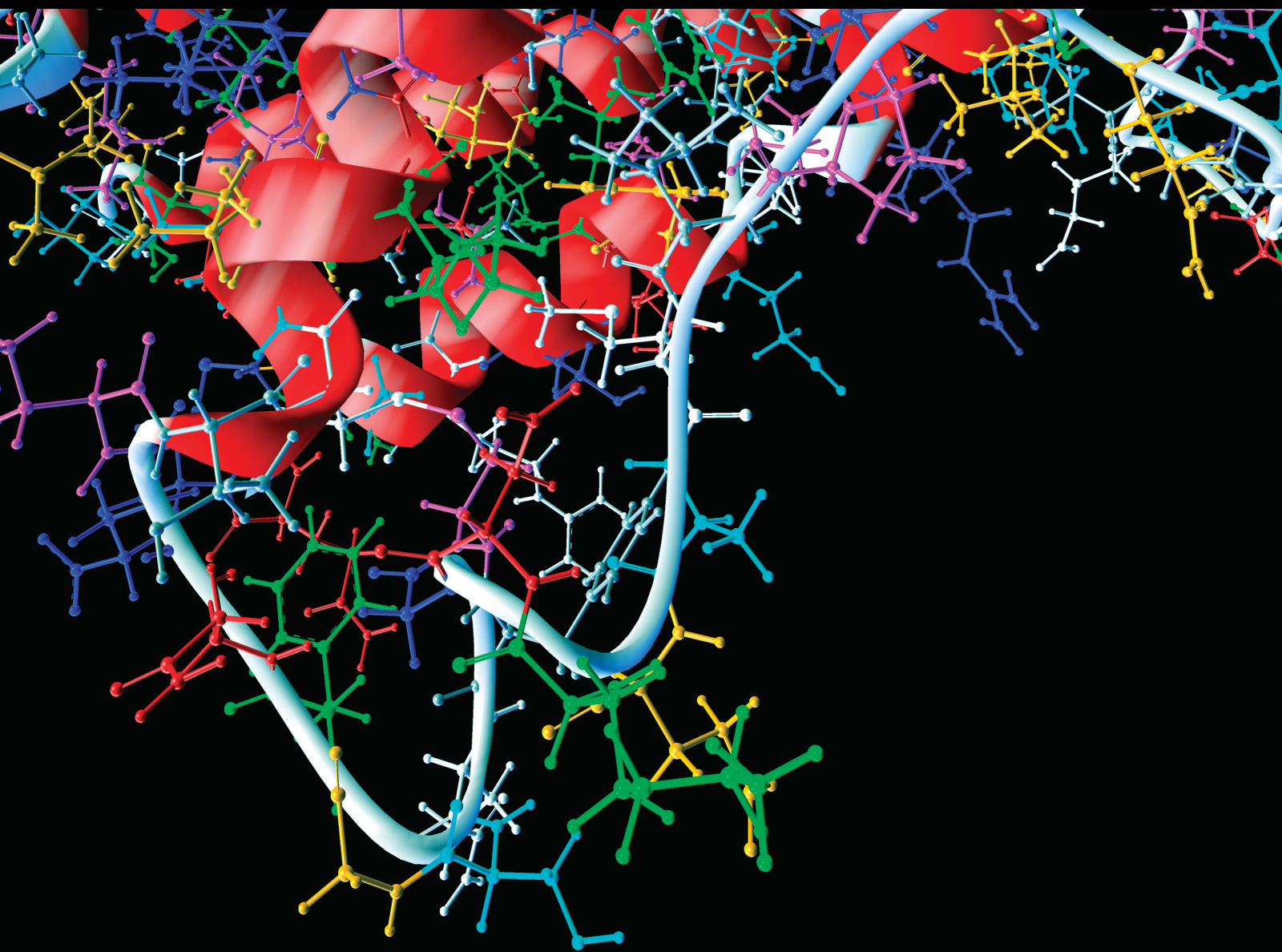


Computational and Mathematical Methods in Medicine

Machine Learning and Network Methods for Biology and Medicine 2020

Lead Guest Editor: Lei Chen

Guest Editors: Tao Huang, Chuan Lu, Lin Lu, and Dandan Li





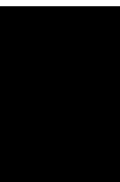
**Machine Learning and Network Methods for
Biology and Medicine 2020**

Computational and Mathematical Methods in Medicine

**Machine Learning and Network
Methods for Biology and Medicine 2020**

Lead Guest Editor: Lei Chen




Guest Editors: Tao Huang, Chuan Lu, Lin Lu, and
Dandan Li



Copyright © 2023 Hindawi Limited. All rights reserved.

This is a special issue published in “Computational and Mathematical Methods in Medicine.” 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.

Associate Editors

Ahmed Albahri, Iraq
Konstantin Blyuss , United Kingdom
Chuangyin Dang, Hong Kong
Farai Nyabadza , South Africa
Kathiravan Srinivasan , India

Academic Editors

Laith Abualigah , Jordan
Yaser Ahangari Nanehkaran , China
Mubashir Ahmad, Pakistan
Sultan Ahmad , Saudi Arabia
Akif Akgul , Turkey
Karthick Alagar, India
Shadab Alam, Saudi Arabia
Raul Alcaraz , Spain
Emil Alexov, USA
Enrique Baca-Garcia , Spain
Sweta Bhattacharya , India
Junguo Bian, USA
Elia Biganzoli , Italy
Antonio Boccaccio, Italy
Hans A. Braun , Germany
Zhicheng Cao, China
Guy Carrault, France
Sadaruddin Chachar , Pakistan
Prem Chapagain , USA
Huiling Chen , China
Mengxin Chen , China
Haruna Chiroma, Saudi Arabia
Watcharaporn Cholamjiak , Thailand
Maria N. D.S. Cordeiro , Portugal
Cristiana Corsi , Italy
Qi Dai , China
Nagarajan Deivanayagam Pillai, India
Didier Delignières , France
Thomas Desaive , Belgium
David Diller , USA
Qamar Din, Pakistan
Irina Doytchinova, Bulgaria
Sheng Du , China
D. Easwaramoorthy , India

Esmaeil Ebrahimie , Australia
Issam El Naqa , USA
Ilias Elmouki , Morocco
Angelo Facchiano , Italy
Luca Faes , Italy
Maria E. Fantacci , Italy
Giancarlo Ferrigno , Italy
Marc Thilo Figge , Germany
Giulia Fiscon , Italy
Bapan Ghosh , India
Igor I. Goryanin, Japan
Marko Gosak , Slovenia
Damien Hall, Australia
Abdulsattar Hamad, Iraq
Khalid Hattaf , Morocco
Tingjun Hou , China
Seiya Imoto , Japan
Martti Juhola , Finland
Rajesh Kaluri , India
Karthick Kanagarathinam, India
Rafik Karaman , Palestinian Authority
Chandan Karmakar , Australia
Kwang Gi Kim , Republic of Korea
Andrzej Kloczkowski, USA
Andrei Korobeinikov , China
Sakthidasan Sankaran Krishnan, India
Rajesh Kumar, India
Kuruva Lakshmana , India
Peng Li , USA
Chung-Min Liao , Taiwan
Pinyi Lu , USA
Reinoud Maex, United Kingdom
Valeri Makarov , Spain
Juan Pablo Martínez , Spain
Richard J. Maude, Thailand
Zahid Mehmood , Pakistan
John Mitchell , United Kingdom
Fazal Ijaz Muhammad , Republic of Korea
Vishal Nayak , USA
Tongguang Ni, China
Michele Nichelatti, Italy
Kazuhisa Nishizawa , Japan
Bing Niu , China

Hyuntae Park , Japan
Jovana Paunovic , Serbia
Manuel F. G. Penedo , Spain
Riccardo Pernice , Italy
Kemal Polat , Turkey
Alberto Policriti, Italy
Giuseppe Pontrelli , Italy
Jesús Poza , Spain
Maciej Przybyłek , Poland
Bhanwar Lal Puniya , USA
Mihai V. Putz , Romania
Suresh Rasappan, Oman
Jose Joaquin Rieta , Spain
Fathalla Rihan , United Arab Emirates
Sidheswar Routray, India
Sudipta Roy , India
Jan Rychtar , USA
Mario Sansone , Italy
Murat Sari , Turkey
Shahzad Sarwar, Saudi Arabia
Kamal Shah, Saudi Arabia
Bhisham Sharma , India
Simon A. Sherman, USA
Mingsong Shi, China
Mohammed Shuaib , Malaysia
Prabhishek Singh , India
Neelakandan Subramani, India
Junwei Sun, China
Yung-Shin Sun , Taiwan
Min Tang , China
Hongxun Tao, China
Alireza Tavakkoli , USA
João M. Tavares , Portugal
Jlenia Toppi , Italy
Anna Tsantili-Kakoulidou , Greece
Markos G. Tsipouras, North Macedonia
Po-Hsiang Tsui , Taiwan
Sathishkumar V E , Republic of Korea
Durai Raj Vincent P M , India
Gajendra Kumar Vishwakarma, India
Liangjiang Wang, USA
Ruisheng Wang , USA
Zhouchao Wei, China
Gabriel Wittum, Germany
Xiang Wu, China

KI Yanover , Israel
Xiaojun Yao , China
Kaan Yetilmezsoy, Turkey
Hiro Yoshida, USA
Yuhai Zhao , China

Contents

Retracted: circFAT1(e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9820923, Volume 2023 (2023)

Retracted: mir-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9895645, Volume 2023 (2023)

Retracted: miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9789647, Volume 2023 (2023)

Retracted: lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9894214, Volume 2023 (2023)

Retracted: Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9875915, Volume 2023 (2023)

Retracted: Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9870132, Volume 2023 (2023)





Retracted: Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9826041, Volume 2023 (2023)



Retracted: miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase

Computational and Mathematical Methods in Medicine
Retraction (1 page), Article ID 9765408, Volume 2023 (2023)

Deep Learning-Based Acute Ischemic Stroke Lesion Segmentation Method on Multimodal MR Images Using a Few Fully Labeled Subjects



Bin Zhao , Zhiyang Liu , Guohua Liu, Chen Cao, Song Jin, Hong Wu , and Shuxue Ding 
Research Article (13 pages), Article ID 3628179, Volume 2021 (2021)

[Retracted] miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1

Jianfeng Li, Xi He, Xiaotang Wu, Xiaohui Liu, Yixiong Huang , and Yuchen Gong 

Research Article (10 pages), Article ID 2953598, Volume 2020 (2020)

Comprehensive Analysis of Differently Expressed and Methylated Genes in Preeclampsia

Wenyi Xu, Ping Ru, Zhuorong Gu, Ruoxi Zhang, Xixia Pang, Yi Huang, Zhou Liu , and Ming Liu 

Research Article (10 pages), Article ID 2139270, Volume 2020 (2020)

Gene Expression Profiling of Type 2 Diabetes Mellitus by Bioinformatics Analysis

Huijing Zhu, Xin Zhu, Yuhong Liu, Fusong Jiang , Miao Chen, Lin Cheng, and Xingbo Cheng 


Research Article (10 pages), Article ID 9602016, Volume 2020 (2020)

[Retracted] circFAT1 (e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis

Jiazhe Liu, Hongchang Li, Chuanchao Wei, Junbin Ding, Jingfeng Lu, Gaofeng Pan , and Anwei Mao 


Research Article (9 pages), Article ID 1459368, Volume 2020 (2020)

Imrecoxib Inhibits Paraquat-Induced Pulmonary Fibrosis through the NF- κ B/Snail Signaling Pathway

Haihao Jin 

Research Article (9 pages), Article ID 6374014, Volume 2020 (2020)

Effects of Bronchoalveolar Lavage with Ambroxol Hydrochloride on Treating Pulmonary Infection in Patients with Cerebral Infarction and on Serum Proinflammatory Cytokines, MDA and SOD

Fanhua Meng, Jing Cheng, Peng Sang, and Jianhui Wang 




Research Article (6 pages), Article ID 7984565, Volume 2020 (2020)

Wavelet Scattering Transform for ECG Beat Classification

Zhishuai Liu, Guihua Yao, Qing Zhang , Junpu Zhang, and Xueying Zeng 


Research Article (11 pages), Article ID 3215681, Volume 2020 (2020)

Texture Feature-Based Classification on Transrectal Ultrasound Image for Prostatic Cancer Detection

Xiaofu Huang, Ming Chen , Peizhong Liu , and Yongzhao Du 



Research Article (9 pages), Article ID 7359375, Volume 2020 (2020)

Transcriptome Analysis Identifies Novel Prognostic Genes in Osteosarcoma

Junfeng Chen, Xiaojun Guo, Guangjun Zeng, Jianhua Liu, and Bin Zhao 

Research Article (8 pages), Article ID 8081973, Volume 2020 (2020)

A Medical Decision Support System to Assess Risk Factors for Gastric Cancer Based on Fuzzy Cognitive Map

Seyed Abbas Mahmoudi , Kamal Mirzaie , Maryam Sadat Mahmoudi, and Seyed Mostafa Mahmoudi

Research Article (13 pages), Article ID 1016284, Volume 2020 (2020)


Contents

A Benign and Malignant Breast Tumor Classification Method via Efficiently Combining Texture and Morphological Features on Ultrasound Images

Mengwan Wei, Yongzhao Du , Xiuming Wu, Qichen Su, Jianqing Zhu, Lixin Zheng, Guorong Lv , and Jiafu Zhuang





Research Article (12 pages), Article ID 5894010, Volume 2020 (2020)

Comprehensive Analysis of Immunoinhibitors Identifies LGALS9 and TGFBR1 as Potential Prognostic Biomarkers for Pancreatic Cancer

Yue Fan, Tianyu Li, Lili Xu, and Tiantao Kuang 



Research Article (13 pages), Article ID 6138039, Volume 2020 (2020)

Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence

Ilker Ozsahin , Boran Sekeroglu , Musa Sani Musa , Mubarak Taiwo Mustapha, and Dilber Uzun Ozsahin 


Research Article (10 pages), Article ID 9756518, Volume 2020 (2020)

Identification of the Key Genes Involved in the Effect of Folic Acid on Endothelial Progenitor Cell Transcriptome of Patients with Type 1 Diabetes

Yi Lu, Qianhong Yang, Wei Hu , and Jian Dong 



Research Article (7 pages), Article ID 4542689, Volume 2020 (2020)

[Retracted] Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis

Chuanwu Fang, Xiaohong Wang, Dongliang Guo, Run Fang, and Ting Zhu 



Research Article (10 pages), Article ID 1367576, Volume 2020 (2020)

[Retracted] Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma

Kaifeng Zhou, Jun Xu, Xiaofan Yin , and Jiangni Xia 



Research Article (9 pages), Article ID 7236245, Volume 2020 (2020)

Comprehensive Analysis of Differentially Expressed circRNAs Reveals a Colorectal Cancer-Related ceRNA Network

Feng Que, Hua Wang, Yi Luo , Li Cui, Lanfu Wei, Zhaohong Xi, Qiu Lin, Yongsheng Ge, and Wei Wang 


Research Article (14 pages), Article ID 7159340, Volume 2020 (2020)

A Semantic Analysis and Community Detection-Based Artificial Intelligence Model for Core Herb Discovery from the Literature: Taking Chronic Glomerulonephritis Treatment as a Case Study

Yun Zhang, Yongguo Liu , Jiating Zhu, Shuangqing Zhai, Rongjiang Jin, and Chuanbiao Wen 


Research Article (23 pages), Article ID 1862168, Volume 2020 (2020)

[Retracted] miR-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis

Xiaoyi Zhu, Zhan Shen, Da Man, Hang Ruan, and Sha Huang 



Research Article (10 pages), Article ID 8209504, Volume 2020 (2020)

Construction of circRNA-Associated ceRNA Network Reveals Novel Biomarkers for Esophageal Cancer

Yunhao Sun, Limin Qiu, Jinjin Chen, Yao Wang, Jun Qian, Lirong Huang, and Haitao Ma 

Research Article (12 pages), Article ID 7958362, Volume 2020 (2020)

Discovery of Prognostic Signature Genes for Overall Survival Prediction in Gastric Cancer

Changyuan Meng, Shusen Xia , Yi He, Xiaolong Tang, Guangjun Zhang, and Tong Zhou 




Research Article (9 pages), Article ID 5479279, Volume 2020 (2020)

[Retracted] lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p

Baoliang Zhu, Jing Liu, Ying Zhao, and Jing Yan 





Research Article (10 pages), Article ID 3128053, Volume 2020 (2020)

Identification of Key mRNAs and lncRNAs in Neonatal Sepsis by Gene Expression Profiling

Lin Bu, Zi-wen Wang, Shu-qun Hu , Wen-jing Zhao, Xiao-juan Geng, Ting Zhou, Luo Zhuo , Xiao-bing Chen, Yan Sun, Yan-li Wang, and Xiao-min Li 

Research Article (13 pages), Article ID 8741739, Volume 2020 (2020)

Comparison of the Therapeutic Effects of Tension Band with Cannulated Screw and Tension Band with Kirschner Wire on Patella Fracture

Chengwu Liu , Haitao Ren , Chunyan Wan , and Jianlin Ma 


Research Article (7 pages), Article ID 4065978, Volume 2020 (2020)

Application of Deep Learning for Early Screening of Colorectal Precancerous Lesions under White Light Endoscopy

Junbo Gao , Yuanhao Guo , Yingxue Sun , and Guoqiang Qu 


Research Article (8 pages), Article ID 8374317, Volume 2020 (2020)

Functional Modular Network Identifies the Key Genes of Preoperative Inhalation Anesthesia and Intravenous Anesthesia in Off-Pump Coronary Artery Bypass Grafting

Hongfei Zhao, Weitian Wang, Liping Liu, Junlong Wang, and Quanzhang Yan 




Research Article (12 pages), Article ID 4574792, Volume 2020 (2020)

Integrated Genome-Wide Methylation and Expression Analyses Reveal Key Regulators in Osteosarcoma

Fei Wang, Guoqing Qin, Junzhi Liu, Xiunan Wang, and Baoguo Ye 



Research Article (11 pages), Article ID 7067649, Volume 2020 (2020)

A Comparative Analysis of Visual Encoding Models Based on Classification and Segmentation Task-Driven CNNs

Ziya Yu , Chi Zhang , Linyuan Wang , Li Tong , and Bin Yan 

Research Article (15 pages), Article ID 5408942, Volume 2020 (2020)



A Simple Method to Train the AI Diagnosis Model of Pulmonary Nodules

Zhehao He , Wang Lv, and Jian Hu 

Research Article (6 pages), Article ID 2812874, Volume 2020 (2020)

Contents

Comparison of Common Methods for Precision Volume Measurement of Hematoma

Minhong Chen, Zhong Li , Jianping Ding, Xingqi Lu, Yinan Cheng, and Jiayun Lin 

Research Article (11 pages), Article ID 6930836, Volume 2020 (2020)

CT-TEE Image Registration for Surgical Navigation of Congenital Heart Disease Based on a Cycle Adversarial Network

Yunfei Lu, Bing Li, Ningtao Liu, Jia-Wei Chen , Li Xiao, Shuiping Gou , Linlin Chen, Meiping Huang , and Jian Zhuang



Research Article (8 pages), Article ID 4942121, Volume 2020 (2020)

Systematic Identification of lncRNA-Associated ceRNA Networks in Immune Thrombocytopenia

Zhenwei Fan, Xuan Wang , Peng Li , Chunli Mei, Min Zhang , Chunshan Zhao, and Yan Song

Research Article (8 pages), Article ID 6193593, Volume 2020 (2020)

[Retracted] miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase

Xinhua Xu, Yan Ding, Jun Yao, Zhiping Wei, Haipeng Jin, Chen Chen, Jun Feng , and Rongbiao Ying 






Research Article (10 pages), Article ID 5807836, Volume 2020 (2020)

Prediction of High-Risk Types of Human Papillomaviruses Using Reduced Amino Acid Modes

Xinnan Xu, Rui Kong, Xiaoqing Liu, Pingan He , and Qi Dai 


Research Article (10 pages), Article ID 5325304, Volume 2020 (2020)

Construction and Comprehensive Analysis of Dysregulated Long Noncoding RNA-Associated Competing Endogenous RNA Network in Moyamoya Disease

Xuefeng Gu , Dongyang Jiang, Yue Yang, Peng Zhang , Guoqing Wan, Wangxian Gu, Junfeng Shi, Liying Jiang, Bing Chen, Yanjun Zheng, Dingsheng Liu , Sufen Guo , and Changlian Lu 


Research Article (12 pages), Article ID 2018214, Volume 2020 (2020)

Influences of Daily Life Habits on Risk Factors of Stroke Based on Decision Tree and Correlation Matrix

Zeguo Shao, Yuhong Xiang, Yingchao Zhu, Aiqin Fan, and Peng Zhang 

Research Article (12 pages), Article ID 3217356, Volume 2020 (2020)

CUL1-Mediated Organelle Fission Pathway Inhibits the Development of Chronic Obstructive Pulmonary Disease

Ran Li, Feng Xu, Xiao Wu, Shaoping Ji, and Ruixue Xia 


Research Article (11 pages), Article ID 5390107, Volume 2020 (2020)

Interpretable Learning Approaches in Resting-State Functional Connectivity Analysis: The Case of Autism Spectrum Disorder


Jinlong Hu , Lijie Cao , Tenghui Li , Bin Liao , Shoubin Dong , and Ping Li

Research Article (12 pages), Article ID 1394830, Volume 2020 (2020)




[Retracted] Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle

Weifen Lu, Qianli Pan, Yinxin Zhou, Wenyu Chen, Hongyan Zhang, and Weibo Qi 
Research Article (4 pages), Article ID 6896517, Volume 2020 (2020)


Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy

Haiyan Liang, Lei Chen , Xian Zhao, and Xiaolin Zhang
Research Article (16 pages), Article ID 1573543, Volume 2020 (2020)




Fusion of FDG-PET Image and Clinical Features for Prediction of Lung Metastasis in Soft Tissue Sarcomas

Jin Deng , Weiming Zeng , Yuhu Shi , Wei Kong , and Shunjie Guo 
Research Article (11 pages), Article ID 8153295, Volume 2020 (2020)






In Silico Analysis Identifies Differently Expressed lncRNAs as Novel Biomarkers for the Prognosis of Thyroid Cancer

Yuansheng Rao, Haiying Liu, Xiaojuan Yan, and Jianhong Wang 
Research Article (10 pages), Article ID 3651051, Volume 2020 (2020)

ACNNT3: Attention-CNN Framework for Prediction of Sequence-Based Bacterial Type III Secreted Effectors

Jie Li , Zhong Li , Jiesi Luo, and Yuhua Yao 
Research Article (7 pages), Article ID 3974598, Volume 2020 (2020)

HMMPred: Accurate Prediction of DNA-Binding Proteins Based on HMM Profiles and XGBoost Feature Selection

Xiuzhi Sang , Wanyue Xiao , Huiwen Zheng , Yang Yang , and Taigang Liu 
Research Article (10 pages), Article ID 1384749, Volume 2020 (2020)

Retraction

Retracted: circFAT1(e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis

Computational and Mathematical Methods in Medicine

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] J. Liu, H. Li, C. Wei et al., "circFAT1(e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1459368, 9 pages, 2020.

Retraction

Retracted: mir-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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. Zhu, Z. Shen, D. Man, H. Ruan, and S. Huang, "miR-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 8209504, 10 pages, 2020.

Retraction

Retracted: miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] J. Li, X. He, X. Wu, X. Liu, Y. Huang, and Y. Gong, "miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 2953598, 10 pages, 2020.

Retraction

Retracted: lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p

Computational and Mathematical Methods in Medicine

Received 19 September 2023; Accepted 19 September 2023; Published 20 September 2023

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] B. Zhu, J. Liu, Y. Zhao, and J. Yan, "lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 3128053, 10 pages, 2020.

Retraction

Retracted: Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] W. Lu, Q. Pan, Y. Zhou, W. Chen, H. Zhang, and W. Qi, "Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 6896517, 4 pages, 2020.

Retraction

Retracted: Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] K. Zhou, J. Xu, X. Yin, and J. Xia, "Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 7236245, 9 pages, 2020.

Retraction

Retracted: Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] C. Fang, X. Wang, D. Guo, R. Fang, and T. Zhu, "Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1367576, 10 pages, 2020.

Retraction

Retracted: miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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. Xu, Y. Ding, J. Yao et al., "miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 5807836, 10 pages, 2020.

Research Article

Deep Learning-Based Acute Ischemic Stroke Lesion Segmentation Method on Multimodal MR Images Using a Few Fully Labeled Subjects

Bin Zhao ¹, Zhiyang Liu ¹, Guohua Liu,¹ Chen Cao,² Song Jin,² Hong Wu ¹,
and Shuxue Ding ^{1,3}

¹Tianjin Key Laboratory of Optoelectronic Sensor and Sensing Network Technology, College of Electronic Information and Optical Engineering, Nankai University, Tianjin 300350, China

²Key Laboratory for Cerebral Artery and Neural Degeneration of Tianjin, Department of Medical Imaging, Tianjin Huanhu Hospital, Tianjin 300350, China

³School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin Guangxi 541004, China

Correspondence should be addressed to Hong Wu; wuhong@nankai.edu.cn and Shuxue Ding; sding@guet.edu.cn

Received 18 June 2020; Revised 17 December 2020; Accepted 10 January 2021; Published 30 January 2021

Academic Editor: Lei Chen

Copyright © 2021 Bin Zhao 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.

Acute ischemic stroke (AIS) has been a common threat to human health and may lead to severe outcomes without proper and prompt treatment. To precisely diagnose AIS, it is of paramount importance to quantitatively evaluate the AIS lesions. By adopting a convolutional neural network (CNN), many automatic methods for ischemic stroke lesion segmentation on magnetic resonance imaging (MRI) have been proposed. However, most CNN-based methods should be trained on a large amount of fully labeled subjects, and the label annotation is a labor-intensive and time-consuming task. Therefore, in this paper, we propose to use a mixture of many weakly labeled and a few fully labeled subjects to relieve the thirst of fully labeled subjects. In particular, a multifeature map fusion network (MFMF-Network) with two branches is proposed, where hundreds of weakly labeled subjects are used to train the classification branch, and several fully labeled subjects are adopted to tune the segmentation branch. By training on 398 weakly labeled and 5 fully labeled subjects, the proposed method is able to achieve a mean dice coefficient of 0.699 ± 0.128 on a test set with 179 subjects. The lesion-wise and subject-wise metrics are also evaluated, where a lesion-wise F1 score of 0.886 and a subject-wise detection rate of 1 are achieved.

1. Introduction

Stroke has been one of the most serious threats to human health, which can lead to long-term disability or even death [1]. In general, stroke can be divided into ischemia and hemorrhage based on the types of cerebrovascular accidents, where ischemic stroke accounts for 87% [2]. In clinical practice, multimodal magnetic resonance images (MRIs), including the diffusion-weighted imaging (DWI) and the apparent diffusion coefficient (ADC) maps derived from multiple DWI images with different b values, have been used in diagnosing acute ischemic stroke (AIS), thanks to the short acqui-

sition time and high sensitivity [3]. As AIS progresses rapidly and may lead to severe outcomes, it is of paramount importance to quickly diagnose and quantitatively evaluate the AIS lesions from the multimodal MRIs, which is, however, time-consuming and requires experienced medical imaging clinicians. Therefore, it is quite necessary to develop automatic methods in analyzing the images.

Many automatic stroke lesion segmentation methods have been developed in the literature. For instance, Nabizadeh et al. [4] proposed a gravitational histogram optimization by identifying the abnormal intensity. To reduce the false positive rate, Mitra et al. [5] used the random forest to

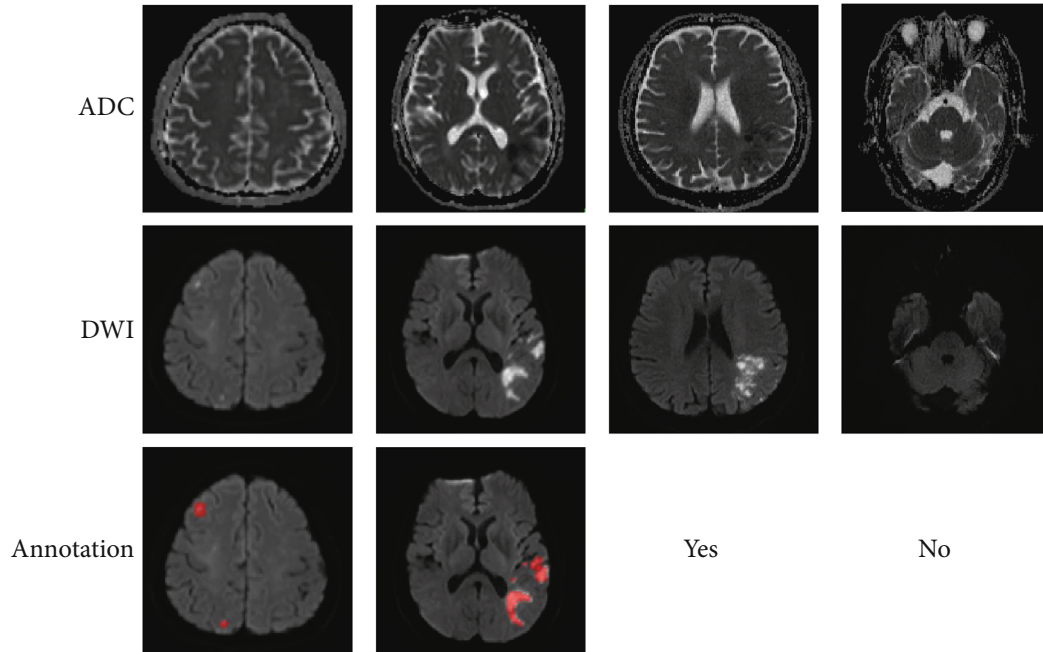


FIGURE 1: Examples of fully labeled and weakly labeled subjects. The first two columns show fully labeled examples, and the last two are weakly labeled ones, where the label “yes” indicates that the slice has a lesion and “no” indicates the opposite. Best viewed in color.

extract features and identify the lesions based on multimodal MRIs. Maier et al. [6] adopted the support vector machine based on the local features extracted from multimodal MRIs. Although such methods achieved high performance on ischemic stroke lesion segmentation, their modeling capabilities were significantly limited due to their heavy dependence on handcrafted features.

A convolutional neural network (CNN) has recently presented an exceptional performance in computer vision. By training on a large number of fully labeled subjects where the stroke lesions were annotated in a pixel-by-pixel manner, the CNN-based methods have shown their great potentials in segmenting ischemic stroke lesions on the MRIs [7–11]. As a CNN typically has millions of parameters, such methods require hundreds of fully labeled subjects to train the CNN. Figure 1 presents some examples of fully labeled subjects. It is obvious that annotating pixel-by-pixel labels is a tedious task and would take a significant amount of time to establish a large dataset with fully labeled subjects, which makes it impossible to establish a medical imaging dataset with a comparable size to the commonly used datasets in computer vision. This motivates us to develop segmentation methods while reducing the annotation burden for medical imaging clinicians.

Few-shot learning has recently been adopted in image semantic segmentation [12–15]. By fine-tuning the network parameters with a few samples, the CNN can achieve high segmentation accuracy in many tasks. Typically, the few-shot learning methods require ImageNet [16] pretrained parameters to help extract features. In the medical image segmentation task, however, it is not possible to find a dataset as large as ImageNet to obtain pretrained parameters. Therefore, it is necessary to design an auxiliary task with easily obtained labels to pretrain the network.

In particular, we make use of many weakly labeled subjects and propose to use weakly supervised learning method to facilitate the AIS lesion segmentation. Different from the other AIS lesion segmentation methods [17–21], the weakly labeled subjects are annotated as whether each slice of a subject incorporates lesion or not, as shown in Figure 1, which significantly reduces the cost on annotation.

Our proposed method consists of three processes: classification, segmentation, and inference. In the classification process, the network is trained on the weakly labeled subjects as a classifier to obtain a set of pretrained parameters. In the segmentation process, the network freezes the pretrained parameter and is further trained on the fully labeled subjects. In the inference process, the classification branch generates class activation mapping (CAM) [22] and the segmentation branch predicts the segmentation result. A postprocessing algorithm is adopted to combine the CAM with the segmentation result to generate a final prediction. By using 398 weakly labeled subjects and 5 fully labeled ones, the proposed method is able to achieve a dice coefficient of 0.699 ± 0.128 . The lesion-wise and subject-wise performances are also evaluated, where a lesion-wise F1 score of 0.886 and a subject-wise detection rate of 1 are achieved.

2. Materials and Methods

In this section, we propose a deep learning-based method using a few fully labeled subjects for AIS segmentation on two-modal MR images, and the pipeline is presented in Figure 2. In particular, our proposed method consists of three processes: classification, segmentation, and inference. In the classification process, the network is trained on the weakly

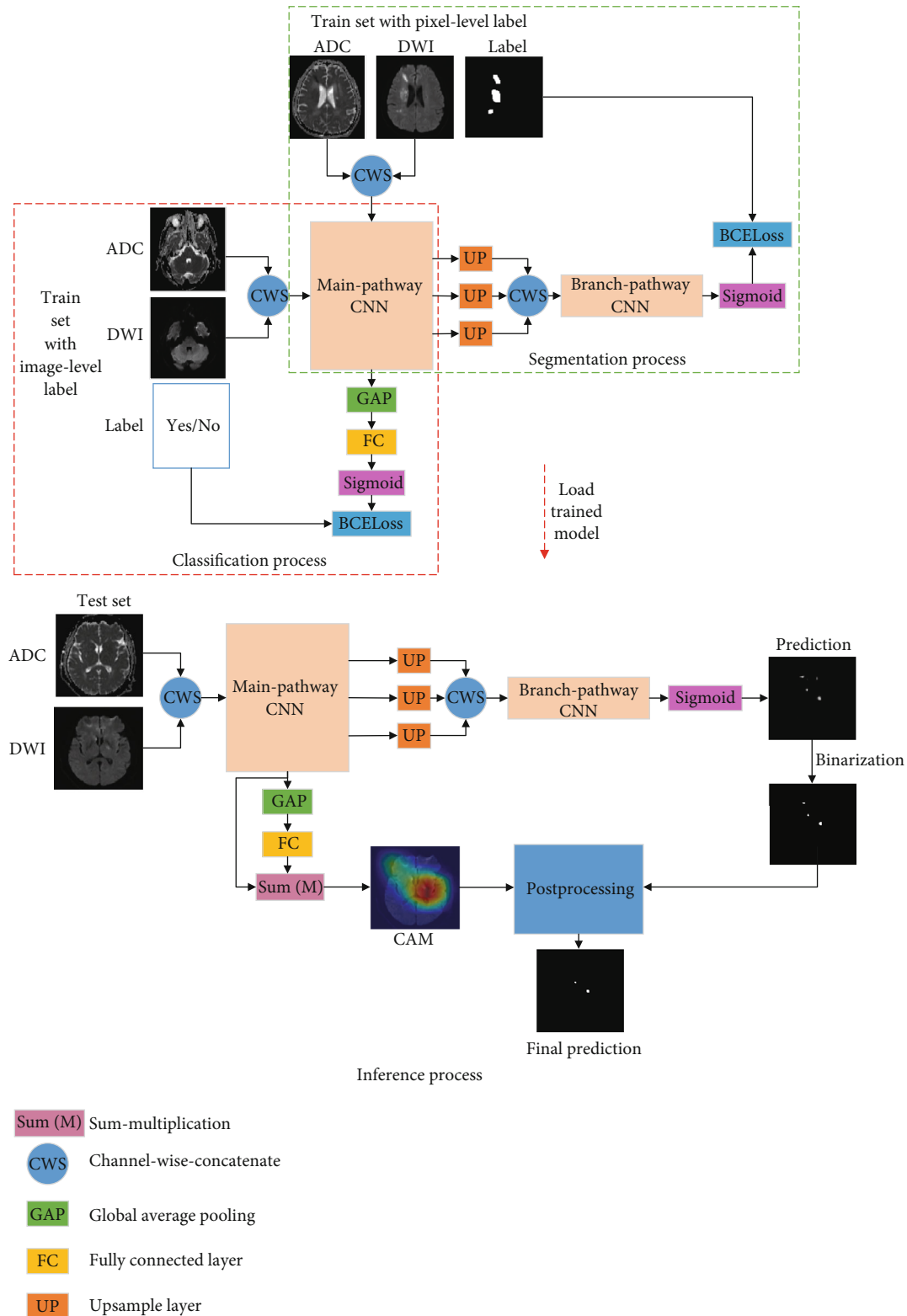


FIGURE 2: Whole pipeline of the proposed method. Best viewed in color.

labeled subjects as a classifier. This process obtains a set of pretrained parameters. In the segmentation process, the network is trained end-to-end on the fully labeled subjects by freezing the pretrained parameters. That is to say, in order to avoid overfitting, only the decoder is trained using a few

fully labeled subjects. In the inference process, the classification branch generates class activation mapping (CAM) [22] and the segmentation branch predicts the segmentation result. Then, a postprocessing method is adopted to combine the CAM with the segmentation result to generate a final

prediction. As we will show in this paper, only 5 fully labeled subjects are adequate to achieve accurate segmentation.

2.1. Multifeature Map Fusion Network. Different from the few-shot semantic segmentation on natural images where the ImageNet pretrained parameters were easily obtained, there is no available large dataset for brain MRIs. A multifeature map fusion network (MFMF-Network) is proposed and trained on the weakly labeled subjects to extract features whose architecture is presented in Figure 3. The proposed MFMF-Network is a two-branch CNN, where the backbone CNN is a VGG16 [23] truncated before the 5th MaxPooling layer.

As Figure 2 shows, we add a global average pooling (GAP) followed by a fully connected (FC) layer at the top of the main-pathway CNN as the classification branch, which is trained by the weakly labeled subjects at the classification process. On the other hand, the segmentation branch fuses the upsampled feature maps from convolutional blocks 4, 7, and 10, which is used to generate a pixel-wise segmentation map.

Intuitively, the feature maps of the deeper convolutional block have much lower spatial resolution than the original input images but with better semantic information. We further incorporate the squeeze-and-excitation (SE) module [24] into the upsample layer as depicted in Figure 3(b), such that the network can focus on the feature maps that contribute most to AIS segmentation.

The training of the MFMF-Network takes two steps. In the classification process, the backbone CNN, together with the classification branch, is trained on the weakly labeled subjects as a classifier. In the segmentation process, the segmentation branch is trained on a few fully labeled subjects, while the parameters of the backbone CNN are frozen.

2.2. Postprocessing. In the inference process, as Figure 2 shows, the classification branch generates CAM [22] as

$$M_c(x, y) = \sum_k w_k^c \cdot f_k(x, y), \quad (1)$$

where $f_k(x, y)$ represents the activation of unit k in the last convolutional layer of main-pathway CNN at the spatial location (x, y) and w_k is the weight corresponding to the class c for unit k . Note that as the AIS lesion segmentation is a binary segmentation task, that is, $c=2$, therefore, we only consider the CAM of the lesion class. The CAM is normalized to generate a segmentation probability map, and a binary segmentation result $M_c(x, y; \delta)$ is further obtained by using a threshold of $\delta=0.5$. Simultaneously, the segmentation branch predicts the segmentation probability map. The binary segmentation result $S_c(x, y; \delta)$ at the spatial location (x, y) is also obtained by using the same threshold δ .

Nevertheless, since few fully labeled subjects are used to train the segmentation branch, it is inevitable to generate some false positives. To fully utilize the rich semantic information from the weakly labeled data, we further fuse the CAM generated from the classification branch with the seg-

mentation branch output to reduce the FPs, which is computed as

$$P_c(x, y) = M_c(x, y; \delta) \cdot S_c(x, y; \delta). \quad (2)$$

2.3. Evaluation Metrics. In this subsection, we introduce a number of metrics to evaluate our proposed method. First, the dice coefficient (DC) is used to evaluate the pixel-level segmentation performance. It measures the overlap between the predicted segmentation P and the ground truth G and is formulated as

$$DC = \frac{2|G \cap P|}{|G| + |P|}, \quad (3)$$

where $|\bullet|$ denotes the number of pixels in the set.

In addition, we further propose the lesion-wise precision rate P_L , the lesion-wise recall rate R_L , and the lesion-wise F1 score as metrics, which are defined as

$$P_L = \frac{m\#TP}{m\#TP + m\#FP}, \quad (4)$$

$$R_L = \frac{m\#TP}{m\#TP + m\#FN}, \quad (5)$$

$$F1 = \frac{2P_L \cdot R_L}{P_L + R_L}, \quad (6)$$

where $m\#TP$, $m\#FP$, and $m\#FN$ are the mean number of true positives (TPs), false positives (FPs), and false negatives (FNs), respectively, which are calculated in a lesion-wise manner. In this paper, a 3D connected component is performed on both the ground truth and the predicted segmentation map. A TP is defined as a connected region on the predicted segmentation map that overlaps with that on the ground truth. The number of TPs is counted on each subject, and the mean number of TPs ($m\#TP$) is then obtained by averaging the number of TPs over all subjects. A FP is counted if a region on the predicted segmentation has no overlap with any region on the ground truth. While a FN is counted if a region on the ground truth has no overlap with any region on the predicted segmentation.

We further use the detection rate (DR) to measure missed subjects as a subject-wise metric, which is defined as

$$DR = \frac{N_{TP}}{N}, \quad (7)$$

where N denotes the number of all subjects and N_{TP} denotes the number of subjects with any TP lesion detection.

3. Experiments

In this section, we will introduce the experimental data, the implementation details, and the results.

3.1. Data and Preprocessing. The experimental data includes 582 subjects with AIS lesions, which were collected from a retrospective database of Tianjin Huanhu Hospital (Tianjin,

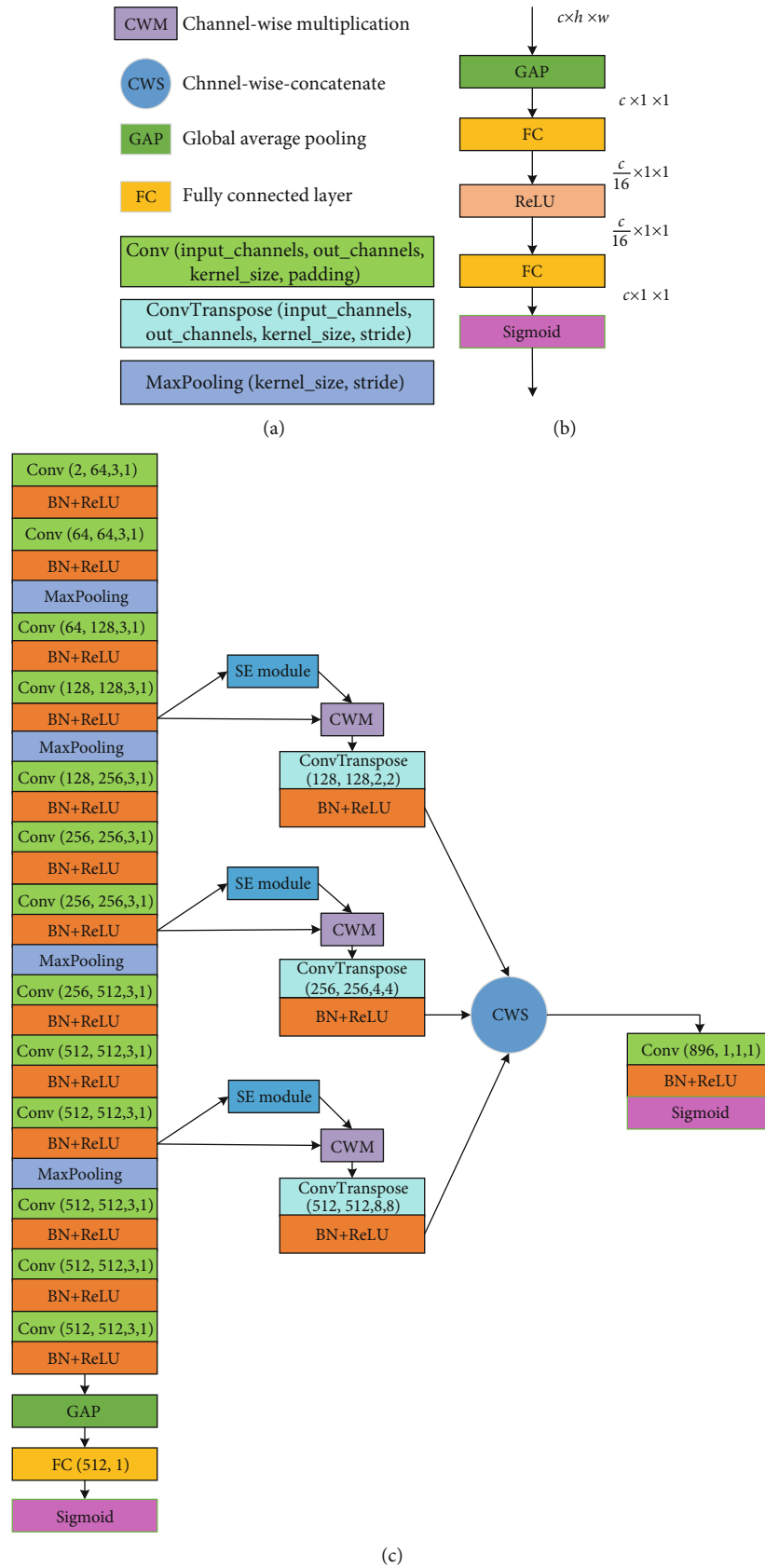


FIGURE 3: Our proposed network architecture. (a) Unit parameter description. (b) SE module. (c) Multifeature map fusion network (MFMF-Network). Best viewed in color.

TABLE 1: Parameters used in DWI acquisition.

MR scanners	Skyra	Trio	Avanto
Repetition time (ms)	5200	3100	3800
Echo time (ms)	80	99	102
Flip angle (°)	150	120	150
Number of excitations	1	1	3
Field of view (mm ²)	240 × 240	200 × 200	240 × 240
Matrix size	130 × 130	132 × 132	192 × 192
Slice thickness (mm)	5	6	5
Slice spacing (mm)	1.5	1.8	1.5
Number of slices	21	17	21

China) and anonymized prior to the use of researchers. Ethical approval was granted by the Tianjin Huanhu Hospital Medical Ethics Committee. MR images were acquired from three MR scanners, with two 3T MR scanners (Skyra, Siemens, and Trio, Siemens) and one 1.5T MR scanner (Avanto, Siemens). DWIs were acquired using a spin echo-type echo planner imaging (SE-EPI) sequence with b values of 0 and 1000 s/mm². The parameters used in DWI acquisition are shown in Table 1. ADC maps were calculated from the scan raw data in a pixel-by-pixel manner as

$$\text{ADC} = \frac{\ln S_1 - \ln S_0}{b_1 - b_0}, \quad (8)$$

where b characterizes the diffusion-sensitizing gradient pulses, with $b_1 = 1000$ s/mm² and $b_0 = 0$ s/mm² in our data. S_1 is the diffusion-weighted signal intensity with $b_1 = 1000$ s/mm². S_0 is the signal with no diffusion gradient applied, i.e., with $b_0 = 0$ s/mm².

The AIS lesions were manually annotated by two experienced experts (Dr. Song Jin and Dr. Chen Cao) from Tianjin Huanhu Hospital. The entire dataset includes 398 weakly labeled subjects and 184 fully labeled subjects, and they are divided into the training set and test set. The training set includes 398 weakly labeled subjects and 5 fully labeled subjects, which are used to train the network parameters. The test set includes the remaining 179 fully labeled subjects to evaluate the generalization capacities on unknown samples. For the sake of simplicity, we name the weakly labeled and fully labeled subjects in the training set as *cla-data* and *seg-data*, respectively.

As the MR images were acquired on the three different MR scanners, their matrix sizes are different, as shown in Table 1. Therefore, we resample all the MR images to the same size of 192 × 192 using linear interpolation. The pixel intensity of each MR image is normalized into that of zero mean and unit variance, and the DWI and ADC slices are channel-wise concatenated as dual-channel images and fed into the MFMF-Network. Data augmentation technique is adopted in both the classification process and the segmentation process. In particular, each input image is randomly rotated by a degree ranging from 1 to 360 degrees, flipped

vertically and horizontally on the fly, so as to augment the dataset and reduce memory footprint.

3.2. Implementation Details. The parameters of the proposed MFMF-Network are shown in Figure 3. In the classification process, we initialize the main-pathway CNN using the pre-trained parameters of VGG16 on ImageNet [16]. The FC layer parameters are initialized from zero-mean Gaussian distributions with a standard deviation of 0.1. After training the classification branch, we freeze the main-pathway CNN and initialize the other parameters in the segmentation branch, as suggested in [25]. In both the classification and segmentation processes, the RAdam method [26] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used as the optimizer and the initial learning rate is set as 10^{-3} . The loss function used in this paper is binary cross-entropy (BCELoss).

We randomly select 0.1 of the *cla-data* as the validation set, which is used to fine-tune the hyperparameters in the classification process. During training, the learning rate is scaled down by a factor of 0.1 if no progress is made for 15 epochs on validation loss, and the training stops after 30 epochs with no progress on the validation loss. For the segmentation process, we pick all slices with lesions from the *seg-data* to train the segmentation branch. Dynamic learning rate scheduling is also adopted, where the learning rate is scaled down by a factor of 0.1 if no progress is made for 15 epochs on training loss. We stop the training of the segmentation process if the learning rate is 10^{-9} or no progress after 30 epochs on the training loss.

The experiments are performed on a computer with an Intel Core i7-6800K CPU, 64 GB RAM, and Nvidia GeForce 1080Ti GPU with 11 GB memory. The network is implemented in PyTorch. The MR image files are stored as Neuroimaging Informatics Technology Initiative (NIFTI) format and processed using Simple Insight ToolKit (SimpleITK) [27]. We use ITK-SNAP [28] for visualization.

3.3. Results. The proposed method is evaluated on the test set with 179 fully labeled subjects. For the sake of comparison, we also train and evaluate U-Net [29], FCN-8s [30], ResUNet [21], and the method proposed in [31] on our dataset. For fairness consideration, the encoder parts of these methods are also pretrained as a classifier on our weakly labeled data. In particular, for the few-shot segmentation method proposed in [31], we split the slices of the *seg-data* with AIS lesions into the support set and query set. Other experimental details are the same as our proposed method except for freezing the pretrained parameters.

Figure 4 visualizes some examples of AIS segmentation. As Figure 4 shows, our proposed method, i.e., column (h), is accurate on both the large and small AIS lesions. Even though U-Net and Res-UNet have more multifeature fusion, they overestimate the lesion but ignore the details of adjacent lesions. On the other hand, FCN-8s uses three-scale feature fusion, which is the same as our method, but the outputs of its last convolutional layer resampled to the size of input images require interpolation of 32 times, which inevitably leads to an overestimated lesion region. For the few-shot segmentation method proposed in [31], the multifeature fusion

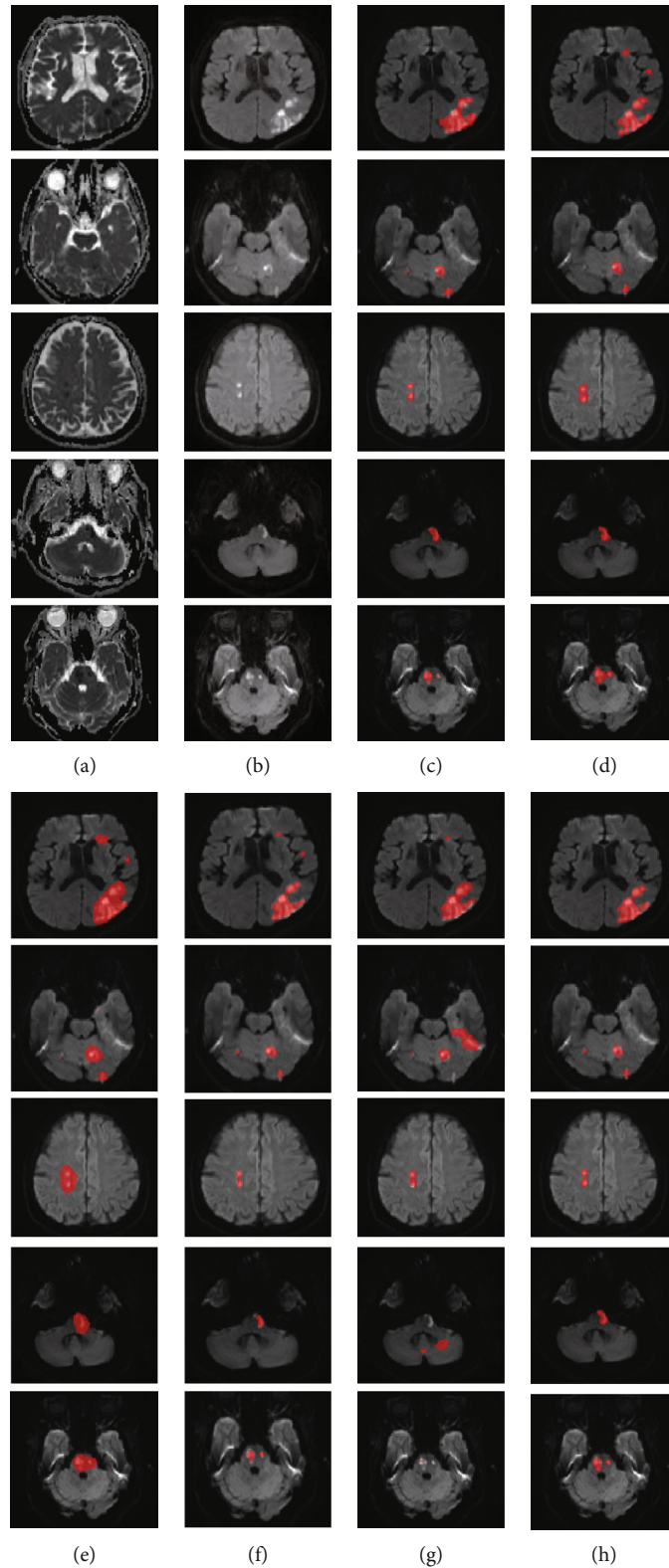


FIGURE 4: Visualization examples of the MRI slices and lesion segmentation results. (a–c) The original ADC map, DWI, and ground truth, respectively. (d–h) The segmentation results of U-Net, FCN-8s, Res-UNet, the method in [31], and the proposed method, respectively. The segmentation results are overlaid on the DWIs and highlighted in red.

combines the support set with the query set to train the parameters. Nevertheless, the proportion of positive pixels in the medical slice is typically smaller than that of the natu-

ral image, making the few-shot segmentation method in [31] tend to ignore small lesions or misclassify the artifact regions as lesions, as shown in Figure 4.

TABLE 2: Evaluation results on the test set. In particular, the mean DC is presented in the way of mean \pm standard deviation. The best result has been highlighted in italic.

Method	DC	P_L	R_L	F1	DR
U-Net [29]	0.629 \pm 0.152	0.285	0.942	0.437	1.000
FCN-8s [30]	0.289 \pm 0.222	0.234	0.938	0.374	1.000
Res-UNet [21]	0.557 \pm 0.227	0.494	0.901	0.638	0.972
Few-shot [31]	0.239 \pm 0.253	0.191	0.591	0.288	0.642
Ours	<i>0.699 \pm 0.128</i>	<i>0.852</i>	<i>0.923</i>	<i>0.886</i>	1.000

The quantitative evaluation results are summarized in Table 2. As Table 2 shows, our proposed method achieves the best results on all of the metrics except for the recall rate. Specifically, our proposed method achieves a mean dice coefficient of 0.699 ± 0.128 from the aspect of the pixel-level metric, which is much higher than the results obtained by FCN-8s [30] and the few-shot segmentation method [31] and is also higher than that of U-Net [29] and Res-UNet [21]. For the lesion-wise metrics, our proposed method achieves the highest precision rate of 0.852 and the highest F1 score of 0.886 over the competitors. The recall rate of 0.923, however, is slightly worse than U-Net and FCN-8s due to the fact that they tend to cover a larger area than the real lesion size, which reduces the number of FNs when many small lesions gathered together. Furthermore, for the subject-wise metric, all of the methods achieve a detection rate of 1 except for the few-shot segmentation method in [31] and Res-UNet.

Figure 5 further plots the scatter map between the volumes of the manual annotation and the predicted segmentation, where the purple line indicates a perfect match between the predicted volumes and the ground truth volumes. As Figure 5 shows, the predicted volumes of our proposed method are closer to the true volumes than the competitors.

4. Discussions

4.1. How Many Weakly Labeled Subjects Do We Need? So far, we have shown that our proposed method can achieve high segmentation accuracy by using 398 weakly labeled and 5 fully labeled subjects. It is worth investigating whether we can further reduce the number of weakly labeled subjects. In particular, we randomly select proportions of 0.8, 0.6, 0.4, and 0.2 from the 398 subjects to train the classification branch.

Table 3 summarizes the evaluation results with different numbers of weakly labeled subjects. As we can see from Table 3, we can achieve a DR of 1 when more than 238 subjects are used to train the classification branch; besides, we can also achieve a higher mean dice coefficient and recall rate as the number of weakly labeled subjects increased. The other metrics, including the precision rate and F1 score, generally rise accompanied by small fluctuations.

4.2. Effect of Postprocessing. From Table 3, we can also see that our proposed method uses 159 subjects to obtain the pretrained parameters achieving a detection rate of 0.966, which means that it fails to detect 6 subjects in the test set.

In fact, the detection rate is 1 when the segmentation branch directly predicts the segmentation results without using post-processing. However, the precision rate and the F1 score are much lower than those using postprocessing. To investigate the importance of postprocessing, we summarize the comparison results with different numbers of weakly labeled subjects, as shown in Table 4. As Table 4 shows, postprocessing greatly improves the dice coefficient, precision rate, and F1 score but reduces the detection rate, which is because of the CAM generated by the classification branch. Figure 6 presents some samples of CAM. As Figure 6 shows, the CAM shows a higher probability in the suspected lesion region with the increasing number of weakly labeled subjects used in the classification branch. In particular, the CAM shows a probability of 0 or a probability below the threshold of $\delta = 0.5$ in some subjects when less than 159 weakly labeled subjects are used to train the classification branch, which leads to missed diagnosis when postprocessing is used in the inference process. In a word, our postprocessing is critical for AIS lesion segmentation in this research.

4.3. Single Modal vs. Multimodal. In this subsection, we explore the effect of different modalities of MR images on our results. We use single-modal and multimodal subjects to train and test our proposed method. The dataset for training the classification branch includes all the 398 subjects regardless of the modal combination. As Table 5 shows, the multimodal subjects achieve the best results. The DWI also achieves competitive results compared with the multimodal. The DWI achieves competitive results due to the fact that the AIS lesions appear as hyperintense on the DWIs, which is more prominent to be recognized than that on the ADC maps. The combinational use of the DWI and ADC map, on the other hand, helps in reducing the FPs and FNs, which largely improves the segmentation results.

4.4. Impact of Using Lesion Slices Only. Note that we only extract slices with AIS lesions from the 5 fully labeled subjects in the seg-data to train the segmentation branch. In this subsection, we would like to further discuss whether the slices without any lesion should be included. Table 6 summarizes the evaluation results after training on all subjects and only lesion slices. As Table 6 shows, the network trained on lesion slices shows superior performance over that trained on all slices on all metrics except the recall rate, which means that training on both the normal and lesion slices will reduce the number of FNs but increase the number of FPs. Intuitively, including the normal slices will make the class imbalance problem more severe, leading to inadequate learning on the lesion features. In fact, as the AIS lesion volume is much smaller than the normal tissues in most cases, the lesion slices have included much information about the normal tissue appearance. We can then conclude that to improve the segmentation accuracy, it is necessary to only include the lesion slices when training the segmentation branch.

4.5. Performance on Large and Small Lesions. Clinically, an AIS lesion is classified as a lacunar infarction (LI) lesion if its diameter is smaller than 1.5 cm [32]. LI is much difficult

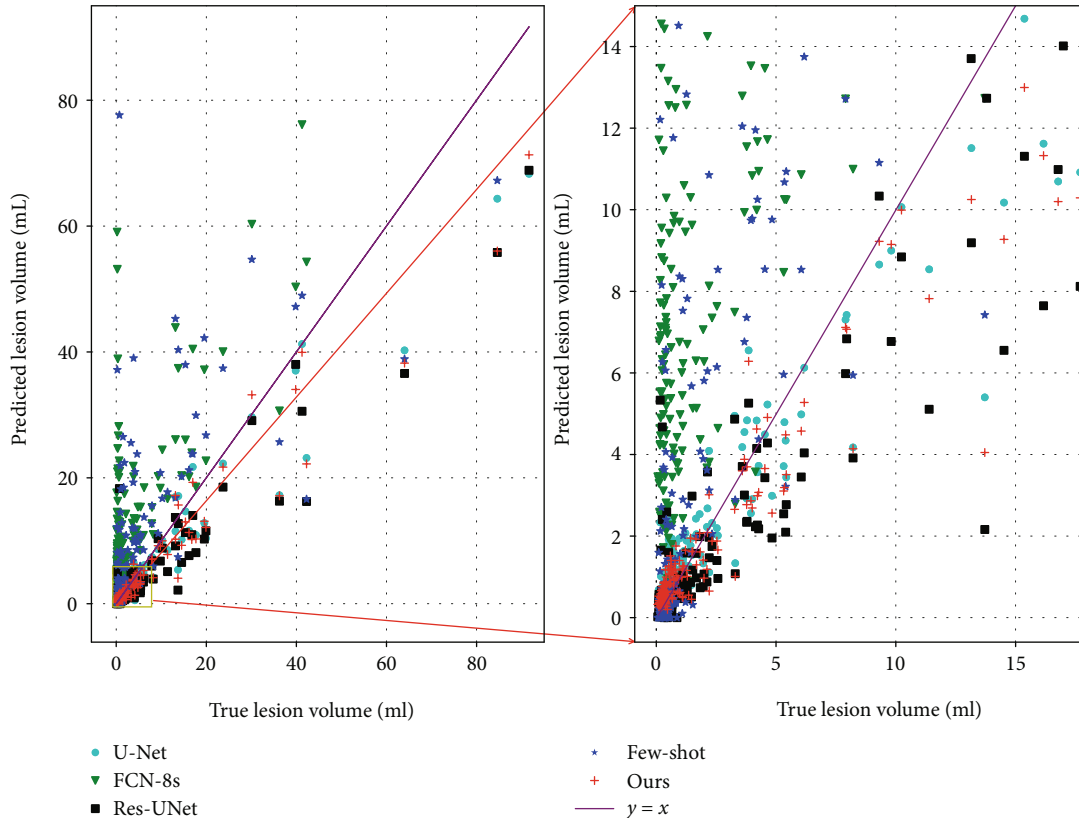


FIGURE 5: Predicted lesion volume versus ground truth volume.

TABLE 3: Evaluation results obtained by using different numbers of weakly labeled subjects on the training set. The mean DC is presented in the way of mean \pm standard deviation. The best result has been highlighted in italic.

Scale of the dataset	DC	P_L	R_L	F1	DR
79 subjects	0.557 \pm 0.250	0.793	0.741	0.766	0.922
159 subjects	0.665 \pm 0.181	<i>0.854</i>	0.872	0.863	0.966
238 subjects	0.675 \pm 0.138	0.843	0.901	0.871	1.000
318 subjects	<i>0.700 \pm 0.134</i>	0.821	0.920	0.867	1.000
398 subjects	0.699 \pm 0.128	0.852	<i>0.923</i>	<i>0.886</i>	1.000

to be diagnosed in clinical practice, especially when it is too small to be noticed. Therefore, it is very necessary to evaluate the performance on LI.

In this subsection, we divide the test set into the small lesion set and large lesion set. A subject is categorized into a small lesion subject only if all of the lesions are LI lesions. Otherwise, it will be included in the large lesion set. In the test set, there are 118 subjects and 61 subjects included in the small lesion set and the large lesion set, respectively. As Table 7 shows, we achieve a mean dice coefficient of 0.718 ± 0.120 on the large lesion set, while a mean dice coefficient of 0.689 ± 0.222 on the small lesion set. On other metrics, our proposed method achieves higher performance on the small lesion set.

In clinical diagnosis, large lesions are more easily diagnosed, while small lesions are not. Our proposed method achieves high performance not only on large lesions but also on small lesions.

4.6. Performance on the Public Dataset. To demonstrate the effectiveness of the proposed method, the performance on an external public dataset is further evaluated. In particular, we choose to use the training set of SPES in the ISLES2015 challenge [33]. Even though the SPES task is originally designed for ischemic stroke outcome prediction, the training set includes the ADC maps (known as DWI in SPES) and the corresponding AIS lesion annotations. We randomly split the subjects in the SPES training set into three sets, i.e., training set, validation set, and test set, with 5, 5, and 20 subjects, respectively.

The classification branch is trained on our institutional weakly labeled images with 398 weakly labeled ADC subjects, and the segmentation branch is trained on the new training set and the validation set. By noting that the public dataset and our institutional dataset were acquired from various MRI scanners with different parameters, the statistical property varies, which is known as domain adaption. As the classification branch is trained on our institutional data, the threshold of CAM has to be further tuned by using the validation set to adapt the SPES data.

For the sake of comparison, we also train and evaluate the methods used in Section 3.3. For fairness consideration, the

TABLE 4: Evaluation results by using different numbers of weakly labeled subjects with and without postprocessing. In particular, the mean dice coefficient is presented in the way of mean \pm standard deviation.

Scale of the dataset	Postprocessing	DC	P_L	R_L	F1	DR
398 subjects	No	0.651 ± 0.158	0.403	0.956	0.567	1.000
318 subjects		0.649 ± 0.157	0.391	0.949	0.554	1.000
238 subjects		0.630 ± 0.165	0.383	0.949	0.546	1.000
159 subjects		0.593 ± 0.184	0.297	0.949	0.452	1.000
79 subjects		0.620 ± 0.209	0.487	0.898	0.632	0.979
398 subjects	Yes	0.699 ± 0.128	0.852	0.923	0.886	1.000
318 subjects		0.700 ± 0.134	0.821	0.920	0.867	1.000
238 subjects		0.675 ± 0.138	0.843	0.901	0.871	1.000
159 subjects		0.665 ± 0.181	0.854	0.872	0.863	0.966
79 subjects		0.557 ± 0.250	0.793	0.741	0.766	0.922

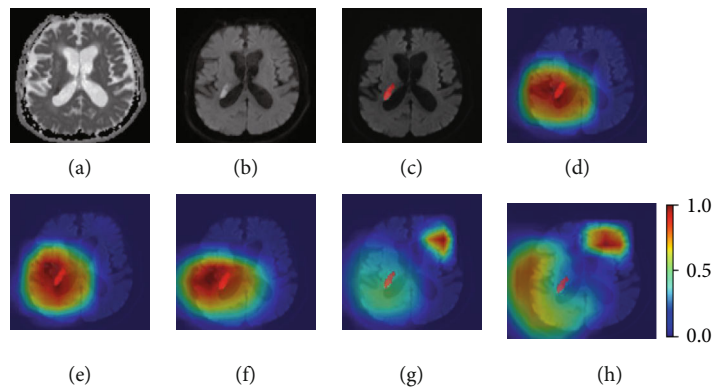


FIGURE 6: Examples of CAM. (a) ADC slice. (b) DWI slice. (c) Ground truth. (d) 398 subjects. (e) 318 subjects. (f) 238 subjects. (g) 159 subjects. (h) 79 subjects. The CAM and ground truth are depicted on the DWI. Best viewed in color.

TABLE 5: Evaluation results of single-modal and multimodal MR images. The mean DC is presented in the way of mean \pm standard deviation.

Modality	DC	P_L	R_L	F1	DR
ADC+DWI	0.699 ± 0.128	0.852	0.923	0.886	1.000
DWI	0.665 ± 0.166	0.743	0.876	0.804	0.989
ADC	0.451 ± 0.278	0.599	0.600	0.570	0.804

TABLE 6: Evaluation results of the MFMF-Network whose segmentation branch is trained on different data, where “all slices” means both the normal and lesion slices are used, and “lesion slices” means that only lesion slices are used. The best result has been highlighted in italic.

	DC	P_L	R_L	F1	DR
All slices	0.659 ± 0.124	0.702	<i>0.931</i>	0.801	1.000
Lesion slices	<i>0.699 ± 0.128</i>	<i>0.852</i>	0.923	<i>0.886</i>	1.000

TABLE 7: Evaluation results on large and small lesions. The best result has been highlighted in italic.

	DC	P_L	R_L	F1	DR
Large lesion set	<i>0.718 ± 0.120</i>	0.846	0.887	0.866	1.000
Small lesion set	0.689 ± 0.222	<i>0.858</i>	<i>0.962</i>	<i>0.907</i>	1.000

encoder parts of these methods are also pretrained as a classifier on our 398 weakly labeled ADC subjects. In particular, for the few-shot segmentation method proposed in [31], we split the slices of the new training set with AIS lesions into the support set and query set. Other experimental details are the same as used in Section 3.3 except that the validation loss determines when to stop the training.

Figure 7 plots some visualized examples on the test set. Similar to the results obtained on our institutional data, the proposed method achieves the best segmentation accuracy. As Figure 8 shows, the proposed method is able to achieve a mean dice coefficient of 0.651 ± 0.183 , which highlights the better capacity of our proposed method even in the cross-domain case.

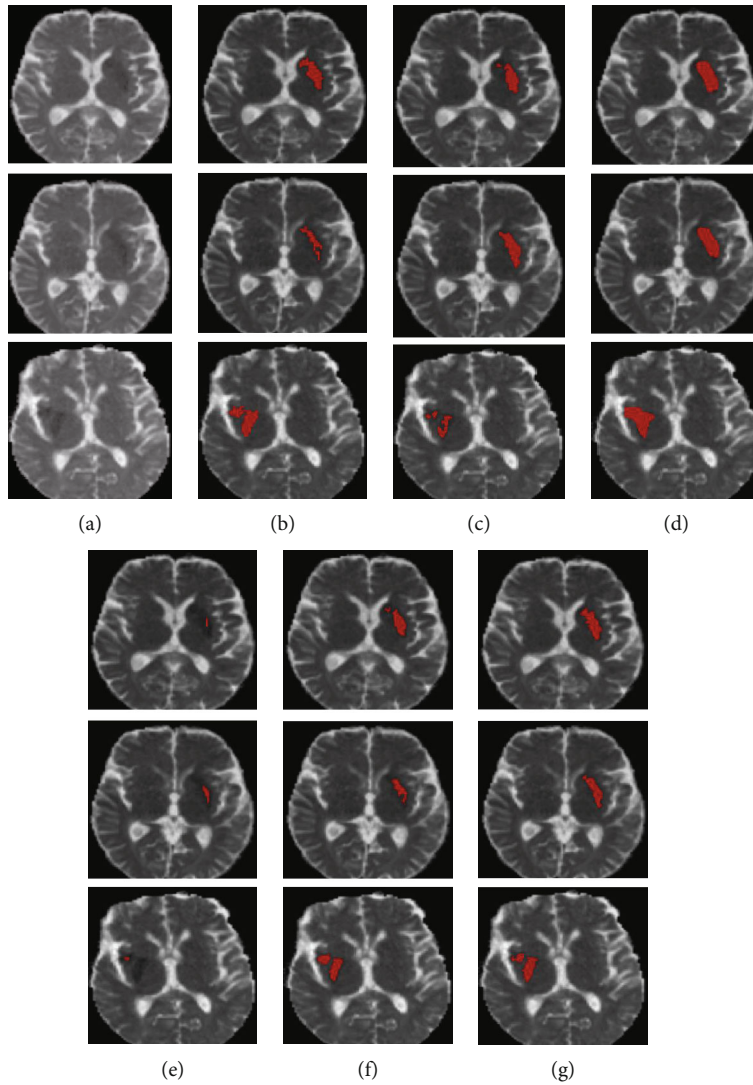


FIGURE 7: Visualization examples of the MRI slices and lesion segmentation results. (a, b) The original ADC map and ground truth, respectively. (c–g) The segmentation results of U-Net, FCN-8s, Res-UNet, the method in [31], and the proposed method, respectively. The segmentation results are overlaid on the ADCs and highlighted in red.

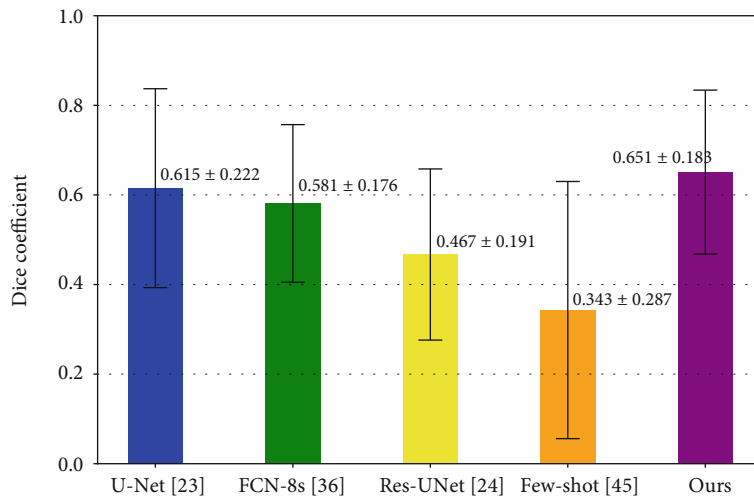


FIGURE 8: Bar plots of the dice coefficient for different methods.

5. Conclusion

In this paper, we proposed a deep learning-based method using a few fully labeled subjects for AIS lesion segmentation. Our proposed method consists of three processes: classification, segmentation, and inference. Since there are no pre-trained parameters available for processing medical images using CNN, some weakly labeled subjects are used to train the MFMF-Network to obtain a set of pretrained parameters in the classification process. Then, only 5 fully labeled subjects are used to train the segmentation branch.

The proposed method presents high performance on the clinical MR images with a mean dice coefficient of 0.699 ± 0.128 from the aspect of the pixel-level metric. More importantly, it presents a very high precision rate of 0.852 and recall rate of 0.923 from the lesion-wise metrics. Therefore, the proposed method can greatly reduce the expense of obtaining a large number of fully labeled subjects in a supervised setting, which is more meaningful in terms of engineering maneuverability.

Data Availability

The patient data used to support the findings of this study were supplied by Tianjin Huanhu Hospital, so they cannot be made freely available. The public dataset used in this paper is available at <http://www.isles-challenge.org/ISLES2015/>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Bin Zhao and Zhiyang Liu contributed equally to this work.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (61871239, 62076077) and the Natural Science Foundation of Tianjin (20JCQNJC0125).

References

- [1] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.
- [2] E. J. Benjamin, P. Muntner, A. Alonso et al., "Heart disease and stroke statistics-2019 update a report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.
- [3] J. Yang, A. Wong, Z. Wang et al., "Risk factors for incident dementia after stroke and transient ischemic attack," *Alzheimer's & Dementia*, vol. 11, no. 1, pp. 16–23, 2015.
- [4] N. Nabizadeh, M. Kubat, N. John, and C. Wright, "Automatic ischemic stroke lesion segmentation using single MR modality and gravitational histogram optimization based brain segmentation," in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCVR)*, p. 1, Las Vegas Nevada, USA, 2013.
- [5] J. Mitra, P. Bourgeat, J. Fripp et al., "Lesion segmentation from multimodal MRI using random forest following ischemic stroke," *NeuroImage*, vol. 98, pp. 324–335, 2014.
- [6] O. Maier, M. Wilms, J. von der Gableutz, U. Krämer, and H. Handels, "Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers," in *Medical Imaging 2014: Computer-Aided Diagnosis*, vol. 9035, p. 903504, San Diego, California, USA, March 2014.
- [7] J. Dolz, I. B. Ayed, and C. Desrosiers, "Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities," *International MICCAI Brainlesion Workshop*, 2018, pp. 271–282, Springer, 2018.
- [8] Z. Liu, C. Cao, S. Ding, Z. Liu, T. Han, and S. Liu, "Towards clinical diagnosis: automated stroke lesion segmentation on multi-spectral MR image using convolutional neural network," *IEEE Access*, vol. 6, pp. 57006–57016, 2018.
- [9] K. Kamnitsas, C. Ledig, V. F. J. Newcombe et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.
- [10] R. Karthik, R. Menaka, M. Hariharan, and D. Won, "Ischemic lesion segmentation using ensemble of multi-scale region aligned CNN," *Computer Methods and Programs in Biomedicine*, no. article 105831, 2020.
- [11] L. Liu, L. Kurgan, F.-X. Wu, and J. Wang, "Attention convolutional neural network for accurate segmentation and quantification of lesions in ischemic stroke disease," *Medical Image Analysis*, vol. 65, article 101791, 2020.
- [12] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 221–230, Honolulu, HI, USA, July 2017.
- [13] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," 2017, <https://arxiv.org/abs/1709.03410>.
- [14] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning," *BMVC*, vol. 3, no. 4, 2018.
- [15] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Few-shot segmentation propagation with guided networks," 2018, <https://arxiv.org/abs/1806.07373>.
- [16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, June 2009.
- [17] L. Chen, P. Bentley, and D. Rueckert, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *NeuroImage: Clinical*, vol. 15, pp. 633–643, 2017.
- [18] R. Zhang, L. Zhao, W. Lou et al., "Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional DenseNets," *IEEE Transactions on Medical Imaging*, vol. 37, no. 9, pp. 2149–2160, 2018.
- [19] O. Öman, T. Mäkelä, E. Salli, S. Savolainen, and M. Kangasniemi, "3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke," *European Radiology Experimental*, vol. 3, no. 1, p. 8, 2019.
- [20] C. Lucas, A. Kemmling, A. M. Mamlouk, and M. P. Heinrich, "Multi-scale neural network for automatic segmentation of ischemic strokes on acute perfusion images," in *2018 IEEE*

15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1118–1121, Washington, DC, USA, April 2018.

- [21] L. Liu, S. Chen, F. Zhang, F.-X. Wu, Y. Pan, and J. Wang, “Deep convolutional neural network for automatically segmenting acute ischemic stroke lesion in multi-modality MRI,” *Neural Computing and Applications*, vol. 32, pp. 1–14, 2019.
- [22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, Las Vegas Nevada, USA, 2016.
- [23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, <https://arxiv.org/abs/1409.1556>.
- [24] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, Utah, USA, 2018.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: surpassing human-level performance on ImageNet classification,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, Santiago, Chile, 2015.
- [26] L. Liu, H. Jiang, P. He et al., “On the variance of the adaptive learning rate and beyond,” 2019, <https://arxiv.org/abs/1908.03265>.
- [27] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, “The design of SimpleITK,” *Frontiers in Neuroinformatics*, vol. 7, p. 45, 2013.
- [28] P. A. Yushkevich, J. Piven, H. C. Hazlett et al., “User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability,” *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, 2006.
- [29] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241, Springer, 2015.
- [30] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, Massachusetts, USA, 2015.
- [31] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, “Fss-1000: a 1000-class dataset for few-shot segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2869–2878, Virtual, 2020.
- [32] J. Lodder, “Size criterion for lacunar infarction,” *Cerebrovascular Diseases*, vol. 24, no. 1, p. 156, 2007.
- [33] O. Maier, B. H. Menze, J. von der Gabelentz et al., “ISLES 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI,” *Medical Image Analysis*, vol. 35, pp. 250–269, 2017.

Retraction

Retracted: miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] J. Li, X. He, X. Wu, X. Liu, Y. Huang, and Y. Gong, "miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 2953598, 10 pages, 2020.

Research Article

miR-139-5p Inhibits Lung Adenocarcinoma Cell Proliferation, Migration, and Invasion by Targeting MAD2L1

Jianfeng Li,¹ Xi He,¹ Xiaotang Wu,² Xiaohui Liu,¹ Yixiong Huang ,³ and Yuchen Gong ⁴

¹Department of Thoracic Surgery, Tangshan People's Hospital, Tangshan, China

²Shanghai Engineering Research Center of Pharmaceutical Translation, Shanghai, China

³Department of Thoracic Surgical Oncology, Fujian Cancer Hospital & Fujian Medical University Cancer Hospital, Fujian, China

⁴Department of Respiration, China Coast Guard of the Chinese People's Armed Police Force Hospital, Zhejiang Province, China

Correspondence should be addressed to Yuchen Gong; gongyuchennn@163.com

Received 24 June 2020; Accepted 22 July 2020; Published 4 November 2020

Guest Editor: Tao Huang

Copyright © 2020 Jianfeng Li 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.

Background. miR-139-5p is lowly expressed in various human cancers and exerts its antitumor effect through different molecular mechanisms, yet the molecular mechanism of miR-139-5p in lung adenocarcinoma (LUAD) remains to be further elucidated. The study is aimed at investigating the role and the regulatory mechanism of miR-139-5p in LUAD progression. **Methods.** Differential analysis was performed on miRNA expression data in the TCGA-LUAD dataset. qRT-PCR was employed to detect the transcription levels of miR-139-5p and MAD2L1 in LUAD cells, while western blot was carried out for the detection of MAD2L1 protein expression. CCK-8 and Transwell assays were implemented to assess LUAD cell proliferation, migration, and invasion. A dual-luciferase reporter gene assay was conducted to verify the direct targeting relationship between miR-139-5p and MAD2L1. **Results.** miR-139-5p was significantly downregulated in LUAD cells in comparison with that in human normal bronchial epithelial cells. Overexpressing miR-139-5p inhibited LUAD cell proliferation, migration, and invasion, while opposite results could be observed when miR-139-5p was inhibited. MAD2L1 was identified as a direct target of miR-139-5p in LUAD. Besides, the inhibitory effect of miR-139-5p overexpression on LUAD cell proliferation, migration, and invasion was attenuated by overexpressing MAD2L1. **Conclusion.** Our study suggests that miR-139-5p is lowly expressed in LUAD cells and inhibits LUAD cell proliferation, migration, and invasion by targeted suppressing MAD2L1 expression. It is of potential significance for the prognosis and treatment of LUAD.

1. Introduction

Lung adenocarcinoma (LUAD), the most common type of nonsmall cell lung cancer (NSCLC), is characterized by dense lymphocyte infiltration and is prone to metastasize at early stages [1]. Different medical interventions, such as chemotherapy, surgical removal, and radiotherapy, are the conventional treatments for LUAD. However, these treatments lack specificity and will also do harm to adjacent normal cells [2], which make the treatment for LUAD evolve from cytotoxic chemotherapy to personalized treatment based on molecular alterations [3]. In recent years, the identification of oncogenes and the use of immunotherapy have already changed the treatment strategies for LUAD, but the survival rate still remains low [4]. Therefore, it is of paramount importance

to find a novel therapeutic target to improve the treatment for LUAD.

Noncoding RNAs have always been a hot topic in the cancer field for years, especially microRNAs (miRNAs). The reason is that they are key players in mediating different molecular processes and they participate in tumorigenesis more often than protein-coding genes [5]. miRNAs are involved in the control of several cancer-related processes, such as proliferation, apoptosis, migration, and invasion. Additionally, miRNAs are also involved in many other diseases, such as metabolic disorders [6]. miR-197-3p, as reported, serves as an oncogene in LUAD to promote LUAD cell proliferation and inhibit cell apoptosis by downregulating lysine 63 deubiquitinase (CYLD) [7]. miR-938 exerts its cancer-promoting role in LUAD by targeting RBM5 [8]. As

a tumor suppressor, miR-144-3p inhibits LUAD cell proliferation and invasion by increasing the EZH2 expression [9]. These findings indicate that miRNAs have a great potential in the diagnosis and targeted therapy of LUAD.

It is reported that miR-139-5p is downregulated in various cancers and exerts its antitumor role by different molecular mechanisms. For example, miR-139-5p plays an antitumor role in cervical cancer and inhibits Wnt/ β -catenin signal transduction by targeting transcription factor 4 (TCF4) [10]. In oral squamous cell carcinoma, miR-139-5p inhibits cell proliferation and metastasis by suppressing HOXA9 expression [11]. Moreover, miR-139-5p acts as a tumor suppressor by regulating SOX5 in prostate cancer cells [12]. However, the mechanism of miR-139-5p underlying LUAD cell proliferation, migration, and invasion remains to be improved and supplemented.

In this study, we made an attempt to explore the expression and role of miR-139-5p in LUAD and the underlying molecular mechanism of miR-139-5p in regulating LUAD cell proliferation, migration, and invasion. Our study may bring additional insights into the molecular mechanism underlying LUAD progression and provide potential indicators for the diagnosis and prognosis of LUAD.

2. Materials and Methods

2.1. Bioinformatics Analysis. Expression data of miRNAs and mRNAs of the TCGA-LUAD dataset were downloaded from the TCGA database (<https://portal.gdc.cancer.gov/>), of which miRNA expression data were obtained from 46 normal tissue samples and 521 tumor tissue samples, and mRNA expression data were obtained from 59 normal tissue samples and 535 tumor tissue samples. Expression analysis was performed on miR-139-5p according to the obtained data. Differential analysis was carried out using “edgeR” package with threshold set as $|\log_{2}FC| > 2.0$, $p < 0.01$, and then differentially expressed mRNAs (DEmRNAs) were obtained. Three databases miRDB (<http://mirdb.org/>), miDIP (<http://ophid.utoronto.ca/mirDIP/index.jsp#r>), and starBase (<http://starbase.sysu.edu.cn/>) were employed to predict the target genes of miR-139-5p. Candidate genes obtained from the intersection of DEmRNAs and predicted target genes of miR-139-5p were subjected to Pearson correlation analysis, and the mRNA showing the highest negative correlation coefficient was selected as the object of the study.

2.2. Cell Culture. LUAD cell lines A549 (BNCC337696), PC-9 (BNCC340767), H1975 (BNCC340345), H1650 (BNCC100260), and human normal bronchial epithelial cell line BEAS-2B (BNCC338205) were all purchased from BeNa Culture Collection (Beijing, China). All cell lines were cultured in 100 U RPMI-1640 mediums supplemented with 10% fetal bovine serum (FBS), 100 U/ml penicillin (Invitrogen, Grand Island, NY, USA), and 100 μ g/ml streptomycin (Invitrogen, Grand Island, NY, USA), and maintained in an incubator with 5% CO₂ at 37°C.

2.3. Cell Transfection. miR-139-5p mimic, miR-139-5p inhibitor, oe-MAD2L1, and their corresponding negative controls

(NC) were accessed from RiboBio (Guangzhou, China). Cells were grown in antibiotic-free complete mediums 24 h before transfection. Lipofectamine 2000 (Thermo Fisher Scientific, Inc.) was employed to transfect cells at a concentration of 50 nM according to the manufacturer’s protocol.

2.4. qRT-PCR. After 48 h of transfection, total RNA was extracted from LUAD cells using a Trizol kit (Invitrogen Life Technologies, Carlsbad, CA, USA). miRNA and mRNA were reversely transcribed into cDNA by the PrimeScript RT kit (Takara, Japan). qRT-PCR was performed by SYBR Premix Ex Taq (Takara) under the Applied Biosystems ABI 7500 Real-Time PCR System (Thermo Fisher Scientific, Inc.). Primers were shown in Table 1. U6 and GAPDH were applied as endogenous references of miRNA and mRNA, respectively. The relative expression was analyzed by the $2^{-\Delta\Delta Ct}$ method.

2.5. Western Blot. Transfected cells were washed with phosphate-buffered saline (PBS) twice and then were lysed on ice with RIPA loading buffer (Thermo Fisher Scientific, MA, USA) containing protease and phosphatase inhibitor (Solarbio). Protein concentration was assayed by the BCA kit (Beyotime). The protein samples were separated by 10% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE; 50 μ g/lane) and then transferred onto polyvinylidene fluoride (PVDF) membranes (ZY-160FP, Zeye Bio Co., Ltd., Shanghai, China). After being blocked with 5% skim milk for 2 h at 37°C and washed with Tris-Buffered Saline Tween (TBST) three times, the membranes were incubated with rabbit polyclonal anti-MAD2L1 antibody (1:1000, ab97777, Abcam, Cambridge, UK) at 4°C overnight. Rabbit monoclonal anti-GAPDH antibody (1:2500, ab9485, Abcam, Cambridge, UK) was taken as control. After being washed with TBST, the membranes were incubated with goat anti-rabbit IgG H&L (1:2000, ab205718, Abcam, Cambridge, UK) for 2 h. Finally, protein bands were visualized using an electrochemiluminescence kit (ECL; Pierce Biotechnology) and analyzed by imaging system (ZG11SCIBRIGHTCL, Bio-Rad, CA, USA).

2.6. CCK-8. Cell-counting kit-8 (CCK-8) assay was used for detection of cell proliferation in different transfection groups. Cells (2×10^3) were seeded into 96-well plates (Corning Costar, Corning, NY), and then 10 μ l of CCK-8 solution (Beyotime, Nanjing, Jiangsu, China) was added into plates at 0 h, 24 h, 48 h, and 72 h for 2 h of incubation, respectively. The absorbance at 450 nm was measured by an enzyme-labeled instrument (BioTek Company, Winooski, VT, USA) to evaluate cell viability.

2.7. Colony Formation Assay. After 24 h of transfection, A549 cells were inoculated into a 6-well plate with 1×10^3 cells/well, and each treatment group was made in triplicate. Cells were cultured in a complete medium for one week until clear colonies were formed. Cell colonies were fixed with 70% methanol for 5 min and stained with 0.5% crystal violet (Thermo Fisher, USA). Each well was washed with sterile water to remove residual crystal violet. Colonies with more

TABLE 1: Primer sequences in qRT-PCR.

Gene	Forward	Reverse
miR-139-5p	5'-TCTACAGTGCACGTGTCTCCA G-3'	5'-GTGCAGGGTCCGAGGT-3'
U6	5'-TGCGGG TGCTCGTTCGGCAG C-3'	5'-GTGCAGGGTCCGAGGT-3'
MAD2L1	5'-GTTCTTCTCATTTCGGCATCAACA-3'	5'-GAGTCCGTATTTCTGCACTCG-3'
GAPDH	5'-GGAGCGAGATCCCTCCAAAAT-3'	5'-GGCTGTTGTCATACTTCTCATGG-3'

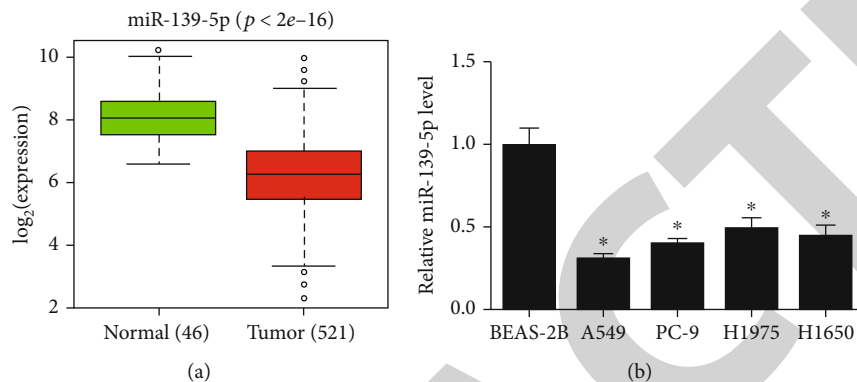


FIGURE 1: miR-139-5p is downregulated in LUAD cells. (a) Box plots of miR-139-5p expression in the TCGA-LUAD dataset; (b) The expression level of miR-139-5p in LUAD cell lines A549, PC-9, H1975, H1650, and human normal bronchial epithelial cell line BEAS-2B was detected by qRT-PCR; * $p < 0.05$.

than 50 cells were identified, and the number of colonies per well was calculated.

2.8. Transwell Migration and Invasion Assays. Transwell assay was applied to evaluate cell migration and invasion. Trypsinized cells were collected after 48 h of transfection. For cell invasion assay, 88 μm pore size inserts (Transwell; Costar, High Wycombe, UK) were placed into 24-well plates to separate the upper chambers from the lower chambers. A total of 200 μl cell suspension containing $(3-5) \times 10^4$ cells was added into the upper chambers precoated with Matrigel (BD Biosciences, San Jose, CA, USA). 500 μl RPMI-1640 medium supplemented with 20% FBS was added into the lower chambers. Cells invading the lower chambers were then fixed with 70% methanol and treated by 0.5% crystal violet after 24 h. The invaded cells were counted under an inverted microscope (Olympus IX83; Olympus Corporation, Tokyo, Japan). The procedure of cell migration assay was similar to the invasion assay, except that the upper chambers were not precoated with Matrigel.

2.9. Dual-Luciferase Reporter Gene Assay. Amplified wild type (WT) and mutant (MUT) MAD2L1 3'UTR were accessed from Shanghai GenePharma Co., Ltd. and inserted into luciferase vector PmirGLO. PmirGLO-MAD2L1-Wt/PmirGLO-MAD2L1-Mut and miR-139-5p mimic/NC mimic were cotransfected into LUAD cell lines by using LipofectamineTM2000 kit (Invitrogen). After 48 h of transfection, the relative activity of luciferase was assayed by dual-luciferase reporter gene system (Promega Corporation) according to the manufacturer's instructions.

2.10. Statistical Analysis. All data were analyzed by GraphPad Prism 6.0 (La Jolla, CA). Each experiment was repeated three times. The results were presented by means \pm standard deviation (SD). Student's t -test was used for comparison between two groups. $p < 0.05$ was considered statistically significant.

