NSGA-II. On the vertical axis, in the interval $[0, 0.4]$, the algorithm UCEMOA in this article searches for more and uniformly distributed noninferior solutions than NSGA-II. Therefore, the algorithm UCEMOA can effectively search for the Pareto optimal solution, not only for the convex Pareto optimal

solution interface problem, but also for the nonconvex Pareto optimal solution interface problem.

Based on the above analysis, the effect of the collaborative development model of new rural production and life civilization construction proposed in this article is evaluated, and the experimental results shown in Figure 11 are obtained.

It can be seen from the above research that the collaborative development model of new rural production and living civilization construction has a certain

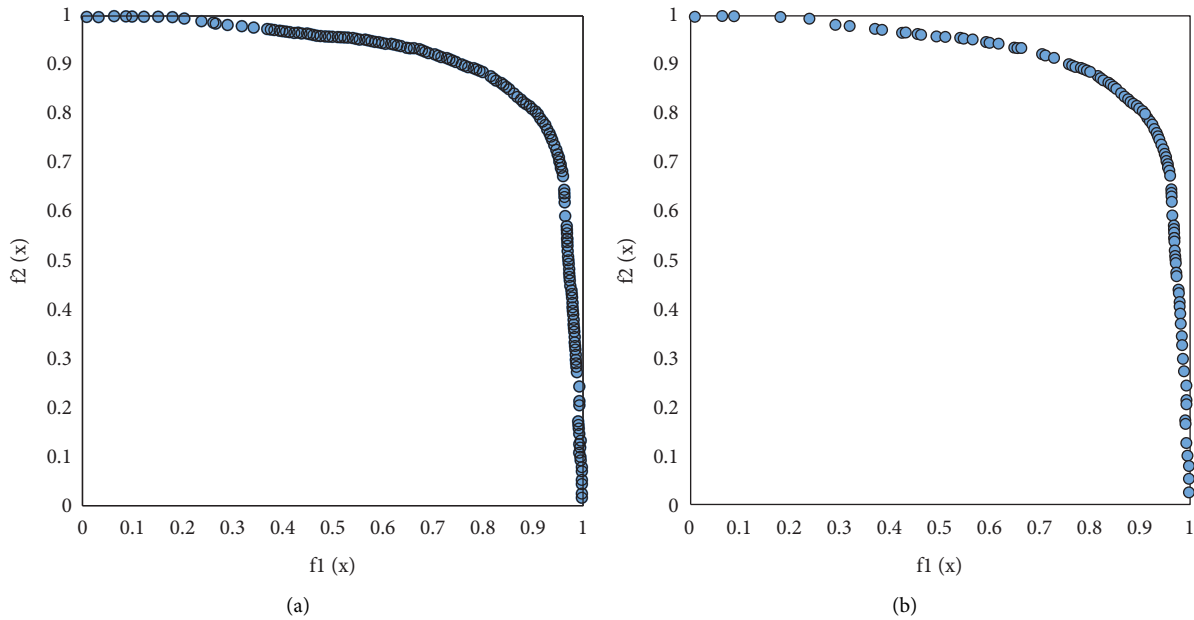


FIGURE 10: Simulation results for problem 4. (a) The running result of UCEMOA on problem 4. (b) The running results of NSGA-II on problem 4.

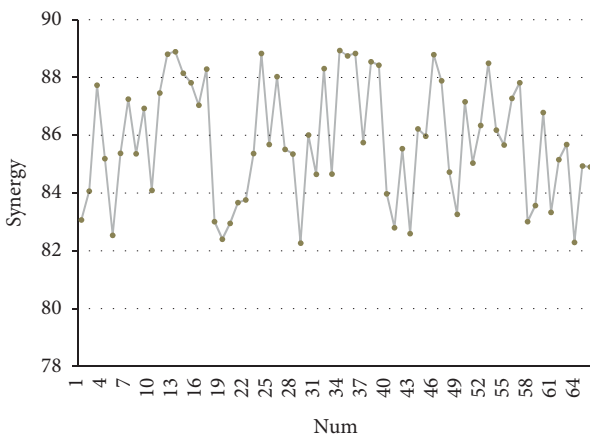


FIGURE 11: Evaluation of the effect of the collaborative development model of new rural production and life civilization construction.

effect, and has a certain role in promoting the collaborative development of new rural production and living civilization construction.

5. Conclusion

Scientific research is always carried out under certain pre-conditions. The study of rural revitalization and the construction of rural civilization also has its own profound motivation. From the perspective of rural revitalization, the construction of rural-style civilization contains major theoretical problems faced by contemporary society, and the construction of rural-style civilization from the perspective of rural revitalization is a major academic issue worthy of in-depth study. Moreover, the construction of rural civilization

from the perspective of rural revitalization is a major practical problem faced by contemporary society. In addition, the problem of rural development is a worldwide problem with universal significance. Therefore, being familiar with the research on the development of the country's rural society in foreign academic circles undoubtedly has important comparison and reference value for the writing of this article. This article studies the coordinated development of rural production and living civilization construction combined with intelligent data processing algorithms, and builds a model from the perspective of green transformation and development. The research shows that the collaborative development model of new rural production and living civilization construction has a certain effect, and has a certain role in promoting the collaborative development of new rural production and living civilization construction.

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This work was supported by Xinyang Vocational and Technical College.