3. Results

3.1. miR-139-5p Is Downregulated in LUAD Cells. miR-139-5p expression was searched in the TCGA-LUAD dataset, and it was found that miR-139-5p was significantly downregulated in LUAD tissues (Figure 1(a)). qRT-PCR was employed to detect the expression of miR-139-5p in LUAD cell lines A549, PC-9, H1975, H1650, and human normal bronchial epithelial cell line BEAS-2B, exhibiting that the expression of miR-139-5p was remarkably downregulated in LUAD cell lines relative to that in human normal bronchial epithelial cell line (Figure 1(b)). A549 cell line with the lowest expression level of miR-139-5p was selected for subsequent experiments.

3.2. miR-139-5p Inhibits LUAD Cell Proliferation, Migration, and Invasion. To verify the biological function of miR-139-5p in LUAD, miR-139-5p mimic or miR-139-5p inhibitor and their corresponding NC were transfected into A549 cells for evaluation of cell proliferation, migration, and invasion. First of all, qRT-PCR was used to detect the expression level of miR-139-5p in different groups, and the results displayed that the expression of miR-139-5p after transfection met the requirements (Figure 2(a)), and the successfully transfected cells could be used in subsequent experiments. Cell

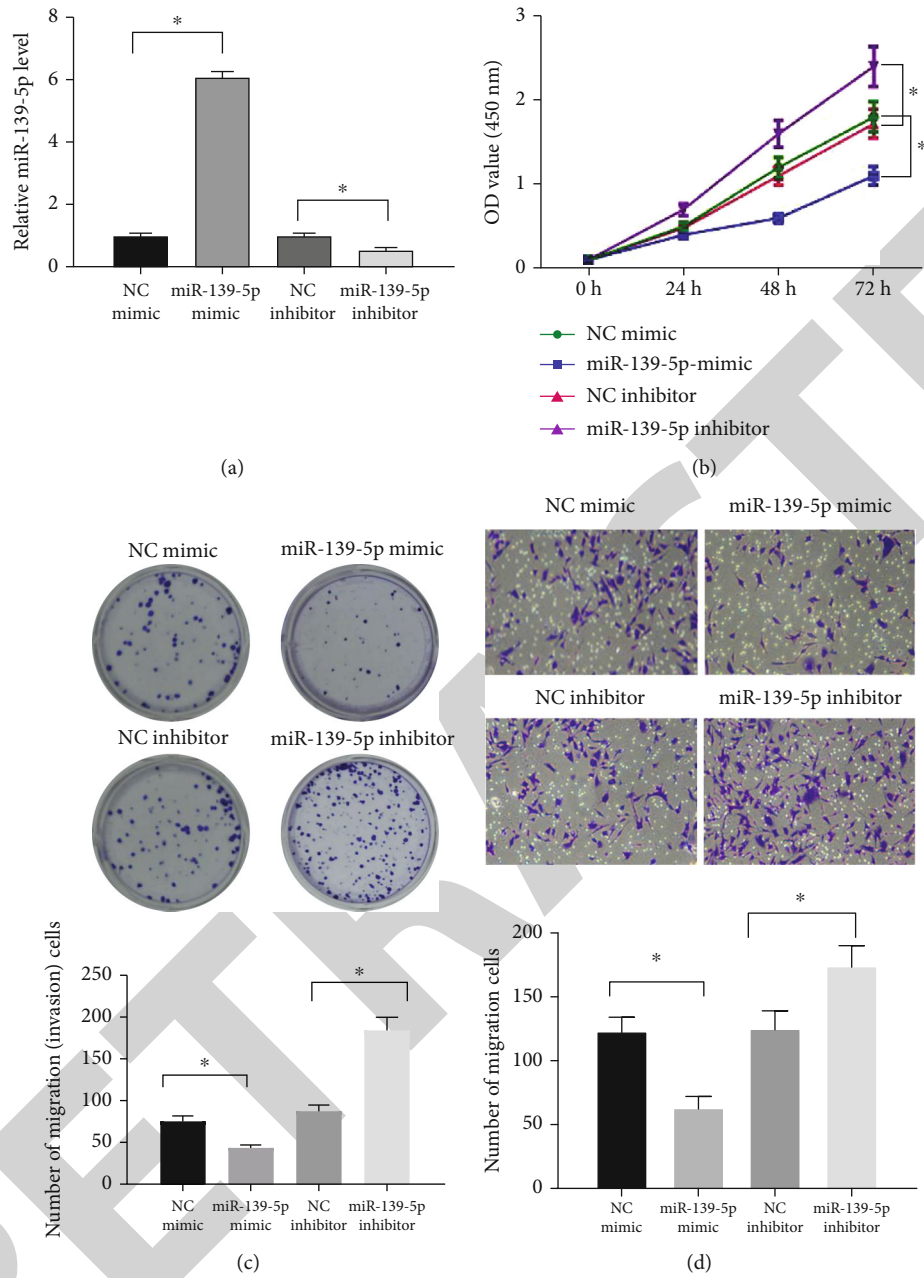


FIGURE 2: Continued.

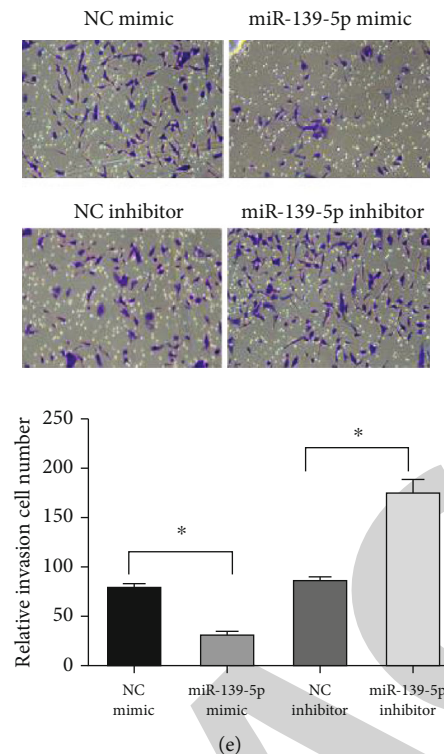


FIGURE 2: miR-139-5p inhibits LUAD cell proliferation, migration, and invasion. (a) qRT-PCR was used to detect the transfection efficiency of miR-139-5p in A549 cells; (b) CCK-8 was performed for the detection of cell viability in different transfection groups; (c) Colony formation assay was performed to detect the colony formation ability of A549 cells in different transfection groups; (d, e) Transwell assay (100 \times) was carried out for assessment of cell (d) migration and (e) invasion in each transfection group; * $p < 0.05$.

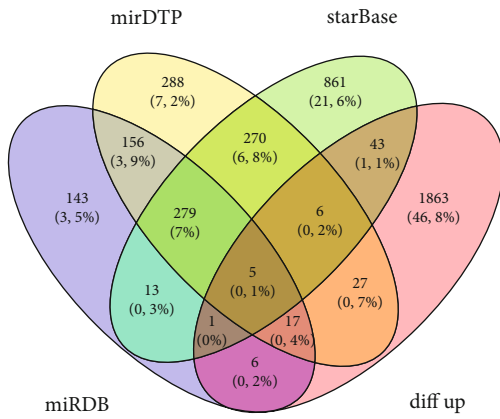
biological behaviors in each group were sequentially assayed. As revealed by CCK-8, we found that the viability of A549 cells transfected with miR-139-5p mimic was markedly lower than that of A549 cells transfected with NC mimic, while the cell proliferative ability was significantly increased in the miR-139-5p inhibitor than that in the NC inhibitor group (Figure 2(b)). The results of colony formation assay indicated that the colony formation ability of miR-139-5p overexpressed cells was significantly inhibited, while that of LUAD cells was remarkably improved after miR-139-5p was inhibited (Figure 2(c)). Next, the Transwell assay illustrated that the migration and invasion of A549 cells transfected with miR-139-5p mimic were considerably inhibited, while those of A549 cells transfected with miR-139-5p inhibitor were increased (Figures 2(d) and 2(e)). Collectively, these findings suggested that miR-139-5p inhibited A549 cell proliferation, migration, and invasion as a tumor suppressor in LUAD.

3.3. MAD2L1 Is a Direct Target of miR-139-5p. We then explored the underlying molecular mechanism of miR-139-5p in LUAD. Databases including miRDB, miDIP, and starBase were firstly implemented for target prediction for miR-139-5p, and five candidate genes (NPTX1, ELAVL2, FBN2, GPR37, and MAD2L1) obtained from the intersection of predicted genes and upregulated DE mRNAs were subjected to Pearson correlation analysis with miR-139-5p. The result showed that MAD2L1 had the highest negative correlation coefficient with miR-139-5p (Figures 3(a)–3(c)). Expression analysis was performed on MAD2L1 in the

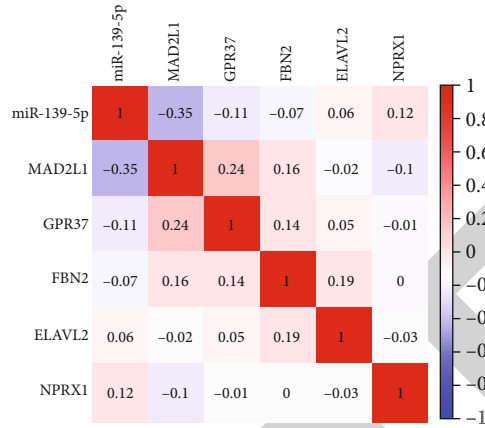
TCGA-LUAD dataset, which discovered that MAD2L1 was noticeably highly expressed in LUAD tissue, and LUAD patients with high MAD2L1 expression had relatively low overall survival (OS) (Figures 3(d) and 3(e)). We speculated that MAD2L1 might be a direct target of miR-139-5p based on the result of bioinformatics analysis.

To validate our speculation, miR-139-5p mimic or miR-139-5p inhibitor and their corresponding NC were firstly transfected into A549 cells, and then qRT-PCR and western blot were conducted to determine the transcription level and protein expression level of MAD2L1. The results indicated that the mRNA and protein expression levels of MAD2L1 were significantly downregulated after miR-139-5p was overexpressed, whereas opposite results were observed when miR-139-5p was suppressed (Figures 3(f) and 3(g)). Furthermore, the binding sites of miR-139-5p on MAD2L1 3'UTR were predicted by the starBase database (Figure 3(h)) and then verified by dual-luciferase assay. We found that the luciferase activity of A549 cells transfected with miR-139-5p mimic and MAD2L1-Wt was decreased, while that of A549 cells cotransfected with miR-139-5p mimic and MAD2L1-Mut exhibited no marked change (Figure 3(i)). Taken together, these findings elucidated that MAD2L1 was a direct target of miR-139-5p and was negatively regulated by miR-139-5p.

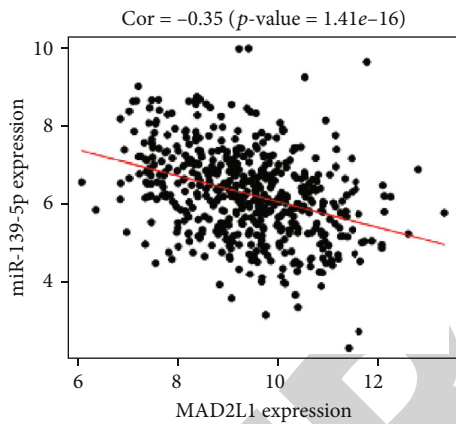
3.4. MAD2L1 Mediates the Effect of miR-139-5p on LUAD Cells. To investigate whether miR-139-5p inhibited LUAD cell proliferation, migration, and invasion by regulating



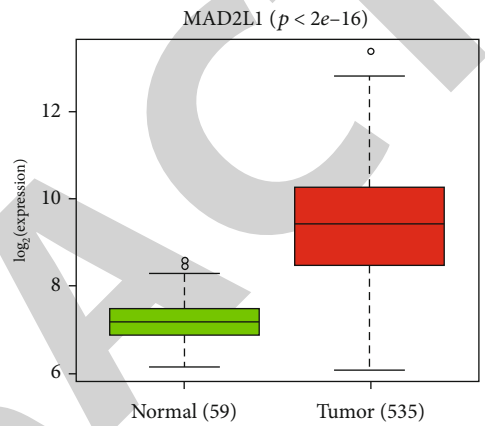
(a)



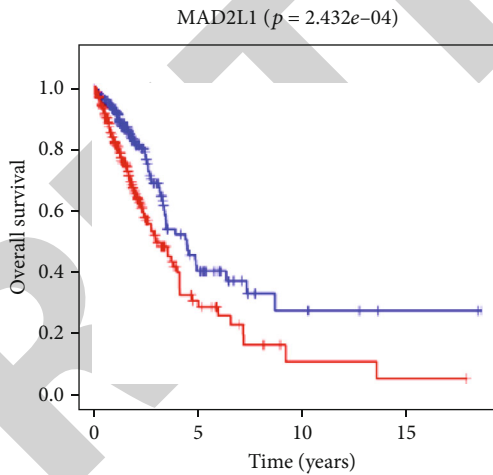
(b)



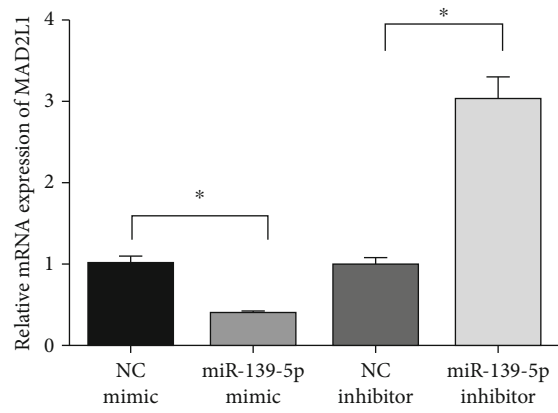
(c)



(d)



(e)



(f)

FIGURE 3: Continued.

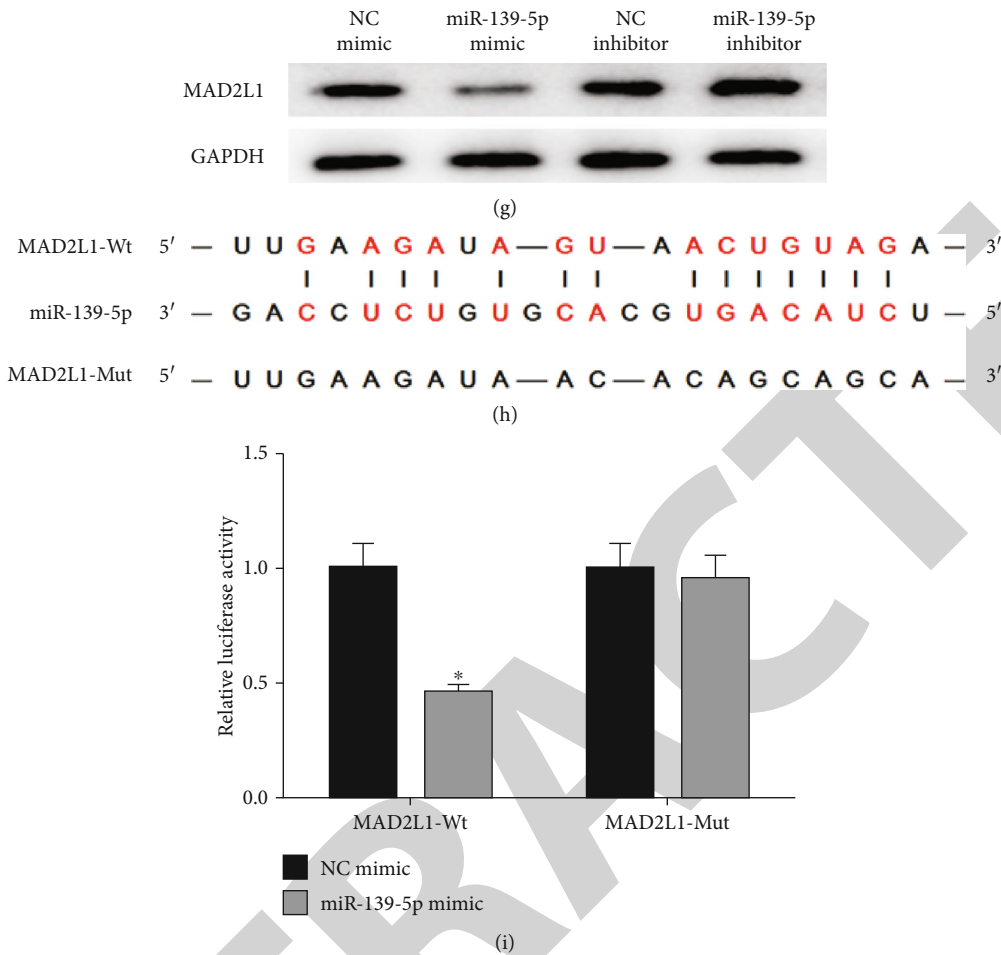


FIGURE 3: miR-139-5p targeted binds to MAD2L1 and negatively regulates MAD2L1 expression. (a) Venn diagram of target genes of miR-139-5p predicted by miRDB, miDIP, and starBase databases and upregulated DEMRNAs in the TCGA-LUAD; (b) Pearson correlation of miR-139-5p and five candidate genes NPTX1, ELAVL2, FBN2, GPR37, and MAD2L1; (c) Pearson correlation of miR-139-5p and MAD2L1; (d) Box plots of MAD2L1 expression in LUAD tissue and normal tissue in the TCGA-LUAD dataset; (e) Survival curves of patients with high expression of MAD2L1 (red) and low expression of MAD2L1 (blue). The abscissa refers to the time (in years) and the ordinate refers to survival rate; (f) qRT-PCR was used to detect the mRNA expression level of MAD2L1 after transfection of miR-139-5p mimic or miR-139-5p inhibitor; (g) Western blot was employed to examine the protein expression of MAD2L1 after transfection; (h) starBase database was used to predict the binding sites of miR-139-5p on MAD2L1 3'UTR; (i) Dual-luciferase reporter gene assay was used for verification of the targeted binding relationship between miR-139-5p and MAD2L1; * $p < 0.05$.

MAD2L1, we designed three groups: NC mimic+oe-NC, miR-139-5p mimic+oe-NC, and miR-139-5p mimic+oe-MAD2L1. We found that the inhibitory effect of miR-139-5p overexpression on MAD2L1 expression could be reversed by overexpressing MAD2L1 (Figures 4(a) and 4(b)). CCK-8 and colony formation assays suggested that the overexpression of miR-139-5p significantly inhibited the proliferation of LUAD cells, while the overexpression of MAD2L1 reversed the inhibitory effect of miR-139-5p on cell proliferation (Figures 4(c) and 4(d)). Transwell assay was conducted for testing cell migration and invasion. The results illustrated that overexpressing miR-139-5p markedly inhibited cell migration and invasion, whereas overexpressing MAD2L1 reversed the inhibitory effect of miR-139-5p on cell behaviors (Figures 4(e) and 4(f)). Therefore, miR-139-5p inhibited LUAD cell proliferation, migration, and invasion by regulating MAD2L1.

4. Discussion

miRNAs are capable of interfering transcriptional signal transduction and regulating the key processes of cells, thus playing vital roles in the occurrence and development of cancers [13]. It has been reported that the differential expression of miRNAs between normal lung and cancerous lung leads to the emergence of novel biomarkers, which is conducive to the screening of high-risk groups and helps the diagnosis and treatment of lung cancer [14]. Up to now, the potential miRNAs that regulate the progression of LUAD have not been fully identified.

In this study, miR-139-5p expression was searched in the TCGA-LUAD dataset, finding that miR-139-5p was down-regulated in LUAD tissue. miR-139-5p was lowly expressed in LUAD cell lines as evidenced by qRT-PCR, and the result was consistent with the expression of miR-139-5p in

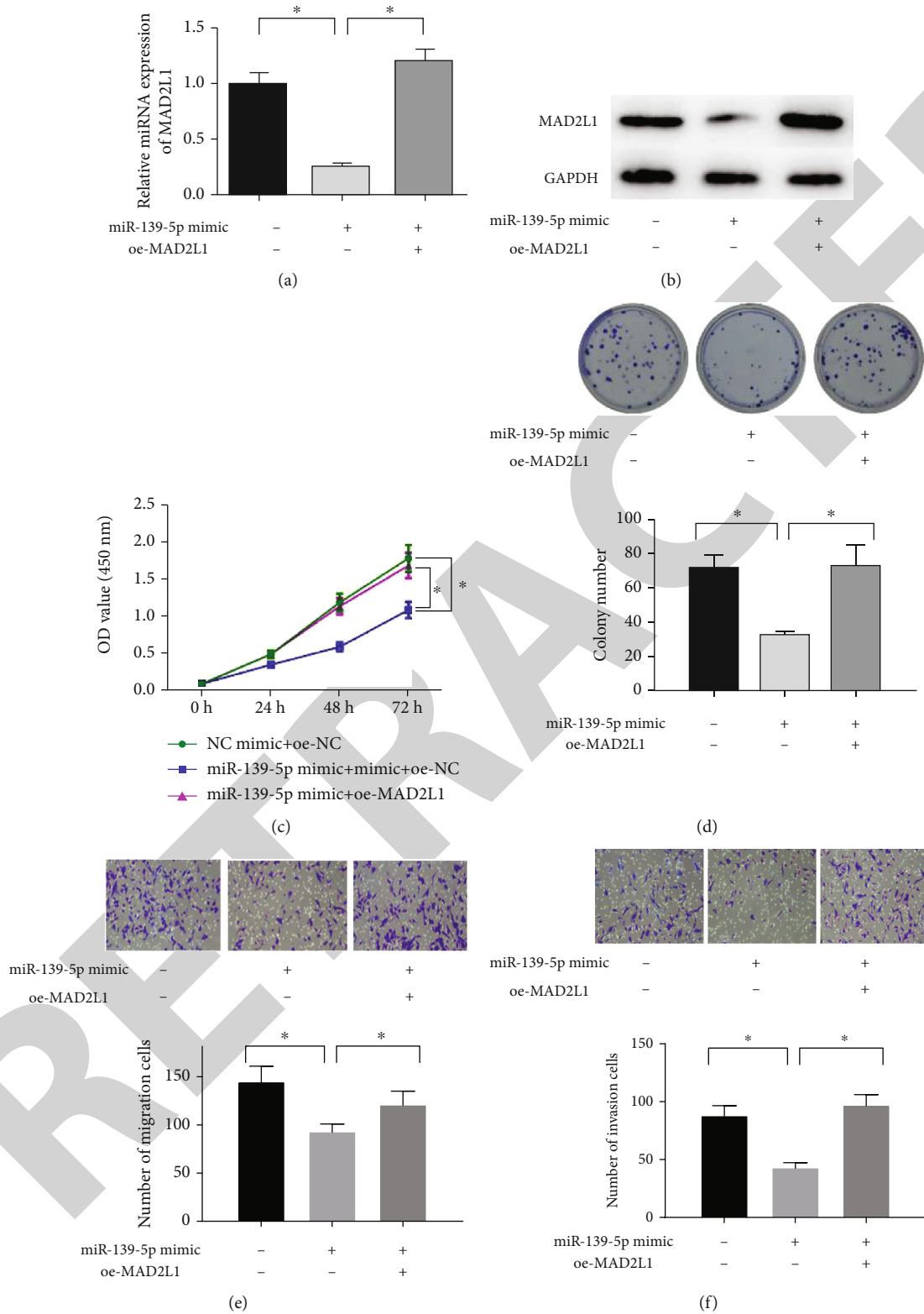


FIGURE 4: MAD2L1 mediates the effect of miR-139-5p on LUAD cells. (a, b) qRT-PCR and western blot were carried out to assess the mRNA and protein expression of MAD2L1 in each treatment group; (c, d) CCK-8 and colony formation assays were conducted for the assessment of cell proliferation in different treatment groups; (e, f) Transwell assay (100 \times) was performed to assess (e) cell migration and (f) invasion in different treatment groups; * $p < 0.05$.

hepatocellular carcinoma (HCC) [15], endometrial carcinoma [16], and gallbladder carcinoma [17]. In osteosarcoma, overexpressing miR-139-5p inhibits cell proliferation, migration, and invasion, while loss of miR-139-5p facilitates cell proliferation, migration, and invasion [18], and similar trends could be observed in LUAD in this study. These findings indicate that miR-139-5p acts as a tumor suppressor in LUAD.

Due to the fact that miRNAs can regulate the growth and metastasis of tumor by various molecular mechanisms [19], we performed bioinformatics analysis to predict the target gene of miR-139-5p, and MAD2L1 was identified as a potential target of miR-139-5p. A dual-luciferase reporter gene assay was conducted and confirmed the targeting relationship between the two genes. Besides, MAD2L1 was found to be highly expressed in LUAD tissue, and LUAD patients with high MAD2L1 expression had relatively low OS, which were consistent with the results of the study made by Li et al. regarding to MAD2L1 expression in HCC [20]. MAD2L1 is reported to be a key player in maintaining the function of spindle assemble checkpoint. The overexpression of MAD2L1 in spindle assemble checkpoint can result in the instability and aneuploidy of chromosome, and the genetic variation of MAD2L1 can lead to lung cancer susceptibility [21, 22]. Therefore, we further confirmed whether miR-139-5p exerted its antitumor role in LUAD by suppressing MAD2L1. miR-139-5p and MAD2L1 were simultaneously overexpressed in LUAD cells, and we found that the inhibitory effect of miR-139-5p overexpression on cell proliferation, migration, and invasion was reversed by MAD2L1 overexpression. Hence, we believed that miR-139-5p regulated LUAD cell proliferation, migration, and invasion by targeting MAD2L1.

Generally speaking, we elucidated that miR-139-5p was lowly expressed in LUAD and inhibited LUAD cell proliferation, migration, and invasion. Besides, MAD2L1 was identified as a direct target of miR-139-5p in LUAD, and miR-139-5p exerted its antitumor role by inhibiting MAD2L1 expression. Our discovery not only lays a molecular foundation for exploration of the mechanism of miR-139-5p in LUAD but also provides a potential target for the treatment of LUAD.

Data Availability

The data and materials in the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no potential conflicts of interest.

Authors' Contributions

Jianfeng Li contributed to the study design. Xi He conducted the literature search and performed data analysis. Xiaotang Wu acquired the data and wrote the article. Xiaohui Liu drafted. Yuchen Gong revised the article and gave the final approval of the version to be submitted. All authors read and approved the final manuscript.

References

- [1] C. Luo, M. Lei, Y. Zhang et al., "Systematic construction and validation of an immune prognostic model for lung adenocarcinoma," *Journal of Cellular and Molecular Medicine*, vol. 24, no. 2, pp. 1233–1244, 2020.
- [2] P. Sharma, M. Mehta, D. S. Dhanjal et al., "Emerging trends in the novel drug delivery approaches for the treatment of lung cancer," *Chemico-Biological Interactions*, vol. 309, article 108720, 2019.
- [3] L. Sorber, K. Zwaenepoel, V. Deschoolmeester et al., "Circulating cell-free nucleic acids and platelets as a liquid biopsy in the provision of personalized therapy for lung cancer patients," *Lung Cancer*, vol. 107, pp. 100–107, 2017.
- [4] E. K. Kleczko, J. W. Kwak, E. L. Schenk, and R. A. Nemenoff, "Targeting the complement pathway as a therapeutic strategy in lung cancer," *Frontiers in Immunology*, vol. 10, p. 954, 2019.
- [5] T. Guo, J. Li, L. Zhang et al., "Multidimensional communication of microRNAs and long non-coding RNAs in lung cancer," *Journal of Cancer Research and Clinical Oncology*, vol. 145, no. 1, pp. 31–48, 2019.
- [6] M. Acunzo and C. M. Croce, "MicroRNA in cancer and cachexia—a mini-review," *The Journal of Infectious Diseases*, vol. 212, Supplement 1, pp. S74–S77, 2015.
- [7] Y. Chen and C. Yang, "miR1973pinduced downregulation of lysine 63 deubiquitinase promotes cell proliferation and inhibits cell apoptosis in lung adenocarcinoma cell lines," *Molecular Medicine Reports*, vol. 17, pp. 3921–3927, 2017.
- [8] T. Qian, S. Shi, L. Xie, and Y. Zhu, "miR-938 promotes cell proliferation by regulating RBM5 in lung adenocarcinoma cells," *Cell Biology International*, vol. 44, no. 1, pp. 295–305, 2019.
- [9] C. Liu, Z. Yang, Z. Deng et al., "Downregulated miR-144-3p contributes to progression of lung adenocarcinoma through elevating the expression of EZH2," *Cancer Medicine*, vol. 7, no. 11, pp. 5554–5566, 2018.
- [10] X. Ji, H. Guo, S. Yin, and H. Du, "miR-139-5p functions as a tumor suppressor in cervical cancer by targeting TCF4 and inhibiting Wnt/ β -catenin signaling," *Oncotargets and Therapy*, vol. 12, pp. 7739–7748, 2019.
- [11] K. Wang, J. Jin, T. Ma, and H. Zhai, "miR-139-5p inhibits the tumorigenesis and progression of oral squamous carcinoma cells by targeting HOXA9," *Journal of Cellular and Molecular Medicine*, vol. 21, no. 12, pp. 3730–3740, 2017.
- [12] B. Yang, W. Zhang, D. Sun et al., "Downregulation of miR-139-5p promotes prostate cancer progression through regulation of SOX5," *Biomedicine & Pharmacotherapy*, vol. 109, pp. 2128–2135, 2019.
- [13] C. Braicu, D. Gulei, R. Cojocneanu et al., "miR-181a/b therapy in lung cancer: reality or myth?," *Molecular Oncology*, vol. 13, no. 1, pp. 9–25, 2019.
- [14] R. Sheervalilou, K. Ansarin, S. Fekri Aval et al., "An update on sputum MicroRNAs in lung cancer diagnosis," *Diagnostic Cytopathology*, vol. 44, no. 5, pp. 442–449, 2016.
- [15] P. Li, Z. Xiao, J. Luo, Y. Zhang, and L. Lin, "miR-139-5p, miR-940 and miR-193a-5p inhibit the growth of hepatocellular carcinoma by targeting SPOCK1," *Journal of Cellular and Molecular Medicine*, vol. 23, no. 4, pp. 2475–2488, 2019.
- [16] J. Liu, C. Y. Li, Y. Jiang, Y. C. Wan, S. L. Zhou, and W. J. Cheng, "Tumor-suppressor role of miR-139-5p in endometrial cancer," *Cancer Cell International*, vol. 18, no. 1, p. 51, 2018.

Research Article

Comprehensive Analysis of Differently Expressed and Methylated Genes in Preeclampsia

Wenyi Xu,¹ Ping Ru,¹ Zhuorong Gu,¹ Ruoxi Zhang,¹ Xixia Pang,² Yi Huang,³ Zhou Liu ,⁴ and Ming Liu ¹

¹Department of Obstetrics and Gynecology, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200032, China

²Department of Obstetrics and Gynecology, Kongjiang Hospital, Shanghai 200093, China

³Department of Life Science, Sichuan Agricultural University, Sichuan 625014, China

⁴Department of Health Sciences Affiliated Zhoupu Hospital, Shanghai University of Medicine, Shanghai 200032, China

Correspondence should be addressed to Zhou Liu; zpyfck@126.com and Ming Liu; ming_l2016@126.com

Received 23 June 2020; Revised 18 August 2020; Accepted 6 September 2020; Published 2 November 2020

Academic Editor: Lei Chen

Copyright © 2020 Wenyi Xu 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.

Preeclampsia (PE) is one of the mainly caused maternal and infant incidences and mortalities worldwide. However, the mechanisms underlying PE remained largely unclear. The present study identified 1716 high expressions of gene and 2705 low expressions of gene using GSE60438 database, and identified 7087 hypermethylated and 15120 hypomethylated genes in preeclampsia using GSE100197. Finally, 536 upregulated genes with hypomethylation and 322 downregulated genes with hypermethylation were for the first time revealed in PE. Gene Ontology (GO) analysis revealed that these genes were associated with peptidyl-tyrosine phosphorylation, skeletal system development, leukocyte migration, transcription regulation, T cell receptor and IFN- γ -involved pathways, innate immune response, signal transduction, cell adhesion, angiogenesis, and hemopoiesis. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis demonstrated that aberrantly methylated differentially expressed genes were involved in regulating adherens junction, pluripotency of stem cell regulation, immune processing, T cell receptor and NF- κ B pathways, HTLV-I and HSV infections, leishmaniasis, and NK-induced cytotoxicity. Protein-protein interaction (PPI) network analysis identified several hub networks and key genes, including MAPK8, CCNF, CDC23, ABL1, NF1, UBE2E3, CD44, and PIK3R1. We hope these findings will draw more attention to these hub genes in future PE studies.

1. Background

As a kind of pregnancy-induced hypertension, preeclampsia (PE) is one of the mainly caused maternal and infant incidences and mortalities worldwide [1, 2]. Numerous body organs and functional systems could be affected by PE, followed by emerging renal failure, ischemic heart, type II diabetes, etc. [1–3]. Several researches have shown a part of external and internal factors that had been identified to induce PE [4]. Currently, trophoblast invasion and failure of spiral artery transformation have been considered to be one inducer of PE [5]. Even though perinatal care was improved, the ratio occurrence of PE has not been reduced

[6, 7]. Up to date, the inherent mechanism of PE taken part in many physiological disorders stayed elusive.

Many studies have identified a large number of differentially expressed genes (DEGs) and differentially methylated genes (DMGs) in PE based on advanced technologies [8–12]. Liu et al. reported that 268 dysfunctional genes were identified in PE, which were related to hormone activity and immune response. Besides, this study revealed TLR2, GSTO1, and mapk13 functioned importantly in the progression of PE [10, 11]. Presently, no studies to investigate the regulated role of gene expression implicated in PE.

Epigenetics indicated that the change of gene expression was heritable, but did not turn out to be in DNA [13, 14].

Among them, DNA methylation was the mostly generated modification in biological metabolism [15]. DNA methyltransferases (DNMTs) were responsible for transmitting DNA methylation to target sites [16]. Nevertheless, the details towards the methylation are not fully understood.

Here, we wanted to explore the association of gene expression with DNA methylation and potential signal pathway in PE development. Therefore, we evaluated the unknown interaction and related signaling pathways of DEG and DMGs in PE by gene expression microarray data (GSE60438) [12] and gene methylation microarray data (GSE100197) [17]. To this end, we attempted to uncover the potential indicator for early diagnosis and prognosis of PE, and also give a hint of probing the involved pathways of DEG/DMGs in PE.

2. Materials and Methods

2.1. Microarray Data. Differently expressed genes (DEGs)/differently methylated genes (DMGs) were individually analyzed by GSE60438 [12] (including 47 preeclampsia and 48 normal samples) and GSE100197 (including 22 preeclampsia and 51 normal samples) [17]. The details could be seen in the website <https://www.ncbi.nlm.nih.gov/geo/>.

2.2. Data Processing. GEO2R is an online tool that allows users to perform comparisons between different groups in GEO series, which depends on the GEOquery and the Linear Models for Microarray Analysis (LIMMA) R packages [18, 19]. The raw data in TXT format were checked in Venn software online to detect the commonly DEGs among the three datasets. The cutoff standards of DEGs were defined as $P < 0.05$ and fold change > 2 , while those of DMGs were indicated as $FDR < 0.05$ and a fold change > 2 .

2.3. The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Analysis. DAVID [20] was conducted to do bioinformatics analysis. Significant difference was indicated as $P < 0.01$.

2.4. Construction of Protein-Protein Interaction (PPI) Network. PPI network, including highly methylated and lowly methylated genes, was constructed by STRING database. Interaction score of 0.4 was regarded as cutoff. Cytoscape and the Molecular Complex Detection (MCODE) algorithm were separately applied to visualize PPI network and screen modules. The Molecular Complex Detection (MCODE) app was used to analyze PPI network modules [21], and MCODE scores > 3 and the number of nodes > 5 were set as cutoff criteria with the default parameters (degree cutoff ≥ 2 , node score cutoff ≥ 2 , K-core ≥ 2 , and max depth = 100). DAVID was utilized to perform pathway enrichment analysis of gene modules. Finally, cytoHubba, a Cytoscape plugin, was utilized to explore PPI network hub genes; it provides a user-friendly interface to explore important nodes in biological networks and computes using eleven methods, of which MCC has a better performance in the PPI network [22].

3. Results

3.1. Identification of Aberrantly Methylated DEGs in PE. After microarray analysis, our data have shown upregulated and downregulated 3378 DEGs which were 1663 and 1715, respectively. We identified 7087 highly methylated and 15120 lowly methylated genes in PE after relative to normal samples. 829 highly methylated genes (Figure 1(c)) with enhanced level and 408 lowly methylated genes (Figure 1(d)) with weak level were classified after overlapping DEGs and aberrantly methylated genes. Figure 1(a) shows DEGs in GSE60438 and Figure 1(b) illustrates DMGs of PE and normal tissue. The top 10 upregulated and downregulated genes in PE are shown in Tables 1 and 2.

3.2. Functional Analysis. GO analysis indicated that high methylation of genes with increasing expression was generally concentrated in peptidyl-tyrosine phosphorylation, skeletal system development, regulation of bone resorption, mitotic cell cycle, peptidyl-serine phosphorylation pathway, movement of cell or subcellular component, axonogenesis, retina layer formation, calcium ion homeostasis, and cell proliferation (Figure 2(a)).

Low methylation of genes with reduced expression was abundant in leukocyte migration, transcription regulation, T cell receptor and IFN- γ -involved pathways, innate immune response, signal transduction, cell adhesion, angiogenesis, and hemopoiesis (Figure 2(b)).

3.3. Analysis of Pathway. Upregulated genes with high methylation were dramatically enriched in adherens junction, pluripotency of stem cell regulation, proteoglycans in cancer, the ErbB and sphingolipid signaling pathways, actin cytoskeleton process, ovarian steroidogenesis, carbon metabolism, renal carcinoma, and metabolic pathways (Figure 3(a)).

Downregulated genes with hypermethylation were enriched in cell adhesion, immune processing, T cell receptor and NF- κ B pathways, HTLV-I and HSV infection, leishmaniasis, and NK-induced cytotoxicity (Figure 3(b)).

3.4. PPI Network Establishment and cytoHubba Analysis. For strong expression of genes with hypomethylation, 264 nodes and 456 edges were elected. For weak expression of genes with hypermethylation, 159 nodes and 290 edges were obtained (Figure 4). For upregulated oncogenes with hypomethylation, 380 nodes and 1170 edges are shown in Figures 4 and 5. Downregulated TSGs with hypermethylation are indicated in (Figure 5). Totally, 212 nodes and 458 edges were included in TSGs. MCODE plugin detection revealed that FLNA and PRKCB were reduced with hypermethylation, and AKT1, PRDM10, CCND1, and FASN 4 were heightened with hypomethylation.

3.5. Key Module and Gene Analysis. There is obvious difference between three modules with hypomethylation of upregulated genes and three modules with hypermethylation of downregulated genes (Figure 4). The hub network 1 of overexpressed hypomethylated genes included CCNF, RNF14, UBE2B, SH3RF1, UBE2V1, FBXO30, FBXW7, FBXO17, PJA2, UBE2M, TRIM36, HECW2, UBE2E3, SOCS1, MYLIP,

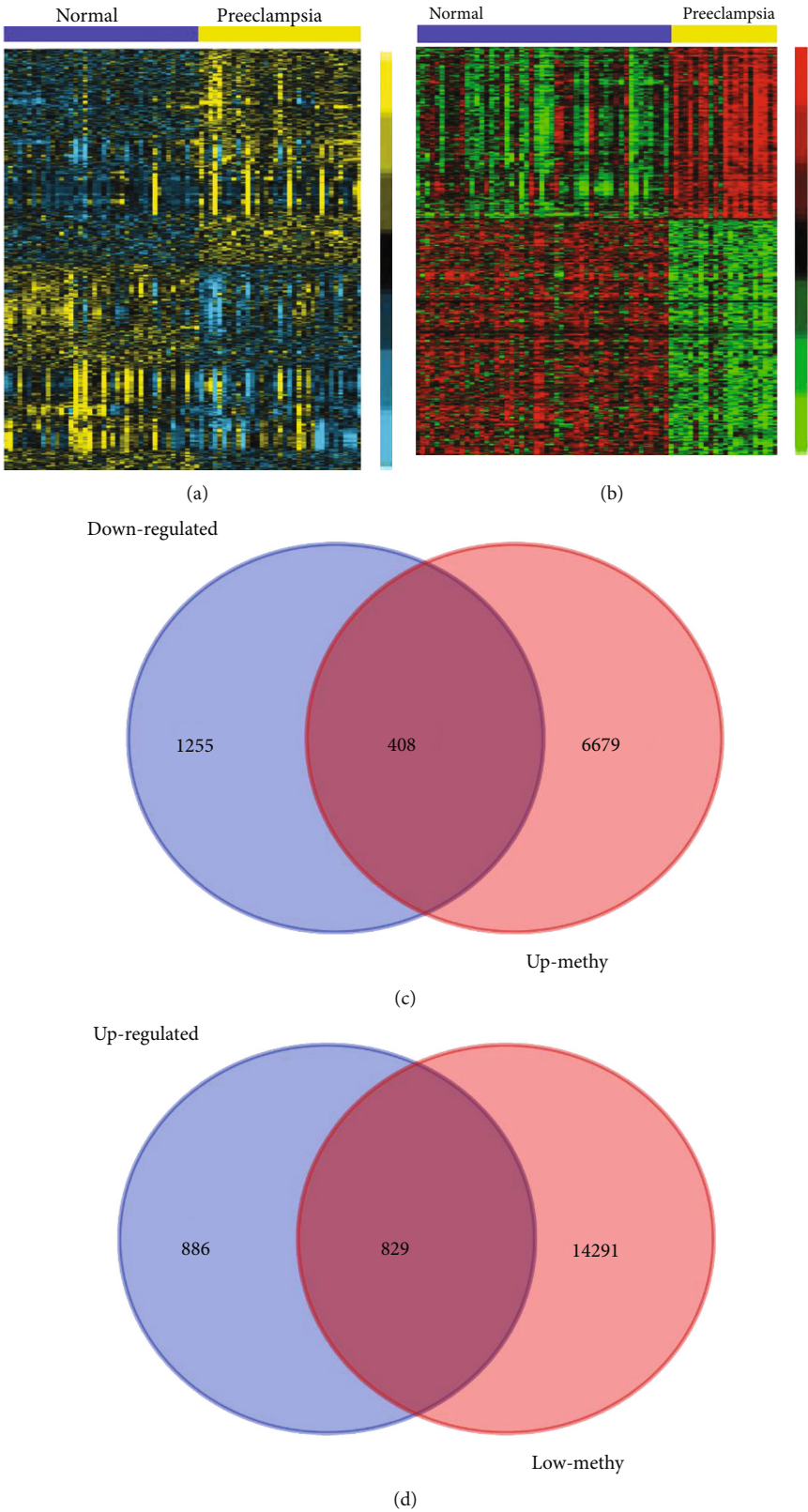


FIGURE 1: Identification of aberrantly methylated differentially expressed genes in PC. (a) Identification of differentially expressed genes in PC using GSE60438. (b) Identification of differentially methylated genes in PC using GSE100197. (c) A total of 829 upregulated hypomethylated genes were identified in PC. (d) A total of 409 downregulated hypermethylated genes were identified in PC.

TABLE 1: The top 10 upregulated genes in PE compared to normal samples.

Gene	AVE NC	AVE PE	FC	P value
CGB5	7.140694905	8.555772571	2.666740913	0.000524413
CRH	7.35786819	8.743007971	2.611972633	0.00019934
CGB1	7.330520262	8.540119829	2.312734358	0.00097466
KISS1	7.882624452	9.119308971	2.356563438	0.00381749
ADAM12	8.732927738	10.04238443	2.478481844	0.002319055
DLK1	7.393902548	8.500453743	2.153302782	0.010929138
CGA	8.495501333	9.765370743	2.41139737	0.003638636
PSG6	8.514054476	9.723809743	2.312983969	0.007442872
CGB8	7.176896333	8.174694886	1.996950473	0.001320154
PAGE4	7.430149214	8.450650829	2.028624174	0.009079707

TABLE 2: The top 10 downregulated genes in PE compared to normal samples.

Gene	AVE NC	AVE PE	FC	P value
LOC647169	8.7158875	8.070096229	0.639142146	0.012472714
FCN1	10.84868995	10.03952706	0.570712911	0.023728745
LYZ	12.23783017	11.29681926	0.520867776	0.002834118
CCL2	10.03269452	9.2354408	0.575443535	0.000263349
CX3CR1	8.787585643	8.068285229	0.607391905	0.007852246
CCL18	8.994379333	8.217637057	0.583683311	0.000247839
GSTA1	8.678950238	7.926187143	0.593465844	0.006579547
PI3	8.461040238	7.713865943	0.595769307	0.013088748
LTB	9.728131619	8.857907943	0.547062027	0.00330664
GSTA1	8.871995429	8.0231716	0.555237214	0.004279029

and CDC23. The hub network 2 of overexpressed hypomethylated genes included GPER1, OPN4, GPR17, PLCB4, MCHR2, MCHR1, TAS2R14, PTGER3, CCL4, NPS, KISS1, and ADCY8. The hub network 3 of overexpressed hypomethylated genes included SEC22B, LHB, CGA, HNRNPA3, NEIL3, TAAR6, SLC30A5, GOLIM4, BAG4, ABCB1, GOLGA5, MAN1A2, CRH, PTPN6, PREB, SEC24B, FOLR1, DEPDC1B, TPX2, SLC30A2, CEP152, FGFR1, SGOL2, LIMK1, PSG3, CDC25C, KHSRP, DHX9, SYNCRIP, PAK4, ERBB2, SDC3, SDC1, PSG6, JUP, DCTN3, RPL22L1, KRT19, NUF2, PSG11, NCAPG, QPCT, RHOTB1, RPL34, SRP19, YWHAE, MATR3, NTF3, LMAN1, PSG4, ERBB3, SPCS3, SEC11A, ARHGEF11, SLC30A1, SLC39A1, TROAP, MAN1C1, MAP2K1, RRAS2, AKT3, SLC39A8, PSG9, TRIP13, TIMP2, TRIM24, and PSG1.

The hub network 1 of downregulated hypermethylated genes included ATG7, UBA7, RNF213, ARIH2, FBXL19, FBXO44, HERC4, and ASB15. The hub network 2 of downregulated hypermethylated genes included SRSF4, RBM5, PRPF3, SF3B1, HNRNPU, CPSF2, and CSTF3. The hub network 3 of downregulated hypermethylated genes included ADCY7, ZAP70, GPR18, LY9, NPBWR1, CD4, ITGA4, CD44, FPR3, SSTR1, GABBR1, GNB4, CCR3, and SLAMF1 (Figure 5).

Among these genes, MAPK8, CCNF, CDC23, ABL1, NF1, UBE2E3, CD44, and PIK3R1 were identified as key reg-

ulators in PE by connecting with more than 20 different genes in the network.

4. Discussion

Preeclampsia was reported to be largely related to increasing incidence and death of maternal organ, dysfunction of maternal organ, or restricted growth of foetal organ [23]. However, the mechanisms related to this disease remained largely unclear. Emerging studies demonstrated that the aberrant changes in DNA methylation contributed to the abnormal expression of key genes in multiple diseases, such as preeclampsia [24]. Therefore, conclusive delineation of gene level and methylation could provide novel insights to identify novel predictive and therapeutic targets for preeclampsia. The present study identified 1716 high expressions of gene and 2705 low expressions of gene using GSE60438 database, and identified 7087 hypermethylated and 15120 hypomethylated genes in preeclampsia using GSE100197 database. Finally, 536 upregulated genes with hypomethylation and 322 downregulated genes with hypermethylation were for the first time revealed in PE.

Furthermore, bioinformatics analysis was performed to reveal the potential functions of these aberrantly methylated DEGs in preeclampsia. Meanwhile, we identified aberrantly methylated DEGs in preeclampsia that were

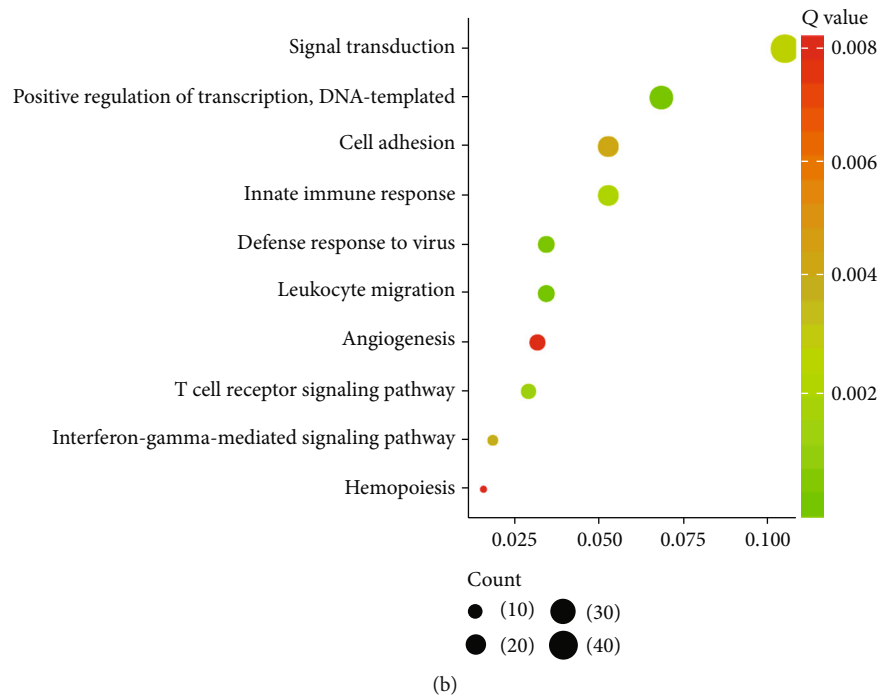
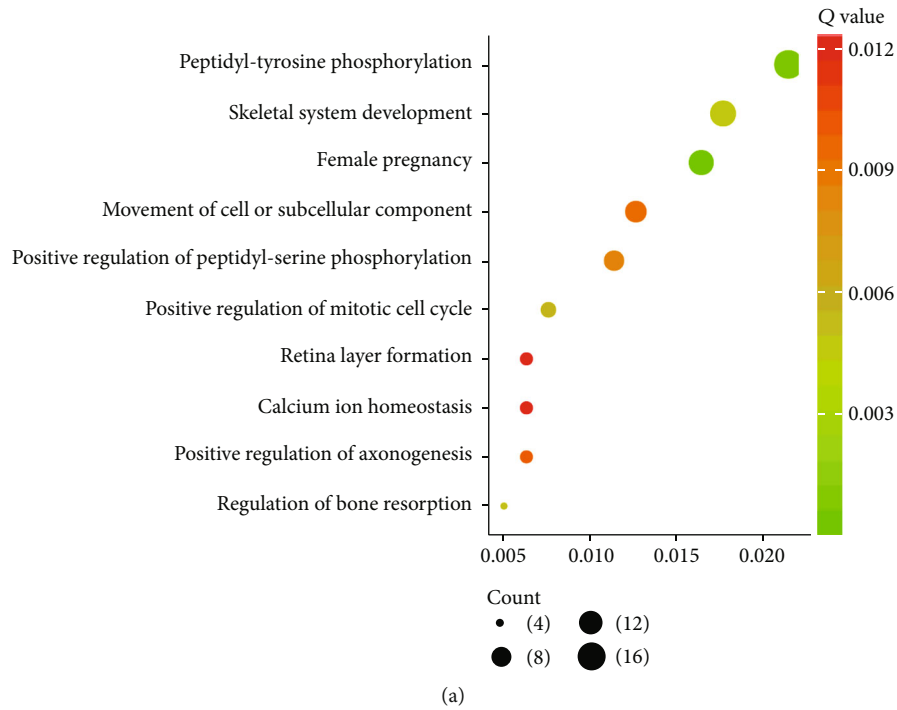


FIGURE 2: GO analysis of aberrantly methylated differentially expressed genes in PC. GO analysis of upregulated hypomethylated genes (a) and downregulated hypermethylated genes (b) in PC.

associated with transcription level, cell defense, cell immunity response, IFN- γ -involved pathway, and T cell receptor pathway. These findings were consistent with previous reports that abnormal regulation of immune functions was related to preeclampsia progression [25]. Our results showed that hypomethylated highly expressed genes were related to the regulation of multiple key signalings in cell biology, such as cell mitosis, axonogenesis, Ca²⁺ homeosta-

sis, cell proliferation, the ErbB signaling pathway, ovarian steroidogenesis, and the sphingolipid signaling pathway. As a second messenger, Ca²⁺ acts as a primary role in cell growth, cell death, etc. [26]. Downstream pathway was activated by Ca²⁺ via exporting intracellular organelles or importing extracellular depots [27–29]. As the foremost form of Ca²⁺ pathway, downstream effectors of intracellular Ca²⁺ oscillations included transcription factors, kinases, and

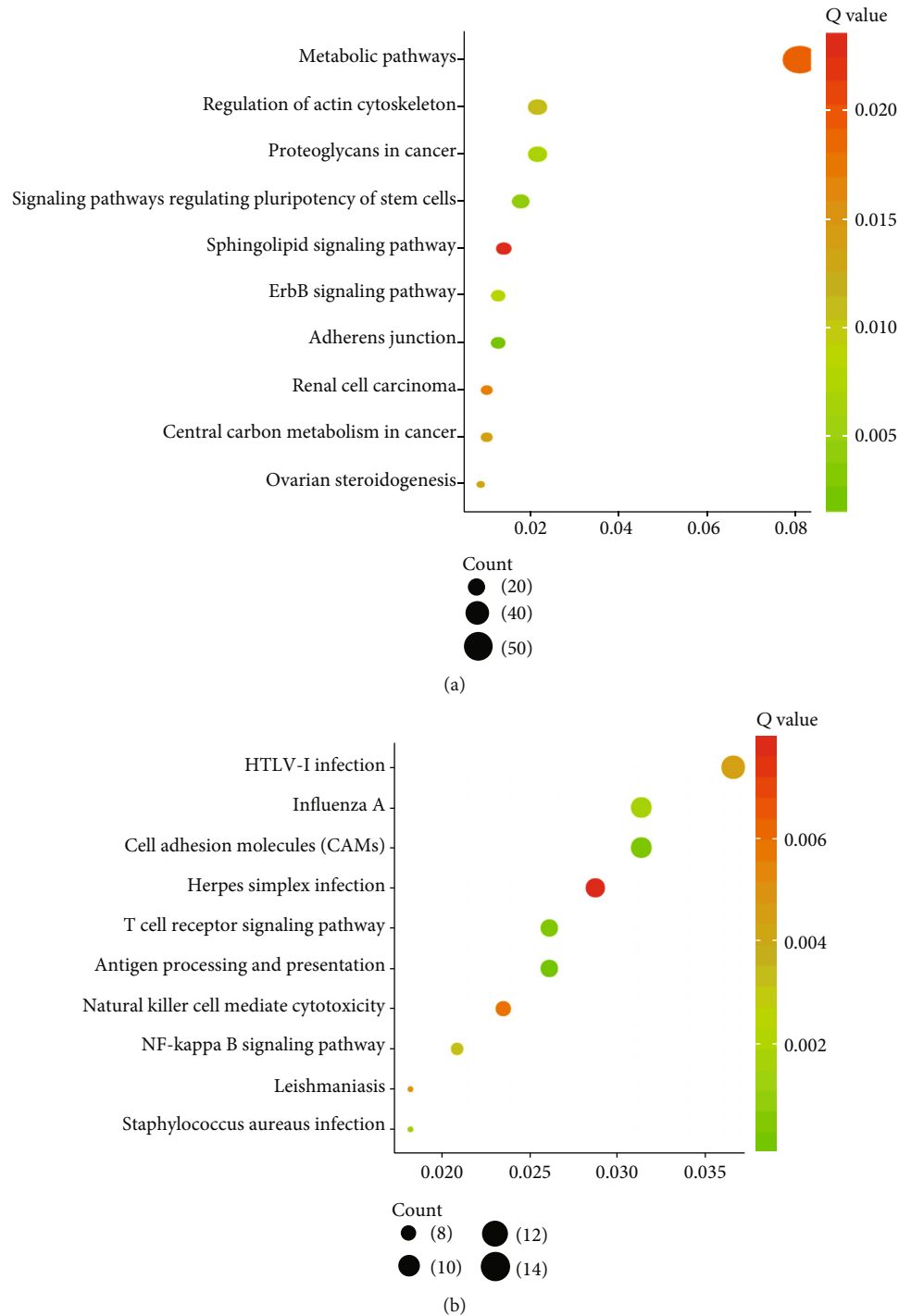


FIGURE 3: KEGG pathway analysis of aberrantly methylated differentially expressed genes in PC. KEGG pathway analysis of upregulated hypomethylated genes (a) and downregulated hypermethylated genes (b) in PC.

other functional proteins [30–32]. Our data suggests that the imbalance of Ca^{2+} in homeostatic cells may be linked to the progression of PE. A very interesting finding is that a recent study showed that Ca^{2+} signaling is related to the activation of the ErbB pathway, involving lots of tyrosine kinases, and is resistant to radiation and chemotherapy in many tumors. Two tyrosine residues were dimerized and phosphorylated by EGFR after conjugating to ligands [33,

34]. Conversely, these phosphorylated tyrosines could be regarded as binding sites for some signal transmitters which participated in biological pathways.

Moreover, we revealed that hypermethylated genes with low expression were associated with cell adhesion, angiogenesis, hemopoiesis, and the NF-kappa B signaling pathway. A recent study showed that the genes of cell adhesion signaling in the preeclamptic placentas were observed to be

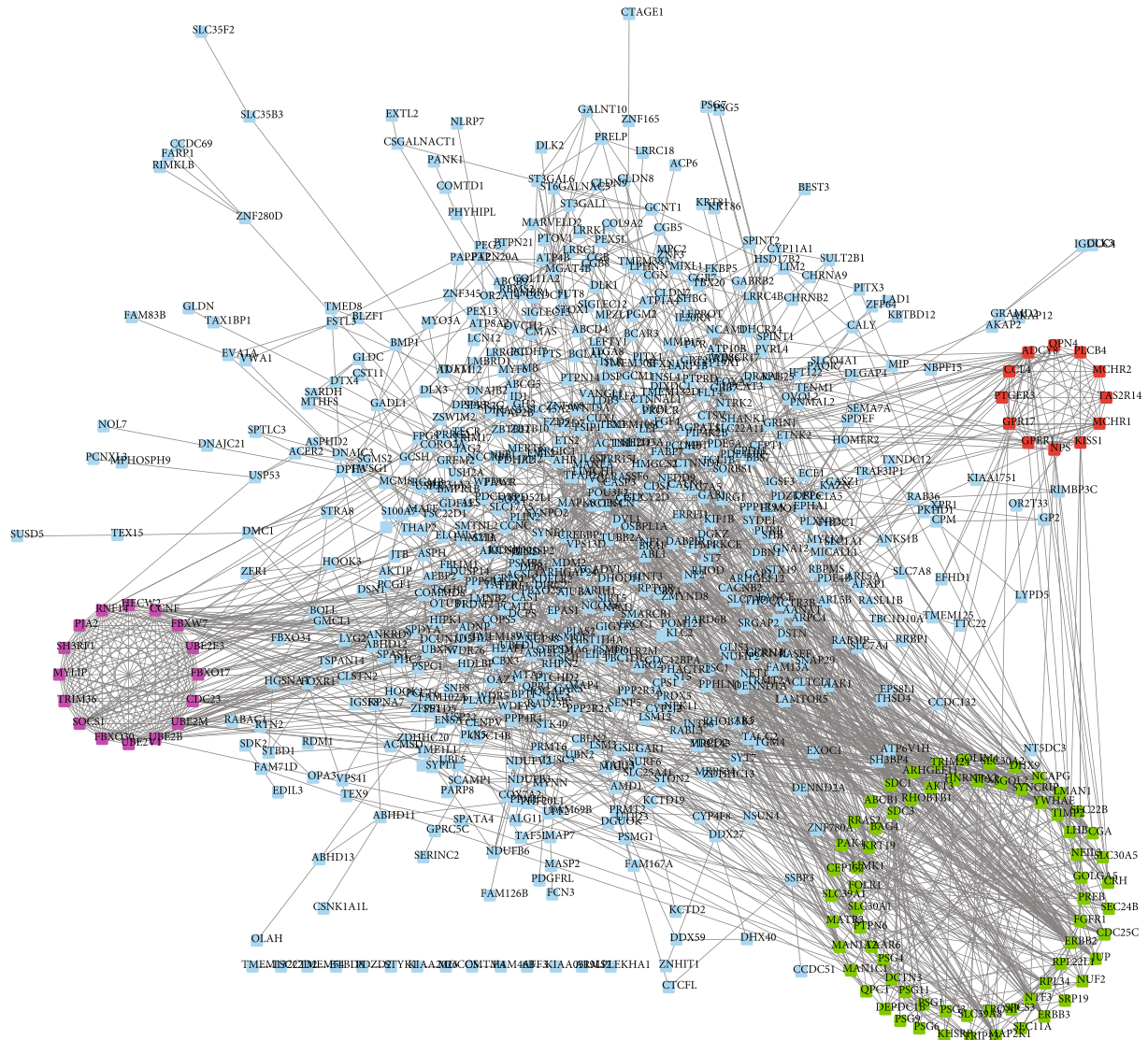


FIGURE 4: Protein-protein interaction network analysis of upregulated hypomethylated genes in PC. We constructed PPI networks of upregulated hypomethylated genes in PC.

differentially methylated [35]. Endothelial cells have been confirmed to be acted as the key inducer to angiogenesis via cell-promoting cell metastasis [36]. Notedly, EPCs (endothelial progenitor cells) functioned importantly in the generation of the postnatal blood vessel and vascular homeostasis [37]. The endothelial dysfunction in PE probably led to the destructive fetoplacental angiogenesis and neovascuogenesis [38]. The decreasing level of some proangiogenic factors in the placenta was observed in the early-stage PE not the late-stage PE [38]. There were more than 2 angiogenesis-related genes with the reduced level in the early-stage PE after comparison with the late-stage PE or control [39]. Currently, our data revealed that the growth/migration of human umbilical vein endothelial cells was suppressed in the early-stage PE compared to that in the late-stage PE or control, suggesting negative regulation of angiogenesis in PE.

In order to identify the hub genes and networks in PE, we conducted a PPI network analysis. The upregulated hypomethylated PPI network was composed of 380 nodes and

1170 edges, while the downregulated hypermethylated PPI network consisted 380 nodes and 1170 edges. Furthermore, we identified 6 hub networks using MCODE plugin in Cytoscape software. Among these genes, MAPK8, CCNF, CDC23, ABL1, NF1, UBE2E3, CD44, and PIK3R1 were identified as key regulators in PE. MAPK8 belonged to mitogen-activated protein kinase (MAPK) family which is critical for cellular function through regulating numerous signaling pathways [40]. A recent study showed that MAPK8, which is necessary for epithelial-mesenchymal transition, is responsible for regulating transcription [41]. CDC23 is a cell cycle regulator, exhibiting importantly in both initiation and elongation of DNA replication [42, 43]. Loss of NF1 results in dysregulation of MAPK, PI3K, and other signaling cascades, to promote cell proliferation and to inhibit cell apoptosis. UBE2E3 have a key role in regulation of cell aging which was essential for homeostasis of tissues. Cells' absence of UBE2E3 will be senescent even though without DNA damage [44]; meanwhile, accumulated mitochondrial and lysosomal

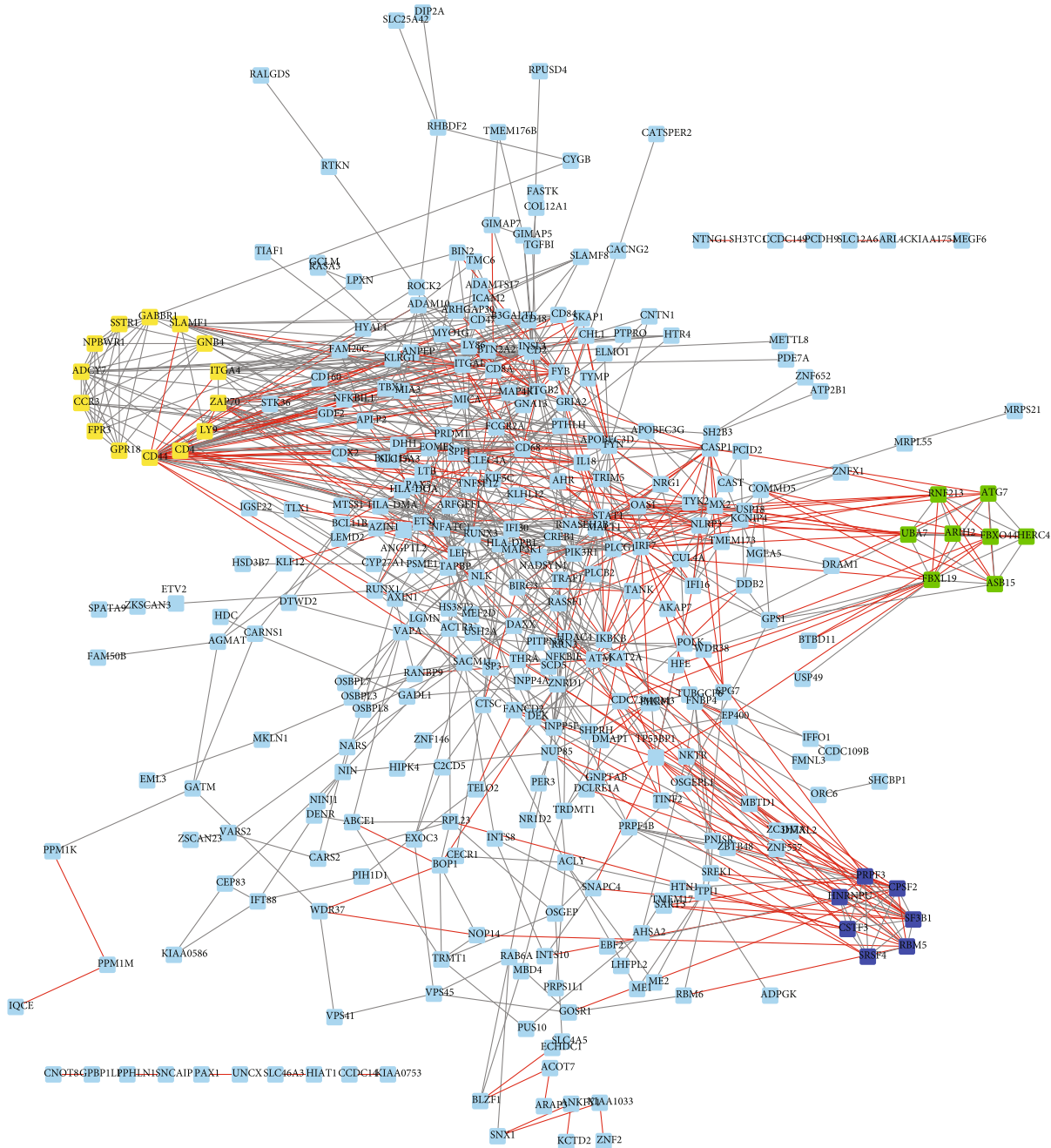


FIGURE 5: Protein-protein interaction network analysis of downregulated hypermethylated genes in PC. We constructed PPI networks of downregulated hypermethylated genes in PC.

mass and raised basal autophagic flux were shown in UBE2E3 absent cells. CD44 as a member of CAM family mostly takes part in cell movement and proliferation [45]. PIK3R1-encoded PI3K, p85 α , could conjugate, maintain, and suppress catalytic subunit of PI3K p110 [46]. Not only did mutated PIK3R1 reduce the subtype of P110 inhibition but also destroyed the new regulatory effect of p85 α on PTEN or activated a new signal pathway.

Nevertheless, our studies still had some limitations. Firstly, our researches concentrated on the classification of DEG with different methylations. Secondly, our researches should broaden the analysis datasets so as to acquire com-

prehensive data. Thirdly, we needed to conduct qRT-PCR or western blot to further ensure the selected gene level in PE samples. Finally, the function and mechanism of biomarkers in PE need to be further studied in vivo and in vitro.

5. Conclusion

Collectively, we identified some oncogene expression patterns and their links with corresponding pathways in PE, providing a hint of exploring the mechanisms implicated in PE onset and development.

Data Availability

The datasets used during the present study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no competing interests.

Acknowledgments

This work has been supported by the Key Discipline Construction of National Health Commission in Pudong New Area of Shanghai Obstetrics and Gynecology (PWZxk2017-14) and the Establishment and Popularization of Standardized System for Screening and Comprehensive Prevention of Preterm Labor (2019SY044).

References

- [1] T. A. Gudeta and T. M. Regassa, "Pregnancy induced hypertension and associated factors among women attending delivery service at Mizan-Tepi University Teaching Hospital, Tepi General Hospital and Gebretsadik Shawo Hospital, Southwest, Ethiopia," *Ethiopian journal of health sciences*, vol. 29, no. 1, pp. 831–840, 2019.
- [2] M. C. Remich and E. Q. Youngkin, "Factors associated with pregnancy-induced hypertension," *The Nurse Practitioner*, vol. 14, no. 1, pp. 20–24, 1989.
- [3] K. Watanabe, C. Kimura, A. Iwasaki et al., "Pregnancy-induced hypertension is associated with an increase in the prevalence of cardiovascular disease risk factors in Japanese women," *Menopause*, vol. 22, no. 6, pp. 656–659, 2015.
- [4] C. E. Powe, R. J. Levine, and S. A. Karumanchi, "Preeclampsia, a disease of the maternal endothelium: the role of antiangiogenic factors and implications for later cardiovascular disease," *Circulation*, vol. 123, no. 24, pp. 2856–2869, 2011.
- [5] F. Lyall, J. N. Bulmer, E. Duffie, F. Cousins, A. Theriault, and S. C. Robson, "Human trophoblast invasion and spiral artery transformation," *The American Journal of Pathology*, vol. 158, no. 5, pp. 1713–1721, 2001.
- [6] J. Mayrink, M. L. Costa, and J. G. Cecatti, "Preeclampsia in 2018: revisiting concepts, physiopathology, and prediction," *Scientific World Journal*, vol. 2018, article 6268276, pp. 1–9, 2018.
- [7] R. Townsend, P. O'Brien, and A. Khalil, "Current best practice in the management of hypertensive disorders in pregnancy," *Integrated Blood Pressure Control*, vol. 9, pp. 79–94, 2016.
- [8] K. Liu, Q. Fu, Y. Liu, and C. Wang, "An integrative bioinformatics analysis of microarray data for identifying hub genes as diagnostic biomarkers of preeclampsia," *Bioscience Reports*, vol. 39, no. 9, 2019.
- [9] S. Liu, X. Xie, H. Lei, B. Zou, and L. Xie, "Identification of key circRNAs/lncRNAs/miRNAs/mRNAs and pathways in preeclampsia using bioinformatics analysis," *Medical Science Monitor*, vol. 25, pp. 1679–1693, 2019.
- [10] E. Tejera, M. Cruz-Monteagudo, G. Burgos et al., "Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis," *BMC Medical Genomics*, vol. 10, no. 1, p. 50, 2017.
- [11] E. Tejera, J. Bernardes, and I. Rebelo, "Preeclampsia: a bioinformatics approach through protein-protein interaction networks analysis," *BMC Systems Biology*, vol. 6, no. 1, p. 97, 2012.
- [12] H. E. J. Yong, P. E. Melton, M. P. Johnson et al., "Genome-wide transcriptome directed pathway analysis of maternal pre-eclampsia susceptibility genes," *PLoS One*, vol. 10, no. 5, article e0128230, 2015.
- [13] M. Trerotola, V. Relli, P. Simeone, and S. Alberti, "Epigenetic inheritance and the missing heritability," *Human Genomics*, vol. 9, no. 1, p. 17, 2015.
- [14] M. Slatkin, "Epigenetic inheritance and the missing heritability problem," *Genetics*, vol. 182, no. 3, pp. 845–850, 2009.
- [15] W. Xu, F. Wang, Z. Yu, and F. Xin, "Epigenetics and cellular metabolism," *Genetics & Epigenetics*, vol. 8, pp. 43–51, 2016.
- [16] X. Cao and S. E. Jacobsen, "Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing," *Current Biology*, vol. 12, no. 13, pp. 1138–1144, 2002.
- [17] S. L. Wilson, K. Leavey, B. J. Cox, and W. P. Robinson, "Mining DNA methylation alterations towards a classification of placental pathologies," *Human Molecular Genetics*, vol. 27, no. 1, pp. 135–146, 2018.
- [18] S. Davis and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.
- [19] I. Diboun, L. Wernisch, C. A. Orengo, and M. Koltzenburg, "Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma," *BMC Genomics*, vol. 7, no. 1, p. 252, 2006.
- [20] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.
- [21] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.
- [22] C. H. Chin, S. H. Chen, H. H. Wu, C. W. Ho, M. T. Ko, and C. Y. Lin, "cytoHubba: identifying hub objects and sub-networks from complex interactome," *BMC Systems Biology*, vol. 8, Supplement 4, p. S11, 2014.
- [23] D. B. Nelson, L. F. Chalak, D. D. McIntire, and K. J. Leveno, "Is preeclampsia associated with fetal malformation? A review and report of original research," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 28, no. 18, pp. 2135–2140, 2014.
- [24] L. Han, Y. Liu, S. Duan, B. Perry, W. Li, and Y. He, "DNA methylation and hypertension: emerging evidence and challenges," *Briefings in Functional Genomics*, vol. 15, no. 6, pp. 460–469, 2016.
- [25] K. A. Pennington, J. M. Schlitt, D. L. Jackson, L. C. Schulz, and D. J. Schust, "Preeclampsia: multiple approaches for a multifactorial disease," *Disease Models & Mechanisms*, vol. 5, pp. 9–18, 2011.
- [26] R. R. Resende, L. M. Andrade, A. G. Oliveira, E. S. Guimaraes, S. Guatimosim, and M. F. Leite, "Nucleoplasmic calcium signaling and cell proliferation: calcium signaling in the nucleus," *Cell Communication and Signaling: CCS*, vol. 11, no. 1, p. 14, 2013.
- [27] B. W. Poovaiah and L. Du, "Calcium signaling: decoding mechanism of calcium signatures," *The New Phytologist*, vol. 217, no. 4, pp. 1394–1396, 2018.

- [28] G. Antunes, "Modelling intracellular competition for calcium: kinetic and thermodynamic control of different molecular modes of signal decoding," *Scientific Reports*, vol. 6, no. 1, p. 23730, 2016.
- [29] A. G. Oliveira, E. S. Guimaraes, L. M. Andrade, G. B. Menezes, and M. Fatima Leite, "Decoding calcium signaling across the nucleus," *Physiology (Bethesda)*, vol. 29, no. 5, pp. 361–368, 2014.
- [30] P. Maroni, P. Bendinelli, and R. Piccoletti, "Intracellular signal transduction pathways induced by leptin in C2C12 cells," *Cell Biology International*, vol. 29, no. 7, pp. 542–550, 2005.
- [31] W. E. Muller, D. Ugarkovic, V. Gamulin, B. E. Weiler, and H. C. Schroder, "Intracellular signal transduction pathways in sponges," *Electron Microscopy Reviews*, vol. 3, no. 1, pp. 97–114, 1990.
- [32] F. C. Tsai, G. H. Kuo, S. W. Chang, and P. J. Tsai, "Ca²⁺ signaling in cytoskeletal reorganization, cell migration, and cancer metastasis," *BioMed Research International*, vol. 2015, Article ID 409245, 13 pages, 2015.
- [33] E. R. Purba, E. I. Saita, and I. N. Maruyama, "Activation of the EGF receptor by ligand binding and oncogenic mutations: the "rotation model"," *Cells*, vol. 6, no. 2, p. 13, 2017.
- [34] G. Ambrosini, J. Plescia, K. C. Chu, K. A. High, and D. C. Altieri, "Activation-dependent exposure of the inter-EGF sequence Leu83-Leu88 in factor Xa mediates ligand binding to effector cell protease receptor-1," *The Journal of Biological Chemistry*, vol. 272, no. 13, pp. 8340–8345, 1997.
- [35] L. Anton, A. G. Brown, M. S. Bartolomei, and M. A. Elovitz, "Differential methylation of genes associated with cell adhesion in preeclamptic placentas," *PLoS One*, vol. 9, no. 6, article e100148, 2014.
- [36] R. Heidenreich, M. Rocken, and K. Ghoreschi, "Angiogenesis drives psoriasis pathogenesis," *International Journal of Experimental Pathology*, vol. 90, no. 3, pp. 232–248, 2009.
- [37] M. Shibuya, "Differential roles of vascular endothelial growth factor receptor-1 and receptor-2 in angiogenesis," *Journal of Biochemistry and Molecular Biology*, vol. 39, no. 5, pp. 469–478, 2006.
- [38] C. Escudero, J. M. Roberts, L. Myatt, and I. Feoktistov, "Impaired adenosine-mediated angiogenesis in preeclampsia: potential implications for fetal programming," *Frontiers in Pharmacology*, vol. 5, p. 134, 2014.
- [39] K. Junus, M. Centlow, A. K. Wikstrom, I. Larsson, S. R. Hansson, and M. Olovsson, "Gene expression profiling of placentae from women with early- and late-onset pre-eclampsia: down-regulation of the angiogenesis-related genes ACVRL1 and EGFL7 in early-onset disease," *Molecular Human Reproduction*, vol. 18, no. 3, pp. 146–155, 2012.
- [40] M. Cargnello and P. P. Roux, "Activation and function of the MAPKs and their substrates, the MAPK-activated protein kinases," *Microbiology and Molecular Biology Reviews*, vol. 75, no. 1, pp. 50–83, 2011.
- [41] N. Tiwari, N. Meyer-Schaller, P. Arnold et al., "Klf4 is a transcriptional regulator of genes critical for EMT, including Jnk1 (Mapk8)," *PLoS One*, vol. 8, no. 2, article e57329, 2013.
- [42] B. Singh, K. K. Bisht, U. Upadhyay et al., "Role of Cdc23/Mcm10 in generating the ribonucleotide imprint at the mat1 locus in fission yeast," *Nucleic Acids Research*, vol. 47, no. 7, pp. 3422–3433, 2019.
- [43] S. Prinz, E. S. Hwang, R. Visintin, and A. Amon, "The regulation of Cdc20 proteolysis reveals a role for APC components Cdc23 and Cdc27 during S phase and early mitosis," *Current Biology*, vol. 8, no. 13, pp. 750–760, 1998.
- [44] K. S. Plafker, K. Zyla, W. Berry, and S. M. Plafker, "Loss of the ubiquitin conjugating enzyme UBE2E3 induces cellular senescence," *Redox Biology*, vol. 17, pp. 411–422, 2018.
- [45] A. Ouhtit, B. Rizeq, H. A. Saleh, M. D. M. Rahman, and H. Zayed, "Novel CD44-downstream signaling pathways mediating breast tumor invasion," *International Journal of Biological Sciences*, vol. 14, no. 13, pp. 1782–1790, 2018.
- [46] L. M. Thorpe, J. M. Spangle, C. E. Ohlson et al., "PI3K-p110 α mediates the oncogenic activity induced by loss of the novel tumor suppressor PI3K-p85 α ," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 27, pp. 7095–7100, 2017.

Research Article

Gene Expression Profiling of Type 2 Diabetes Mellitus by Bioinformatics Analysis

Huijing Zhu,^{1,2} Xin Zhu,² Yuhong Liu,² Fusong Jiang ,³ Miao Chen,^{1,4} Lin Cheng,² and Xingbo Cheng ¹

¹Department of Endocrinology and Metabolism, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China

²Department of Endocrinology and Metabolism, Heze Municipal Hospital, Heze, Shandong, China

³Department of Endocrinology and Metabolism, The Affiliated Sixth People's Hospital of Shanghai Jiao Tong University, Shanghai, China

⁴Department of ICU, Heze Municipal Hospital, Heze, Shandong, China

Correspondence should be addressed to Xingbo Cheng; fancycarp158@163.com

Received 18 June 2020; Revised 22 July 2020; Accepted 3 August 2020; Published 23 October 2020

Guest Editor: Lei Chen

Copyright © 2020 Huijing Zhu 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.

Objective. The aim of this study was to identify the candidate genes in type 2 diabetes mellitus (T2DM) and explore their potential mechanisms. **Methods.** The gene expression profile GSE26168 was downloaded from the Gene Expression Omnibus (GEO) database. The online tool GEO2R was used to obtain differentially expressed genes (DEGs). Gene Ontology (GO) term enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed by using Metascape for annotation, visualization, and comprehensive discovery. The protein-protein interaction (PPI) network of DEGs was constructed by using Cytoscape software to find the candidate genes and key pathways. **Results.** A total of 981 DEGs were found in T2DM, including 301 upregulated genes and 680 downregulated genes. GO analyses from Metascape revealed that DEGs were significantly enriched in cell differentiation, cell adhesion, intracellular signal transduction, and regulation of protein kinase activity. KEGG pathway analysis revealed that DEGs were mainly enriched in the cAMP signaling pathway, Rap1 signaling pathway, regulation of lipolysis in adipocytes, PI3K-Akt signaling pathway, MAPK signaling pathway, and so on. On the basis of the PPI network of the DEGs, the following 6 candidate genes were identified: PIK3R1, RAC1, GNG3, GNAI1, CDC42, and ITGB1. **Conclusion.** Our data provide a comprehensive bioinformatics analysis of genes, functions, and pathways, which may be related to the pathogenesis of T2DM.

1. Introduction

Type 2 diabetes mellitus (T2DM), a disease with significant morbidity, disability, and mortality, has affected increasing numbers of people worldwide. The World Health Organization (WHO) projected that diabetes would be the 7th leading cause of death in 2030. In addition, it has been predicted that by 2030, developing countries would account for 77.6% of all diabetic patients [1]. Although diabetes is a chronic disease that often causes various complications, in terms of financial burden, the cost of diabetes is 2-4 times more than that of the average patient in all medical systems [2]. Early detection and diagnosis of diabetes to prevent diabetes-associated compli-

cations and to reduce the economic costs on medical care are therefore of significant importance.

T2DM, which is characterized by hyperglycemia in the case of insulin resistance and impaired insulin secretion, is also a multigene heterogeneous disease that is the result of the interaction of genetic and environmental factors [3]. Although genetic factors play an important role in the occurrence and development of T2DM, the elaboration of its exact mechanism depends on the identification of susceptibility genes for T2DM.

At present, most of the gene research on T2DM mainly uses gene chip technology to detect and analyze model animals or clinical patient samples alone. Through this single

analysis method, some valuable genes can be screened out for research and analysis. Gene expression analysis based on microarray technology is a powerful and high-throughput research method. Through gene expression profiling, some studies have found that hundreds of differentially expressed genes (DEGs) are involved in multiple molecular functions, biological processes, and signaling pathways [4], which played an important role in the occurrence and development of diseases and could be used as a potential molecular target and diagnostic marker. In the current study, the GSE26168 dataset [5] was downloaded from the Gene Expression Omnibus (GEO) database to identify T2DM-associated DEGs between T2DM and normal samples. Subsequently, GO term enrichment analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, and PPI network analysis were performed to discover candidate genes as T2DM biomarkers and therapeutic targets worthy of further progress.

2. Methods

2.1. Microarray Data. The dataset GSE26168 based on the GPL6883 platform (Illumina HumanRef-8 v3.0 expression bead chip) was downloaded from GEO. A total of 9 T2DM samples and 8 normal samples were analyzed.

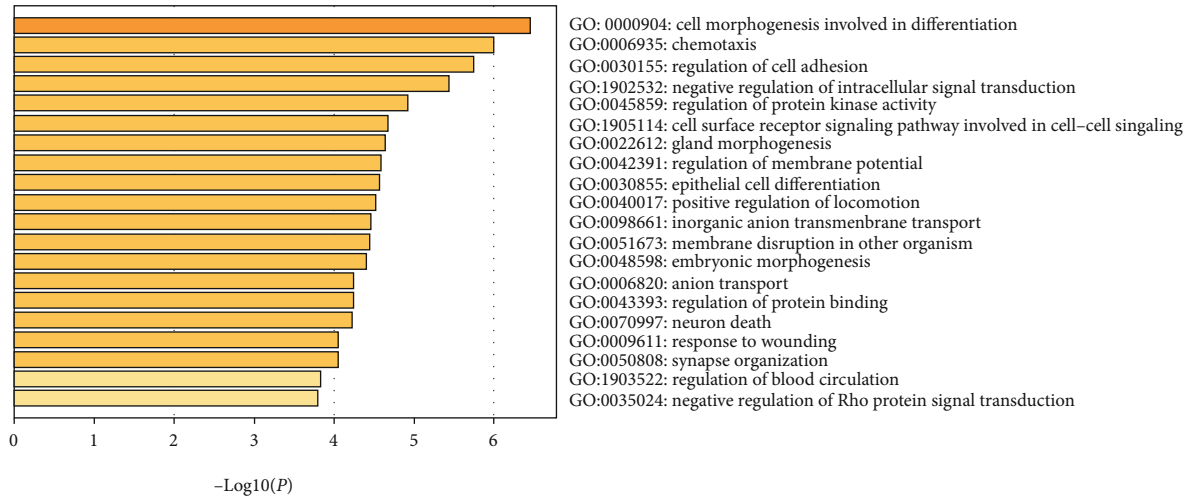
2.2. Identification of DEGs. GEO2R (<http://www.ncbi.nlm.nih.gov/geo/geo2r/>) is an interactive web tool for comparing two sets of data under the same experimental conditions and can analyze any geo series [6]. GEO2R was applied to explore DEGs between T2DM and normal blood samples. Statistically significant DEGs were defined with $|\log_{2}FC| \geq 2$, and the P value < 0.05 was the cut-off criterion.

2.3. Functional and Pathway Enrichment Analysis of DEGs. GO is a common way to annotate genes, gene products, and sequences as potential biological phenomena, mainly including biological process (BP), cellular component (CC), and molecular function (MF); the Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive database resource for the biological interpretation of genomic sequences and other high-throughput data. GO and KEGG analyses were performed using the Metascape database to analyze the DEGs at the functional level. A P value < 0.01 and min overlap > 3 were set as the cut-off criterion.

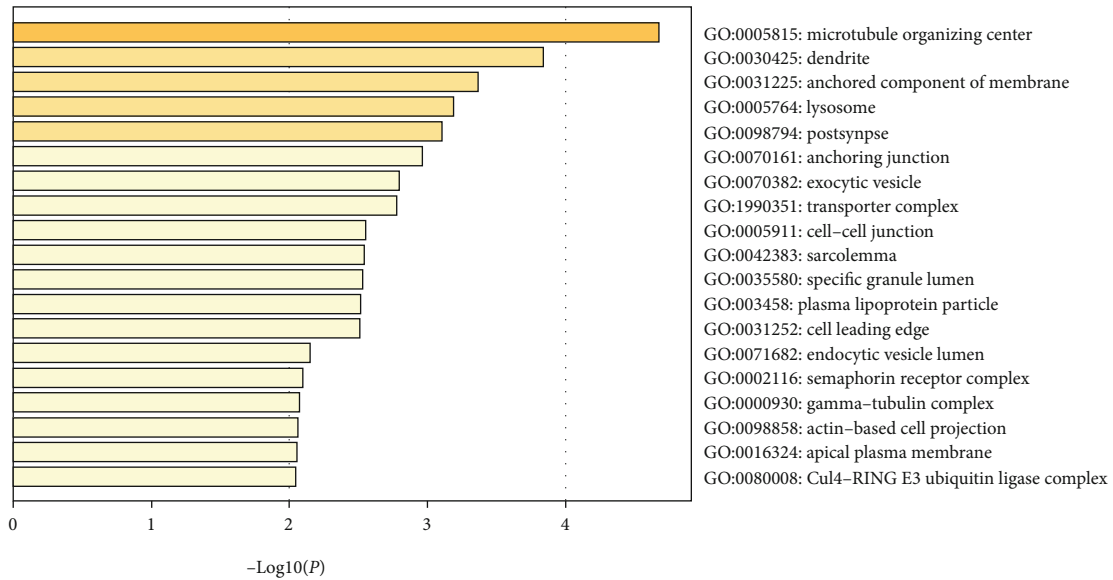
2.4. Integration of the Protein-Protein Interaction (PPI) Network. The PPI network of DEGs was constructed by STRING. A confidence score ≥ 0.9 was set as significant. Subsequently, the PPI networks were visualized using Cytoscape software (3.7.1). MCODE was used to screen out the core genes that constitute the stable structure of the PPI network with degree cut-off = 3, haircut on, node score cut-off = 0.2, k -core = 4, and maximum depth = 100. Moreover, the CentiScape plug-in was used to calculate the centrality index and topological properties for the identification of the most important nodes of a network, including undirected, directed, and weighted networks. The key (hub) genes were defined with degree value $\geq \text{mean} + 2SD$, while the bottleneck genes were defined with betweenness value $\geq \text{mean} + 2SD$. Then, using Venn diagram analysis, the genes in the



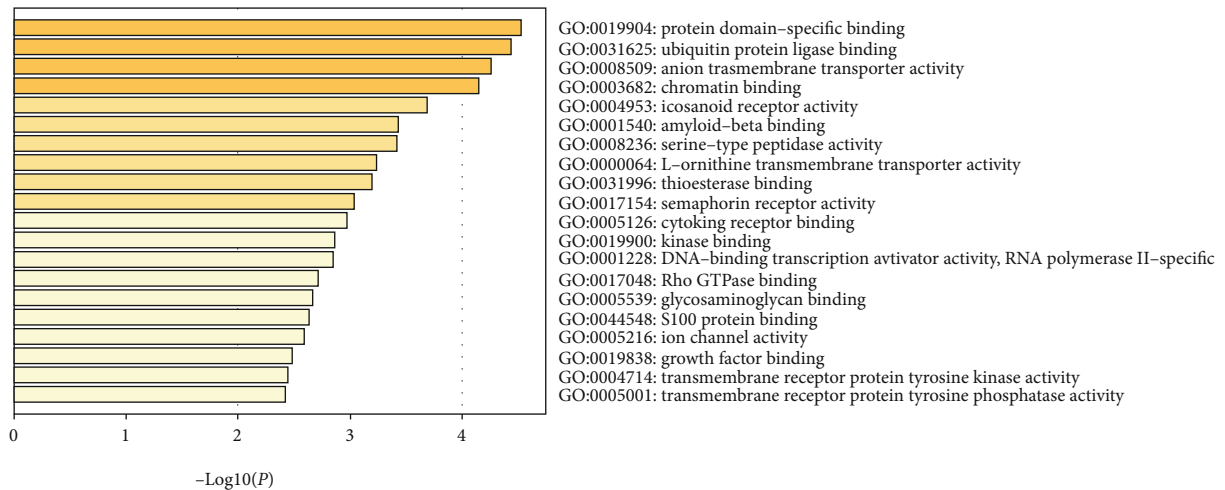
FIGURE 1: Cluster analysis of DEGs. The abscissa represents different samples; the vertical axis represents clusters of DEGs. Red indicates that expression of the gene is relatively upregulated while green indicates that expression of the gene is relatively downregulated; black indicates no significant changes in gene expression.



(a) GO BP analysis



(b) GO CC analysis



(c) GO MF analysis

FIGURE 2: Enriched GO functions of DEGs. DEGs: differentially expressed genes; GO: Gene Ontology; BP: biological process; CC: cellular component; MF: molecular function.

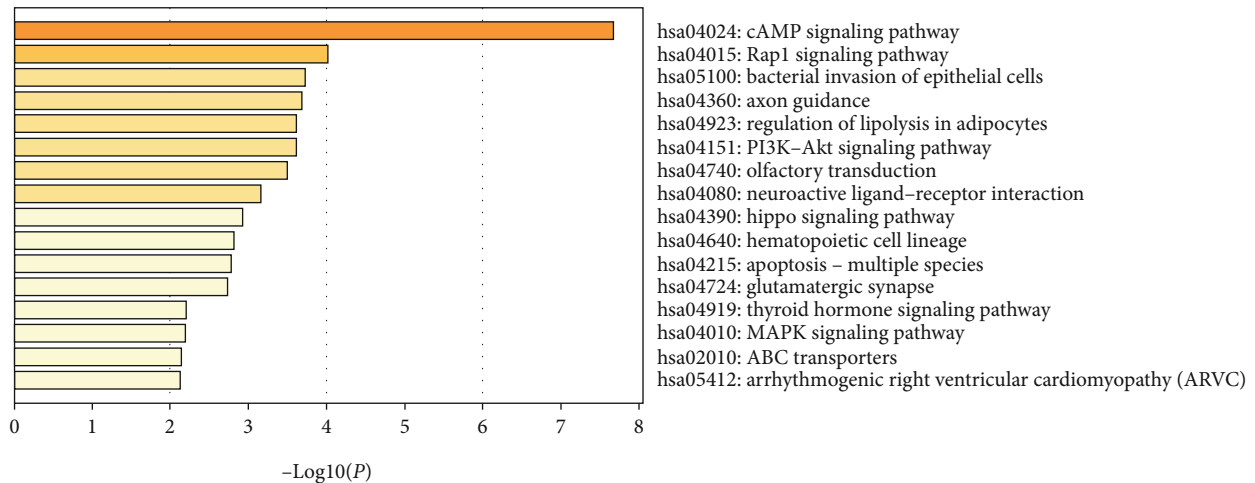


FIGURE 3: KEGG pathway analysis of differentially expressed genes. KEGG: Kyoto Encyclopedia of Genes and Genomes.

intersection of the above three datasets were selected as candidate genes for the diagnosis of T2DM.

3. Results

3.1. Identification of DEGs. Based on the aforementioned threshold ($|\log_{2}FC| \geq 2$ and $P < 0.05$), a total of 981 DEGs including 301 upregulated DEGs and 680 downregulated DEGs were filtered with GEO2R (Figure 1).

3.2. Functional and Pathway Enrichment Analysis. Three GO category results are presented in Figures 2(a)–2(c). As to the biological process (BP), DEGs were significantly enriched in cell morphogenesis involved in differentiation, chemotaxis, and regulation of cell adhesion (Figure 2(a)). For the cell component (CC), DEGs were enriched in the microtubule organizing center, dendrite, and anchored component of the membrane (Figure 2(b)). In terms of the molecular function (MF), DEGs were enriched in protein domain-specific binding, ubiquitin protein ligase binding, and anion transmembrane transporter activity (Figure 2(c)). The KEGG pathway analysis revealed that DEGs were highly associated with the cAMP signaling pathway, Rap1 signaling pathway, and bacterial invasion of epithelial cells (Figure 3).

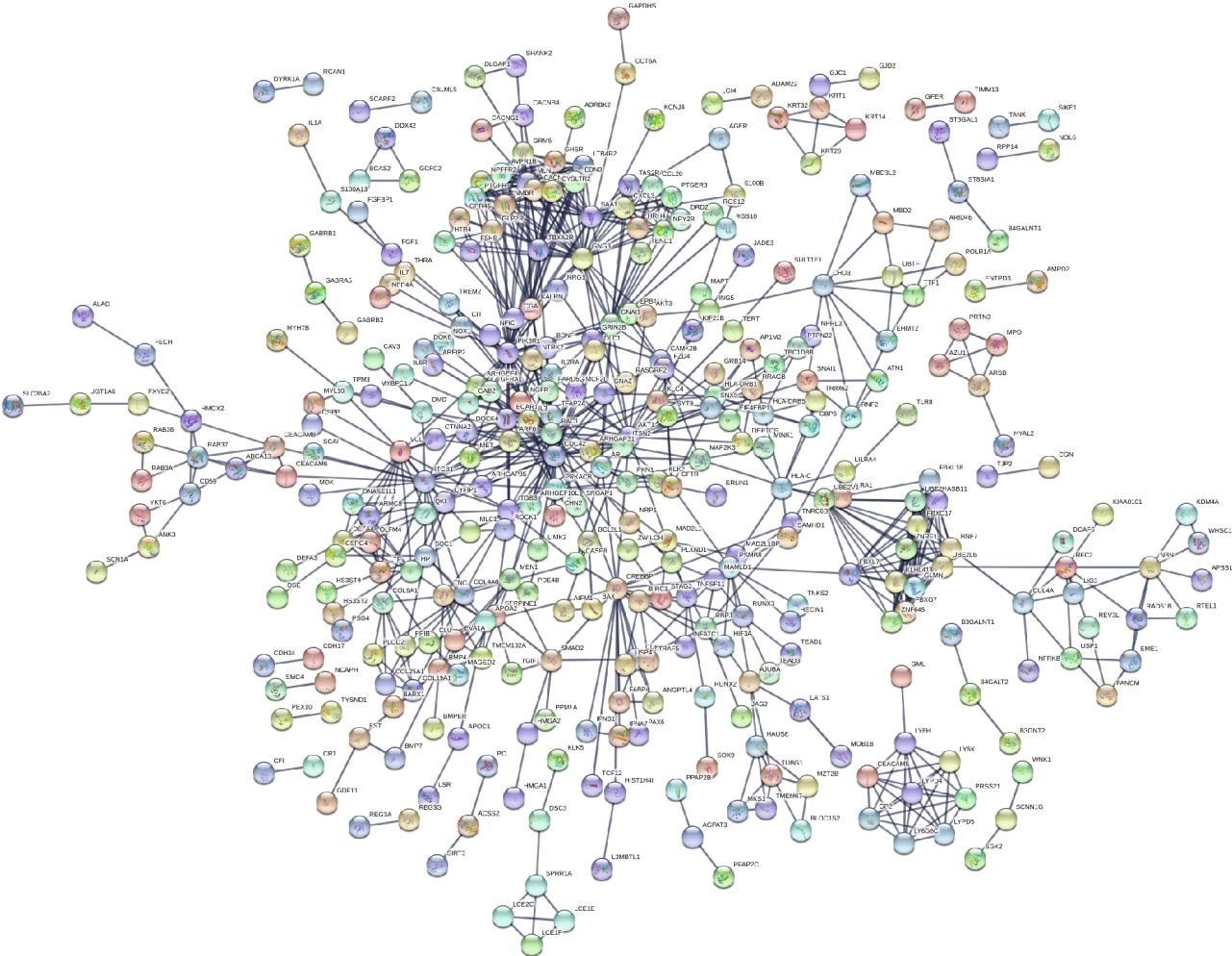
3.3. PPI Network Construction and Analysis of Modules. The PPI network consisted of 945 nodes and 835 edges after hiding nodes which could not interact with other nodes (Figure 4(a)). Then, we used MCODE to perform K kernel analysis of the string network, a total of 9 clusters were generated, and 90 core genes were screened out (Table 1, Figures 4(b)–4(j)). Besides, the topology characteristics of the string network and each node were computed with CentiScape. 12 hub genes and 14 bottleneck genes were obtained (Table 1). Moreover, using Venn diagram analysis, 6 candidate genes in the intersection of the above three datasets were selected for further analysis, including phosphoinositide-3-kinase regulatory subunit 1 (PIK3R1), ras-related C3 botulinum toxin substrate 1 (rho family, small GTP-binding protein Rac1) (RAC1), G protein subunit gamma 3 (GNG3), G

protein subunit alpha i1 (GNAI1), cell division cycle 42 (CDC42), and integrin subunit beta 1 (ITGB1) (Figure 5).

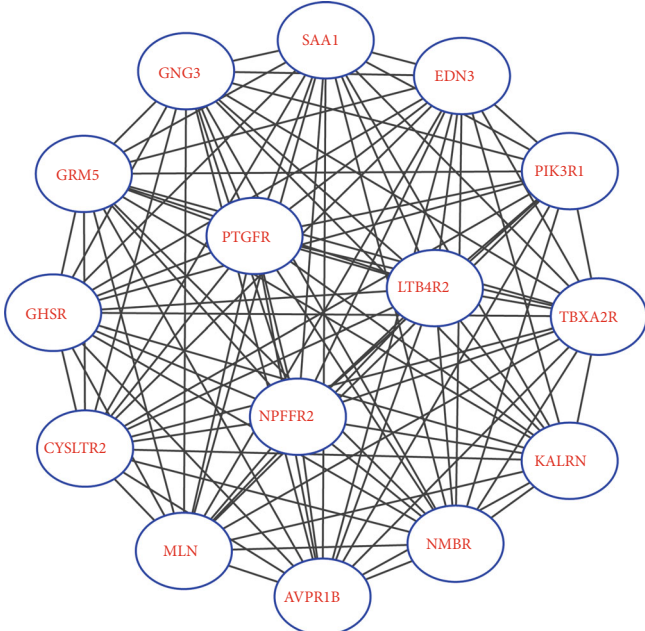
4. Discussion

In this study, we identified a total of 981 significant DEGs between T2DM and normal samples, including 301 upregulated genes and 680 downregulated genes, and conducted a series of bioinformatics analysis to screen candidate genes and pathways related to T2DM. DEGs were investigated in both GO term enrichment analysis and KEGG pathway analysis for functional annotation. As the outcomes of GO term enrichment analysis, DEGs might play critical roles in T2DM through cell differentiation, cell adhesion, intracellular signal transduction, and regulation of protein kinase activity. Meanwhile, KEGG pathway analysis revealed that DEGs were mainly enriched in the cAMP signaling pathway, Rap1 signaling pathway, regulation of lipolysis in adipocytes, PI3K–Akt signaling pathway, and MAPK signaling pathway. Moreover, by constructing the PPI, 6 candidate genes were identified, which exerted a momentous effect on the T2DM initiation, progression, and intervention strategy from different sides, including PIK3R1, RAC1, GNG3, GNAI1, CDC42, and ITGB1. The regulatory network consisting of microRNAs (miRNAs), long noncoding RNA (lncRNA), and mRNAs has attracted increasing attention to elucidate the mechanism of action in various diseases. In this study, mirDIP and starBase were used to analyze and predict the upstream miRNA interacting with candidate genes and the upstream lncRNA interacting with miRNA. A total of 22 miRNAs and 5 lncRNA were screened, which may play crucial parts in the development of T2DM.

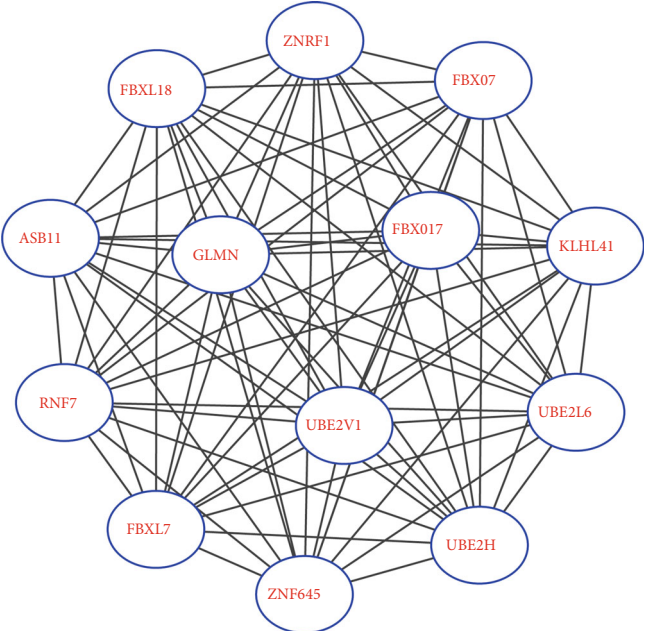
PIK3R1 encodes the p85 α regulatory subunit of the phosphatidylinositol-3-kinase (PI3K), which connects firmly with the p110 catalytic subunit, and together, they form the PI3K protein. PI3K plays a key role in insulin signaling by binding to phosphorylated insulin receptor substrates (IRS), producing phosphatidylinositol-4,5-trisphosphate (PIP3), which then activates several downstream targets such as AKT serine-threonine kinase [7]. AKT regulates cell survival,



(a) DEG PPI network

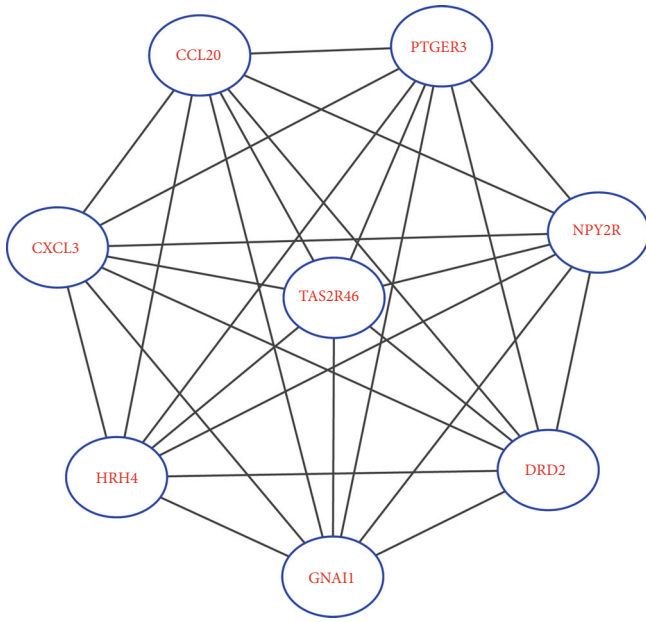


(b) Module 1

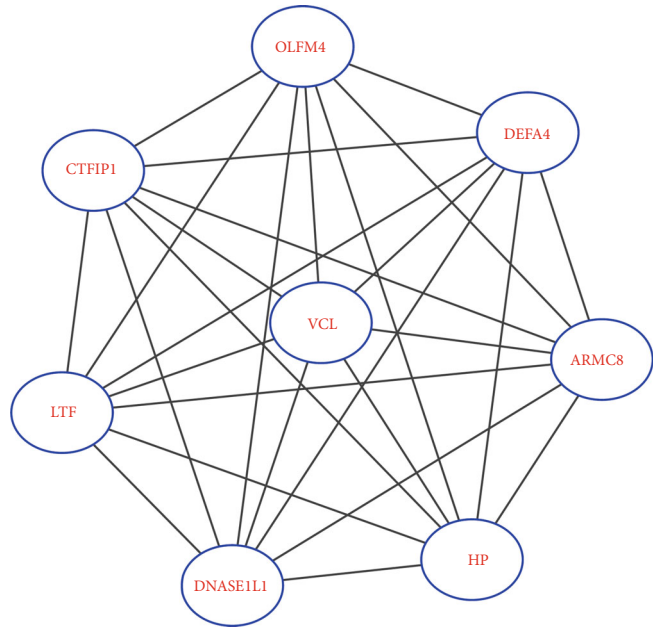


(c) Module 2

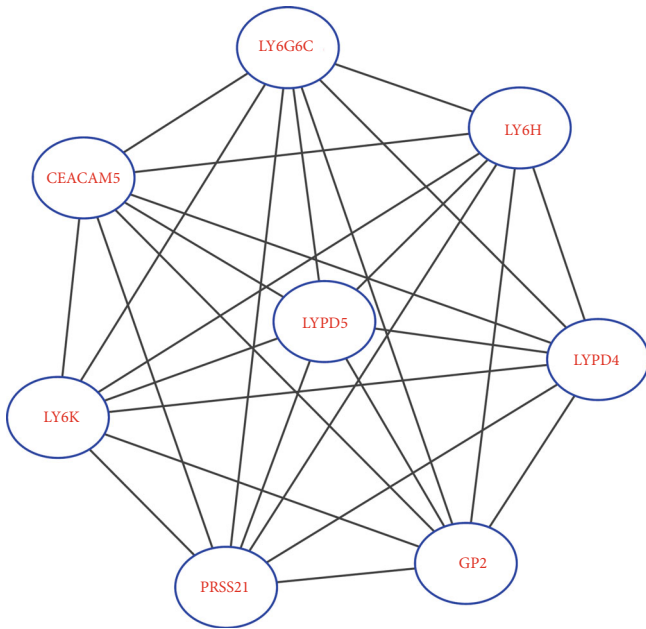
FIGURE 4: Continued.



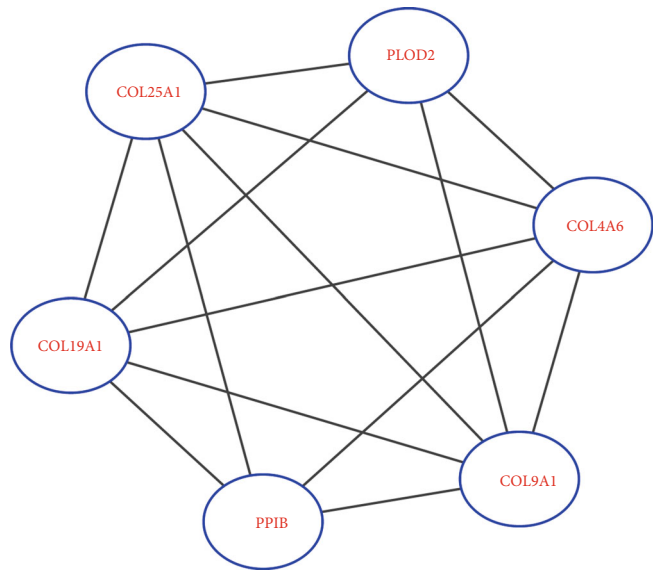
(d) Module 3



(e) Module 4



(f) Module 5



(g) Module 6

FIGURE 4: Continued.

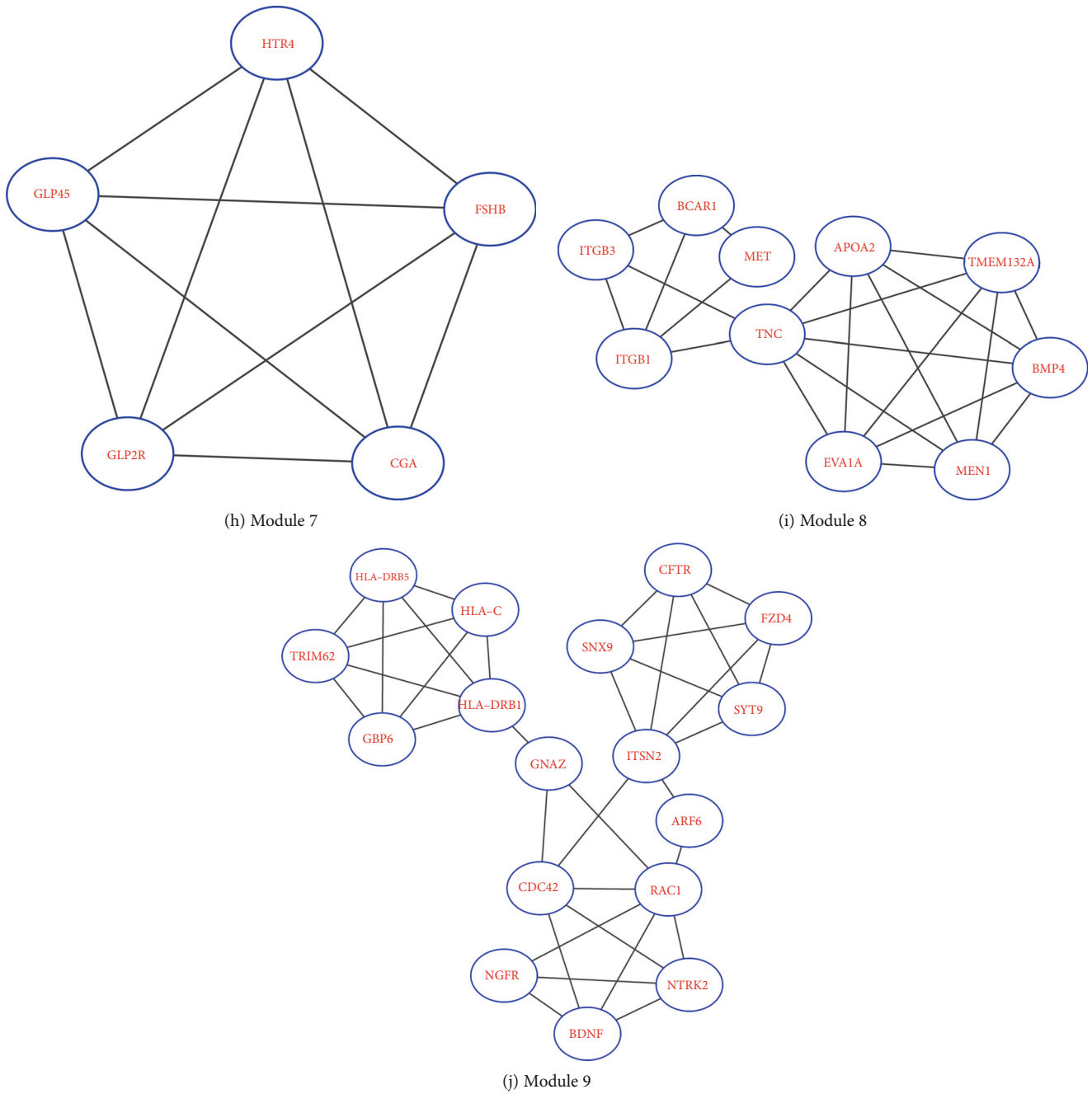


FIGURE 4: Protein-protein interaction (PPI) networks constructed by STRING and modular analysis.

growth, differentiation, glucose transporter type 4 (GLUT-4) trafficking, and glucose utilization [8]. Mouse studies have shown that mice lacking PIK3R1 display enhanced insulin sensitivity and glucose tolerance, due to an improved stoichiometry of the p85 α /p110 complex for binding to IRS and enhanced insulin-stimulated Akt activity [9, 10]. The overexpression of p85 α weakens signal transmission and causes insulin resistance by disrupting the activity of the p85 α /p110 complex and the connection between PI3K and IRS [11, 12], which indicates that p85 α subunits play a negative role in PI3K signaling downstream of the insulin receptor. Thus, PIK3R1 is a logical candidate gene involved in

the development of T2DM. Regulation of p85 α expression in insulin-sensitive tissues may be a new strategy to increase insulin sensitivity and may also become a new target for the treatment of T2DM.

CDC42 and RAC1 are members of the Rho GTPase family, which regulate signaling pathways that control a variety of cellular functions, including cell morphology, migration, endocytosis, and cell cycle progression. Both CDC42 and RAC1 regulate the second phase of glucose-stimulated insulin secretion (GSIS), and the circulation of these proteins between the activated state (GTP-bound) and the inactive state (GDP-bound) is important for insulin secretion [13, 14].

TABLE 1: The candidate genes selected from the protein-protein interaction network.

Cluster	Core genes		Gene IDs
	Nodes	Edges	
1	15	104	AVPR1B CYSLTR2 EDN3 GHSR GNG3 GRM5 KALRN LTB4R2 MLN NMBR NPFFR2 PIK3R1 PTGFR SAA1 TBXA2R
2	13	78	ASB11 FBXL18 FBXL7 FBXO17 FBXO7 GLMN KLHL41 RNF7 UBE2H UBE2L6 UBE2V1 ZNF645 ZNRF1
3	8	28	CCL20 CXCL3 DRD2 GNAI1 HRH4 NPY2R PTGER3 TAS2R46
4	8	28	ARMC8 CYFIP1 DEFA4 DNASE1L1 HP LTF OLFM4 VCL
5	8	28	CEACAM5 GP2 LY6G6C LY6H LY6K LYPD4 LYPD5 PRSS21
6	6	14	COL19A1 COL25A1 COL4A6 COL9A1 PLOD2 PPIB
7	5	10	CGA FSHB GLP2R GPR45 HTR4
8	10	22	APOA2 BCAR1 BMP4 EVA1A ITGB1 ITGB3 MEN1 MET TMEM132A TNC
9	17	35	ARF6 BDNF CDC42 CFTR FZD4 GBP6 GNAZ HLA-C HLA-DRB1 HLA-DRB5 ITS2 NGFR NTRK2 RAC1 SNX9 SYT9 TRIM62
Hub genes			AKT1 CDC42 CREBBP GNAI1 GNG3 GRM5 ITGB1 KALRN PIK3R1 RAC1 SAA1 TBXA2R
Bottleneck genes			AKT1 CDC42 CEACAM8 CREBBP GNAI1 GNG3 ITGB1 MAP2K3 PIK3R1 PSMB8 RAC1 RFC2 UBE2L6 UBE2V1

Candidate genes are shown in bold.

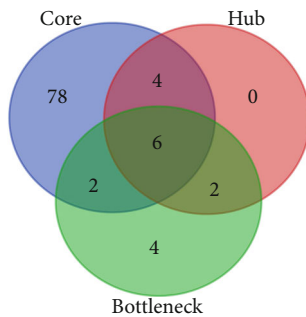


FIGURE 5: Venn diagram analysis of candidate genes. The blue circle represents core genes, the red circle represents hub genes, and the green circle represents bottleneck genes.

CDC42 plays critical roles in the process of insulin synthesis by regulating granule fusion and cytoskeletal rearrangement [15, 16] and also regulates mobilization and cell membrane exocytosis and endocytosis of insulin granules via activating a series of downstream factors [17–20]. One study showed that upregulation of miR-330-3p reduces the expression of CDC42 and E2F1 in patients with gestational diabetes (GDM), resulting in impaired β -cell proliferation [21]. P21-activated kinase 1 (PAK1) is a downstream factor of CDC42 and an important promoter of cell proliferation. Another study showed an 80% reduction in PAK1 in patients with T2DM [2]. Thus, CDC42 is an important member in the progress of T2DM, and targeted therapy for CDC42 may be one of the effective methods for treating T2DM and related diseases.

RAC1, which can stimulate actin cytoskeleton reorganization [22], is required for insulin-stimulated translocation of glucose transporter 4 (GLUT-4) in muscle cells [23]. RAC1 also can activate PAK [24]. The study indicated that RAC1 and its downstream target protein PAK were reduced in insulin-resistant mice and human skeletal muscle [25]. In addition, RAC1 can activate NADPH oxygenase (NOX),

which produces reactive oxygen species (ROS) and activates p38MAPK under high glucose conditions, leading to mitochondrial disorders and islet β -cell apoptosis [26, 27]. This mechanism also plays a crucial role in diabetes-induced vascular injuries, such as diabetic retinopathy [28] and diabetic cardiomyopathy [29]. Thus, RAC1 can be a novel molecular candidate of T2DM and provide new insight to improve therapeutic strategies for T2DM and diabetic complications.

GNG3, a member of signal-transducing molecules, a signal transduction molecule encoding the G protein gamma 3 subunit, plays a variety of roles during signal transduction, from membrane targeting of the α subunit [30] to receptor recognition [31], to activation of effectors [32], and then to effect signaling regulation of various proteins of intensity or duration [33]. The research found that inhibition of G-protein $\beta\gamma$ signaling produces the changes in the cytokine mRNA levels, which can benefit the autoimmune diseases [34]. In addition, the mice lacking the G protein $\gamma 3$ subtype show decreased weight gain, reduced fat intake, and defective Oprm1 signaling [35], when maintained on a high-fat diet. These results suggest that GNG3 may be involved in the pathogenesis of T2DM, and further research on GNG3 may provide new targets for the development of drugs to treat obesity and relevant diseases.

GNAI1, also known as Gi, an adenylate cyclase inhibitor that inhibits the conversion of ATP to cAMP [36], can interact with other proteins. T cell differentiation may change its structure [37]. Studies showed that altered expression of GNAI1 was associated with the progression of inflammation and immune disease [38, 39]. Thus, GNAI1 may be considered to be a novel biomarker for T2DM.

ITGB1, a member of the integrin family, consists of 18 α and 8 β transmembrane subunits that form at least 24 different heterodimeric receptors allowing cells to adhere to extracellular matrix (ECM) proteins [40]. Integrins play an important role in mediating cell-to-cell and cell-to-ECM adhesion [41], especially between the extracellular environment and platelets, inflammatory cells, and the

vasculature [42]. The previous study has confirmed that integrin-mediated adhesion was preferentially mediated by ITGB1. Integrins have been shown to be involved in angiogenesis [43], which is a key pathological characteristic of diabetic microvascular complications and also is essential for homeostasis of adipose tissues. ITGB1 may be a therapeutic target for obesity [44]. Consistently, as a significant membrane gene identified in our study, ITGB1 was expressed differentially between T2DM and normal groups. ITGB1 may play central roles in all DEGs and have a close relationship with the development of obesity, T2DM, and its complications.

5. Conclusion

Our study tried to identify some candidate genes and pathway regulatory network closely related to T2DM by a series of bioinformatics analysis on DEGs between T2DM samples and normal samples. The findings in the current work may help us understand the underlying molecular mechanisms of T2DM. DEGs such as PIK3R1, RAC1, GNG3, GNAI1, CDC42, and ITGB1 have the potential to be used as targets for T2DM diagnosis and treatments. However, the lack of experimental validation is a limitation of this study. In the future, these prediction results obtained through bioinformatics analysis can be verified by further experimental studies such as qRT-PCR and Western blot.

Data Availability

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

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

We greatly appreciate the encouragements and supports from Prof. Yongzhi Lun (Department of Laboratory Medicine, School of Pharmacy and Medical Technology, Putian University, China). The study was supported by the Special Fund for Information Development of Shanghai Municipal Commission of Economy and Informatization (grant agreement number 201701014).

References

- [1] J. E. Shaw, R. A. Sicree, and P. Z. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," *Diabetes Research and Clinical Practice*, vol. 87, no. 1, pp. 4–14, 2010.
- [2] J. S. Williams, K. Bishu, C. E. Dismuke, and L. E. Egede, "Sex differences in healthcare expenditures among adults with diabetes: evidence from the medical expenditure panel survey, 2002-2011," *BMC Health Service Research*, vol. 17, no. 1, p. 259, 2017.
- [3] S. Cuschieri, "The genetic side of type 2 diabetes – a review," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 13, no. 4, pp. 2503–2506, 2019.
- [4] M. Nannini, M. A. Pantaleo, A. Maleddu, A. Astolfi, S. Formica, and G. Biasco, "Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives," *Cancer Treatment Reviews*, vol. 35, no. 3, pp. 201–209, 2009.
- [5] D. S. Karolina, A. Armugam, S. Tavintharan et al., "MicroRNA 144 impairs insulin signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus," *PLoS One*, vol. 6, no. 8, article e22839, 2011.
- [6] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 2012.
- [7] C. M. Taniguchi, B. Emanuelli, and C. R. Kahn, "Critical nodes in signalling pathways: insights into insulin action," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 2, pp. 85–96, 2006.
- [8] A. Klippel, C. Reinhard, W. M. Kavanaugh, G. Apell, M. A. Escobedo, and L. T. Williams, "Membrane localization of phosphatidylinositol 3-kinase is sufficient to activate multiple signal-transducing kinase pathways," *Molecular and Cellular Biology*, vol. 16, no. 8, pp. 4117–4127, 1996.
- [9] F. Mauvais-Jarvis, K. Ueki, D. A. Fruman et al., "Reduced expression of the murine p85alpha subunit of phosphoinositide 3-kinase improves insulin signaling and ameliorates diabetes," *Journal of Clinical Investigation*, vol. 109, no. 1, pp. 141–149, 2002.
- [10] K. Ueki, C. M. Yballe, S. M. Brachmann et al., "Increased insulin sensitivity in mice lacking p85 subunit of phosphoinositide 3-kinase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 1, pp. 419–424, 2002.
- [11] F. Giorgino, M. T. Pedrini, L. Matera, and R. J. Smith, "Specific increase in p85 α expression in response to dexamethasone is associated with inhibition of insulin-like growth factor-1 stimulated phosphatidylinositol 3-kinase activity in cultured muscle cells," *Journal of Biological Chemistry*, vol. 272, no. 11, pp. 7455–7463, 1997.
- [12] L. A. Barbour, J. Shao, L. Qiao et al., "Human placental growth hormone increases expression of the P85 regulatory unit of phosphatidylinositol 3-kinase and triggers severe insulin resistance in skeletal muscle," *Endocrinology*, vol. 145, no. 3, pp. 1144–1150, 2004.
- [13] A. K. Nevins and D. C. Thurmond, "Glucose regulates the cortical actin network through modulation of Cdc42 cycling to stimulate insulin secretion," *American Journal of Physiology-Cell Physiology*, vol. 285, no. 3, pp. C698–C710, 2003.
- [14] Z. Wang and D. C. Thurmond, "Mechanisms of biphasic insulin-granule exocytosis - roles of the cytoskeleton, small GTPases and SNARE proteins," *Journal of Cell Science*, vol. 122, no. 7, pp. 893–903, 2009.
- [15] M. Malacombe, M. Ceridono, V. Calco et al., "Intersectin-1L nucleotide exchange factor regulates secretory granule exocytosis by activating Cdc42," *The EMBO Journal*, vol. 25, no. 15, pp. 3494–3503, 2006.
- [16] Z. Wang, E. Oh, and D. C. Thurmond, "Glucose-stimulated Cdc42 signaling is essential for the second phase of insulin secretion," *Journal of Biological Chemistry*, vol. 282, no. 13, pp. 9536–9546, 2007.
- [17] A. Kowluru, S. E. Seavey, G. Li et al., "Glucose- and GTP-dependent stimulation of the carboxyl methylation of

- CDC42 in rodent and human pancreatic islets and pure beta cells. Evidence for an essential role of GTP-binding proteins in nutrient-induced insulin secretion,” *Journal of Clinical Investigation*, vol. 98, no. 2, pp. 540–555, 1996.
- [18] E. N. Rittmeyer, S. Daniel, S. C. Hsu, and M. A. Osman, “A dual role for IQGAP1 in regulating exocytosis,” *Journal of Cell Science*, vol. 121, no. 3, pp. 391–403, 2008.
- [19] V. L. Tokarz, P. E. MacDonald, and A. Klip, “The cell biology of systemic insulin function,” *The Journal of Cell Biology*, vol. 217, no. 7, pp. 2273–2289, 2018.
- [20] Q. Y. Huang, X. N. Lai, X. L. Qian et al., “Cdc42: a novel regulator of insulin secretion and diabetes-associated diseases,” *International Journal of Molecular Sciences*, vol. 20, no. 1, p. 179, 2019.
- [21] G. Sebastiani, E. Guarino, G. E. Grieco et al., “Circulating microRNA (miRNA) expression profiling in plasma of patients with gestational diabetes mellitus reveals upregulation of miRNA miR-330-3p,” *Frontiers in Endocrinology*, vol. 8, 2017.
- [22] N. Tapon and A. Hall, “Rho, Rac and Cdc42 GTPases regulate the organization of the actin cytoskeleton,” *Current Opinion in Cell Biology*, vol. 9, no. 1, pp. 86–92, 1997.
- [23] S. Nozaki, S. Ueda, N. Takenaka, T. Kataoka, and T. Satoh, “Role of RalA downstream of Rac1 in insulin-dependent glucose uptake in muscle cells,” *Cellular Signalling*, vol. 24, no. 11, pp. 2111–2117, 2012.
- [24] E. Manser, T. Leung, H. Salihuddin, Z. Zhao, and L. Lim, “A brain serine/threonine protein kinase activated by Cdc42 and Rac1,” *Nature*, vol. 367, no. 6458, pp. 40–46, 1994.
- [25] L. Sylow, T. E. Jensen, M. Kleinert et al., “Rac1 signaling is required for insulin-stimulated glucose uptake and is dysregulated in insulin-resistant murine and human skeletal muscle,” *Diabetes*, vol. 62, no. 6, pp. 1865–1875, 2013.
- [26] V. Sidarala, R. Veluthakal, K. Syeda, C. Vlaar, P. Newsholme, and A. Kowluru, “Phagocyte-like NADPH oxidase (Nox2) promotes activation of p38MAPK in pancreatic β -cells under glucotoxic conditions: evidence for a requisite role of Ras-related C3 botulinum toxin substrate 1 (Rac1),” *Biochemical Pharmacology*, vol. 95, no. 4, pp. 301–310, 2015.
- [27] A. Kowluru, “Role of G-proteins in islet function in health and diabetes,” *Obesity and Metabolism*, vol. 19, pp. 63–75, 2017.
- [28] R. A. Kowluru, M. Mishra, and B. Kumar, “Diabetic retinopathy and transcriptional regulation of a small molecular weight G-protein, Rac1,” *Experimental Eye Research*, vol. 147, pp. 72–77, 2016.
- [29] E. Shen, Y. Li, Y. Li et al., “Rac1 is required for cardiomyocyte apoptosis during hyperglycemia,” *Diabetes*, vol. 58, no. 10, pp. 2386–2395, 2009.
- [30] D. S. Evanko, M. M. Thiyagarajan, D. P. Siderovski, and P. B. Wedegaertner, “G $\beta\gamma$ isoforms selectively rescue plasma membrane localization and palmitoylation of mutant Gas and G α_q ,” *Journal of Biological Chemistry*, vol. 276, no. 26, pp. 23945–23953, 2001.
- [31] W. K. Lim, C. S. Myung, J. C. Garrison, and R. R. Neubig, “Receptor–G protein γ Specificity: γ 11 shows unique potency for A1Adenosine and 5-HT_{1A} Receptors†,” *Biochemistry*, vol. 40, no. 35, pp. 10532–10541, 2001.
- [32] D. E. Clapham and E. J. Neer, “G protein $\beta\gamma$ subunits,” *Annual Review of Pharmacology and Toxicology*, vol. 37, no. 1, pp. 167–203, 1997.
- [33] D. T. Lodowski, J. A. Pitcher, W. D. Capel, R. J. Lefkowitz, and J. J. Tesmer, “Keeping G proteins at bay: a complex between G protein-coupled receptor kinase 2 and Gbetagamma,” *Science*, vol. 300, no. 5623, pp. 1256–1262, 2003.
- [34] T. R. Hynes, E. A. Yost, C. M. Hartle, B. J. Ott, and C. H. Berlot, “Inhibition of G-protein $\beta\gamma$ signaling decreases levels of messenger RNAs encoding proinflammatory cytokines in T cell receptor-stimulated CD4+ T helper cells,” *Journal of Molecular Signaling*, vol. 10, 2015.
- [35] W. F. Schwindinger, B. M. Borrell, L. C. Waldman, and J. D. Robishaw, “Mice lacking the G protein γ 3-subunit show resistance to opioids and diet induced obesity,” *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 297, no. 5, pp. R1494–R1502, 2009.
- [36] V. V. Pineda, J. I. Athos, H. Wang et al., “Removal of G(ialpha1) constraints on adenylyl cyclase in the hippocampus enhances LTP and impairs memory formation,” *Neuron*, vol. 41, no. 1, pp. 153–163, 2004.
- [37] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016.
- [38] S. Rivetti, M. Lauriola, M. Voltattorni et al., “Gene expression profile of human colon cancer cells treated with cross-reacting material 197, a diphtheria toxin non-toxic mutant,” *International Journal of Immunopathology and Pharmacology*, vol. 24, no. 3, pp. 639–649, 2011.
- [39] S. A. Diehl, B. McElvany, R. Noubade et al., “G proteins G α i1/3 are critical targets for Bordetella pertussis toxin-induced vasoactive amine sensitization,” *Infection and Immunity*, vol. 82, no. 2, pp. 773–782, 2014.
- [40] R. O. Hynes, “Integrins,” *Cell*, vol. 110, no. 6, pp. 673–687, 2002.
- [41] Y. Takada, X. Ye, and S. Simon, “The integrins,” *Genome Biology*, vol. 8, no. 5, p. 215, 2017.
- [42] S. Weng, L. Zemany, K. N. Standley et al., “ α 3 integrin deficiency promotes atherosclerosis and pulmonary inflammation in high-fat-fed, hyperlipidemic mice,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 11, pp. 6730–6735, 2003.
- [43] P. C. Brooks, A. M. P. Montgomery, M. Rosenfeld et al., “Integrin α v β 3 antagonists promote tumor regression by inducing apoptosis of angiogenic blood vessels,” *Cell*, vol. 79, no. 7, pp. 1157–1164, 1994.
- [44] R. Cheng and J. Ma, “Angiogenesis in diabetes and obesity,” *Reviews in Endocrine and Metabolic Disorders*, vol. 16, no. 1, pp. 67–75, 2015.

Retraction

Retracted: circFAT1(e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis

Computational and Mathematical Methods in Medicine

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] J. Liu, H. Li, C. Wei et al., "circFAT1(e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1459368, 9 pages, 2020.

Research Article

circFAT1(e2) Promotes Papillary Thyroid Cancer Proliferation, Migration, and Invasion via the miRNA-873/ZEB1 Axis

Jiazhe Liu, Hongchang Li, Chuanchao Wei, Junbin Ding, Jingfeng Lu, Gaofeng Pan , and Anwei Mao 

Minhang Hospital, Fudan University, 170 Xin-Song Road, Shanghai 201199, China

Correspondence should be addressed to Gaofeng Pan; panda_gaofeng@fudan.edu.cn and Anwei Mao; anwei_mao@fudan.edu.cn

Received 23 June 2020; Revised 14 September 2020; Accepted 17 September 2020; Published 20 October 2020

Academic Editor: Tao Huang

Copyright © 2020 Jiazhe Liu 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.

Circular RNAs (circRNAs) play an extremely important regulatory role in the occurrence and development of various malignant tumors including papillary thyroid cancer (PTC). circFAT1(e2) is a new type of circRNA derived from exon 2 of the FAT1 gene, which is distributed in the cytoplasm and nucleus of PTC cells. However, so far, the role of circFAT1(e2) in PTC is still unclear. In this study, circFAT1(e2) was found to be highly expressed in PTC cell lines and tissues. circFAT1(e2) knockdown suppressed PTC cell growth, migration, and invasion. Also, circFAT1(e2) acted as a sponge for potential microRNAs (miRNAs) to modulate cancer progression. A potential miRNA target was discovered to be miR-873 which was targeted by circFAT1(e2) in PTC. The dual-luciferase assay conducted later also confirmed that there was indeed a direct interaction between circFAT1(e2) and miR-873. This study also confirmed that circFAT1(e2) inhibited the miR-873 expression and thus promoted the ZEB1 expression, thus affecting the proliferation, metastasis, and invasion of PTC cells. In conclusion, the results of this study indicated that circFAT1(e2) played a carcinogenic role by targeting the miR-873/ZEB1 axis to promote PTC invasion and metastasis, which might become a potential novel target for therapy of PTC.

1. Introduction

Circular RNA (circRNA) is a special type of alternative splicing (called reverse splicing) that produces a specifically and covalently closed-loop structural RNA [1]. Recently, several studies have shown that circRNAs are differentially expressed in various diseases [2], including cancer, atherosclerotic vascular diseases, and neurological diseases. This may suggest that circRNA could play a potential regulatory role in the progression of some diseases [3]. For example, circRNAs could bind to RNA polymerase II in a variety of ways to form complexes that could regulate the activity of RNA polymerase II and then affect parental gene transcription thereby [4]. According to the study of Abdelmohsen et al., circular PABPN1 (polyadenylate-binding nuclear protein 1) and HuR (human antigen R) can be extensively bound to form the HuR-circRNA complex, which competitively inhibits the binding of HuR and PABPN1 mRNA, lead-

ing to a decrease in the PABPN1 translation, thus disrupting the normal metabolism of cells [5]. Studies have also shown that the binding of circRNA UBAP2 to miR-143 can abolish the inhibition of Bcl-2 and the caspase apoptosis pathway [6].

Papillary thyroid cancer (PTC) is one of the most popular cancers in females [7]. The global incidence of PTC ranks 9th among all cancers [8]. The factors contributing to the progression of PTC are unclear. Despite several pieces of evidence have indicated that obesity, smoking, hormonal exposure, and certain environmental contaminants may be related to PTC [9–12], the only risk factor validated in PTC is ionizing radiation [13]. Therefore, there was an urgent need to identify novel regulators for the understanding of molecular mechanisms and biomarkers for the prognosis of this cancer [14]. circRNA plays an important role in thyroid cancer. For example, circRNA circZFR promotes the expression of C8orf4 by acting as a sponge on miR-1261 and facilitates the growth of PTC cells [15]. circRNA circRNA_

102171 regulates CTNNBIP1-dependent activation through the β -catenin pathway to promote PTC progression [16]. circRNA circ-ITCH inhibits the progression of PTC via the miR-22-3p/CBL/ β -catenin pathway.

circFAT1(e2) is a type of circRNA derived from exon 2 of the FAT1 gene and mainly exists in the cytoplasm of gastric cancer cells. Studies have shown that circFAT1(e2) is reduced in tissues and cell lines in gastric cancer (GC). Furthermore, mechanism analysis indicated overexpressed circFAT1(e2) could hinder the proliferation and metastasis of GC cells [17]. In this study, we focused on the role of circFAT1(e2) in PTC, and this finding might promote the prognosis and treatment of PTC.

2. Materials and Methods

2.1. Cell Lines and Cell Culture. The human thyroid normal cell line Nthy-ori 3-1 and the PTC cell lines CAL-62, TPC-1, and K1 were all from ATCC (Manassas, VA, USA). All cells were placed in an incubator at 37°C containing 5% CO₂ and cultured using DMEM (BI, Israel) supplemented with 10% FBS (Invitrogen).

2.2. Cell Fractionation Assay. The cell fractionation assay was conducted according to a previous report [18]. Approximately 3×10^6 cells were grown on 10 cm dishes (Corning), trypsinised, washed in cold 1x PBS, and centrifuged (1200 rpm, 5 min). Pellets were lysed in 1 mL hypotonic lysis buffer (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM MgCl₂, 1 mM β -mercaptoethanol, 0.075% NP-40, 1x murine RNase inhibitor, 1x protease/phosphatase inhibitor cocktails, Roche) and incubated for 15 min at 4°C with rotation. Nuclei were pelleted by centrifugation (1200 rpm, 4°C) for 15 min. The cytoplasm was collected from the supernatant. Nuclei were washed three times in 800 μ L PBS and collected as the pelleted nuclear fraction. Fractionated cytoplasmic and nuclear lysates were confirmed by localization of GAPDH and histone 3.1, respectively.

2.3. Dual-Luciferase Reporter Assay. For the sake of constructing luciferase reporter vectors, the whole circFAT1(e2) or the 3'UTR fragment of ZEB1 that contained the expected latent binding sites was cloned into the pmiR-RB-REPORTTM luciferase reporter vector (RiboBio, Guangzhou, China) at the XhoI and NotI sites; the same was true for the construction of a mutant sequence of circFAT1(e2) or ZEB1 at the 3'UTR.

For the dual-luciferase activity assay, we applied Lipofectamine 2000 (Invitrogen) to cotransfect each construct with marked miRNAs (RiboBio) in PTC cells for 48 h. The Dual-Luciferase Reporter Assay System (Promega, WI, USA) was performed under the instructions of the manufacturer. Besides, the BioTek Synergy HTX multimode reader was used to obtain the luminescent signals, and the luciferase activities were displayed according to the relative hRluc/hluc ratio.

2.4. qRT-PCR. Total RNA was extracted with the TRIzol reagent (Invitrogen, USA). For circRNA and mRNA, the reverse transcription of total RNA to cDNA was performed

by reverse transcriptase (Vazyme, Nanjing, China). And then, we conducted qPCR with a SYBR Green PCR Kit (Vazyme, Nanjing, China). The fluorescence quantitative PCR instrument was QuantStudioTM 6 Flex manufactured by Thermo Fisher Scientific (USA). Besides, Sangon (Shanghai, China) was used to construct all the primer sequences. GAPDH was selected as the reference gene for circRNA and mRNA, and the internal control for miRNA was set as U6. We also used the $2^{-\Delta\Delta Ct}$ method to quantify the gene expression.

2.5. CCK-8 Assay. According to the instruction of the manufacturer (Dojindo Laboratories, Kumamoto, Japan), the proliferative ability of PTC cells was assessed with a CCK-8 assay. Next, we plated the CAL-62 and TPC-1 cells (1×10^3 cell/well) in 96-well plates, treated them with 10 μ L of CCK-8 solution for 2 hours, and then analyzed spectrophotometrically at 450 nm by an automatic microplate reader.

2.6. Transwell Migration and Invasion Assays. The transwell chamber (Corning, NY, USA) was used to conduct the assays of cell migration and invasion. It could be directly used for migration assay, or with Matrigel mix (BD Biosciences, San Jose, CA, USA) for invasion assay. After incubation for 48 h, we used cotton swabs to scrape the cells which settled on the upper layers of the transwell chambers and fixed the cells settled on the lower surfaces. Next, we observed the number of cells in the transwell chambers under a fluorescent inverted microscope and took a photo.

2.7. Statistical Analysis. We used the form average value \pm SD to present a continuous variable. We used one-way ANOVA [19] and Student's *t*-test [20] for multiple comparisons. All the analyses were performed in the software GraphPad Prism, v5.0 (GraphPad, La Jolla, CA, USA). In this paper, a significant difference was identified by a *p* value < 0.05.

3. Results

3.1. circFAT1(e2) Increased Abnormally in PTC Samples and Cell Lines. Compared with Nthy-ori 3-1 cells, the expression levels of circFAT1(e2) in K1, TPC-1, and CAL-62 cells were increased by onefold, twofold, and fourfold, respectively (Figure 1(a)). Furthermore, the circFAT1(e2) expression levels in 12 pairs of PTC tissues were also detected using qRT-PCR. The results indicated that the circFAT1(e2) expression in PTC tissues was higher than that in matched normal tissues remarkably (Figure 1(b)).

3.2. Silencing circFAT1(e2) Suppressed PTC Cell Proliferation. To determine the biological roles of circFAT1(e2) in TC, we firstly detected its subcellular location in CAL-62 and TPC-1 cells. Our results indicated that circFAT1(e2) is mainly located in the cytoplasm of TC cells (Figure 2(c)). Meanwhile, nuclear-located U6 and cytoplasm-located GAPDH were also detected as a positive control (Figures 2(a) and 2(b)). Next, we knocked down the expression levels of circFAT1(e2) in TC cells using a specific siRNA targeting of this circRNA (Figures 2(d) and 2(f)). In order to search the function of circFAT1(e2) knockdown on cell proliferation in TC cells, CCK-8 assays were performed. Our results

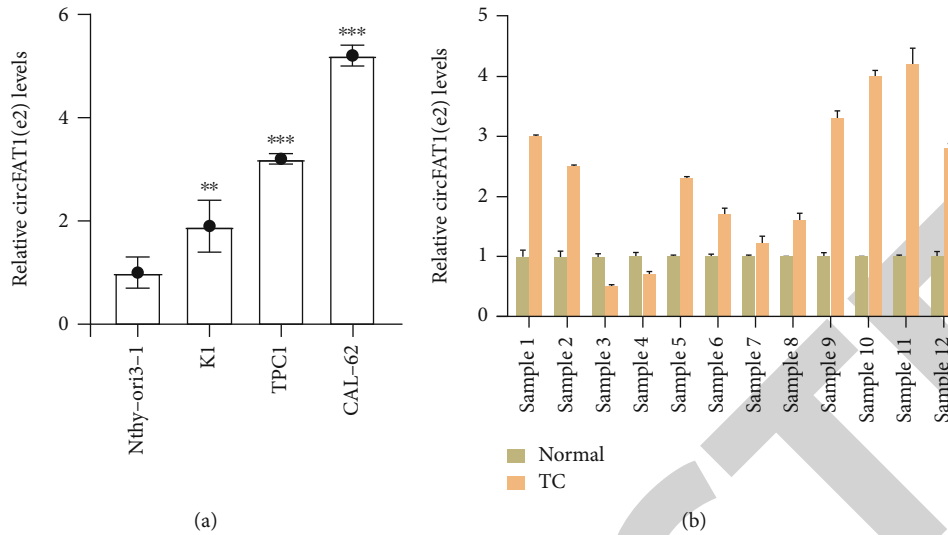


FIGURE 1: circFAT1(e2) increased abnormally in PTC samples and cells. (a) The circFAT1(e2) was more expressed in PTC cells. (b) PTC tissues had higher expression level of circFAT1(e2). * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

showed that the proliferation rate of cells with si-circFAT1(e2) was remarkably downregulated compared to the control group in both CAL-62 (Figure 2(e)) and TPC-1 cells (Figure 2(g)).

3.3. circFAT1(e2) Suppressed PTC Cell Metastasis In Vitro after Knockdown. Furthermore, we then tested whether circFAT1(e2) would affect the metastatic abilities of PTC cells. Migration assay results indicated that circFAT1(e2) knockdown remarkably weakened the migratory capability of CAL-62 (Figures 3(a) and 3(b)) and TPC-1 cells (Figures 3(c) and 3(d)). Besides, we detected the invasion ability of PTC cells by estimating the penetration of cells through Matrigel in a transwell chamber. As illustrated in Figure 4, we found that circFAT1(e2) knockdown inhibited cell invasion. The numbers of invading cells were, respectively, decreased by 80 percent and 85 percent in CAL-62 (Figures 4(a) and 4(b)) and TPC-1 (Figures 4(c) and 4(d)) cells transfected with si-circFAT1(e2) compared with the negative control group. Altogether, these results reveal that circFAT1(e2) is a positive regulator of TC metastasis.

3.4. circFAT1(e2) Served as a Sponge of miR-873 to Promote ZEB1. Previous studies have shown that circRNA could serve as sponges for miRNAs. In this study, we proposed that circFAT1(e2) might also serve as a miRNA sponge in TC. circFAT1(e2) is mainly located in the cytoplasm of TC cells. circFAT1(e2) may regulate the expression of target proteins at the posttranscriptional level, indicating that circFAT1(e2) may be the ceRNA of miRNAs. To test this hypothesis, we used miRanda (<https://www.microrna.org/microrna/home.do>) to predict the potential interaction between miRNAs and circFAT1(e2). We found that circFAT1(e2) and miR-873 had binding sites. Using HumanTargetScan, ZEB1 was predicted by bioinformatics as a potential target of miR-873 (http://www.targetscan.org/cgi-bin/targetscan/vert_71/). Previous studies have found that miR-873 is involved in the

disease by targeting ZEB1 progress [21, 22]. ZEB1 was reported to be an oncogene in human cancers. qRT-PCR assay indicated that circFAT1(e2) knockdown significantly suppressed the expression levels of ZEB1 in both TPC-1 and CAL-62 (Figures 5(a) and 5(b)).

In order to further validate these findings, we detected the expression levels of circFAT1(e2) and ZEB1 after overexpressing miR-873 in PTC cells and found that overexpression of miR-873 remarkably suppressed the RNA levels of circFAT1(e2) and ZEB1 in TC cells (Figures 5(c)–5(f)). Dual-luciferase assays were performed to verify whether miR-873 could straightly interact with circFAT1(e2) and ZEB1. Luciferase reporters were cotransfected with miR-873 into PTC cells. The pmiR-RB-Report vector containing 3'-UTR regions of ZEB1 or circFAT1(e2) cotransfected with miR-873 mimic significantly reduced the Renilla/firefly ratio in contrast to the control vector (Figures 5(g) and 5(i)). However, overexpression of miR-873 did not significantly influence the luciferase activity of the pmiR-RB-Report vector, which contained 3'-UTR-mut regions of ZEB1 or circFAT1(e2)-mut (Figures 5(h) and 5(j)). These results indicated that miR-873 could directly target circFAT1(e2) and ZEB1.

4. Discussion

In this study, circFAT1(e2) was found to be higher in papillary thyroid tumors and cell lines than that in normal thyroid tissues and cells. Interestingly, the downregulation of circFAT1(e2) modulated miRNA-873/ZEB1 signaling pathways. More importantly, circFAT1(e2) knockdown inhibited some physiological functions of PTC cells.

Many studies also indicated that circRNA played a crucial role in multiple types of human cancers, such as breast cancer, liver cancer, gastric cancer, and thyroid cancer. For instance, in Wei et al.'s study, it is showed that circZFR contributes to the growth of PTC cells by the miR-

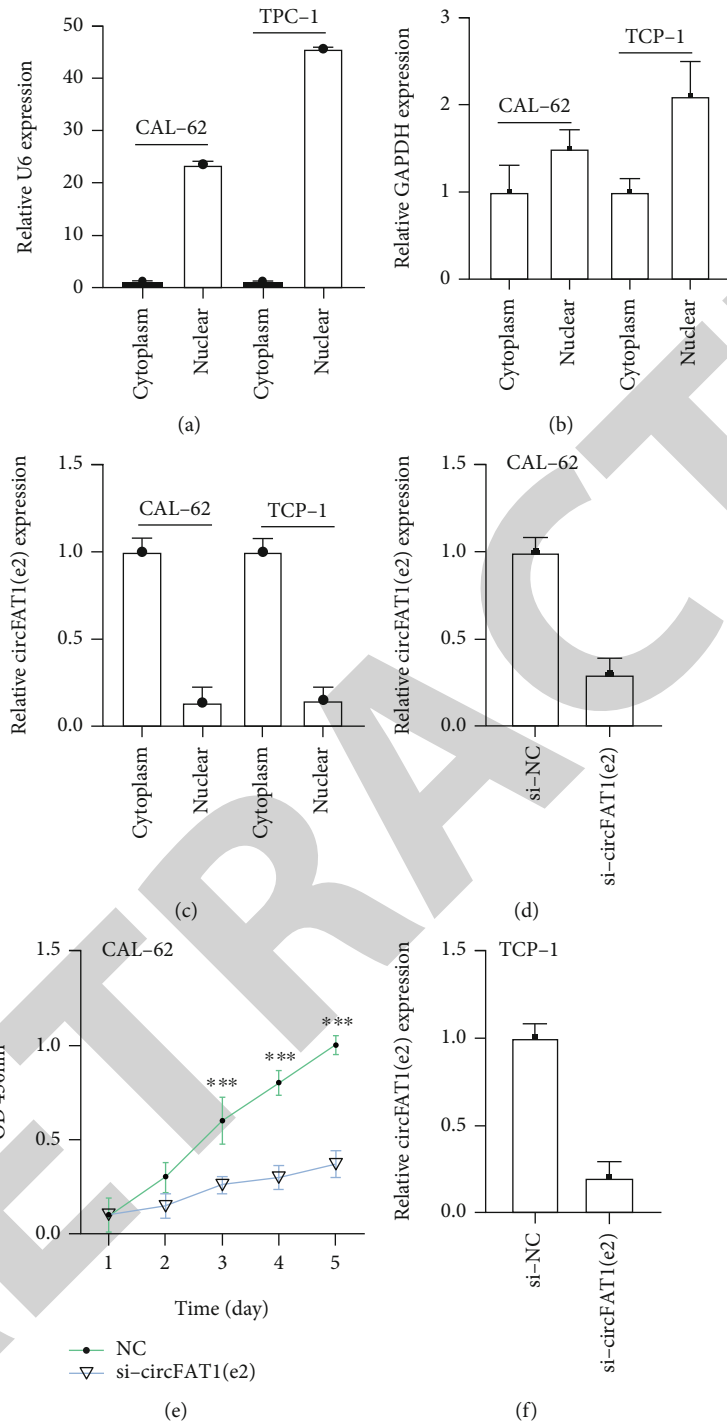


FIGURE 2: Continued.

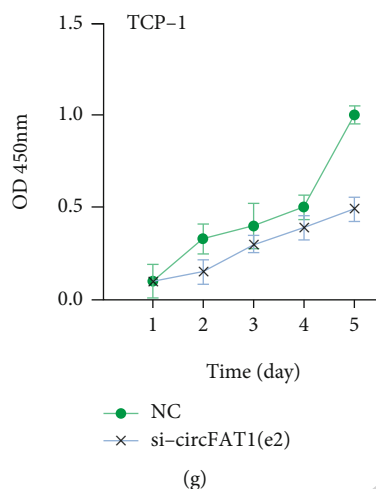


FIGURE 2: Silencing circFAT1(e2) suppressed PTC cell proliferation. (a-c) circFAT1(e2) mainly located in the cytoplasm of TC cells. The efficiency of circFAT1(e2) knockdown was detected by qRT-PCR in CAL-62 (d) and TPC-1 (f) cells. CCK-8 experiments showed that circFAT1(e2) knockdown attenuated the proliferation capacity of CAL-62 (e) and TPC-1 (g) cells. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

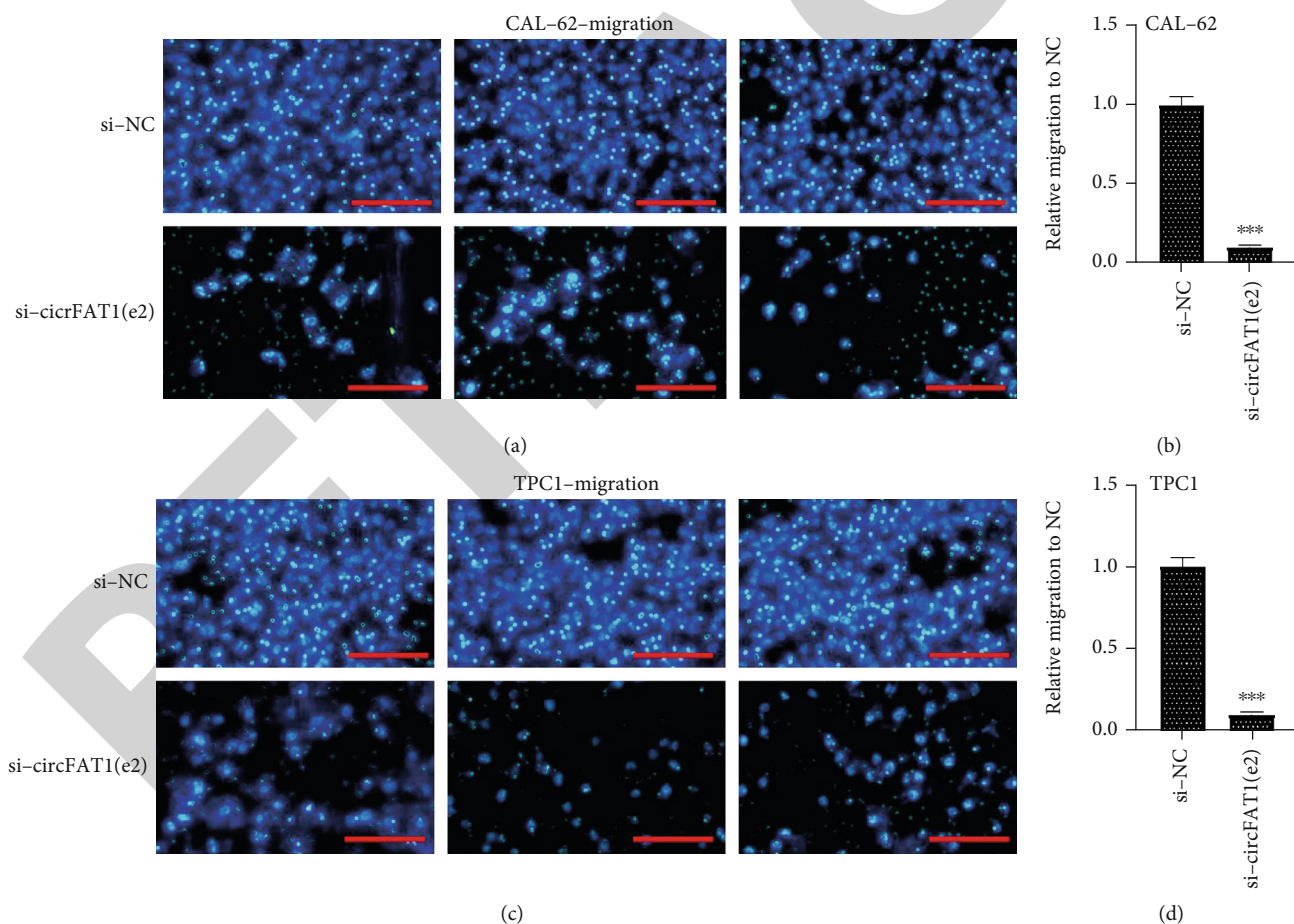


FIGURE 3: circFAT1(e2) knockdown remarkably weakened the migratory capability of CAL-62 (a, b) and TPC-1 (c, d) cells. (a) Representative images of transwell migration assays in CAL-62 cells (scale bar, 45 μm). (b) The quantification of transwell migration assays in CAL-62 cell. (c) Representative images of transwell migration assays in TPC-1 cells (scale bar, 45 μm). (d) The quantification of transwell migration assays in TPC-1 cell. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

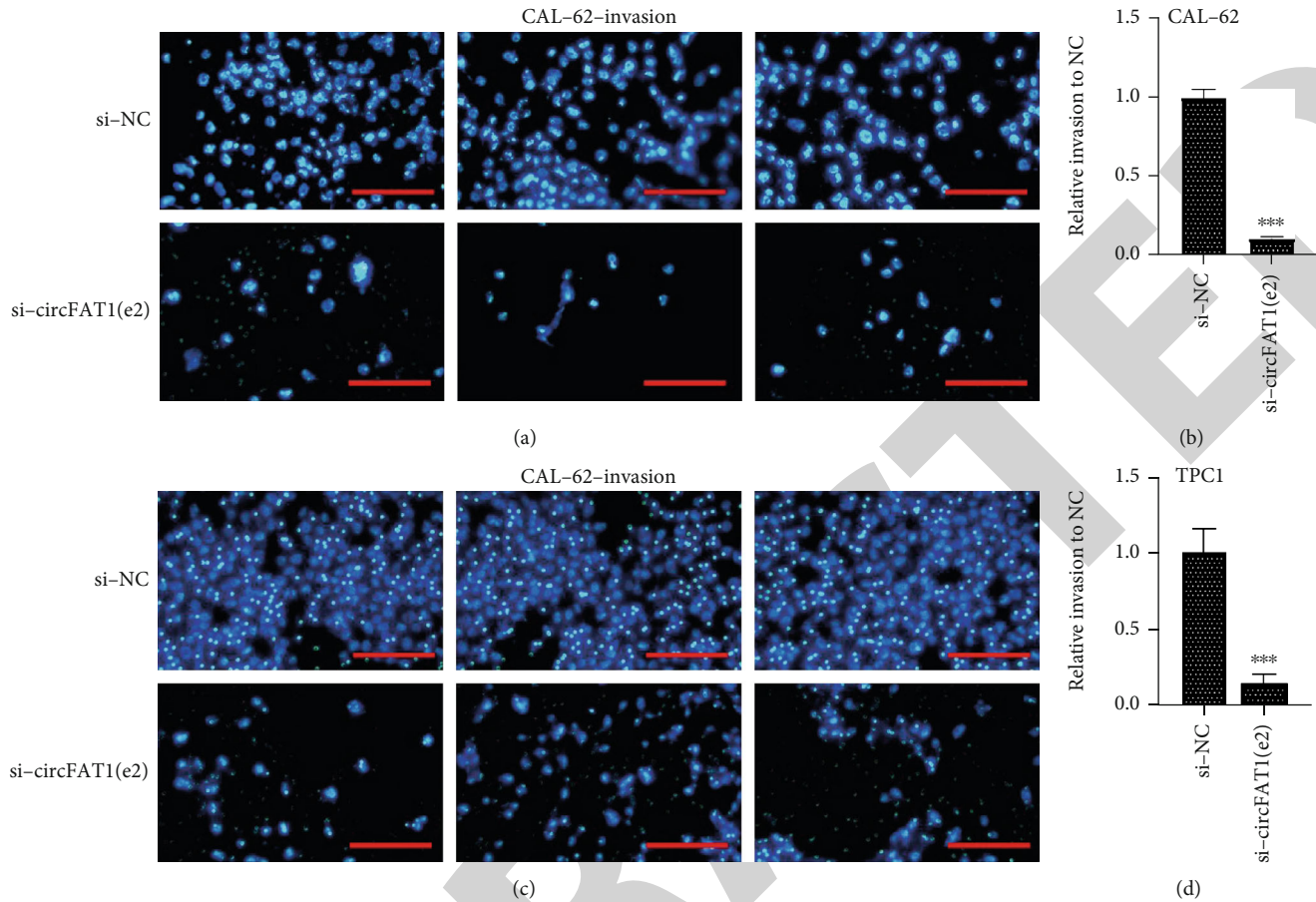


FIGURE 4: circFAT1(e2) silenced inhibited invasive capability of CAL-62 (a, b) and TPC-1 (c, d) cells. (a) Representative images of transwell invasion assays in CAL-62 cells (scale bar, 45 μm). (b) The quantification of transwell invasion assays in CAL-62 cell. (c) Representative images of transwell invasion assays in TPC-1 cells (scale bar, 45 μm). (d) The quantification of transwell invasion assays in TPC-1 cell. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

1261/C8orf4 axis [15]. By using the microarray analysis of circRNA in thyroid cancer, 98 circRNAs were found to be dysregulated. It was also found that circRNA-100395 had a significant potential for interaction with cancer-associated miRNAs [23]. Wang et al. revealed that circ-ITCH could inhibit the progression of PTC by regulating miR-22-3p/CBL/ β -catenin cascade [24]. While it was unclear how circFAT1(e2) functions in PTC. This study showed elevated levels of circFAT1(e2) in PTC. And it displayed that circFAT1(e2) promoted the development of PTC cells by affecting proliferation, invasion, and migration. The results in our study also suggested that circFAT1(e2) played a carcinogenic role in thyroid cancer, suggesting that circFAT1(e2) might be a potential therapeutic target.

In gastric cancer (GC), a novel circRNA circFAT1(e2) has been identified. circFAT1(e2) is significantly downregulated in GC tissue and is related to the overall survival rate of GC patients. circFAT1(e2) is distributed in the cytoplasm and nucleus of GC cells. circFAT1(e2) in the nucleus can directly interact with Y-box-binding protein 1 (YBX1) and inhibit its function. circFAT1(e2) in the cytoplasm exerts a tumor suppressor effect by regulating the miR-548g/RUNX1 axis. Moreover, this study also demonstrates that overexpressed

circFAT1(e2) induces the proliferation, migration, and invasion of GC cells, and it plays a tumor-suppressive role in GC [17]. Through our research, circFAT1(e2) is upregulated in PTC cells and tissues and may participate in cell proliferation, cell metastasis, cell invasion, and other biological processes. Also, we proved that the knockdown of circFAT1(e2) could inhibit the expression of ZEB1. miR-873 overexpression strongly suppressed the expression levels of ZEB1 and circFAT1(e2). Luciferase assay showed that miR-873 could reduce the activity of pmir-RB-Report vector containing 3'-UTR regions of ZEB1 or circFAT1(e2). These results demonstrated that both ZEB1 and circFAT1(e2) were the direct targets of miR-873.

It has been demonstrated that miR-873 played different roles in different cancers [25]. It has been identified as an oncogene in lung adenocarcinoma [26], however, as a suppressor in breast cancer, ovarian cancer, and glioblastoma [27]. In glioma cells, overexpression of miR-873 leads to a decrease in Bcl-2, which in turn inhibits cell proliferation, cell metastasis, and cell invasion [28, 29]. In breast cancer cells, miR-873 not only depresses breast cancer cell proliferation but also enhances tamoxifen resistance [30]. Similarly, miR-873 has been shown to bind to IGF2BP1 in glioblastoma cells and depress the growth and metastasis of glioblastoma cells

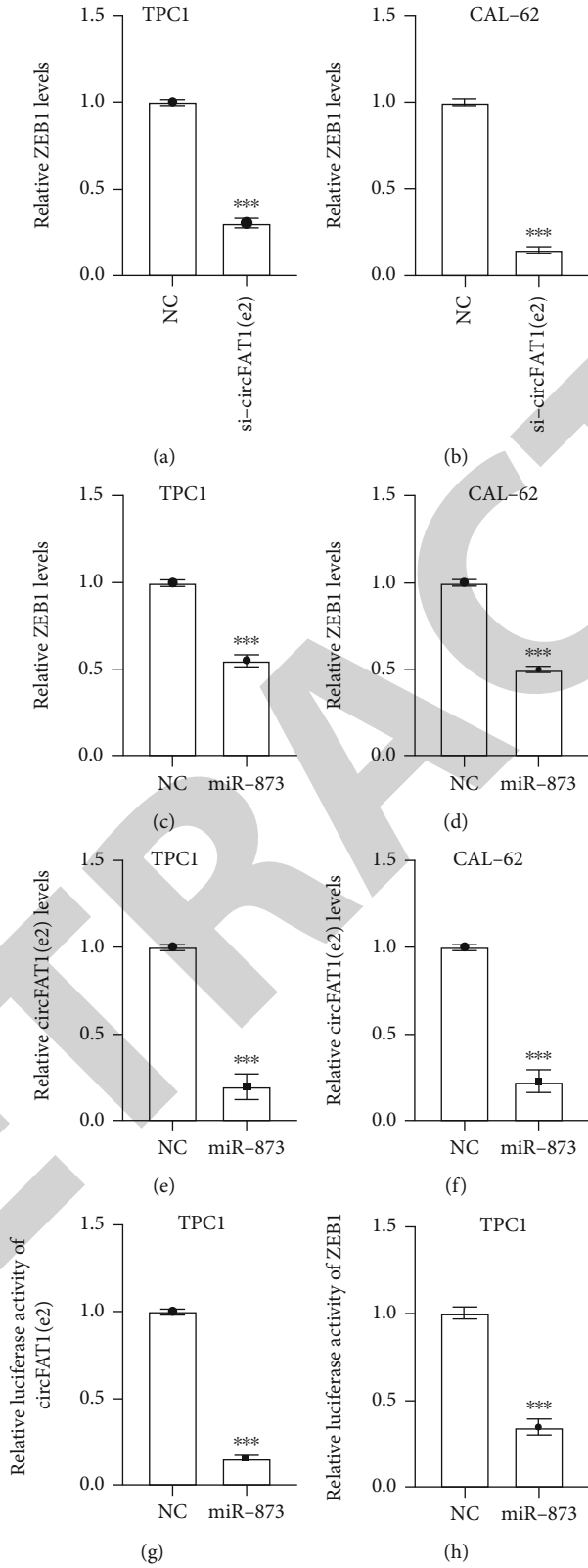


FIGURE 5: Continued.

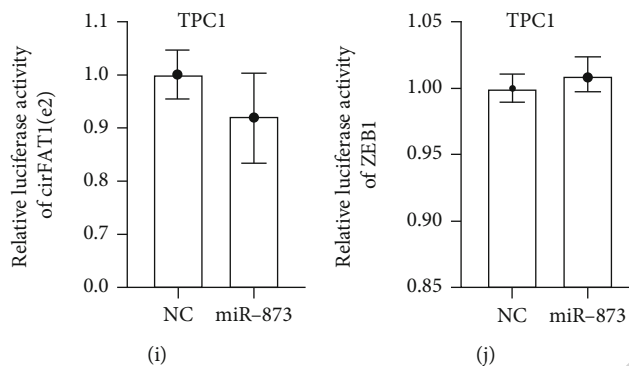


FIGURE 5: circFAT1(e2) served as a sponge of miR-873 to augment ZEB1. ZEB1 expression levels were lower in TPC-1 (a) and CAL-62 (b) cells with circFAT1(e2) knockdown by qRT-PCR. TPC-1 (c, e) and CAL-62 (d, f) cells transfected with miR-873 had less expression of circFAT1(e2), ZEB1. The luciferase activities of circFAT1(e2) and ZEB1 were measured in TPC-1 (g, i) and CAL-62 (h, j) cells transfected miR-873 mimics or miR-NC with dual-luciferase reporter assay. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

[31]. The functions of miR-873 in different cancer types could be different relying on the particular cellular environment. Consistently, it was revealed that miR-873 was down-regulated in PTC cells according to our qRT-PCR results. ZEB1 is a transcription factor known for its ability to induce EMT carcinogenesis, which plays a role in cells through various mechanisms including Wnt, NF- κ B, and miRNAs [32]. In breast cancer cells, the transfection of miR-873 mimics decreased ZEB1 expression. ZEB1 can act as a transcriptional activator to activate YAP1 target genes, including the expression of AXL, CTGF, and CYR61, and also as a transcriptional repressor to inhibit the expression of the target gene (E-cadherin) consistently [33]. This study also confirmed that circFAT1(e2) inhibited the miR-873 expression and thus promoted the ZEB1 expression, thus affecting the proliferation, metastasis, and invasion of PTC cells.

In conclusion, we have proved that circFAT1(e2) promotes the tumorigenesis and invasiveness of PTC. It participates in the regulation of PTC by competitively binding with miR-873 and upregulating the expression of its target gene ZEB1. These findings suggest that the circFAT1(e2)-miR-873-ZEB1 axis may be a promising target for the prognosis and therapy of PTC.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

Gaofeng Pan and Anwei Mao conceptualized and designed the study. Chuanchao Wei developed the methodology. Junbin Ding collected the sample. Jingfeng Lu analyzed and interpreted the data. Jiazhe Liu and Hongchang Li wrote, reviewed, and revised the manuscript. Jiazhe Liu and Hongchang Li contributed equally to this work.

References

- [1] V. M. Conn, V. Hugouvieux, A. Nayak et al., "A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation," *Nature Plants*, vol. 3, no. 5, p. 17053, 2017.
- [2] Z. Zhu, Y. Li, W. Liu et al., "Comprehensive circRNA expression profile and construction of circRNA-associated ceRNA network in fur skin," *Experimental Dermatology*, vol. 27, no. 3, pp. 251–257, 2018.
- [3] S. Meng, H. Zhou, Z. Feng et al., "CircRNA: functions and properties of a novel potential biomarker for cancer," *Molecular Cancer*, vol. 16, no. 1, p. 94, 2017.
- [4] Z. Zhou, X. Li, C. Deng, P. A. Ney, S. Huang, and J. Bungert, "USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the beta-globin gene locus," *The Journal of Biological Chemistry*, vol. 285, no. 21, pp. 15894–15905, 2010.
- [5] K. Abdelmohsen, A. C. Panda, R. Munk et al., "Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1," *RNA Biology*, vol. 14, no. 3, pp. 361–369, 2017.
- [6] H. Zhang, G. Wang, C. Ding et al., "Increased circular RNA UBAP2 acts as a sponge of miR-143 to promote osteosarcoma progression," *Oncotarget*, vol. 8, no. 37, pp. 61687–61697, 2017.
- [7] A. Chong, H. C. Song, J. J. Min et al., "Improved detection of lung or bone metastases with an I-131 whole body scan on the 7th day after high-dose I-131 therapy in patients with thyroid cancer," *Nuclear Medicine and Molecular Imaging*, vol. 44, no. 4, pp. 273–281, 2010.
- [8] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [9] J. Ma, M. Huang, L. Wang, W. Ye, Y. Tong, and H. Wang, "Obesity and risk of thyroid cancer: evidence from a meta-analysis of 21 observational studies," *Medical Science Monitor*, vol. 21, pp. 283–291, 2015.
- [10] K. N. Kim, Y. Hwang, K. Kim et al., "Active and passive smoking, BRAFV600E mutation status, and the risk of papillary thyroid cancer: a large-scale case-control and case-only study," *Cancer Research and Treatment*, vol. 51, no. 4, pp. 1392–1399, 2019.

Research Article

Imrecoxib Inhibits Paraquat-Induced Pulmonary Fibrosis through the NF- κ B/Snail Signaling Pathway

Haihao Jin 

Department of Traditional Chinese Medicine, Liuzhou People's Hospital, Liuzhou, 545000 Guangxi Zhuang Minority Autonomous Region, China

Correspondence should be addressed to Haihao Jin; jingben44638363@163.com

Received 26 June 2020; Revised 3 August 2020; Accepted 4 August 2020; Published 14 October 2020

Guest Editor: Tao Huang

Copyright © 2020 Haihao Jin. 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.

Objective. In recent years, pulmonary fibrosis caused by paraquat poisoning is still concerned. However, no effective drugs have been developed yet to treat paraquat-induced pulmonary fibrosis. The aim of our research is to investigate whether imrecoxib can inhibit paraquat-induced pulmonary fibrosis and its possible mechanism. **Methods.** Extraction of primary pulmonary fibrosis cells (PPF cells) in vitro by the method of trypsin digestion. RT-qPCR and western blot were employed to measure the transcription level and protein expression of EMT related markers in paraquat-induced A549 cells. MTT, wound-healing, and Transwell experiments were used to verify the effect of imrecoxib on the proliferation, migration, and invasion of PPF and HFL1 cells. **Results.** Firstly, our results confirmed that paraquat can induce EMT and activate the NF- κ B/snail signal pathway in lung epithelial cell A549. Furthermore, experimental results showed that imrecoxib could repress the proliferation, migration, and invasion of PPF and HFL1 cells. Finally, our study found that imrecoxib can inhibit EMT of paraquat-induced A549 cells by the NF- κ B/snail signal pathway. **Conclusion.** Imrecoxib can inhibit EMT of paraquat-induced A549 cells and alleviate paraquat-caused pulmonary fibrosis through the NF- κ B/snail signal pathway. Therefore, imrecoxib is a drug worthy of study in the treatment of paraquat-induced pulmonary fibrosis.

1. Introduction

Human pulmonary fibrosis (PF), especially idiopathic pulmonary fibrosis (IPF), is a fatal and chronic disease that causes fibroblast proliferation and excessive deposition of relative protein, which in turn disrupts the lung structure and function and eventually leads to respiratory failure [1–3]. In recent years, the mortality of pulmonary fibrosis is increasing, but the current drugs have no obvious therapeutic effect [4, 5]. At present, the drugs used to treat pulmonary fibrosis are either ineffective or have too many side effects. For example, the drug nintedanib can be used to treat IPF, but it can cause great hepatotoxicity to patients. Clinically, pirfenidone can be used to treat mild or moderate idiopathic pulmonary fibrosis, but it has many side effects, such as skin photosensitivity, liver injury, vomiting, and gastrointestinal reaction. Using glucocorticoids, such as dexamethasone, to treat pulmonary fibrosis through anti-inflammation is a classic treatment, but it can cause serious side effects such as the

liver and kidney dysfunction and osteoporosis [6–10]. Therefore, it is urgent to find a drug that can not only effectively treat pulmonary fibrosis but also produce fewer side effects on patients.

Paraquat is a nonselective, contact, low pollution, low residue, broad-spectrum, and efficient herbicide, which is widely used in agricultural production. Pulmonary fibrosis is the main cause of death of paraquat poisoning. In recent years, pulmonary fibrosis caused by paraquat poisoning is still concerned. Many studies have proved that EMT is one of the main causes of paraquat induced pulmonary fibrosis and paraquat poisoning in humans and animals. Epithelial-mesenchymal transition (EMT) is a specific procedure, causing cells phenotypes transformed. It plays an indispensable function in embryonic, tissue reconstruction, cancer metastasis, inflammation, and many fibrotic diseases. Its obvious characteristics are the decrease of the expression of E-cadherin, the increase of vimentin, and the change of cells morphology. During fibrosis, an EMT-like process occurs

[11, 12]. As a very important factor of EMT, NF- κ B can promote the release of a large number of inflammatory related factors, such as TNF- α , IL, and TGF- β . Many studies have shown that TGF- β is related to pulmonary fibrosis. For example, some studies have confirmed that doxycycline can regulate the balance between epithelial cells and mesenchymal cells and affect the TGF- β signal pathway to achieve the purpose of treating paraquat-induced pulmonary fibrosis [13]. The epithelial-to-mesenchymal transition of pulmonary fibrosis is induced through the TGF- β /smad signaling pathway [14]. Parthenolide is able to influence the level of EMT-related proteins and to treat pulmonary fibrosis by affecting the progression of pulmonary fibrosis through the NF- κ B/Snail signaling pathway [15]. Therefore, EMT, NF- κ B, and related inflammatory factors (TGF- β , TNF- α , and IL) are very important in the evaluation of pulmonary fibrosis.

Imrecoxib is a cyclooxygenase-2 (COX-2) inhibitor with anti-inflammatory effect, mainly used in the treatment of arthritis pain [16]. The drug was approved for marketing by the State Food and Drug Administration (SFDA) on May 20, 2011. Imrecoxib can not only inhibit inflammation and pain but also reduce the risk of gastrointestinal tract stimulation and cardiovascular injury [17]. The therapeutic effect of imrecoxib arthritis has been confirmed, but there are few studies on the effect of imrecoxib on pulmonary disease. There are research reports that cyclooxygenase-2 (COX-2) is associated with angiogenesis and lymphatic metastasis of NSCLC [18]. For example, celecoxib and nimesulide, as COX-2 inhibitors, can repress the progression of lung cancer [19, 20]. Only studies have shown that at the animal level, imrecoxib has a certain effect on lung adenocarcinoma cells by affecting the invasion and metastasis. However, there is no study on the effect of imrecoxib on pulmonary fibrosis. The purpose of this research is to study the effect of imrecoxib on pulmonary fibrosis and further study the mechanism.

2. Materials and Methods

2.1. Reagents. Imrecoxib was acquired from Biofount Biotechnology Co., Ltd. (Beijing, China). The antibodies were all obtained from Abcam.

2.2. Isolation of Primary Pulmonary Fibrosis Cells. Mice treated with paraquat were killed, and their lungs were taken out and cut into small pieces. Then, we digested the lung samples with pancreatin and isolated the pulmonary fibrosis cells. Then, we cultured them and screened the primary pulmonary fibrosis (PPF) cells which can proliferate stably in vitro.

2.3. Cell Culture. HFL1 cells were originated from Shanghai Sixin Biotechnology Co., Ltd. HFL1 cells and isolated primary pulmonary fibrosis cells were cultured in DMEM medium containing 10% fetal bovine serum (FBS) in a sterile incubator at 37°C and in a 5% CO₂ atmosphere. Human lung adenocarcinoma epithelial cell line A549 was cultured in RPMI-1640 medium.

2.4. Paraquat-Induced Lung Epithelial Cell A549 Assay. Lung adenocarcinoma epithelial cell A549 was inoculated in a 6-well plate and set to the control group and paraquat-induced group. The control group was cultured with normal RPMI-1640 medium, but the paraquat-induced group was cultured in RPMI-1640 medium with 20 μ mol/L paraquat. A549 cells were cultured for 5 days at 37°C and in a 5% CO₂ condition. On the third day, the medium of the control group was substituted with fresh medium, and the medium of the paraquat-induced group was substituted with fresh culture medium with 20 μ mol/L paraquat. A549 cells were cultured until the 5th day.

2.5. MTT Assay. HFL1 cells and PPF cells with imrecoxib or without imrecoxib were inoculated in 96-well plates and cultured certain time at 37°C and in a 5% CO₂ condition. After 24 or 48 hours, MTT reagent was put in, and then the cells were maintained for another 4-6 hours in dark. Thereafter, DMSO was added to react for 10 minutes, and the 96-well plates were shaken sufficiently to dissolve the crystal in the cells. Then, at the wavelength of 570 nm, the light absorption value of each well was measured by Microplate Reader. The cell viability of the control group and the imrecoxib-treated group was calculated by graphpad prism 5 software.

2.6. Wound-Healing Assays. Mark with a marker pen at the bottom of sterile 6-well plate. Then, HFL1 or PPF cells were added in a sterile 6-well plate and cultured overnight in an incubator. When the cell confluency was close to 100%, a straight line was scraped on the cell surface with a sterile pipette tips. Then, the cells were cultured in serum-free medium with or without imrecoxib for 24 hours, and the 6-well plate were taken at 0, 6, 12, and 24 hours, respectively, and the migration distance of cells in different groups at different time points was recorded.

2.7. Transwell Assay. Prepare Matrigel diluted with serum-free medium in advance, put it into the upper chamber of Transwell, and then add the cell suspension with imrecoxib or without imrecoxib after the coagulation of Matrigel. Serum-free medium was put into the upper part, and medium containing serum was put into the lower chamber. Culture the cells in the condition of 37°C, 5% CO₂ for 24 hours. Transwell chamber was removed and fixed with glutaraldehyde, and then the excess glutaraldehyde solution was washed by PBS. Thereafter, cells were stained by hexamethylparosaniline and wiped off the upper cells with cotton swabs. Finally, the invading cells were photographed under a microscope, and the number of invading cells in different group was recorded.

2.8. Real-Time Quantitative PCR Assay. Prepare treated or untreated cell samples and extract RNA with Trizol reagent. After the RNA concentration was determined, the PCR reaction system was formulated based on the protocol of the fluorescent RT-qPCR kit, and the PCR cycle was carried out. The sequences of snail are F: 5'-TCGGAAGCCTAACTACAGC GA', R: 5'-AGATAGCATTTGGCAGCGAG-3'. The primer sequences of β -actin are F: 5'-CCTGTACGCCAACACA

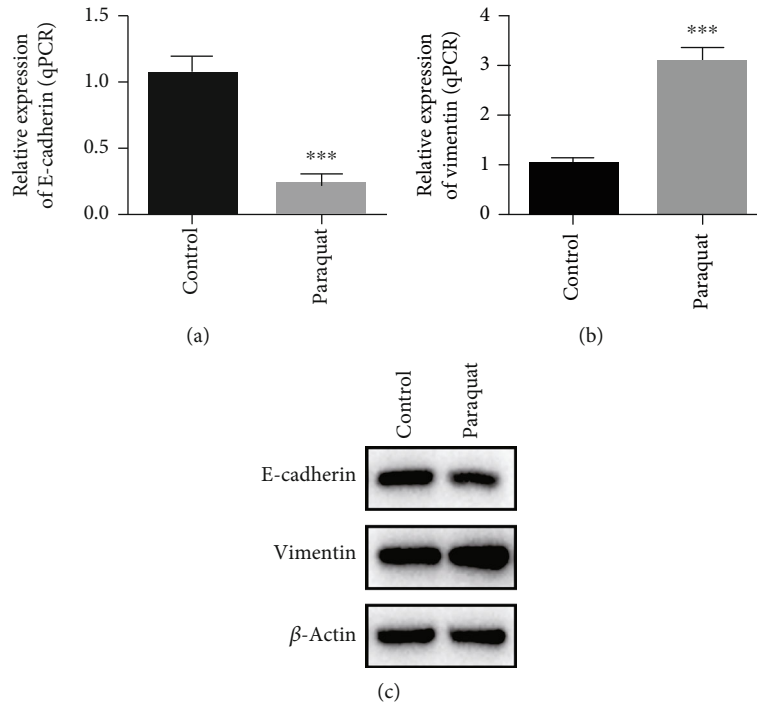


FIGURE 1: mRNA and protein expression of EMT markers in paraquat-induced lung epithelial cell A549. (a). The mRNA level of E-cadherin in A549 cells of the control group and paraquat-induced group was measured by the RT-qPCR method. (b). The mRNA level of vimentin in A549 cells of the control group and paraquat-induced group was measured by the RT-qPCR method. (c). The protein expression of E-cadherin and vimentin in A549 cells of the control group and paraquat-induced group was measured by western blot. *** $P < 0.001$.

GTGC-3', R: 5'-ATACTCCTGCTTGCTGATCC-3'; the primer sequences of E-cadherin are F: 5'-TGGACAGGAGGATTTTGGAG-3', R: 5'-ACCTGAGGCTTTGGATTCC T-3'. The primer sequences of vimentin are F: 5'-GAGA ACTTTGCCGTTGAAGC-3', R: 5'-CTCAATGTCAAGGG CCATCT-3'. The reaction conditions were as follows: 95°C predenaturation for 10 min, 95°C for 10 s, 60°C for 60 s, and 40 cycles were performed. Three replicate wells were set for each gene, and the gene expression was calculated using the $2^{-\Delta\Delta Ct}$ method. The experiment was repeated three times.

2.9. Western Blot Assay. The treated or untreated cells were placed on ice and treated with RIPA lysate for 20 minutes. And the cells were collected in EP tubes and centrifuged to obtain protein samples. All protein samples were adjusted to the same concentration, and the polyacrylamide gel electrophoresis step was performed to separate the proteins with different molecular weights. Then, the protein on the gel was transferred to the nitrocellulose (NC) film. The hydrophobic binding sites on the nitrocellulose film were blocked with 5% BSA. After blocking, the membrane was added with primary antibody NF- κ B (Abcam, 1: 1000), snail (Abcam, 1: 500), and GAPDH (Abcam, 1: 2000) and incubate at 4°C overnight. Then, added the corresponding horseradish peroxidase labeled secondary antibody to react in the dark for 1 hour. Finally, the protein was detected and photographed according to the operation of the BeyoECL Plus chemiluminescence kit (Beyotime).

2.10. Statistical Analysis. All the recorded data are analyzed by SPSS19.0 software, and all the data are recorded by means of mean \pm SD. A P value less than 0.05 is considered statistically significant.

3. Results

3.1. Paraquat Can Induce EMT in A549 Cells. To investigate the effect of paraquat on epithelial-mesenchymal transitions of A549 cells, we divided the cells of A549 into the control group and paraquat-induced group. The RNA of two groups of A549 cells was extracted for the RT-qPCR assay, and the content of E-cadherin and vimentin in the two groups of A549 cells was determined by western blot. The data demonstrated that the expression of E-cadherin in the paraquat-induced group was much lower, but on the contrary, the expression of vimentin was significantly higher (Figures 1(a)–1(c)). These results indicated that paraquat can inhibit the expression of E-cadherin in A549 cells, while promote the vimentin expression, that is paraquat can induce the epithelial-mesenchymal transitions of A549 cells.

3.2. Paraquat Can Induced the NF- κ B/Snail Signaling Pathway in Lung Epithelial Cell A549. NF- κ B/snail is a very important signaling pathway in the EMT process. In order to study the efficacy of paraquat on the NF- κ B/snail signaling pathway in A549 cells, we extracted mRNA and protein from the control group and the paraquat-induced group of A549 cells, respectively, and performed RT-qPCR and western blot

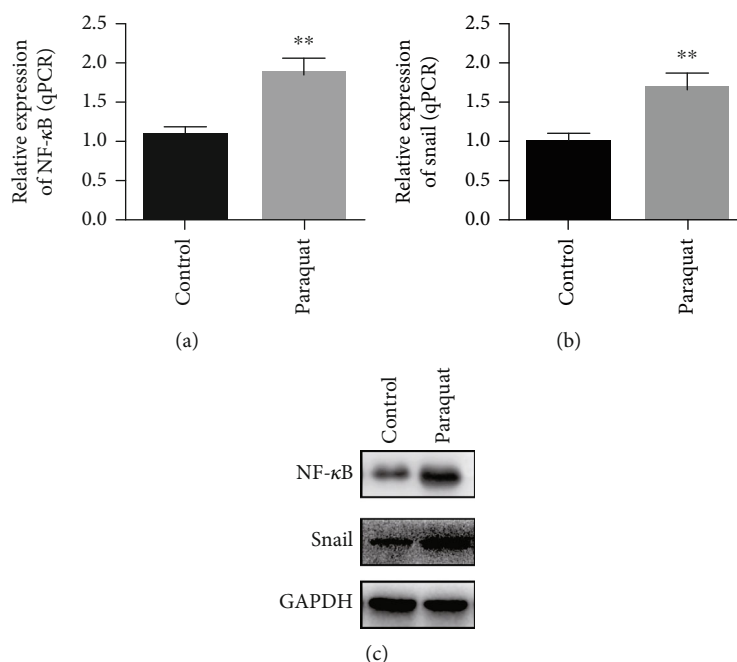


FIGURE 2: mRNA and protein expression of NF- κ B and snail in paraquat-induced lung epithelial cell A549. (a). The mRNA level of NF- κ B in A549 cells of the control group and paraquat-induced group was measured by the RT-qPCR method. (b). The mRNA level of snail in A549 cells of the control group and paraquat induced group was measured by the RT-qPCR method. (c). The protein expression of NF- κ B and snail in A549 cells of the control group and paraquat-induced group was measured by the western blot method. ** $P < 0.01$.

assay to study the effects of paraquat on mRNA and protein of NF- κ B and snail in A549 cells. Our data showed that the level of mRNA of NF- κ B and snail in A549 cells of the paraquat-treated group was significantly higher than that of the control group (Figures 2(a) and 2(b)). Furthermore, the results of western blot also demonstrated that the protein expression is a consistent trend with mRNA (Figure 2(c)). The above results all confirmed that paraquat could induce the NF- κ B/snail signaling pathway in A549 cells.