References

[1] A. Mdee, A. Ofori, M. Chasukwa, and S. Manda, "Neither sustainable nor inclusive: a political economy of agricultural

- policy and livelihoods in Malawi, Tanzania and Zambia,” *Journal of Peasant Studies*, vol. 48, no. 6, pp. 1260–1283, 2021.
- [2] N. Gontard, U. Sonesson, M. Birkved et al., “A research challenge vision regarding management of agricultural waste in a circular bio-based economy,” *Critical Reviews in Environmental Science and Technology*, vol. 48, no. 6, pp. 614–654, 2018.
 - [3] A. Sakhno, I. Salkova, A. Broyaka, and N. Priamukhina, “A methodological analysis for the impact assessment of the digitalisation of economy on agricultural growth,” *International Journal of Advanced Science and Technology*, vol. 29, no. 8s, pp. 242–249, 2020.
 - [4] S. Fortunati, D. Morea, and E. M. Mosconi, “Circular economy and corporate social responsibility in the agricultural system: cases study of the Italian agri-food industry,” *Agricultural Economics*, vol. 66, no. 11, pp. 489–498, 2020.
 - [5] J. Hua and D. Wang, “Research on relationship between agricultural water and agricultural economy based on growth drag of water resources,” *Acta Agriculturae Jiangxi*, vol. 30, no. 6, pp. 129–139, 2018.
 - [6] Y. O. Lupenko, A. O. Gutorov, and O. I. Gutorov, “Investment ensuring for development of integration relations in the agricultural sector of Ukrainian economy,” *Financial and credit activity: Problems of Theory and Practice*, vol. 4, no. 27, pp. 381–389, 2018.
 - [7] V. Trukhachev, A. Bobrishev, E. Khokhlova, V. Ivashova, and O. Fedisko, “Personnel training for the agricultural sector in terms of digital transformation of the economy: trends, prospects and limitations,” *International Journal of Civil Engineering & Technology*, vol. 10, no. 1, pp. 2145–2155, 2019.
 - [8] O. Radchenko, M. Matveyeva, H. Holovanova, K. Makhyboroda, and Y. Haibura, “Information and analytical provision of budget support of institutional sectors of the economy (on the example of the agricultural sector of Ukraine),” *Independent Journal of Management & Production*, vol. 11, no. 9, pp. 2355–2378, 2020.
 - [9] V. Astolfi, A. L. Astolfi, M. A. Mazutti et al., “Cellulolytic enzyme production from agricultural residues for biofuel purpose on circular economy approach,” *Bioprocess and Biosystems Engineering*, vol. 42, no. 5, pp. 677–685, 2019.
 - [10] A. M. Featherstone, “The farm economy: future research and education priorities,” *Applied Economic Perspectives and Policy*, vol. 40, no. 1, pp. 136–154, 2018.
 - [11] R. V. Zharikov, V. V. Bezpалov, S. A. Lochan, M. V. Barashkin, and A. R. Zharikov, “Economic security of regions as a criterion for formation and development of agricultural clusters by means of innovative technologies,” *Scientific Papers Series Management, Economic Engineering in Agriculture and Rural Development*, vol. 18, no. 4, pp. 431–439, 2018.
 - [12] C. Brannstrom, “Feeding the world: Brazil’s transformation into a modern agricultural economy; agribusiness and the neoliberal food system in Brazil: frontiers and fissures of agri-neoliberalism,” *The AAG Review of Books*, vol. 8, no. 2, pp. 76–80, 2020.
 - [13] M. Bergius, T. A. Benjaminsen, and M. Widgren, “Green economy, Scandinavian investments and agricultural modernization in Tanzania,” *Journal of Peasant Studies*, vol. 45, no. 4, pp. 825–852, 2018.
 - [14] I. Ahmed, C. Socci, F. Severini, Q. R. Yasser, and R. Pretaroli, “The structures of production, final demand and agricultural output: a Macro Multipliers analysis of the Nigerian economy,” *Economia Politica*, vol. 35, no. 3, pp. 691–739, 2018.
 - [15] H. Noviar, R. Masbar, S. S. Aliasuddin, T. Zulham, and J. Saputra, “The agricultural commercialisation and its impact on economy management: an application of duality-neo-classic and stochastic frontier approach,” *Industrial Engineering & Management Systems*, vol. 19, no. 3, pp. 510–519, 2020.
 - [16] M. Debela, S. Diriba, and H. Bekele, “Impact of cooperatives membership on economy in eastern oromia: the case of haramaya agricultural FARMERS’COOPERATIVE UNION (hafcu),” *Annals of Public and Cooperative Economics*, vol. 89, no. 2, pp. 361–376, 2018.
 - [17] D. Ramsey and C. D. Malcolm, “The importance of location and scale in rural and small town tourism product development: the case of the Canadian Fossil Discovery Centre, Manitoba, Canada,” *The Canadian Geographer/Le Géographe canadien*, vol. 62, no. 2, pp. 250–265, 2018.
 - [18] F. Drummond and J. Snowball, “Cultural clusters as a local economic development strategy in rural, small town areas: the sarah baartman district in South Africa,” *Bulletin of Geography. Socio-Economic Series*, vol. 43, no. 43, pp. 107–119, 2019.
 - [19] I. Abreu, J. M. Nunes, and F. J. Mesias, “Can rural development be measured? design and application of a synthetic index to Portuguese municipalities,” *Social Indicators Research*, vol. 145, no. 3, pp. 1107–1123, 2019.
 - [20] J. A. González Díaz, R. Celaya, F. Fernández García, K. Osoro, and R. Rosa García, “Dynamics of rural landscapes in marginal areas of northern Spain: past, present, and future,” *Land Degradation & Development*, vol. 30, no. 2, pp. 141–150, 2019.