3.3. Imrecoxib Can Inhibit the Proliferation and Migration of PPF Cells. Primary pulmonary fibrosis cells (PPF cells) are a kind of cells with mesenchymal phenotype, which are extracted from the lungs of paraquat-treated mice. PPF cells were set into two groups: control group and imrecoxib-treated group. According to the results of MTT, the cell viability of PPF cells treated with imrecoxib was lower than that without imrecoxib (Figure 3(a)). The results showed that imrecoxib could inhibit the proliferation of PPF cells. Besides, the findings of cell wound-healing assays and transwell revealed that the invasion and migration ability of PPF cells in the imrecoxib-treated group was much weaker when compared with THE control group. In other words, imrecoxib can significantly inhibit the invasion and migration ability of PPF cells (Figure 3(b)). The above experiments showed that imrecoxib can suppressed the proliferation, migration, and invasion functions of PPF cells.

3.4. Imrecoxib Can Inhibit the Proliferation and Migration of HFL1 Cells. HFL1 cell is a human fetal pulmonary fibrosis cell with mesenchymal phenotype. Similar to the result of imre-

coxib on PPF cells, MTT test showed that imrecoxib could inhibit the proliferation of HFL1 cells (Figure 4(a)), and the wound-healing test showed that imrecoxib could inhibit the migration of HFL1 cells (Figure 4(b)). These results indicated that the drug imrecoxib can also inhibit the proliferation and migration HFL1 cells.

3.5. Imrecoxib Can Inhibit Paraquat-Induced EMT in Lung Epithelial Cell A549. We further studied the effect of imrecoxib on EMT of A549 cells induced by paraquat. We set three groups: control group (untreated A549 cells), imrecoxib-treated group (paraquat-induced A549 cells were treated with imrecoxib), paraquat-induced group (paraquat-induced A549 cells). We measured the content of E-cadherin and vimentin by western blot in three groups, respectively. The results displayed that the E-cadherin level of A549 cells in the control group was the highest, followed by that in the imrecoxib-treated group and the lowest in the paraquat-induced group. In contrast to the E-cadherin expression, the vimentin expression of A549 cells in the control group was the lowest, the imrecoxib-treated group was the second highest, and the vimentin expression of A549 cells in the paraquat-induced group was the highest (Figure 5(a)). The above experimental results show that imrecoxib can affect the expression of EMT-related proteins; specifically, imrecoxib can inhibit paraquat-induced EMT in lung epithelial cell A549.

3.6. Imrecoxib Can Inhibit Paraquat-Induced the NF- κ B/Snail Signaling Pathway in Lung Epithelial Cell A549. Our experiments have confirmed that paraquat can induce

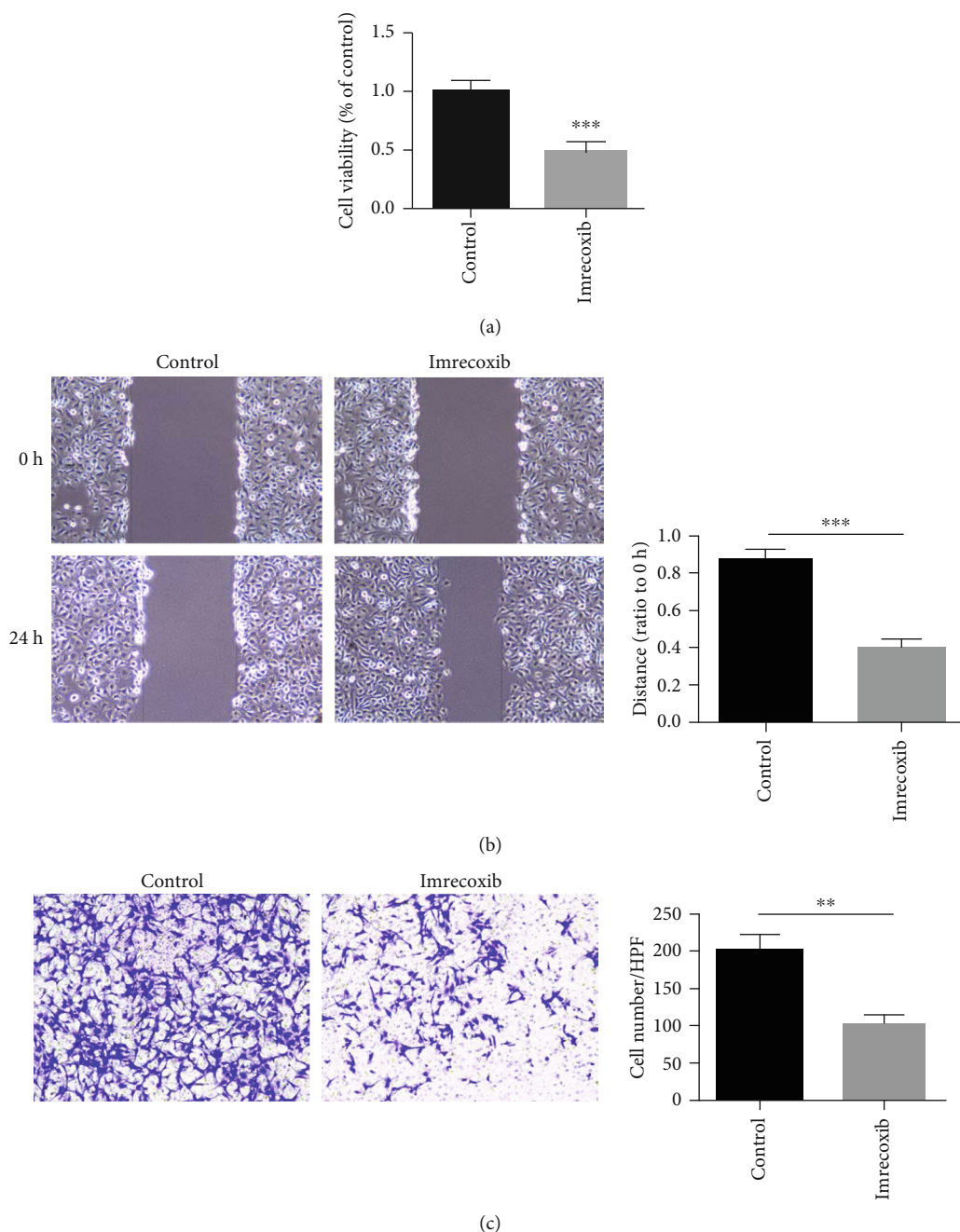


FIGURE 3: Effects of imrecoxib on the proliferation and migration and of PPF cells in vitro. (a) The effect of imrecoxib on the proliferation ability of PPF cells was measured by the MTT assay. (b) The effect of imrecoxib on the migration ability of PPF cells was measured by the scratch test. (c) The effect of imrecoxib on the invasion ability of PPF cells was measured by transwell. ** $P < 0.01$; *** $P < 0.001$.

the NF- κ B/snail signal pathway of A549 cells. What we need to study next is the relationship between imrecoxib and the signal pathway. We divided A549 cells into three groups: control group (untreated A549 cells), imrecoxib-treated group (paraquat-induced A549 cells were treated with imrecoxib), paraquat-induced group (paraquat-induced A549 cells). First, we measured the miRNA transcription level of NF- κ B and snail of A549 cells in three groups by RT-qPCR. The results displayed the mRNA level of NF- κ B and snail in paraquat-induced group was the highest, that of the mRNA level of the control group was the lowest, and that

of the mRNA level in the imrecoxib-treated group was between that of the control group and paraquat-induced group. The above results confirmed that imrecoxib could suppress the mRNA level of NF- κ B and snail in paraquat-induced A549 cells (Figure 6(a)). We further studied the effect of imrecoxib on NF- κ B and snail protein in paraquat-induced A549 cells by western blot, which showed that the two protein content was the highest in the paraquat-induced A549 cell group and the lowest in A549 cells of the control group. The two protein levels in the imrecoxib-treated group were between the control group and

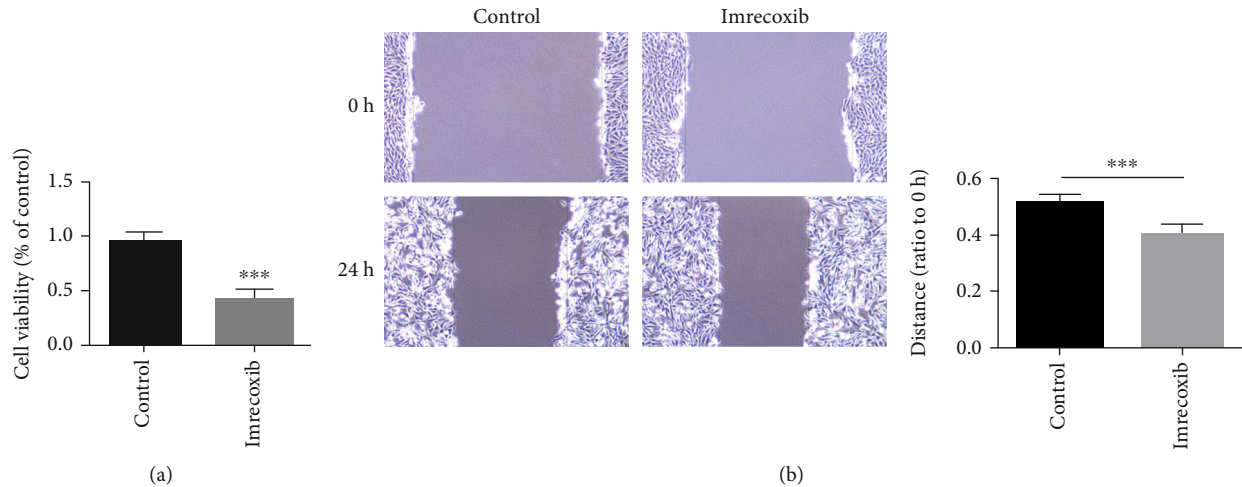


FIGURE 4: Effects of imrecoxib on the proliferation and migration of HFL1 cells in vitro. (a) The effect of imrecoxib on the proliferation ability of HFL1 cells was measured by the MTT assay. (b) The effect of imrecoxib on the migration ability of HFL1 cells was measured by the scratch test. *** $P < 0.001$.

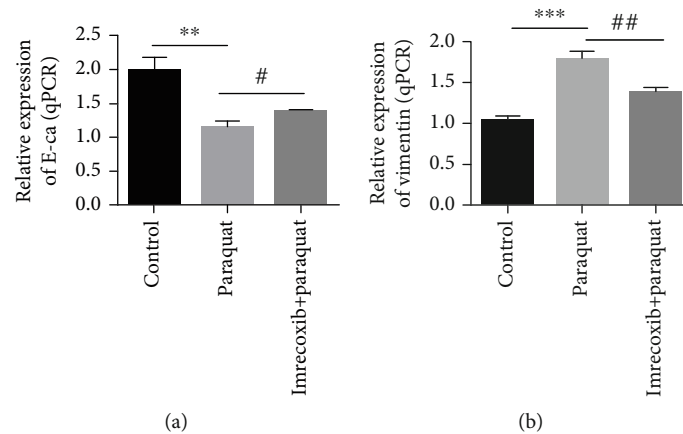


FIGURE 5: The effect of imrecoxib on the mRNA expression of E-cadherin and vimentin in paraquat-induced A549 cells was assessed by RT-qPCR. # $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

paraquat-induced group. Our data showed that imrecoxib could affect the expression of NF- κ B and snail protein in paraquat-induced A549 cells; specifically, imrecoxib could inhibit the expression of NF- κ B and snail protein in paraquat-induced A549 cells (Figure 6(b)). The above experimental results confirmed that imrecoxib can not only inhibit the transcription levels of NF- κ B and snail in paraquat-induced A549 cells but also can inhibit the protein expression levels.

4. Discussion

Imrecoxib is a kind of drug that can moderately inhibit COX-2 to play an anti-inflammatory role. It has a good inhibition of inflammation and pain, but also reduces the chance of gastrointestinal tract stimulation and cardiovascular damage [16]. At present, only a few studies have confirmed that imrecoxib can repress the invasion and metastasis of NSCLC, but the mechanism is not clear. It is only speculated that the possible mechanism is to inhibit the inactivation of PTEN pro-

tein, interfere with PI3K/Akt signal transduction, block cell cycle in the G1 phase, or promote apoptosis [21]. But at present, there is almost no research on imrecoxib in the pulmonary fibrosis field. For the sake of research on the effect of imrecoxib on pulmonary fibrosis cells, we carried out MTT and wound-healing assays. The current experimental results demonstrated that imrecoxib can inhibit the proliferation, migration, and invasion of PPF and HFL1 cells.

The main cause of paraquat poisoning is that it can cause pulmonary fibrosis. There are many studies on the mechanism of paraquat inducing pulmonary fibrosis. EMT plays an important role in paraquat-induced pulmonary fibrosis [22]. It has been confirmed that paraquat can activate the Wnt/ β -Catenin signal pathway and then induce EMT of type II alveolar epithelial cells, which is mainly manifested as the decrease of the E-cadherin expression and the increase of the vimentin expression [23]. More and more evidence showed that cytokines play an essential role in EMT. For example, transforming growth factor β (TGF- β) is considered to be one of the main cytokines in paraquat-induced

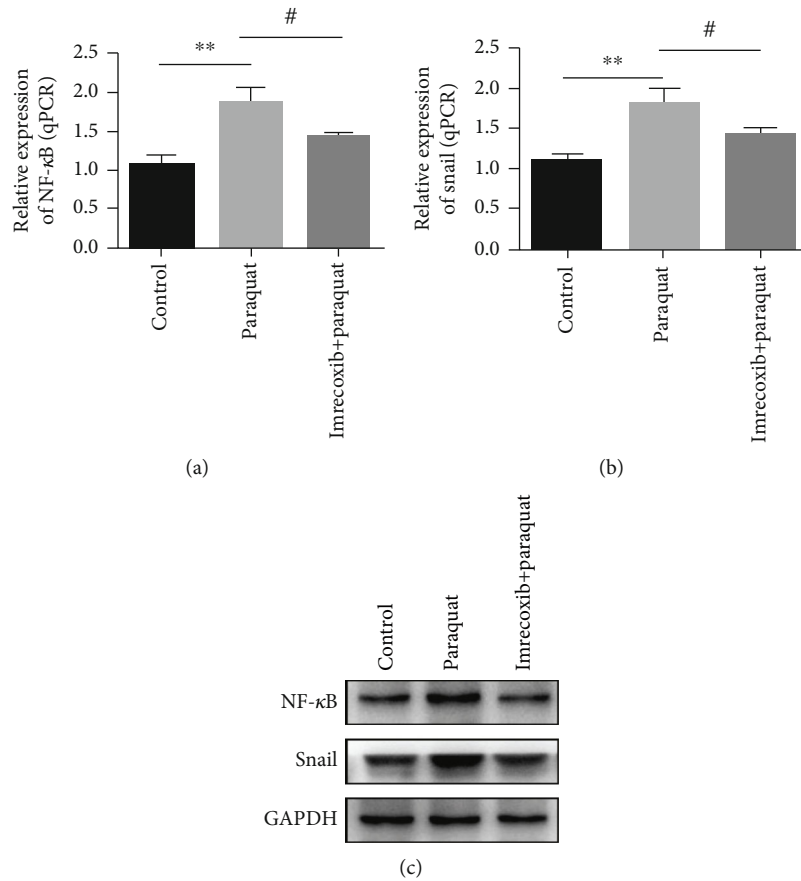


FIGURE 6: The effect of imrecoxib on NF- κ B and snail mRNA transcription and protein expression in paraquat-induced A549 cells. (a, b) The effect of imrecoxib on the mRNA level of NF- κ B and snail in paraquat-induced A549 cells was determined by the RT-qPCR method. (c) The effect of imrecoxib on the protein expression of NF- κ B and snail in paraquat-induced A549 cells was measured by the western blot method. # $P < 0.05$; ** $P < 0.01$.

pulmonary fibrosis, and it is also a research hotspot of scholars at home and abroad. Paraquat can induce EMT and active TGF- β /Smad signal pathway [24, 25]. Our results showed that paraquat can activate EMT by inhibiting E-cadherin mRNA and protein expression of A549 cells and promoting vimentin mRNA and protein expression. And our experimental results indicated that imrecoxib can reduce the inhibition on E-cadherin and promotion on vimentin in paraquat-induced A549 cells and then inhibit the EMT activated by paraquat, so as to treat paraquat-induced pulmonary fibrosis.

Moreover, studies have shown that paraquat can activate inflammatory related factors including TNF α , NF- κ B, interleukin 1 β , and IL-6, which in turn contribute to the development of pulmonary fibrosis [26]. NF- κ B is a crucial transcription regulator, which plays an extremely important function in the process of inflammation, immunity, and apoptosis. In addition, NF- κ B is also involved in the occurrence and development of fibrosis by regulating transcription factors related to fiber growth, such as PDGF and TGF- β 1. The expression product of the snail gene is a transcription factor that plays an important effect in the process of EMT. We evaluated the expression of NF- κ B and snail mRNA and protein in paraquat-induced A549 cells. The results

showed that paraquat could activate the transcription of NF- κ B and snail mRNA and protein in A549 cells. Moreover, we further evaluated the effect of imrecoxib on NF- κ B and snail in paraquat-induced A549 cells. The results proved that the level of mRNA transcription and protein expression of NF- κ B and snail in A549 cells activated by paraquat was alleviated after treatment with imrecoxib. These results suggest that imrecoxib can attenuate the activation of NF- κ B and snail induced by paraquat and thus play a role in the treatment of pulmonary fibrosis. The NF- κ B pathway activates the snail expression through transcriptional and post-translational mechanisms. Moreover, NF- κ B can be bind to the promoter of snail, further to increase the transcription of the snail. Therefore, imrecoxib can inhibit paraquat-induced pulmonary fibrosis by inhibiting the NF- κ B/snail signal pathway.

To sum up, we first established the A549 cell model induced by paraquat and confirmed that paraquat can change the transcription and protein expression ability of E-cadherin and vimentin in A549 cells at the level of gene transcription and protein expression, indicating that paraquat can activate the EMT of A549 cells. Then, by measuring the mRNA and protein expression of NF- κ B and snail in A549 cells induced by paraquat, we proved that paraquat

activated the NF- κ B/snail signal pathway in A549 cells. Next, we assessed the influence of imrecoxib on PPF and HFL1 cells. The results showed that imrecoxib could suppress the proliferation and migration of PPF and HFL1 cells. Finally, we found that imrecoxib can activate the EMT in A549 cells induced by paraquat, accompanied by the activation of the NF- κ B/snail signal pathway. Our study shows that imrecoxib are expected to be an effective drug for paraquat-induced pulmonary fibrosis. Therefore, our research not only provides a new idea to treat paraquat-induced pulmonary fibrosis but also intends the treatment methods of imrecoxib. However, the effect or mechanism of imrecoxib in treating pulmonary fibrosis needs further study.

5. Conclusion

Our results confirm that imrecoxib can inhibit EMT of paraquat-induced A549 cells and alleviate paraquat-induced pulmonary fibrosis through the NF- κ B/snail signal pathway.

Data Availability

All the data could be provided if qualified authors required it.

Conflicts of Interest

The author declares that they have no conflicts of interest.

References

- [1] R. P. Charbeneau and M. Peters-Golden, "Eicosanoids: mediators and therapeutic targets in fibrotic lung disease," *Clinical Science*, vol. 108, no. 6, pp. 479–491, 2005.
- [2] Y. Ji, T. Wang, Z. F. Wei et al., "Paeoniflorin, the main active constituent of *Paeonia lactiflora* roots, attenuates bleomycin-induced pulmonary fibrosis in mice by suppressing the synthesis of type I collagen," *Journal of Ethnopharmacology*, vol. 149, no. 3, pp. 825–832, 2013.
- [3] B. Ley, H. R. Collard, and T. E. King Jr., "Clinical course and prediction of survival in idiopathic pulmonary fibrosis," *American Journal of Respiratory and Critical Care Medicine*, vol. 183, no. 4, pp. 431–440, 2011.
- [4] M. W. Foster, L. D. Morrison, J. L. Todd et al., "Quantitative proteomics of bronchoalveolar lavage fluid in idiopathic pulmonary fibrosis," *Journal of Proteome Research*, vol. 14, no. 2, pp. 1238–1249, 2015.
- [5] K. K. Kim, M. C. Kugler, P. J. Wolters et al., "Alveolar epithelial cell mesenchymal transition develops *in vivo* during pulmonary fibrosis and is regulated by the extracellular matrix," *Proceedings of the National Academy of Sciences*, vol. 103, no. 35, pp. 13180–13185, 2006.
- [6] A. Azuma, T. Nukiwa, E. Tsuboi et al., "Double-blind, placebo-controlled trial of pirfenidone in patients with idiopathic pulmonary fibrosis," *American Journal of Respiratory & Critical Care Medicine*, vol. 171, no. 9, pp. 1040–1047, 2005.
- [7] S. Ikeda, A. Sekine, T. Baba et al., "Hepatotoxicity of nintedanib in patients with idiopathic pulmonary fibrosis: A single-center experience," *Respiratory Investigation*, vol. 55, no. 1, pp. 51–54, 2017.
- [8] W. A. Dik, M. A. Versnel, B. A. Naber, D. J. Janssen, A. H. van Kaam, and L. J. I. Zimmermann, "Dexamethasone treatment does not inhibit fibroproliferation in chronic lung disease of prematurity," *European Respiratory Journal*, vol. 21, no. 5, pp. 842–847, 2003.
- [9] F. Chen, L. Gong, L. Zhang et al., "Short courses of low dose dexamethasone delay bleomycin-induced lung fibrosis in rats," *European Journal of Pharmacology*, vol. 536, no. 3, pp. 287–295, 2006.
- [10] Y. Han, L. Han, M. Dong et al., "Comparison of a loading dose of dexmedetomidine combined with propofol or sevoflurane for hemodynamic changes during anesthesia maintenance: a prospective, randomized, double-blind, controlled clinical trial," *Bmc Anesthesiology*, vol. 18, no. 1, p. 12, 2018.
- [11] H. Corvol, F. Flamein, R. Epaud, A. Clement, and L. Guillot, "Lung alveolar epithelium and interstitial lung disease," *The International Journal of Biochemistry & Cell Biology*, vol. 41, no. 8–9, pp. 1643–1651, 2009.
- [12] J. M. Lee, S. Dedhar, R. Kalluri, and E. W. Thompson, "The epithelial-mesenchymal transition: new insights in signaling, development, and disease," *Journal of Cell Biology*, vol. 172, no. 7, pp. 973–981, 2006.
- [13] X.-F. Hua, X.-H. Li, M.-M. Li et al., "Doxycycline attenuates paraquat-induced pulmonary fibrosis by downregulating the TGF- β signaling pathway," *Journal of Thoracic Disease*, vol. 9, no. 11, pp. 4376–4386, 2017.
- [14] T. Chen, H. Nie, X. Gao et al., "Epithelial-mesenchymal transition involved in pulmonary fibrosis induced by multi-walled carbon nanotubes via TGF-beta/Smad signaling pathway," *Toxicology Letters*, vol. 226, no. 2, pp. 150–162, 2014.
- [15] X. H. Li, T. Xiao, J. H. Yang et al., "Parthenolide attenuated bleomycin-induced pulmonary fibrosis via the NF- κ B/Snail signaling pathway," *Respiratory Research*, vol. 19, no. 1, p. 111, 2018.
- [16] X. H. Chen, J. Y. Bai, F. Shen, A. P. Bai, Z. R. Guo, and G. F. Cheng, "Imrecoxib: a novel and selective cyclooxygenase 2 inhibitor with anti-inflammatory effect," *Acta Pharmacologica Sinica*, vol. 25, no. 7, pp. 927–931, 2004.
- [17] H. Jian-lin, G. Jie-ruo, P. Yun-feng et al., "A Multicenter, double-blind and randomized controlled phase II trial of imrecoxib in treatment of knee osteoarthritis," *Chinese Pharmaceutical Journal*, vol. 22, 2011.
- [18] H. Liu, Y. Yang, J. Xiao et al., "COX-2-mediated regulation of VEGF-C in association with lymphangiogenesis and lymph node metastasis in lung Cancer," *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, vol. 293, no. 11, pp. 1838–1846, 2010.
- [19] L. Liu, F. Ren, M. Fan et al., "Interferon- γ and celecoxib inhibit lung-tumor growth through modulating M2/M1 macrophage ratio in the tumor microenvironment," *Drug Design, Development and Therapy*, vol. 8, pp. 1527–1538, 2014.
- [20] B. M. Kim, J. Won, K. A. Maeng, Y. S. Han, Y. Yun, and S. H. Hong, "Nimesulide, a selective COX-2 inhibitor, acts synergistically with ionizing radiation against A549 human lung cancer cells through the activation of caspase-8 and caspase-3," *International Journal of Oncology*, vol. 34, pp. 1467–1473, 2009.
- [21] F. Yun, Y. Jia, X. Li et al., "Clinicopathological significance of PTEN and PI3K/AKT signal transduction pathway in non-small cell lung cancer," *International Journal of Clinical and Experimental Pathology*, vol. 6, pp. 2112–2120, 2013.
- [22] A. Yamada, T. Aki, K. Unuma, T. Funakoshi, and K. Uemura, "Paraquat induces epithelial-mesenchymal transition-like

cellular response resulting in fibrogenesis and the prevention of apoptosis in human pulmonary epithelial cells.” *PloS ONE*, R. W. Dettman, Ed., vol. 10, no. 3, article e0120192, 2015.

- [23] S. D. Su, S. G. Cong, Y. K. Bi, and D. D. Gao, “Paraquat promotes the epithelial-mesenchymal transition in alveolar epithelial cells through regulating the Wnt/ β -catenin signal pathway,” *European Review for Medical and Pharmacological Sciences*, vol. 22, no. 3, pp. 802–809, 2018.
- [24] L. Xie, D. Zhou, J. Xiong, J. You, Y. Zeng, and L. Peng, “Paraquat induce pulmonary epithelial–mesenchymal transition through transforming growth factor- β 1-dependent mechanism,” *Experimental and Toxicologic Pathology*, vol. 68, no. 1, pp. 69–76, 2016.
- [25] Y. Y. Han, P. Shen, and W. X. Chang, “Involvement of epithelial-to-mesenchymal transition and associated transforming growth factor- β /Smad signaling in paraquat-induced pulmonary fibrosis,” *Molecular Medicine Reports*, vol. 12, no. 6, pp. 7979–7984, 2015.
- [26] Z. Khalighi, A. Rahmani, J. Cheraghi et al., “Perfluorocarbon attenuates inflammatory cytokines, oxidative stress and histopathologic changes in paraquat-induced acute lung injury in rats,” *Environmental Toxicology and Pharmacology*, vol. 42, pp. 9–15, 2016.

Research Article

Effects of Bronchoalveolar Lavage with Ambroxol Hydrochloride on Treating Pulmonary Infection in Patients with Cerebral Infarction and on Serum Proinflammatory Cytokines, MDA and SOD

Fanhua Meng,¹ Jing Cheng,² Peng Sang,³ and Jianhui Wang³ 

¹Stroke Unit, The Affiliated Hospital of Beihua University, Jilin 132011, China

²Respiratory Department, The Affiliated Hospital of Beihua University, Jilin 132011, China

³Department of Rehabilitation Medicine, The Affiliated Hospital of Beihua University, Jilin 132011, China

Correspondence should be addressed to Jianhui Wang; jiangfanliangpg@163.com

Received 21 June 2020; Accepted 23 July 2020; Published 9 October 2020

Guest Editor: Tao Huang

Copyright © 2020 Fanhua Meng 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.

Objective. This paper was aimed at investigating the effects of bronchoalveolar lavage (BAL) with ambroxol hydrochloride (AH) on treating pulmonary infection and on serum proinflammatory cytokines and oxidative stress responses in patients with cerebral infarction (CI). **Methods.** One hundred and two patients with cerebral infarction complicated with pulmonary infection (CIPI) who were treated in our hospital were enrolled as research objects, divided into an observation group (52 cases; AH combined with BAL) and a control group (50 cases; single AH) based on therapeutic schemes. They were compared in terms of the therapeutic effect and pre- and posttreatment serum inflammatory cytokines, pulmonary function, and serum indices of oxidative stress. Their adverse reactions during treatment were also recorded and compared. **Results.** The therapeutic effect in the observation group was remarkably better than that in the control group ($P < 0.05$). After treatment, the serum inflammatory cytokines, pulmonary function, and serum indices of oxidative stress were remarkably improved in the two groups ($P < 0.05$), but the improvement was remarkably better in the observation group ($P < 0.05$). The differences were not significant in intratreatment adverse reactions between the two groups ($P > 0.05$). **Conclusion.** For CIPI patients, BAL with AH has a better therapeutic effect and higher safety and can control the patients' systemic inflammatory responses and oxidative stress responses, so it is worthy of further promotion in clinical practice.

1. Introduction

As a disease with a high incidence among the elderly, cerebral infarction (CI) is affected by many factors, with its clinical symptoms mostly accompanied by dysphagia and pharyngeal secretions that cannot be excluded [1, 2]. Many CI patients have low immunity, prone to be complicated with pulmonary infection [3]. Easy to cause further damages to the body, the complicated diseases affect the therapeutic effect on and the rehabilitation speed of the patients [4]. Moreover, due to the abuse of antibiotics and the emergence of drug-resistant strains, it is more difficult to treat CI complicated with pulmonary infection (CIPI) [5]. Therefore, it

is of great clinical significance to seek effective therapeutic methods for CIPI patients.

At present, anti-infection treatment is mainly used for pulmonary infection, and the airway of patients should be kept unobstructed in order to improve pulmonary function [6]. CIPI patients experience expectoration disorders after pulmonary infection, and the viscous sputum in the lung is not easy to cough out, which easily results in pulmonary retention and even death in serious cases. Therefore, the rapid and effective control of pulmonary infection is crucial to treat CI patients [7]. Ambroxol hydrochloride (AH) is an expectorant, which is effective in dissolving viscous phlegm and lubricating the respiratory tract [8]. Having been widely

used in the clinical treatment of pulmonary infection at present, bronchoalveolar lavage (BAL) is a therapeutic method to improve respiratory function and infection control via directly infusing drugs into diseased regions of pulmonary segments through a bronchoscope [9]. According to a previous study, AH combined with BAL has a better curative effect on patients with severe pneumonia and can improve their pulmonary function [10].

For seeking active and effective therapeutic methods for CIPI patients, we have explored the therapeutic effect of AH combined with BAL on the patients and obtained positive clinical results.

2. Materials and Methods

2.1. Clinical Data. A prospective analysis was made on 102 CIPI patients admitted to our hospital from February 2016 to October 2018, with an average age of 55.21 ± 2.62 years. Fifty cases in the control group received single AH, while 52 cases in the observation group received AH combined with BAL. All treatments were performed on the basis of conventional anti-infection treatment. Inclusion criteria include patients who were confirmed with CI by head CT or MRI and confirmed with pulmonary infection based on clinical manifestations, signs, laboratory examinations, and chest X-rays. Exclusion criteria include patients who were allergic to AH; patients who had used glucocorticoids for a long time; patients with severe hepatic and renal insufficiency; patients complicated with other malignant tumors; and patients who did not cooperate in treatment. All patients have consented to participation in the experiment and signed the informed consent. The Hospital Ethics Committee has agreed with this experiment.

2.2. Therapeutic Methods. All patients received conventional oxygen inhalation and anti-infection treatment. On this basis, those in the control group were intravenously dripped with 30 mg of AH (Sinopharm Group Guorui Pharmaceutical Co., Ltd., SFDA Approval Number: H20143385) twice per day. On the basis of the control group, those in the observation group were given BAL by an electronic bronchoscope once/day. Specific steps were as follows: before the operation, 2% lidocaine was atomized and inhaled to perform local anesthesia on the throat. After the flexible bronchofiberscope was connected with a negative pressure aspirator, high-concentration oxygen was inhaled for approximately 5 min under ECG monitoring. After blood oxygen saturation was $\geq 95\%$, the electronic bronchoscope was inserted through the nose, mouth, or artificial airway, with secretions in the airway sucked. Then, 0.9% sodium chloride solution (100 mL) + ambroxol hydrochloride injection (90 mg) was prepared for BAL, 10-20 mL once. During the operation, the negative pressure was controlled at ≤ 100 mmHg (1 mmHg = 0.133 kPa), and the actions should be gentle and rapid. Additionally, parameters of the ECG monitoring should be closely observed. The operation should be immediately stopped with the bronchoscope exited and oxygen inhaled, if the blood oxygen saturation was $< 85\%$. The operation should be continued if the blood oxygen saturation was $\geq 95\%$. Repeated lavage was

carried out on each diseased pulmonary segment until the bronchoalveolar lavage fluid was clean. The lavage was conducted for ≤ 3 times, and the lobes on each side were lavaged with approximately 50 mL of the fluid. The patients in both groups were consecutively treated for 1 week.

2.3. Outcome Measures. (1) After treatment, the therapeutic effects on the patients were evaluated, which were divided into cured (the patients' clinical signs, laboratory examinations, and pathogen examinations showed recovery), markedly effective (the clinical symptoms were obviously relieved, but laboratory or pathogen examinations showed incomplete recovery), effective (the clinical symptoms were relieved but the relief was not very obvious), and ineffective (the clinical symptoms were not obviously relieved). Total effective rate = (number of cured cases + number of markedly effective cases) / total number of cases $\times 100\%$. (2) The patients' pulmonary function before and after treatment was assessed and compared between the two groups. MasterScreen PFT System was used to evaluate pulmonary function indices, which included forced vital capacity (FVC), forced expiratory volume in 1 s (FEV1), and FEV1 to FVC (FEV1/FVC). (3) ELISA was used to detect and compare levels of TNF- α , IL-8, and IL-6 before and after treatment between the two groups. (4) ELISA was also used to detect contents of serum indices of oxidative stress malondialdehyde (MDA) and superoxide dismutase (SOD) before and after treatment between the two groups. (5) The adverse reactions of the patients during treatment were recorded and compared, including increased heart rate, small amount of hemoptysis, decreased blood oxygen, and decreased heart rate

2.4. Statistical Methods. In this study, SPSS19.0 was applied to analyze the experimental data statistically. We use the chi-squared test to count data. Measurement data were expressed by mean \pm standard deviation, and *t*-test was applied for the comparison between two groups and paired *t*-test for the comparison between before and after treatment. GraphPad Prism 6 was used for plotting figures in this experiment. *P* value < 0.05 was recognized as statistically significant.

3. Results

3.1. General Information. The differences were not significant in gender, age, body mass index (BMI), and history of smoking between the observation and control groups ($P > 0.05$) (see Table 1).

3.1.1. Comparison of Therapeutic Effects. After treatment, the therapeutic effects were compared between the observation and control groups. There were 22 cured cases, 20 markedly effective cases, 7 effective cases, and 3 ineffective cases in the observation group, with a total effective rate of 80.77%. There were 15 cured cases, 12 markedly effective cases, 15 effective cases, and 8 ineffective cases in the control group, with a total effective rate of 56.86%. The effective rate of treatment in the observation group was remarkably higher than that in the control group ($P < 0.05$) (see Table 2).

TABLE 1: General information.

Factors	Observation group (<i>n</i> = 52)	Control group (<i>n</i> = 50)	<i>t</i> / <i>X</i> ²	<i>P</i>
Gender			0.007	0.933
Male	36 (69.23)	35 (70.00)		
Female	16 (30.77)	15 (30.00)		
Age (years)			0.197	0.657
≤56	22 (42.31)	19 (38.00)		
>56	30 (57.69)	31 (62.00)		
BMI (kg/m ²)			0.001	0.988
≤23	28 (53.85)	27 (54.00)		
<23	24 (46.15)	23 (46.00)		
APACHEII score	22.15 ± 3.04	22.21 ± 3.08	0.099	0.921
NIHSS score	36.59 ± 3.86	36.86 ± 3.77	0.357	0.722
History of smoking			0.002	0.981
Yes	29 (55.77)	28 (56.00)		
No	23 (44.23)	22 (44.00)		
Family history of CI			0.009	0.925
Yes	15 (28.85)	14 (28.00)		
No	37 (71.15)	36 (72.00)		

3.2. Comparison of Pulmonary Function Indices before and after Treatment. Before treatment, the differences were not significant in FVC, FEV1, and FEV1/FVC between the observation and control groups ($P > 0.05$). After treatment, the three indices in the two groups were remarkably improved ($P < 0.05$), but the improvement was remarkably better in the observation group ($P < 0.05$) (see Figure 1).

3.3. Comparison of Serum Inflammatory Cytokines before and after Treatment. Before treatment, the difference was not significant in the expression of serum TNF- α , IL-8, and IL-6 between the observation and control groups ($P > 0.05$). After treatment, the expression in the two groups was remarkably improved ($P < 0.05$), but the improvement was remarkably better in the observation group ($P < 0.05$) (see Figure 2).

3.4. Comparison of Indices of Oxidative Stress before and after Treatment. We compared contents of serum MDA and SOD before and after treatment between the observation and control groups. Before treatment, the differences were not statistically significant in the contents between the two groups ($P > 0.05$). At one week after treatment, MDA content reduced but SOD content rose in the two groups; MDA content was lower but SOD content was higher in the observation group ($P < 0.05$) (see Figure 3).

3.5. Comparison of Adverse Reactions. The adverse reactions of the patients during treatment were recorded and compared. In the observation group, the number of patients suffering from increased heart rate, small amount of hemoptysis, decreased blood oxygen, and decreased heart rate was 2, 2, 4, and 2, respectively, with the incidence of adverse reac-

TABLE 2: Comparison of therapeutic effects.

Efficacy	Observation group (<i>n</i> = 52)	Control group (<i>n</i> = 50)	<i>X</i> ²	<i>P</i>
Cured	22 (42.31)	15 (30.00)	—	—
Markedly effective	20 (38.46)	12 (24.00)	—	—
Effective	7 (13.46)	15 (30.00)	—	—
Ineffective	3 (5.77)	8 (16.00)	—	—
Total effective rate	42 (80.77)	27 (54.00)	8.346	0.004

tions of 18.51%. In the control group, the number of patients suffering from the four adverse reactions was 2, 3, 1, and 3, respectively, with the incidence of 17.64%. The difference was not significant in the incidence of adverse reactions between the two groups ($P > 0.05$) (see Table 3).

4. Discussion

In recent years, with the improvement of social environment and living standards, the incidence of CI has been rising, and many CI patients usually suffer from some complications among which pulmonary infection is one of the most common ones [1, 11]. CI patients have poor respiratory and expectoration abilities, which makes their treatment more difficult when pulmonary infection occurs, so more active therapeutic methods need to be sought urgently in clinical practice [12].

In our study, the therapeutic effects of AH combined with BAL on CIPI patients were deeply analyzed. In recent years, BAL has been gradually applied to the clinical treatment of pulmonary infection, and a positive therapeutic effect of it has been achieved [13]. AH as a mucolytic can dissolve viscous phlegm and promote the secretion of pulmonary surfactant, thereby promoting sputum excretion and relieving dyspnea symptoms [14]. In this study, the combined application of AH and BAL effectively increased the effective rate of treatment and relieved the clinical symptoms of the patients. According to previous studies, the application of high-dose AH significantly improves the therapeutic effect on pulmonary infection without obvious adverse reactions [15]. BAL with a bronchoscope can directly enter the diseased region and remove inflammatory secretions in the region, thus reducing airway resistance and respiratory consumption, increasing the blood oxygenation level of the pulmonary alveolus, and improving pulmonary function [16]. In our study, AH can directly reach the diseased region through BAL, further improving the efficiency of treatment. As reported by a previous study, BAL relieves pulmonary atelectasis caused by inflammatory responses and phlegm and blood stasis obstruction in patients with pulmonary infection in a short time, thus improving pulmonary gas transfer and ventilation functions [17]. This is also consistent with our research results.

Patients with pulmonary infection suffer from more serious systemic inflammatory responses [18]. In our study, before treatment, patients in the two groups had relatively

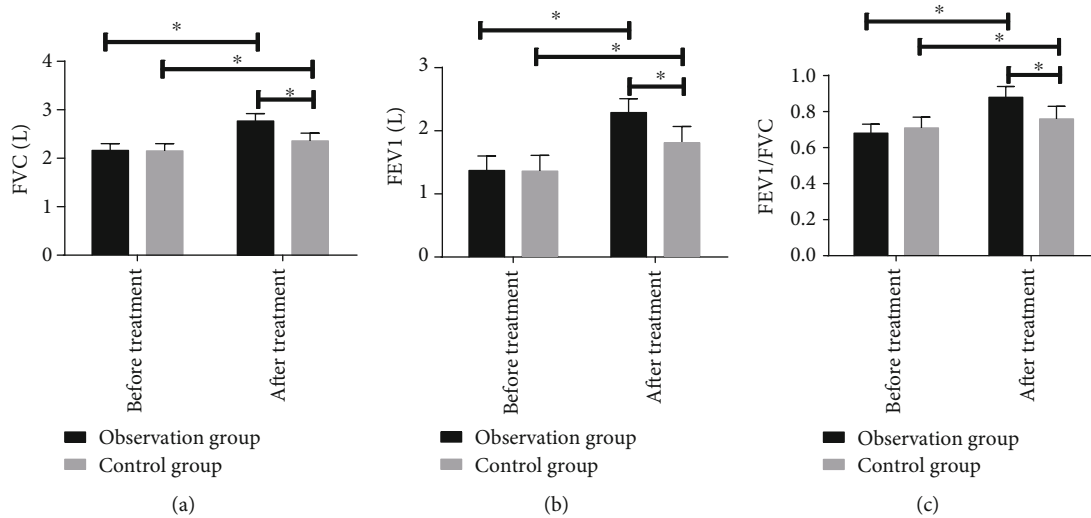


FIGURE 1: Comparison of pulmonary function indices before and after treatment: (a) the comparison of FVC before and after treatment between the observation and control groups; (b) the comparison of FEV1 before and after treatment between the observation and control groups; (c) the comparison of FEV1/FVC before and after treatment between the observation and control groups. * indicates $P < 0.05$.

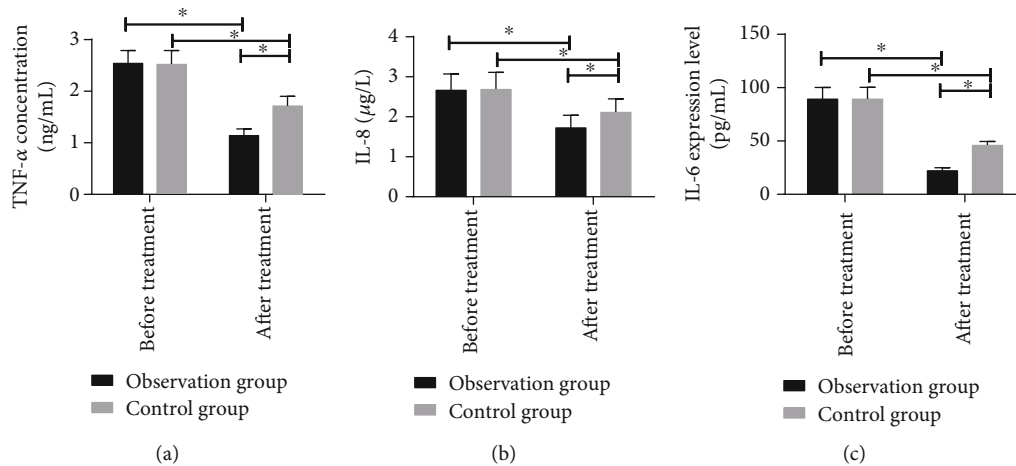


FIGURE 2: Comparison of serum inflammatory cytokines before and after treatment: (a) the comparison of TNF- α before and after treatment between the observation and control groups; (b) the comparison of IL-8 before and after treatment between the observation and control groups; (c) the comparison of IL-6 before and after treatment between the observation and control groups. * indicates $P < 0.05$.

serious inflammatory responses, and the expression of TNF- α , IL-8, and IL-6 was remarkably higher; after treatment, the expression in the serum remarkably reduced. This indicates that the patients' inflammatory responses are remarkably reduced and that BAL with AH can further reduce the severity of inflammation in CIPI patients. Some studies have shown that the body is in an oxidative stress state when acute inflammation occurs, and oxygen consumption increases when patients suffer from pulmonary infection. At this time, the body is in a relatively hypoxic state. During this process, it produces anaerobic metabolites such as MDA, which causes further damage to the pulmonary tissue [19, 20]. SOD is a cytokine with an antioxidant effect. Its content reduces when the body has oxidative

stress responses, thus weakening the body's antioxidant ability [21]. Before treatment, the oxidative stress responses in the observation and control groups were relatively strong, but they were remarkably inhibited after treatment. After treatment, compared with the control group, serum MDA content was remarkably lower but SOD content was remarkably higher in the observation group. This reveals that conventional treatment combined with AH and BAL can inhibit oxidative stress responses of CIPI patients more effectively. Finally, for further confirming the safety of BAL with AH, we compared the incidence of adverse reactions. The difference was not significant in the incidence between the observation and control groups, suggesting that BAL with AH has relatively high safety.

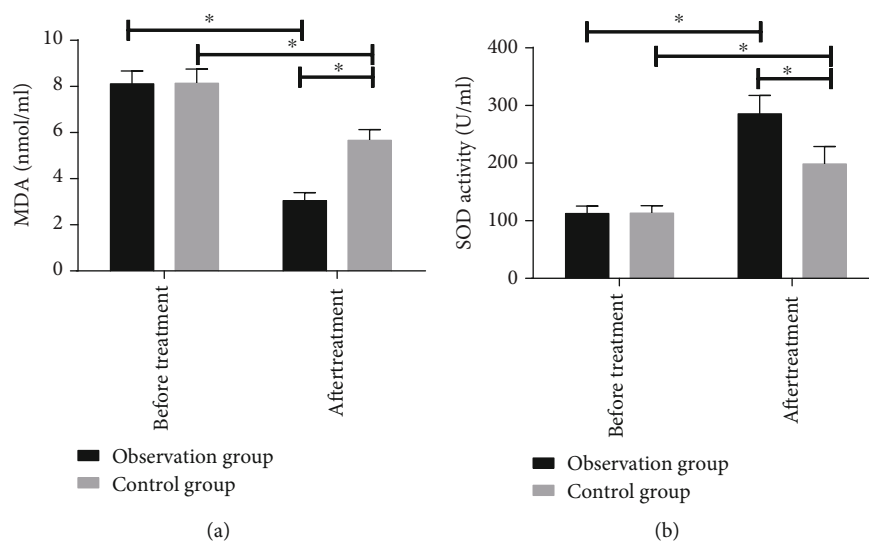


FIGURE 3: Comparison of indices of oxidative stress before and after treatment: (a) the comparison of serum MDA content between the observation and control groups; (b) the comparison of serum SOD content between the observation and control groups. * indicates $P < 0.05$.

TABLE 3: Comparison of adverse reactions.

Factors	Observation group ($n = 52$)	Control group ($n = 50$)	χ^2	P
Increased heart rate	2 (3.85)	2 (4.00)	—	—
Small amount of hemoptysis	2 (3.85)	3 (6.00)	—	—
Decreased blood oxygen	4 (7.69)	1 (2.00)	—	—
Decreased heart rate	2 (3.85)	3 (6.00)	—	—
Total incidence	10 (19.23)	9 (18.00)	0.025	0.873

However, in our study, the sample size is relatively small and needs further validation.

In summary, for CIPI patients, BAL with AH has a better therapeutic effect and higher safety and can control the patients' systemic inflammatory responses and oxidative stress responses, which is helpful to control and stabilize the patients' conditions, so it is worthy of further promotion in clinical practice.

Data Availability

All the raw data could be accessed by contacting the corresponding author if any qualified researcher needs.

Conflicts of Interest

We have no conflict of interest to declare.

Acknowledgments

We would like to acknowledge funding from the Science and technology project of the 13th five-year plan of Education Department of Jilin Province (Contract No. JJKH20180366KJ).

References

- [1] Y.-X. Liu, Q.-M. Cao, and B.-C. Ma, "Pathogens distribution and drug resistance in patients with acute cerebral infarction complicated with diabetes and nosocomial pulmonary infection," *BMC Infectious Diseases*, vol. 19, no. 1, article 603, 2019.
- [2] T. Ni, Y. Fu, W. Zhou et al., "Carotid plaques and neurological impairment in patients with acute cerebral infarction," *PLoS ONE*, vol. 15, no. 1, article e0226961, 2020.
- [3] R. Jin, X. Zhu, L. Liu, A. Nanda, D. N. Granger, and G. Li, "Simvastatin attenuates stroke-induced splenic atrophy and lung susceptibility to spontaneous bacterial infection in mice," *Stroke*, vol. 44, no. 4, pp. 1135–1143, 2013.
- [4] Y. T. Guan, L. L. Mao, J. Jia et al., "Postischemic administration of a potent PTEN inhibitor reduces spontaneous lung infection following experimental stroke," *CNS Neuroscience & Therapeutics*, vol. 19, no. 12, pp. 990–993, 2013.
- [5] K. Prass, J. S. Braun, U. Dirnagl, C. Meisel, and A. Meisel, "Stroke propagates bacterial aspiration to pneumonia in a model of cerebral ischemia," *Stroke*, vol. 37, no. 10, pp. 2607–2612, 2006.
- [6] S. Kim, M. J. Kim, C. H. Kim et al., "The superiority of IFN- λ as a therapeutic candidate to control acute influenza viral lung infection," *American Journal of Respiratory Cell and Molecular Biology*, vol. 56, no. 2, pp. 202–212, 2017.
- [7] W. Lou, S. Venkataraman, G. Zhong et al., "Antimicrobial polymers as therapeutics for treatment of multidrug-resistant

- Klebsiella pneumoniae lung infection,” *Acta Biomaterialia*, vol. 78, pp. 78–88, 2018.
- [8] P. C. Curti and H. D. Renovanz, “Therapeutic study on ambroxol in chronic bronchopulmonary diseases (author’s transl),” *Arznei-mittel-Forschung*, vol. 28, no. 5a, pp. 922–925, 1978.
- [9] A. Pandit, N. Gupta, K. Madan, S. J. Bharti, and V. Kumar, “Anaesthetic considerations for whole lung lavage for pulmonary alveolar proteinosis,” *Ghana Medical Journal*, vol. 53, pp. 248–251, 2019.
- [10] Y. Hisashi, “Ambroxol hydrochloride, bronchoalveolar lavage,” *alveolar proteinosis, atherosclerosis*, vol. 151, no. 1, pp. 165–165, 1999.
- [11] Y. X. Liu, Q. M. Cao, and B. C. Ma, “Pathogens distribution and drug resistance in patients with acute cerebral infarction complicated with diabetes and nosocomial pulmonary infection,” *BMC Infectious Diseases*, vol. 19, no. 1, article 603, 2019.
- [12] B. Kang, D. H. Kim, Y. J. Hong et al., “Complete occlusion of the right middle cerebral artery associated with *Mycoplasma pneumoniae pneumonia*,” *Korean Journal of Pediatrics*, vol. 59, no. 3, pp. 149–152, 2016.
- [13] E. C. Panagiotopoulou, S. Fouzas, K. Douros et al., “Increased β -glucuronidase activity in bronchoalveolar lavage fluid of children with bacterial lung infection: a case-control study,” *Respirology*, vol. 20, no. 8, pp. 1248–1254, 2015.
- [14] K. I. Hayashi, H. Hosoe, T. Raise, and K. Ohmori, “Protective effect of erdosteine against hypochlorous acid-induced acute lung injury and lipopolysaccharide-induced neutrophilic lung inflammation in mice,” *The Journal of Pharmacy and Pharmacology*, vol. 52, no. 11, pp. 1411–1416, 2000.
- [15] I. Kang, M. Y. Chang, T. N. Wight, and C. W. Frevert, “Proteoglycans as immunomodulators of the innate immune response to lung infection,” *The Journal of Histochemistry and Cytochemistry*, vol. 66, no. 4, pp. 241–259, 2018.
- [16] C. Wang, S. Ye, X. Wang, Y. Zhao, Q. Ma, and L. Wang, “Clinical efficacy and safety of mechanical ventilation combined with fiberoptic bronchoalveolar lavage in patients with severe pulmonary infection,” *Medical Science Monitor*, vol. 25, pp. 5401–5407, 2019.
- [17] Z. Shi, Y. Qin, Y. Zhu et al., “Effect of bronchoalveolar lavage with fiberoptic bronchoscopy combined with vibration sputum drainage on mechanically ventilated patients with severe pneumonia: a prospective randomized controlled trial in 286 patients,” *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*, vol. 29, no. 1, pp. 66–70, 2017.
- [18] M. K. KK, F. Alam, V. Flores-Malavet et al., “Memory CD4 T cell-derived IL-2 synergizes with viral infection to exacerbate lung inflammation,” *PLoS Pathogens*, vol. 15, article e1007989, 2019.
- [19] S. Bansal and S. Chhibber, “Curcumin alone and in combination with augmentin protects against pulmonary inflammation and acute lung injury generated during *Klebsiella pneumoniae* B5055-induced lung infection in BALB/c mice,” *Journal of Medical Microbiology*, vol. 59, no. 4, pp. 429–437, 2010.
- [20] S. H. Huang, X. J. Cao, W. Liu, X. Y. Shi, and W. Wei, “Inhibitory effect of melatonin on lung oxidative stress induced by respiratory syncytial virus infection in mice,” *Journal of Pineal Research*, vol. 48, no. 2, pp. 109–116, 2010.
- [21] E. Bortz, T. T. Wu, P. Patel, J. P. Whitelegge, and R. Sun, “Proteomics of bronchoalveolar lavage fluid reveals a lung oxidative stress response in murine herpesvirus-68 infection,” *Viruses*, vol. 10, no. 12, p. 670, 2018.

Research Article

Wavelet Scattering Transform for ECG Beat Classification

Zhishuai Liu,¹ Guihua Yao,² Qing Zhang¹,² Junpu Zhang,¹ and Xueying Zeng¹

¹School of Mathematical Sciences, Ocean University of China, 238 Songling Road, Qingdao, Shandong 266100, China

²Department of Cardiology, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, 758 Hefei Road, Qingdao, Shandong 266035, China

Correspondence should be addressed to Qing Zhang; zhangqing2199@163.com and Xueying Zeng; zxying@ouc.edu.cn

Received 6 June 2020; Revised 9 August 2020; Accepted 20 September 2020; Published 9 October 2020

Academic Editor: Lin Lu

Copyright © 2020 Zhishuai Liu 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.

An electrocardiogram (ECG) records the electrical activity of the heart; it contains rich pathological information on cardiovascular diseases, such as arrhythmia. However, it is difficult to visually analyze ECG signals due to their complexity and nonlinearity. The wavelet scattering transform can generate translation-invariant and deformation-stable representations of ECG signals through cascades of wavelet convolutions with nonlinear modulus and averaging operators. We proposed a novel approach using wavelet scattering transform to automatically classify four categories of arrhythmia ECG heartbeats, namely, nonectopic (N), supraventricular ectopic (S), ventricular ectopic (V), and fusion (F) beats. In this study, the wavelet scattering transform extracted 8 time windows from each ECG heartbeat. Two dimensionality reduction methods, principal component analysis (PCA) and time window selection, were applied on the 8 time windows. These processed features were fed to the neural network (NN), probabilistic neural network (PNN), and k -nearest neighbour (KNN) classifiers for classification. The 4th time window in combination with KNN ($k = 4$) has achieved the optimal performance with an averaged accuracy, positive predictive value, sensitivity, and specificity of 99.3%, 99.6%, 99.5%, and 98.8%, respectively, using tenfold cross-validation. Thus, our proposed model is capable of highly accurate arrhythmia classification and will provide assistance to physicians in ECG interpretation.

1. Introduction

Cardiovascular diseases (CVDs) are the main causes of death globally. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths [1]. There are many factors that lead to CVDs, including smoking and tobacco use, physical inactivity, poor dietary habit, overweight and obesity, etc. [2]. One broad group of complication of CVDs is arrhythmia, which expresses the electrical dysfunction of the heart.

An arrhythmia refers to the abnormal rate or rhythm of heartbeat. During an arrhythmia, the heart can beat too fast, too slowly, or with an irregular rhythm [3]. An electrocardiogram (ECG) monitors the electrical activity of the heart, and cardiac arrhythmias can be detected through any change in the morphological pattern over a recorded ECG waveform. There are many arrhythmia categories, and each contains different pathological information. Figure 1 shows the patterns of ECG signals for different arrhythmia categories. It

is of vital importance to accurately classify ECG signals into those categories in time. For cardiologists, relying on large amount of expertise and experience in their field, they visually observe the ECG waveform and obtain diagnostic results. However, this visual assessment may lead to subjective interpretations due to the presence of noise and minute morphological parameter values in ECG signals [4]. Moreover, it is also time-consuming and exhausting for cardiologists to interpret ECG signals, which may delay the best treatment opportunity for patients.

To address these drawbacks, various computer-aided diagnosis (CAD) systems have been developed recently. The CAD systems can be used as an adjunct tool for physicians in their interpretation of ECG signals to improve the accuracy and diagnostic speed. It plays an important role in the management of CVDs [5]. Table 1 summarizes some selected state-of-the-art studies of CAD systems. Most of them focused on conventional machine learning approaches. Feature extraction and classification are essential steps for

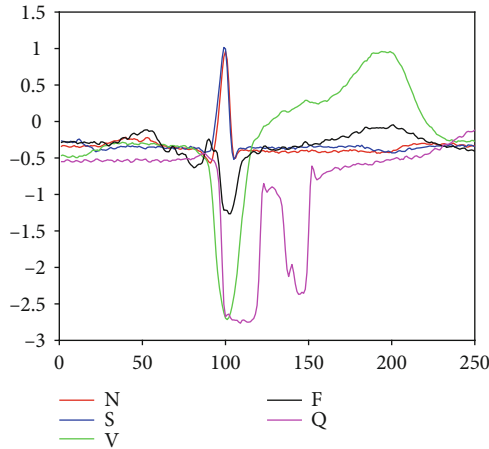


FIGURE 1: ECG signals for different arrhythmia categories.

these methods. The features extracted, including parametric and visual pattern features [6–8], from ECG signals and the classifiers designed for classification directly influence the performance of arrhythmia detection. Although some of these studies have achieved great classification performances, they might have two main drawbacks: firstly, they require a well-designed feature extractor and the features need to be manually optimized before feeding into classifiers; secondly, they usually suffer from overfitting. Moreover, few of these methods provided the confusion matrix recommended by the ANSI/AAMI EC57:1998 standard [9]. Hence, it is difficult to compare their classification performances on different arrhythmia categories in detail.

Since 2016, the methods based on deep learning approaches such as convolution neural network (CNN) have been proposed to identify abnormal ECG heartbeats including arrhythmias. Both of the feature extraction and classification are embedded together in the model. These methods have the ability to extract self-learn features [10]. However, they might have three main drawbacks: lack of strong theoretical support, requiring large amount of training data to achieve good performance, and consuming huge computational costs to train the model. Due to these drawbacks, one has to take a large number of numerical experiments to empirically conduct hyperparameter optimization as well as set up the optimal architecture, and the features extracted may be unexplainable in practical applications. Further, the performances of these methods remain to be improved.

The wavelet transform is an efficient tool for analyzing nonstationary ECG signals due to its time-frequency localization properties [11–13]. However, it is not invariant to translation. Recently, Mallat proposed a novel signal-processing method, the wavelet scattering transform, by cascading the wavelet transform with a nonlinear modulus and averaging operators [14]. The wavelet scattering transform can provide time and frequency resolutions, which is invariant to translation, stable to deformations, and preserves high frequency information for classification [15]. Moreover, Mallat characterized three properties that deep learning architectures pos-

sess for extracting useful features from data [16]: multiscale contractions, linearization of hierarchical symmetries, and sparse representation. The wavelet scattering transform also possesses these properties and, hence, has both advantages of conventional and deep learning approaches. It has achieved state-of-the-art performances in the tasks of art authentication, musical genre classification, audio recognition, and handwriting classification [17–20].

Motivated by the excellent property of wavelet scattering transform, we aim to explore the performance of the wavelet scattering transform in extracting the features from ECG signals for automated classification of arrhythmias. Specifically, we get data from the MIT-BIH Arrhythmia Database and classify the arrhythmias into four classes; more details are shown in Section 2. Then, we use wavelet scattering transform combined with some dimension reduction methods to extract features. Several existing classifiers, k -nearest neighbour (KNN), neural network (NN), and probabilistic neural network (PNN), are used to test the performances of the wavelet scattering transform on arrhythmia identification. In the end, our results are compared to some existing approaches listed in Table 1.

The paper is organized as follows: Section 2 introduces the database and data preprocessing methods. Section 3 presents the wavelet scattering transform as well as its properties and introduces the classifiers used in this study. Section 4 shows the detailed architecture and numerical experimental results, which are discussed in Section 5. We conclude the paper in Section 6.

2. Materials Used

In this section, we will briefly introduce the database that we used for ECG classification and describe our data preprocessing and augmentation methods.

2.1. MIT-BIH Database. We used the MIT-BIH Arrhythmia Database [21] to train and test our method. This database is widely used for ECG classification and is publicly available. The MIT-BIH database contains 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979 [22]. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. These records were first annotated by at least two cardiologists independently. After reaching an agreement for all annotations, the agreed annotations were marked in a computer-readable format. The annotation for every beat on ECG includes the position of R-peak and the type of arrhythmia it belongs to. The database includes 15 types of arrhythmias such as ventricular premature, atrial premature, and atrial flutter. Figure 2 shows a fragment of record 100. As shown in Figure 2, each record contains two leads, say, two channels of the ECG signal.

2.2. Data Preprocessing. According to the ANSI/AAMI EC57:1998 standard [9], the 15 types of arrhythmia beats can be classified into five categories including nonectopic (N) beats, supraventricular ectopic (S) beats, ventricular

TABLE 1: Selected automated ECG classification methods on the MIT-BIH Arrhythmia Database.

Author	Year	Method	Class	Performance
Conventional machine learning approaches				
Inan et al. [35]	2006	Feature extraction: classifier	3	ACC: 95.16%
				ACC: 99.10%
Sayadi et al. [36]	2010	Feature extraction: classifier	2	SEN: 98.77%
				SPEC: 97.47%
				ACC: 98.11%
Martis et al. [32]	2012	Feature extraction: classifier	5	SEN: 99.90%
				SPEC: 99.10%
				ACC: 97.65%
Prasad et al. [37]	2013	Feature extraction: classifier	3	SEN: 98.75%
				SPEC: 99.53%
				ACC: 99.5%
Martis et al. [38]	2013	Feature extraction: classifier	3	SEN: 100%
				SPEC: 99.22%
				ACC: 93.48%
Martis et al. [7]	2013	Feature extraction: classifier	3	SEN: 99.27%
				SPEC: 98.31%
				ACC: 94.52%
Martis et al. [39]	2013	Feature extraction: classifier	5	SEN: 98.61%
				SPEC: 98.41%
				ACC: 99.52%
Martis et al. [32]	2012	Feature extraction: classifier	5	SEN: 98.69%
				SPEC: 99.91%
				ACC: 99.45%
Martis et al. [40]	2014	Feature extraction: classifier	3	SEN: 99.61%
				SPEC: 100%
				ACC: 99.69%
Kaya and Pehlivan [41]	2015	Feature extraction: classifier	5	SEN: 99.46%
				SPEC: 99.91%
				ACC: 99.63%
Kaya and Pehlivan [8]	2015	Feature extraction: classifier	5	SEN: 99.29%
				SPEC: 99.89%
Li and Zhou [33]	2016	Feature extraction: classifier	5	ACC: 94.61%
				ACC: 94.5%
Mondjar-Guerra et al. [42]	2018	Feature extraction: classifier	5	SEN: 66.4%
				SPEC: 70.3%
Yang and Wei [6]	2020	Feature extraction: classifier	5	ACC: 97.70%
				ACC: 99.3%
This work	2020	Feature extraction: classifier	4	SEN: 99.5%
				SPEC: 98.8%
Deep learning approaches				
				ACC: 93.47%
Martis et al. [40]	2014	9-layer deep convolution neural network	5	SEN: 96.01%
				SPEC: 91.64%

ACC: accuracy; SEN: sensitivity; SPEC: specificity; WT: wavelet transform; EKF: extended Kalman filter; DCT: discrete cosine transform; DWT: discrete wavelet transform; HOS: higher order statistics; IC: independent component; ICA: independent component analysis; RR: RR intervals; WPE: wavelet packet entropy; LBP: local binary patterns; RF: random forest; LS-SVM: least square-support vector machine.

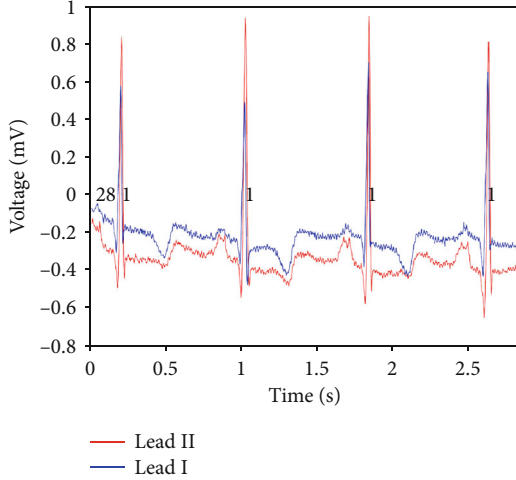


FIGURE 2: A fragment of record 100.

ectopic (V) beats, fusion (F) beats, and unknown (Q) beats. Table 2 shows the subdivisions of these categories.

Complying with the ANSI/AAMI EC57:1998 recommended practice [9], we excluded 4 records which are from patients with pacemakers, because records containing paced beats do not retain sufficient signal quality. For the remaining records, only modified-lead II signals were used. Then, we detected the R-peak in each record to segment heartbeats. The R-peak detection algorithm is not the focus of our study, as many excellent algorithms have been proposed in literatures [11, 23]. Moreover, we directly used the raw data and no denoising technique was applied. Further details are available in [9].

A total of 100507 heartbeats were segmented from the 44 records. Each beat is 250 samples long, centered around the R-peak, containing 99 samples before the R-peak and 150 samples after the R-peak. Then, they were sorted into five categories according to their annotations. Table 3 shows the number of heartbeats in each category. Similar to [6, 24, 25], the class Q was discarded since it is marginally represented (0.012%) in the database. Figure 1 shows some segments in the considered four categories.

2.3. Data Augmentation. There are huge imbalances between the number of heartbeats in classes N, S, V, and F, which will lead to inferior classification performance [10, 26]. Following the data augmentation method in [26], we augmented the data by adding Gauss white noise with zero mean and 0.05 variance. Specifically, as class N has enough heartbeats, we randomly chose 90000 heartbeats from it and did not add noise. The number of beats in the remaining classes was increased to 90000 separately to match that in class N. Consequently, the augmented database includes 360000 heartbeats.

3. Methodology

In this section, we will present our methods for ECG classification. In Section 3.1, we describe the wavelet scattering transform that we used to learn the feature representation of ECG signals. We then introduce the used classifiers in Section 3.2.

3.1. Wavelet Scattering Transform. A wavelet scattering transform builds translation invariant, stable, and informative signal representations. It is stable to deformations and preserves class discriminability, which makes it particularly effective for classification. We refer to [17–20] for its excellent practical performance for classification.

We will follow the notations in [19]. Let $f(t)$ be the signal under analysis. The low-pass filter ϕ and the wavelet function ψ are designed to build filters which cover the whole frequencies contained in the signal. Let $\phi_j(t)$ be the low-pass filter that provides locally translation invariant descriptions of f at a predefined scale T . We denote by Λ_k the family of wavelet indices having an octave frequency resolution Q_k . The multiscale high-pass filter banks $\{\psi_{j_k}\}_{j_k \in \Lambda_k}$ can be constructed by dilating the wavelet ψ .

A wavelet scattering transform is implemented with a deep convolution network that iterates over traditional wavelet transform, nonlinear modulus, and averaging operators. The convolution $S_0 f(t) = f \star \phi_j(t)$ generates a locally translation invariant feature of f , but also results in the loss of high-frequency information. These lost high frequencies can be recovered by a wavelet modulus transform

$$|W_1|f = \left\{ S_0 f(t), \left| f \star \psi_{j_1}(t) \right| \right\}_{j_1 \in \Lambda_1}. \quad (1)$$

The first-order scattering coefficients are obtained by averaging the wavelet modulus coefficients with ϕ_j :

$$S_1 f(t) = \left\{ \left| f \star \psi_{j_1} \right| \star \phi_j(t) \right\}_{j_1 \in \Lambda_1}. \quad (2)$$

To recover the information lost by averaging, noting that $S_1 f(t)$ can be seen as the low-frequency component of $|f \star \psi_{j_1}|$, we can extract complementary high-frequency coefficients by

$$|W_2|f \star \psi_{j_1} = \left\{ S_1 f(t), \left| \left| f \star \psi_{j_1} \right| \star \psi_{j_2}(t) \right| \right\}_{j_2 \in \Lambda_2}. \quad (3)$$

It further defines the second-order scattering coefficients

$$S_2 f(t) = \left\{ \left| \left| f \star \psi_{j_1} \right| \star \psi_{j_2} \right| \star \phi_j(t) \right\}_{j_1 \in \Lambda_1}, \quad i = 1, 2. \quad (4)$$

Iterating the above process defines wavelet modulus convolutions

$$U_m f(t) = \left\{ \left| \left| \left| f \star \psi_{j_1} \right| \star \dots \star \psi_{j_m} \right| \right| \right\}_{j_i \in \Lambda_i}, \quad i = 1, 2, \dots, m. \quad (5)$$

Averaging $U_m f(t)$ with ϕ_j gives the m -th-order scattering coefficients

$$S_m f(t) = \left\{ \left| \left| \left| f \star \psi_{j_1} \right| \star \dots \star \psi_{j_m} \right| \star \phi_j(t) \right| \right\}_{j_i \in \Lambda_i}, \quad i = 1, 2, \dots, m. \quad (6)$$

TABLE 2: MIT-BIH Arrhythmia Database beats classified as per ANSI/AAMI EC57:1998 standard [9].

N	S	V	F	Q
Normal	Atrial premature	Premature ventricular contraction	Fusion of ventricular and normal	Paced
Left bundle branch	Aberrant atrial			Fusion of paced and normal
Right bundle branch block	Nodal (junctional) premature	Ventricular escape		Unclassifiable
Atrial escape	Supraventricular premature			
Nodal (junctional) escape				

TABLE 3: The breakdown of five arrhythmia categories.

Class	Number of ECG heartbeats
N	90023
S	2758
V	6914
F	800
Q	12
Total	100507

This scattering process is illustrated in Figure 3. The final scattering matrix

$$Sf(t) = \{S_m f(t)\}_{0 \leq m \leq l}, \quad (7)$$

aggregates scattering coefficients of all orders to describe the features of input signal, where l is the maximal decomposition order.

The network is invariant to translations up to the invariance scale, which can be potentially large, due to the average operation determined by the low-pass filter ϕ_J . As a property inherited from wavelet transform, the features $Sf(t)$ are stable to local deformations. The scattering decomposition can capture subtle changes in amplitude and duration of ECG signals, which are hard to measure but reflect the condition of the heart. Therefore, we use the wavelet scattering network to produce robust representations of ECG heartbeats that minimize differences within one arrhythmia category while maintaining enough discriminability between different categories.

Though the structure of the wavelet scattering network is similar to CNN, they have two main differences: the filters are not learned but set in advance and the features are not only the output of the last convolution layer but also the combination of all those layers. It has been shown that the energy of scattering coefficients decreases rapidly as the layer level increases, with almost 99% of the energy contained in the first two layers [18, 19]. Therefore, we used a two-order scattering network to extract the features of ECG signals. This also reduces the computational complexity significantly.

3.2. Classifier. We next briefly introduce the used classifiers that combine features to predict the class membership of

the ECG signal. We choose classifiers according to two criteria. First, the classifier must be widely used in existing literatures, such as NN, KNN, PNN, and support vector machine (SVM). Second, it must be capable of efficiently processing high dimension and large size training data. NN, KNN, and PNN satisfy both of the requirements, while SVM is ruled out for the low computational efficiency. Thus, we use NN, KNN, and PNN for classification in this work.

3.2.1. Neural Network. The feedforward NN is the most widely used artificial neural network for classification [27, 28]. We set the architecture as follows. There are 75 neurons in the input layer, corresponding to the 75 dimensions of the feature vector extracted by wavelet scattering transform. Six hidden layers contain 70, 60, 45, 30, 20, and 10 neurons, respectively, and the first five hidden layers are activated by the ReLU function: $f(x) = \max(0, x)$. The output layer has 4 neurons, each of which represents an arrhythmia category and is activated by the Softmax function:

$$g(z)_i = \frac{\exp z_i}{\sum_{j=1}^4 \exp z_j}, \quad i = 1, \dots, 4. \quad (8)$$

We used the cross-entropy cost function [10] and employed error backpropagation algorithm to solve the weights. The Adam algorithm [29] was used to adaptively update the learning rate. We set the iteration number to 50 which is enough for training the network.

The above architecture was set up through trial and error. We have tried several combinations of different numbers of hidden layers, different activation functions, different numbers of neurons in each layer, different numbers of sample sizes in minibatch, and different epochs of parameter update, etc. Considering the computational cost and classification accuracy comprehensively, the network we present achieves the optimal performance compared to other tested architectures. Once the neural network was trained, all the testing data were fed into the network to measure its classification performance.

3.2.2. Probabilistic Neural Network. The PNN [30] is widely used in classification and pattern recognition problems. In the PNN algorithm, the class probability of a new input data is estimated and the Bayesian rule is then employed to allocate the class with the highest posterior probability to new

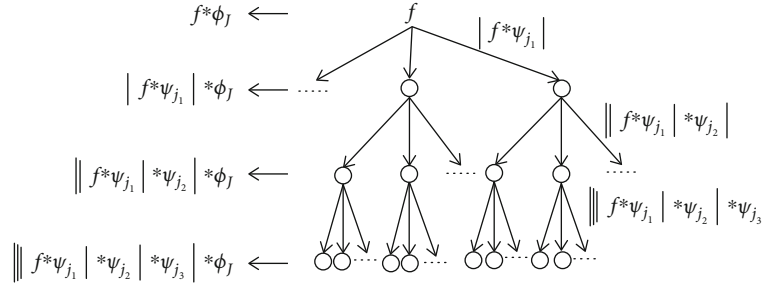


FIGURE 3: The tree view of wavelet scattering network.

input data. The operations in a PNN are organized into a feedforward network with four layers: input layer, pattern layer, summation layer, and output layer. The input layer has the same number of neurons as the dimension of feature vector. Each neuron represents a predictor variable and feeds the values to each of the neurons in the pattern layer. The pattern layer contains one neuron for each sample in the training data. Each hidden neuron computes the Euclidean distance of the test sample from the neuron's center point. The summation layer has the same number of neurons as that of the categories of the input data. The weight coming out of a hidden neuron is fed only to the pattern neuron that corresponds to the hidden neuron's category. The output layer compares the weighted votes for each target category accumulated in the summation layer and uses the largest vote to predict the target category. PNN is more accurate than the multilayer neural network. It can approach the Bayesian optimal classification as long as the training data is enough. In this study, four layers in the trained PNN contain 75, 324000, 4, and 1 neurons, respectively.

3.2.3. *k*-Nearest Neighbours. The KNN is a nonparametric method widely used for classification. The input consists of the k closest training samples in the feature space. An unlabeled data is classified by assigning the label which is most frequent among the k training samples nearest to that query data. The commonly used distance metric for KNN is the Euclidean distance. As for the selection of k values, we use the brute-force method. Specifically, $k = 1, 2, 3, 4, 5$ have been tested and $k = 4$ is the most appropriate value for the classification. Thus, we only present the results of $k = 4$ in Section 4.

4. Experimental Results

In this section, we will discuss the features extracted by scattering transform and our classification process. Specifically, two methods will be introduced for dimensionality reduction based on the pattern of features.

The wavelet scattering transform, PNN, and KNN classifiers were implemented by MATLAB 2018b. We used Python 3.7 to implement the NN classifier.

4.1. Feature Extraction. We used the Gabor wavelets to perform wavelet decomposition. The corresponding low-pass filter ϕ is a Gaussian function. We set the invariance scale to 0.5 second. The constructed wavelet scattering network

includes two layers. We set $Q_1 = 8$ and $Q_2 = 1$ wavelets per octave at the first and second layers, respectively. We had tried other different settings for the invariance scale and wavelet octave resolution, but this architecture preserves the signal information best for classification. Figure 4 shows the used Gabor wavelets and its low-pass filter $\phi_j(t)$. Note that the coarsest-scale wavelet does not exceed the invariance scale determined by the time support of the low-pass filter $\phi_j(t)$.

The output of the wavelet scattering network forms a tensor with the size of $75 \times 8 \times 36000$. Each slice of the tensor is the scattering coefficients of one ECG heartbeat. The scattering coefficients are critically downsampled in time based on the bandwidth of the low-pass filter, which results in 8 time windows for each of the 75 scattering paths. To obtain a data structure compatible with the used classifiers, we reshaped the tensor into a 2880000×75 matrix where each column and row corresponds to a scattering path and a time window, respectively. We obtained 2880000 rows because there are 8 time windows for each of the 360000 signals in the database. Figure 5 shows the scattering coefficients of the 8 time windows for one ECG heartbeat.

4.2. Classification with NN. The NN classifier is capable of classification task for big data, so we used it to preliminarily test the classification performances of 8 time windows. For each heartbeat, we created labels to match the number of time windows. The decision for each time window was aggregated by majority vote to generate a label for the input ECG heartbeat.

We employed a 10-fold cross-validation [31]. Firstly, the 360000 ECG heartbeats were divided into 10 equal parts. Then, 90% of them were used to train the network, and the remaining 10% were used for testing. This process was repeated 10 times, and the overall performance was the averaged value over the 10 folds.

The AAMI has provided the standards and recommended practices for reporting performance results of automated arrhythmia detection algorithms [9]. We followed those practices so that the methods in this paper can be compared with those in Table 1. The positive predictive value (PPV), sensitivity (SEN), and specificity (SPEC) were used to measure the classification performances of our methods. Table 4 presents the confusion matrix across 10 folds. Table 5 presents the accuracy of each time window.

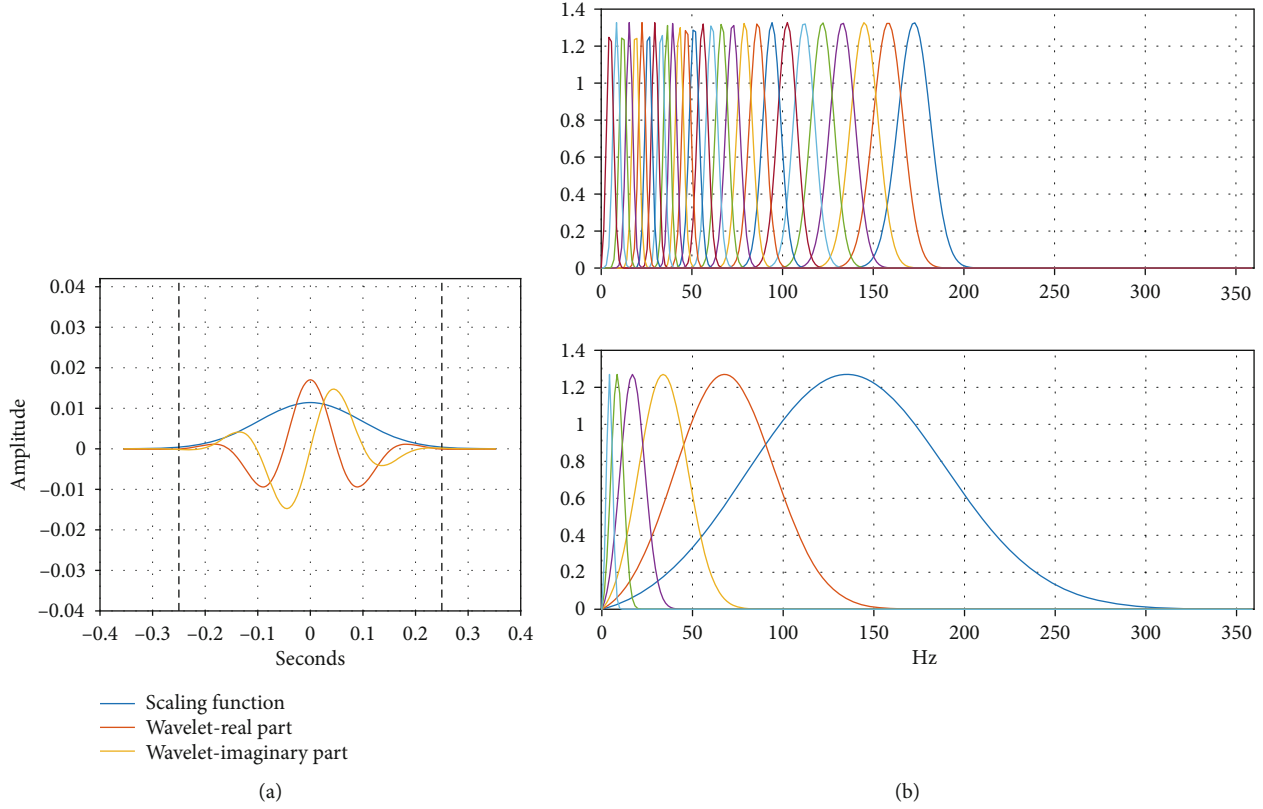


FIGURE 4: Wavelet filters. (a) The low pass filter with 0.5 s invariance scale. (b) The first filter bank with 8 wavelets per octave and the second filter bank with 1 wavelet per octave.

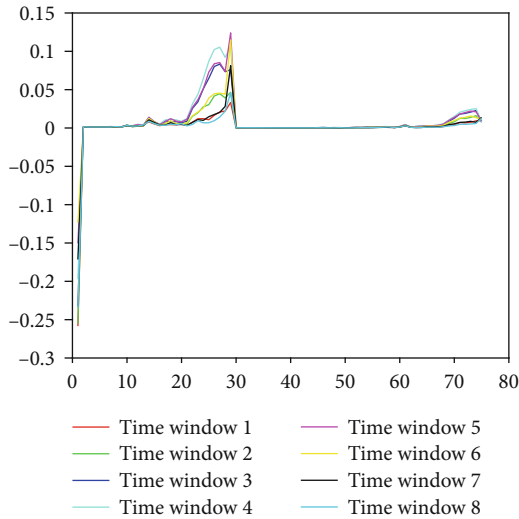


FIGURE 5: Scattering coefficients of 8 time windows for one ECG heartbeat.

4.3. *PCA and Time Window Selection.* As shown in Figure 5, the 8 time windows are significantly correlated with each other. Table 5 illustrates that the 3th, 4th, and 5th time windows have better discrimination than the others. We can also see from Figure 5 that these three time windows have larger amplitude and more fluctuations, which means they contain more and clearer details of ECG heartbeat, especially the 4th

TABLE 4: The confusion matrix for 8 time windows combined with the NN across 10 folds.

Original	Predicted						
	N	S	V	F	PPV (%)	SEN (%)	SPEC (%)
N	88681	369	737	213	90.9	98.5	96.7
S	4106	82994	2006	894	93.9	92.2	98.0
V	2669	2132	83701	1498	91.7	93.0	97.2
F	2124	2872	4843	80161	96.9	89.1	96.5

TABLE 5: The accuracy of each time window classified by the NN.

Time window	1	2	3	4	5	6	7	8
ACC (%)	88.9	90.3	92.2	92.8	92.2	91.2	89.8	87.3

time window. In order to get better performance and reduce computational cost, we used two methods to reduce the redundancy of the 8 time windows.

- (i) *Principal component analysis (PCA):* PCA projects features in the directions of the highest variance to reduce the dimensionality of features [32]. The first few principal components can represent the most variability in features. The contribution rate of a principal component is the percentage of the total variability it represents. In this study, there are 8 time

windows for each node in the scattering network. However, the 8 time windows have collinearity to some extent, which may lead to low classification performance. In order to remove the collinearity and generate more concise features, we used PCA to extract principal components of the 8 time windows for each node. The averaged contribution rate of the first and second principal components is approximately 84% and 15%, respectively. Hence, for each ECG heartbeat, we took the first principal component of 8 time windows as the new feature, which is a 75-dimensional vector with each dimension corresponding to a node.

- (ii) *Time window selection*: as described in Section 4.3, majority vote was used to predict the label for each testing ECG heartbeat. However, as shown in Table 5, the performance can be affected by those time windows with low accuracy. Moreover, the pathological information of ECG signals mainly concentrates around the R-peak, which has a very short duration. The discrimination between different arrhythmia categories may be involved in one particular time window. This motivates us to test the performance of each time window separately using different classifiers and find the time windows that generate the best classification results.

The NN classifier is capable of using any number of time windows as features. While limited by their computational ability, the PNN and KNN classifiers are suitable for the case of using one time window as features. To test the PCA method, we fed the first principal component of 8 time windows into the NN, PNN, and KNN classifiers, respectively. The confusion matrices across 10 folds are shown in Table 6. To test the time window selection method, we conducted two experiments. Firstly, we fed the 8 time windows into the NN, PNN, and KNN classifiers separately and found that the 4th time window generates the best performance. The confusion matrices are shown in Table 7. Secondly, we tried different time window combinations and classified them by the NN classifier and found that the combination of the 3th, 4th, and 5th time windows performs better than the others. Table 8 presents the confusion matrix.

5. Discussion

In this section, we will discuss the classification results presented in Section 4.3 and compare our methods with those state-of-the-art studies.

NN: among all methods using the NN classifier, the one using the 4th time window as feature provides the maximum averaged ACC of 98.1% and averaged PPV, SEN, and SPEC of 99.3%, 98.2%, and 97.8%, respectively. Comparing Tables 4, 7, and 8, we can confirm that removing some time windows improves the classification performance. This indicates that there is some redundancy among the 8 time windows and the differences between the four categories (N, S, V, and F) are mainly reflected in the 3th, 4th, and 5th time

windows. The performance of the 4th time window is close to that of the combination of the 3th, 4th, and 5th time windows. However, the training time of the latter is three times that of the single 4th time window. Moreover, the performance of the first principal component of 8 time windows is unsatisfactory, which is much worse than that of the 4th time window.

PNN: comparing Tables 6 and 7, the 4th time window and the first principal component provide almost the same results in combination with the PNN (spread = 0.01) classifier. The former is slightly better, yielding an averaged ACC, PPV, SEN, and SPEC of 99.0%, 98.7%, 99.9%, and 96.0%, respectively. We set the spread value by the brute-force method. The PNN classifiers with a spread value of 0.005, 0.01, 0.02, 0.03, 0.04, 0.1, and 1 have been tested, and the one with the spread value of 0.01 produces the best results. Table 7 shows that the SEN of supraventricular ectopic beats (SVEB) and ventricular ectopic beats (VEB) are 99.8% and 99.9%, respectively; it means that almost all the SVEB and VEB have been correctly detected. Therefore, the PNN classifier has excellent performance in classifying the SVEB and VEB, which should be paid more attention in clinical diagnosis.

KNN: the best performance of this work is achieved by KNN with $k = 4$ and using the 4th time window as the feature. The averaged ACC, PPV, SEN, and SPEC are 99.3%, 99.6%, 99.5%, and 98.8%, respectively, and are much better than those of the PCA features. However, this result only measures the performance in classifying normal (N) and abnormal (S, V, and F) ECG heartbeats. From Table 7, we can find that the PNN classifier performs better in classifying different arrhythmia categories, especially the VEB and SVEB.

Table 1 summarizes recent advances in automated classification of ECG beats using the MIT-BIH Arrhythmia Database. Only four of them have the same arrhythmia categories as this work, which are N, S, V, F, and Q. Martis et al. [32] used PCA on discrete cosine transform (DCT) coefficients computed from the segmented beats of ECG. The dimensionality-reduced features in combination with the KNN classifier yield the highest averaged ACC, SEN, and SPEC of 99.52%, 98.69%, and 99.91%, respectively. However, the confusion matrix was not provided in [32]. Li and Zhou [33] used wavelet packet entropy (WPE) and random forests (RF) to classify ECG signals into 5 categories; they obtained an ACC of 94.61%. Acharya et al. [26] used a 9-layer convolution neural network and achieved an averaged ACC, SEN, and SPEC of 93.47%, 96.01%, and 91.64%, respectively. Yang and Wei [6] combine parametric and visual pattern features and use KNN for classification. They obtain an overall ACC of 97.70%. The accuracies of V and S are not satisfying and reduce the overall accuracy significantly.

Table 9 summarizes the performances achieved by our methods. We can conclude from Tables 9 and 1 that the performance of this work is better than those state-of-the-art studies which classify ECG heartbeats into 5 categories (N, S, V, F, and Q). This demonstrates that wavelet scattering transform performs well in extracting the features of ECG

TABLE 6: The confusion matrices for the first principal component combined with the NN, PNN, and KNN across 10 folds.

	Original	Predicted						
		N	S	V	F	PPV (%)	SEN (%)	SPEC (%)
NN	N	83508	2621	2799	1072	92.4	92.8	97.5
	S	3144	84182	1946	728	94.0	93.5	98.0
	V	1434	996	86100	1460	91.7	95.7	97.1
	F	2270	1803	3069	82858	96.2	92.1	97.4
PNN	N	86738	1225	1459	578	97.8	96.4	99.3
	S	1646	88085	214	55	98.6	97.9	99.5
	V	50	13	89797	140	98.1	99.8	99.4
	F	228	27	21	89724	99.1	99.7	99.9
KNN	N	87816	690	1091	403	96.5	97.6	98.8
	S	2418	86520	627	435	98.9	96.1	99.7
	V	156	62	89612	170	98.0	99.6	99.3
	F	572	186	112	89130	98.9	99.0	99.7

TABLE 7: The confusion matrices for the 4th time window combined with the NN, PNN, and KNN across 10 folds.

	Original	Predicted						
		N	S	V	F	PPV (%)	SEN (%)	SPEC (%)
NN	N	88146	811	789	254	93.8	97.9	97.8
	S	3181	85641	901	277	94.9	95.2	98.3
	V	1553	1323	85084	2040	94.4	94.5	98.1
	F	1108	2439	3343	83110	97.0	92.3	97.5
PNN	N	86415	2540	615	430	99.8	96.0	99.9
	S	171	89828	0	1	97.2	99.8	99.0
	V	1	71	89879	49	99.3	99.9	99.8
	F	0	1701	1645	86654	96.0	96.3	98.8
KNN	N	88915	644	281	160	98.5	98.8	99.5
	S	893	87177	1155	775	92.7	96.9	97.4
	V	496	4545	82247	2712	96.4	91.4	98.9
	F	0	3	0	89997	99.5	100	100

TABLE 8: The confusion matrix for the 3th, 4th, and 5th time windows combined with the NN across 10 folds.

Original	Predicted						
	N	S	V	F	PPV (%)	SEN (%)	SPEC (%)
N	88156	560	1042	242	95.3	98.0	98.4
S	2662	86180	721	437	96.3	95.8	98.8
V	909	933	86431	1727	94.6	96.0	98.2
F	794	1836	3163	84207	97.2	93.6	97.9

heartbeats that minimize intraclass differences and maintain interclass discriminability. Moreover, the scattering coefficients in particular time windows contain more representative information for different categories than those in the other time windows. The dimensionality reduction of the 8 time windows eliminates the redundancy of features, which not only improves the classification performance

but also reduces the computational cost. In this study, our results show that the scattering coefficients of the 4th time window contain sufficient information for the classification of arrhythmias.

6. Conclusion

In this study, we discussed the automated ECG classification using the nonlinear features extracted by wavelet scattering transform from ECG beats. Combined with proper classifiers, this study demonstrates that the wavelet scattering coefficients can be well utilized for classification and yield highly accurate classification results. Our results showed that the scattering coefficients of the 4th time window combined with the KNN classifier achieve the best performance. The averaged ACC, PPV, SEN, and SPEC are 99.3%, 99.6%, 99.5%, and 98.8%, respectively. In our future work, we will attempt to combine all time windows by a proper method and then feed them into a sparse classifier to improve the classification

TABLE 9: Summary of classification results achieved by all the methods in this paper.

Feature extraction	Classifier	TP	TN	FP	FN	ACC (%)	PPV (%)	SEN (%)	SPEC (%)
WSN	NN	261101	88681	1319	8899	97.2	99.5	96.7	98.5
	NN	263152	83508	6492	6848	96.3	97.6	97.5	92.3
WSN+PCA	PNN	268076	86738	3262	1924	98.6	98.8	99.3	96.4
	KNN	266854	87816	2184	3146	98.5	99.2	98.8	97.6
WSN+the 3th, 4th, and 5th time window	NN	264158	88146	1854	5842	97.9	99.3	97.8	97.9
	NN	265091	87999	2001	4909	98.1	99.3	98.2	97.8
WSN+the 4th time window	PNN	269828	86415	3585	172	99.0	98.7	99.9	96.0
	KNN	268611	88915	1085	1389	99.3	99.6	99.5	98.8

TP: true positive; TN: true negative; FP: false positive; FN: false negative; WSN: wavelet scattering network.

performance and reduce the computational cost. Moreover, all the work presented in Table 1 are patient independent, that is, ECG beats are collected from a patient pool and experiments are conducted without considering the autocorrelation of ECG beats from the same patient. Nascimento et al. [34] propose an innovation in the configuration of the structural cooccurrence matrix. It is also of great interest to expand the wavelet scattering transform to the patient-dependent classification of arrhythmias using ECG signals.

Data Availability

All the data utilized in our research can be accessed from <http://ecg.mit.edu/dbinfo.html>.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

Zhishuai Liu, Junpu Zhang, and Xueying Zeng were supported by the National Natural Science Foundation of China [Nos. 11701538, 11771408 and 11871444] and the Fundamental Research Funds for the Central Universities [No. 201964006]. Qing Zhang and Guihua Yao were supported by the National Natural Science Foundation of China [No. 81671703], the Key Research and Development Project of Shandong Province [No. 2015GSF118026], the Qingdao Key Health Discipline Development Fund, and the People's Livelihood Science and Technology Project of Qingdao [No. 18-6-1-62-nsh].

References

- [1] World Health Organization, "Cardiovascular diseases (CVDs)," May 2019, [https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
- [2] E. J. Benjamin, S. S. Virani, C. W. Callaway et al., "Heart disease and stroke statistics-2018 update: a report from the American Heart Association," *Circulation*, vol. 137, no. 12, 2018.
- [3] Nation Heart Lung and Blood Institute, "Arrhythmia," May 2019, <https://www.nhlbi.nih.gov/health-topics/arrhythmia>.
- [4] R. J. Martis, U. R. Acharya, and H. Adeli, "Current methods in electrocardiogram characterization," *Computers in Biology and Medicine*, vol. 48, no. 3, pp. 133–149, 2014.
- [5] M. M. A. Hadhoud, M. I. Eladawy, and A. Farag, "Computer aided diagnosis of cardiac arrhythmias," in *2006 International Conference on Computer Engineering and Systems*, pp. 262–265, Cairo, Egypt, 2006.
- [6] H. Yang and Z. Wei, "Arrhythmia recognition and classification using combined parametric and visual pattern features of ECG morphology," *IEEE Access*, vol. 8, no. 99, pp. 47103–47117, 2020.
- [7] R. J. Martis, U. R. Acharya, K. M. Mandana, A. K. Ray, and C. Chakraborty, "Cardiac decision making using higher order spectra," *Biomedical Signal Processing and Control*, vol. 8, no. 2, pp. 193–203, 2013.
- [8] Y. Kaya and H. Pehlivan, "Classification of premature ventricular contraction in ECG," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 34–40, 2015.
- [9] Association for the Advancement of Medical Instrumentation, "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," vol. 1998, 1998ANSI/AAMI EC38.
- [10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT press, 2016.
- [11] C. Li, C. Zheng, and C. Tai, "Detection of ECG characteristic points using wavelet transforms," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 1, pp. 21–28, 1995.
- [12] T. Ince, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ECG signals," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 5, pp. 1415–1426, 2009.
- [13] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 437–448, 2013.
- [14] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.
- [15] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.
- [16] S. Mallat, "Understanding deep convolutional networks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, article 20150203, 2016.
- [17] R. Leonarduzzi, H. Liu, and Y. Wang, "Scattering transform and sparse linear classifiers for art authentication," *Signal Processing*, vol. 150, pp. 11–19, 2018.

- [18] J. Bruna and S. Mallat, "Classification with scattering operators," in *Computer Vision and Pattern Recognition*, pp. 1561–1566, Providence, RI, USA, June 2011.
- [19] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *International Society for Music Information Retrieval Conference*, pp. 657–662, Miami, Florida, USA, 2011.
- [20] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.
- [21] R. Mark and G. Moody, "MIT-BIH Arrhythmia Database," May 2019, <http://ecg.mit.edu/dbinfo.html>.
- [22] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2002.
- [23] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. -BME-32, no. 3, pp. 230–236, 1985.
- [24] M. M. A. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Information Sciences*, vol. 345, pp. 340–354, 2016.
- [25] K. Luo, J. Q. Li, Z. G. Wang, and A. Cuschieri, "Patient-specific deep architectural model for ECG classification," *Journal of Healthcare Engineering*, vol. 2017, Article ID 4108720, 13 pages, 2017.
- [26] U. R. Acharya, S. L. Oh, Y. Hagiwara et al., "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017.
- [27] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford university press, 1995.
- [28] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall PTR, 1994.
- [29] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, <http://arxiv.org/abs/1412.6980>.
- [30] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.
- [31] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2012.
- [32] R. J. Martis, U. R. Acharya, K. M. Mandana, A. K. Ray, and C. Chakraborty, "Application of principal component analysis to ECG signals for automated diagnosis of cardiac health," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11792–11800, 2012.
- [33] T. Li and M. Zhou, "ECG classification using wavelet packet entropy and random forests," *Entropy*, vol. 18, no. 8, p. 285, 2016.
- [34] N. M. M. Nascimento, L. B. Marinho, S. A. Peixoto, J. P. do Vale Madeiro, V. H. C. de Albuquerque, and P. P. R. Filho, "Heart arrhythmia classification based on statistical moments and structural co-occurrence," *Circuits Systems and Signal Processing*, vol. 39, no. 2, pp. 631–650, 2020.
- [35] O. T. Inan, L. Giovangrandi, and G. T. A. Kovacs, "Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2507–2515, 2006.
- [36] O. Sayadi, M. B. Shamsollahi, and G. D. Clifford, "Robust detection of premature ventricular contractions using a wave-based Bayesian framework," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 2, pp. 353–362, 2010.
- [37] H. Prasad, R. J. Martis, U. R. Acharya, L. C. Min, and J. S. Suri, "Application of higher order spectra for accurate delineation of atrial arrhythmia," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 57–60, Osaka, Japan, 2013.
- [38] R. J. Martis, U. R. Acharya, H. Prasad, C. K. Chua, C. M. Lim, and J. S. Suri, "Application of higher order statistics for atrial arrhythmia classification," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 888–900, 2013.
- [39] R. J. Martis, U. R. Acharya, C. M. Lim, K. M. Mandana, A. K. Ray, and C. Chakraborty, "Application of higher order cumulant features for cardiac health diagnosis using ECG signals," *International Journal of Neural Systems*, vol. 23, no. 4, 2013.
- [40] R. J. Martis, U. R. Acharya, H. Adeli et al., "Computer aided diagnosis of atrial arrhythmia using dimensionality reduction methods on transform domain representation," *Biomedical Signal Processing and Control*, vol. 13, pp. 295–305, 2014.
- [41] Y. Kaya and H. Pehlivan, "Feature selection using genetic algorithms for premature ventricular contraction classification," in *2015 9th International Conference on Electrical and Electronics Engineering (ELECO)*, pp. 1229–1232, Bursa, Turkey, November 2015.
- [42] V. Mondjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, and M. Ortega, "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers," *Biomedical Signal Processing and Control*, vol. 47, pp. 41–48, 2019.

Research Article

Texture Feature-Based Classification on Transrectal Ultrasound Image for Prostatic Cancer Detection

Xiaofu Huang,¹ Ming Chen ,² Peizhong Liu ,^{1,3} and Yongzhao Du ^{1,3}

¹College of Engineering, Huaqiao University, Quanzhou 362021, China

²Zhangzhou Affiliated Hospital of Fujian Medical University, Zhangzhou 363000, China

³Collaborative Innovation Center for Application Technology of Maternal and Infant Health Services, Quanzhou Medical College, Quanzhou 362022, China

Correspondence should be addressed to Ming Chen; 304552532@qq.com and Peizhong Liu; pzliu@hqu.edu.cn

Received 25 June 2020; Revised 21 August 2020; Accepted 20 September 2020; Published 6 October 2020

Academic Editor: Lei Chen

Copyright © 2020 Xiaofu Huang 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.

Prostate cancer is one of the most common cancers in men. Early detection of prostate cancer is the key to successful treatment. Ultrasound imaging is one of the most suitable methods for the early detection of prostate cancer. Although ultrasound images can show cancer lesions, subjective interpretation is not accurate. Therefore, this paper proposes a transrectal ultrasound image analysis method, aiming at characterizing prostate tissue through image processing to evaluate the possibility of malignant tumours. Firstly, the input image is preprocessed by optical density conversion. Then, local binarization and Gaussian Markov random fields are used to extract texture features, and the linear combination is performed. Finally, the fused texture features are provided to SVM classifier for classification. The method has been applied to data set of 342 transrectal ultrasound images obtained from hospitals with an accuracy of 70.93%, sensitivity of 70.00%, and specificity of 71.74%. The experimental results show that it is possible to distinguish cancerous tissues from noncancerous tissues to some extent.

1. Introduction

Prostate cancer is one of the most common malignant tumours in the male genitourinary system. In recent years, its incidence and mortality rate in China has been increasing. In 2018, the China Cancer Center released the latest issue of national cancer statistics, pointing out that prostate cancer ranked sixth in the incidence rate of men in China, only lower than lung cancer, gastric cancer, liver cancer, esophageal cancer, and intestinal cancer [1]. The following year, the American Cancer Society released a data analysis report, showing that prostate cancer was the first in male morbidity and the second in deaths [2]. Prostate cancer has posed a serious threat to men's health, so early detection is particularly important.

At present, digital rectal examination, prostate-specific antigen, nuclear magnetic resonance imaging, and transrectal ultrasound are commonly used methods to examine prostate cancer [3]. Among them, a digital rectal examination is the most common and cheapest method to examine prostate

cancer. However, the digital rectal examination cannot reach tumours in the anterior part of the prostate, which is easy to miss diagnosis [3]. Prostate-specific antigen concentration is a sensitive indicator for the diagnosis of prostate cancer, but some patients with benign prostate diseases will also have increased prostate-specific antigen concentration [4]. Therefore, prostate-specific antigen examination is easy to cause overdiagnosis, leading to unnecessary biopsy and potential overtreatment [5]. Nuclear magnetic resonance imaging (MRI) is an important examination technique for noninvasive evaluation of the prostate and its surrounding tissues, which has a relatively high diagnostic accuracy for prostate. Some studies have shown that compared with transrectal ultrasound-guided prostate biopsy, MRI-guided prostate biopsy can puncture targeted nodules with higher accuracy [6]. However, because MRI-guided biopsy requires special equipment, which is time-consuming and expensive, it cannot be popularized at present.

Transrectal ultrasound is generally used to guide prostate biopsy because of its visualization of prostate, nondamage,

low cost, and real-time characteristics. A transrectal ultrasound-guided prostate biopsy is the gold standard for diagnosing prostate cancer. Although transrectal ultrasound is currently the most widely used imaging method, unfortunately, the visual interpretation of transrectal ultrasound images is poor and is not very reliable in distinguishing prostate cancer from normal glandular tissue. The diagnosis process will inevitably take the form of a tissue biopsy. However, a transrectal ultrasound-guided biopsy is to uniformly sample glands, not prostate cancer [7], and its positive diagnostic rate is shallow, especially for early prostate cancer lesions [8]. To obtain reliable results in histological analysis, multiple puncture biopsies are often required [9]. However, the increase of puncture times will bring a lot of pain to the patient (the probability of complications such as postoperative infection, hematuria, hematochezia, and the like will increase). At the same time, more clinically meaningless cancers will also be detected, resulting in excessive diagnosis and treatment [10].

Although the predictive value of positive results of transrectal ultrasound examination is very low (even trained urologists can hardly detect prostate cancer from ultrasound images), it is currently the most commonly used image detection method for diagnosing prostate cancer [3]. Improving the detection accuracy of transrectal ultrasound is helpful to reduce the number of puncture biopsies. Therefore, one possible way to improve the transrectal ultrasound-guided prostate puncture is to use computer-assisted analysis of transrectal ultrasound images.

Due to speckle noise, artifacts, attenuation, and signal loss inherent in transrectal ultrasound images, it is difficult for ultrasound doctors to analyze the image from the texture level to determine whether the image is positive (suffering from prostate cancer) or negative (normal). Therefore, this study uses a texture feature analysis method to try to obtain useful information from transrectal ultrasound images so as to improve the accuracy of prostate cancer detection.

We realize that our research is very difficult because specific pixels are not correctly marked. Ultrasound doctors are unable to analyze images at the microtexture level to determine whether pixels are positive or negative, and histological analysis of extracted tissues cannot be converted into pixel marker maps. Therefore, we can only use an imperfect label for all pixels in the biopsy area. Despite this problem, we have achieved some results, showing that it is possible to distinguish cancer tissues from noncancer tissues to a certain extent.

2. Related Work

Texture features consider the distribution of pixel intensity and the relationship between adjacent pixels [11, 12]. Different texture measurements often describe the corresponding texture from different angles. In medical imaging, since the internal structure of lesions can be quantitatively described, each texture feature is considered as an important indication feature for image pattern recognition [13, 14]. Previous studies have shown that heterogeneity reflected by texture fea-

tures can be used to identify the nature of lesions with high diagnostic accuracy [15, 16].

In recent years, with the vigorous development of artificial intelligence, the technology of machine science has become increasingly mature. Many diagnostic methods based on computer-aided diagnosis provide convenience for medical diagnosis. In the field of vision research, medical image research mostly trains classic machine learning separators (such as support vector machines and decision trees) to extract human engineering-based features (such as texture and shape). So far, these algorithms have been successful applied to various medical applications such as liver [17], thyroid [18], and bladder cancer [19, 20]. However, although early detection of prostate cancer is very important, there is still little research on computer-aided detection of prostate cancer. Moreover, due to the number of patients used and the experimental techniques adopted, many of them are limited in scope, or the results cannot be considered representative. Llobet et al. [3] proposed a method of transrectal ultrasound image analysis for computer-aided diagnosis of prostate cancer. The best classification result of this research method reached a 61.6% area under the ROC curve. However, the recognition ability of urologists using the computer-aided system is only slightly improved compared with that of experts who do not use the system. Huynen et al. [21] developed a system for automatic analysis and interpretation of prostate ultrasound images. The system extracts five parameters of the cooccurrence matrix from ultrasound images to classify benign and cancerous prostate tissues. The sensitivity and specificity of this method are 80% and 88%, respectively, with good results. Kratzik et al. [22] published a study on prostate testing. The study used texture feature analysis to obtain good results (specificity and sensitivity both exceed 80%) but did not specify how to evaluate. Han et al. [23] proposed a prostate cancer detection method, which uses multiresolution autocorrelation texture features and clinical features. The method can effectively detect cancer tissues with a specificity of about 90% and a sensitivity of about 92%. However, this method is only applicable to similar databases. If other database data are used, such high sensitivity and specificity may not be achieved. Glotsos et al. [24] established a computer-aided diagnosis system based on texture analysis of transrectal ultrasound images. The system extracts 23 texture features from the region of interest in each image and uses exhaustive search (combining up to 5 features) and omission method to select and train the features of the classifier. In terms of overall system performance, the best classification accuracy rates for identifying normal, infectious, and cancer cases are 89.5%, 79.6%, and 82.7%. However, this research method is only suitable for use when data are scarce. Gomez-Ferrer and Arlandis [25] recorded 288 cases of transrectal ultrasound-guided biopsy and extracted three still images from the biopsy area. The texture features of ultrasonic images are obtained by "simple mapping" on grey and spatial grey correlation matrices. Finally, two methods based on nearest neighbour and Markov hidden model are used for classification. The nearest neighbour of the ROC curve is 59.7%, and the classification of Markov hidden model is 61.6%; ROC curve



FIGURE 1: Biopsy tissue was recorded before the examination. In the image, the needle track is visible but has not been inserted into the prostate. Tissue and corresponding ultrasonic texture are not disturbed, and this image is used for image processing and texture analysis.

area of cooccurrence matrix is 60.1% nearest neighbour, and Markov hidden model is 60.0%. To solve the problems of unclear prostate boundary and insufficient data, Zhu et al. [26] proposed a boundary-weighted domain adaptive neural network (BOWDA-Net).

3. Materials and Methods

3.1. Data Procurement. In Zhangzhou Hospital, a transrectal ultrasound-guided prostate biopsy is usually performed for all patients with prostate cancer-related symptoms (such as high PSA value and abnormal DRE results). The inserted transrectal ultrasound probe displays sagittal images of the prostate. When suspicious areas are found, the biopsy needle connected to the probe will be triggered for tissue extraction and later histological analysis. Generally, multiple biopsies will be performed if no particularly suspicious area is found. According to the guidance of ultrasound doctors, our experimental data are mainly pictures before biopsy (Figure 1). Because histology can only be determined from the resected tissue, the puncture location must be accurately known. To achieve this, we used the second image, which was recorded before the biopsy needle was retracted from the gland (Figure 2). There is a white needle track in each image. In each biopsy, we define a point for the first needle where the probe appears in front of the prostate and define a second point for the position where the probe is inserted into the prostate about two scales from the first point. We define a rectangle based on these two points (as shown in Figure 2). Since there is no obvious patient movement during the biopsy, pixels in the previous image are marked with the defined rectangle. The image of Figure 1 is our experimental data. In Figure 1,

we manually cut a rectangular image at the same position as that of Figure 2, which is the region of interest for our experiment.

Histological analysis can indicate whether the extracted tissue has prostate cancer, and if so, its location can also be known. However, in a clinical environment, it is difficult to carry out reliable physical labelling on the extracted tissue and then map the physical labelling to pixel labelling. Therefore, we will use a label definition for pixels in all biopsy areas, which means that some images marked as positive samples may contain normal tissues. Fortunately, however, an image pixel marked negative always corresponds to normal tissue, because histological examination can ensure that the entire biopsy area is free of prostate cancer.

3.2. Method. Figure 3 shows the flow of the proposed classification method. The whole method flow mainly includes four parts: image preprocessing, feature extraction, feature fusion, and classification. The main contribution lies in the use of optical density conversion technology to increase the contrast of the image and reduce its noise. Gaussian Markov random field and local binarization are used to extract the two texture features of the image, and then the two features are linearly combined. Finally, the fused features are put into SVM classifier for classification.

3.2.1. Optical Density Image Processing Technology. Removing noise from original images is still a challenging research topic in image processing. Generally speaking, there is no commonly used noise reduction enhancement method. Usually, the appropriate noise reduction method is selected according to the image characteristics. Limited by the

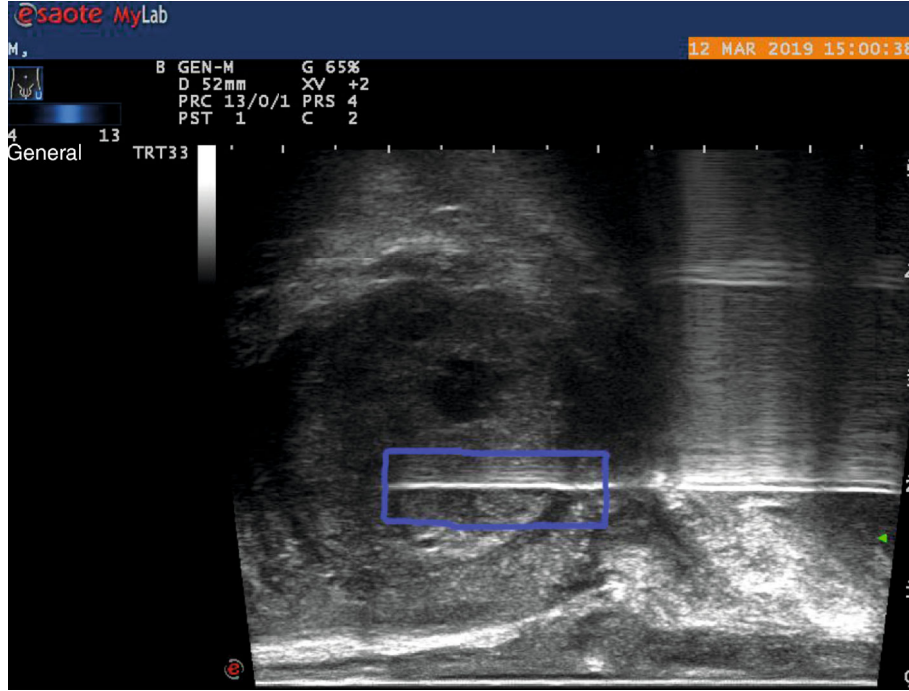


FIGURE 2: Biopsy needle track can still be seen in the prostate gland. In this picture, the puncture position is determined. The extracted tissue is analyzed by a pathologist, and the puncture position determines the analysis position in a clean image (Figure 1).

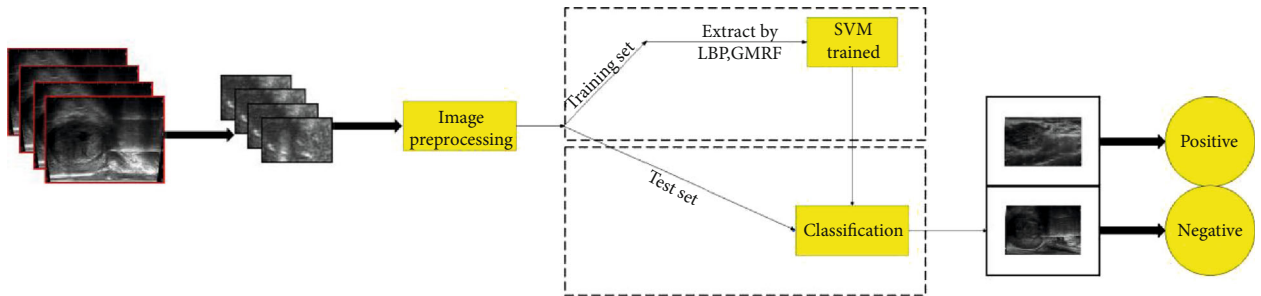


FIGURE 3: Method flow chart.

principle of ultrasonic imaging and the hardware itself, the quality of ultrasonic images is not satisfactory. The main manifestations are of low contrast and speckle noise. Therefore, this paper uses an optical density conversion technology [27] to dry the selected region of interest and enhance the contrast and to make the details of the image clearer and more obvious, which is conducive to the subsequent analysis and processing of the image. The optical density transformation for each pixel (i, j) of an object region is defined as follows:

$$OD_{ij} = \log \left(\frac{I_{ij}}{I_o} \right), \quad (1)$$

where I_{ij} is the intensity value of pixel, and I_o is the average intensity. In this method, the intensity of gray image is converted into optical density, and each optical density value is linearly mapped to the image with 8-bit depth

information, so that the optical density image can be obtained. As shown in Figure 4.

3.2.2. Feature Extraction. Texture features can reflect the overall change of grey pixel values in the image, and different tissues have different textures. Therefore, by distinguishing and identifying texture features in transrectal ultrasound prostate images, suspected case samples with similar texture structures to confirmed case samples can be detected, thus providing decision support for doctors. As one of the most widely used and basic image global features, there are many texture feature extraction methods, which are often used in medical ultrasound image analysis: grey level cooccurrence matrices (GLCM) [28], histogram of oriented gradient (HOG) [29], local binary pattern (LBP) [30], etc. Each method has its advantages and disadvantages. In actual use, it is often necessary to select the corresponding feature extraction method according to the practical application requirements.

According to the relevant research results in recent years, different types of texture features are generally

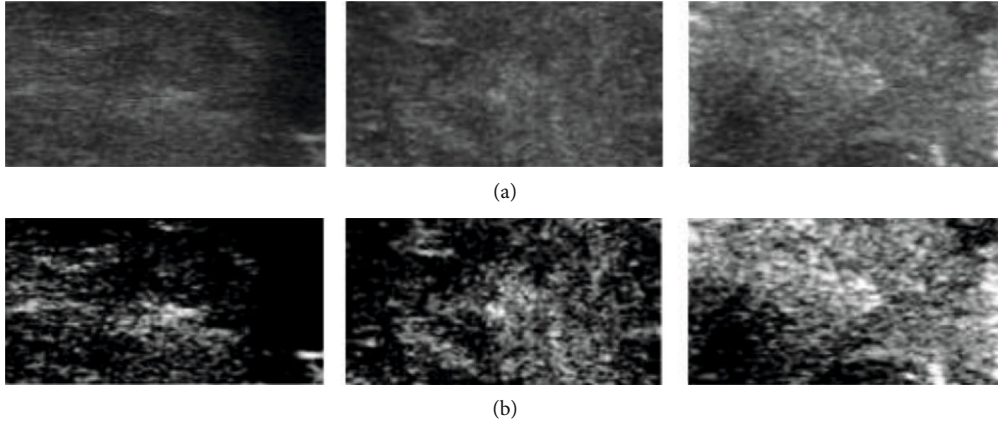


FIGURE 4: (a) Transrectal ultrasound prostate image and (b) optical density image.

complementary. In the image classification task, the combination of different feature extraction methods can often achieve higher classification accuracy than when used alone. Therefore, we use local binarization and Gaussian Markov random field model to extract texture features.

- (1) LBP. LBP (local binary pattern) [30] is an operator used to describe local texture features of images. Its basic idea is defined in the eight fields of pixels (3×3 window). The grey value of the central pixel is taken as the threshold, and the values of the surrounding 8 pixels are compared with it. If the surrounding pixel value is less than the grey value of the central pixel, the pixel value is marked as 0; otherwise, it is marked as 1. In this way, 8 points in the domain size of 3×3 can be compared to generate 8-bit binary numbers (usually converted into a decimal, i.e., LBP code, 256 kinds in total). Each pixel obtains a binary combination, i.e., LBP value of pixel point in the centre of the window, and this value is used to reflect texture information of the region. However, as the image rotates, the pixels in the neighbourhood will recombine, and the LBP value will change. To keep LBP rotation unchanged, Ojala et al. [30] improved the LBP operator. The formula is as follows:

$$\begin{cases} \text{LBP}_{N,R}(a, b) = \sum_{i=0}^{N-1} s(G_i - G_0) \bullet 2^i \\ s(t) = \begin{cases} 1, t \geq 0 \\ 0, \text{else} \end{cases} \end{cases} \quad (2)$$

where R is the radius of the neighbourhood circle, and N is the number of pixels evenly distributed in the neighbourhood. G_i represents N pixels centred on G_0 .

- (2) GMRF. Gaussian Markov random field (GMRF) model [31] is a probability model to describe the image structure and is a better method to describe the texture. It was originally described by Leonard

E. Baum and other authors in a series of statistical papers in the second half of the 1960s. There is a certain correlation between the category of a pixel in an image and the category of pixels in its surrounding areas. This correlation is called Markov correlation. An image can be regarded as a two-dimensional random process, and the distribution of image data can be described by conditional probability. MRF assumes that the pixel value of each pixel in the image depends only on the pixel value of the pixel in its domain. A Markov random field is usually described by the following local conditional probability density (PDF):

$$\begin{aligned} p((m, n) | f(k, l), (k, l) \neq (m, n), (k, l) \in \Lambda) \\ = p(f(m, n) | f(k, l), (k, l) \in N_{(m, n)}). \end{aligned} \quad (3)$$

$N(m, n)$ is the neighbourhood pixel point of the centre pixel. If PDF follows Gaussian distribution, MRF is called GMRF. Its prominent feature is that it introduces structural information through a properly defined neighbourhood system and provides a model commonly used to express the interaction between spatially related random variables. The parameters generated from this model can describe the aggregation characteristics of textures in different directions and forms.

3.2.3. Feature Fusion. Both LBP texture features and GMRF texture features have strong capabilities in feature extraction of transrectal ultrasound prostate images, but they have some limitations in practical application. In the process of feature extraction, LBP only considers the grey values of other surrounding pixels, but does not fully consider the interaction and interdependence between the central pixel and the surrounding pixels. These dependencies can be random, functional, or structural and can be represented by Gaussian Markov random field model. Therefore, this paper first extracts LBP features from transrectal ultrasound prostate images and then calculates the conditional probability density of the extracted LBP feature images.

3.2.4. Classifier Design. Image classification is an important research field and has practical applications in many areas such as pattern recognition, artificial intelligence medicine, and visual analysis. For image classification, we adopt SVM classifier, which is described in detail as follows.

SVM is based on statistical learning technology and is the foundation of modern statistical learning theory. It was proposed by Cortes and Vapnik in 1995 [32]. SVM algorithm is a supervised machine learning algorithm by minimizing empirical errors and maximizing geometric edges to complete pattern classification and regression analysis. It is widely used in statistical classification and regression analysis. It has unique advantages in solving small samples and nonlinear high-dimensional pattern recognition and can be widely applied to machine learning problems such as function fitting. The basic principle of the modified method is to find the fractal hyperplane of the training set n in the sample space and to separate the categories to the maximum extent. Besides, SVM, as a quadratic programming problem, can find a globally unique optimal solution, thus avoiding the occurrence of local minima. The principle and solving process are as follows:

Given a data set:

$$N\{(x_i, y_i) \mid x_i \in R^n, y_i \in \{-1, +1\}, i = 1, \dots, n\}. \quad (4)$$

Then, the discriminant function of SVM is as follows:

$$f(x) = \text{sign} \left(\sum_{i=1}^n A_i y_i K(x, x_i) + B \right), \quad (5)$$

where $K(x, x_i)$ is the kernel function, and n is the number of support vector machines. The kernel function is vital in support vector training. It can effectively overcome the dimension disaster problem. Proper kernel function can improve the prediction accuracy of the classification model. Common kernel functions include Gaussian function, polynomial function, sigmoid function, and linear function. In this paper, input vectors composed of texture features are selected as Gaussian functions. The classification results of SVM are used to distinguish positive samples from negative samples in transrectal ultrasound images.

In this paper, SVM classification data are using a linear hyperplane that separates data into two isolated classes. This hyperplane is calculated using Gaussian kernel function. The number of neighbours in k-nearest neighbour (KNN) [33] is set to 5. The confidence factor of decision tree (DT) [34] is set to 0.25, and the minimum case tree of each leaf is set to 2. Random forest (RF) [35] is using matlab random forest toolbox, with trees selection of 500 and mtry of 61.

4. Experimental Results

4.1. Experimental Data. This research has been approved and reviewed by the local ethics committee, and all relevant topics have been notified with permission. Transrectal ultrasound prostate images used in this experiment were from Zhangzhou Hospital affiliated to Fujian Medical University, with a

TABLE 1: Definition of evaluation index.

Evaluations	Definition
ACC	$(TP + TN)/(TP + TN + FP + FN)$
SEN	$TP/(TP + FN)$
SPEC	$TN/(TN + FP)$

total of 48 cases. All pathological cases were biopsied under ultrasound guidance by experienced pathologists and confirmed histologically. The data collection time is from March 2019 to November 2019, and each patient file contains multiple images. The data are classified according to pathological results. There were 36 cases in training set and 12 cases in test set. Experiments were conducted on prostate diagnosis to distinguish whether transrectal ultrasound images have prostate cancer. Therefore, the negative samples of training data were 18 cases (126 images), and the positive samples were 18 cases (130 images). The remaining 12 cases were used as experimental test sets, of which 6 cases (40 images) were positive samples, and 6 cases (46 images) were negative samples.

4.2. Experimental Setup and Performance Evaluation. The experiment is completed based on Windows10 operating system. The computer hardware is configured as follows: Intel(R) Core(TM) i7-8700 is used for CPU, NVIDIA GeForce GTX-1080Ti is used for GPU, and the video memory is 11G and the memory is 32G. The programming environment is Matlab2017a.

Disease classification results are true positive, true negative, false positive, and false negative. In order to facilitate comparative analysis with the existing methods, we have considered three indicators: accuracy (ACC), sensitivity (SEN), and specificity (SPEC) [36], as shown in Table 1. Among them, TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively, in the classification results. Accuracy is a direct measure of comparison between methods. Sensitivity and specificity describe how diagnostic tests capture the real presence or absence of disease. These evaluation indexes together describe the accuracy and error rate of image classification and recognition methods. Among them, the higher the accuracy, sensitivity, and specificity, the lower the error rate of the method.

4.3. Comparison of Characteristic Combination Performance. In order to test the effectiveness of the combination of Gaussian Markov random field and local binarization, we respectively use a variety of texture features to carry out experiments and compare the accuracy with the proposed method. In all experiments, we use support vector machine to classify. The classification performance of different methods is shown in Table 2. All the values in Table 2 are obtained using our data set. As can be seen from the table, compared with individual features, the classification accuracy of feature fusion has been significantly improved, and other indicators have also been improved to varying degrees, especially the specificity indicators are more obvious. Compared with the classification results of different texture features in

TABLE 2: Classification accuracy with different types of features.

Method	ACC	SEN	SPEC
GLCM [28]	61.63%	67.50%	56.52%
HOG [29]	66.28%	65.00%	67.39%
LBP [30]	60.47%	67.50%	54.35%
GMRF [31]	53.49%	57.50%	50.00%
GLDS [37]	61.63%	62.50%	60.87%
Our method	70.93%	70.00%	71.74%

TABLE 3: Classification performance of all comparison methods.

Method	ACC	SEN	SPEC
KNN [33]	63.95%	57.50%	69.57%
DT [34]	63.96%	55.00%	71.72%
RF [35]	62.78%	62.50%	63.04%
Our method	70.93%	70.00%	71.74%

Table 2, the texture feature fusion classification results proposed in this paper are the best, with the classification accuracy rate reaching 70.93%, sensitivity 70.00%, and specificity 71.74%. Using Gaussian Markov random fields to extract texture features alone does not provide meaningful results for our data set. As can be seen from Table 2, our method has high specificity while maintaining high sensitivity.

4.4. Performance Comparison of the Methods. To test the effectiveness of our approach, we compared our method with the following three ways: (a) KNN classifier [33], (b) decision tree (DT) classifier [34], and (c) random forest (RF) classifier [35]. Specifically, in each experiment, the image is preprocessed by optical density conversion technology, and then the Gaussian Markov random field and local binarization features are extracted and fused. Finally, the above three classifiers are used for classification.

Compared with the classification results of different classifiers in Table 3, the classification results of support vector machine are higher than those of other classifiers, with the classification accuracy rate reaching 70.93%, sensitivity reaching 70.00%, and specificity reaching 71.74%. The second is DT, with a classification accuracy of 63.96%, sensitivity of 55%, and specificity of 71.72%. The classification accuracy of KNN was 63.95%, sensitivity 57.50%, and specificity 69.57%. The classification accuracy of RF was 62.78%, sensitivity 62.50%, and specificity 63.04%. As can be seen from the results shown in Table 3, our proposed method has better performance than other methods.

By comparing the experimental results in Tables 2 and 3, it can be found that LBP+GMRF+SVM proposed in this paper gives full play to the complementarity of texture features. The classification accuracy of this method is 4.65% higher than the highest accuracy of single feature. At the same time, the accuracy of this method is 6.97% higher than highest accuracy of other classifiers.

4.5. Effect Analysis of Image Preprocessing. The experimental data are preprocessed, and the texture features are analyzed by the SVM classifier. As shown in Table 4, preprocessing helps to extract more useful features from images and effectively improves classification accuracy.

4.6. Method Performance Evaluation. Aiming at the problem of small amount of data sets, 5-fold cross-validation is used to verify the effectiveness of the proposed method. That is, the whole data set is divided into five different subsets. Every time one subset is used as the test set and the other four subsets are used as the training set, this process is repeated five times. Finally, the average of five experimental results is calculated to evaluate the performance of the classifier.

By comparing Table 3 with Table 5, it can be found that the error between the results of 5-fold cross-validation and the experimental results of dividing the data set into training set and test set is not great. This verifies the effectiveness of the proposed method.

5. Discussion

Since TRUS cannot reliably identify prostate cancer [8], 6-18 or more puncture biopsies [9] are used to detect cancerous lesions. However, some biopsy samples taken from some male patients will not contain cancer. Also, clinically significant PSA does not necessarily have prostate cancer [4].

Prostate cancer is hypoechoic in ultrasound images [38]. Therefore, TRUS has poor visual interpretation and cannot accurately identify the tumour area. Gomez-Ferrer and Arlandis [25] found only 12.6% hypoechoic lesions in their work, which were found in most (68%) benign tissues. These data confirm the need to try to analyze transrectal ultrasound images with computer assistance. Therefore, this paper proposes an image analysis method based on texture feature fusion.

Sensitivity is an essential criterion for medical diagnosis, especially in the early stage of disease examination. Positive samples of clinical examination should avoid missed diagnosis as much as possible. The texture feature fusion method is used in this experiment. The sensitivity of the fused texture feature is 70.00%, which is better than 67.50% of LBP and 57.50% of GMRF. This shows that there is a correlation between texture description and sensitivity in the image: the more texture descriptions, the more obvious the features of positive lesions. However, for transrectal ultrasound images with small differences between classes, the overall classification performance will also decrease. Experiments show that texture feature fusion has a significant effect on the classification of transrectal ultrasound images.

According to our experimental methods and results, it is quite difficult to develop software for real-time image recognition in the future. Because images will have to be analyzed in real-time and suspicious areas identified, we believe that this may be due to several factors: the method or the disease itself. It may be that prostate cancer and its histological changes have different structures from normal glands. Another problem we are facing is incomplete labelling when conducting such studies because it is almost impossible to

TABLE 4: The classification accuracy (%) based on transrectal ultrasound image preprocessing.

ACC (%)	LBP	HOG	GMRF	GLDS	GLCM	LBP+GMRF
Before preprocessing	55.81	58.14	50.00	60.47	53.49	58.14
After preprocessing	60.47	66.28	53.49	61.63	61.63	70.93

TABLE 5: Average test results of 5-fold cross-validation.

Method	ACC	SEN	SPEC
KNN [33]	62.73%	58.84%	66.81%
DT [34]	60.83%	56.30%	64.99%
RF [35]	64.04%	66.47%	65.14%
SVM [32]	70.11%	68.26%	71.97%

accurately determine the exact location of cancer in transrectal ultrasound images with current technologies. In our research, this annotation was obtained by studying histological analysis and puncture site location. However, this may be affected more or less because the length of the region of interest we extract rarely corresponds to cancer.

6. Conclusion

This paper proposes a texture feature analysis method to improve the classification accuracy of transrectal ultrasound prostate images. Firstly, the transrectal ultrasound image is preprocessed by optical density conversion technology, and then Gaussian Markov random fields and local binarization features are extracted. The two features are linearly combined, and then SVM classifier is used for classification experiments. Finally, several comparative experiments were carried out on the data set we collected, and the experimental results were given and analyzed. The experimental results show that the method has good classification accuracy (70.93%), sensitivity (70%), and specificity (71.74%). This provides a low cost, rapid, and repeatable analysis method for transrectal ultrasound-guided prostate puncture. In the future work, we plan to carry out more effective cooperation with hospitals to obtain more data sets, and then we will improve the proposed method to make it more suitable for the actual needs of the medical field.

Data Availability

The (transrectal ultrasound images) data used to support the findings of this study were supplied by Zhangzhou Affiliated Hospital of Fujian Medical University, China, under license and so cannot be made freely available.

Conflicts of Interest

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

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province, China (Grant No. 2017J01386), in part by

Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (No. ZQN-PY518), in part by Fujian Provincial Big Data Research Institute of Intelligent Manufacturing, and in part by the Quanzhou scientific and technological planning projects (No. 2018C113R and No. 2017G024) and the grants from National Natural Science Foundation of China (No. 61605048), in part by the Subsidized Project for Postgraduates' Innovative Fund in Scientific Research of Huaqiao University under Grant 18014084002.

References

- [1] W. Chen, K. Sun, R. Zheng et al., "Cancer incidence and mortality in China, 2014," *Chinese Journal of Cancer Research*, vol. 30, no. 1, pp. 1–12, 2018.
- [2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [3] R. Llobet, J. Perezcortes, A. Toselli, and A. Juan, "Computer-aided detection of prostate cancer," *International Journal of Medical Informatics*, vol. 76, no. 7, pp. 547–556, 2007.
- [4] S. Loeb and W. J. Catalona, "Prostate-specific antigen in clinical practice," *Cancer Letters*, vol. 249, no. 1, pp. 30–39, 2007.
- [5] G. S. Sandhu and G. L. Andriole, "Overdiagnosis of prostate cancer," *Journal of the National Cancer Institute Monographs*, vol. 2012, no. 45, pp. 146–151, 2012.
- [6] J. P. Rادتke, C. Schwab, M. B. Wolf et al., "Multiparametric magnetic resonance imaging (MRI) and MRI-transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen," *European Urology*, vol. 70, no. 5, pp. 846–853, 2016.
- [7] G. J. Kelloff, P. Choyke, D. S. Coffey, and Prostate Cancer Imaging Working Group, "Challenges in clinical prostate cancer: role of imaging," *American Journal of Roentgenology*, vol. 192, no. 6, pp. 1455–1470, 2009.
- [8] H. S. Han, H. J. Lu, L. Zhang, H. Q. Zhong, and S. B. Wang, "The value of transrectal real-time tissue elastography combined with multi parameter magnetic resonance imaging in prostate biopsy," *Chinese Medical Ultrasonic Magazine*, vol. 14, no. 9, pp. 706–710, 2017.
- [9] V. Scattoni, A. Russo, E. di Trapani, U. Capitanio, G. la Croce, and F. Montorsi, "Repeated biopsy in the detection of prostate cancer: when and how many cores," *Archivio Italiano di Urologia e Andrologia*, vol. 86, no. 4, pp. 311–313, 2014.
- [10] A. Anastasiadis, L. Zapala, E. Cordeiro, A. Antoniewicz, and T. D. Reijke, "Complications of prostate biopsy," *Expert Review Anti Infective Therapy*, vol. 13, no. 7, pp. 829–837, 2014.
- [11] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes, "Texture analysis of medical images," *Clinical Radiology*, vol. 59, no. 12, pp. 1061–1069, 2004.
- [12] S. Alobaidli, S. McQuaid, C. South, V. Prakash, P. Evans, and A. Nisbet, "The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment

- planning,” *The British Institute of Radiology*, vol. 87, no. 1042, article 20140369, 2014.
- [13] T. McInerney and D. Terzopoulos, “Deformable models in medical image analysis: a survey,” *Medical Image Analysis*, vol. 1, no. 2, pp. 91–108, 1996.
- [14] B. Ganeshan and K. A. Miles, “Quantifying tumour heterogeneity with CT,” *Cancer Imaging*, vol. 13, no. 1, pp. 140–149, 2013.
- [15] Z. Li, L. Yu, X. Wang et al., “Diagnostic performance of mammographic texture analysis in the differential diagnosis of benign and malignant breast tumors,” *Clinical Breast Cancer*, vol. 18, no. 4, pp. e621–e627, 2018.
- [16] R. Xu, S. Kido, K. Suga et al., “Texture analysis on 18F-FDG PET/CT images to differentiate malignant and benign bone and soft-tissue lesions,” *Annals of Nuclear Medicine*, vol. 28, no. 9, pp. 926–935, 2014.
- [17] L. Saba, N. Dey, A. S. Ashour et al., “Automated stratification of liver disease in ultrasound: an online accurate feature classification paradigm,” *Computer Methods and Programs in Biomedicine*, vol. 130, pp. 118–134, 2016.
- [18] U. R. Acharya, G. Swapna, S. V. Sree et al., “A review on ultrasound-based thyroid cancer tissue characterization and automated classification,” *Technology in Cancer Research and Treatment*, vol. 13, no. 4, pp. 289–301, 2014.
- [19] Q. Zhu, B. Du, P. Yan, H. Lu, and L. Zhang, “Shape prior constrained PSO model for bladder wall MRI segmentation,” *Neurocomputing*, vol. 294, pp. 19–28, 2018.
- [20] X. Li, B. Du, C. Xu, Y. Zhang, L. Zhang, and D. Tao, “Robust learning with imperfect privileged information,” *Artificial Intelligence*, vol. 282, article 103246, 2020.
- [21] A. L. Huynen, R. J. B. Giesen, J. J. M. C. H. de la Rosette, R. G. Aarnink, F. M. J. Debruyne, and H. Wijkstra, “Analysis of ultrasonographic prostate images for the detection of prostatic carcinoma: the automated urologic diagnostic expert system,” *Ultrasound in Medicine and Biology*, vol. 20, no. 1, pp. 1–10, 1994.
- [22] C. Kratzik, E. Schuster, A. Hainz, W. Kuber, and G. Lunglmayr, “Texture analysis — a new method of differentiating prostatic carcinoma from prostatic hypertrophy,” *Urological Research*, vol. 16, no. 5, pp. 395–397, 1988.
- [23] S. M. Han, H. J. Lee, and J. Y. Choi, “Computer-aided prostate cancer detection using texture features and clinical features in ultrasound image,” *Journal of Digital Imaging*, vol. 21, no. S1, pp. 121–133, 2008.
- [24] D. Glotsos, I. Kalatzis, P. Theocharakis et al., “A multi-classifier system for the characterization of normal, infectious, and cancerous prostate tissues employing transrectal ultrasound images,” *Computer Methods and Programs in Biomedicine*, vol. 97, no. 1, pp. 53–61, 2010.
- [25] A. Gómez-Ferrer and S. Arlandis, “Computer-aided analysis of transrectal ultrasound images of the prostate,” *Actas Urológicas Españolas (English Edition)*, vol. 35, no. 7, pp. 404–413, 2011.
- [26] Q. Zhu, B. Du, and P. Yan, “Boundary-weighted domain adaptive neural network for prostate MR image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 3, pp. 753–763, 2020.
- [27] S.-C. Tai, Z.-S. Chen, and W.-T. Tsai, “An automatic mass detection system in mammograms based on complex texture features,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 618–627, 2014.
- [28] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *Studies in Media and Communication*, vol. SMC-3, no. 6, pp. 610–621, 1973.
- [29] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, pp. 886–893, San Diego, CA, USA, 2005.
- [30] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [31] D. Ming, J. C. Luo, and Z. F. Shen, “Research on region partition in high resolution remote sensing image based on GMRF-SVM,” *Science of Surveying and Mapping*, vol. 34, no. 2, pp. 33–37, 2009.
- [32] C. Cortes and V. Vapnik, “Support vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [33] R. Min, D. A. Stanley, Z. Yuan, A. Bonner, and Z. Zhang, “A deep non-linear feature mapping for large-margin knn classification,” in *2009 Ninth IEEE International Conference on Data Mining*, pp. 357–366, Miami, FL, USA, 2009.
- [34] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Transactions on Systems Man and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
- [35] L. Yu and Pavlov, “Random forests,” *Karelian Centre Russian Academy Petrozavodsk*, vol. 45, no. 1, pp. 5–32, 1997.
- [36] P. Qiu, “The statistical evaluation of medical tests for classification and prediction,” *Journal of the American Statistical Association*, vol. 100, no. 470, p. 705, 2005.
- [37] H. Liu, Y. S. Zhang, Y. H. Zhang, and H. E. Zi-Fen, “Texture feature extraction of flame image based on gray difference statistics,” *Control Engineering of China*, vol. 20, no. 2, pp. 213–218, 2013.
- [38] K. Shinohara, T. M. Wheeler, and P. T. Scardino, “The appearance of prostate cancer on transrectal ultrasonography: correlation of imaging and pathological examinations,” *The Journal of Urology*, vol. 142, no. 1, pp. 76–82, 1989.

Research Article

Transcriptome Analysis Identifies Novel Prognostic Genes in Osteosarcoma

Junfeng Chen,¹ Xiaojun Guo,¹ Guangjun Zeng,¹ Jianhua Liu,¹ and Bin Zhao^{1,2} 

¹Department of Orthopedics, Tianmen First People's Hospital, Tianmen, Hubei 431700, China

²Department of Foot and Ankle Surgery, Wuhan Fourth Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China

Correspondence should be addressed to Bin Zhao; ganqiaogaox@163.com

Received 25 June 2020; Accepted 22 July 2020; Published 6 October 2020

Guest Editor: Tao Huang

Copyright © 2020 Junfeng 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.

Osteosarcoma (OS), a malignant primary bone tumor often seen in young adults, is highly aggressive. The improvements in high-throughput technologies have accelerated the identification of various prognostic biomarkers for cancer survival prediction. However, only few studies focus on the prediction of prognosis in OS patients using gene expression data due to small sample size and the lack of public datasets. In the present study, the RNA-seq data of 82 OS samples, along with their clinical information, were collected from the TARGET database. To identify the prognostic genes for the OS survival prediction, we selected the top 50 genes of contribution as the initial candidate genes of the prognostic risk model, which were ranked by random forest model, and found that the prognostic model with five predictors including *CD180*, *MYC*, *PROSER2*, *DNAI1*, and *FATE1* was the optimal multivariable Cox regression model. Moreover, based on a multivariable Cox regression model, we also developed a scoring method and stratified the OS patients into groups of different risks. The stratification for OS patients in the validation set further demonstrated that our model has a robust performance. In addition, we also investigated the biological function of differentially expressed genes between two risk groups and found that those genes were mainly involved with biological pathways and processes regarding immunity. In summary, the identification of novel prognostic biomarkers in OS would greatly assist the prediction of OS survival and development of molecularly targeted therapies, which in turn benefit patients' survival.

1. Introduction

Osteosarcoma (OS), a malignant primary bone tumor often seen in young adults, is highly aggressive [1]. According to previous studies, OS patients without metastatic diseases present a survival of 70%, yet evidence suggests that metastases that take place at early stages result in worse prognosis [2]. OS can be further categorized into different groups as intramedullary and surface subtypes according to their histologic characteristics and is considered to be associated with multifactorial causes, and both genetic and environmental influences seem to have an impact on this disease [3]. However, for the majority of OS patients, its etiology still remains veiled. Patients' physique [4–6] and their genetic background [7], along with hormone secretion that could affect skeletal development [8], are all risk factors for OS. Currently,

patients with OS mostly receive surgery and chemotherapy, which brings dramatic improvement in their long-term survival, yet accurate prognosis prediction is still required in making therapeutic decisions [9]. However, the only about 15–17% OS patients treated with only surgery could survive [10, 11]. In the early 1970s, the adjuvant chemotherapy was introduced and applied in the treatment of OS patients without metastatic disease [12]. Combined with surgery resection, current combinational chemotherapy could cure ~70% of OS patients. However, the five-year overall survival for patients with metastasis or relapse was still only about 20% [11, 13], voicing an urgent call for new therapies aimed at these patients.

With the advances in sequencing technologies, such as microarray, next-generation sequencing, and proteomics mass spectrum, the prognostic biomarkers for cancer

survival prediction have been proposed by several studies [14–16]. Mutations in *TP53*, *RBI*, *CDKN2A*, *PTEN*, and *YAP1* [17] have been identified and widely observed using whole-genome sequencing (WGS) or whole-exome sequencing (WES) in OS patients, which greatly improved our understanding of the genomic landscape of OS. With next-generation sequencing, the identification of novel biomarkers becomes possible, which can not only broaden our insights into the pathogenic mechanism of OS but also provide the resource to build machine learning models to predict the prognosis of OS patients. For instance, *KRT5*, *HIPK2*, *MAP3K5*, and *CD5* were identified to serve as prognostic factors of osteosarcoma patients [18]. Moreover, risk predictive models based on one eight-gene [19] and one two-gene [20] (*PML-EPB41*) signatures have been built to predict overall survival of patients with osteosarcoma. However, only few studies focus on the prediction of prognosis in OS patients due to a small sample size using gene expression data and the lack of public datasets. In this study, we collected 82 OS samples with RNA-seq data and corresponding clinical data from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) database, selected prognostic genes, and built a prognostic risk model to assess and predict the overall survival of OS. The identification of novel prognostic biomarkers in OS would greatly assist the prediction of OS survival and the development of molecularly targeted therapies, which in turn benefit patients' survival.

2. Material and Methods

2.1. Data Sources. We obtained osteosarcoma RNA-seq data (TPM) and matched clinical data of OS patients from the TARGET database (<https://ocg.cancer.gov/programs/target>). A total of 82 patients from this dataset were constructed as a training set of our prognostic risk model. The dataset of GSE21257 [21] for further validation consisted of 34 osteosarcoma patients.

2.2. Screening of Genes for the Prognostic Risk Model of Osteosarcoma. First, a list of genes yielding TPM > 0.1 in more than half of the total samples was chosen for feature selection. Based on the clinical information and expression profiles of each patient, genes significantly associated with patients' survival were obtained by performing Cox regression analysis with the R Survival package. We further narrowed down this gene list based on the differential levels of prognostic outcomes, and genes whose *P* value is less than 0.01 were selected. Subsequently, these genes were ranked by the random forest algorithm in the R package randomForestSRC based on their relative contribution. Consequently, the top 50 genes were identified as the candidate genes to construct the risk model of osteosarcoma prognosis.

2.3. Model Construction and Evaluation. Utilizing the expression profiles of candidate prognostic genes and the survival data of patients, we built a prognosis risk model for OS using the multivariate Cox regression as previous studies described

[22], and a list of genes that contributed significantly to this model was obtained, which consisted of our final candidates. We established a scoring formula for these finalized candidate genes to evaluate the risks for OS patients and used the median score to divide them into two subgroups, namely, high-risk and low-risk groups. Kaplan-Meier survival curves were plotted, respectively, for each group, and the differences in their survival were further assessed by the log-rank test.

2.4. Validation of the Prognostic Risk Model by an Independent Dataset. Validation dataset consisted of 34 osteosarcoma patients obtained from the GSE21257 dataset [21]. The prognostic risk scoring formula obtained from the training set was applied to evaluate the risk for each patient according to the expression of finalized candidate genes in each sample, accordingly. These patients were then labeled as those of high risk and of low risk based on the scores assigned to them, and their prognostic difference was further analyzed.

2.5. Gene Set Enrichment Analysis. Our prognostic risk model divided osteosarcoma patients into two categories, termed as high- and low-risk groups, and then differentially expressed genes (DEGs) between two groups were selected with two thresholds at $|\log_2(\text{fold})| > 1$ and *P* value < 0.05. Gene Ontology- (GO-) based enrichment analyses of these significantly differentially expressed genes were carried out in R with package clusterProfiler, as described in previous studies [23–25].

3. Results

3.1. Screening of Prognostic Genes for OS Survival Prediction. The gene expression data and corresponding clinical data of 82 patients were retrieved from the TARGET database. A total of 16,034 genes were introduced as variables in our prognostic risk model under the condition that these genes exhibited TPM > 0.1 in more than half of the total samples. Univariable Cox regression was performed, and Kaplan-Meier curves were plotted accordingly on all these genes, out of which 50 genes significantly related to the patient's survival were obtained (*P* values < 0.01). Subsequently, the contribution of these genes was ranked by the random forest algorithm, and the top 50 genes were selected as the initial candidate genes for the construction of the prognostic risk model.

Using those 50 genes, the prognostic risk model of osteosarcoma was developed, based on a multivariable cox regression. Among them, five genes, including *CD180*, *MYC*, *PROSER2*, *DNAI1*, and *FATE1*, with significant contribution to the model were selected as the candidate genes in the optimal prognostic risk model of osteosarcoma. Notably, the low expression of *CD180* and high expressions of *MYC*, *PROSER2*, *DNAI1*, and *FATE1* were identified to result in worse prognostic outcomes in OS (Figures 1(a)–1(e)). These results indicated that the five prognostic genes were highly associated with the OS prognosis.

3.2. Construction of Multivariable Cox Model Using Five Prognostic Genes. Given the 5 genes, a multivariable Cox

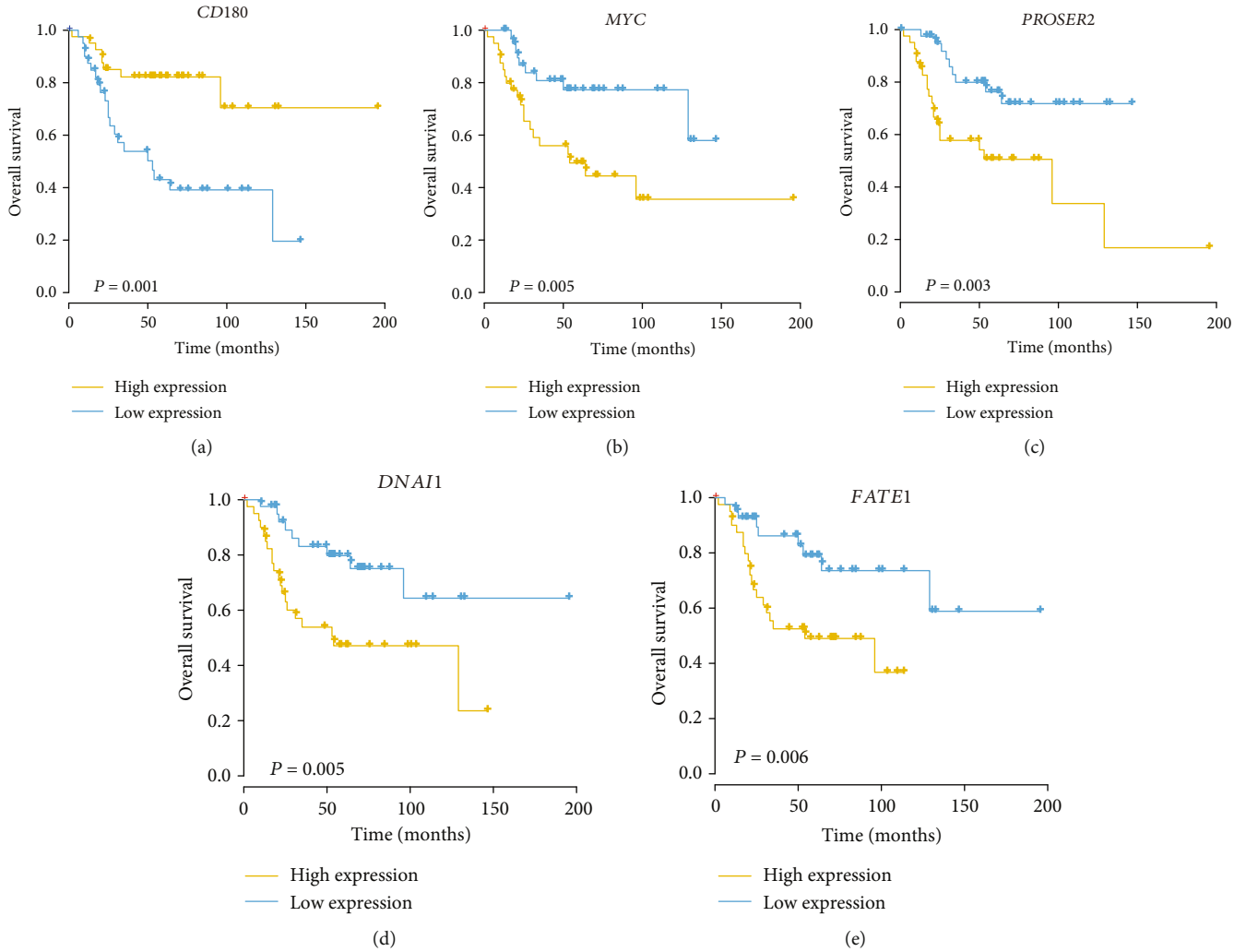


FIGURE 1: The Kaplan-Meier curves for samples stratified based on the expression levels of *CD180*, *MYC*, *PROSER2*, *DNAI1*, and *FATE1*, respectively. (a–e) Survival of OS patients stratified by expression of *CD180*, *MYC*, *PROSER2*, *DNAI1*, and *FATE1*, respectively, depicted by Kaplan-Meier plots.

TABLE 1: The summary of prognostic values for the five genes in univariable and multivariable Cox regression models.

Features	Univariable cox regression		Multivariable cox regression	
	Hazard ratio	P value	Hazard ratio	P value
<i>CD180</i>	0.43	2.19E-03	0.44	4.98E-03
<i>MYC</i>	1.01	9.60E-05	1.01	1.18E-05
<i>PROSER2</i>	1.10	6.52E-06	1.09	4.60E-03
<i>DNAI1</i>	1.53	8.54E-06	1.42	2.19E-03
<i>FATE1</i>	6.08	2.62E-05	7.05	4.05E-06

model was built to evaluate the risk of the OS patients. Specifically, the five genes showed significant association with the OS overall survival in both univariable and multivariable Cox models (Table 1). All 82 patients in the TARGET-osteosarcoma dataset were used as a training set, and we then estimated their risk scores according to

the model. They were divided into high- and low-risk groups by the median risk score. Consistently, the overall survival of the high-risk group was significantly shorter than that of the low-risk group (Figure 2(a)). From this stratification, we noticed that the deceased patients of the high-risk group were found to be more than those in the low-risk group (Figure 2(b), P value < 0.05). In accordance with the Cox regression analyses, *CD180* was downregulated and another four genes were upregulated in the high-risk group (Figure 2(b)). These results indicated that the five genes acquired good fitting effect on the overall survival in OS.

3.3. Validation for the Prognostic Risk Model of OS. To validate the prognostic value of the five-gene-based Cox model, we collected an independent gene expression dataset with 34 OS samples from Gene Expression Omnibus (GEO) with accession GSE21257. The risk scores for 34 OS samples were estimated using the expression levels of the five

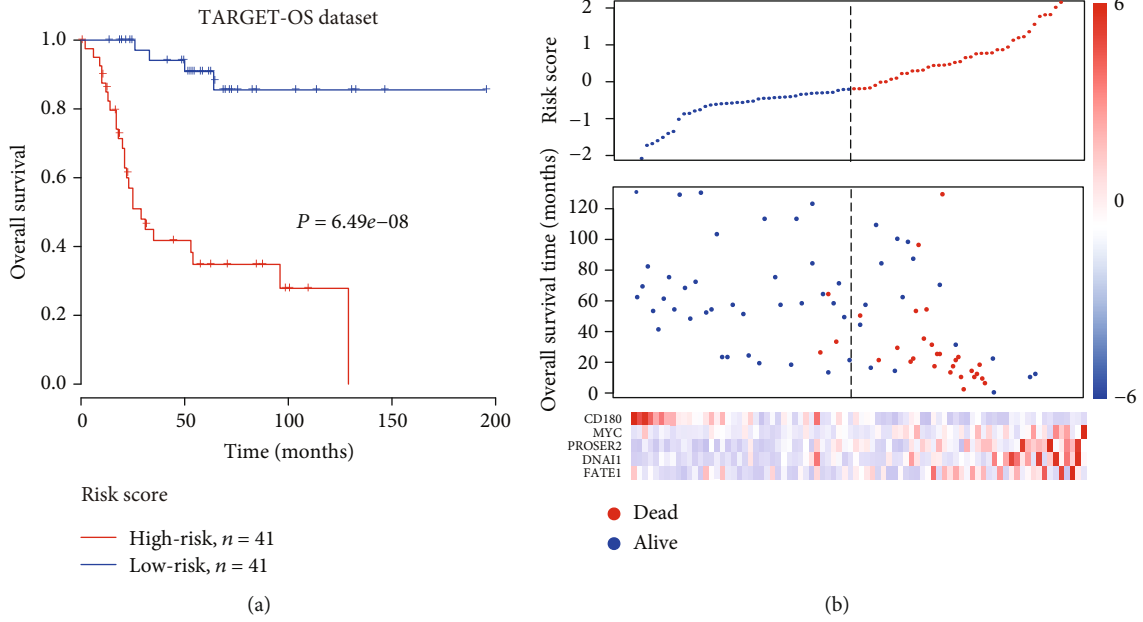


FIGURE 2: The risk stratification of the OS patients by the risk score in the training set (TARGET cohort): (a) Kaplan-Meier curves for patients in high-risk and low-risk groups in the training set stratified by risk scores; (b) the association of the risk scores with the survival time and status and expression levels of five genes. The samples were ranked by the risk scores.

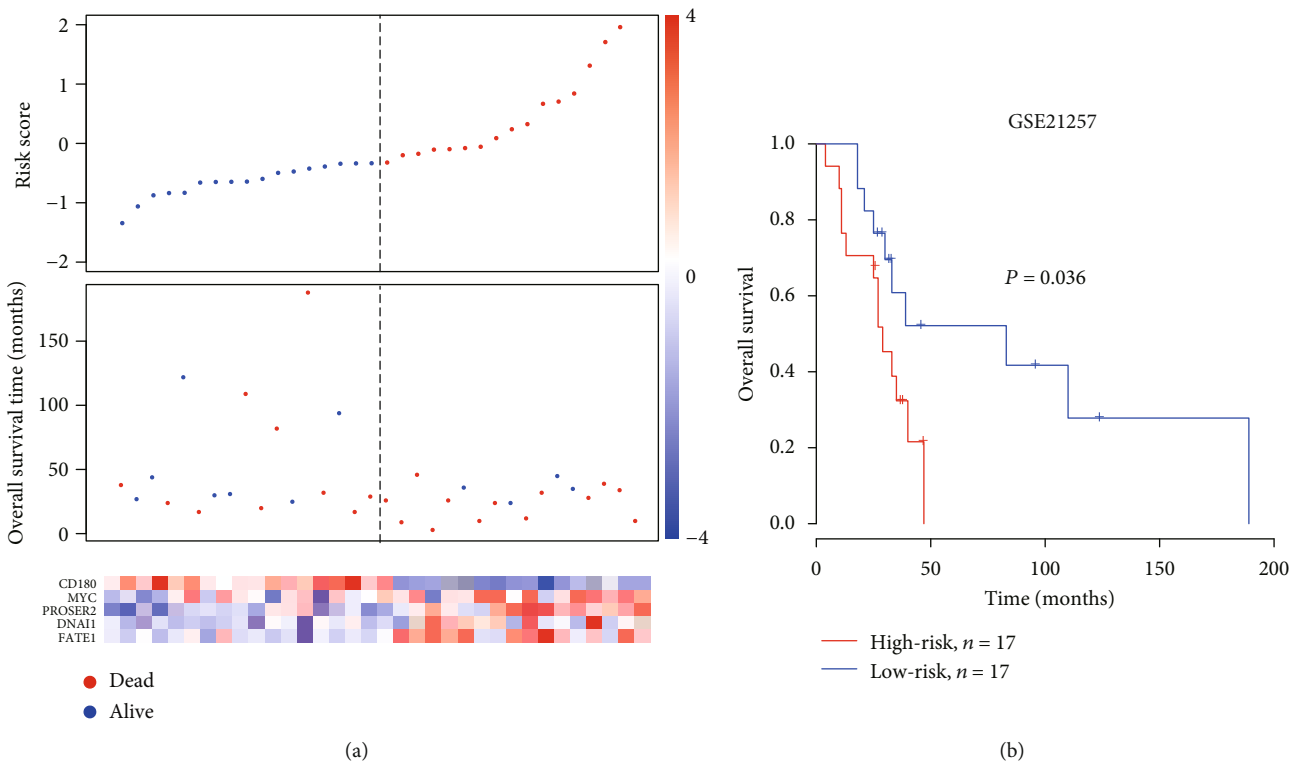


FIGURE 3: Prognostic risk model performance on the validation set (GSE21257 cohort). (a) The association of the risk scores with the survival time and status and expression levels of five genes in the validation set (GSE21257). The samples were ranked by the risk scores. (b) Kaplan-Meier curves for patients in the high-risk and low-risk groups in the validation set stratified by risk scores.

genes in each individual. Similarly, these samples were also assigned into high-risk and low-risk groups by the median score. Consistently, in the high-risk group, we observed a

greater number of deceased patients and shorter overall survival than the low-risk group (P value < 0.05 , Figure 3(a)). The KM curves illustrated that patients of high

TABLE 2: The statistical significance of the risk score in univariable and in multivariable Cox regression models with other clinical parameters.

Features	P value	Univariable cox regression			P value	Multivariable cox regression		
		HR	Lower 95% CI	Upper 95% CI		HR	Lower 95% CI	Upper 95% CI
Risk score	4.42E-12	19.7	4.64	15.6	4.34E-11	12.3	5.85	26.1
Gender (female/male)	0.30	0.68	0.33	1.41	0.17	0.54	0.23	1.29
Race (white/other)	0.23	0.64	0.30	1.34	0.07	0.47	0.21	1.08
Age	0.82	1	1	1	0.82	1	1	1

risk exhibited significantly worse prognosis than those of the low-risk group (P value < 0.05 , Figure 3(b)). Such significant difference in the overall survival time between two risk groups in the validation set suggested that the five-gene-signature Cox model could efficiently predict the prognostic risk in OS.

3.4. The Risk Score Based on the Five-Gene-Based Cox Model Serves as an Independent Prognostic Factor in OS. In order to evaluate the independence of this risk score, we conducted both univariable and multivariable Cox regression, using the risk score and three other clinical variables. In both univariable and multivariable Cox analyses, risk score was the only variable that correlated with the survival time (Table 2). Moreover, we found that white OS patients might have a lower risk than other ethnic groups in the multivariable Cox model with a lower statistical significance (P value < 0.1). These results suggested that risk score could function as an independent prognostic indicator in OS.

3.5. Biological Differences between the Two Risk Groups. The differential expression analysis was performed for patients from the TARGET OS dataset, where patients were labeled as of high- and low-risk ones. A total of 351 significant differentially expressed genes were identified (thresholds: $|\log_2(\text{fold change})| > 1$ and P value < 0.05), of which, compared with the low-risk group, 138 and 213 genes exhibited increased and decreased expression in the high-risk group (Figure 4(a)), respectively.

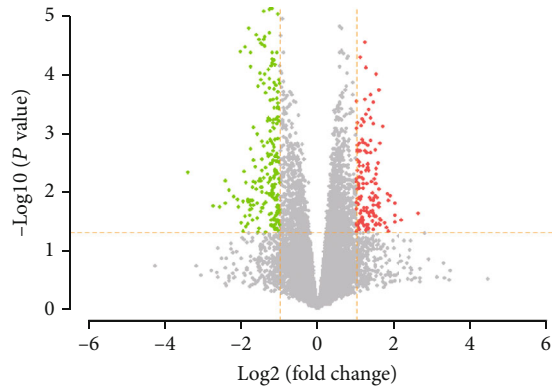
Gene Ontology- (GO-) based gene enrichment analysis revealed that immune-related GO terms, including leukocyte cell-cell adhesion and its T cell activation and its regulation, positive regulation of leukocyte cell-cell adhesion, and positive regulation of T cell activation, were highly enriched by these DEGs (Figure 4(b)), suggesting that the differed immune microenvironment of patients in high-risk and low-risk groups may be responsible for the difference in their prognostic outcomes. Further analysis revealed that major histocompatibility complex (MHC) class II genes were downregulated in the high-risk group (Figure 4(c)), suggesting that the lack of antigen processing and presentation might be associated with reduced immunity against tumor cells, thereby resulting in worse prognosis in OS. Collectively, the results suggested that the immune microenvironment of patients with osteosarcoma plays an essential role in OS patients' prognosis.

4. Discussion

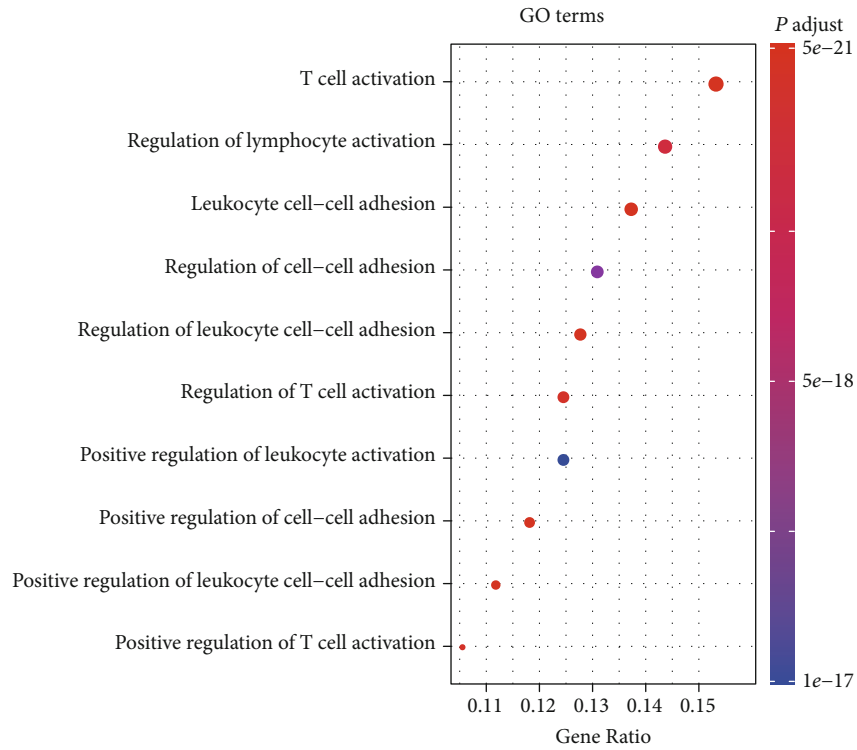
The molecular basis behind OS tumorigenesis, progression, and metastasis has attracted growing attention. Though extensive researches have demonstrated the biological function of certain genes in OS, there is still a lack of effective biomarkers for OS survival prediction. Meanwhile, the prediction and comparison of the OS patients with different clinical outcomes could help clinicians make improvements in diagnostic or therapeutic strategies.

In the present study, 82 OS samples with RNA-seq data and matched clinical data were collected from the TARGET database. To identify the prognostic genes for the OS survival prediction, we selected the top 50 genes of contribution as the initial candidate genes of the prognostic risk model, which were ranked by the random forest model. Multivariable Cox regression analysis suggested that the prognostic model with five predictors including CD180, MYC, PROSER2, DNAIL1, and FATE1 was the optimal multivariable Cox regression model. Among the five prognostic genes, only CD180, which could lead to NF-kappa-B activation [26], was negatively correlated with the OS survival. As we know, CD180 is a cell surface molecule of lymphocytes, and its high expression may indicate the high anticancer activity of lymphocytes, thereby suppressing the growth of OS tumors. CD180, as well as CCR2, has been identified as robust pharmacodynamic tumor and blood biomarkers for clinical use with BRD4/BET inhibitors [27]. In addition to DNAIL1, another three genes, MYC [28], PROSER2 [29], and FATE1 [30], have been reported to be associated with several cancers.

Moreover, we also stratified the OS patients into high-risk and low-risk groups according to the risk score estimated by the multivariable Cox regression model. The stratification for OS patients in the validation set based on risk scores predicted by the multivariable Cox regression model further demonstrated that our model was significant and robust (log-rank test, $P < 0.05$). It should be noted that only 34 primary tissues were used in the validation set, which might be a major limitation for the five-gene-based predictive model. In addition, we also investigated the biological differences between the high-risk and low-risk groups and found that biological processes regarding immunity were highly enriched by the differentially expressed genes between the two risk groups. The expression of MHC II class genes was reduced in high-risk OS samples (Figure 4(c)), suggesting that these samples might lose the abilities of presenting and processing extracellular pathogens. Consistently, MHC class

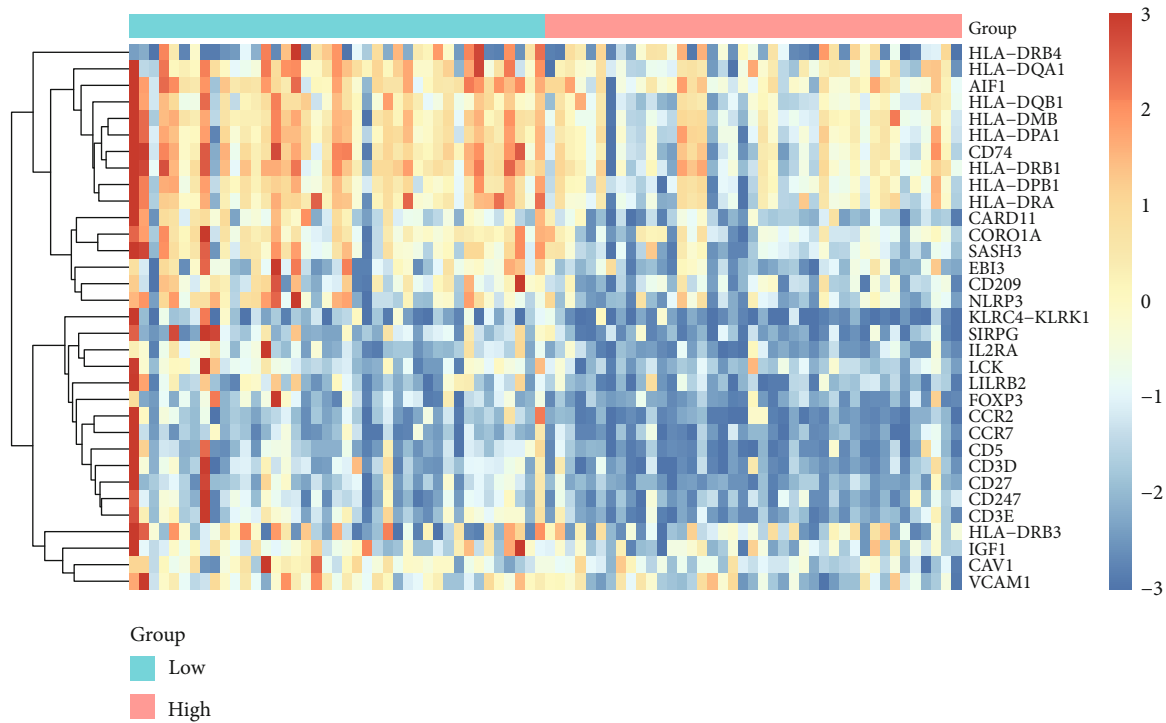


(a)



(b)

FIGURE 4: Continued.



(c)

FIGURE 4: The differentially expressed genes (DEGs) and GO terms enriched by the DEGs between high- and low-risk groups. (a) The volcano plot displays the significantly up- and downregulated genes, which are colored by red and blue. (b) GO terms enriched by the differentially expressed genes. (c) The GO term-related genes downregulated in the high-risk group.

II genes were downregulated and associated with unfavorable outcome in OS [31, 32].

In summary, we established a prognostic risk model of five genes in osteosarcoma, stratified the osteosarcoma samples into high-risk and low-risk groups, and uncovered the underlying molecular mechanism associated with the prognosis, which not only provided some evidence for related researchers but also improved our understanding of OS prognosis.

Data Availability

Osteosarcoma RNA-seq data (TPM) and matched clinical data of OS patients from the TARGET database (<https://ocg.cancer.gov/programs/target>) and the dataset of GSE21257 were used for further validation.

Conflicts of Interest

The authors declare there is no conflict of interest regarding this manuscript.

References

- [1] A. Longhi, C. Errani, M. De Paolis, M. Mercuri, and G. Bacci, "Primary bone osteosarcoma in the pediatric age: state of the art," *Cancer Treatment Reviews*, vol. 32, no. 6, pp. 423–436, 2006.
- [2] M. E. Anderson, "Update on survival in osteosarcoma," *The Orthopedic Clinics of North America*, vol. 47, no. 1, pp. 283–292, 2016.
- [3] D. D. Moore and H. H. Luu, "Osteosarcoma," *Cancer Treatment and Research*, vol. 162, pp. 65–92, 2014.
- [4] L. Mirabello, R. Pfeiffer, G. Murphy et al., "Height at diagnosis and birth-weight as risk factors for osteosarcoma," *Cancer Causes & Control*, vol. 22, no. 6, pp. 899–908, 2011.
- [5] S. Chen, L. Yang, F. Pu et al., "High birth weight increases the risk for bone tumor: a systematic review and meta-analysis," *International Journal of Environmental Research and Public Health*, vol. 12, no. 9, pp. 11178–11195, 2015.
- [6] R. S. Arora, E. Kontopantelis, R. D. Alston, T. Eden, M. Geraci, and J. M. Birch, "Relationship between height at diagnosis and bone tumours in young people: a meta-analysis," *Cancer Causes & Control*, vol. 22, no. 5, pp. 681–688, 2011.
- [7] M. Kansara, M. W. Teng, M. J. Smyth, and D. M. Thomas, "Translational biology of osteosarcoma," *Nature Reviews Cancer*, vol. 14, no. 11, pp. 722–735, 2014.
- [8] J. R. Musselman, T. L. Bergemann, J. A. Ross et al., "Case-parent analysis of variation in pubertal hormone genes and pediatric osteosarcoma: a Children's Oncology Group (COG) study," *International Journal of Molecular Epidemiology and Genetics*, vol. 3, no. 4, pp. 286–293, 2012.
- [9] Z. Wang, M. Tan, G. Chen, Z. Li, and X. Lu, "LncRNA SOX2-OT is a novel prognostic biomarker for osteosarcoma patients and regulates osteosarcoma cells proliferation and motility through modulating SOX2," *IUBMB Life*, vol. 69, no. 11, pp. 867–876, 2017.

- [10] N. M. Bernthal, N. Federman, F. R. Eilber et al., “Long-term results (>25 years) of a randomized, prospective clinical trial evaluating chemotherapy in patients with high-grade, operable osteosarcoma,” *Cancer*, vol. 118, no. 23, pp. 5888–5893, 2012.
- [11] M. P. Link, A. M. Goorin, A. W. Miser et al., “The effect of adjuvant chemotherapy on relapse-free survival in patients with osteosarcoma of the extremity,” *The New England Journal of Medicine*, vol. 314, no. 25, pp. 1600–1606, 1986.
- [12] N. Jaffe, E. Frei 3rd, D. Traggis, and Y. Bishop, “Adjuvant methotrexate and citrovorum-factor treatment of osteogenic sarcoma,” *The New England Journal of Medicine*, vol. 291, no. 19, pp. 994–997, 1974.
- [13] P. A. Meyers, J. H. Healey, A. J. Chou et al., “Addition of pamidronate to chemotherapy for the treatment of osteosarcoma,” *Cancer*, vol. 117, no. 8, pp. 1736–1744, 2011.
- [14] J. Y. Cho, J. Y. Lim, J. H. Cheong et al., “Gene expression signature-based prognostic risk score in gastric cancer,” *Clinical Cancer Research*, vol. 17, no. 7, pp. 1850–1857, 2011.
- [15] H. Tang, S. Wang, G. Xiao et al., “Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies,” *Annals of Oncology*, vol. 28, no. 4, pp. 733–740, 2017.
- [16] J. Tang, D. Kong, Q. Cui et al., “Prognostic genes of breast cancer identified by gene co-expression network analysis,” *Frontiers in Oncology*, vol. 8, p. 374, 2018.
- [17] X. Ma, Y. Liu, Y. Liu et al., “Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours,” *Nature*, vol. 555, no. 7696, pp. 371–376, 2018.
- [18] Y. Li, F. Ge, and S. Wang, “Four genes predict the survival of osteosarcoma patients based on TARGET database,” *Journal of Bioenergetics and Biomembranes*, vol. 52, no. 4, pp. 291–299, 2020.
- [19] G. Wu and M. Zhang, “A novel risk score model based on eight genes and a nomogram for predicting overall survival of patients with osteosarcoma,” *BMC Cancer*, vol. 20, no. 1, p. 456, 2020.
- [20] S. Liu, J. Liu, X. Yu, T. Shen, and Q. Fu, “Identification of a two-gene (PML-EPB41) signature with independent prognostic value in osteosarcoma,” *Frontiers in Oncology*, vol. 9, 2020.
- [21] E. P. Buddingh, M. L. Kuijjer, R. A. Duim et al., “Tumor-infiltrating macrophages are associated with metastasis suppression in high-grade osteosarcoma: a rationale for treatment with macrophage activating agents,” *Clinical Cancer Research*, vol. 17, no. 8, pp. 2110–2119, 2011.
- [22] C. Gu, X. Pan, R. Wang et al., “Analysis of mutational and clinicopathologic characteristics of lung adenocarcinoma with clear cell component,” *Oncotarget*, vol. 7, no. 17, pp. 24596–24603, 2016.
- [23] X. Shi, T. Huang, J. Wang et al., “Next-generation sequencing identifies novel genes with rare variants in total anomalous pulmonary venous connection,” *eBioMedicine*, vol. 38, pp. 217–227, 2018.
- [24] The Gene Ontology Consortium, C. A. Ball, J. A. Blake et al. et al., “The Gene Ontology resource: 20 years and still GOing strong,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D330–D338, 2019.
- [25] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016.
- [26] P. Tieri, A. Termanini, E. Bellavista, S. Salvioli, M. Capri, and C. Franceschi, “Charting the NF- κ B pathway interactome map,” *PLoS One*, vol. 7, no. 3, article e32678, 2012.
- [27] T. C. Yeh, G. O’Connor, P. Petteruti et al., “Identification of CCR2 and CD180 as robust pharmacodynamic tumor and blood biomarkers for clinical use with BRD4/BET inhibitors,” *Clinical Cancer Research*, vol. 23, no. 4, pp. 1025–1035, 2017.
- [28] C. V. Dang, “MYC on the path to cancer,” *Cell*, vol. 149, no. 1, pp. 22–35, 2012.
- [29] L. S. Nguyen, H. G. Kim, J. A. Rosenfeld et al., “Contribution of copy number variants involving nonsense-mediated mRNA decay pathway genes to neuro-developmental disorders,” *Human Molecular Genetics*, vol. 22, no. 9, pp. 1816–1825, 2013.
- [30] K. E. Maxfield, P. J. Taus, K. Corcoran et al., “Comprehensive functional characterization of cancer-testis antigens defines obligate participation in multiple hallmarks of cancer,” *Nature Communications*, vol. 6, no. 1, p. 8840, 2015.
- [31] L. Endo-Munoz, A. Cumming, S. Sommerville, I. Dickinson, and N. A. Saunders, “Osteosarcoma is characterised by reduced expression of markers of osteoclastogenesis and antigen presentation compared with normal bone,” *British Journal of Cancer*, vol. 103, no. 1, pp. 73–81, 2010.
- [32] M. Li, X. Jin, H. Li et al., “Key genes with prognostic values in suppression of osteosarcoma metastasis using comprehensive analysis,” *BMC Cancer*, vol. 20, no. 1, p. 65, 2020.

Research Article

A Medical Decision Support System to Assess Risk Factors for Gastric Cancer Based on Fuzzy Cognitive Map

Seyed Abbas Mahmoodi ¹, Kamal Mirzaie ², Maryam Sadat Mahmoodi,³
and Seyed Mostafa Mahmoudi⁴

¹Department of Computer Engineering, Yazd Science and Research Branch, Islamic Azad University, Yazd, Iran

²Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran

³Department of Electrical and Computer Engineering, Faculty of Sepideh Kashani, Birjand Branch, Technical and Vocational University (TVU), South Khorasan, Iran

⁴Oral and Maxillofacial Pathology Department, School of Dentistry, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

Correspondence should be addressed to Kamal Mirzaie; k.mirzaie@maybodiau.ac.ir

Received 23 March 2020; Revised 19 June 2020; Accepted 14 July 2020; Published 5 October 2020

Guest Editor: Tao Huang

Copyright © 2020 Seyed Abbas Mahmoodi 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.

Gastric cancer (GC), one of the most common cancers around the world, is a multifactorial disease and there are many risk factors for this disease. Assessing the risk of GC is essential for choosing an appropriate healthcare strategy. There have been very few studies conducted on the development of risk assessment systems for GC. This study is aimed at providing a medical decision support system based on soft computing using fuzzy cognitive maps (FCMs) which will help healthcare professionals to decide on an appropriate individual healthcare strategy based on the risk level of the disease. FCMs are considered as one of the strongest artificial intelligence techniques for complex system modeling. In this system, an FCM based on Nonlinear Hebbian Learning (NHL) algorithm is used. The data used in this study are collected from the medical records of 560 patients referring to Imam Reza Hospital in Tabriz City. 27 effective features in gastric cancer were selected using the opinions of three experts. The prediction accuracy of the proposed method is 95.83%. The results show that the proposed method is more accurate than other decision-making algorithms, such as decision trees, Naïve Bayes, and ANN. From the perspective of healthcare professionals, the proposed medical decision support system is simple, comprehensive, and more effective than previous models for assessing the risk of GC and can help them to predict the risk factors for GC in the clinical setting.

1. Introduction

Gastric cancer (GC) which is one of the major cancers around the world with about one million new patients each year is known to be the third cause of cancer deaths [1, 2]. This represents an important public health issue in the world, especially in Central Asian countries, where the incidence of this disease is very high [2]. GC is a multifactorial disease, and its formation is related to various risk factors [3]. Various scientific methods, such as photofluorography and esophago-gastroduodenoscopy, are used to diagnose GC in the early stages and can help reduce the mortality rate of GC with a practical approach [3]. Given that these methods are invasive

and expensive, it is necessary to provide a simple inexpensive and effective tool for the diagnosis of people at risk for GC, which can then be followed by more accurate examinations. Moreover, appropriate prevention efforts can be made to reduce the incidence of this disease.

The initial definitions of the decision support system (DSS) consider it as a system to support decision-makers of the management in the semistructured and unstructured positions and decisions [4]. Accordingly, DSS means helping decision-makers and increasing their ability, not replacing their judgments [4]. Today, the use of DSSs has expanded in a variety of areas, such as management, industry, agriculture, information systems, medicine, and hundreds of other

topics. The medical decision support system (MDSS) is a computer system designed to help physicians or other healthcare professionals in making clinical decisions. Some applications of the medical decision support system are outlined below [5]:

- (i) Preventive care services, for example, screenings for blood pressure and cancer
- (ii) Patient symptom checker
- (iii) Care plan
- (iv) Guide to reducing long hospital stays
- (v) Intelligent health monitoring systems

MDSS contains numerous advantages, of which the most important is to minimize medical failure and make a relatively stable structure for diagnosing and treating the disease, thereby resolving various and conflicting ideas of specialists [5]. Therefore, it is vital to design and implement these models.

FCMs are regarded as soft computing methods that try attempting to act like humans for decision-making and reasoning [6]. In fact, an FCM is an instrument for modeling multifaceted systems, which is attained by integrating neural networks and fuzzy logic [7, 8], and to describe the complex system's performance utilizing concepts. This technique creates a conceptual model where each concept provides a characteristic or a state of a system dynamically interacting with these notions [9]. FCM is a graphical representation of a system structure [10]. According to the artificial intelligence, FCMs are dynamic learning networks; thus, more data to model the problem can help the system with adapting itself and reaching a solution. This conceptual model is not restricted to the exact measurements and quantities. Hence, it is very appropriate for concepts without accurate structures.

FCMs were presented by Kosko as a fuzzy directed graph with sign and feedback loops to illustrate the computational complexity and dependence of a model symbolically and explicitly [11]. In other words, a set of nodes is created by the FCM affecting each other via causal relations. The details and mathematical formulation of this technique are described in Supplementary Materials (available here). Using the benefits of fuzzy systems (if-then rules) and neural networks (teaching and learning), FCM was able to quickly prove its effectiveness in various areas so that we can see its successful presence in politics, economics, engineering, medicine, etc. [12].

In recent years, MDSS using FCM has been developed as one of the main applications of this tool. FCM has emerged as a tool for representing and studying the behavior of systems, and it can deal with complex systems using an argumentative process. This study is aimed at providing an MDSS for assessing the risk of GC using FCM.

In the following, some successful instances of FCM applications regarding decision support systems are provided. Papageorgiou et al. [13] utilized FCM for predicting infectious diseases and infection severity. A novel FCM-based

technique was presented by Amirkhani et al. [14] to screen and isolate UDH from other internal brain lesions. Hence, they examined 86 patients in Shahid Beheshti Hospital in Isfahan City. The pathologist extracted the ten key properties needed to screen these lesions to use them as the key concepts of FCM. The accurateness of the suggested technique was 95.35%. Based on the results, it was indicated that not only the suggested FCM contained a high accuracy level it is also able to preset an acceptable false-negative rate (FNR). A decision support system was proposed by Baena de Moraes Lopes et al. [7] to diagnose the changes in urinary elimination, based on the nursing terminology of North American Nursing Diagnosis Association International (NANDA-I). For 195 cases of urinary incontinence, an FCM model was utilized after the NANDA-I classifications. The high specificity and sensitivity of 0.92 and 0.95, were, respectively, found by the FCM model; however, a low specificity value was provided in the determination of the diagnosis of urge urinary incontinence (0.43) along with a low sensitivity value to overall urinary incontinence (0.42).

Recently, the use of FCM with Hebbian-based learning capabilities has increased. According to [15], a decision-making framework was proposed that can accurately assess the progression of depression symptoms in the elderly people and warn healthcare providers by providing useful information for regulating the patient's treatment. According to [16], a risk management system for familial breast cancer was presented using the NHL-based FCM technique. Data needed for this study were extracted from 40 patients and 18 key features were selected. The results showed that the accuracy is 95%. According to [17], the first specialized diagnostic system for obesity was proposed based on psychological and social characteristics. In this study, a mathematical model based on FCM was presented. According to the proposed model, the effects of different weight-loss treatment methods can be studied.

No certain reason exists for GC. The cause-effect associations are not systematically investigated and understood so far between the integrated impacts of the multiple risk factors on the probability of developing GC. Even the ideas of radiologists and oncologists are greatly subjective in this regard. In such instances, it is considered to use an FCM as a human-friendly and transparent clinical support instrument to determine the cause-effect associations between the factors and the subjectivity can be remarkably eliminated by the degrees of its effects on the risk level. The present work is mainly focused on developing a clinical decision-making instrument in terms of an FCM to evaluate GC risk.

2. Methods

2.1. FCM Model for GC Risk Factors. Addressing GC is a complex process that needs to understand the various parameters, risk factors, and symptoms to make the right decision and assessment. This study assesses the risk of GC by providing a medical decision-making system. The design of this decision-making system is based on a proposed model of FCM, which is presented below. Designing and developing a suitable FCM require human knowledge to describe a

decision support system. In this study, GC specialists are used for the development of the FCM model. The development of the FCM model is divided into three main steps, which is briefly summarized:

- (1) Identify concepts
- (2) Determine the relationships between concepts and initial weights
- (3) Weighting

First, the experts individually identify the factors that contribute to GC. In the following, common concepts among specialists are selected as model nodes. The second step is to identify the relationships between concepts. To this end, experts define the interactions between concepts with respect to fuzzy variables. To do so, determine the relationship and the direction of the relationship (if any). The amounts of these effects are expressed as very low, low, medium, high, and very high. Finally, the linguistic variables expressed by the experts are integrated. Using the SUM technique, these values are aggregated and the total linguistic weight is generated by the “centric” defuzzification method and converted to a numerical value. The corresponding weight matrix is then constructed. Choosing a learning algorithm to teach initial weights is the third step of this method. The purpose of a learning algorithm, setting the initial weight, is the same way as neural networks to improve the modeling FCM.

To better understand, these steps were used step by step to develop an FCM model for GC. For this purpose, the opinions of three specialists were used. In the first phase of the research presented in this article, information on GC risk factors was collected from medical sources, pathologists, and informal sources [18–48]. The collected knowledge was transformed into a well-structured questionnaire and presented to three experts. The questionnaire includes risk factors associated with GC. According to three experts, 27 common features were identified as the major risk factors for end-stage GC. To better understand, we used the mentioned process step by step to develop an FCM model for GC.

Risk factors for gastric cancer may be categorized into four groups (personal features, systemic conditions, stomach condition, and diet food), each of which includes several risk factors. The final features are presented in Figure 1, and their explanations are given in Table 1.

In the second phase, first, the sign for the relationship between the two concepts is determined, and finally, the numerical values of the two concepts are calculated. Five membership functions were used for this purpose. Consider the following example.

- 1st specialist:* C4 has a great impact on C27.
2nd specialist: C4 has a moderate impact on C27.
3rd specialist: C4 has a great impact on C27.

Using the SUM method, the above three linguistic weights (high, very high, and very high) are aggregated. The above three linguistic weights (high, very high, and very high) are aggregated using the SUM method. Figure 2 represents the centroid defuzzification method that is implemented to calculate the numerical value of the weight in the range $[-1, 1]$.

Using this method, the weight of all relationships between the concepts related to FCM for GC was calculated. The developed FCM is shown in Figure 3. In the third step, we used a learning algorithm to train the model, which includes updating the relationship weight, and finally, a fuzzy cognition map for GC risk factors was extracted. For this purpose, data collected from 560 patients referred to Imam Reza Hospital in Tabriz (after the preprocessing steps) were used through a questionnaire. Table 2 shows the features, values, and frequency of patients.

Figure 4 shows the proposed FCM model for risk factors of GC. This FCM has 28 concepts and 38 edges with their weights. Considering the 28 concept nodes, 27 are the ultimate physician-selected features that interfere with the disease and are shown by the values C1 to C27. The central node is the concept of GC, which receives and collects interactions from all other nodes. The positive weight of an edge indicates that it has a positive effect on the incidence of GC, and the negative weight indicates the role of deterrence in the incidence of the disease. The yellow, purple, blue, and green colors were used to specify the category of any feature or concept. The C1 to C8 features specified with yellow were classified as personal features. The violet color was used for the C9 to C17 features of the diet food category. Blue and green were also used for the C18 to C22 features of the systemic condition, respectively, and C23 to C27 features were used for the stomach condition category.

2.2. Learning FCM Using NHL Algorithm. GC specialists were well positioned to create FCM in our method. Nonlinear Hebbian Learning (NHL) is utilized to learn the weights due to no access to a relatively large data set, causal weight optimization, and more accurate results [49]. The Hebbian-based algorithms were used for FCM training to determine the best matrix in terms of expert knowledge [50]. Algorithms set the FCM weights through existing data and a learning formula in terms of repetition and Hebbian rule methods [50]. The NHL algorithm is based on the assumption that all of the concepts of the FCM model are stimulated at each time step and their values change. The value ω_{ji} corresponding to the concepts of c_j and c_i is updated, and the weight ω_{ji} is corrected in iteration k . The value of $A_i^{(k+1)}$ is determined in the $(k+1)$ th iteration. The impact of concepts with values A_j and corrected weighted values $\omega_{ij}^{(k)}$ in iteration k is determined by

$$A_i^{(k+1)} = f \left(A_i^{(k)} + \sum_{j=1, j \neq i}^n \omega_{ji}^{(k)} \cdot A_j^{(k)} \right). \quad (1)$$

Each of the concepts in the FCM model may be input or output concepts. A number of concepts are defined as output concepts (OCs). These concepts are the state of the system in which we want to estimate the value that represents the final state of the system. The classification of concepts as input and output concepts is by the experts of the group and according to the subject under consideration.

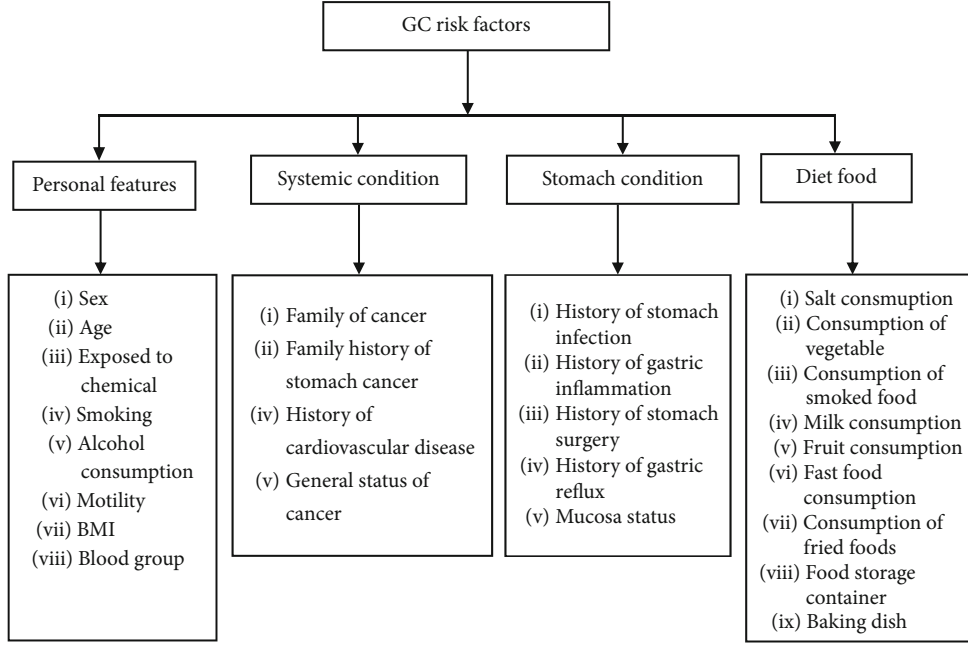


FIGURE 1: Classification of GC risk factors.

The mathematical relations used in the NHL algorithm for learning FCM are shown in equations (1) and (2).

$$\Delta\omega_{ji} = \eta_t A_i^{(K-1)} \left(A_j^{(K-1)} - \omega_{ji}^{(K-1)} A_i^{(K-1)} \right), \quad (2)$$

where η is a scaling parameter called the learning rate. η is a very small positive scalar factor called learning parameter. Its value is obtained through test error.

$$\omega_{ji}^{(k)} = \gamma \cdot \omega_{ji}^{(k-1)} + \eta A_i^{(k-1)} \left(A_j^{(k-1)} - \text{sgn}(\omega_{ji}) \omega_{ji}^{(k-1)} A_i^{(k-1)} \right). \quad (3)$$

Equation (3) is the main equation of the NHL algorithm. γ is the weight decay parameter. The values of concepts and weights $\omega_{ji}^{(k)}$ are calculated by equations (1) and (3), respectively. In fact, the NHL algorithm updates the basic matrix nonzero elements suggested by the experts in each iteration. The following criteria determine when the NHL algorithm ends [50].

(a) The terminating function $F1$ is given as

$$F1 = \|\text{OC}_i - T_i\|, \quad (4)$$

where T_i is the mean value of OC_i .

This kind of metric function is suitable for the NHL algorithm used in the FCMs. In each step, $F1$ calculates the Euclidean distance for OC_i and T_i . Assuming that $\text{OC}_i = [T_i^{\min}, T_i^{\max}]$, T_i is calculated by

$$T_i = \frac{T_i^{\min} + T_i^{\max}}{2}. \quad (5)$$

Given that the FCM model has m -OCs, for calculating $F1$, the sum of the square between m -OCs and m - T_s can be calculated by

$$F1 = \sqrt{\sum_{j=1}^m (\text{OC}_i - T_i)^2}. \quad (6)$$

After $F1$ is minimized, the situation ends. b) The second condition for completing the algorithm is the difference between two consecutive OCs. This value should be less than e . Therefore, the value of the $(k+1)$ th iteration should be less than e based on

$$F2 = \left| \text{OC}_i^{(t+1)} - \text{OC}_i^{(t)} \right| < e = 0.002. \quad (7)$$

In this algorithm, the values of the parameters η and γ are determined through test error. After several tests, the values of η and γ show the best performing algorithm. Finally, when the algorithmic termination conditions are met, the final weight matrix (ω_{NHL}) is obtained.

For the convenience of end-users, a graphical interface is designed using the GUI in MATLAB for the proposed system. The user interface for the GC risk prediction software is shown in Figure 3.

For example, the user enters the requested information into the system. The system displays the risk assessment result after receiving information from the user and using the proposed NHL-FCM model.

For the comparison of classification accuracy, the same data set is used for classification with other machine learning models. Backpropagation neural network, support vector machine, decision tree, and Bayesian classifier were used in the Weka toolkit V3.7 to test other learning algorithms. For

TABLE 1: Risk factors of GC.

Risk factors	Description
C1: sex	Studies show that men around the world are diagnosed with GC almost twice as much as women [18].
C2: blood group	Scientific research shows that there is a significant relationship between blood type and GC. The blood groups A and O have the highest and lowest incidence of GC, respectively [19].
C3: BMI	High BMI increases GC [20]. In 2016, the IACR formed a team of specialists. They reported that GC is one of the diseases caused by excessive fat gain and high BMI [21].
C4: age	The risk of GC increases with age [18, 22, 23].
C5: motility	People with any regular physical activity have a lower risk of GC than nonactive people. According to the US Physical Activity Guidelines Advisory Committee (2018), moderate evidence showed that physical activity reduces the risk of various cancers, including GC [21].
C6: alcohol consumption	Regular alcohol consumption increases the risk of GC [24, 25].
C7: exposed to chemicals	Some jobs exposed to chemicals, such as cement and chromium, increase the risk of GC [26].
C8: smoking	Smoking increases the risk of GC [27, 28].
C9: salt consumption	High salt intake increases the risk of GC [23, 29, 30].
C10: consumption of vegetable	The daily consumption of 200-200 grams of vegetables per day may reduce the risk of GC [31].
C11: consumption of smoked food	The smoked food is a great source of polycyclic aromatic hydrocarbons (PAHs). Scientific research has shown that this biopollutant is one of the factors involved in many cancers, including GC [32, 33].
C12: milk consumption	Increasing dairy consumption, such as milk, is associated with a lower risk of GC [34].
C13: fast food consumption	Fast food consumption is one of the factors affecting the incidence of GC [35].
C14: consumption of fried foods	The results of scientific studies show that people who use a lot of fried foods in their diet are at increased risk of GC [27, 28].
C15: fruit consumption	A daily consumption of 120-150 grams of fruit per day may reduce the risk of GC [31].
C16: food storage container	Today's food containers are often made of chemicals, such as plastics that contain bisphenol A. Thus, it can be the source of various types of cancer and hormonal disorders [36].
C17: baking dish	The use of metal containers, such as aluminum for cooking, can be a factor in the development of diseases because these types of metals, when exposed to heat, emit a small amount of lead [37].
C18: history of allergy	Recent studies indicate that the history of allergic diseases is associated with a lower risk of GC [38].
C19: family history of cancer	A family history of cancer in certain specific sites may be associated with a risk of GC [39].
C20: family of GC	This risk factor is strongly associated with different types of GC [40, 41].
C21: history of cardiovascular disease	People with cardiovascular disease are at a lower risk of GC because of using some drugs [42].
C22: general status of cancer	People with a good general health status are less likely to be at risk of GC [43].
C23: history of gastric reflux	Gastric reflux causes a 3-10% percent increase in being at risk of GC [44].
C24: history of stomach surgery	Gastric surgeries, such as gastric ulcers, may increase the risk of cancer [45].
C25: history of stomach infection	Helicobacter pylorus is the most important risk factor for GC [46-48].
C26: mucosa status	Gastric ulcers are considered as a risk factor for GC [35].
C27: history of gastric inflammation	The history of gastric inflammation is one of the most important factors in the incidence of GC [35].

this purpose, the Excel file containing the collected data collection was converted to .arff format so that it can be read for Weka. Then, the required steps for data preprocessing were performed. In this software, one of the most common methods of evaluating the performance of categories that divide the tagged data set into several subsets is cross-validation. 10-fold cross-validation was used for all the studied algorithms. 10-fold cross-validation divides the data set into 10 parts and performs the test 10 times. In each step, one part is considered as a test and the other 9 parts are considered for training. In this way, each data is used once for

testing and 9 times for training. As a result, the entire data set is covered for training and testing.

The backpropagation neural network with 27 input neurons, 10 neurons, and 3 output nodes was used as the multi-layer perceptron. Also, for classification of the assess risk into three classes, high, medium, and low, the support vector machine, decision tree C4.5, and Naïve Bayesian classifier were used.. Given that the data studied are not linearly separable, we need to use the core technology to implement the SVM algorithm. The core technology is one of the most common techniques for solving problems that are not linearly

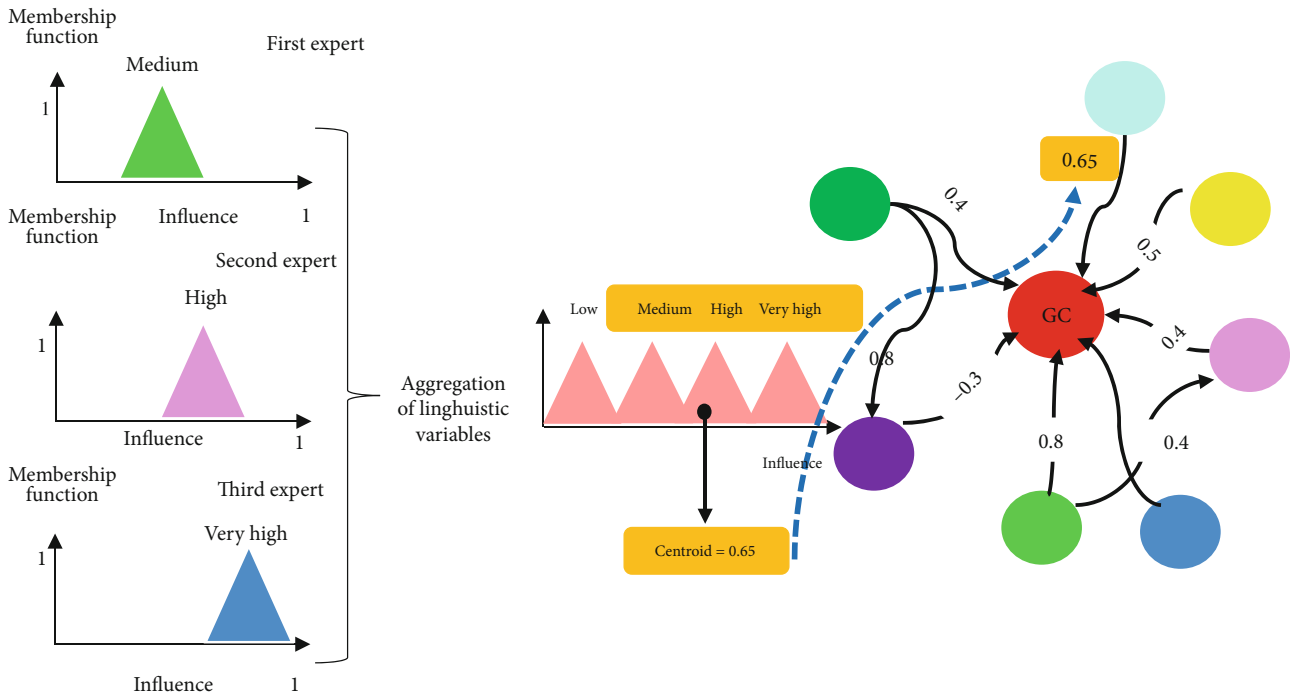


FIGURE 2: Aggregation and defuzzification of linguistic weights.

FIGURE 3: User interface of the proposed MDSS.

TABLE 2: Data sets.

Features	Range	Number	Percent
Sex	Male	256	45.7%
	Female	304	54.3%
	<40	20	3.47%
Age	41-60	210	37.5%
	≥61	330	59.03%
	A	123	21.96%
Blood group	B	78	13.92%
	AB	80	14.28%
	O	279	49.82%
	BMI > 30	69	12.32%
BMI	25 < BMI > 29.5	76	13.57%
	18.5 < BMI > 24.9	120	21.42%
	BMI < 18.5	293	52.32%
Motility	Light	156	27.85%
	Medium	236	42.14%
	High	168	30%
Alcohol consumption	Yes	85	15.17%
	No	475	84.82%
Exposed to chemicals	Yes	54	9.64%
	No	506	90.35%
Smoking	Yes	198	35.35%
	No	362	64.64%
	None	10	1.78%
Salt consumption	Low	175	31.25%
	High	375	66.96%
Consumption of vegetable	Daily	26	4.64%
	1-3 times a week	214	38.21%
	1-3 times a month	320	57.14%
Consumption of smoked food	None	5	0.89%
	Daily	0	0%
	1-3 times a week	149	26.60%
Milk consumption	1-3 times a month	406	72.5%
	Yes	214	38.21%
	No	346	61.78%
Fast food consumption	None	4	0.71%
	1-3 times a week	315	56.25%
	1-3 times a month	241	43.03%
Consumption of fried foods	None	0	0%
	1-3 times a week	191	34.10%
	1-3 times a month	369	65.89%
Fruit consumption	None	6	1.07%
	1-3 times a week	185	33.03%
	1-3 times a month	369	65.89%
Food storage container	Aluminum	216	38.57%
	Plastic	301	53.75%
	Copper	32	5.71%
	Style	9	1.60%
	Chinese	2	0.35%

TABLE 2: Continued.

Features	Range	Number	Percent
Baking dish	Aluminum	10	1.78%
	Teflon	390	69.64%
	Copper	21	3.75%
History of allergy	Yes	89	15.89%
	No	471	84.10%
Family history of cancer	Yes	211	37.67%
	No	349	62.32%
Family of GC	Yes	123	21.965%
	No	437	78.03%
History of cardiovascular disease	Yes	185	33.03%
	No	375	66.96%
General status	Good	79	14.10%
	So-so	190	33.92%
	Poor	291	51.965%
History of gastric reflux	Yes	234	41.78%
	No	326	58.21%
History of stomach surgery	Yes	48	8.57%
	No	512	91.42%
History of stomach infection	Yes	176	31.42%
	No	384	68.57%
Mucosa status	Normal	94	16.78%
	Swollen	126	22.5%
	Red	157	28.03%
	Sore	183	32.67%
History of gastric inflammation	Yes	163	29.10%
	No	397	70.89%
Risk score	High	300	53.57%
	Moderate	186	33.21%
	Low	74	8.39%

separable. In this method, a suitable core function is selected and executed. In fact, the purpose of kernel functions is to linearize nonlinear problems. There are several kernel functions in Weka. The RBF (Radial Basis Function) was used to run the SVM algorithm. By selecting and running the C4.5 algorithm, you can see the results of the classification. Also, the tree created by this algorithm can be seen graphically, which is a large tree. The three categories of high risk, medium risk, and low risk were selected as target variables and other characteristics as predictive variables. The leaves of the tree are the target variables and can be seen as a number of rules according to the model made by the tree. Naïve Bayesian was another classification algorithm that was implemented using Weka on the studied data, and its results were examined. This algorithm uses a possible framework to solve classification problems.

3. Results

To analyze the performance of the proposed method, we divided the data into two categories. The proposed model

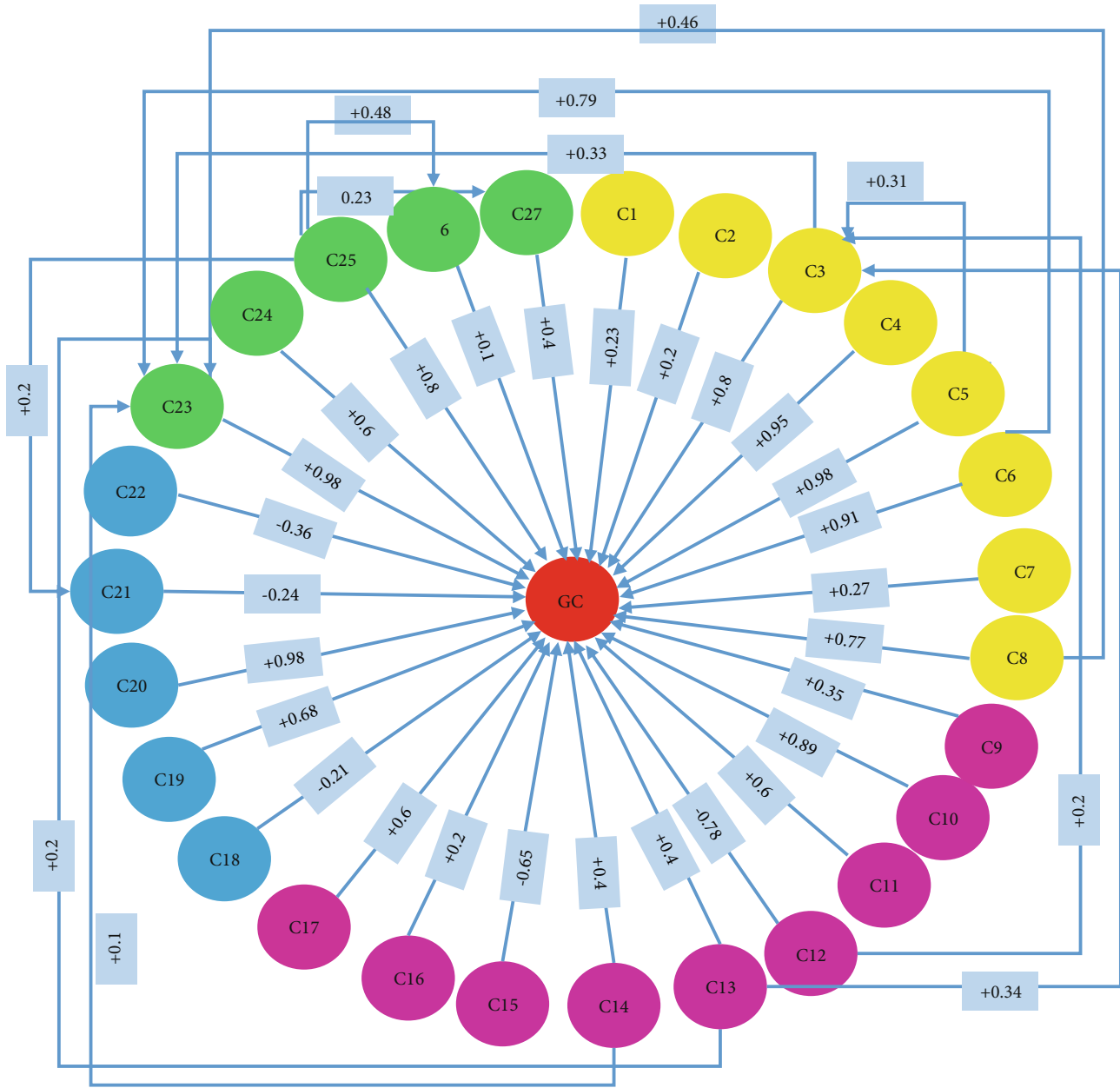


FIGURE 4: FCM model for GC risk factors.

was trained using 70% of the patient records (392 records) based on the NHL algorithm and tested using 30% of the records (168 records). Considering 168 patient records selected for testing randomly, there were 56 records in the high category, 64 records in the medium category, and 48 records in the low category.

Root square error (RMSE) and performance measure accuracy, recall, precision, and mean absolute error (MAE) are the key behavior measures in the medical field [17] widely utilized in the literature. To determine accuracy, recall, and precision, the turbulence matrix was utilized. A confusion matrix is a table making possible to visualize the behavior of an algorithm. Table 3 represents the general scheme of a confusion matrix (with two groups C1 and C2).

TABLE 3: Confusion matrix.

		Predicted class	
		C1	C2
Actual class	C1	True positive (TP)	False positive (FP)
	C2	False negative (FN)	True negative (TN)

The matrix contains two columns and two rows specifying the values including the number of true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP). TP shows the number of specimens for class C1 classified appropriately. FP represents the number of specimens

for group C2 classified inaccurately as C1. TN shows the number of samples for class C2 classified correctly. FN represents the number of specimens for class C1 classified incorrectly as class C2.

- (i) *Accuracy*: accuracy represents the ratio of accurately classified specimens to the total number of tested samples. It is determined by

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{TP} + \text{FN} + \text{FP})}. \quad (8)$$

- (ii) *Recall*: recall is the number of instances of the class C1 that has actually predicted correctly. It is calculated by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (9)$$

- (iii) *Precision*: it represents the classifier's ability not to label a C2 sample as C1. It is calculated by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (10)$$

The MAE performance index is calculated by

$$\text{MAE} = \frac{1}{N} \left(\sum_{L=1}^N \sum_{J=1}^C \left| \text{OC}_{JL}^{\text{Real}} - \text{OC}_{JL}^{\text{Predicted}} \right| \right). \quad (11)$$

In equation (11), N represents the number of training data ($N = 560$), C shows the number of output concepts ($C = 3$), and $\text{OC}_{JL}^{\text{Real}} - \text{OC}_{JL}^{\text{Predicted}}$ denotes the difference between the l th decision output concept (OC) and its equivalent real value (target) by appearing the k th set of input concepts to the input of the tool.

The RMSE evaluation index is defined based on

$$\text{RMSE} = \sqrt{\frac{1}{NC} \left(\sum_{L=1}^N \sum_{J=1}^C \left(\text{OC}_{JL}^{\text{Real}} - \text{OC}_{JL}^{\text{Predicted}} \right)^2 \right)}, \quad (12)$$

where N is the number of training sets and C is the system outputs.

Table 4 shows the accuracy results obtained from the proposed method and other standard categorizers. The proposed method works better than other categories because of the efficiency of the NHL's efficiency for working with very small data to correct FCM weight. As a result, optimal decisions are made for output concepts.

The results show that the highest total accuracy is related to the proposed method (95.83%) which is about 5% higher

than the accuracy of the MLP-ANN algorithm. The highest precision and recall are related to the proposed algorithm, which are, respectively, 96.77% (medium) and 98.21% (high). It also shows that the training error of the proposed method based on NHL is less than the other algorithms used in this study.

As stated, γ and η are two learning parameters in the NHL algorithm. In this algorithm, the upper and lower limits of these parameters are determined by trial and error in order to optimize the final solution. After several simulations with parameters γ and η , it was observed that the use of large amounts of γ causes significant changes in weights and weight marks. Also, simulation with small η also creates significant weight changes, thus preventing the weight of concepts from entering the desired range. For this reason, values γ and η are limited to $0 < \gamma < 0.1$ and $0.9 < \eta < 1$. In each study, a constant value is considered for these parameters.

After several investigations, it was found that the best performance of the category is related to $\eta = 0.045$ and $\gamma = 0.98$. The classification results obtained for the different values of learning parameters are presented in Table 5.

4. Discussion

In this study, we designed a risk prediction model and a GC risk assessment tool using data from a study on a population of patients referring to the gastroenterology unit of Imam Reza Hospital in Tabriz. The proposed model presented in this study is attempting to rationalize beyond the analyses of clinical experts and increase the ability of experts to make logical decisions in a clinical setting for patients with different levels of risk factors for GC and help clinical specialists to make a logical decision about optimal preventive methods for patients.

The 95.8% overall classification accuracy obtained through the Hebbian-based FCM using 560 patients indicates a high level of coordination between the proposed system and medical decisions, and the proposed decision support tool can be trusted for clinical professionals and also helps them in the process of risk assessment of gastric GC.

Specifically, our risk assessment tool is simple and inexpensive to use in the clinical environment, because many other methods to predict the risk of GC are invasive. Therefore, this is an effective instrument for estimating the population at risk of cancer in the future. The results show that this new model can predict the probability of developing GC concerning the characteristics specified in this study with a better accuracy than previous studies.

In recent years, several researches have been carried out on the development and validation of risk assessment tools for various cancers [51, 52]. Recent studies have shown that the combination of *H. pylori* antibody and serum pepsinogen can be a good predictor of GC [53, 54].

We believe that only two other evaluation instruments exist for GC rather than ours. Based on the Japan Public Health Center-based Prospective Study, a device was designed to estimate the cumulative probability of GC incidence including sex, age, smoking status, the mixture of *H.*

TABLE 4: Performance metrics.

Classifiers	+	High	Medium	Low	Class recall	Class precision	Overall accuracy	RMSE	MAE
Decision trees	High	30	10	1	53.57	73.17	76.78	0.5120	0.721
	Medium	16	52	0	81.25	76.47			
	Low	10	2	47	97.91	79.66			
Naïve Bayes	High	40	8	5	71.42	75.47	80.35	0.334	0.645
	Medium	8	56	4	87.5	77.77			
	Low	8	0	39	81.25	82.97			
SVM	High	46	2	4	82.14	88.46	86.9	0.193	0.342
	Medium	0	60	4	93.75	93.75			
	Low	10	2	40	83.3	76.92			
MLP-ANN	High	49	2	7	87.5	84.48	90.47	0.248	0.097
	Medium	4	58	4	90.62	87.87			
	Low	3	4	45	93.75	86.53			
Proposed model	High	55	1	1	98.21	96.49	95.83	0.173	0.0471
	Medium	1	60	1	93.75	96.77			
	Low	0	3	46	95.83	93.87			

TABLE 5: Classification results, based on different values of η and γ .

η	γ	Confusion matrix			Classification accuracy (%)
		High	Medium	Low	
0.01	0.97	50	4	7	88.69
		4	59	1	
		2	1	40	
0.03	0.95	45	6	1	89.28
		5	58	0	
		6	0	47	
0.045	0.98	55	1	1	95.83
		1	60	1	
		0	3	46	
0.05	0.96	54	6	0	94.04
		1	56	0	
		1	2	48	
0.055	0.96	53	2	5	91.6
		2	58	0	
		1	4	43	

pylori antibody and serum pepsinogen, consumption of salty food, and family history of GC as the risk factors [55]. A good performance was found by the model based on calibration and discrimination. Based on [2], a risk evaluation instrument for GC was proposed in the general population of Japan. In this work, gender, age, the combination of *Helicobacter pylori* antibody and pepsinogen status, smoking status, and hemoglobin A1C level were risk factors for GC.

The risk factors chosen in these two studies were very limited to a few specific characteristics and had little similarity to the factors in our study. Risks such as consumption of fruits and vegetables, alcohol consumption, history of cardiovascular disease, blood type, milk consumption, history of

allergy, gastric reflux, storage containers, food intake, and family history of cancer did not exist in both studies in spite of their importance in previous studies. Factors such as salt intake and a history of GC are known as causes of GC that did not exist in [2]. Another remarkable point in our study is that, given the nature of the proposed model, this method addresses the effects of factors that are sometimes related to each other or even the mutual effects that might put each other at risk, but it is not included in the two previous studies.

Another advantage of the proposed method than other algorithms is that other methods cannot provide any explicit causal relationship and the system works as a black box. This problem also makes these algorithms less suited to medical decision support systems. Finally, the new system has the following benefits:

- (i) It examines the factors that have not been taken into account in previous models to assess the risk of GC
- (ii) Because of the use of new factors, this model can be more effective in predicting the risk of GC
- (iii) The proposed model is presented by a software that has a simple, convenient, and user-friendly interface
- (iv) The use of this software by physicians and other researchers can tackle individual healthcare decisions
- (v) It helps healthcare professionals decide on individual risk management mechanisms

The system presented in this study has the following limitations: (1) a small sample of patients used to learn and anticipate GC, (2) the heavy dependence of this model on knowledge of domain specialists, (3) dependence on initial conditions and communication, and (4) the absence of external validation of the forecast system. Although this system has nice results due to the use of an appropriate database

and the important and relevant GC factors, the generalizability of our results cannot be proved without the experiment of the system in another data set. As a result, it is necessary to use a larger statistical population to test the proposed model.

5. Conclusions

Assessing the level of risk for GC is very important and helps make decisions about screening. Given the limited number of GC risk assessment tools that have been proposed so far, there is no tool that comprehensively covers the risk factors in scientific studies on GC. The proposed model based on soft computing covers all the factors influencing the incidence of GC. The classification accuracy of the proposed method is higher than other methods of the machine learning classification, such as the decision tree and SVM. This is due to the useful features of FCM for checking domain knowledge and determining the initial structure of FCM and the initial weights and then using the NHL algorithm to teach the FCM model and adjust these weights. The FCM-based model is comprehensive, transparent, and more effective than previous models for assessing the risk of GC. As a result, this risk assessment tool can help diagnose people with a high risk of GC and help both healthcare providers and patients with the decision-making process. Our future work is to use more features and variations and other learning algorithms to determine the weight of the edges in the FCM.

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 claim no conflicts of interest.

Supplementary Materials

A review of fuzzy cognitive maps [56]. (*Supplementary Materials*)

References

- [1] J. Ferlay, I. Soerjomataram, R. Dikshit et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–E386, 2015.
- [2] H. Charvat, S. Sasazuki, M. Inoue et al., "Prediction of the 10-year probability of gastric cancer occurrence in the Japanese population: the JPHC study cohort II," *International Journal of Cancer*, vol. 138, no. 2, pp. 320–331, 2016.
- [3] M. Iida, F. Ikeda, J. Hata et al., "Development and validation of a risk assessment tool for gastric cancer in a general Japanese population," *Gastric Cancer*, vol. 21, no. 3, pp. 383–390, 2018.
- [4] R. Sharda, D. Delen, and E. Turban, *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support*, Pearson, 2019.
- [5] A. Amirkhani, E. I. Papageorgiou, A. Mohseni, and M. R. Mosavi, "A review of fuzzy cognitive maps in medicine: taxonomy, methods, and applications," *Computer Methods and Programs in Biomedicine*, vol. 142, pp. 129–145, 2017.
- [6] E. I. Papageorgiou, P. P. Spyridonos, C. D. Stylios, P. Ravazoula, P. P. Groumpos, and G. N. Nikiforidis, "Advanced soft computing diagnosis method for tumour grading," *Artificial Intelligence in Medicine*, vol. 36, no. 1, pp. 59–70, 2006.
- [7] M. H. B. de Moraes Lopesa, N. R. S. Ortegab, P. S. P. Silveirab, E. Massadb, R. Higac, and H. de Fátima Marind, "Fuzzy cognitive map in differential diagnosis of alterations in urinary elimination: a nursing approach," *International Journal of Medical Informatics*, vol. 82, no. 3, pp. 201–208, 2013.
- [8] E. I. Papageorgiou, C. D. Stylios, and P. P. Groumpos, "Active Hebbian learning algorithm to train fuzzy cognitive maps," *International Journal of Approximate Reasoning*, vol. 37, no. 3, pp. 219–249, 2004.
- [9] V. K. Mago, R. Mehta, R. Woolrych, and E. I. Papageorgiou, "Supporting meningitis diagnosis amongst infants and children through the use of fuzzy cognitive mapping," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, 2012.
- [10] P. Beena and R. Ganguli, "Structural damage detection using fuzzy cognitive maps and Hebbian learning," *Applied Soft Computing*, vol. 11, no. 1, pp. 1014–1020, 2011.
- [11] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no. 1, pp. 65–75, 1986.
- [12] E. I. Papageorgiou and J. L. Salmeron, "A review of fuzzy cognitive maps research during the last decade," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 1, pp. 66–79, 2013.
- [13] E. I. Papageorgiou, N. I. Papandrianos, G. Karagianni, and D. Sfyas, "Fuzzy cognitive map based approach for assessing pulmonary infections," in *Lecture Notes in Computer Science*, pp. 109–118, Springer, Berlin, Heidelberg, 2009.
- [14] A. Amirkhani, M. R. Mosavi, F. Mohammadzadeh, and S. B. Shokouhi, "Classification of intraductal breast lesions based on the fuzzy cognitive map," *Arabian Journal for Science and Engineering*, vol. 39, no. 5, pp. 3723–3732, 2014.
- [15] A. Mpilllis, E. I. Papageorgiou, C. A. Frantzidis, M. S. Tsatali, A. C. Tsolaki, and P. D. Bamidis, "A decision-support framework for promoting independent living and ageing well," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 199–209, 2015.
- [16] E. I. Papageorgioua, J. Subramanianb, A. Karmegamc, and N. Papandrianosd, "A risk management model for familial breast cancer: a new application using fuzzy cognitive map method," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 2, pp. 123–135, 2015.
- [17] P. J. Giabbanelli, T. Torsney-Weir, and V. K. Mago, "A fuzzy cognitive map of the psychosocial determinants of obesity," *Applied Soft Computing*, vol. 12, no. 12, pp. 3711–3724, 2012.
- [18] G. Murphy, R. Pfeiffer, M. C. Camargo, and C. S. Rabkin, "Meta-analysis shows that prevalence of Epstein–Barr virus-positive gastric cancer differs based on sex and anatomic location," *Gastroenterology*, vol. 137, no. 3, pp. 824–833, 2009.
- [19] B. L. Zhang, N. He, Y. B. Huang, F. J. Song, and K. X. Chen, "ABO blood groups and risk of cancer: a systematic review and meta-analysis," *Asian Pacific Journal of Cancer Prevention*, vol. 15, no. 11, pp. 4643–4650, 2014.

- [20] C. Q. Sun, Y. B. Chang, L. L. Cui et al., "A population-based case-control study on risk factors for gastric cardia cancer in rural areas of Linzhou," *Asian Pacific Journal of Cancer Prevention*, vol. 14, no. 5, pp. 2897–2901, 2013.
- [21] S. M. Gapstur, J. M. Drope, E. J. Jacobs et al., "A blueprint for the primary prevention of cancer: targeting established, modifiable risk factors," *CA: a Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 446–470, 2018.
- [22] P. Karimi, F. Islami, S. Anandasabapathy, N. D. Freedman, and F. Kamangar, "Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 23, no. 5, pp. 700–713, 2014.
- [23] M. Rugge and M. Fassan, Eds. D. Y. Graham, "Epidemiology of gastric cancer," in *Gastric Cancer*, M. Rugge, M. Fassan, and , Eds., pp. 23–33, Springer International Publishing, Switzerland, 2015.
- [24] J. Dong and A. P. Thrift, "Alcohol, smoking and risk of oesophago-gastric cancer," *Best Practice & Research. Clinical Gastroenterology*, vol. 31, no. 5, pp. 509–517, 2017.
- [25] K. A. Moy, Y. Fan, R. Wang, Y. T. Gao, M. C. Yu, and J. M. Yuan, "Alcohol and tobacco use in relation to gastric cancer: a prospective study of men in Shanghai, China," *Cancer Epidemiology Biomarkers & Prevention*, vol. 19, no. 9, pp. 2287–2297, 2010.
- [26] A. R. Yusefi, K. Bagheri Lankarani, P. Bastani, M. Radinmanesh, and Z. Kavosi, "Risk factors for gastric cancer: a systematic review," *Asian Pacific Journal of Cancer Prevention*, vol. 19, no. 3, pp. 591–603, 2018.
- [27] M. Pakseresht, D. Forman, R. Malekzadeh et al., "Dietary habits and gastric cancer risk in north-west Iran," *Cancer Causes & Control*, vol. 22, no. 5, pp. 725–736, 2011.
- [28] K. Jain, V. Sreenivas, T. Velpandian, U. Kapil, and P. K. Garg, "Risk factors for gallbladder cancer: a case-control study," *International Journal of Cancer*, vol. 132, pp. 1660–1666, 2012.
- [29] L. D'Elia, G. Rossi, R. Ippolito, F. P. Cappuccio, and P. Strazzullo, "Habitual salt intake and risk of gastric cancer: a meta-analysis of prospective studies," *Clinical Nutrition*, vol. 31, no. 4, pp. 489–498, 2012.
- [30] M. Verdalet-Olmedo, C. Sampieri, J. M. Romero, H. M. Guevara, Á. M. Machorro-Castaño, and K. L. Córdoba, "Omission of breakfast and risk of gastric cancer in Mexico," *World Journal of Gastrointestinal Oncology*, vol. 4, no. 11, pp. 223–229, 2012.
- [31] M. Ganjavi, B. Faraji, F. Kamangar, and C. Tucker, "Delayed effect of fruits and vegetables on gastric cancer," *Journal of The Academy of Nutrition and Dietetics*, vol. 117, no. 9, p. A21, 2017.
- [32] M. B. Braga-Neto, J. G. Carneiro, A. M. de Castro Barbosa et al., "Clinical characteristics of distal gastric cancer in young adults from Northeastern Brazil," *BMC Cancer*, vol. 18, no. 1, p. 131, 2018.
- [33] X. J. Cheng, J. C. Lin, and S. P. Tu, "Etiology and prevention of gastric cancer," *Gastrointest Tumors*, vol. 3, no. 1, pp. 25–36, 2016.
- [34] Y. Guoa, Z. Shanb, H. Renc, and W. Chena, "Dairy consumption and gastric cancer risk: a meta-analysis of epidemiological studies," *Nutrition and Cancer*, vol. 76, no. 4, pp. 555–568, 2015.
- [35] F. Habibzadeh, "Gastric cancer in the Middle East," *The International Journal of Occupational and Environmental Medicine*, vol. 382, 2013.
- [36] I. Husaina, M. Alalyanib, and A. H. Hanga, "Disposable plastic food container and its impacts on health," *The Journal of Energy and Environmental Science*, vol. 130, pp. 618–623, 2015.
- [37] J. D. Weidenhamer, M. P. Fitzpatrick, A. M. Biro et al., "Metal exposures from aluminum cookware: an unrecognized public health risk in developing countries," *Science of the Total Environment*, vol. 579, pp. 805–813, 2017.
- [38] S. Jo, T. J. Kim, H. Lee et al., "Associations between atopic dermatitis and risk of gastric cancer: a nationwide population-based study," *The Korean Journal of Gastroenterology*, vol. 71, no. 1, pp. 38–44, 2018.
- [39] X. Jiang, C. C. Tseng, L. Bernstein, and A. H. Wu, "Family history of cancer and gastroesophageal disorders and risk of esophageal and gastric adenocarcinomas: a case-control study," *BMC Cancer*, vol. 14, no. 1, 2014.
- [40] M. Song, M. C. Camargo, S. J. Weinstein et al., "Family history of cancer in first-degree relatives and risk of gastric cancer and its precursors in a Western population," *Gastric Cancer*, vol. 21, no. 5, pp. 729–737, 2018.
- [41] C. Y. Yun, N. Kim, J. Lee et al., "Usefulness of OLGA and OLGIM system not only for intestinal type but also for diffuse type of gastric cancer, and no interaction among the gastric cancer risk factors," *Helicobacter*, vol. 23, no. 6, 2018.
- [42] S. A. Mahmoodi, K. Mirzaie, and S. M. Mahmoudi, "A new algorithm to extract hidden rules of gastric cancer data based on ontology," *Springerplus*, vol. 5, no. 1, 2016.
- [43] M. S. Kwak, K. S. Choi, S. Park, and E. C. Park, "Perceived risk for gastric cancer among the general Korean population: a population-based survey," *Psychooncology*, vol. 18, no. 7, pp. 708–715, 2009.
- [44] M. Rugge, R. M. Genta, F. di Mario et al., "Gastric cancer as preventable disease," *Clinical Gastroenterology and Hepatology*, vol. 15, no. 12, pp. 1833–1843, 2017.
- [45] "Causes, risk factors, and prevention, stomach cancer risk factors," 2017, <https://www.cancer.org/cancer/stomach-cancer/causes-risks-prevention/risk-factors/>.
- [46] K. Sugano, "Effect of *Helicobacter pylori* eradication on the incidence of gastric cancer: a systematic review and meta-analysis," *Gastric Cancer*, vol. 22, no. 3, pp. 435–445, 2019.
- [47] V. E. Cokkinides, P. Bandi, R. L. Siegel, and A. Jemal, "Cancer-related risk factors and preventive measures in US Hispanics/Latinos," *CA: a Cancer Journal for Clinicians*, vol. 62, no. 6, pp. 353–363, 2012.
- [48] J. Jiang, Y. Chen, J. Shi, C. Song, J. Zhang, and K. Wang, "Population attributable burden of *Helicobacter pylori*-related gastric cancer, coronary heart disease, and ischemic stroke in China," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 36, no. 2, pp. 199–212, 2017.
- [49] A. E. I. Papageorgiou and W. Froelich, "Application of evolutionary fuzzy cognitive maps for prediction of pulmonary infections," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 1, pp. 143–149, 2012.
- [50] A. Amirkhani, M. R. Mosavi, K. Mohammadi, and E. I. Papageorgiou, "A novel hybrid method based on fuzzy cognitive maps and fuzzy, clustering algorithms for grading celiac disease," *Neural Computing and Applications*, vol. 30, no. 5, pp. 1573–1588, 2018.
- [51] K. G. Yeoh, K. Y. Ho, H. M. Chiu et al., "The Asia-Pacific Colorectal Screening score: a validated tool that stratifies risk for

- colorectal advanced neoplasia in asymptomatic Asian subjects,” *Gut*, vol. 60, no. 9, pp. 1236–1241, 2011.
- [52] B. Rosner, G. A. Colditz, J. D. Iglehart, and S. E. Hankinson, “Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurse’s Health Study,” *Breast Cancer Research*, vol. 10, no. 4, p. R55, 2008.
- [53] T. Terasawa, H. Nishida, K. Kato et al., “Prediction of gastric cancer development by serum pepsinogen test and *Helicobacter pylori* seropositivity in Eastern Asians: a systematic review and meta-analysis,” *PLoS One*, vol. 9, no. 10, p. e109783, 2014.
- [54] H. Watabe, T. Mitsushima, Y. Yamaji et al., “Predicting the development of gastric cancer from combining *Helicobacter pylori* antibodies and serum pepsinogen status: a prospective endoscopic cohort study,” *Gut*, vol. 54, no. 6, pp. 764–768, 2005.
- [55] C. De Martel, J. Ferlay, S. Franceschi et al., “Global burden of cancers attributable to infections in 2008: a review and synthetic analysis,” *The Lancet Oncology*, vol. 13, no. 6, 2012.
- [56] E. I. Papageorgiou, “Learning algorithms for fuzzy cognitive maps—a review study,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 150–163, 2012.

Research Article

A Benign and Malignant Breast Tumor Classification Method via Efficiently Combining Texture and Morphological Features on Ultrasound Images

Mengwan Wei,¹ Yongzhao Du ,^{1,2,3} Xiuming Wu,⁴ Qichen Su,^{3,5} Jianqing Zhu,¹ Lixin Zheng,¹ Guorong Lv ,^{3,5} and Jiafu Zhuang⁶

¹College of Engineering, Huaqiao University, Quanzhou 362021, China

²School of Medicine, Huaqiao University, Quanzhou 362021, China

³Collaborative Innovation Center for Maternal and Infant Health Service Application Technology, Quanzhou Medical College, Quanzhou, China

⁴The First Hospital of Quanzhou, Fujian Medical University, Quanzhou 350005, China

⁵Department of Medical Ultrasonics, The Second Affiliated Hospital of Fujian Medical University, Quanzhou 362000, China

⁶Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, 362216 Quanzhou, China

Correspondence should be addressed to Yongzhao Du; yongzhaodu@126.com and Guorong Lv; lgr_feus@sina.com

Received 13 June 2020; Revised 1 September 2020; Accepted 15 September 2020; Published 1 October 2020

Academic Editor: Lin Lu

Copyright © 2020 Mengwan Wei 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 classification of benign and malignant based on ultrasound images is of great value because breast cancer is an enormous threat to women's health worldwide. Although both texture and morphological features are crucial representations of ultrasound breast tumor images, their straightforward combination brings little effect for improving the classification of benign and malignant since high-dimensional texture features are too aggressive so that drown out the effect of low-dimensional morphological features. For that, an efficient texture and morphological feature combing method is proposed to improve the classification of benign and malignant. Firstly, both texture (i.e., *local binary patterns* (LBP), *histogram of oriented gradients* (HOG), and *gray-level co-occurrence matrixes* (GLCM)) and morphological (i.e., shape complexities) features of breast ultrasound images are extracted. Secondly, a *support vector machine* (SVM) classifier working on texture features is trained, and a *naive Bayes* (NB) classifier acting on morphological features is designed, in order to exert the discriminative power of texture features and morphological features, respectively. Thirdly, the classification scores of the two classifiers (i.e., SVM and NB) are weighted fused to obtain the final classification result. The low-dimensional nonparameterized NB classifier is effectively control the parameter complexity of the entire classification system combine with the high-dimensional parametric SVM classifier. Consequently, texture and morphological features are efficiently combined. Comprehensive experimental analyses are presented, and the proposed method obtains a 91.11% accuracy, a 94.34% sensitivity, and an 86.49% specificity, which outperforms many related benign and malignant breast tumor classification methods.

1. Introduction

Breast cancer is a common cause of death for women worldwide. According to the global cancer statistics 2018 [1], the incidence and mortality of cancer in China rank the first in the world, among which the incidence of breast cancer is the highest among women and the mortality rate ranks the fifth. Early detection, early diagnosis, and early treatment

are the key to improve the recovery rate of breast cancer and reduce the mortality rate [2]. Therefore, it is desired to develop an effective benign and malignant breast tumor classification method.

Commonly, texture and morphological features of breast ultrasound images are used to analyze the benign and malignant of tumors. The straightforward approach is to rely on high-level and experienced radiologists to judge the benign

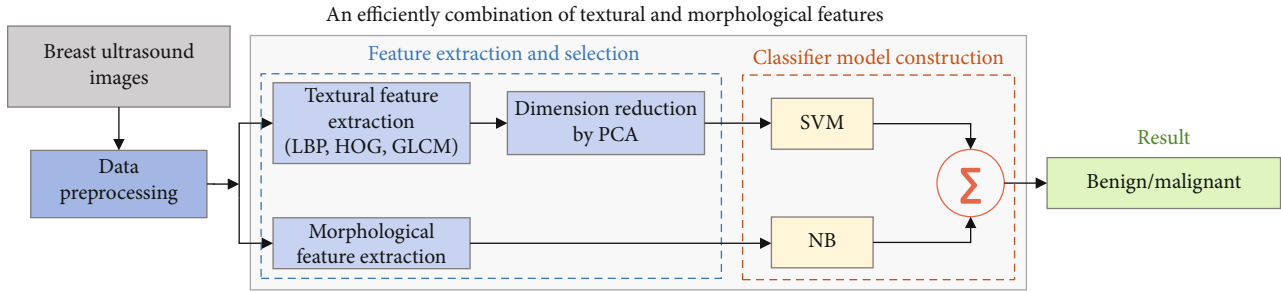


FIGURE 1: A benign and malignant breast tumor classification method via efficiently combining textural and morphological features. Σ represents the weighted fusion of the classification scores of the two classifiers (i.e., SVM and NB).

and malignant of tumors by manually analyzing the texture and morphological features in images [3]. However, the proportion of each feature in the diagnosis in the comprehensive judgment is likely to lead to poor objectivity and repeatability of the diagnosis results due to different doctors' technology and experience. Moreover, ultrasound images themselves also have the disadvantages of high noise and low resolution, which greatly limit the accuracy of artificial ultrasonic detection.

Another straightforward approach is to train classifiers based on texture and morphological features by a computer for classifying benign and malignant tumors automatically to overcome the subjectivity of manually ultrasound image analysis [4]. There are two primary methods of computer automatic analysis. One method is to utilize single features (one of texture features and morphological features [5–8]) with single classifier for computer modeling of breast images. However, this method [5–8] does not fully consider the complementarity of features, and the accuracy of classification is restricted. Another method utilizes multiple features (texture and morphological features) with single classifier [9–12] to take advantage of the complementarity between texture and morphological features. Nevertheless, the direct combination of multiple features will affect the performance of classification such as high-dimensional texture features are too aggressive so that drown out the effect of low-dimensional morphological features [13]. Single classifier cannot solve this problem. Therefore, the main purpose of this article is to effectively combine texture and morphological features to improve the classification performance.

For that, a benign and malignant breast tumor classification method via efficiently combining texture and morphological features is proposed. Figure 1 shows an overview of the proposed method. One can see that two different classifiers are used to train texture and morphological features, respectively, in the proposed method. Firstly, three texture features (*local binary patterns* (LBP) [14], *histogram of gradients* (HOG) [15], *gray-level co-occurrence matrixes* (GLCM) [16]) and three morphological features (compactness, elliptical compactness, and radial distance spectrum) are extracted from 448 collected breast ultrasound images which have been denoised and equalized. Then, the dimensions of texture features are reduced by PCA. Secondly, using *support vector machine* (SVM) [17] classifier and *naive Bayes* (NB) [18] classifier to, respectively, learn texture features and morpho-

logical features. SVM is already a high-dimensional parametric classifier. If one wants to combine multiple classifiers, according to Occam's razor [19], it is reasonable to select a low-dimensional nonparametric classifier to control the parameter complexity of the entire classification system. Thirdly, the outputs of the two classifiers are weighted fused to obtain the final classification result.

This paper is an extension of our preliminary works [20, 21], which improves both methodology and experimental analysis. The contributions of this paper can be summarized as follows. (1) A novel method is proposed to effectively combine multiple features and multiple classifiers to improve the benign and malignant breast tumor classification performance. Specifically, in order to avoid the sharp increase in parameter complexity caused by using multiple classifiers, a nonparameterized NB classifier trained on low-dimensional morphological features is designed to cooperate with a parameterized SVM classifier trained on high-dimensional texture features. (2) Comprehensive experimental analyses are presented to verify the advantage of the proposed method, including data preprocessing, dimension reduction, single feature with single classifier, multiple features with single classifier, and effectively combining multiple features and multiple classifiers.

The rest of this paper is structured as follows. Section 2 introduces the related work. Section 3 describes the feature extraction, the experimental details, and the collected breast ultrasound image dataset. Section 4 presents the experimental results to analyze the effectiveness of the proposed method. Section 5 concludes this paper.

2. Related Work

With the progress of computer technology, medical imaging technology has been greatly developed. It has become a trend to use the computer to classify breast ultrasound images automatically. In this section, an overview of based on hand-crafted features and deep-learned feature methods for breast tumor classification is presented.

2.1. Hand-Crafted Features for Breast Tumor Classification.

In breast ultrasound images, the traditional breast tumor classification technology mainly includes the following four steps [22, 23]: image preprocessing, image segmentation, feature extraction, and tumor classification. Among them,

feature extraction is the main task of breast tumor classification, which has a great impact on the classification results [24]. Texture (i.e., LPB [14], HOG [15], and GLCM [16]) and morphological (i.e., shape complexities) which called hand-crafted features are the key to analyze breast ultrasound images. The hand-crafted feature-based breast tumor classification methods can be roughly divided into two categories.

Firstly, the most common method is to model the breast ultrasound images using single features (one of texture features and morphological features) with single classifier [5–8]. For example, Pomponiu et al. [5] filtered tumors and normal areas based on the histogram of oriented gradients (HOG) descriptor and used SVM to classify the recognized tumors. Mohamed et al. [8] used a superresolution method to preprocess ultrasound images and evaluated the performance of five texture features.

Secondly, many methods are to model the breast ultrasound images using multiple features (texture features and morphological features) with single classifier [9–12]. For example, Menon et al. [10] extracted the textural, morphological, and histogram features of tumor ultrasound images and used SVM to classify tumors. Gonzelezluna et al. [12] extracted 41 morphological features and 96 texture features to analyze the classification effects of 7 classifiers.

In addition, SVM [17], NB [18], *k*-nearest neighbor (KNN) [25], *decision tree* (DT) [26], *linear discriminant analysis* (LDA) [27], and other classifiers are commonly used in hand-crafted feature methods. These classifiers can be divided into two categories: parameterized classifiers and nonparameterized classifiers. Generally, in the process of classification, the calculation of parameterized classifier is complicated which needs to train repeatedly to obtain the best parameters, but this kind of classifier has strong generalization ability on small data sets, such as SVM [17] and KNN [25]. The nonparameterized classifier does not introduce additional parameter complexities although it has poor generalization ability on small data sets, such as NB [18]. When combining multiple features with different classifiers, using two parameterized classifiers will make the training model too complicated, while two nonparameterized classifiers lack strong discrimination learning ability [19]. Therefore, a parameterized classifier with a nonparameterized classifier is proposed to combine multiple features.

2.2. Deep-Learned Features for Breast Tumor Classification. Deep neural networks, powered by advances in computing capability and very large annotated datasets, have achieved revolutionary breakthroughs in computer vision [28]. CNN [29] is the most basic method for classification of breast tumors by deep-learned features. For example, both Zhou et al. [29] and Qi et al. [30] used CNN to extract image features and classify benign and malignant tumors automatically. Other deep-learned features are also applied to the classification of breast tumors. Choi et al. [31] evaluated a computer-aided diagnostic system that combines three deep learning models (Fully Convolutional Network (FCN) [32], AlexNet [33], and GoogLeNet [34]) by comparing the diagnostic results of the doctors and computer.

3. Materials and Methods

3.1. Data Acquisition and Preprocessing. Although there are indeed some breast ultrasound databases, they are not easy to obtain for protecting the privacy of patients. Therefore, a new dataset of breast ultrasound images is collected in Quanzhou First Hospital in Fujian, China, since the public ultrasound images are not easy to obtain and may infringe the patient's privacy. All the images were collected by PHILIPS iu22, PHILIPS iu Elite, and other color ultrasound diagnostic devices with the probe frequency of 12 MHz from 2018 to 2019. The imaging parameters of the ultrasound device were adjusted by radiologists. The images are used with the consent of the relevant patients. Figure 2 shows same examples of the collected ultrasound images.

Cases with previous breast surgery history, poor image quality, and incomplete clinical data were removed, and 448 breast ultrasound images were finally obtained. Among them, 184 are benign tumors, and 264 are malignant tumors. All cases underwent biopsy. According to the definitions of assessment categories in breast imaging reporting and data system (BI-RADS) [3, 35], the final assessment of 448 solid breast tumors on the basis of ultrasound findings is category 2, consider benign changes, for 43 tumors (9.6%); category 3, probably benign tumors, for 50 tumors (11.2%); category 4a, low probability of malignancy, for 91 tumors (20.3%); category 4b, median probability of malignancy, for 77 tumors (17.2%); category 4c, high probability of malignancy, for 66 tumors (14.7%); category 5, highly suspicious of malignancy, for 106 tumors (23.7%); and category 6, malignant tumors, for 15 tumors (3.3%). The collected data covers all tumor categories. Figure 3 shows the distribution of the collected images. For each breast ultrasound image, the *region of interest* (ROI) and outline of tumors are manually annotated by a high-level professional radiologist with more than 10 years of experience. And the annotated results are verified by another experienced radiologist.

The edges of all the images are removed at first. At the same time, due to the presence of speckle noise and low contrast in ultrasound imaging, the ability of the computer to fully extract texture and morphological features will be limited. For this, all the images are denoised by *speckle reducing anisotropic diffusion* (SRAD) filter [10]. Then, the denoised images are equalized using histogram. The result after using SRAD filter and histogram to denoise and equalize the breast ultrasound images is shown in Figure 4. Compared with the original images, the denoised and equalized images show better resolution and contrast.

3.2. Feature Extraction and Selection

3.2.1. Feature Extraction. The feature extraction of breast ultrasound image is a key step in the classification of benign and malignant breast tumors. By extracting a large number of features from ultrasound images and quantifying major diseases such as tumors, the problem of quantitative evaluation of tumor heterogeneity can be effectively solved. It is of certain significance to introduce texture features for tumor analysis since there are significant differences in internal

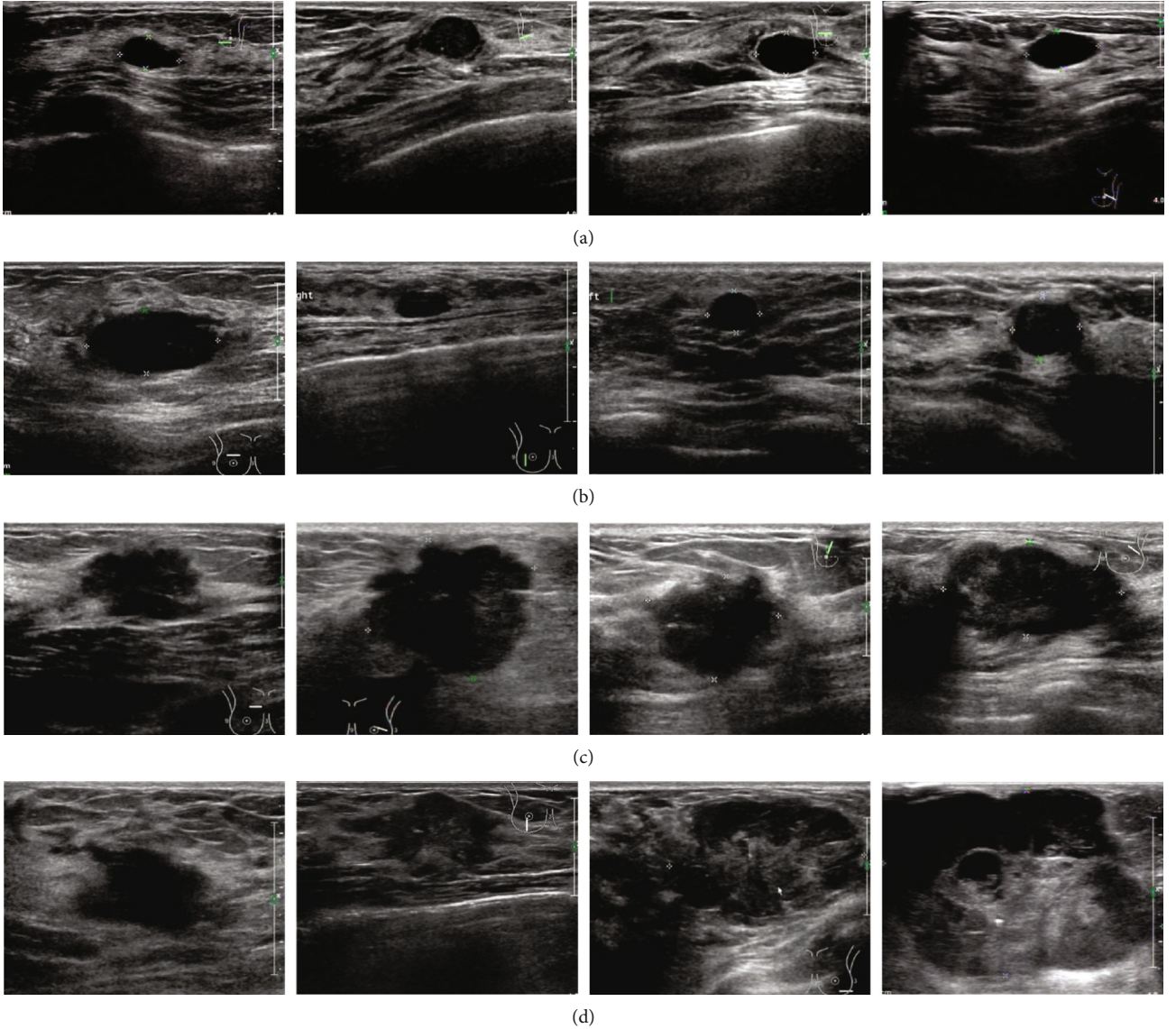


FIGURE 2: Samples of ultrasound images of breast tumors classified according to BI-RADs standard: (a, b) are benign tumors and (c, d) are malignant tumors. Benign tumors are usually well-defined and round or oval in shape. Malignant tumors are usually poorly defined and irregular with lobules.

echoes and boundary echoes of typical benign and malignant tumors in ultrasound imaging. Therefore, the local binary patterns (LBP) [14], histogram of oriented gradients (HOG) [15], and gray-level co-occurrence matrixes (GLCM) [16] features are extracted for classifying. At the same time, benign and malignant tumors often show differences in morphology. It is generally believed that benign tumors are of regular shape, mostly round or oval shape, and the tumor contour itself is relatively smooth. But malignant tumor is on the contrary. Therefore, compactness, elliptical compactness, and radial distance spectrum are extracted to reflect the complexity of tumor contour.

(1) *Texture Features*. The LBP [14] is an operator used to describe local texture features of the image, which has obvious advantages such as rotation invariance and gray invari-

ance. The LBP [14] operator is defined as a 3×3 window. An ordered 8-bit binary number is generated by comparing the size of the central pixel value with the surrounding pixel value (usually converted to LBP [14] code, which is 256 decimal), expressed as follows:

$$\text{LBP}(x, y) = \sum_{p=1}^8 2^{p-1} s(i_p - i_c), \quad (1)$$

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases},$$

where i is the gray value of the center pixel (x, y) , p is the number of the adjacent pixel, i_p is the gray value of the adjacent pixel, and $s(x)$ is the symbolic function.

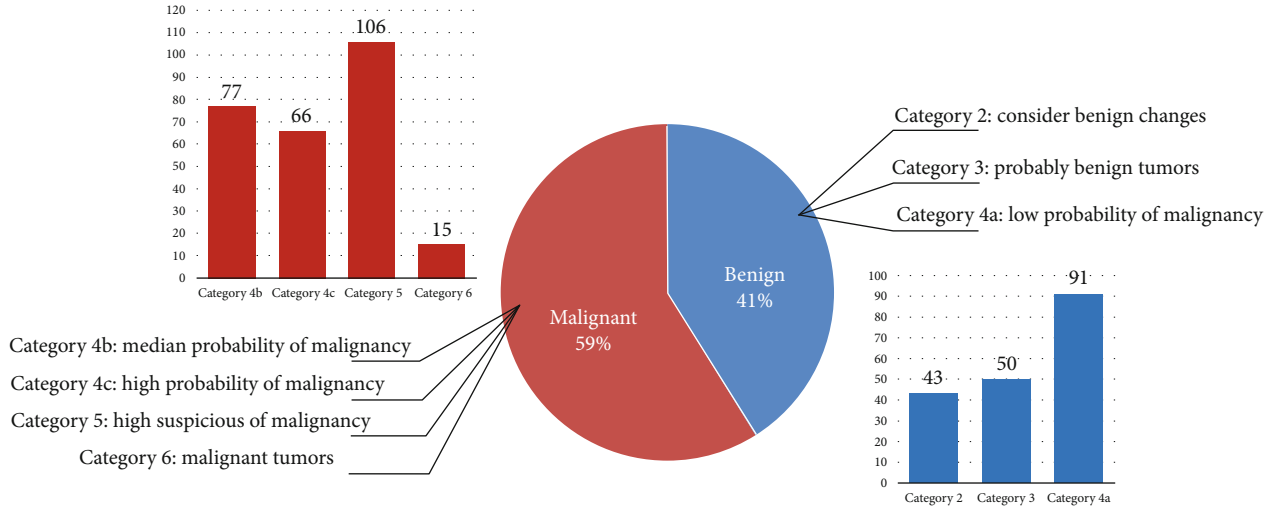


FIGURE 3: Histogram distribution of 448 breast ultrasound images used for texture and morphological analysis.

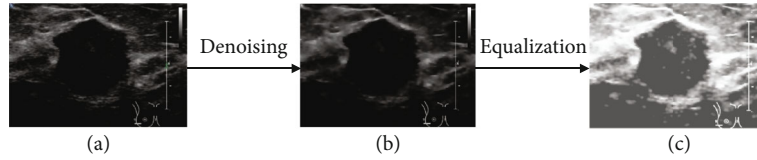


FIGURE 4: The result after using SRAD filter and histogram to denoise and equalize the breast ultrasound images: (a) shows the original image, (b) shows the denoised image, and (c) shows the result after equalization.

The HOG [15] forms the feature by calculating and statistics the histogram of gradient direction in the local area of the image. Firstly, the image is Gamma corrected, and the gradient of each pixel is calculated. Secondly, the image is divided into 32×32 pixel cells in this paper, and the histogram of gradient of each cell is counted to form a descriptor. Finally, every 2×2 cell is concatenated to form a block, and then, all blocks are concatenated to get the HOG [15] feature descriptor.

The GLCM [16] extracts the relationship between the pixel pairs. In this paper, the grayscale level is set to 64. The distance between pixels is adjusted within the range of [1, 10], and the relationship between pixels with a certain distance is calculated from four directions (0, 45, 90, 135). Finally, 40 different matrices are obtained from each image. The energy, contrast, correlation, and homogeneity are extracted from matrices to reflect the roughness of the texture, the local variation, and the uniformity of the gray distribution of the image.

(2) *Morphological Features.* Morphological features are obtained by calculating the compactness (equation (2)), the elliptic compactness (equation (3)), and the mean and variance of the radial distance spectrum (equation (4)) of the tumor. The tumor has the potential to be malignant if the shape of the tumor looks like irregular lobules, rather than just round or oval [8].

Compactness measures the similarity between the shape of a breast tumor and its fitting circle. The closer the com-

pactness value is to 1, the less likely the tumor is to be malignant, expressed as follows:

$$C = \frac{A}{4\pi L^2}, \quad (2)$$

where A represents the area of the tumor and L is the perimeter of the breast tumor contour.

The elliptic compactness is the ratio of the circumference of the fitting ellipse to the circumference of the original tumor contour. It is negatively correlated with the degree of malignancy of the tumor. The elliptic fitting method is to find an ellipse for a given set of tumor contour points and make it as close as possible to these contour points. More generally, the contour points of the tumor are fitted with the elliptic equation as the model so that a certain elliptic equation can satisfy these points as far as possible, and each parameter of the elliptic equation is obtained. Here, used the least square method proposed by Fitzgibbon et al. [36] for ellipse fitting. The effect of ellipse fitting is shown in Figure 5. The blue line is the contour of the tumor, and the red line is the fitting ellipse. According to the fitting ellipse obtained, the features are calculated as follows:

$$EC = \frac{\pi(a+b)}{D}, \quad (3)$$

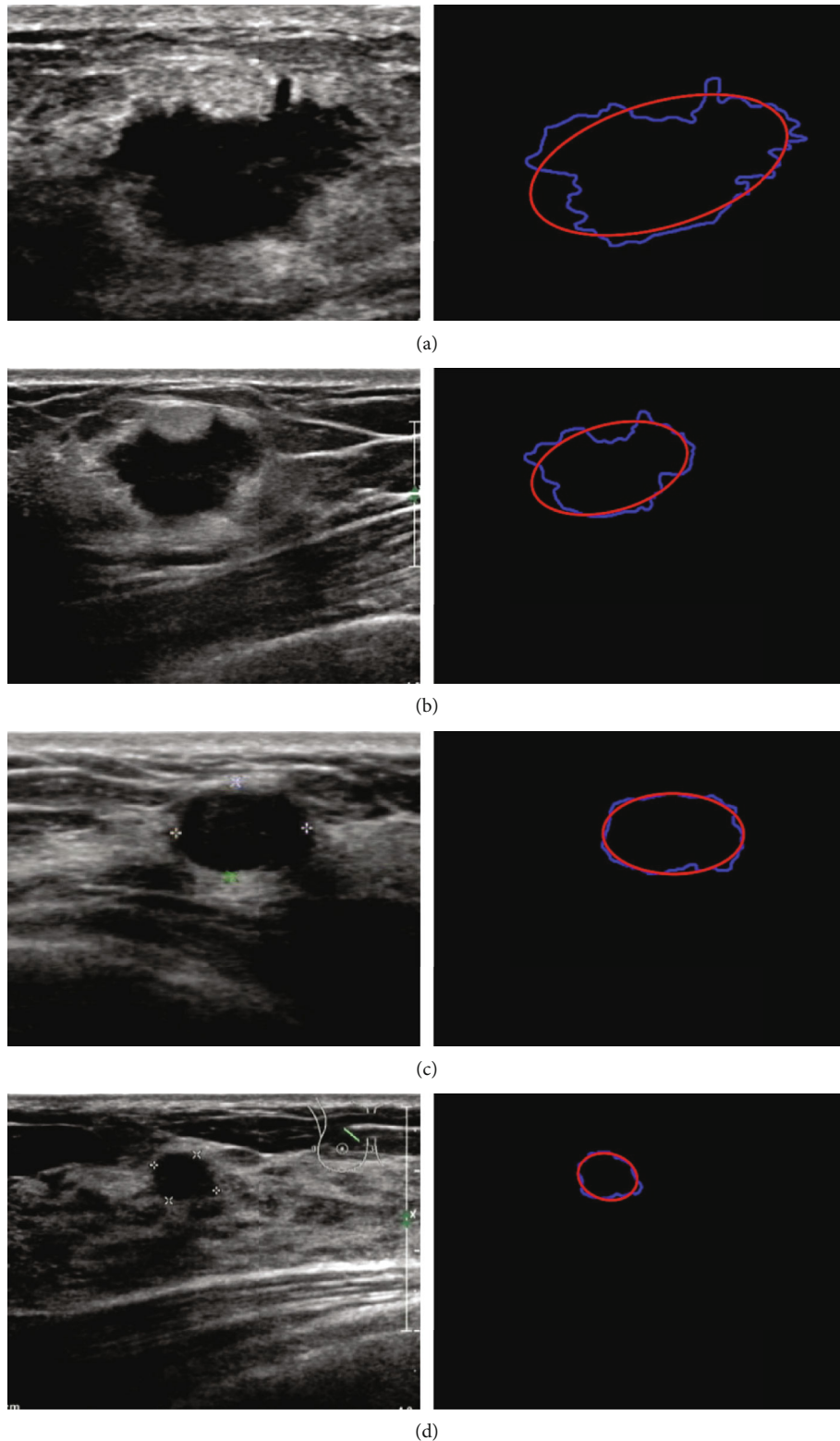


FIGURE 5: The examples of the fitting ellipse that transformed from breast tumor contour: (a, b) malignant tumor and (c, d) benign tumor.

where a represents the semimajor axis of the fitting ellipse, b is the semiminor axis of the fitting ellipse, and D is the perimeter of the breast tumor contour.

Radial distance spectrum method quantified the degree of tumor margin roughness by statistical and analyzing the

radial distance from each point on the tumor margin to the tumor center. In this paper, Fourier transform is applied to the obtained radial distance spectrum, and its logarithm is taken to obtain the logarithmic amplitude spectrum of radial distance. Finally, the mean and variance of harmonic

components in the logarithmic amplitude spectrum are taken as characteristic parameters. Radial distance can be calculated as follows:

$$D(t) = \sqrt{(p_t - x_0)^2 + (q_t - y_0)^2}, \quad (4)$$

where the tumor edge points are denoted as $P_t(p_t, q_t)$ and the center point is denoted as (x_0, y_0) .

3.2.2. Feature Selection. In this paper, the *principal component analysis* (PCA) [37] is used to reduce the dimension of extracted texture features in order to speed up the training and testing time and improve the efficiency of the proposed method.

3.3. Experiments

3.3.1. Experimental Setup. It is well-known that texture and morphological features are complementary in the ultrasound image. However, the classification ability via combining texture and morphological features directly will be limited because of the aggressiveness of high-dimensional texture features. For that, a classification method for benign and malignant breast tumor via efficiently combining texture and morphological features is proposed in this paper. The specific process is shown in Figure 1. The collected breast ultrasound images are randomly divided into training set (80%) and test set (20%); then, all images are preprocessed. Three texture features (i.e., LBP [14], HOG [15], and GLCM [16]) and three morphological features (compactness, elliptical compactness, and radial distance spectrum) are extracted and normalized. The dimensions of the extracted texture features are reduced by PCA [37]. On the account of high-dimensional texture features can easily affect low-dimensional morphological features in the single classifier, support vector machine (SVM) [17] and naive Bayes (NB) [18] classifiers are used to learn texture features and morphological features, respectively, in this paper. SVM is already a high-dimensional parametric classifier. If one wants to combine multiple classifiers, according to Occam's razor [19], it is reasonable to select a low-dimensional nonparametric classifier to control the parameter complexity of the entire classification system. Finally, the classification scores of the two classifiers are weighted fused (equation (5)) to obtain the final classification result:

$$S_c(\lambda) = S_{SVM} \times \lambda + S_{NB} \times (1 - \lambda), \quad (5)$$

where λ represents the weight, ranging from 0 to 1; S_{SVM} is the score of malignant classification output by SVM classifier; S_{NB} is the score of malignant classification output by NB classifier; and $S_c(\lambda)$ represents the weighted fusion of the classification scores of two classifiers (SVM and NB) and its values between 0 and 1. When the value of $S_c(\lambda)$ is greater than or equal to 0.5, the tumor is considered malignant; when the value of $S_c(\lambda)$ is less than 0.5, the tumor is considered benign.

Comprehensive experimental analyses are presented, and the experiment is divided into three parts to compare and

analyze the advantages of the proposed method. In the first part, the classification performance of using single features with single classifier is evaluated and compared. In the second part, the classification performance of using multiple features with single classifier is evaluated and compared. In the third part, the classification performance of using multiple features with multiple classifiers is evaluated and compared. Another three classifiers (k-nearest neighbor (KNN) [25], decision tree (DT) [26], and linear discriminant analysis (LDA) [27]) are used to analyze the three extracted texture features and three morphological features in order to verify the superiority of the proposed method. The methods of analyzing features include single features and combined multiple features.

In this work, the parameters of each classifier are optimized to improve the classification performance. In SVM [17], the radial basis function is used as the kernel function, and the mesh search method is used to perform the 5-fold cross-validation to automatically find the optimal penalty factor c and the kernel parameter g . The number of neighbors in KNN [25] is set to 5.

3.3.2. Evaluation Criterion. The classification performance is quantitatively measured by accuracy, sensitivity, and specificity [38]. In addition, the receiver operating characteristic (ROC) curve analysis is used to evaluate the performance of classifiers. The area under the curve (AUC) is calculated based on the ROC to measure the ability of features to distinguish benign and malignant tumors.

4. Results and Discussion

4.1. Experimental Results. The result of the proposed method is verified through a comparison in the following. Support vector machine (SVM) [17] and naive Bayes (NB) [18] classifiers are used to effectively learn texture features (local binary patterns (LBP) [14], histogram of gradients (HOG) [15], gray-level co-occurrence matrixes (GLCM) [16]) and morphological features (compactness, elliptical compactness, and radial distance spectrum), respectively, in this paper. In order to show the superiority of the proposed method, this paper compares it with the related methods [5, 7, 8, 10, 12]. The experiments are mainly completed on Matlab 2017b.

It can be seen from Table 1 that the hand-crafted feature method can learn a small sample well to get a better classification effect. In addition, the experimental results show that the classification performance of multiple features is often better than single feature, and our method takes full advantage of the complementarity of texture and morphological features to get the better performance than single classifier. The performance of our method is superior to other related methods, with the accuracy of 91.11%, the sensitivity of 94.34%, and the specificity of 86.49%. The effective combination of multiple features and multiple classifiers can effectively improve the classification of benign and malignant breast tumors.

4.2. Discussion. In order to prove the effectiveness of the proposed method, the following analysis and discussion are

TABLE 1: The performance comparison of our method and multiple related methods.

Method		Evaluation (%)		
		Accuracy	Sensitivity	Specificity
Single feature with single classifier (SFSC)	Pomponiu et al. [5]	81.11	84.91	75.68
	Biswas et al. [7]	75.56	67.92	86.49
	Mohamed et al. [8]	84.44	84.91	83.78
Multiple features with single classifier (MFSC)	Menon et al. [10]	87.78	88.68	86.49
	Gonzelezluna et al. [12]	86.67	88.68	83.78
Multiple features with multiple classifiers (MFMC)	Our method	91.11	94.34	86.49

TABLE 2: The classification results based on the methods of single features with single classifier.

Method Feature	Classifier	Accuracy	Evaluation (%)	
			Sensitivity	Specificity
LBP	SVM [17]	85.56	86.79	83.78
	KNN [25]	84.44	84.91	83.78
	DT [26]	66.33	58.49	81.08
	LDA [27]	74.44	77.36	70.27
HOG	SVM [17]	81.11	84.91	75.68
	KNN [25]	61.11	100.00	5.41
	DT [26]	67.78	67.92	67.57
	LDA [27]	70.00	75.47	62.16
GLCM	SVM [17]	78.89	92.45	59.46
	KNN [25]	65.56	75.47	51.35
	DT [26]	71.11	77.36	62.16
	LDA [27]	74.44	84.91	59.46
LBP+HOG+GLCM	SVM [17]	86.67	92.45	78.38
	KNN [25]	64.44	100.00	13.51
	DT [26]	72.22	73.58	70.27
	LDA [27]	75.56	84.91	62.16
Morphological	SVM [17]	75.56	67.92	86.49
	NB [18]	81.11	69.81	97.30
	LDA [26]	75.56	60.38	97.30

carried out. Another three classifiers (k-nearest neighbor (KNN) [25], decision tree (DT) [26], and linear discriminant analysis (LDA) [27]) with SVM [17] and NB [18] are used to analyze the three extracted texture features (LBP [14], HOG [15], GLCM [16]) and three morphological features (compactness, elliptical compactness, and radial distance spectrum). The experimental analysis will be carried out from three subsections as follows.

4.2.1. Experiments Based on Classification Methods Using Single Features with Single Classifier. Based on LBP [14], HOG [15], GLCM [16] texture features, fused texture features, and morphological features, the detailed data of sensitivity, specificity, and accuracy of model prediction are shown in Table 2.

Compared with the classification results of different texture features in Table 2, the classification results based

on the fused texture features are the best, with the accuracy reaching 86.67%, the sensitivity reaching 92.45%, and the specificity reaching 78.38%. The second best feature is LBP [14], which achieves 85.56% in accuracy, 86.79% in sensitivity, and 83.78% in specificity. The accuracy of HOG [15] feature is 81.11%, the sensitivity is 84.91%, and the specificity is 75.68%. The accuracy of GLCM [16] is 78.89%, the sensitivity is 92.45%, and the specificity is 59.46%. For morphological features, the accuracy is 81.11%, the sensitivity is 69.81%, and the specificity is 97.30%.

By comparing the classification results of different classifiers in Table 2, the classification results of SVM [17] classifier are generally higher than those of other classifiers, with the accuracy reaching 86.67%, the sensitivity reaching 92.45%, and the specificity reaching 78.38%. The second is KNN [25] classifier, with accuracy of 84.44%, sensitivity of 84.91%, and specificity of 83.08%. The NB [18] classifier

TABLE 3: The classification results based on the methods of multiple features with single classifier.

Features	Evaluation (%)	Classifier	
		SVM	LDA
LBP+ morphological	Accuracy	80.00	76.67
	Sensitivity	79.25	75.47
	Specificity	81.08	78.38
HOG+ morphological	Accuracy	83.33	72.22
	Sensitivity	81.13	71.70
	Specificity	86.48	72.97
GLCM+ morphological	Accuracy	80.00	80.00
	Sensitivity	69.81	81.13
	Specificity	94.59	78.38
LBP+HOG+GLCM+ morphological	Accuracy	87.78	76.67
	Sensitivity	88.68	75.47
	Specificity	86.49	78.38

achieves 81.11% in accuracy, 69.81% in sensitivity, and 97.30% in specificity. The accuracy of LDA [27] classifier is 75.56%, the sensitivity is 60.38%, and the specificity is 97.30%. The accuracy of DT [26] classifier can reach 72.22%, the sensitivity is 73.58%, and the specificity is 60.27%.

By comparing the experimental results in Table 2, the best results of the single feature and single classifier classification methods are the fused texture features (LBP+HOG+GLCM) with SVM classifier. However, the complementarity of features is not fully considered to restrict the accuracy of classification in these methods of using single features with single classifier.

4.2.2. Experiments Based on Classification Methods Using Multiple Features with Single Classifier. Multiple features provide a good way to identify benign and malignant tumors well [39] by considering the complementarity of texture and morphological features. Therefore, the classification method based on multiple features with single classifier is analyzed and discussed experimentally. The accuracy, sensitivity, and specificity are shown in Table 3.

Compared with the different combination of multiple features (texture features and morphological features) with single classifier (SVM and LDA), the classification result of the method using the fused texture features (LBP+HOG+GLCM) and morphological features with SVM classifier is the best. The accuracy, sensitivity, and specificity are 87.78%, 88.68%, and 86.49%, respectively. From the analysis of the experimental results in Tables 2 and 3, it can be concluded that the classification result is not ideal, although the method of multiple features with single classifier which straightforward combining texture features and morphological features have considered the complementarity of texture features and morphological features. This is because that the method of straightforward combining multiple features with single classifier has not consider the aggressiveness of high-dimensional texture features to low-dimensional morphological features.

TABLE 4: The classification results based on the method of multiple features with multiple classifiers.

Method	Evaluation (%)		
	Accuracy	Sensitivity	Specificity
SVM (LBP+HOG+GLCM) [17]	86.67	92.45	78.38
NB (morphological) [18]	81.11	69.81	97.30
Our method	91.11	94.34	86.49

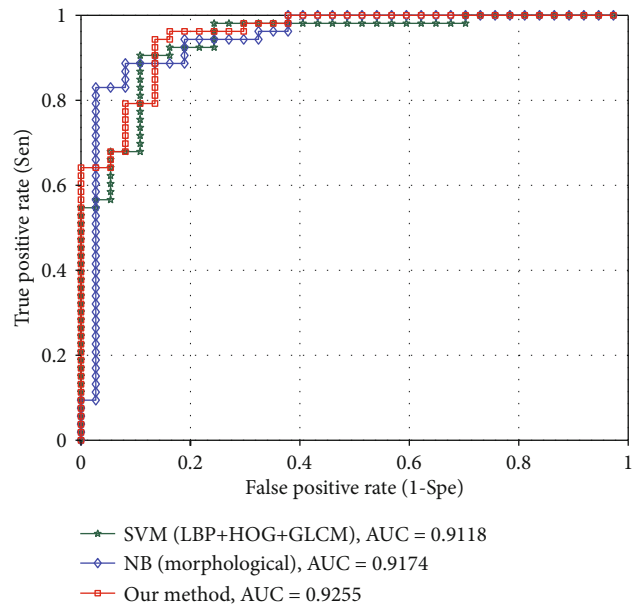


FIGURE 6: The ROC curve of different combinations of texture and morphological features with different classifiers.

4.2.3. Experiments Based on Classification Methods Using Multiple Features with Multiple Classifiers. Based on the above analysis, the method of using SVM classifier and NB

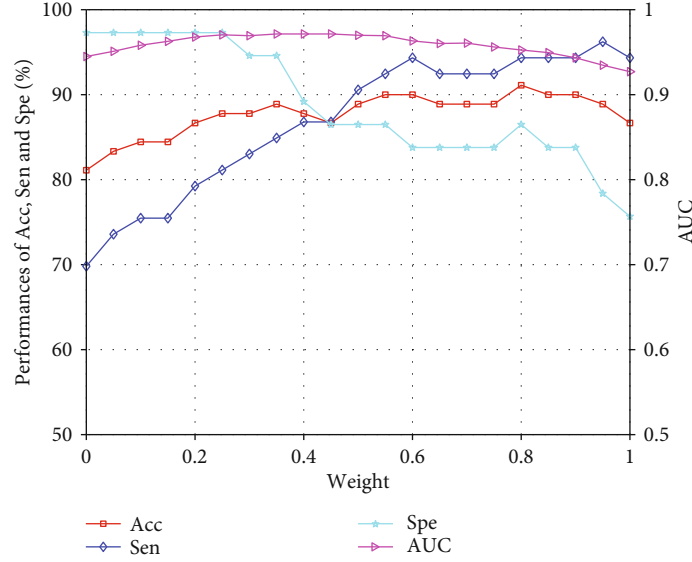


FIGURE 7: The classifier weighted fusion analysis diagram.

classifier working on texture and morphological features, respectively, is proposed in order to exert the discriminative power of texture features and morphological features. SVM is already a high-dimensional parametric classifier. If one wants to combine multiple classifiers, according to Occam's razor [19], it is reasonable to select a low-dimensional non-parametric classifier to control the parameter complexity of the entire classification system. The accuracy, sensitivity, and specificity are shown in Table 4.

From the analysis of the experimental results in Tables 3 and 4, it can be concluded that the proposed method of using SVM and NB classifier to effectively combine texture and morphological features has fully considered the complementarity of texture and morphological features and eliminates the aggressive of high-dimensional texture features to low-dimensional morphological features. The proposed methods are 3.33% and 5.66% higher than the method of using SVM [17] to directly combine texture and morphological features in the accuracy and sensitivity, respectively. At the same time, the accuracy of the proposed method is about 4.44% higher than the highest accuracy of single feature. The ROC curves of the three methods in Table 4 are shown in Figure 6. The AUC of the method based on texture feature and SVM [17] classifier reached 0.9118. The AUC based on morphological features and NB [18] classifier method reached 0.9174. The AUC of the proposed method is 0.9225. The final classification result of the proposed method is obtained from the weighted fused of the classification scores of SVM [17] and NB [18] classifiers as shown in equation (5). Figure 7 shows the weight analysis of weighted fusion. When the weight is set from 0.6 to 0.9, the proposed method performs well. Among them, the accuracy is the highest when the weight is 0.8, and both sensitivity, specificity, and AUC are taken into account.

4.2.4. Effect Analysis of Image Preprocessing and Feature Selection. To confirm that denoising and equalization have

TABLE 5: The accuracy (%) based on breast ultrasound image preprocessing.

Accuracy (%)	Features			
	LBP	HOG	GLCM	LBP + HOG + GLCM
Before preprocessing	81.11	80.00	71.11	82.22
After preprocessing	85.56	81.11	78.89	86.67

TABLE 6: The elapsed time before and after dimension reduction based on PCA.

	Time/s	
	Before dimension reduction	After dimension reduction
LBP+HOG+GLCM (SVM)	0.3601	0.0135

an auxiliary effect on classification of ultrasound image, the classification experiment is also performed using images without denoising and equalization, and the two results are compared. Due to the noise and contrast of the image that have little effect on the morphological features, this paper compares the experiments that only extract the texture features. The SVM classifier with the best classification performance is preprocessed to analyze the texture features. As shown in Table 5, preprocessing helps extract more useful texture features from images, efficiently improving accuracy.

The dimension of texture feature extracted for the first time is too large. The eigenvector matrix can be reduced by using PCA to retain the most effective features. As can be seen from Table 6, after dimension reduction of texture features, the time required for testing is greatly reduced, which improves efficiency of the whole classification method. Dimension reduction reduces the training time to 0.0135s, 1/26 of the training time before dimension reduction.

5. Conclusion

In this paper, an efficient texture and morphological feature combining method is proposed to improve the classification performance of benign and malignant tumors in ultrasound imaging. Firstly, the texture features (i.e., local binary patterns (LBP), histogram of oriented gradients (HOG), and gray-level co-occurrence matrixes (GLCM)) and morphological features (i.e., compactness, elliptical compactness, and radial distance spectrum) are extracted from the collected 448 breast tumor ultrasound images after denoised and equalized. Secondly, support vector machine (SVM) and naive Bayes (NB) classifiers are used to learn texture features and morphological features, respectively, since high-dimensional texture features can easily affect low-dimensional morphological features in the single classifier. Finally, the classification scores of the two classifiers are weighted fused to obtain the final classification result. The low-dimensional nonparameterized NB classifier is effectively control the parameter complexity of the entire classification system combine with the high-dimensional parametric SVM classifier. Comprehensive experimental analyses are presented to verify the effectiveness of the proposed method that another three classifiers (i.e., k-nearest neighbor (KNN), decision tree (DT), and linear discriminant analysis (LDA)) are used to analyze the three extracted texture features and three morphological features in order to verify the superiority of the proposed method. The methods of analyzing features include single features and combined multiple features. Experimental results show that the proposed method has the best accuracy, sensitivity, and specificity. This provides a rapid, low-cost, and repeatable diagnostic method for the ultrasound examination of breast tumors and has certain feasibility and good robustness.

Data Availability

The Breast Ultrasound Image data used to support the findings of this study were supplied by Quanzhou first hospital in Fujian, China, under license and so cannot be made freely available.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (No. ZQN-PY518). The grants from National Natural Science Foundation of China (Grant No. 61605048), in part by the Quanzhou scientific and technological planning projects (No. 2019C028R, 2019C029R, 2018C113R, and 2018N072S), and in part by the Subsidized Project for Postgraduates Innovative Fund in Scientific Research of Huaqiao University under Grant 18014084012.

References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] C. E. DeSantis, J. Ma, A. Goding Sauer, L. A. Newman, and A. Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 6, pp. 439–448, 2017.
- [3] T. Steifer and M. Lewandowski, "Ultrasound tissue characterization based on the Lempel–Ziv complexity with application to breast lesion classification," *Biomedical Signal Processing and Control*, vol. 51, pp. 235–242, 2019.
- [4] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: a survey," *Pattern Recognition*, vol. 43, no. 1, pp. 299–317, 2010.
- [5] V. Pomponiu, H. Hariharan, B. Zheng, and D. Gur, "Improving breast mass detection using histogram of oriented gradients," *Medical Imaging 2014: Computer-Aided Diagnosis*, vol. 9035, article 90351R, 2014 International Society for Optics and Photonics.
- [6] A. A. Ardakani, A. Gharbali, and A. Mohammadi, "Classification of breast tumors using sonographic texture analysis," *Journal of Ultrasound in Medicine*, vol. 34, no. 2, pp. 225–231, 2015.
- [7] R. Biswas, A. Nath, and S. Roy, "Mammogram classification using gray-level co-occurrence matrix for diagnosis of breast cancer," in *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, pp. 161–166, Ghaziabad, India, 2016.
- [8] M. Abdel-Nasser, J. Melendez, A. Moreno, O. A. Omer, and D. Puig, "Breast tumor classification in ultrasound images using texture analysis and super-resolution methods," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 84–92, 2017.
- [9] W. Gómez Flores, W. C. A. Pereira, and A. F. C. Infantosi, "Improving classification performance of breast lesions on ultrasonography," *Pattern Recognition*, vol. 48, no. 4, pp. 1125–1136, 2015.
- [10] R. V. Menon, P. Raha, S. Kothari, S. Chakraborty, I. Chakrabarti, and R. Karim, "Automated detection and classification of mass from breast ultrasound images," in *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pp. 1–4, Patna, India, 2015.
- [11] A. Rodríguez-Cristerna, W. Gómez-Flores, and W. C. de Albuquerque Pereira, "A computer-aided diagnosis system for breast ultrasound based on weighted BI-RADS classes," *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 33–40, 2018.
- [12] F. A. González-Luna, J. Hernández-López, and W. Gomez-Flores, "A performance evaluation of machine learning techniques for breast ultrasound classification," in *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pp. 1–5, Mexico City, Mexico, 2019.
- [13] W. C. Shen, R. F. Chang, W. K. Moon, Y. H. Chou, and C. S. Huang, "Breast ultrasound computer-aided diagnosis using BI-RADS features," *Academic Radiology*, vol. 14, no. 8, pp. 928–939, 2007.

- [14] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, San Diego, CA, USA, 2005.
- [16] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.
- [17] 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.
- [18] I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, pp. 41–46, 2001.
- [19] V. Balasubramanian, "Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions," *Neural Computation*, vol. 9, no. 2, pp. 349–368, 1997.
- [20] M. Wei, X. Wu, J. Zhu et al., "Multi-feature fusion for ultrasound breast image classification of benign and malignant," in *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pp. 474–478, Xiamen, China, 2019.
- [21] M. Wei, Y. Du, X. Wu, and J. Zhu, "Automatic classification of benign and malignant breast tumors in ultrasound image with texture and morphological features," in *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, pp. 126–130, Xiamen, China, 2019.
- [22] K. M. Meiburger, U. R. Acharya, and F. Molinari, "Automated localization and segmentation techniques for B-mode ultrasound images: a review," *Computers in Biology and Medicine*, vol. 92, pp. 210–235, 2018.
- [23] M. Elawady, I. Sadek, A. E. R. Shabayek, G. Pons, and S. Ganau, "Automatic nonlinear filtering and segmentation for breast ultrasound images," in *International Conference on Image Analysis and Recognition*, pp. 206–213, Póvoa de Varzim, Portugal, 2016.
- [24] W. K. Moon, I. L. Chen, J. M. Chang, S. U. Shin, C. M. Lo, and R. F. Chang, "The adaptive computer-aided diagnosis system based on tumor sizes for the classification of breast tumors detected at screening ultrasound," *Ultrasonics*, vol. 76, pp. 70–77, 2017.
- [25] R. Min, D. A. Stanley, Z. Yuan, A. Bonner, and Z. Zhang, "A deep non-linear feature mapping for large-margin knn classification," in *2009 Ninth IEEE International Conference on Data Mining*, pp. 357–366, Miami, FL, USA, 2009.
- [26] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
- [27] L. J. Hargrove, E. J. Scheme, K. B. Englehart, and B. S. Hudgins, "Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 1, pp. 49–57, 2010.
- [28] Q. Huang, F. Zhang, and X. Li, "Machine learning in ultrasound computer-aided diagnostic systems: a survey," *BioMed Research International*, vol. 2018, Article ID 5137904, 10 pages, 2018.
- [29] Y. Zhou, J. Xu, Q. Liu et al., "A radiomics approach with CNN for shear-wave elastography breast tumor classification," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1935–1942, 2018.
- [30] X. Qi, L. Zhang, Y. Chen et al., "Automated diagnosis of breast ultrasonography images using deep neural networks," *Medical Image Analysis*, vol. 52, pp. 185–198, 2019.
- [31] J. S. Choi, B. K. Han, E. S. Ko et al., "Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography," *Korean Journal of Radiology*, vol. 20, no. 5, pp. 749–758, 2019.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, 2015.
- [33] W. Nawaz, S. Ahmed, A. Tahir, and H. A. Khan, "Classification of breast cancer histology images using ALEXNET," in *International Conference Image Analysis and Recognition*, pp. 869–876, Póvoa de Varzim, Portugal, 2018.
- [34] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, 2015.
- [35] C. L. Mercado, "BI-RADS update," *Radiologic Clinics of North America*, vol. 52, no. 3, pp. 481–487, 2014.
- [36] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476–480, 1999.
- [37] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.
- [38] N. Ohuchi, A. Suzuki, T. Sobue et al., "Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan strategic anti-cancer randomized trial (J-START): a randomised controlled trial," *The Lancet*, vol. 387, no. 10016, pp. 341–348, 2016.
- [39] S. K. Alam, E. J. Feleppa, M. J. Rondeau, A. Kalisz, and B. S. Garra, "Computer-aided diagnosis of solid breast lesions using an ultrasonic multi-feature analysis procedure," *Bangladesh Journal of Medical Physics*, vol. 4, no. 1, pp. 1–10, 2013.

Research Article

Comprehensive Analysis of Immunoinhibitors Identifies LGALS9 and TGFBR1 as Potential Prognostic Biomarkers for Pancreatic Cancer

Yue Fan,¹ Tianyu Li,¹ Lili Xu,¹ and Tiantao Kuang^{1,2} 

¹Department of Integrated TCM & Western Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China

²Department of General Surgery, Zhongshan Hospital, Fudan University, Shanghai 200032, China

Correspondence should be addressed to Tiantao Kuang; kuang.tiantao@zs-hospital.sh.cn

Received 8 June 2020; Accepted 21 July 2020; Published 30 September 2020

Guest Editor: Lei Chen

Copyright © 2020 Yue Fan 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.

Pancreatic cancer (PC) is one of the most deadly cancers worldwide. To uncover the unknown novel biomarker used to indicate early diagnosis and prognosis in the molecular therapeutic field of PC is extremely of importance. Accumulative evidences indicated that aberrant expression or activation of immunoinhibitors is a common phenomenon in malignances, and significant associations have been noted between immunoinhibitors and tumorigenesis or progression in a wide range of cancers. However, the expression patterns and exact roles of immunoinhibitors contributing to tumorigenesis and progression of pancreatic cancer (PC) have not yet been elucidated clearly. In this study, we investigated the distinct expression and prognostic value of immunoinhibitors in patients with PC by analyzing a series of databases, including TISIDB, GEPIA, cBioPortal, and Kaplan-Meier plotter database. The mRNA expression levels of IDO1, CSF1R, VTCN1, KDR, LGALS9, TGFBR1, TGFB1, IL10RB, and PVRL2 were found to be significantly upregulated in patients with PC. Aberrant expression of TGFBR1, VTCN1, and LGALS9 was found to be associated with the worse outcomes of patients with PC. Bioinformatics analysis demonstrated that LGALS9 was involved in regulating the type I interferon signaling pathway, interferon-gamma-mediated signaling pathway, RIG-I-like receptor signaling pathway, NF-kappa B signaling pathway, cytosolic DNA-sensing pathway, and TNF signaling pathway. And TGFB1 was related to mesoderm formation, cell matrix adhesion, TGF-beta signaling pathway, and Hippo signaling pathway. These results suggested that LGALS9 and TGFBR1 might serve as potential prognostic biomarkers and targets for PC.

1. Introduction

The mortality of pancreatic cancer (PC), the fourth widely occurred cancer with poor prognosis, has an overall five-year survival rate lower than 10% [1]. Due to the hidden symptoms at early stages, fewer than 15% of patients are diagnosed with PC at a stage when they could be eligible for curative surgical resection [2]. To improve early detection and prognosis and to provide timely and effective treatment for high-risk patients, predictive biomarkers for PC are required [3, 4]. To date, carbohydrate antigen 199 (CA199) and CA242, which are currently used in clinical settings as serum biomarkers for PC, are inadequate for early screening and prognosis [5, 6]. To uncover the unknown novel biomarker used to indicate early diagnosis and prognosis in

the molecular therapeutic field of PC is extremely of importance.

Previously, the application of the immune system to recognize and eradicate tumors has made significant advance in the clinical use of cancer immunotherapy [7–9]. Notably, the emergence of immune checkpoints inhibitors typically interfered negative regulators of T cell immunity including LAG3 [10–12], CTLA-4 [13, 14], PD-1 [15, 16], and TIM3 [17, 18]. The advent of these “checkpoint inhibitors” has thoroughly altered and improved the former therapies for melanoma, lung cancer, and so on [19]. For instance, interference of LAG3 relieved the exhaustion of T cells and heightened immunity against tumor due to the interaction among LAG3 with MHC class II and galectin 3 [20]. Additionally, tumor-infiltrating lymphocyte- (TIL-) produced TIM3 has

been identified to display a key role in maintaining inactive lymphocyte status or inducing lymphocyte apoptosis [21]. LGALS9 is a ligand of TIM-3 and expressed in a variety of cell types, especially in lymphoid organs and monocytes [22]. In addition, LGALS9 could impose unequal effects on immune cells in a tumor microenvironment [22]. TGF- β signaling exhibited importance in biological signal regulation including cell growth and death, differentiation, angiogenesis, and inflammation [23]. Several recent studies demonstrated that TGF- β signaling played a key role in immune response [24]. Understanding the potential functions and expression pattern of immune “checkpoint inhibitors” could be helpful for the identification of novel prognosis and treatment biomarkers for PC.

The occurrence and progression of newly produced strategies, comprising microarray and RNA-sequencing, exerted a positive effect in molecular research and also gave impetus to exploring accurate and safe treatment for PC [25–27]. Here, we expanded PC-related knowledge in view of different databases, thus generating a conclusive analysis of the link between the function of immune checkpoint inhibitors and the diagnosis along with the development of PC.

2. Materials and Methods

2.1. Survival Analysis. Kaplan-Meier plotter (<http://www.kmplot.com/>) is an online database containing microarray gene expression data, and survival information derived from Gene Expression Omnibus, TCGA, and the Cancer Biomedical informatics Grid. Kaplan-Meier (K-M) Plotter database was used to analyze prognostic parameter of expected candidates [28]. K-M survival curves and logrank test were performed to disintegrate correlation, such as gene expression with overall survival (OS) or first progression (FP) or post progression survival (PPS), respectively. Significant difference was indicated as $P < 0.05$.

2.2. Construction of Protein Interaction Network. A functional protein interaction network was constructed as indicated in website (<http://string-db.org/>) [29]. Among them, 50 selected proteins indeed associating with Homo sapiens were selected, followed by calculating confidence score as more than 0.9.

2.3. TISIDB, GEPIA, TCGA, and cBioPortal Analysis. TISIDB is an integrated repository portal for tumor-immune system interactions. The present study used TISIDB (<http://cis.hku.hk/TISIDB>) database to detect the relationship between centromere protein expression and clinical stages, lymphocytes, immunomodulators, and chemokines in PC. Gene Expression Profiling Interactive Analysis (GEPIA) [30] was a powerful tool to determine key interactive and customizable functions including differential expression analysis, profiling plotting, correlation analysis, patient survival analysis, similar gene detection, and dimensionality reduction analysis, which was used to determine mRNA expression in 9,736 tumors and 8,587 normal tissues. The cBioPortal system was used to investigate cancer genomic and clinical-related characters within 105 cancer subjects in

the TCGA pipeline cancer [31]. Besides, the coexpression and interaction of selected proteins were probed referred to cBioPortal guidelines [32].

2.4. Gene Ontology and Pathway Enrichment Analysis. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed using DAVID online tool. $P < 0.01$ was set as the cut-off criterion.

2.5. Statistical Analysis. Student’s *t*-test was analyzed for statistical significance. Statistical analysis was performed by SPSS 21.0 (SPSS Inc., Chicago, IL).

3. Results

3.1. Identification of Immunoinhibitor Expression Pattern in PC. The present study analyzed the expression pattern of 23 immunoinhibitors in PC using TCGA database, including CD160, CD244, KIR2DL1, KIR2DL3, BTLA, CSF1R, HAVCR2, TIGIT, LAG3, PDCD1, VTCN1, PDCD1LG2, LGALS9, CD96, TGFBR1, TGFBI, CTLA4, ADORA2A, PVRL2, IL10, IDO1, IL10RB, and KDR. As shown in Figure 1, we found that IDO1, CSF1R, VTCN1, KDR, LGALS9, TGFBR1, TGFBI, IL10RB, and PVRL2 were highly expressed in PC tissues.

3.2. Increasing Expression of Immunoinhibitors Was Observed in PC Samples. The GEPIA database was used to compare the difference of expression of 9 overexpressed immunoinhibitors in transcription level between cancers and normal tissues (Figure 1). The data demonstrated TGFBI (Figure 2(a)), PVRL2 (Figure 2(b)), CSF1R (Figure 2(c)), TGFBR1 (Figure 2(d)), VTCN1 (Figure 2(e)), LGALS9 (Figure 2(f)), IL10RB (Figure 2(g)), KDR (Figure 2(h)), and IDO1 (Figure 2(i)) mRNA levels were significantly upregulated in patients with PC compared to normal tissues.

3.3. Immunoinhibitors Were Positively Correlated to the Advanced Stage and Grade in PC. Furthermore, the TISIDB database analysis showed TGFBI was positively correlated to the advanced grades of PC samples (Figure 3). However, we did not observe a significant upregulation of PVRL2 (Figure 3(b)), CSF1R (Figure 3(c)), TGFBR1 (Figure 3(d)), VTCN1 (Figure 3(e)), LGALS9 (Figure 3(f)), IL10RB (Figure 3(g)), KDR (Figure 3(h)), and IDO1 (Figure 3(i)) in advanced grades of PC samples.

Interestingly, our data also revealed the correlation between Immunoinhibitors level and stages of PC samples. The results displayed that expression of VTCN1 was raised in grade 2 and grade 3 samples compared to grade 1 PC samples, but expression of VTCN1 was decreased in grade 4 sample after being normalized to that in grade 1/2 PC samples (Figure 4(b)). Meanwhile, our data showed IL10RB was enhanced in grade 2, grade 3, and grade 4 PC samples compared to grade 1 PC samples (Figure 4(g)). However, no obvious difference between the expression of TGFBI (Figure 4(a)), PVRL2 (Figure 4(b)), CSF1R (Figure 4(c)), TGFBR1 (Figure 4(d)), LGALS9 (Figure 4(f)), KDR

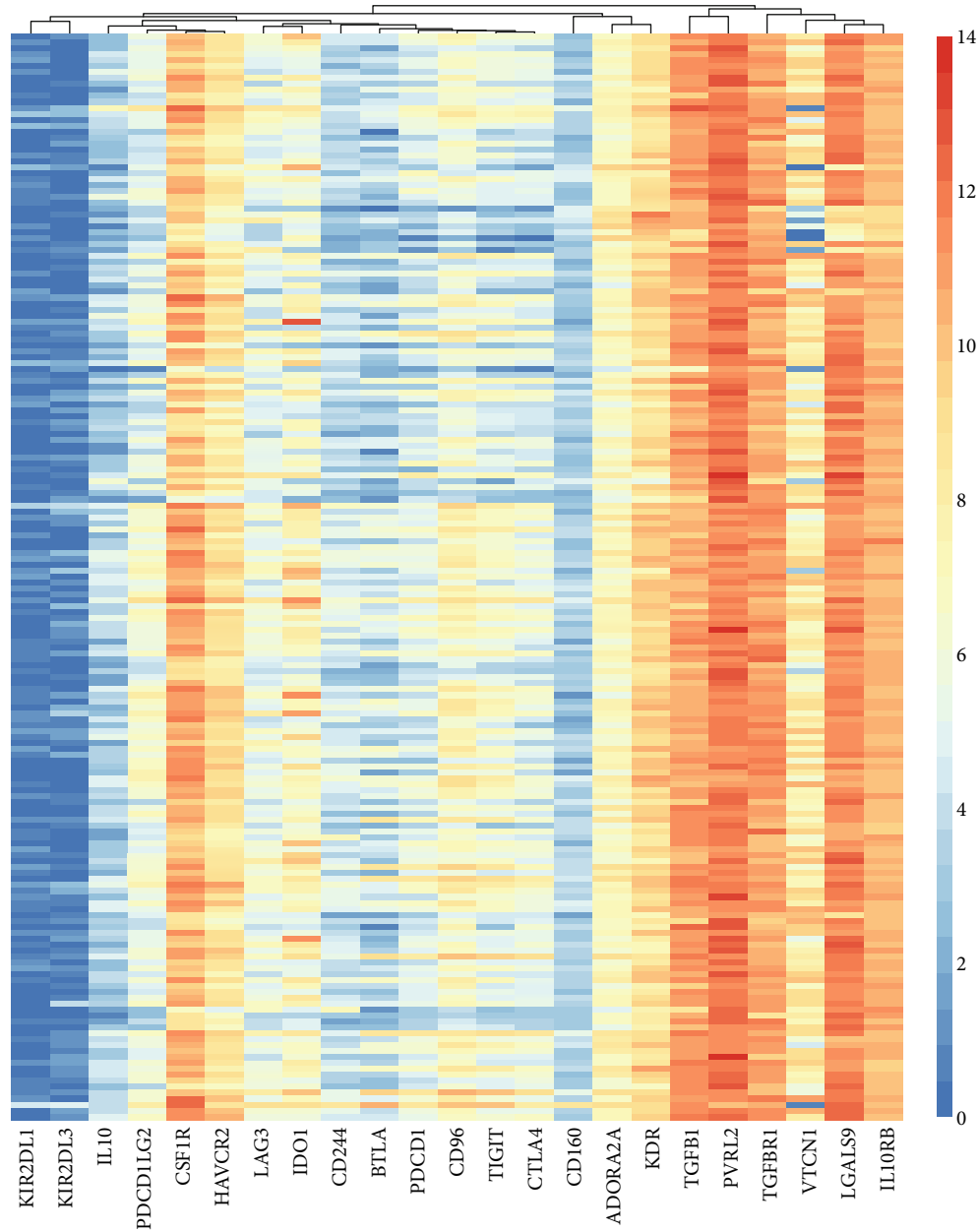


FIGURE 1: Identification of immunoinhibitor expression pattern in PC. The present study analyzed the expression pattern of 23 immunoinhibitors in PC using TCGA database, including CD160, CD244, KIR2DL1, KIR2DL3, BTLA, CSF1R, HAVCR2, TIGIT, LAG3, PDCD1, VTCN1, PDCD1LG2, LGALS9, CD96, TGFBRI, TGFB1, CTLA4, ADORA2A, PVRL2, IL10, IDO1, IL10RB, and KDR.

(Figure 4(h)), and IDO1 (Figure 4(i)) and the stage in the PC patients was taken on.

3.4. Analysis of Immunoinhibitor Feature in Prognostic PC Patients. We deeply explored the profile of immunoinhibitors implicated in prognostic PC patients. Our data revealed that the increasing level of TGFBRI (Figure 5(d)), VTCN1 (Figure 5(e)), LGALS9 (Figure 5(f)), and IDO1 (Figure 5(i)) mRNA was closely pertained to poor OS. However, the dysregulation of TGFB1 (Figure 5(a)), PVRL2 (Figure 5(b)), CSF1R (Figure 5(c)), IL10RB (Figure 5(g)), and KDR (Figure 5(h)) was not related with OS in PC.

3.5. To Assess the Coexpression and Interaction Gene with Immunoinhibitors in PC Patients. We evaluated the association of candidate gene expression with immunoinhibitors by Pearson’s correlation analysis. The immunoinhibitor-target pair with absolute Pearson’s correlation coefficient value > 0.5 was considered significant. The networks were constructed using Cytoscape software. As presented in Figure 6, the coexpression network included 9 immunoinhibitors, 1250 targets, and 1304 edges. From the analysis, we observed LGALS9 may have a primary role in this network and possessed approximately 30% coexpressing targets with IL10RB, PVRL2, and IDO1.

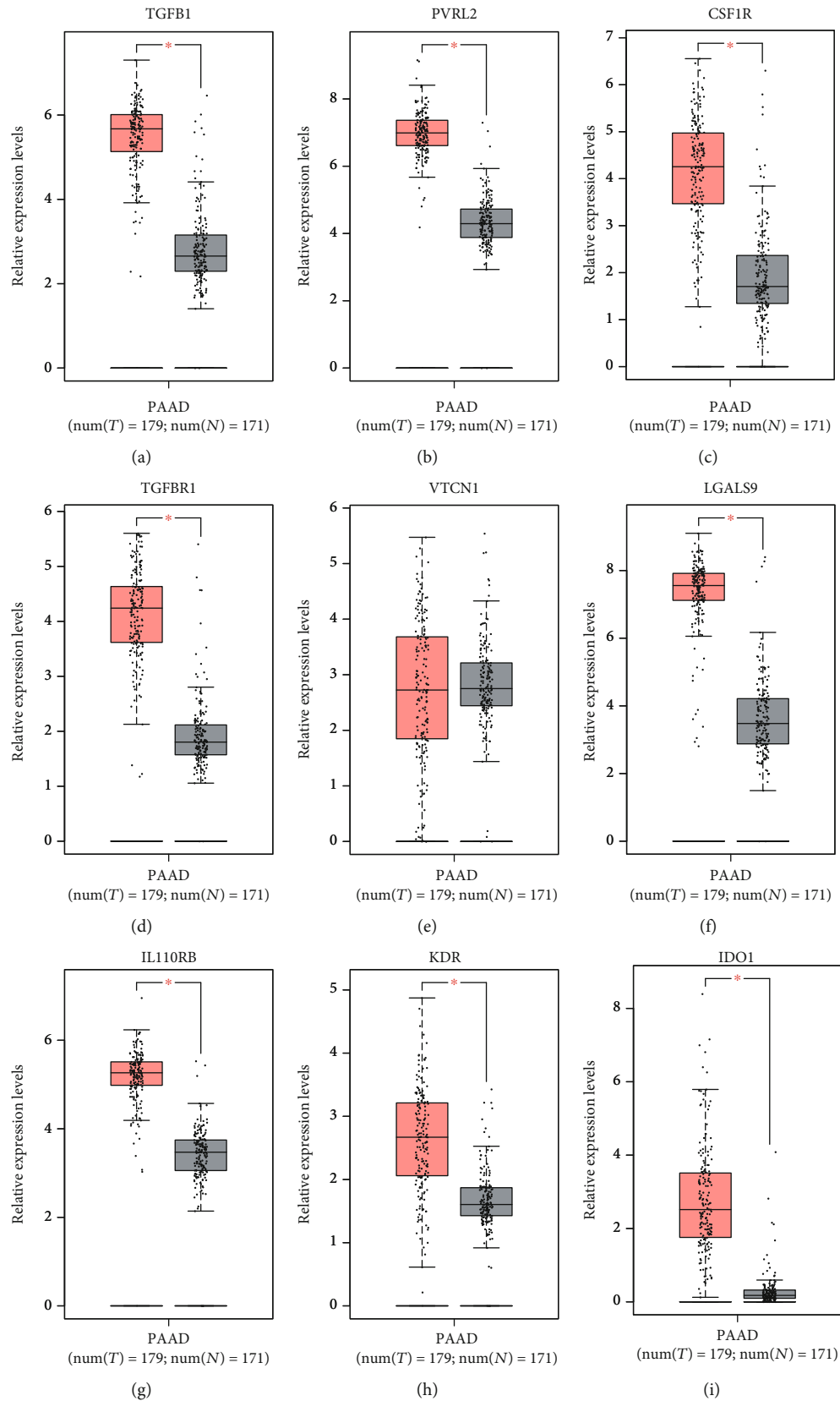


FIGURE 2: Increasing expression of immunoinhibitors was observed in PC samples. TGFB1 (a), PVRL2 (b), CSF1R (c), TGFBR1 (d), VTCN1 (e), LGALS9 (f), IL10RB (g), KDR (h), and IDO1 (i) mRNA levels were significantly upregulated in patients with PC compared to normal tissues.

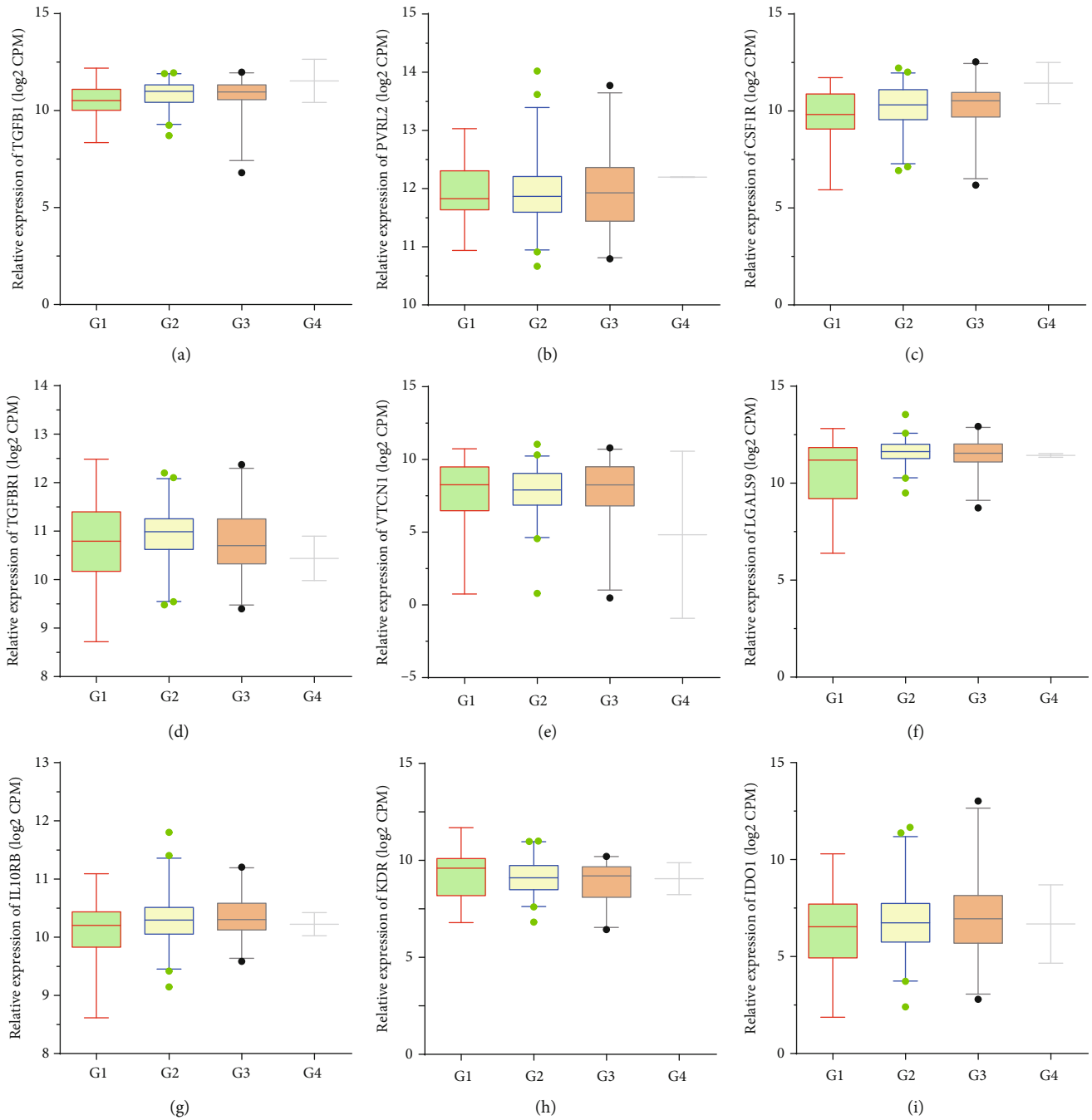


FIGURE 3: Immunoinhibitors were positively correlated to the advanced grade in PC. TISIDB database analysis revealed the expression levels of TGFB1 (a), PVRL2 (b), CSF1R (c), TGFBR1 (d), VTCN1 (e), LGALS9 (f), IL10RB (g), KDR (h), and IDO1 (i) in grade 1, 2, 3, and 4 PC samples.

3.6. Assessment of the Function of LGALS9 and TGFB1 in PC Patients. We finally validated the role of LGALS9 and TGFB1 after the analysis of GO and KEGG in the DAVID system using their genes. After bioinformatics analyzing, LGALS9 was involved in regulating type I interferon signaling pathway, defense response to virus, interferon-gamma-mediated signaling pathway, response to virus, innate immune response, and negative regulation of viral genome replication (Figure 7(a)). KEGG pathway analysis demonstrated that

LGALS9 was related to the RIG-I-like receptor signaling pathway, NF-kappa B signaling pathway, cytosolic DNA-sensing pathway, and TNF signaling pathway (Figure 7(b)).

Also, we found that TGFB1 was related to mesoderm formation, cell matrix adhesion, substrate adhesion-dependent cell spreading, in utero embryonic development, extracellular matrix organization, outflow tract septum morphogenesis, mitral valve morphogenesis, and Hippo signaling (Figure 7(c)). And KEGG pathway analysis showed TGFB1

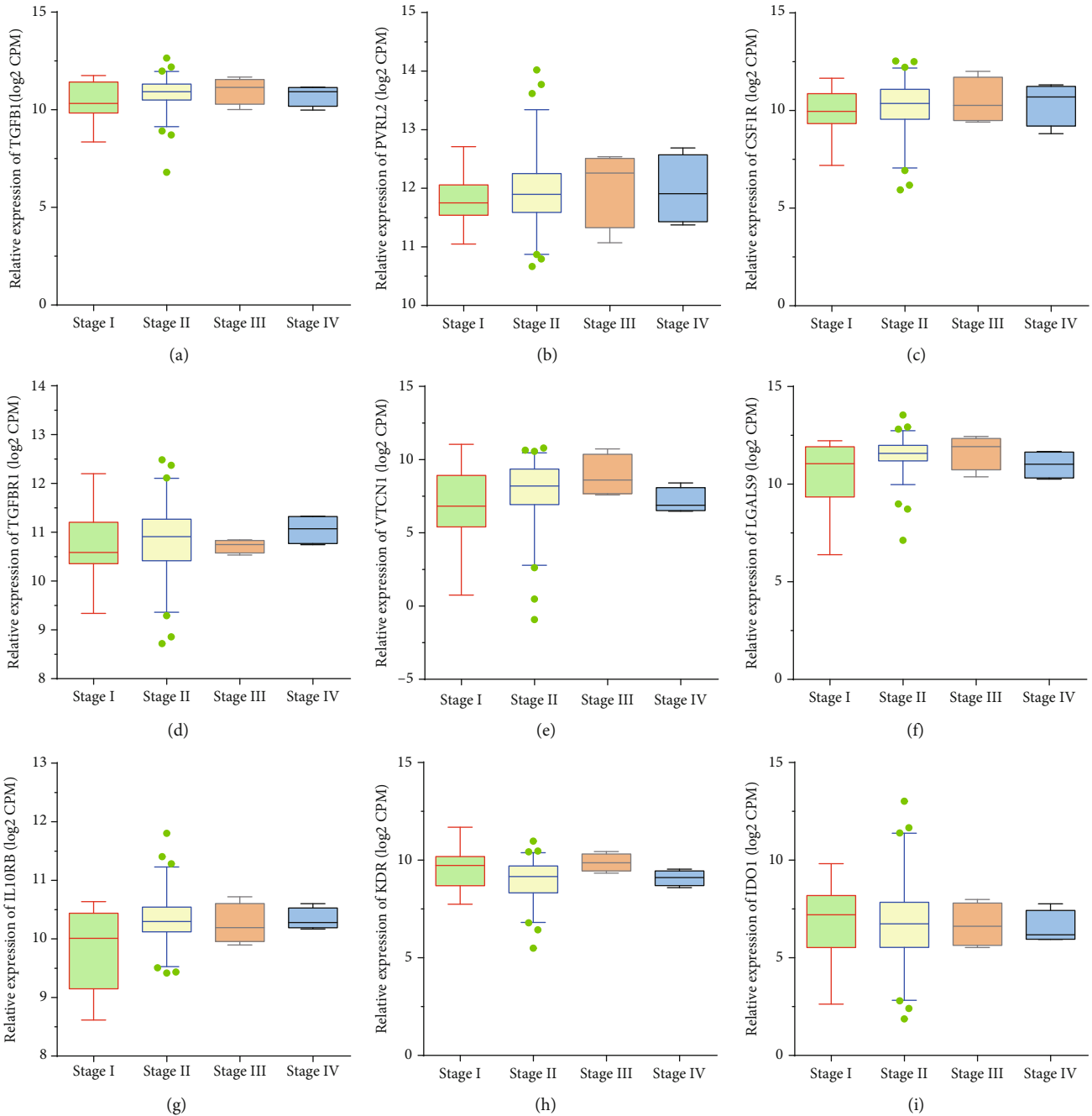


FIGURE 4: Immunoinhibitors were positively correlated to the advanced stages in PC. TISIDB database analysis revealed the expression levels of TGFB1 (a), PVRL2 (b), CSF1R (c), TGFBR1 (d), VTCN1 (e), LGALS9 (f), IL10RB (g), KDR (h), and IDO1 (i) in stage I, II, III, and IV PC samples.

was related to pathways in cancer, TGF-beta signaling pathway, focal adhesion, signaling pathways regulating pluripotency of stem cells, regulation of actin cytoskeleton, Hippo signaling pathway, and shigellosis (Figure 7(d)).

4. Discussion

PC with a poor prognosis was regarded as one of the most deadly carcinomas [33]. The growth and development of Cancer were reported to be involved in the process of immune suppression [34]. Cancer cells could stimulate various

immune checkpoint pathways responsible for curbing immunity [35]. Monoclonal antibodies targeting immune checkpoints exhibited huge advance in cancer therapy. Currently, some researches have revealed that patients with different cancer recovered better after treatment of immunoinhibitors. Developing prospective methods based on immunoinhibitors could be of significance to explore novel biomarkers in the PC diagnosis and prognosis.

In this study, we analyzed the expression pattern of 23 immunoinhibitors in PC using TCGA database and found that IDO1, CSF1R, VTCN1, KDR, LGALS9, TGFBR1, TGFB1,

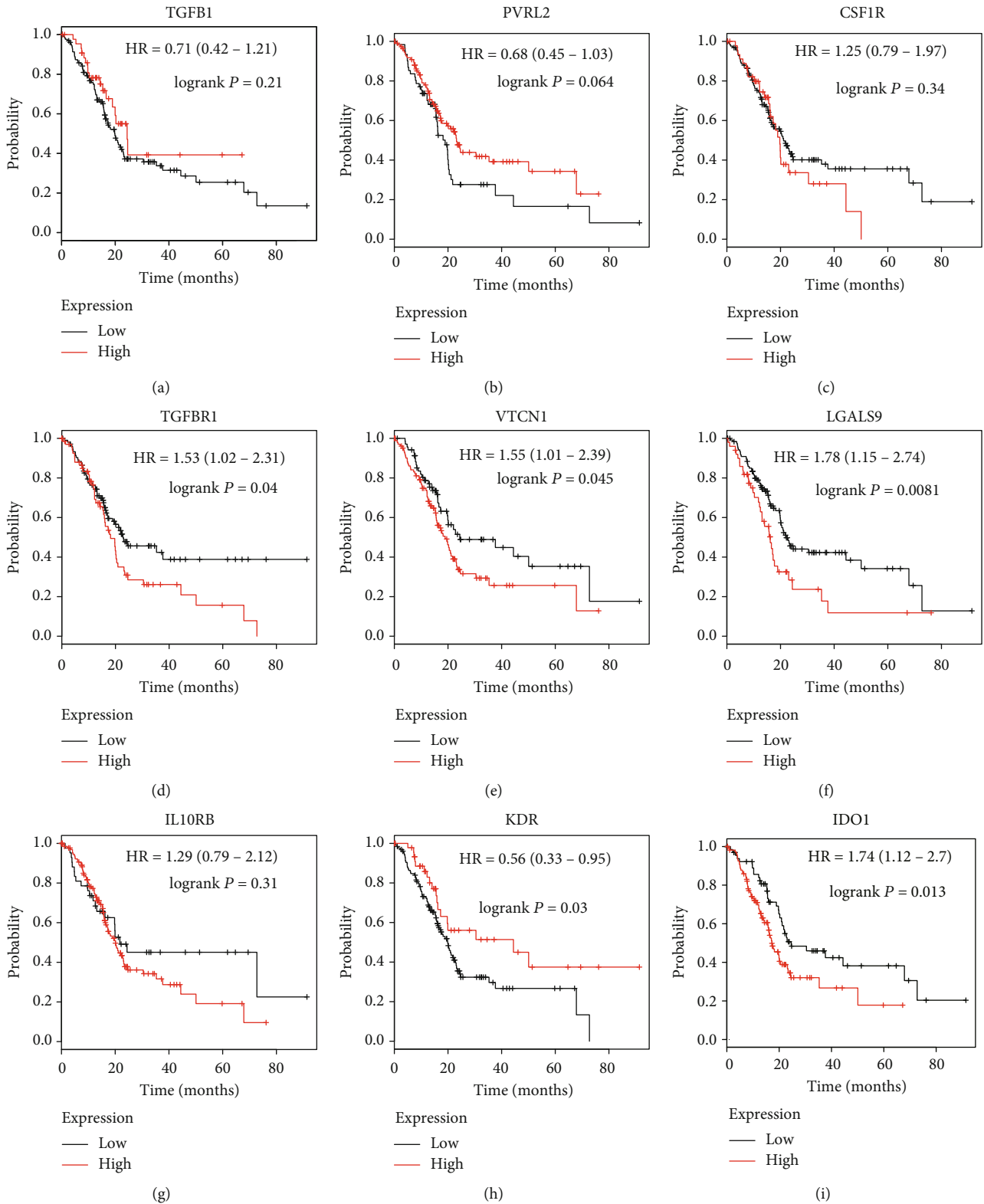


FIGURE 5: Analysis of the correlation between immunoinhibitor expression and survival time in PC patients. Kaplan-Meier plotter database analysis revealed the correlation between the levels of TGFB1 (a), PVRL2 (b), CSF1R (c), TGFBRI (d), VTCN1 (e), LGALS9 (f), IL10RB (g), KDR (h), and IDO1 (i) and overall survival time in PC patients.

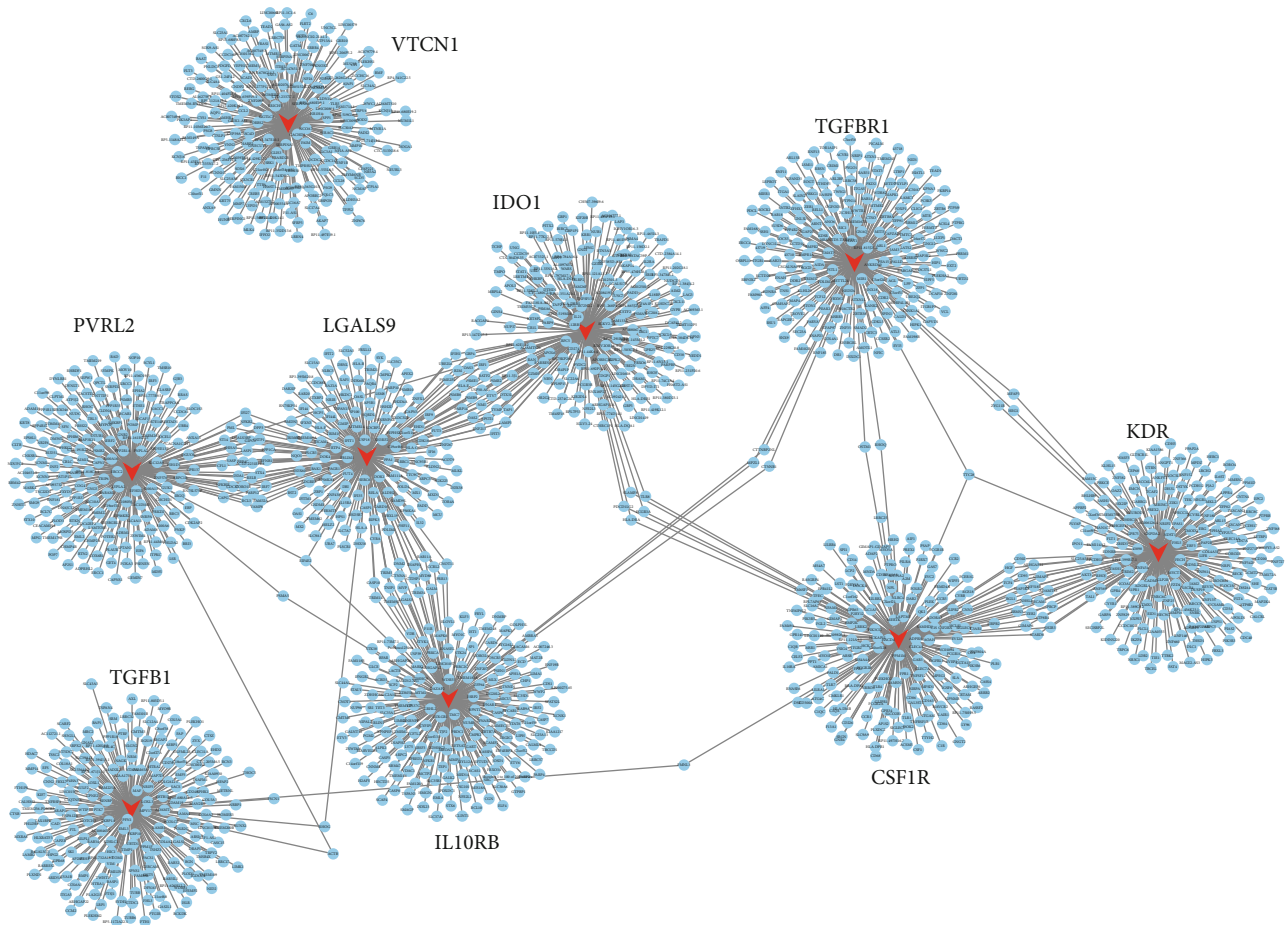


FIGURE 6: Construction of the coexpression network of immunoinhibitors in PC patients.

IL10RB, and PVRL2 were highly expressed in PC tissues. Moreover, the analysis revealed that IDO1, CSF1R, VTCN1, KDR, LGALS9, TGFBR1, TGFBI, IL10RB, and PVRL2 mRNA level was significantly upregulated in patients with PC compared to normal tissues. Kaplan-Meier plotter results demonstrated that the increasing level of TGFBR1, VTCN1, and LGALS9 mRNA was closely pertained to poor OS.

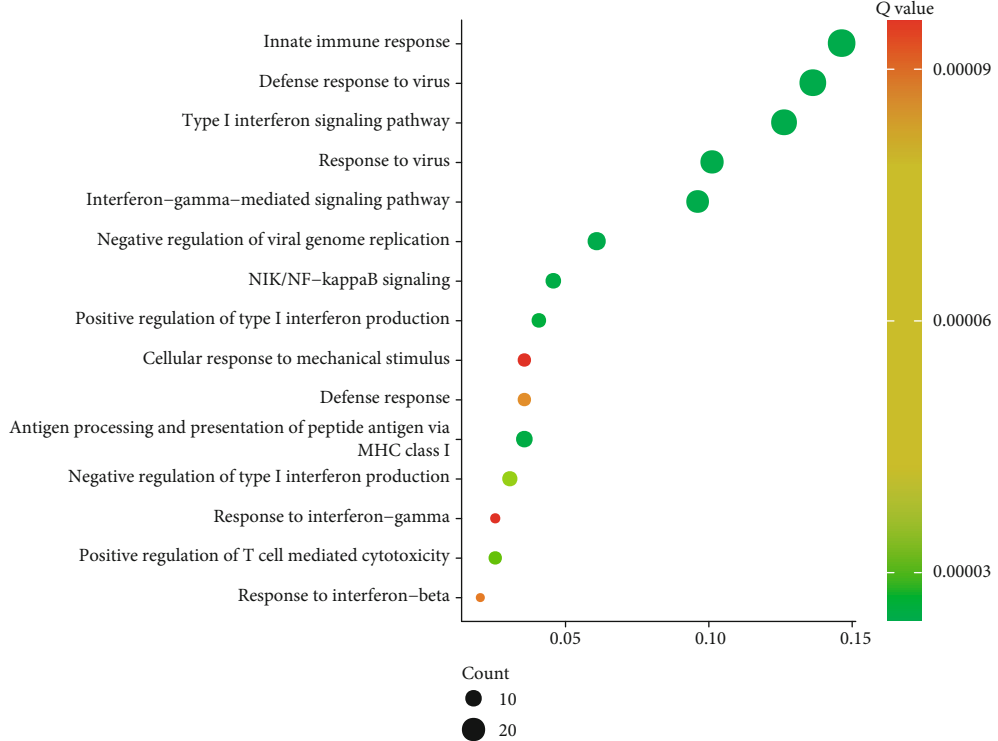
TGF- β was a primary executor of the stability and tolerance of the internal environment of immune system, including controlling many component functions [36]. Thus, disrupting TGF- β signal could result in inflammatory diseases and tumorigenesis. In addition, TGF- β is also a preliminary immunosuppressor in the tumor microenvironment [37]. Current researches have reported TGF- β was engaged in tumor immune evasion and adverse reactions to tumor immunotherapy [37]. Nevertheless, our study confirmed that TGFBR1 and TGFBI are upregulated in PC samples. Kaplan-Meier analysis showed that TGFBR1 was associated with reduced OS and PFS time in PC patients.

VTCN1 exists on the surface of antigen-presenting cells (APC) and interacts with ligands that bind to T-cell surface receptors [38]. The activity of B7-H4 was illustrated to be related to the reduced inflammatory CD4⁺ T cell response as previously described. Studies have indicated VTCN1 expression was positively linked to tumor progression and

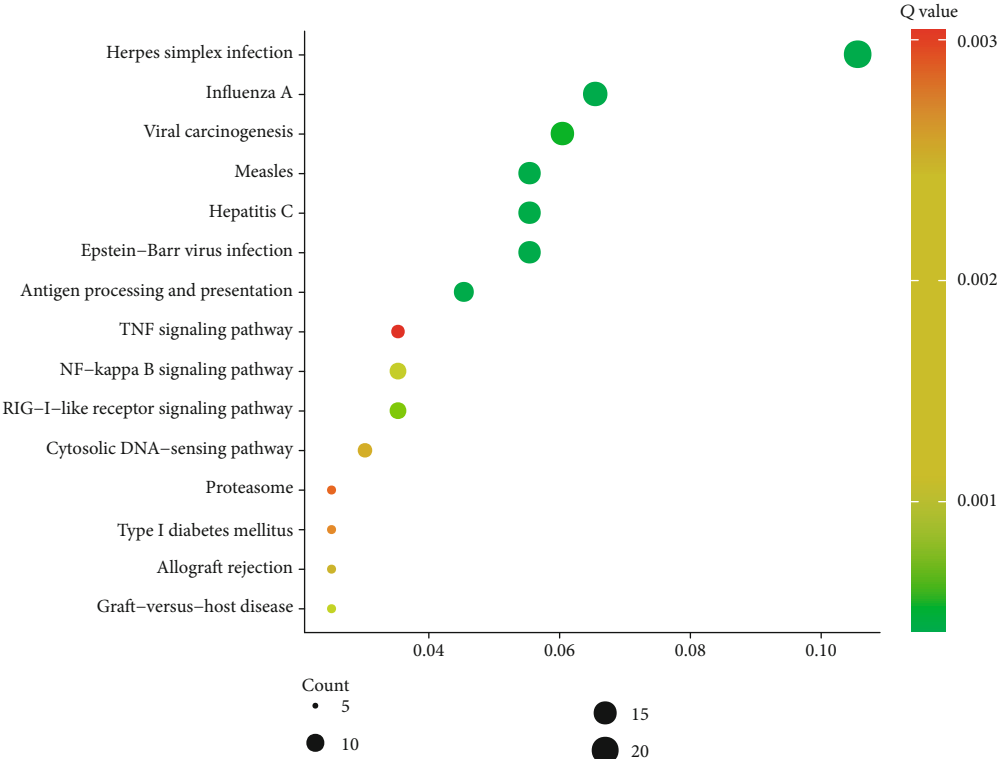
acted as a candidate for the treatment of cancer [39]. The level of B7-H4 on tumor cells with adverse clinical and pathologic features endowed B7-H4 with clinical significance [40]. Moreover, the expression of B7-H4 in tumor-associated macrophages was correlated with Foxp3⁺ regulatory T cells (Tregs) [41]. Because the expression of B7-H4 was on a variety of tumor cells and tumor-related macrophages, blocking of B7-H4 could improve the tumor microenvironment, thus enabling antigen-specific clearance of tumor cells [41]. Our study suggested the enhanced level of VTCN1 was related to OS time and PFS (progression-free survival) time of PC. Nevertheless, no increasing level of VTCN1 was shown in neither PC nor normal tissues.

Transmembrane receptor TIM-3 was encoded by *HAVCR2* and expressed on a variety of cells [42]. The expression of TIM-3 is closely related to exhaustion and impaired function of T cells. The interaction between TIM-3 and galectin 9 has been demonstrated to induce Th1 cell apoptosis, leading to reduced responses from autoimmunity and antitumor immunity [22] and also making TIM-3 as a potential target for ICB. Of note, our study firstly exposed the upregulated level of LGALS9 in PC patients was associated with shorter OS and PFS time.

IDO1 was responsible for converting tryptophan (Trp) into downstream catabolic product, called caninuria.



(a)



(b)

FIGURE 7: Continued.

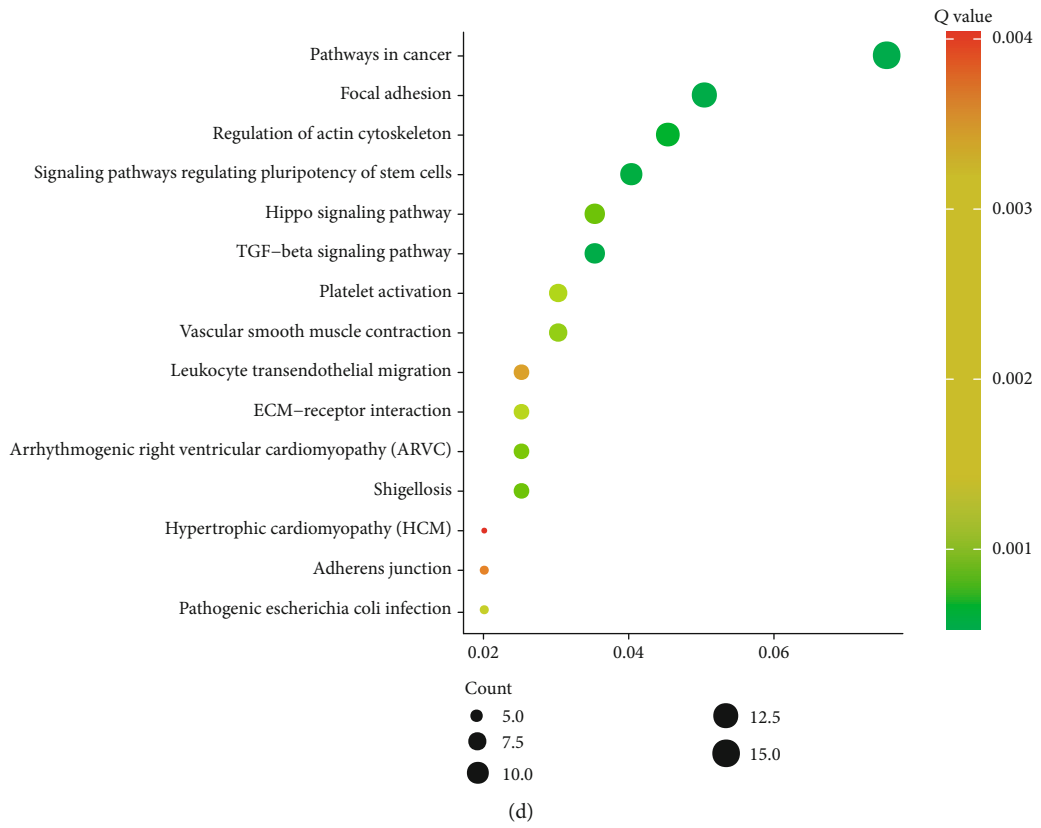
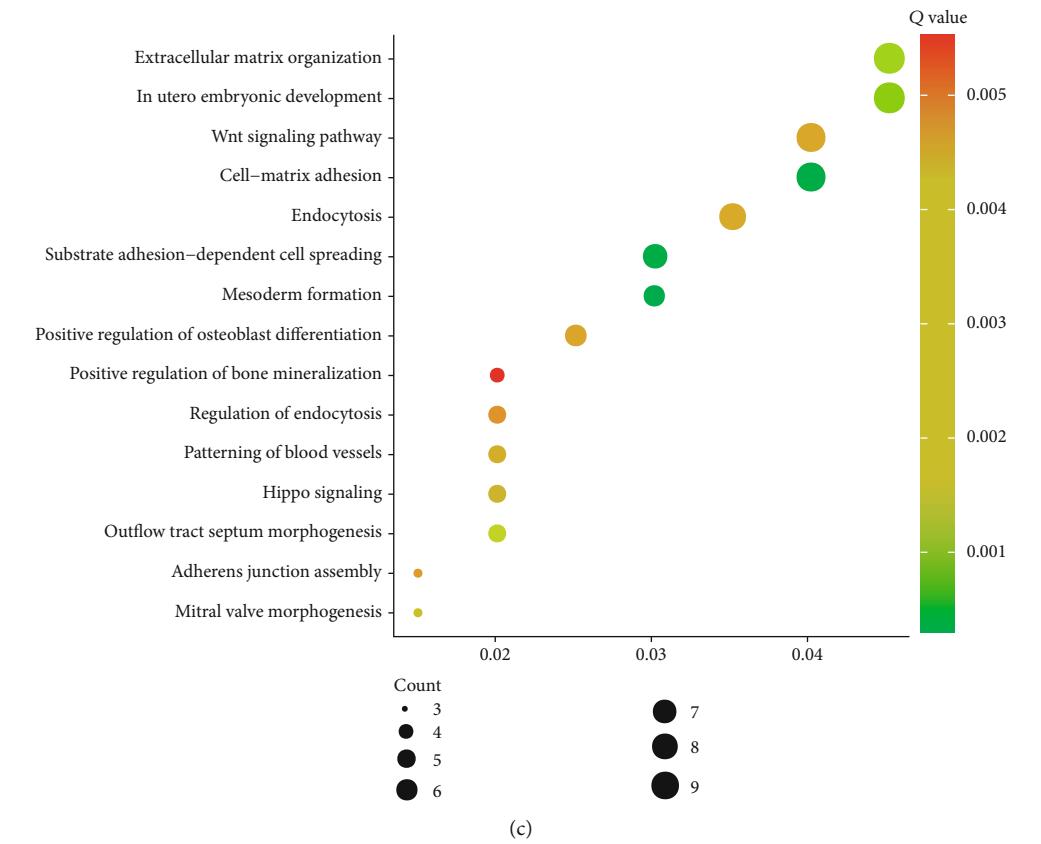


FIGURE 7: Assessment of the function of LGALS9 and TGFBI in PC patients. GO (a) analysis and KEGG pathway (b) analysis of LGALS9 in PC patients. GO analysis (c) and KEGG pathway analysis (d) of LGALS9 in PC patients.

Emerging studies have shown that IDO1 was expressed in a large number of human cancers. In the transcription level, IDO1 displayed powerful relevance with T cell infiltration [43].

Even though the expression of CSFR1, KDR, IL10RB, and PVRL2 was upregulated in PC samples, which was not associated with the prognosis of PC. CSFR1 was mostly found in aggressive cell models and participated in the invasion and migration of tumor cells, and its expression is related to the poor prognosis of cancer patients. Positive feedback existed between the expressions of CSF1 and EGF in tumors [44]. By blocking the signal transduction mediated by the EGF receptor or CSF-1 receptor, incomplete feedback loop would inhibit the migration and invasion of macrophages and tumor cells. Activated VEGF-VEGFR2 could accumulate Treg cells and control the migration of T lymphocytes [45]. The IL-10R signal on effector T cells and Treg cells is essential to keep immune tolerance [46]. Emerging studies have identified PVR2 as a new immune checkpoint [47].

Of note, this study revealed that LGALS9 and TGFBR1 were upregulated in PC compared to normal tissues. Moreover, we showed LGALS9 and TGFBR1 were significantly associated with the prognosis in PC. Despite the fact that LGALS9 and TGFBR1 were not significantly correlated to the grades, we indeed observed LGALS9 had an upregulated trend and TGFBR1 had a downregulated trend. We thought the limited sample size may contribute to this result. Also, the coexpression plus bioinformatics analysis revealed that immunoinhibitors were involved in regulating multiple inflammatory and immune response-related pathways as previously described. Very interestingly, we found LGALS9 was involved in regulating type I interferon signaling pathway, interferon-gamma-mediated signaling pathway, RIG-I-like receptor signaling pathway, NF-kappa B signaling pathway, cytosolic DNA-sensing pathway, and TNF signaling pathway. We also found that TGFBR1 was related to mesoderm formation, cell matrix adhesion, TGF-beta signaling pathway, and Hippo signaling pathway. These pathways had been demonstrated as key regulators of tumorigenesis and immune therapy.

Several limitations should also be noted in this study. First, we showed LGALS9 and TGFBR1 had a crucial role in PC with a series of bioinformatics analysis. The further experimental validations of their functions in PC could strengthen our conclusion. Second, more clinical samples should be collected to detect the expression of these immunoinhibitors in PC, which could provide more evidences to confirm their prognostic value.

5. Conclusion

Conclusively, our data suggested that immunoinhibitor mRNA level was dramatically upregulated, but negatively correlated with OS for PC. All the data suggested these genes could be used as an emerging prognostic indicator and targets in PC patients. Our findings would give a hint to have a better understanding of the mechanism implicated in PC and stretched out more precise immunotherapeutic treatments for PC prognosis. Nevertheless, more researches and

efforts should be contributed to our findings, followed by providing a much more promising clinical strategy for an early diagnosis and prognostic marker in PC therapy.

Data Availability

The datasets used during the present study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

Yue Fan is the first author.

Acknowledgments

This work has been supported by the Development Project of Shanghai Peak Disciplines-Integrative Medicine (20180101).

References

- [1] X. I. Zhao, D. C. Li, X. G. Zhu et al., "B7-H3 overexpression in pancreatic cancer promotes tumor progression," *International Journal of Molecular Medicine*, vol. 31, no. 2, pp. 283–291, 2013.
- [2] A. Lambert, C. Gavaille, and T. Conroy, "Current status on the place of FOLFIRINOX in metastatic pancreatic cancer and future directions," *Therapeutic Advances in Gastroenterology*, vol. 10, no. 8, pp. 631–645, 2017.
- [3] B. Wu, K. Wang, J. Fei et al., "Novel three-lncRNA signature predicts survival in patients with pancreatic cancer," *Oncology Reports*, vol. 40, no. 6, pp. 3427–3437, 2018.
- [4] T. E. Newhook, E. M. Blais, J. M. Lindberg et al., "A thirteen-gene expression signature predicts survival of patients with pancreatic cancer and identifies new genes of interest," *PLoS One*, vol. 9, no. 9, article e105631, 2014.
- [5] X. Du, X. Zheng, Z. Zhang et al., "A label-free electrochemical immunosensor for detection of the tumor marker CA242 based on reduced graphene oxide-gold-palladium nanocomposite," *Nanomaterials*, vol. 9, no. 9, p. 1335, 2019.
- [6] W. Yan, L. Xu, Q. Wu et al., "A case report of spontaneous rupture of intracranial epidermoid cyst with dramatic increase of serum carbohydrate antigen 199: a three-year follow-up study," *BMC Neurology*, vol. 15, no. 1, p. 198, 2015.
- [7] C. S. Hinrichs, A. Kaiser, C. M. Paulos et al., "Type 17 CD8+ T cells display enhanced antitumor immunity," *Blood*, vol. 114, no. 3, pp. 596–599, 2009.
- [8] B. Wang, C. Sun, S. Wang et al., "Image-guided dendritic cell-based vaccine immunotherapy in murine carcinoma models," *American Journal of Translational Research*, vol. 9, no. 10, pp. 4564–4573, 2017.
- [9] H. L. Kaufman, T. Amatruda, T. Reid et al., "Systemic versus local responses in melanoma patients treated with talimogene laherparepvec from a multi-institutional phase II study," *Journal for Immunotherapy of Cancer*, vol. 4, no. 1, p. 12, 2016.
- [10] Y. Imai, K. Hasegawa, H. Matsushita et al., "Expression of multiple immune checkpoint molecules on T cells in malignant

- ascites from epithelial ovarian carcinoma," *Oncology Letters*, vol. 15, no. 5, pp. 6457–6468, 2018.
- [11] H. C. Pühr and A. Ilhan-Mutlu, "New emerging targets in cancer immunotherapy: the role of LAG3," *ESMO Open*, vol. 4, no. 2, article e000482, 2019.
 - [12] L. P. Andrews, A. E. Marciscano, C. G. Drake, and D. A. Vignali, "LAG3 (CD223) as a cancer immunotherapy target," *Immunological Reviews*, vol. 276, no. 1, pp. 80–96, 2017.
 - [13] X. Du, F. Tang, M. Liu et al., "A reappraisal of CTLA-4 checkpoint blockade in cancer immunotherapy," *Cell Research*, vol. 28, no. 4, pp. 416–432, 2018.
 - [14] E. D. Kwon, B. A. Foster, A. A. Hurwitz et al., "Elimination of residual metastatic prostate cancer after surgery and adjunctive cytotoxic T lymphocyte-associated antigen 4 (CTLA-4) blockade immunotherapy," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 26, pp. 15074–15079, 1999.
 - [15] D. E. Dolan and S. Gupta, "PD-1 pathway inhibitors: changing the landscape of cancer immunotherapy," *Cancer Control*, vol. 21, no. 3, pp. 231–237, 2014.
 - [16] L. Zitvogel and G. Kroemer, "Targeting PD-1/PD-L1 interactions for cancer immunotherapy," *Oncoimmunology*, vol. 1, pp. 1223–1225, 2014.
 - [17] A. Friedlaender, A. Addeo, and G. Banna, "New emerging targets in cancer immunotherapy: the role of TIM3," *ESMO Open*, vol. 4, article e000497, Suppl 3, 2019.
 - [18] I. Herrera-Camacho, M. Anaya-Ruiz, M. Perez-Santos, L. Millan-Perez Pena, C. Bandala, and G. Landeta, "Cancer immunotherapy using anti-TIM3/PD-1 bispecific antibody: a patent evaluation of EP3356411A1," *Expert Opinion on Therapeutic Patents*, vol. 29, no. 8, pp. 587–593, 2019.
 - [19] A. H. Sharpe, "Introduction to checkpoint inhibitors and cancer immunotherapy," *Immunological Reviews*, vol. 276, no. 1, pp. 5–8, 2017.
 - [20] S. Buisson and F. Triebel, "MHC class II engagement by its ligand LAG-3 (CD223) leads to a distinct pattern of chemokine and chemokine receptor expression by human dendritic cells," *Vaccine*, vol. 21, no. 9-10, pp. 862–868, 2003.
 - [21] C. Kyi and M. A. Postow, "Immune checkpoint inhibitor combinations in solid tumors: opportunities and challenges," *Immunotherapy*, vol. 8, no. 7, pp. 821–837, 2016.
 - [22] M. Nishino, N. H. Ramaiya, H. Hatabu, and F. S. Hodi, "Monitoring immune-checkpoint blockade: response evaluation and biomarker development," *Nature Reviews Clinical Oncology*, vol. 14, no. 11, pp. 655–668, 2017.
 - [23] I. Fabregat, J. Fernando, J. Mainez, and P. Sancho, "TGF-beta signaling in cancer treatment," *Current Pharmaceutical Design*, vol. 20, no. 17, pp. 2934–2947, 2014.
 - [24] S. Sanjabi, S. A. Oh, and M. O. Li, "Regulation of the immune response by TGF- β : from conception to autoimmunity and infection," *Cold Spring Harbor perspectives in biology*, vol. 9, no. 6, 2017.
 - [25] Y. Pan, F. Lu, Q. Fei et al., "Single-cell RNA sequencing reveals compartmental remodeling of tumor-infiltrating immune cells induced by anti-CD47 targeting in pancreatic cancer," *Journal of Hematology & Oncology*, vol. 12, no. 1, p. 124, 2019.
 - [26] L. Zhang, S. Yu, C. Wang, C. Jia, Z. Lu, and J. Chen, "Establishment of a non-coding RNAomics screening platform for the regulation of KRAS in pancreatic cancer by RNA sequencing," *International Journal of Oncology*, vol. 53, no. 6, pp. 2659–2670, 2018.
 - [27] Y. Mao, J. Shen, Y. Lu et al., "RNA sequencing analyses reveal novel differentially expressed genes and pathways in pancreatic cancer," *Oncotarget*, vol. 8, no. 26, pp. 42537–42547, 2017.
 - [28] A. Nagy, A. Lanczky, O. Menyhart, and B. Györfy, "Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets," *Scientific Reports*, vol. 8, no. 1, p. 9227, 2018.
 - [29] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, pp. D561–D568, 2010.
 - [30] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic Acids Research*, vol. 45, no. W1, pp. W98–W102, 2017.
 - [31] P. Sidaway, "Pancreatic cancer: TCGA data reveal a highly heterogeneous disease," *Nature Reviews Clinical Oncology*, vol. 14, p. 648, 2017.
 - [32] J. Gao, B. A. Aksoy, U. Dogrusoz et al., "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Science signaling*, vol. 6, p. pl1, 2013.
 - [33] B. Y. Li, L. J. He, X. L. Zhang, H. Liu, and B. Liu, "High expression of RAB38 promotes malignant progression of pancreatic cancer," *Molecular Medicine Reports*, vol. 19, no. 2, pp. 909–918, 2019.
 - [34] R. J. McKallip, M. Nagarkatti, and P. S. Nagarkatti, "Delta-9-tetrahydrocannabinol enhances breast cancer growth and metastasis by suppression of the antitumor immune response," *Journal of Immunology*, vol. 174, no. 6, pp. 3281–3289, 2005.
 - [35] V. Chandramohan, X. Bao, X. Yu et al., "Improved efficacy against malignant brain tumors with EGFRwt/EGFRvIII targeting immunotoxin and checkpoint inhibitor combinations," *Journal for Immunotherapy of Cancer*, vol. 7, no. 1, p. 142, 2019.
 - [36] S. H. Wrzesinski, Y. Y. Wan, and R. A. Flavell, "Transforming growth factor- and the immune response: implications for anticancer therapy," *Clinical Cancer Research*, vol. 13, no. 18, pp. 5262–5270, 2007.
 - [37] M. Pickup, S. Novitskiy, and H. L. Moses, "The roles of TGF β in the tumour microenvironment," *Nature Reviews Cancer*, vol. 13, no. 11, pp. 788–799, 2013.
 - [38] J. R. Podojil and S. D. Miller, "Potential targeting of B7-H4 for the treatment of cancer," *Immunological Reviews*, vol. 276, no. 1, pp. 40–51, 2017.
 - [39] C. Chen, Q. X. Qu, Y. Shen et al., "Induced expression of B7-H4 on the surface of lung cancer cell by the tumor-associated macrophages: a potential mechanism of immune escape," *Cancer Letters*, vol. 317, no. 1, pp. 99–105, 2012.
 - [40] N. Xie, J. B. Cai, L. Zhang et al., "Upregulation of B7-H4 promotes tumor progression of intrahepatic cholangiocarcinoma," *Cell Death & Disease*, vol. 8, no. 12, p. 3205, 2017.
 - [41] I. Kryczek, S. Wei, G. Zhu et al., "Relationship between B7-H4, regulatory T cells, and patient outcome in human ovarian carcinoma," *Cancer Research*, vol. 67, no. 18, pp. 8900–8905, 2007.
 - [42] T. A. W. Holderried, L. de Vos, E. G. Bawden et al., "Molecular and immune correlates of TIM-3 (HAVCR2) and galectin 9 (LGALS9) mRNA expression and DNA methylation in melanoma," *Clinical Epigenetics*, vol. 11, no. 1, p. 161, 2019.
 - [43] L. Zhai, E. Ladomersky, K. L. Lauing et al., "Infiltrating T cells increase IDO1 expression in glioblastoma and contribute to

- decreased patient survival,” *Clinical Cancer Research*, vol. 23, no. 21, pp. 6650–6660, 2017.
- [44] M. Yang, D. McKay, J. W. Pollard, and C. E. Lewis, “Diverse functions of macrophages in different tumor microenvironments,” *Cancer Research*, vol. 78, no. 19, pp. 5492–5503, 2018.
- [45] Y. Tada, Y. Togashi, D. Kotani et al., “Targeting VEGFR2 with Ramucirumab strongly impacts effector/ activated regulatory T cells and CD8(+) T cells in the tumor microenvironment,” *Journal for Immunotherapy of Cancer*, vol. 6, no. 1, p. 106, 2018.
- [46] G. Cheng, A. Yu, and T. R. Malek, “T-cell tolerance and the multi-functional role of IL-2R signaling in T-regulatory cells,” *Immunological Reviews*, vol. 241, no. 1, pp. 63–76, 2011.
- [47] H. Stamm, F. Klingler, E. M. Grossjohann et al., “Immune checkpoints PVR and PVRL2 are prognostic markers in AML and their blockade represents a new therapeutic option,” *Oncogene*, vol. 37, no. 39, pp. 5269–5280, 2018.

Research Article

Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence

Ilker Ozsahin ^{1,2} **Boran Sekeroglu** ^{2,3} **Musa Sani Musa** ¹ **Mubarak Taiwo Mustapha**,^{1,2}
and Dilber Uzun Ozsahin ^{1,2}

¹Department of Biomedical Engineering, Near East University, Nicosia / TRNC, Mersin-10, 99138, Turkey

²DESAM Institute, Near East University, Nicosia / TRNC, Mersin-10, 99138, Turkey

³Department of Artificial Intelligence Engineering, Near East University, Nicosia / TRNC, Mersin-10, 99138, Turkey

Correspondence should be addressed to Ilker Ozsahin; ilker.ozsahin@neu.edu.tr

Received 26 June 2020; Revised 28 August 2020; Accepted 16 September 2020; Published 26 September 2020

Academic Editor: Lin Lu

Copyright © 2020 Ilker Ozsahin 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 COVID-19 diagnostic approach is mainly divided into two broad categories, a laboratory-based and chest radiography approach. The last few months have witnessed a rapid increase in the number of studies use artificial intelligence (AI) techniques to diagnose COVID-19 with chest computed tomography (CT). In this study, we review the diagnosis of COVID-19 by using chest CT toward AI. We searched ArXiv, MedRxiv, and Google Scholar using the terms “deep learning”, “neural networks”, “COVID-19”, and “chest CT”. At the time of writing (August 24, 2020), there have been nearly 100 studies and 30 studies among them were selected for this review. We categorized the studies based on the classification tasks: COVID-19/normal, COVID-19/non-COVID-19, COVID-19/non-COVID-19 pneumonia, and severity. The sensitivity, specificity, precision, accuracy, area under the curve, and F1 score results were reported as high as 100%, 100%, 99.62, 99.87%, 100%, and 99.5%, respectively. However, the presented results should be carefully compared due to the different degrees of difficulty of different classification tasks.

1. Introduction

Coronaviruses have been around for many decades, and it has affected many animals/mammal species and human being. By March 11, 2020, the World Health Organization (WHO) [1] declared the new coronavirus called the COVID-19, a pandemic, and it has brought the entire globe into a compulsory lockdown. Coronavirus is a family of RNA viruses that is capable of causing significant viral pathogens in humans and animals. Corona is medium-sized viruses with the largest viral RNA genome known. Coronavirus infects both birds and mammals, but the bat is host to the largest number of the viral genotype of coronavirus. So, the bat is the host and does not get infected. It can, however, spread the virus to a human. As of 24th of August 2020, there have been more than 23 million confirmed cases of coronavirus worldwide, with about 800,000 of such cases resulting in

the death of the infected patient. This is spread around 216 countries, areas, or territories. However, around five million infected patients have recovered worldwide [2]. The USA, Brazil, India, and Russia are the top four countries with the highest number of cases. Around 90 million tests have conducted in China, followed by the USA, Russia, and India, with 72 million, 33 million, and 32 million tests, respectively [2].

Testing COVID-19 involves analyzing samples that indicate the present or past presence of severe acute respiratory syndrome-associated coronavirus 2 (SARS-CoV-2). The test is done to detect either the presence of the virus or of antibodies produced in response to infection. COVID-19 diagnostic approach is mainly divided into two broad categories, a laboratory-based approach, which includes point of care-testing, nucleic acid testing, antigens tests, and serology (antibody) tests. The other approach is using medical

imaging diagnostic tools such as X-ray and computed tomography (CT) [3].

The laboratory-based tests are performed on samples obtained via nasopharyngeal swab, throat swabs, sputum, and deep airway material [4]. The most common diagnostic approach is the nasopharyngeal swab, which involves exposing a swab to paper strips containing artificial antibodies designed to bind to coronavirus antigens. Antigens bind to the strips and give a visual readout [4]. The process is pretty fast and is employed at the point of care. The nucleic acid test has low sensitivity between 60-71% [4]. On the other hand, Fang et al. [5] showed that radiologic methods could provide higher sensitivity than that of lab tests.

The use of medical imaging tools is the second approach of COVID-19 virus detection. These tools are playing an important role in the management of patients that are confirmed or suspected to be infected with the virus. It is worthy of note that without clinical suspicion, findings from X-ray, or CT images are nonspecific as many other diseases could have a similar pattern [6].

Thoracic CT scan is the imaging modality of choice that plays a vital role in the management of COVID-19. Thoracic CT has a high sensitivity for diagnosis of COVID-19 which makes it a primary tool for COVID-19 detection [5]. CT scan involves transmitting X-rays through the patient's chest, which are then detected by radiation detectors and reconstructed into high-resolution medical images. There are certain patterns to look out for in a chest CT scans which present themselves in different characteristic manifestations. The potential findings with 100% confidence for COVID-19 in thoracic CT images are ground – glass opacity (GGO) \pm crazy – paving and consolidation, air bronchograms, reverse halo, and perilobular pattern [6].

The abovementioned findings are reports presented by a radiologist who specializes in interpreting medical images. Interpretation of these findings by expert radiologists does not have a very high sensitivity [4]. Artificial intelligence (AI) has been employed as it plays a key role in every aspect of COVID-19 crisis management. AI has proven to be useful in medical applications since its inception, and it became widely accepted due to its high prediction and accuracy rates. In the diagnosis stage of COVID-19, AI can be used to recognize patterns on medical images taken by CT. Other applications of AI include, but not limited to, virus detection, diagnosis and prediction, prevention, response, recovery, and to accelerate research [7]. AI can be used to segment regions of interest and capture fine structures in chest CT images, self-learned features can easily be extracted for diagnosis and other applications as well. A recent study showed that AI accurately detected COVID-19 and was also able to differentiate it from other lung diseases and community-acquired pneumonia [8]. In this study, we review the diagnosis of COVID-19 by using chest CT toward AI.

2. Materials and Methods

We searched ArXiv, MedRxiv, and Google Scholar for AI for COVID-19 diagnosis with chest CT. At the time of writing (August 24, 2020), there have been nearly 100 studies and

only 17 of them were peer-reviewed papers. In total, 30 studies (17 peer-reviewed and 13 non-peer-reviewed papers) were selected for this review. We noticed that very different classification terms are reported by the authors such as “normal”, “healthy”, “other”, “COVID-19”, “non-COVID-19”, “without COVID-19”, “community-acquired pneumonia (CAP)”, “other pneumonia”, “bacterial pneumonia”, “SARS”, “lung cancer”, “type A influenza (influenza-A)”, and “severity”. Therefore, we categorized the studies into four main tasks as follows: COVID-19/normal, COVID-19/non-COVID-19, COVID-19/non-COVID-19 pneumonia, and COVID-19 severity classification. COVID-19 group consists of COVID-19 patients. The normal group includes only healthy subjects. Non-COVID-19 group includes either one of the cases which is not COVID-19 or a combination of all other cases. The non-COVID-19 pneumonia group includes other types of pneumonia, which is not caused by COVID-19, such as viral or bacterial pneumonia, as well as influenza A and SARS. Lastly, COVID-19 severity classification aims at classifying the COVID-19 cases as severe or nonsevere.

Since the rapid studies on the detection of COVID-19 in CT scans continue, the researchers who take into account the peer-review period in the journals share the results they obtained in their studies with other researchers and scientists as preprints in different publication environments. Machine learning is used to make decisions on tasks that people have difficulty making decisions or problems that require more stable decisions using both numerical and image-based data. A deep convolutional neural network (CNN) is the most widely used among machine learning methods. It is one of the first preferred neural networks, especially in image-based problems, since it contains both feature extraction and classification stages and produces very effective results. In image-based COVID-19 researches, the CNN model or different models produced from CNN are widely encountered. In the researches, a generally hold-out method and a few k -fold cross-validation were used during the training phase. In the hold-out method, while training is done by dividing the data into two parts as test and train, in k -fold cross-validation, the data is divided into k -folds, and the folds are trained k -times by shifting the testing fold in each training so that each fold is used in the test phase. It is used as a better method for model evaluation.

3. Results

3.1. COVID-19/Normal Classification Studies. Alom et al. [9] implemented two deep learning models for COVID-19 detection and segmentation. Inception Recurrent Residual Neural Network (IRRCNN), which is based on transfer learning, was used for the COVID-19 detection task, and the NABLA-N model was for the segmentation task. They considered different datasets to detect COVID-19 on CT images, by using an additional chest X-ray dataset. The publicly available dataset was considered for the segmentation procedure of CT images, and the dataset that consists of 425 CT image samples, with 178 pneumonia, and 247 normal images were considered for the COVID-19 detection purpose. All images were resized to the dimensions of $192 \times$

192 pixels, and 375 of total images were used for training and validation with a data augmentation procedure. The training was performed using Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 16. The COVID-19 detection and segmentation accuracy were achieved by 98.78% and 99.56%, respectively.

Hu et al. [10] constructed an AI model on ShuffleNet V2 [11], which provides fast and accurate training in transfer learning applications. The considered CT dataset consists of 521 COVID-19 infected images, 397 healthy images, 76 bacterial pneumonia images, and 48 SARS images. The data augmentation procedure as flip, rotation, translation, brightness adjustment, and flip+brightness adjustment was applied in this study to increase the number of training images. The first experiment was performed on the classification of COVID-19 images from normal healthy images. The average sensitivity, specificity, and area under the curve (AUC) score were obtained as 90.52%, 91.58%, and 0.9689, respectively.

Gozes et al. [12] proposed a comprehensive system to detect COVID-19 from normal cases. The proposed system included lung segmentation, COVID-19 detection in CT slices, and marking case as COVID-19 using a predetermined threshold based on the counted COVID-19 positive slices. Several datasets were considered in training and testing phases, and pretrained network ResNet50 was used for the detection of COVID-19. The sensitivity, specificity, and the AUC score were achieved as 94%, 98%, and 0.9940, respectively.

In another study for differentiation of COVID-19 from normal cases, Kassani et al. [13] used several pretrained networks such as MobileNet [14], DenseNet [15], Xception [16], InceptionV3 [17], InceptionResNetV2 [18], and ResNet [19] to extract the features of images within the publicly available dataset. Then, extracted features were trained using six machine learning algorithms, namely, decision tree, random forest, XGBoost, AdaBoost, Bagging, and LightGBM. Kassani et al. [13] concluded that the Bagging classifier obtained the optimal results with a maximum of $99.00\% \pm 0.09$ accuracy on features extracted by pretrained network DenseNet121.

Jaiswal et al. [20] implemented a pretrained network DenseNet201-based deep model on classifying 2,492 CT-scans (1,262 positive for COVID-19, and the rest 1,230 are negative) as positive or negative. They compared their results with VGG16, ResNet152V2, and Inception-ResNetV2. They concluded that their model outperformed other considered models and achieved an overall accuracy of 96.25%. Table 1 summarizes the studies on COVID-19 vs. normal cases.

3.2. COVID-19/Non-COVID-19 Classification Studies. Jin et al. [30] considered 496 COVID-19 positive and 260 negative images collected in Wuhan Union Hospital, Western Campus of Wuhan Union Hospital, and Jiangnan Mobile Cabin Hospital in Wuhan. Besides, they used two publicly available international databases, LIDC-IDRI [28] and ILD-HUG [31] (1012 and 113 subjects, respectively) as negative cases to develop the system. A 2D convolutional neural network was used for the segmentation of CT slices, and then, a model was trained for positive and negative cases. Jin et al. reported that the proposed system achieved the AUC

score of 0.9791, sensitivity of 94.06%, and specificity of 95.47% for the external test cohort.

Singh et al. [32] proposed a multiobjective differential evolution- (MODE-) based convolutional neural networks to detect COVID-19 in chest CT images. It was concluded that the proposed method outperformed the CNN, ANFIS, and ANN models in all considered metrics between 1.6827% and 2.0928%.

Amyar et al. [33] developed another model architecture that included image segmentation, reconstruction, and classification tasks, which was based on the encoder and convolutional layer. The experiments were performed on three datasets that included 1044 CT images, and the obtained results showed that the proposed architecture achieved the highest results in their experiment, with 0.93% of the AUC score.

Ahuja et al. [34] used data augmentation and pretrained networks to classify COVID-19 images. Data augmentation was performed using stationary wavelets, and the random rotation, translation, and shear operations were applied to the CT scan images. ResNet18, ResNet50, ResNet101, and SqueezeNet were implemented for the classification task, and Ahuja et al. concluded that ResNet18 outperformed other models by obtaining a 0.9965 AUC score.

Liu et al. [35] proposed another deep neural network model, namely, lesion-attention deep neural networks, where the backbone of the model used the weights of pretrained networks such as VGG16, ResNet18, and ResNet50. The proposed model was capable of classifying COVID-19 images, which was the main aim of the study, with 0.94 of the AUC score using VGG16 as the backbone model. Besides this, the model was able to make a multilabel prediction on the five lesions.

Instead of deep learning approaches, Barstugan et al. [36] considered machine learning algorithms to classify 150 COVID-19 and non-COVID-19 images. Several feature extraction methods such as grey-level size zone matrix (GLSZM) and discrete wavelet transform (DWT) were considered in the feature extraction process, and the extracted features were classified using a support vector machine. K -fold cross-validations were performed in the experiments with 2, 5, and 10 folds. Barstugan et al. concluded that 99.68% of accuracy was achieved by SVM using the GLSZM feature extraction method.

Wang et al. [37] conducted another study on differentiating COVID-19 from non-COVID-19 CT scans. In their proposed network, UNet was first trained for lung region segmentation, and then, they used a pretrained UNet to test CT volumes to obtain all lung masks. They concatenated CT volumes with corresponding lung masks and sent them to the proposed DeCoVNet for the training. Wang et al. concluded that the proposed network achieved a 0.959 ROC AUC score.

Chen et al. [38] performed a study on collected 46,096 images from 106 patients (Renmin Hospital of Wuhan University–Wuhan, Hubei province, China). The proposed system was based on segmenting CT scans using UNet++ and predicting the COVID-19 lesions. The prediction was performed by dividing an image into four segments and

TABLE 1: COVID-19/normal classification results. Class.: classification; bac. pneu.: bacterial pneumonia; Sens.: sensitivity; Spec.: specificity; Prec.: precision; Acc.: accuracy; AUC: area under the curve; Ref.: reference.

Class.	Subjects	Dataset	Method	Sens. (%) or recall	Spec. (%)	Prec. (%)	Acc. (%)	AUC (%)	F1-score	Ref.
COVID-19/normal	178 pneumonia 247 normal	Private + [21–23]	DL IRRCNN	N/A	N/A	N/A	98.78	N/A	98.85	Alom et al. [9] <i>Preprint</i>
COVID-19/normal	521 COVID-19 397 normal 76 bac. pneu. 48 SARS	[24–26]	DL ShuffleNet V2	90.52	91.58	N/A	91.21	96.89	N/A	Hu et al. [10] <i>Preprint</i>
COVID-19/normal	106 COVID-19 100 normal	Private + [27, 28]	DL ResNet50	98.2	92.2	N/A	N/A	99.6	N/A	Gozes et al. [12] <i>Preprint</i>
COVID-19/normal	COVID-19: X-ray:117; CT:20 normal: X-ray:117; CT:20	[21, 22, 29]	DenseNet121 + Bagging	99.00	N/A	99.00	99.00	N/A	99.00	Kassani et al. [13] <i>Preprint</i>
COVID-19/normal	1,262 COVID-19 1,230 normal	[23]	DenseNet201	96.29	96.21	96.29	96.25	97.0	96.29	Jaiswal et al. [20] <i>Peer-reviewed</i>

counting the consecutive images. If three consecutive images were classified as containing lesions, the case was classified as positive for COVID-19. The proposed system was evaluated using five different metrics, and it achieved 92.59% and 98.85% of accuracy in prospective and retrospective testing, respectively.

Jin et al. [39] considered the segmentation and pretrained models to classify COVID-19, healthy images, and inflammatory and neoplastic pulmonary diseases. Initially, preprocessing was applied to CT scan images to standardize images that were collected from five hospitals in China. Several segmentation models such as V-Net and 3D U-Net++ were considered, and segmented images were trained using pretrained network ResNet50 [19], Inception networks [17], DPN-92 [40], and Attention ResNet-50 [41]. Jin et al. concluded that the ResNet50 achieved the highest classification rates by 0.9910 of AUC score, 97.40% of sensitivity, and 92.22% of specificity with the images segmented by 3D U-Net++ segmentation model.

Pathak et al. [42] proposed a system for the detection of COVID-19 in CT scans that considered a preproposed transfer learning. The system used the ResNet50 to extract the features from CT images, and a 2D convolutional neural network was considered for the classification. The proposed system was tested on 413 COVID-19 and 439 non-COVID-19 images with 10-fold cross-validation, and it achieved 93.01% of accuracy.

Polsinelli et al. [43] proposed a light architecture by modifying the CNN. The proposed model was tested on two different datasets, and several experiments with different combinations were performed. The proposed CNN achieved 83.00% of accuracy and 0.8333 of F1 score.

Han et al. [44] proposed a patient-level attention-based deep 3D multiple instance learning (AD3D-MIL) that learns Bernoulli distributions of the labels obtained by a pooling approach. They used a total of 460 chest CT examples, 230

CT examples from 79 COVID-19 confirmed patients, 100 CT examples from 100 patients with pneumonia, and 130 CT examples from 130 people without pneumonia. Their proposed model achieved an accuracy, AUC, and the Cohen kappa score of 97.9%, 99.0%, and 95.7%, respectively, in the classification of COVID-19 and non-COVID-19.

Harmon et al. [45] considered 2724 CT scans from 2617 patients in their study. Lung regions were segmented by using 3d anisotropic hybrid network architecture (AH-Net), and the classification of segmented 3D lung regions was performed by using pretrained model DenseNet121. The proposed algorithm achieved an accuracy, specificity, and AUC score of 0.908, 0.930, and 0.949, respectively. Table 2 shows the summary of the COVID-19/non-COVID-19 classification results.

3.3. COVID-19/Non-COVID-19 Pneumonia Classification Studies. Xu et al. [52] proposed a method that consisted of preprocessing, CT image segmentation using ResNet18, and the classification of CT scans performed by adding location-attention that provides the relative location information of the patch on the pulmonary image. The proposed method tested on the considered 618 CT samples (219 with COVID-19, 224 CT images with influenza-A viral, and 175 CT images for healthy people), and Xu et al. concluded that the overall accuracy rate of the proposed method was 86.7%.

Wang et al. [53] proposed another deep learning method to distinguish COVID-19 and other pneumonia types. The segmentation, suppression of irrelevant area, and COVID-19 analysis were the processes of the proposed method. DenseNet121-FPN [15] was implemented for lung segmentation, and COVID19Net that had a DenseNet-like structure was proposed for classification purposes. Two validation sets were considered, and the authors reported 0.87 and 0.88 ROC AUC scores for these validation sets.

TABLE 2: COVID-19/non-COVID-19 classification results. Class.: classification; Sens.: sensitivity; Spec.: specificity; Prec.: precision; Acc.: accuracy; AUC: area under the curve; Ref.: reference.

Class.	Subjects	Dataset	Method	Sens. (%) or recall	Spec. (%)	Prec. (%)	Acc. (%)	AUC (%)	F1-score	Ref.
COVID-19/ non-COVID-19	496 COVID-19 1385 others	Private + [28, 31]	CNN	94.06	95.47	N/A	94.98	97.91	NA	Jin et al. [30] <i>Preprint</i>
COVID-19/ non-COVID-19	N/A	[46]	CNN	~90	~90	N/A	~90	Not clear	~90	Singh et al. [32] <i>Peer-reviewed</i>
COVID-19/ non-COVID-19	449 COVID-19 100 normal 98 lung cancer 397 other	Private + [47, 48]	DL multitask	94	79	N/A	86	93	N/A	Amyar et al. [33] <i>Preprint</i>
COVID-19/ non-COVID-19	349 COVID-19 397 non-COVID-19	Private + [47, 49]	ResNet18	100.0	98.6	N/A	99.4	99.65	99.5	Ahuja et al. [34] <i>Peer-reviewed</i>
COVID-19/ non-COVID-19	564 COVID-19 660 non-COVID-19	[50]	VGG16 based lesion-attention DNN	88.8	N/A	87.9	88.6	94.0	87.9	Liu et al. [35] <i>Conference proceeding</i>
COVID-19/ non-COVID-19	53 COVID-19 97 other	Not clear	SVM	97.56	99.68	99.62	98.71	N/A	98.58	Barstugan et al. [36] <i>Preprint</i>
COVID-19/ non-COVID-19	313 COVID-19 229 without COVID-19	Private	UNet	90.7	91.1	N/A	90.1	95.9	N/A	Wang et al. [37] <i>Peer-reviewed</i>
COVID-19/ non-COVID-19	51 COVID-19 55 control	Private	UNet++	94.34	99.16	N/A	98.85	N/A	N/A	Chen et al. [38] <i>Preprint</i>
COVID-19/ non-COVID-19	723 COVID-19 413 others	Private	UNet++ + ResNet-50	97.4	92.2	N/A	N/A	99.1	N/A	Jin et al. [39] <i>Preprint</i>
COVID-19/ non-COVID-19	413 COVID-19 439 non-COVID-19	[32, 51]	ResNet-50 + 2D CNN	91.46	94.78	95.19	93.02	N/A	N/A	Pathak et al. [42] <i>Peer-reviewed</i>
COVID-19/ non-COVID-19	460 COVID-19 397 non-COVID-19	[26, 47]	CNN SqueezeNet	85.00	81.00	81.73	83.00	N/A	83.33	Polinelli et al. [43] <i>Preprint</i>
COVID-19/ non-COVID	230 COVID-19 130 normal	Private	AD3D-MIL	97.9	NA	97.9	97.9	99.0	97.9	Han et al. [44] <i>Peer-reviewed</i>
COVID-19/ non-COVID	1029 COVID-19 1695 non-COVID-19	Private	AH-Net DenseNet121	84.0	93.0	NA	90.8	94.9	NA	Harmon et al. [45] <i>Peer-reviewed</i>

In addition to classify COVID-19 and normal cases, Hu et al. [10] performed another experiment to differentiate COVID-19 cases from other cases as bacterial pneumonia and SARS. The average sensitivity, specificity, and the AUC score were obtained as 0.8571, 84.88%, and 92.22%, respectively.

Bai et al. [54] implemented the deep learning architecture EfficientNet B4 [55] to classify COVID-19 and pneumonia slices of CT scans. The diagnosis of the six radiologists on the corresponding patients were used to evaluate the efficiency of the results obtained by an AI model. The AI model achieved 96% of accuracy, while the average accuracy of the diagnosis of radiologists was obtained at 85%.

Kang et al. [56] proposed a pipeline and multiview representation learning technique for COVID-19 classification using different types of features extracted from CT images. They used 2522 CT images (1495 are from COVID-19 patients, and 1027 are from community-acquired pneumonia) for the classification purpose. The comparison was performed using the benchmark machine learning models, namely, support vector machine, logistic regression, Gaussian-naive-Bayes classifier, K -nearest-neighbors, and neural networks. The proposed method outperformed the considered ML models with 95.5%, 96.6%, and 93.2% in terms of accuracy, sensitivity, and specificity, respectively.

Another study was performed by Shi et al. [57] to classify COVID-19 and pneumonia. They considered 1658 and 1027 confirmed COVID-19 and CAP cases. Shi et al. proposed a model that is based on random forest and automatically extracted a series of features as volume, infected lesion number, histogram distribution, and surface area from CT images. The proposed method and considered machine learning models (logistic regression, support vector machine, and neural network) were then trained by the selected features with 5-fold cross-validation. The authors reported that the proposed method outperformed other models and produced the optimal AUC score (0.942).

Ying et al. [58] designed a network named as DRE-Net, which is based on the modifications on pretrained ResNet-50. The CT scans of 88 COVID-19 confirmed patients, 101 patients infected with bacteria pneumonia, and 86 healthy persons. The designed network was compared by the pretrained models, ResNet, DenseNet, and VGG16. The presented results showed that the designed network outperformed other models by achieving 0.92 and 0.95 of AUC scores for the image and human levels.

In addition to COVID-19/non-COVID-19 classification, Han et al. [44] performed experiments to classify COVID-19, common pneumonia, and no pneumonia cases as three classes classification. Their proposed AD3D-MIL model achieved an accuracy, AUC, and the Cohen kappa score of 94.3%, 98.8%, and 91.1%, respectively.

Ko et al. [59] proposed a model, a fast-track COVID-19 classification network (FCONet) that used VGG16, ResNet-50, InceptionV3, and Xception as a backbone to classify images as COVID-19, other pneumonia, or nonpneumonia. They considered 1194 COVID-19, 264 low-quality COVID-19 (only for testing), and 2239 pneumonia, normal, and other

disease CT scans in their study. All images were converted into grayscale image format with dimensions of 256×256 . They used rotation and zoom data augmentation procedures to maximize the number of training samples. It was concluded that FCONet based on ResNet-50 outperformed other pretrained models and achieved 96.97% of accuracy in the external validation data set of COVID-19 pneumonia images.

Li et al. [8] proposed a COVNet that used ResNet50 as a backbone to differentiate COVID-19, nonpneumonia, and community-acquired pneumonia. In their study, 4352 chest CT scans from 3322 patients were considered. A max-pooling operation was applied to the features obtained from COVNet using the slices of the CT series, and the resultant feature map was fed to a fully connected layer. This led to generate a probability score for each considered class. It was concluded that the proposed model achieved a sensitivity, specificity, and ROC AUC scores of 90%, 96%, and 0.96, respectively, for the COVID-19 class.

Ni et al. [60] considered a total of 19,291 CT scans from 14,435 individuals for their proposed model to detect COVID-19 in CT scans. Their proposed model included the combination of Multi-View Point Regression Networks (MVPNet), 3D UNet, and 3D UNet-based network for lesion detection, lesion segmentation, and lobe segmentation, respectively. Their algorithm analyzed the volume of abnormalities and the distance between lesion and pleura to diagnose the COVID-19, and it was concluded that the proposed algorithm outperformed three radiologists in terms of accuracy and sensitivity by achieving 94% and 100%, respectively. Table 3 summarizes the classification results for COVID-19/non-COVID-19 pneumonia cases.

3.4. COVID-19 Severity Classification Studies. Xiao et al. [61] implemented a pretrained network ResNet34 to diagnose COVID-19 severity. The experiments were performed using five-fold cross-validation, and 23,812 CT images of 408 patients were considered. They concluded that the model achieved the ROC AUC score of 0.987, and the prediction quality of detecting severity and nonseverity of 87.50% and 78.46%.

Zhu et al. [62] proposed a model that was optimized by traditional CNN and VGG16 to stage the COVID-19 severity. A publicly available dataset was considered, and 113 COVID-19 confirmed cases were used to test their hypothesis. Obtained scores were compared by scores given by radiologists, and it was concluded that the top model achieved a correlation coefficient (R^2) and mean absolute error of 0.90 and 8.5%, respectively.

Pu et al. [63] proposed an approach that initially segmented lung boundary and major vessels at two t points using UNet and registered these two images using a bidirectional elastic registration algorithm. Then, the average density of the middle of the lungs was used to compute a threshold to detect regions associated with pneumonia. Finally, the radiologist used to rate heat map accuracy in representing progression. In their study, two datasets that consisted of 192 CT scans were considered. Table 4 summarizes the key findings of the severity quantification studies.

TABLE 3: COVID-19/non-COVID-19 pneumonia classification results. Class.: classification; bac. pneu.: bacterial pneumonia; Sens.: sensitivity; Spec.: specificity; Prec.: precision; Acc.: accuracy; AUC: area under the curve; Ref.: reference.

Class.	Subjects	Dataset	Method	Sens. (%) or recall	Spec. (%)	Prec. (%)	Acc. (%)	AUC (%)	F1-score	Ref.
COVID-19/influ-A/normal	219 COVID-19 224 influ-A 175 normal	Private	CNN ResNet	86.7	N/A	81.3	N/A	N/A	83.9	Xu et al. [52] <i>Peer-reviewed</i>
COVID-19/CT-EGFR	1266 COVID-19 4106 CT-EGFR	Private	COVID19Net (DenseNet-like str.)	79.35	71.43	N/A	85.00	86.00	90.11	Wang et al. [53] <i>Peer-reviewed</i>
COVID-19/other pneu.	521 COVID-19 397 normal 76 bac. pneu. 48 SARS	[26, 47]	DL ShuffleNet V2	85.71	84.88	N/A	85.40	92.22	N/A	Hu et al. [10] <i>Preprint</i>
COVID-19/other pneu.	521 COVID-19 665 non-COVID-19 pneu.	Private	DNN EfficientNet B4	95	96	N/A	96	95	N/A	Bai et al. [54] <i>Peer-reviewed</i>
COVID-19/CAP	1495 COVID-19 1027 CAP	Private	Multiview representation learning	96.6	93.2	N/A	95.5	NA	N/A	Kang et al. [56] <i>Peer-reviewed</i>
COVID-19/CAP	1658 COVID-19 1027 CAP	Private	RF-based ML model	90.7	83.3	N/A	87.9	94.2	N/A	Shi et al. [57] <i>Preprint</i>
COVID-19/bac. pneu./ normal	88 COVID-19 101 bac. pneu. 86 normal	Private	DRE-Net	96	N/A	79	86	95	87	Ying et al. [58] <i>Preprint</i>
COVID-19/other pneu./ non-pneu.	230 COVID-19 100 normal	Private	AD3D-MIL	90.5	NA	95.9	94.3	98.8	92.3	Han et al. [44] <i>Peer-reviewed</i>
COVID-19/other pneu./ non-pneu.	1194 COVID-19 1357 other pneu. 998 normal 444 lung cancer	Private + [26, 47]	FCoNet ResNet50	99.58	100.0	NA	99.87	100.0	NA	Ko et al. [59] <i>Peer-reviewed</i>
COVID-19/other pneu./ non-pneu.	1292 COVID-19 1735 pneumonia 713 non-pneu.	Private	COVNet ResNet50	90	96	NA	NA	96.0	NA	Li et al. [8] <i>Peer-reviewed</i>
COVID-19/other pneu./healthy	3854 COVID-19 6871 other pneu. 8566 healthy	Private	MVPNet 3D UNet 3D UNet-based network	100	25	NA	94	NA	97.0	Ni et al. [60] <i>Peer-reviewed</i>

TABLE 4: COVID-19 severity classification results. Class.: classification; Sens.: sensitivity; Spec.: specificity; Prec.: precision; AUC: area under the curve; Ref.: reference.

Class.	Subjects	Dataset	Method	Sens. (%) or recall	Spec. (%)	Prec. (%)	AUC (%)	Ref.
COVID-19 severe/ nonsevere	23,812 COVID-19	Private	ResNet34	N/A	N/A	81.3	98.7	Xiao et al. [61] <i>Peer-reviewed</i>
COVID-19 severity score	131 COVID-19	[21]	CNN VGG16	N/A	N/A	NA	NA	Zhu et al. [62] <i>Peer-reviewed</i>
COVID-19 severity and progression	72 COVID-19 120 others	Private	UNet BER Algorithm	95	84	N/A	N/A	Pu et al. [63] <i>Peer-reviewed</i>

4. Discussion

The 13 of the 30 published articles considered in this review have been published as preprints, while the 17 of them have been published in journals after the peer-review process. Regardless of its form of publication, machine learning and deep learning have been the focus of these studies. In particular, deep learning approaches such as CNN, which performed the feature extraction process automatically, were widely used in these researches.

Besides, pretrained networks were commonly used for the segmentation, feature extraction, and classification stages. Especially DenseNet121, ResNet50, ShuffleNet V2 were successfully reported by the researchers in the classification stages, while successful results were obtained with the images produced by UNet ++ at the segmentation stage. It was pointed out by the researchers that many of the developed systems were modeled using the modifications or improvements pretrained networks to improve the classification accuracy of COVID-19 in CT images after preprocessing and segmentation stages. This has shown that widely used pretrained networks can be used very successfully at every stage of image classification. Some researchers classified COVID-19 cases using machine learning techniques instead of using deep learning approaches by extracting the features from the images and achieved high recognition results. This brings essential advantages in terms of learning speed.

However, while the images used are not standard and performing experiments on different image databases in each research does not make it possible to make a comprehensive comparison, it contributes to deduce general opinion. While the k -fold cross-validation is time-consuming, a few of the researches used it, and most of the researchers performed experiments using a hold-out method, which is based on dividing the dataset into training and testing set with defined percentages. However, this makes it challenging to analyze the consistency of the models, but it does not reduce the importance of performed experiments, obtained results, and the role of artificial intelligence in the fight against COVID-19.

5. Conclusions

COVID-19 continues to spread around the globe. New classification and prediction models using AI, together with more publicly available datasets, have been arising increasingly. However, the majority of the studies are from the preprint literature and have not peer-reviewed. Furthermore, many of

them have different classification tasks. Some of the studies have been conducted with very limited data. The data used in the studies might have come from different institutions and different scanners. Therefore, preprocessing of the data to make the radiographic images more similar and uniform is important in terms of providing more efficient analysis and consistency. The lack of demographic and clinical information of the patients is another limitation of these studies. We believe as the more dataset on COVID-19 with are available, the more accurate studies will be conducted. These findings are promising for AI to be used in the clinic as a supportive system for physicians in the detection of COVID-19.

Data Availability

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

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Ilker Ozsahin and Boran Sekeroglu contributed equally to this work.

References

- [1] World Health Organization August 2020, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>.
- [2] Worldometer August 2020, <https://www.worldometers.info/coronavirus/>.
- [3] F. Shi, J. Wang, J. Shi et al., "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Reviews in Biomedical Engineering*, 2020.
- [4] H. Bai, B. Hsieh, Z. Xiong et al., "Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT," *Radiology*, vol. 296, no. 2, pp. E46–E54, 2020.
- [5] Y. Fang, H. Zhang, J. Xie et al., "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology*, vol. 296, no. 2, pp. E115–E117, 2020.
- [6] S. S. Hare, A. N. Tavare, and V. Dattani, "Validation of the British Society of Thoracic Imaging guidelines for COVID-19 chest radiograph reporting," *Clinical Radiology*, vol. 75, no. 9, 2020.

- [7] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," *European Radiology Experimental*, vol. 2, no. 1, 2018.
- [8] L. Li, L. Qin, Z. Xu et al., "Using Artificial intelligence to Detect COVID-19 and Community-acquired pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.
- [9] M. Z. Alom, M. M. S. Rahman, M. S. Nasrin, T. M. Taha, and V. K. Asari, "COVID MTNet: COVID-19 detection with multi-task deep learning approaches," *arXiv preprint arXiv*, 2004, <https://arxiv.org/abs/2004.03747>.
- [10] R. Hu, G. Ruan, S. Xiang, M. Huang, Q. Liang, and J. Li, "Automated Diagnosis of COVID-19 Using Deep Learning and Data Augmentation on Chest CT," *medRxiv*, 2020, <https://medRxiv.org/abs/2020.04.24.20078998>.
- [11] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: practical guidelines for efficient CNN architecture design," *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV*, Springer International Publishing, vol. 11218, pp. 122–138, 2018.
- [12] O. Gozes, M. Frid-Adar, H. Greenspan et al., "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis," *arXiv*, 2020, <https://arxiv:2003.05037>.
- [13] H. Kassani, P. H. K. Sara, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach," *arXiv*, vol. 10641, 2004.
- [14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, 2018.
- [15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, "Densely connected convolutional networks," in *Proceedings of The Ieee Conference On Computer Vision And Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, 2017.
- [16] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, Honolulu, HI, USA, 2017.
- [17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, 2016.
- [18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning, AAAI'17," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4278–4284, San Francisco, CA, USA, 2017.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, 2016.
- [20] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *Journal of Biomolecular Structure & Dynamics*, pp. 1–8, 2020.
- [21] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection," *arXiv*, vol. 11988, 2006 <https://github.com/ieee8023/covid-chestxray-dataset>.
- [22] Kaggle, "RSNA Pneumonia Detection Challenge," 2020, <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge>.
- [23] Kaggle, "SARS-COV-2 CT-Scan Dataset," <https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset>.
- [24] K. Hun and R. Wen, "COVID19 Dataset," <https://github.com/KevinHuRunWen/COVID-19>.
- [25] <https://github.com/UCSD-AI4H/COVID-CT>.
- [26] COVID-19 Database, "Italian Society of Medical and Interventional Radiology (SIRM)," <https://www.sirm.org/en/category/articles/covid-19-database/>.
- [27] ChainZ <http://www.ChainZ.cn>.
- [28] S. G. Armato, G. McLennan, L. Bidaut et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.
- [29] Kaggle, "Chest X-Ray Images (Pneumonia) dataset," 2020, <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>.
- [30] C. Jin, W. Chen, Y. Cao et al., "Development and evaluation of an AI system for COVID-19 diagnosis," *MedRxiv*, 2020, <https://medRxiv.org/abs/2020.03.20.20039834>.
- [31] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P. A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Medical Imaging and Graphics*, vol. 36, no. 3, pp. 227–238, 2012.
- [32] D. Singh, V. Kumar, M. K. Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 39, no. 7, pp. 1379–1389, 2020.
- [33] A. Amyar, R. Modzelewski, and S. Ruan, "Multi-task deep learning based ct imaging analysis for covid-19: classification and segmentation," *medRxiv*, 2020, <https://medRxiv.org/abs/2020.04.16.20064709>.
- [34] S. Ahuja, B. K. Panigrahi, N. Dey, T. Gandhi, and V. Rajinikanth, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," *Applied Intelligence*, 2020.
- [35] B. Liu, X. Gao, M. He, L. Liu, and G. Yin, "A fast online COVID-19 diagnostic system with chest CT scans," in *Proceedings of KDD 2020*, New York, NY, USA, 2020.
- [36] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning methods," *ArXiv*, 2020, <https://arxiv:2003.09424>.
- [37] X. Wang, X. Deng, Q. Fu et al., "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.
- [38] J. Chen, L. Wu, J. Zhang et al., "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, 2020, <https://medRxiv.org/abs/2020.02.25.20021568>.
- [39] S. Jin, B. Wang, H. Xu et al., "AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI

- system in four weeks,” *medRxiv*, 2020, <https://medRxiv.org/abs/2020.03.19.20039354>.
- [40] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, “Dual path networks,” in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 4467–4475, Long Beach, CA, USA, 2017.
- [41] F. Wang, M. Jiang, C. Qian et al., “Residual attention network for image classification,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, Honolulu, HI, USA, 2017.
- [42] Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin, S. Singh, and P. K. Shukla, “Deep transfer learning based classification model for COVID-19 disease,” *IRBM*, 2020.
- [43] M. Polsinelli, L. Cinque, and G. Placidi, “A light CNN for detecting COVID-19 from CT scans of the chest,” *arXiv preprint arXiv*, vol. 12837, 2004.
- [44] Z. Han, B. Wei, Y. Hong et al., “Accurate screening of COVID-19 using attention based deep 3D multiple instance learning,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2584–2594, 2020.
- [45] S. A. Harmon, T. H. Sanford, S. Xu et al., “Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets,” *Nature Communications*, vol. 11, no. 1, p. 4080, 2020.
- [46] X. Li, X. Zeng, B. Liu, and Y. Yu, “COVID-19 infection presenting with CT halo sign,” *Radiol Cardiothorac Imaging*, vol. 2, no. 1, article e200026, 2020.
- [47] J. Zhao, Y. Zhang, X. He, and P. Xie, “Covid-ct-dataset: a ct scan dataset about covid-19,” *arXiv preprint arXiv*, vol. 13865, 2003.
- [48] “COVID-19 CT segmentation dataset,” <http://medicalsegmentation.com/covid19/>.
- [49] COVID-CT Dataset <https://github.com/UCSD-AI4H/COVID-CT>.
- [50] X. He, X. Yang, S. Zhang et al., “Sample-efficient deep learning for covid-19 diagnosis based on ct scans,” *medRxiv*, 2020, <https://medRxiv.org/abs/2020.04.13.20063941>.
- [51] M. E. H. Chowdhury, T. Rahman, A. Khandakar et al., “Can AI help in screening viral and covid-19 pneumonia? arXiv preprint arXiv,” vol. 13145, 2003.
- [52] X. Xu, X. Jiang, C. Ma et al., “A deep learning system to screen novel coronavirus disease 2019 pneumonia,” *Engineering*, 2020.
- [53] S. Wang, Y. Zha, W. Li et al., “A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis,” *European Respiratory Journal*, vol. 56, no. 2, article 2000775, 2020.
- [54] H. X. Bai, R. Wang, Z. Xiong et al., “AI augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT,” *Radiology*, vol. 296, no. 3, pp. E156–E165, 2020.
- [55] M. Tan and Q. V. Le, “EfficientNet: rethinking model scaling for convolutional neural networks,” *arXiv ePrints*, vol. 11946, 1905.
- [56] H. Kang, L. Xia, F. Yan et al., “Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2606–2614, 2020.
- [57] F. Shi, L. Xia, F. Shan et al., “Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification,” *arXiv*, 2003, <https://arXiv:2003.09860>.
- [58] Y. Song, S. Zheng, L. Li et al., “Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images,” *MedRxiv*, 2020, <https://medRxiv.org/abs/2020.02.23.20026930>.
- [59] H. Ko, H. Chung, W. S. Kang et al., “COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT Image: Model Development and Validation,” *Journal of Medical Internet Research*, vol. 22, no. 6, article e19569, 2020.
- [60] Q. Ni, Z. Sun, L. Qi et al., “A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images,” *European Radiology*, pp. 1–11, 2020.
- [61] L. Xiao, P. Li, F. Sun et al., “Development and Validation of a Deep Learning-Based Model Using Computed Tomography Imaging for Predicting Disease Severity of Coronavirus Disease 2019,” *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020.
- [62] J. Zhu, B. Shen, A. Abbasi, M. Hoshmand-Kochi, H. Li, and T. Q. Duong, “Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs,” *PLoS One*, vol. 15, no. 7, article e0236621, 2020.
- [63] J. Pu, J. K. Leader, A. Bandos et al., “Automated quantification of COVID-19 severity and progression using chest CT images,” *European Radiology*, 2020.

Research Article

Identification of the Key Genes Involved in the Effect of Folic Acid on Endothelial Progenitor Cell Transcriptome of Patients with Type 1 Diabetes

Yi Lu,¹ Qianhong Yang,² Wei Hu ,¹ and Jian Dong ¹

¹Department of Cardiology, Minhang Hospital, Fudan University, 170 Xin-Song Road, Shanghai 201199, China

²Department of Geriatrics, Minhang Hospital, Fudan University, 170 Xin-Song Road, Shanghai 201199, China

Correspondence should be addressed to Wei Hu; 18918169120@163.com and Jian Dong; dongjian19780413@163.com

Received 16 June 2020; Revised 19 August 2020; Accepted 11 September 2020; Published 24 September 2020

Academic Editor: Tao Huang

Copyright © 2020 Yi Lu 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.

Type 1 diabetes (T1D) is one of the most common autoimmune diseases in children. Previous studies have suggested that endothelial progenitor cells (EPCs) might be engaged in the regulating of the biological processes in T1D and folic acid (FA) might be engaged in regulating EPC function. The present study has identified 716 downregulated genes and 617 upregulated genes in T1D EPC cases after treated with FA. Bioinformatics analysis has shown that these DEGs were engaged in regulating metabolic processes, cell proliferation-related processes, bone marrow development, cell adhesion, platelet degranulation, and cellular response to growth factor stimulus. Furthermore, we have conducted and identified hub PPI networks. Importantly, we have identified 6 upregulated genes (POLR2A, BDNF, CDC27, LTN1, RAB1A, and CUL2) and 8 downregulated genes (SHC1, GRIN2B, TTN, GNAL, GNB2, PTK2, TF, and TLR9) as key regulators involved in the effect of FA on endothelial progenitor cell transcriptome of patients with T1D. We think that this study could provide novel information to understand the roles of FA in regulating EPCs of T1D patients.

1. Introduction

Type 1 diabetes (T1D) belongs to a type of autoimmune diseases featuring the destruction of insulin-producing pancreatic β -cells caused by the immune systems [1]. Type 1 diabetes is regarded as one of the most frequent chronic diseases in children and teenagers. It has contributed to a series of symptoms [2]. Insufficient control of hyperglycemia can help develop diabetic nephropathy, neuropathy, and retinopathy, which are the major causes of kidney failure, blindness, and nontraumatic amputation [3]. Patients suffering from T1D are insulin dependent and highly prone to develop vascular diseases, end-stage renal disease, and neurological damages [3]. The detailed mechanisms regulating T1D and novel therapeutic strategies for this disease remain to be further explored. Endothelial progenitor cells (EPCs) stem from the bone marrow and are critical in regulating revascularization and endothelial homeostasis [3]. Increasing evidence has shown that EPCs are significantly decreased in

diabetes patients compared with normal samples, suggesting that EPCs may be involved in the regulating of the biological processes in T1D [4].

With the development of RNA-sequence and microarray methods, emerging studies have explored the pathological mechanisms of human diseases using these novel methods and a lot of data are produced. By analyzing the big data, the researchers could find novel and useful information to understand the progression of human diseases. For example, Safari et al. have reported that YBX1, SRPK1, PSMA1/3, and XRCC6 were key regulators of T1D by using protein-protein interaction network analysis. Jia et al. have identified 329 downregulated genes and 192 upregulated genes in childhood-onset type 2 diabetes [5]. Van et al. have reported that against healthy subjects, there were 1591 genes differently expressed in T1D samples [6].

Folic acid (FA) has been reported to be important in human cell proliferation [7]. Several previous studies have shown that FA was involved in regulating endothelial

progenitor cell function and associated with the progression of coronary artery disease, hypercholesterolemia, and diabetes. However, the mechanisms of FA in regulating T1D remain unclear. This study has tried to determine differentially expressed mRNAs after treated with FA by analyzing a public dataset (GSE17635) [8]. Furthermore, coexpression analysis and bioinformatics analysis have been used to identify hub genes involved in the effect of FA on endothelial progenitor cell transcriptome of patients with T1D.

2. Material and Methods

2.1. Microarray Data. The microarray data of GSE17635 are accessible in the National Center of Biotechnology Information (NCBI) Gene Expression Omnibus database (GEO, <http://www.ncbi.nlm.nih.gov/geo/>). This dataset is aimed at investigating the difference between the gene expression profiles of endothelial progenitor cells from T1D patients before ($n = 11$) and after a four-week duration of FA supplementation ($n = 10$) and that from healthy subjects ($n = 11$). Patients with T1D ($n = 20$) were diagnosed no less than one year prior to the participation in this study. They were all from the outpatient clinic of the Department of Internal Medicine of the University Medical Centre Utrecht, The Netherlands. In manifest liver disease, macrovascular disease, creatinine $> 120 \mu\text{mol/L}$, homocysteine $> 15 \mu\text{mol/L}$, and untreated thyroid disease, the exclusion criteria were present. If those patients were receiving retreatment of vasoactive medication (angiotensin II antagonists, angiotensin-converting enzyme inhibitors, nonsteroidal anti-inflammatory drugs (NSAIDs), statins, vitamins, or FA), then the treatment ceased no less than three weeks before starting this study. Twenty both age-matched and gender-matched participants who were in healthy conditions acted as controls. A questionnaire was used to appraise the cardiovascular risk, and the measurement of some clinical parameters like blood pressure, length and weight was carried out.

The collection of the peripheral blood samples from twenty subjects with T1D and twenty age-matched and gender-matched healthy controls (CTR) at baseline was conducted. A 4-week treatment with FA (Ratiopharm) 5 mg/day was served to T1D subjects, and then, the collection of peripheral blood samples (19/20 patients) was conducted again. The protocol of this study has obtained approval from the Medical Ethical Committee of the University Medical Centre Utrecht. The written informed consent [9] has been provided by all the participants in this study.

GEO provided the downloads of the original datasets, and the \log_2 transformation was employed to preprocess them. The use of the limma package in R software version 3.3.0 (<https://www.r-project.org/>) has helped normalize all the sample data. By employing the linear models for microarray analysis (Limma) method [10], the identification of the differentially expressed mRNA and lncRNAs was achieved. An unpaired t -test was employed to count the P value of each gene, and the Benjamini-Hochberg (BH) method [11] was employed to adjust the P value into a false discovery rate (FDR). Only those genes, the FDR of which was less than 0.01, were selected as DEGs.

2.2. Construction of the PPI Network and the Module Analysis. As the protein interactions (physical and functional associations) were to be predicted, this study constructed the PPI network for DEGs (the minimum required interaction score > 0.4) [9] employing the Search Tool for the Retrieval of Interacting Genes (STRING). Following this construction of the PPI network, the Mcode plugin (degree cut-off ≥ 2 and the nodes with edges ≥ 2 core) [12] was employed to conduct a module analysis of the network. Besides, in order to visualize the PPI networks [11], Cytoscape software version 3.4.0 (http://cytoscape.org/download_old_versions.html) was employed.

2.3. GO and KEGG Pathway Analyses. To figure out how DEGs function, this study has used DAVID system [13] (<https://david.ncifcrf.gov/tools.jsp>) to perform the analysis of the GO function enrichment and the KEGG pathway enrichment. The P value (hypergeometric P value) denotes the significance of the pathway associated with the conditions. $P < 0.05$ was considered to indicate a statistically significant difference.

3. Results

3.1. Identification of DEGs in EPC of T1D Patients after Treated with FA. This study has conducted the analysis of a public expression profiling (GSE17635) in order to determine differently expressed genes (DEG) in endothelial progenitor cells after treated with PA. This dataset has included a total of 11 non-treated T1D EPC samples and 10 PA treated T1D EPC samples. This study has shown that 617 genes were overexpressed and 716 genes were downregulated in T1D EPC samples after treated with PA. Hierarchical clustering analysis of the DEGs is presented in Figure 1. The top 10 upregulated and downregulated genes after FA treatment were shown in Table 1.

3.2. Functional Annotation of DEGs in EPC of T1D Patients after Treated with FA. Furthermore, in Figure 2, we have performed GO analysis for these DEGs. Bioinformatics analysis has shown that upregulated genes were related to the regulation of a smoothed signaling pathway, ventricular system development, negative regulation of cell growth, meiotic cell cycle, very long-chain fatty acid metabolic process, collateral sprouting, glycosaminoglycan metabolic process, bone marrow development, response to pain, and ER to Golgi vesicle-mediated transport.

Meanwhile, this study has also shown that downregulated genes were associated with the regulation of positive regulation of transcription from RNA polymerase I promoter, homophilic cell adhesion via plasma membrane adhesion molecules, cell-cell signaling, cell adhesion, embryonic skeletal system development, platelet degranulation, organ morphogenesis, cellular response to growth factor stimulus, regulation of potassium ion transport, and ion transmembrane transport.

3.3. PPI Network Analysis of DEGs. The prediction of the interaction relationship between 617 upregulated DEGs and 716 downregulated DEGs has been achieved by using the

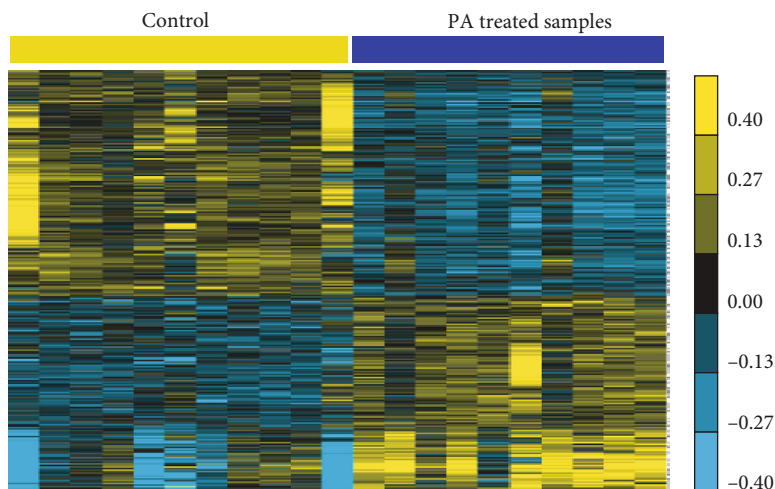


FIGURE 1: Identification of the significantly differentially expressed mRNAs in T1D patients after treated with FA. Heatmaps of the differentially expressed mRNAs in GSE17635, upregulated mRNAs, and downregulated mRNAs between control and treated sample with FA.

TABLE 1: The top 10 upregulated and downregulated genes after FA treatment.

Gene name	<i>P</i> value	Ave nontreatment	Ave treatment	Fc	Regulation
TFRC	0.006293	11.00542	12.23194	2.340016	Upregulated
ZFAND5	0.008766	10.48779	11.68049	2.285802	Upregulated
PPA2	0.008097	10.19328	11.31317	2.173302	Upregulated
LIMS1	0.004958	9.447019	10.47183	2.034697	Upregulated
SPTLC1	0.007231	9.948845	10.89993	1.933331	Upregulated
GPR183	0.008343	10.03497	10.92472	1.852857	Upregulated
STRAP	0.006873	9.901829	10.78517	1.844646	Upregulated
XPO1	0.003324	9.896716	10.76951	1.831212	Upregulated
RAB1A	0.002892	11.08631	11.94823	1.81746	Upregulated
UGP2	0.002848	10.00758	10.86298	1.80927	Upregulated
C19orf24	0.008067	11.64242	11.01402	0.646894	Downregulated
PBX2	0.0019	10.39015	9.723368	0.62991	Downregulated
CCDC106	0.007927	10.40376	9.722742	0.623726	Downregulated
SPATA20	0.00734	10.03424	9.332935	0.615015	Downregulated
PPP6R1	3.43E-05	9.85842	9.156729	0.614851	Downregulated
S100A10	0.007334	11.94692	11.22914	0.608032	Downregulated
PDLIM1	0.006748	9.121195	8.402882	0.607808	Downregulated
HVCN1	0.009486	10.39443	9.670834	0.605587	Downregulated
LHPP	0.000472	9.838647	9.111007	0.603891	Downregulated
LTBP2	0.002257	9.206243	8.462812	0.597317	Downregulated
TMEM156	0.008418	10.49832	9.638557	0.551042	Downregulated

STRING database. This study first sets up the PPI network by the use of these DEGs. After constructing the PPI network, the Mcode plugin (degree cut-off ≥ 3 and the nodes with edges ≥ 3 core) was employed to carry out a module analysis of it. The identification of 23 hub-networks was found in the downregulated DEG-mediated PPI networks and that of 17 hub-networks in the upregulated DEG-mediated PPI networks.

Figure 3 has presented the top 3 hub-networks in upregulated DEG-mediated PPI networks. Figure 3(a) shows that

there are 18 nodes and 183 edges in Hub-network 1. Figure 3(b) shows that there are 35 nodes and 142 edges in Hub-network 2. Figure 3(c) shows that there are includes 37 nodes and 94 edges hub-network 3. Six DEGs, including POLR2A, BDNF, CDC27, LTN1, RAB1A, and CUL2, have been identified as key upregulated regulators by interacting with more than 20 DEGs.

Figure 4 has presented the top 3 hub-networks in downregulated DEG-mediated PPI networks. Figure 4(a) makes it clear that 13 nodes and 78 edges exist in Hub-network 1.

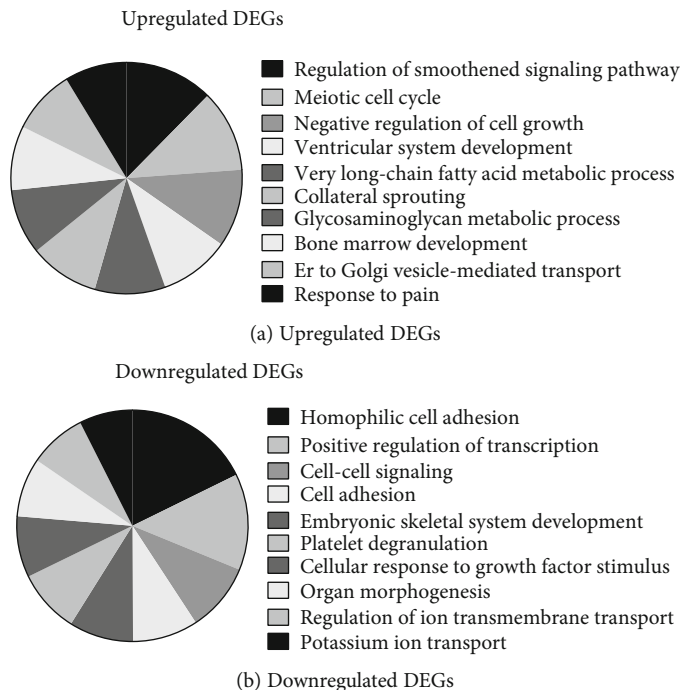


FIGURE 2: GO analysis of DEGs in EPC of T1D patients after treated with FA. (a) The analysis of the biological processes of the upregulated expressed mRNAs in EPC of T1D patients by treated with FA. (b) The analysis of the biological processes of the downregulated expressed mRNAs in EPC of T1D patients by treated with FA.

Figure 4(b) makes it clear that 8 nodes and 28 edges exist in Hub-network 2. Figure 4(c) makes it clear that 15 nodes and 49 edges exist in Hub-network 3. Eight DEGs, including SHC1, GRIN2B, TTN, GNAL, GNB2, PTK2, TF, and TLR9, have been identified as key downregulated regulators by interacting with more than 20 DEGs.

4. Discussion

Endothelial progenitor cells (EPCs) are critical in regulating the revascularization and endothelial homeostasis. Increasing evidence has shown that EPCs were notably decreased in diabetes patients compared with normal samples, suggesting that EPCs might be involved in the regulating of the biological processes in T1D. FA has been reported to act significantly in human cell proliferation. Several previous studies have shown that FA was involved in regulating endothelial progenitor cell function and associated with the progression of diabetes. For example, Anna et al. have reported that metabolic control in overweight T1D patients can be improved through DCI plus FA oral supplementation [14]. Alian et al. have found that FA administration decreased the level of endothelial dysfunction measured [15]. However, the mechanisms of FA in regulating T1D remain unclear. This study has identified DEGs involved in endothelial progenitor cells after treated with FA. The present study has also discovered that 617 genes were overexpressed and 716 genes were downregulated in T1D EPC samples after treated with FA. Furthermore, in order to identify hub genes, two PPI networks have been constructed.

Moreover, we have conducted the bioinformatics analysis for these DEGs in T1D. Our results have shown the involvement of upregulated genes in regulating multiple metabolic and cell proliferation processes, such as cell cycle and very long-chain fatty acid metabolic process. The study has also shown that these DEGs were associated with bone marrow development, which may be modulated by EPC cells. Meanwhile, this study has suggested that downregulated DEGs were involved in regulating cell adhesion, platelet degranulation, and cellular response to growth factor stimulus. These growth factors had been demonstrated to have a crucial role in T1D disease progression. For example, insulin-like growth factor-1 activates AMPK to augment mitochondrial function and correct neuronal metabolism in sensory neurons in type 1 diabetes [16]. Fibroblast growth factor 21 ameliorates diabetes-induced endothelial dysfunction in mouse aorta via activation of the CaMKK2/AMPK α signaling pathway [17]. Inhibition of epidermal growth factor receptor activation is associated with improved diabetic nephropathy in type 2 diabetes [18].

By conducting PPI network analysis, we have identified 3 downregulated hub-networks and 3 upregulated hub-networks involved in the effect of FA on endothelial progenitor cell transcriptome of patients with T1D. Importantly, we have identified 6 upregulated genes (POLR2A, BDNF, CDC27, LTN1, RAB1A, and CUL2) and 8 downregulated genes (SHC1, GRIN2B, TTN, GNAL, GNB2, PTK2, TF, and TLR9) as key regulators in this progression. Among these regulators, BDNF has been considered to be linked with the prognosis and progression of diabetes. For instance, the reduction of BDNF is regarded to partly lead to cognitive

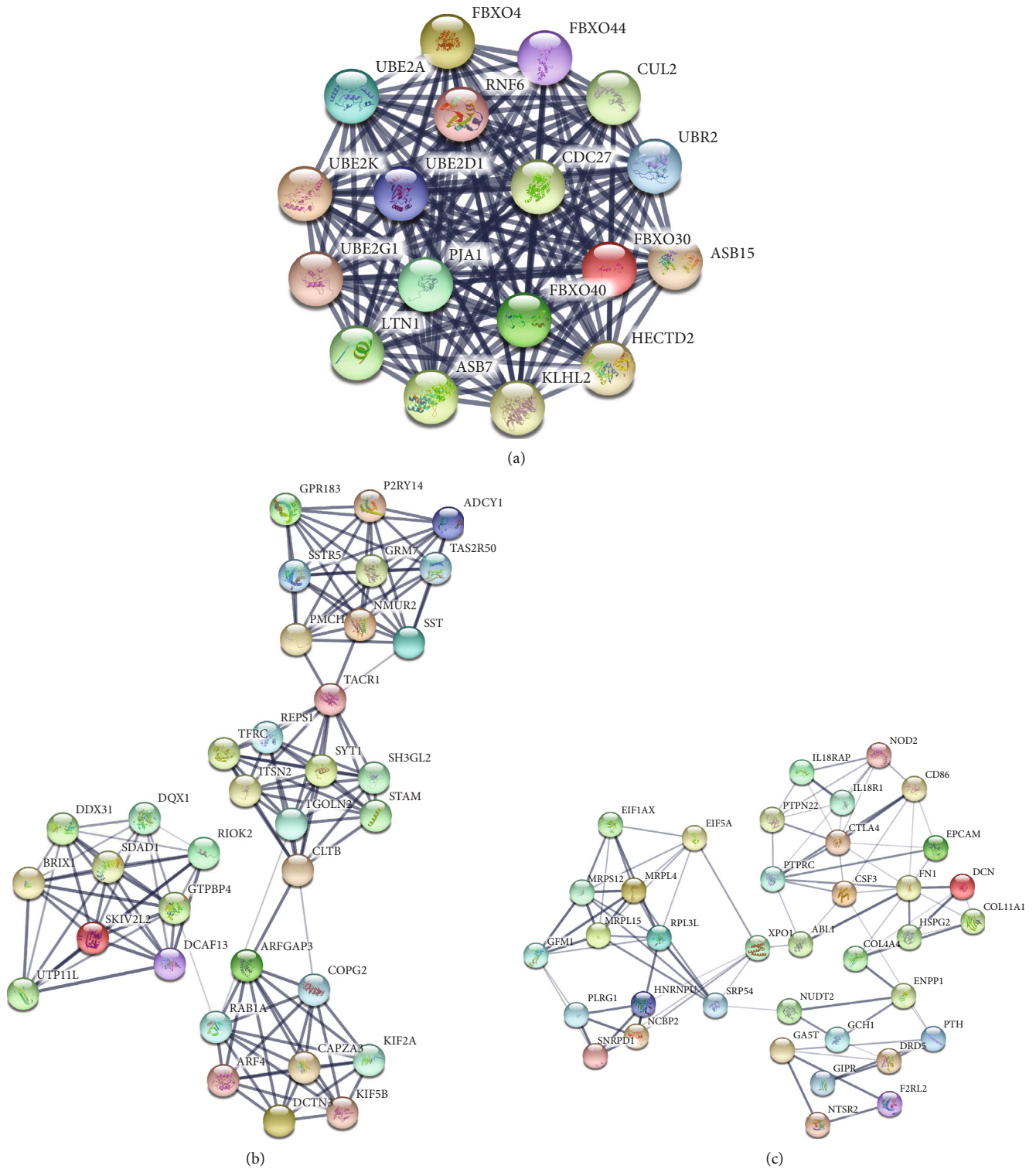


FIGURE 3: Construction of PPI network by upregulated DEGs. The PPI networks of upregulated DEGs in the top 3 hub-networks: (a) 18 nodes, 183 edges in hub-network 1; (b) 35 nodes, 142 edges in hub-network 2; and (c) 37 nodes, 94 edges in hub-network 3.

impairment in type 2 diabetes mellitus (T2DM) [19]. Proteome profiling of mitochondria analysis has shown that RAB1A was upregulated in T2DM. SHC1 has been identified as a key regulator in T1D. In a mouse model of T1D, TLR9

has been found to negatively regulate pancreatic islet beta cell growth and function. These results have suggested that the effect of FA on EPC cells in T1D may depend on these key regulators.

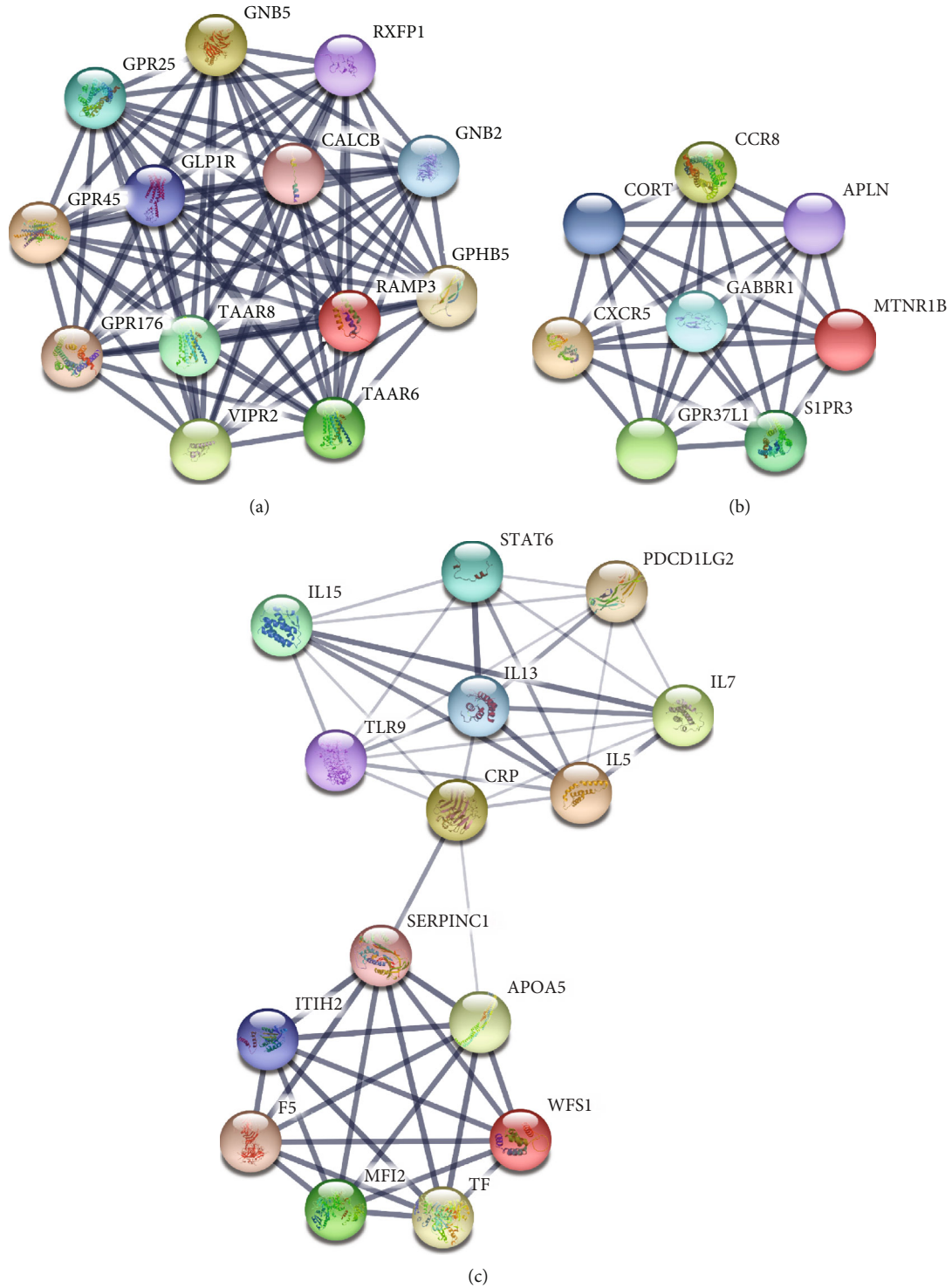


FIGURE 4: Construction of PPI network by downregulated DEGs. The PPI networks of downregulated DEGs in the top 3 hub-networks: (a) 13 nodes, 78 edges in hub-network 1; (b) 8 nodes, 28 edges in hub-network 2; and (c) 15 nodes, 49 edges in hub-network 3.

5. Conclusion

In conclusion, we have identified 716 downregulated and 617 upregulated genes in T1D EPC cases after treated with FA. Bioinformatics analysis has shown the involvement of these DEGs in regulating metabolic processes, cell proliferation-

related processes, bone marrow development, cell adhesion, platelet degranulation, and cellular response to growth factor stimulus. Furthermore, we have conducted and identified hub PPI networks. Importantly, 6 upregulated genes (POLR2A, BDNF, CDC27, LTN1, RAB1A, and CUL2) and 8 downregulated genes (SHC1, GRIN2B, TTN, GNAL,

GNB2, PTK2, TF, and TLR9) have been identified as key regulators involved in the effect of FA on endothelial progenitor cell transcriptome of patients with T1D. We think that this study could provide novel information to understand the roles of FA in regulating EPC of T1D patients.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declared that they had no conflicts of interest to this work.

Authors' Contributions

Yi Lu and Qianhong Yang contributed equally to this work.

Acknowledgments

This work is supported by The Special Plan for Clinical Research of Health Industry of Shanghai Municipal Health Commission (Grant Number 14411972900), applicant: Wei Hu; and The Shanghai Science and Technology Innovation Action Plan 2019 Natural Science Fund Project (Grant Number 19ZR1446000), applicant: Wei Hu.

References

- [1] K. M. Simmons and A. Michels, "Type 1 diabetes: a predictable disease," *World Journal of Diabetes*, vol. 6, no. 3, pp. 380–390, 2015.
- [2] K. Pippitt, M. Li, and H. E. Gurgle, "Diabetes mellitus: screening and diagnosis," *American Family Physician*, vol. 93, no. 2, pp. 103–109, 2016.
- [3] J. Tongers, J. Roncalli, and D. W. Losordo, "Role of endothelial progenitor cells during ischemia-induced vasculogenesis and collateral formation," *Microvascular Research*, vol. 79, no. 3, pp. 200–206, 2010.
- [4] M. Félétou, *The endothelium: part 1: multiple functions of the endothelial cells-focus on endothelium-derived vasoactive mediators*, Morgan & Claypool Life Sciences, San Rafael (CA), 2011.
- [5] Z. Xie, C. Chang, and Z. Zhou, "Molecular mechanisms in autoimmune type 1 diabetes: a critical review," *Clinical Reviews in Allergy & Immunology*, vol. 47, no. 2, pp. 174–192, 2014.
- [6] R. Planas, J. Carrillo, A. Sanchez et al., "Gene expression profiles for the human pancreas and purified islets in type 1 diabetes: new findings at clinical onset and in long-standing diabetes," *Clinical and Experimental Immunology*, vol. 159, no. 1, pp. 23–44, 2010.
- [7] K. S. Crider, L. B. Bailey, and R. J. Berry, "Folic acid food fortification-its history, effect, concerns, and future directions," *Nutrients*, vol. 3, no. 3, pp. 370–384, 2011.
- [8] O. van Oostrom, D. P. de Kleijn, J. O. Fledderus et al., "Folic acid supplementation normalizes the endothelial progenitor cell transcriptome of patients with type 1 diabetes: a case-control pilot study," *Cardiovascular Diabetology*, vol. 8, no. 1, pp. 47–47, 2009.
- [9] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, pp. D561–D568, 2010.
- [10] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, Springer, 2005.
- [11] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [12] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.
- [13] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.
- [14] A. R. Maurizi, M. Menduni, R. del Toro et al., "A pilot study of D-chiro-inositol plus folic acid in overweight patients with type 1 diabetes," *Acta Diabetologica*, vol. 54, no. 4, pp. 361–365, 2017.
- [15] Z. Alian, M. Hashemipour, E. Dehkordi et al., "The effects of folic acid on markers of endothelial function in patients with type 1 diabetes mellitus," *Medicinski Arhiv*, vol. 66, no. 1, pp. 12–15, 2012.
- [16] M.-R. Aghanoori, D. R. Smith, S. Shariati-Ievari et al., "Insulin-like growth factor-1 activates AMPK to augment mitochondrial function and correct neuronal metabolism in sensory neurons in type 1 diabetes," *Molecular Metabolism*, vol. 20, pp. 149–165, 2019.
- [17] L. Ying, N. Li, Z. He et al., "Fibroblast growth factor 21 ameliorates diabetes-induced endothelial dysfunction in mouse aorta via activation of the CaMKK2/AMPK α signaling pathway," *Cell Death & Disease*, vol. 10, no. 9, p. 665, 2019.
- [18] Z. Li, Y. Li, J. M. Overstreet et al., "Inhibition of epidermal growth factor receptor activation is associated with improved diabetic nephropathy and insulin resistance in type 2 diabetes," *Diabetes*, vol. 67, no. 9, pp. 1847–1857, 2018.
- [19] G. A. Moy and E. C. Mcnay, "Caffeine prevents weight gain and cognitive impairment caused by a high-fat diet while elevating hippocampal BDNF," *Physiology & Behavior*, vol. 109, pp. 69–74, 2013.

Retraction

Retracted: Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] C. Fang, X. Wang, D. Guo, R. Fang, and T. Zhu, "Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1367576, 10 pages, 2020.

Research Article

Circular RNA CircITGA7 Promotes Tumorigenesis of Osteosarcoma via miR-370/PIM1 Axis

Chuanwu Fang,¹ Xiaohong Wang,¹ Dongliang Guo,¹ Run Fang,¹ and Ting Zhu^{1,2} 

¹Department of Orthopedic Surgery, The Third Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China

²Department of Oncology, The Third Affiliated Hospital of Anhui Medical University, 390 Huaihe Road, Hefei, Anhui, China

Correspondence should be addressed to Ting Zhu; dr_zhuting@126.com

Received 3 June 2020; Revised 20 August 2020; Accepted 30 August 2020; Published 10 September 2020

Academic Editor: Chuan Lu

Copyright © 2020 Chuanwu Fang 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.

Many studies have shown that there are many circular RNA (circRNA) expression abnormalities in osteosarcoma (OS), and this abnormality is related to the development of osteosarcoma. But at present, it is unclear as to what circITGA7 has in the OS and what it does. In this study, qRT-PCR was used to detect the expression of circITGA7, miR-370, and PIM1 mRNA in OS tissues and cells. The CCK-8 assay was used to detect the effect of circITGA7 on cell proliferation. Later, the transwell assay was used to detect cell migration and invasion. The dual-luciferase reporter assay confirmed the existence of the targeting relationship between circITGA7 and miR-370, and miR-370 and PIM1. We found that circITGA7 was upregulated in OS tissues and cell lines. Knockdown of circITGA7 weakened the cell's ability to proliferate and metastasize. Furthermore, we observed that miR-370 was negatively regulated by circITGA7, while PIM1 was positively regulated by it. A functional assay validated that circITGA7 promoted OS progression via suppressing miR-370 and miR-370 affected OS proliferation and migration via PIM6 in OS. In summary, this study shows that circITGA7 promotes OS proliferation and metastasis via miR-370/PIM1.

1. Background

Osteosarcoma (OS) is a malignant cancer with a worldwide incidence of 0.2% [1]. It is, nevertheless, one of the subjects of huge interest to surgeons. One of the characteristics of osteosarcoma is that it tends to occur in adolescents, representing a bimodal pattern [2]. At the same time, osteosarcomas often occur in the distal femur or proximal tibia [3]. Because osteosarcomas do not have precancerous lesions or in situ cancers, when the diagnosis of osteosarcomas is made, the lesions that have occurred are often already malignant tumors [4]. It is, therefore, necessary to carry out the research of early diagnosis and explore the biomarker of early diagnosis and treatment [5].

Screening for molecular markers of cancer depends on whether there are differences in the molecular content between cancer cells and normal cells [6]. Alpha-fetoprotein, for example, is a glycoprotein found to be expressed at very low levels in the serum of normal adults, but high concentrations in cancer samples [7]. Similarly, some small molecules, such as circular RNA (circRNA), show differential expres-

sions between cancer cells and normal cells. CircRNAs, usually produced in eukaryotic cells, participated in the regulation of various intracellular metabolisms in cells and have differential expression in various cancer cells [8]. For example, according to Wang and Li, circ_0067934 gene is highly expressed in human non-small-cell lung cancer (NSCLC). In in vitro experiments, the inhibition of the expression of this gene can significantly reduce the proliferation and invasion of NSCLC cells [9]. At the same time, due to the circular structure of circRNA, without free ports, it is not easy to be degraded by RNA exonuclease, so it has different stability and conservation compared to linear RNA in cells. Therefore, it is considered a promising biomarker for precision medicine of tumors [8].

One of a class of small molecules closely related to cell metabolism and process is miRNA [10]. miRNAs are involved in regulating gene expression in a variety of ways [11]. For example, miRNA competitively binds RNA-binding proteins, thereby blocking the inhibition of mRNA and RNA-binding proteins and thus regulating the expression of related genes [12]. In addition, certain miRNAs may

be degraded when stimulated by specific stimulus factors, leading to the reactivation of related repressed mRNAs, thereby regulating the expression of related genes in response to the stimulus factors [11]. Similarly, miRNAs can act as oncogenes, or in some cases as tumor suppressors, to play a role in the progression of cancer [13]. Studies have found that miR-223 is shown to be upregulated in metastatic gastric cancer cells. Studies have shown that the transcription factor twist-stimulated miR-223 downregulates the expression of EPB41L3 after transcription by directly targeting its 3'-nontranslation regions, thus enhancing the migration and subsequent invasion ability of nonmetastatic gastric cancer cells as a regulatory factor [14].

CircITGA7 is a novel circular RNA, and ITGA7 is located on the human chromosome 12q13. Related studies have shown that circITGA7 is significantly dysregulated in a variety of human cancers, such as colorectal cancer and thyroid cancer. For example, previous studies have shown that circITGA7 is found to be significantly downregulated in colorectal cancer (CRC), inhibiting the proliferation and metastasis of CRC cells by inhibiting the Ras signaling pathway and promoting the transcription of ITGA7 [15]. At the same time, it is found that circITGA7 can also inhibit the proliferation of CRC by sponging miR-3187-3p and increasing the expression of ASXL1. Therefore, circ-ITGA7 may be a potential diagnostic biomarker and treatment target for CRC [16]. CircITGA7, upregulated in thyroid cancer cells, can directly bind to miR-198, reduce the inhibitory effect of miR-198 on the expression of target FGFR1, and regulate the metastasis and proliferation of TC cells, and it can be a potential marker for TC diagnosis or progression [17]. However, the role of circITGA7 in osteosarcomas is still unclear. Therefore, the focus of this research is to explore the potential functions of osteosarcomas.

In this study, the expression pattern of circITGA7 in osteosarcoma cells and the role of circITGA7 in promoting the proliferation, migration, and invasion of osteosarcoma cells were examined. The mechanism of circITGA7 in competitively binding miR-370 to regulate downstream target gene PIM1 was investigated. Our study demonstrates the potential of the circITGA7/miR-370/PIM1 axis as biomarkers for early screening, diagnosis, and monitoring of treatment progression for osteosarcoma.

2. Material and Methods

2.1. Human Tissue Samples. The osteosarcoma tissue and related normal tissues of 15 osteosarcoma patients from the Third Affiliated Hospital of Anhui Medical University were collected. The study period was between January 2010 and January 2018. Whole osteosarcoma samples were gathered from 15 OS patients and healthy controls. The patients had not received any prior chemotherapy, radiotherapy, or any other adjuvant treatment before surgery. The Research Ethics Committee of the Third Affiliated Hospital of Anhui Medical University approved this study, and all patients gave their written informed consent in this study.

2.2. Cell Lines and Cell Culture. The human OS cell lines (MG-63, HOS, U2OS, and SW1353) were procured from ATCC (Manassas, USA). The cells were cultured in a 37°C, 5% CO₂ incubator using RPMI-1640 medium. 10% FBS (BI, Israel), 100 U/mL penicillin, and 100 µg/mL streptomycin were used as supplements in the medium.

2.3. RNA Extraction and qRT-PCR. According to the manufacturer's instructions, total RNA was isolated from osteosarcoma tissue/cell line and related normal tissue/cell line using the TRIzol reagent. NanoDrop 2000c (Thermo Scientific, USA) was used for RNA quantification and quality examination. Using a SYBR Green PCR Kit (Vazyme, Nanjing, China), 2 µg of total RNA was reverse-transcribed to cDNA. Subsequently, a qRT-PCR assay was conducted using QuantStudio™ 6 Flex (Thermo Fisher, USA) on an ABI 7500 according to the company's protocol. Primers involved in this study were designed and purchased from Sangon Biotech (Shanghai, China). U6 was used to normalize the mRNA expression of miR-370, and the mRNA expression of PIM1 was normalized to GAPDH.

2.4. Cell Counting Kit-8 Proliferation Assay. The cell proliferation assay was done according to the company's protocol of the CCK-8 kit (Dojindo Laboratories, Japan). In brief, nearly 1000 transfected cells in 100 µL were grown in each well of 96-well plates. As indicated time points in the figure, cells in each well were treated with 10 µL CCK-8 solution and subsequently incubated at 37°C for 2 hours. Then, a Varioskan Flash Multimode Reader MB-580 (HEALES, China) was employed to measure the optical density at 450 nm. Five replicates were used in each group of cells.

2.5. Transwell Migration and Invasion Assays. In this assay, transwell permeable supports were employed, with a polycarbonate membrane which was coated with Matrigel (Corning Inc., USA) (for invasion tests) or without Matrigel (for migration tests). 24 hours after transfection, the cells were seeded in the upper layer in 100 µL serum-starved RPMI-1640, and then, 600 µL of medium with 10% FBS was added to the bottom layer. After 48 hours of incubation, the cells on the top layer of the transwell chamber were wiped with cotton swabs, and methanol was used to fix the cells in the lower surface for 10 minutes, and subsequently, cells were stained with 10 µg/mL of diamidino-2-phenylindole (DAPI) (Solarbio, Beijing, China) at room temperature for 15 minutes. A 200x fluorescence inverted microscope (Mshot, China) was used to photograph and count the stained cells in ten randomly selected fields.

2.6. Dual-Luciferase Reporter Assay. Approximately 5000 cells were seeded into each well of 96-well plates and subsequently cotransfected with associated plasmids and miRNA mimics or inhibitors. Lipofectamine 2000 was employed as a transfection reagent. After 48 hours of incubation, a dual-luciferase reporter assay (Promega, USA) was used to measure the luciferase activity. Independent tests were done in triplicate. For the normalizations of relative luciferase activity, it was renilla luciferase that was used as an internal control. siRNAs were designed and synthesized by RiboBio

(Guangzhou, China) as follows: si-NC: 5'-UUCUCCGAA CGUGUCACGUTT-3'; si-circITGA7: 5'-CCUAUAAUU GGAAGGACCUTT-3'. The plasmid was synthesized by RiboBio (Guangzhou, China).

2.7. Statistical Analysis. Results are shown as the mean \pm SD. To measure the difference between two groups, Student's *t*-test was employed, and one-way ANOVA was used to assess differences between more than two groups. $P < 0.05$ was considered significant.

3. Results

3.1. CircITGA7 Was Upregulated in OS Tissues and Cell Lines. To investigate the functions of circITGA7 in osteosarcoma, at first, we studied its expression in osteosarcoma tissues and cell lines. Using qRT-PCR analysis, it was demonstrated that compared with normal tissues, the expression of circITGA7 in osteosarcoma tissue had increased significantly (Figure 1(a)). Moreover, the circITGA7 expression level was significantly elevated in two OS cell lines (U2OS and SW1353) compared to the normal human cell line hFOB 1.19. Therefore, these two cell lines were selected for subsequent research (Figure 1(b)). The discoveries implied that circITGA7 could play a carcinogenic role in OS.

3.2. CircITGA7 Facilitated OS Cell Proliferation, Migration, and Invasion In Vitro. To explore the impact of circITGA7 on the biological function of OS cells, we used siRNA (si-circITGA7) to silence the circITGA7 expression in U2OS cells and the circITGA7 level was reduced by 43 percent (Figure 1(c)). CCK-8 assays were used to measure its effect. It showed that knockdown of circITGA7 could attenuate the proliferation ability of the SW1353 and U2OS cells (Figures 1(d) and 1(e)). Transwell assays illustrated that the invasive and migratory capabilities of SW1353 and U2OS cells were attenuated by silencing circITGA7 (Figures 2(a)–2(d)).

3.3. CircITGA7 Could Upregulate the Expression of PIM1 by miR-370. It has been shown that circRNAs can function as competing RNAs to bind miRNAs in the cytoplasmic space. We assessed binding between circITGA7 and miRNAs using online bioinformatics databases (miRDB, miRTarbase, and miRmap). The miR-370 was shown to have binding sites for circITGA7 as demonstrated by software prediction methods, and the expression of miR-370 in OS cells SW1353 and U2OS increased by 2.5 times and 4 times after the knockdown of circITGA7, respectively (Figure 3(c)). In order to validate the interaction between miR-370 and circITGA7, wild-type (wt) or mutant (mut) targeted sites of miR-370 in circITGA7 were cloned into pGL3 plasmid. Findings obtained from the dual-luciferase reporter assay showed that miR-370 mimics considerably reduced the luciferase activity driven by wild-type circITGA7 in SW1353 and U2OS cells (Figures 3(a) and 3(b)). But the mutant circITGA7 was not impacted by miR-370. Furthermore, the binding site of miR-370 in 3'-UTR of PIM1 was obtained by analysis. We then examined the interaction between

miR-370 and PIM1 by a dual-luciferase reporter assay (Figure 3(d)). PIM1 expression was evidently shown to be lower in SW1353 cells transfected with miR-370 than those cells with other processing groups (Figure 3(e)). Further investigation of PIM1 expression identified that anti-miR-370 could reverse the downregulation of PIM1 induced by circITGA7 knockdown in OS cells (Figure 3(f)). Taken as a whole, our results indicated that circITGA7 could upregulate PIM1 expression by miR-370.

3.4. miR-370 Reversed the Effect of CircITGA7. Using CCK-8 and transwell assay, we demonstrated that miR-370 acted as a tumor suppressor in OS. The results showed that miR-370 overexpression suppressed cell proliferation and migration in U2OS (Figures 4(a) and 4(c)); however, miR-370 knockdown enhanced cell proliferation and migration in SW1353 (Figures 4(b) and 4(d)).

In order to further explore the role of the circITGA7/miR-370 axis in OS, anti-miR-370 was arranged to be transfected into circITGA7 silenced SW1353 cells. CCK-8 and transwell assays were used to examine the effect of this treatment on OS cell proliferation and metastasis. Based on this, it was found that the proliferation and metastasis of SW1353 cells were inhibited due to downregulation of circITGA7 and that transfection with anti-miR-370 could partially reverse this inhibitory effect (Figures 4(e)–4(h)).

3.5. The Knockdown of PIM1 Partially Impeded the Effects of miR-370 Suppression on OS Cells. The above studies indicate that PIM1 is the target gene of miR-370. In order to continue to deeply investigate whether the effect of miR-370 on the physiological function of OS cells depends on PIM1 expression, we used siRNA (si-PIM1) to knock down PIM1 on OS cells that had been transfected with anti-miR-370 (Figures 4(i) and 4(j)). As shown in the figure, after knocking down PIM1, the cell proliferation and metastasis of SW1353 and U2OS were severely hindered. In conclusion, circITGA7 affected OS cell proliferation and migration through the miR-370/PIM1 axis.

4. Discussion

The results of our study showed that circITGA7 expression in osteosarcoma tissues or cells was significantly higher than that in corresponding normal tissues or cells. In vitro experiments, si-circITGA7 was used to knock down circITGA7 in SW1353 and U2OS cells. The results showed that circITGA7 knockdown reduced the proliferation of SW1353 and U2OS cells and dramatically checked the migration and invasion of osteosarcoma cells. To investigate the regulatory mechanism of circITGA7 in osteosarcoma, we found that circITGA7 promoted the progression of osteosarcoma cells by regulating the miR-370/PIM1 axis.

CircRNAs lack free 3' or 5' ends due to their circular structure, which makes it difficult for conventional mechanisms to degrade them. This structural characteristic results in a long half-life and intracellular stability and conservation of circRNAs. CircRNAs play more than one role in cells, and some regulate gene expression by regulating mRNA

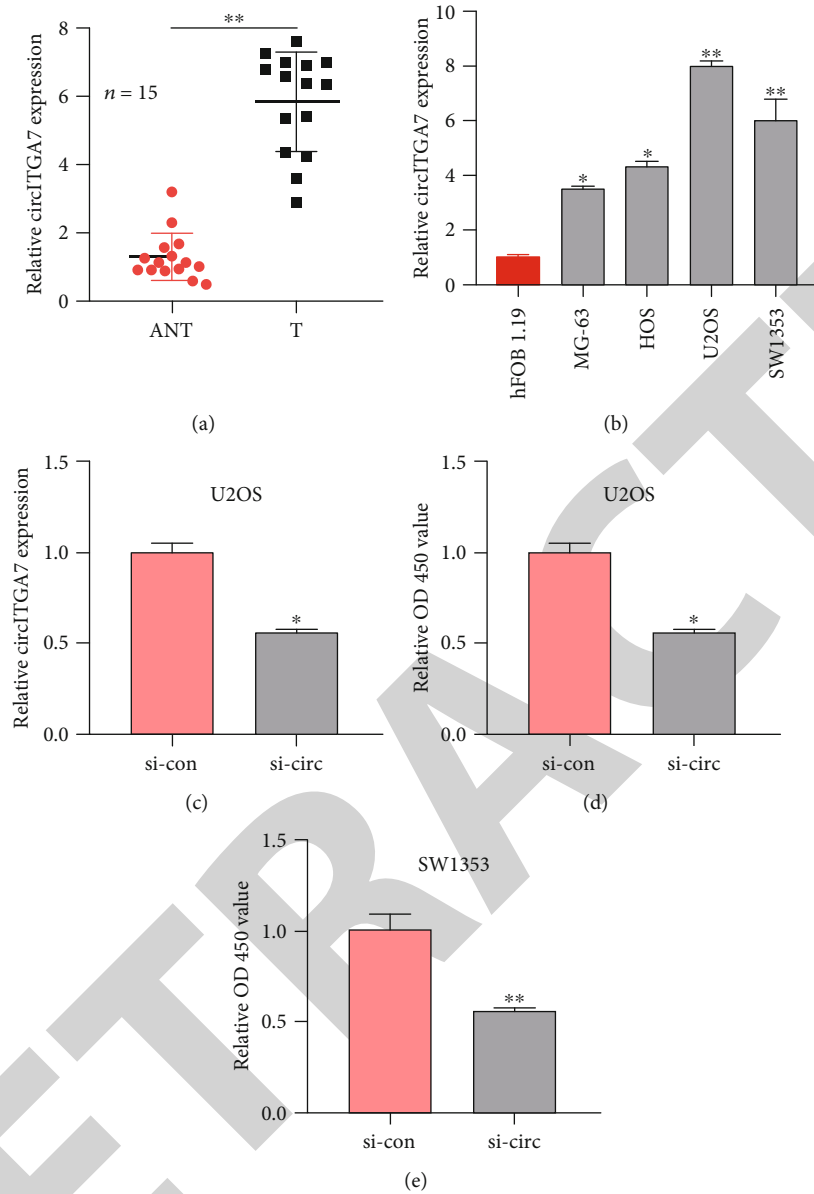


FIGURE 1: Expression and influence of circITGA7 in OS. (a) The expression of circITGA7 in 15 pairs of OS tissues and corresponding normal tissues. (b) The relative expression of circITGA7 in four OS cell lines (MG-63, HOS, U2OS, and SW1353) and one normal cell line (hFOB 1.19). (c) The relative expression of circITGA7 in cells after circITGA7 knockdown. (d, e) Relative OD 450 value of U2OS and SW1353 with si-circ or si-con. * $P < 0.05$, ** $P < 0.01$.

expression of specific genes [8]. Other circRNAs act as inclusions. Other circRNAs may bind transcription factors that are involved in muscle development or viral transcription. Similarly, circRNAs have different mechanisms involved in the development and progression of cancer [18]. Firstly, it has been reported that a genomic translocation leads to the production of a new circRNA that facilitates cell transformation, stimulates cell activity, and promotes tumor development and progression [19]. Secondly, the downregulation of certain circRNA expression levels also affects tumor progression directly or indirectly [8]. It has been found that overexpressed circITCH may regulate the Wnt/attenuated catenin pathway by regulating the activity of miR-7 and miR-214 to inhibit the development of cancer [20].

Finally, a few circRNAs have tumor-promoting effects, and their upregulation may lead to cancer. According to Yang et al., circAmotl1 is highly expressed in cancer tissue samples and related cancer cell lines, which can promote the proliferation of cancer cells [21]. Similarly, circRNA also plays a regulatory role in osteosarcoma. For example, it has been found that circRNA-0008717 may promote the occurrence and development of OS by targeting and competitively inhibiting the regulatory effects of miR-203 [22]. The circRNA circITGA7 studied in this study was found to have differential expression between cancer cells and normal cells. This suggests that circITGA7 may be involved in the development and progression of tumor tissues.

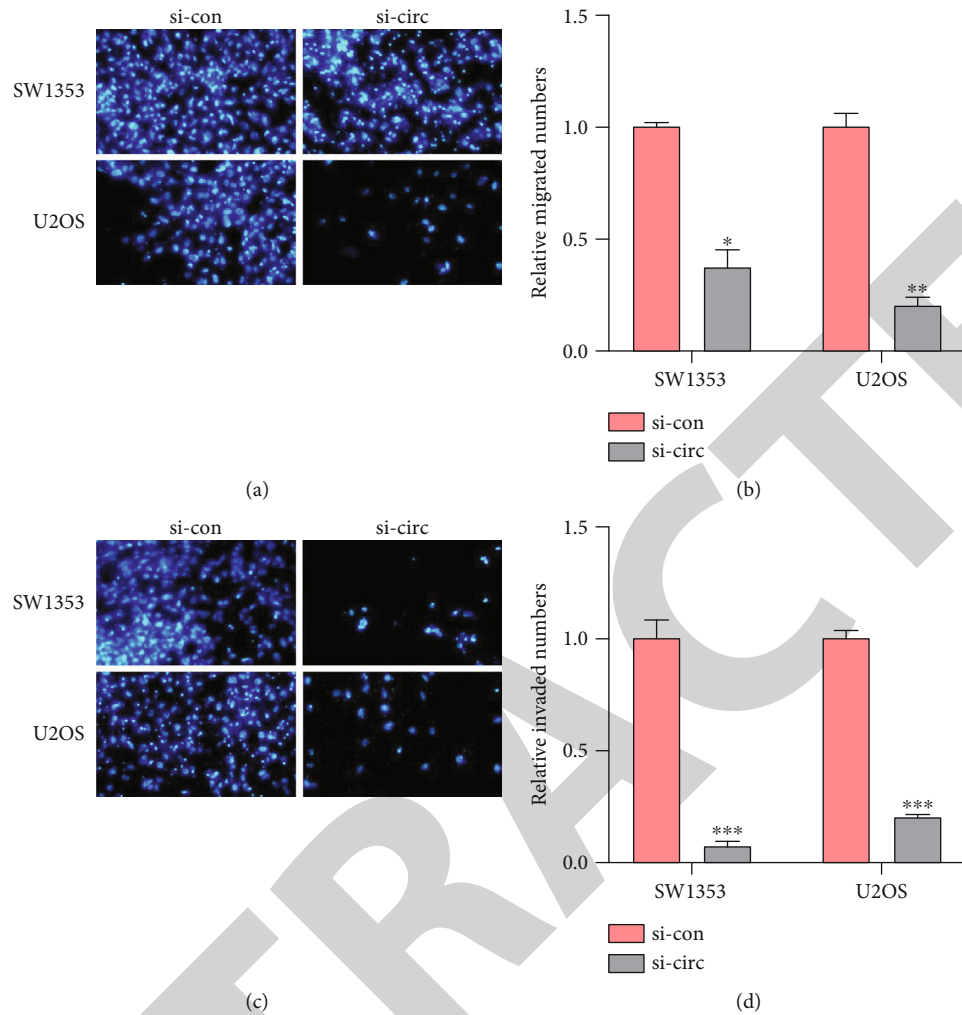


FIGURE 2: Cell migration and invasion by transwell assay. (a, b) Number of migrations in SW1353 and U2OS cells after different treatments. (c, d) The relative invaded numbers of SW1353 and U2OS cells with si-circITGA7 were less. * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.

In this study, we mainly discussed the function and mechanism of circITGA7 in OS. Studies have clarified that the expression of circITGA7 in osteosarcoma tissues is significantly lower than that in paracancer tissues. Similarly, the expression level of circITGA7 in the osteosarcoma cell lines is also lower than that of the normal colorectal mucosal epithelial cell line FHC, and its expression level is correlated negatively with OS tumor size, lymph node metastasis, distant metastasis, and overall TNM staging. In vitro experiments showed that circITGA7 could bind to hsa-miR-370-3p to upregulate its target gene NF1 level and then inhibit Ras pathway, further reduce Ras protein level and the phosphorylation level of Erk and Akt, and finally play a role in inhibiting the growth and metastasis of osteosarcoma. Interestingly, previous studies have shown that circITGA7 is significantly downregulated in colorectal cancer (CRC), inhibiting the proliferation and metastasis of CRC cells in a variety of ways [15, 16]. Our results differ from those of the previous studies in which circITGA7 inhibited CRC cells. We found that in vitro knockout of the circITGA7 gene in the cell line inhibited the proliferation of osteosarcoma cells and reduced the ability of osteosarcoma cells to migrate

and invade. Some studies show that the function of miR-370 was achieved mainly by regulating the activity of the target miR-370. For example, Yungang et al. found that miR-370 targeted FoxM1 functions as a tumor suppressor in large square cell carcinoma (LSCC) [23]; Zhang et al. showed that miRNA-370 has the tumor suppressive role by targeting FoxM1 in acute myeloid leukemia [24].

MicroRNAs are small noncoding RNAs of less than 30-nucleotide long. However, miRNA plays an extremely important role in the life activities of cells, affecting various physiological and metabolic activities of cells, including proliferation, differentiation, and apoptosis [25]. There is increasing evidence that abnormal upregulation or downregulation of miRNAs is associated with the occurrence and development of a variety of human tumors. For example, miRNA-29c inhibits the ability of lung cancer cells to migrate and invade [26]. Similarly, studies have shown that miR-370 has a potential carcinogenic effect by directly targeting the PDHB gene to promote the development of melanoma [27]. Studies have shown that miR-370 inhibits liver cancer or hepatocellular carcinoma by directly targeting PIM1, and similar results have been found in esophageal squamous cell

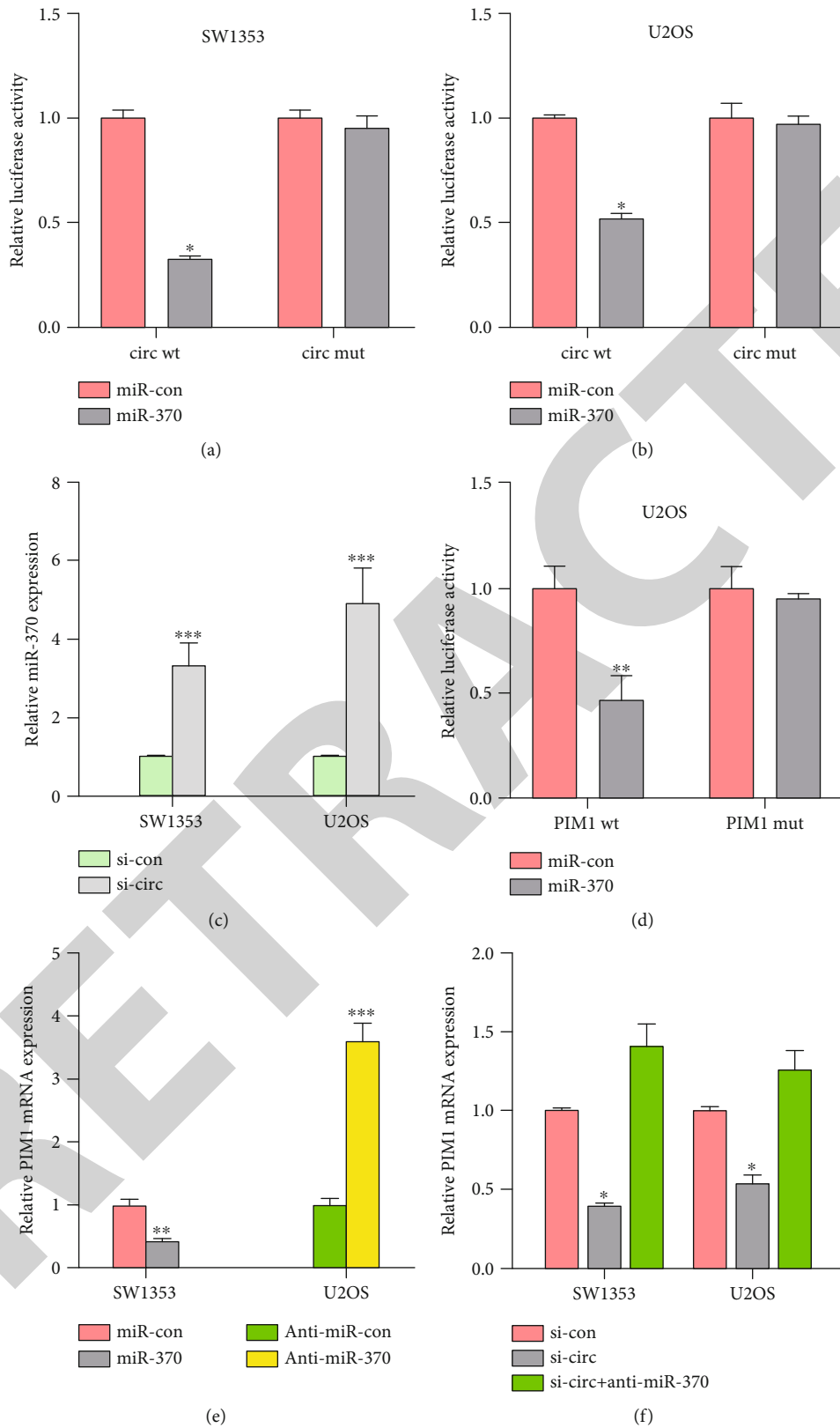


FIGURE 3: miR-370 is the target of circITGA7. (a, b) The relative luciferase activity in SW1353 and U2OS cells cotransfected with miR-370 and circ wt or miR-370 and circ mut. (c) After knocking down circITGA7, the miR-370 expression in OS cells. (d) The relative luciferase activity of PIM1 wt but not PIM1 mut was reduced by miR-370. (e) miR-370 overexpression lowered the PIM1 mRNA expression, while anti-miR-370 increased the PIM1 mRNA expression. (f) Anti-miR-370 compensated for the decreased PIM1 mRNA expression caused by si-circITGA7. * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.

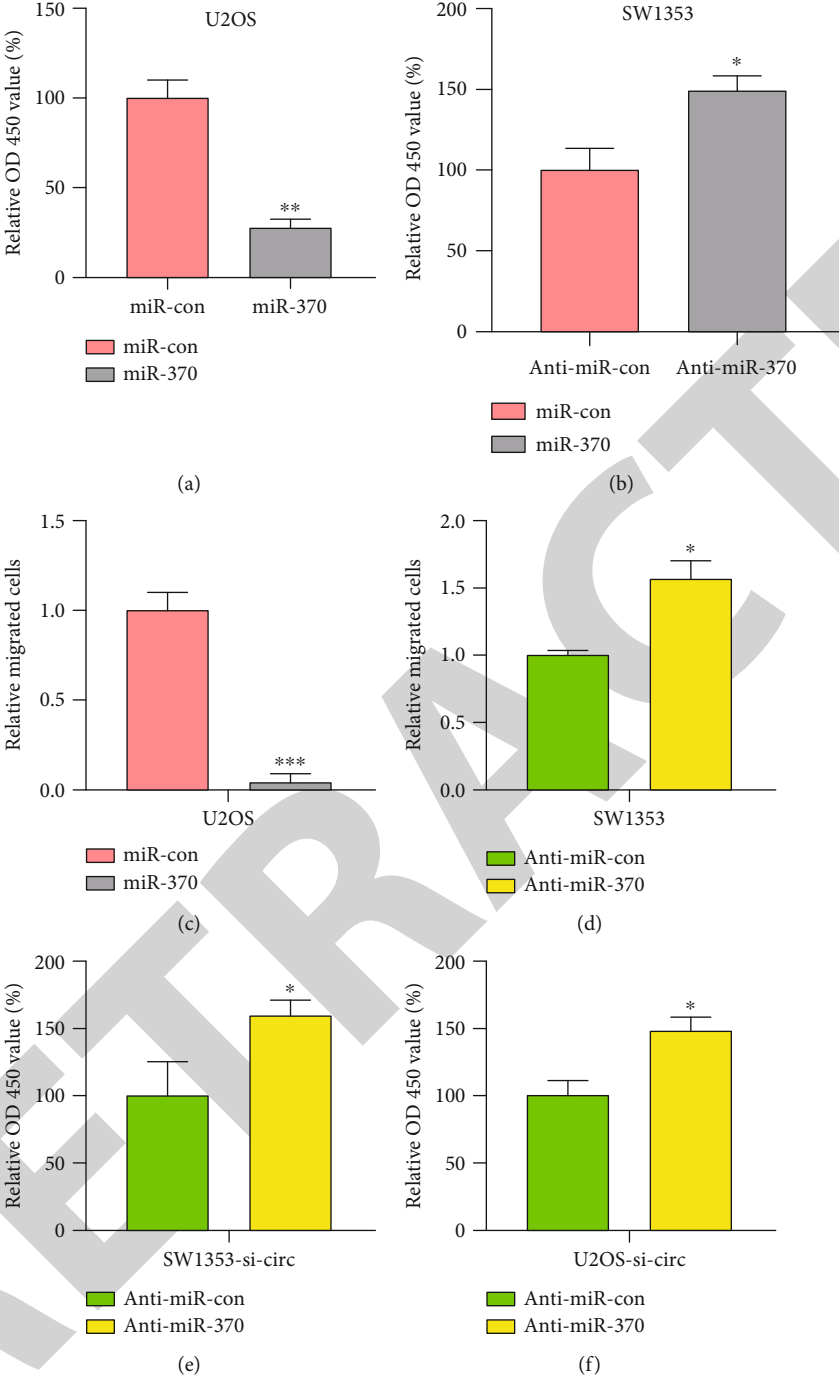


FIGURE 4: Continued.

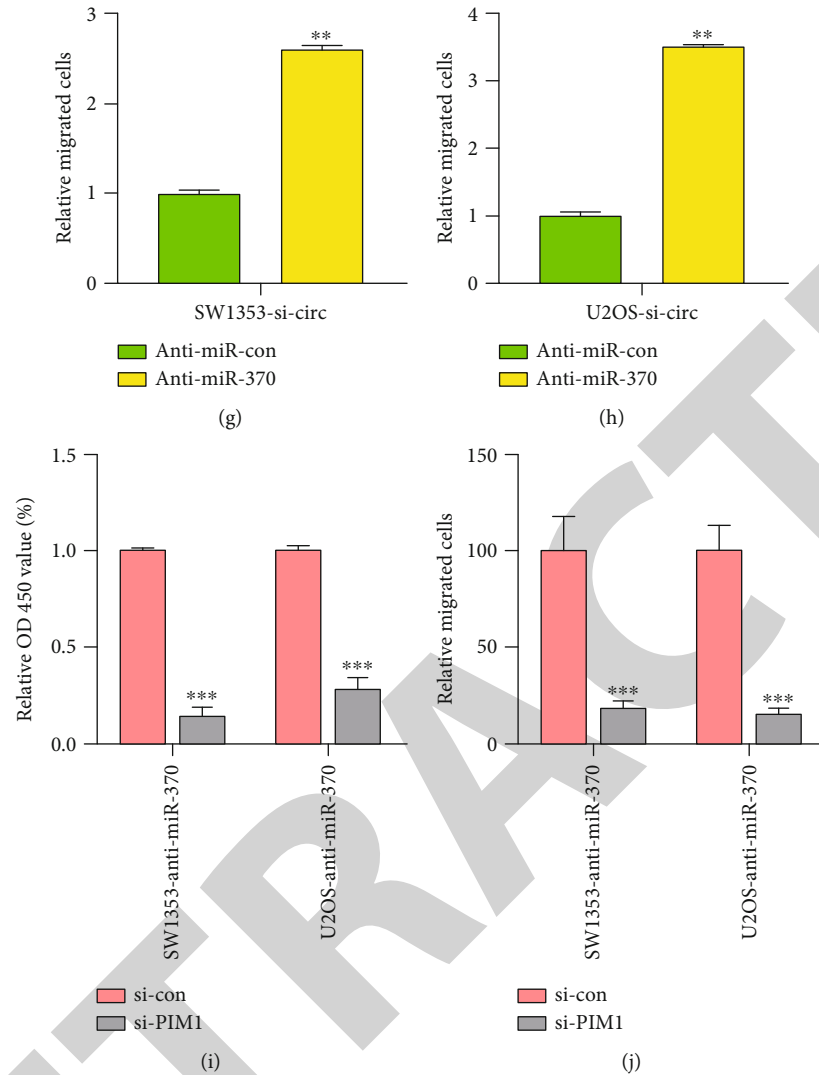


FIGURE 4: miR-370 reversed the effect of circITGA7 and the effect of PIM1 on OS cells. (a) Overexpression of miR-370 in U2OS cells decreased OD 450 value. (b) OD 450 value was higher in SW1353 cells with anti-miR-370. (c) Relative migration cells of U2OS overexpressed miR-370. (d) The relative migration of SW1353 cells with low miR-370 expression. (e, f) OD 450 value under the low expression of miR-370 after the OS cells knocked down the circITGA7. (g, h) Migrated cells under the downregulation of miR-370 in OS cells after circITGA7 knockdown. (i) OD 450 value of OS cells with si-PIM1 and anti-miR-370. (j) Relative migrated cells of OS cells with si-PIM1 and anti-miR-370. * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.

carcinoma (ESCC) [28]. miR-370 impedes cell proliferation and tumor growth by directly targeting PIM1. PIM1 is a serine/threonine-specific kinase 1 involved in the development of cancer cells [29]. PIM1 is an oncogene that can promote the growth and metastasis of colorectal cancer (CRC). Its expression is positively correlated with the progression of CRC and can predict the prognosis of patients with CRC [30]. PIM1 is significantly overexpressed in gallbladder cancer (GBC) tissue, and its expression level is positively correlated with clinical malignancies and poor prognosis. PIM1 is a promising therapeutic target for the treatment of human GBC [31]. In triple-negative breast cancer (TNBCs) tumors and their cell models, PIM1 expression is related to several transcription signals involving the transcription factor MYC. PIM1 is a malignant-cell-selective target in TNBC, and PIM1 inhibitors have potential use in sensitizing TNBC to chemotherapy-induced apoptotic cell death [32].

Our results show that circITGA7 regulates the activity of miR-370, thereby affecting the activity of PIM1, which promotes the proliferation and metastasis of OS cells.

There are still some limitations in this study. First of all, the number of osteosarcoma tissue samples we used in the study was small. Secondly, the expression pattern and prognostic value of miR-370 and PIM1 should be further explored. Therefore, further research is needed to clarify the potential mechanism of miR-370 and PIM1 in osteosarcoma.

In summary, in this study, we found through a series of experiments that circITGA7 could act as a sponge to competitively inhibit the activity of miR-370 and regulate the physiological activity of osteosarcoma cells through miR-370/PIM1, thus promoting the development of osteosarcoma cells. This provides the possibility of biomarkers for the early screening, diagnosis, and later treatment monitoring of OS.

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.

Authors' Contributions

Chuanwu Fang took part in study design, data collection, statistical analysis, data interpretation, manuscript preparation, literature search, and fund collection. Xiaohong Wang participated in data collection, statistical analysis, data interpretation, manuscript preparation, literature search, and fund collection. Dongliang Guo and Run Fang handled study design, data collection, data interpretation, manuscript preparation, and literature search. Ting Zhu conducted data design, literature search, and fund collection.

References

- [1] P. Picci, "Classic osteosarcoma," in *Atlas of Musculoskeletal Tumors and Tumorlike Lesions: The Rizzoli Case Archive*, P. Picci, M. Manfrini, N. Fabbri, M. Gambarotti, and D. Vanel, Eds., pp. 147–152, Springer International Publishing, Cham, 2014.
- [2] D. D. Moore and H. H. Luu, "Osteosarcoma," in *Orthopaedic Oncology: Primary and Metastatic Tumors of the Skeletal System*, T. D. Peabody and S. Attar, Eds., pp. 65–92, Springer International Publishing, Cham, 2014.
- [3] H. A. Finn and M. A. Simon, "Limb-salvage surgery in the treatment of osteosarcoma in skeletally immature individuals," *Clinical Orthopaedics and Related Research*, vol. 262, pp. 108–118, 1991.
- [4] R. Gorlick, "Current concepts on the molecular biology of osteosarcoma," in *Pediatric and Adolescent Osteosarcoma*, N. Jaffe, O. S. Bruland, and S. Bielack, Eds., pp. 467–478, Springer US, Boston, MA, 2010.
- [5] Y. Hua, X. Jia, M. Sun et al., "Plasma membrane proteomic analysis of human osteosarcoma and osteoblastic cells: revealing NDRG1 as a marker for osteosarcoma," *Tumor Biology*, vol. 32, no. 5, pp. 1013–1021, 2011.
- [6] D. Sidransky, "Emerging molecular markers of cancer," *Nature Reviews Cancer*, vol. 2, no. 3, pp. 210–219, 2002.
- [7] B. Adamczyk, T. Tharmalingam, and P. M. Rudd, "Glycans as cancer biomarkers," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1820, no. 9, pp. 1347–1353, 2012.
- [8] J. Li, J. Yang, P. Zhou et al., "Circular RNAs in cancer: novel insights into origins, properties, functions and implications," *American Journal of Cancer Research*, vol. 5, no. 2, pp. 472–480, 2015.
- [9] J. Wang and H. Li, *Circ RNA circ_0067934 silencing inhibits the proliferation, migration and invasion of NSCLC cells and correlates with unfavorable prognosis in NSCLC*, pp. 3053–3060, 2018.
- [10] M. Boehm and F. J. Slack, "Micro RNA control of lifespan and metabolism," *Cell Cycle*, vol. 5, no. 8, pp. 837–840, 2006.
- [11] A. Wilczynska and M. Bushell, "The complexity of miRNA-mediated repression," *Cell Death & Differentiation*, vol. 22, no. 1, pp. 22–33, 2015.
- [12] S. A. Ciafrè and S. Galardi, "MicroRNAs and RNA-binding proteins," *RNA Biology*, vol. 10, no. 6, pp. 934–942, 2014.
- [13] B. Zhang, X. Pan, G. P. Cobb, and T. A. Anderson, "MicroRNAs as oncogenes and tumor suppressors," *Developmental Biology*, vol. 302, no. 1, pp. 1–12, 2007.
- [14] X. Li, Y. Zhang, H. Zhang et al., "miRNA-223 promotes gastric cancer invasion and metastasis by targeting tumor suppressor EPB41L3," *Molecular Cancer Research*, vol. 9, no. 7, pp. 824–833, 2011.
- [15] X. Li, J. Wang, C. Zhang et al., "Circular RNA circITGA7 inhibits colorectal cancer growth and metastasis by modulating the Ras pathway and upregulating transcription of its host gene ITGA7," *The Journal of Pathology*, vol. 246, no. 2, pp. 166–179, 2018.
- [16] G. Yang, T. Zhang, J. Ye et al., "Circ-ITGA7 sponges miR-3187-3p to upregulate ASXL1, suppressing colorectal cancer proliferation," *Cancer Management and Research*, vol. Volume 11, pp. 6499–6509, 2019.
- [17] S. Li, J. Yang, X. Liu, R. Guo, and R. Zhang, "circITGA7 functions as an oncogene by sponging miR-198 and upregulating FGFR1 expression in thyroid cancer," *BioMed Research International*, vol. 2020, Article ID 8084028, 8 pages, 2020.
- [18] L. M. Holdt, A. Kohlmaier, and D. Teupser, "Molecular roles and function of circular RNAs in eukaryotic cells," *Cellular and Molecular Life Sciences*, vol. 75, no. 6, pp. 1071–1098, 2018.
- [19] J. Guarnerio, M. Bezzi, J. C. Jeong et al., "Oncogenic role of fusion-circRNAs derived from cancer-associated chromosomal translocations," *Cell*, vol. 165, no. 2, pp. 289–302, 2016.
- [20] L. Wan, L. Zhang, K. Fan, Z. X. Cheng, and J. J. Wang, "Circular RNA-ITCH suppresses lung cancer proliferation via inhibiting the Wnt/ β -catenin pathway," *BioMed Research International*, vol. 2016, no. 1, 11 pages, 2016.
- [21] Q. Yang, W. W. du, N. Wu et al., "A circular RNA promotes tumorigenesis by inducing c-myc nuclear translocation," *Cell death and differentiation*, vol. 24, no. 9, pp. 1609–1620, 2017.
- [22] X. Zhou, D. Natino, Z. Qin et al., "Identification and functional characterization of circRNA-0008717 as an oncogene in osteosarcoma through sponging miR-203," *Oncotarget*, vol. 9, no. 32, pp. 22288–22300, 2018.
- [23] W. Yungang, L. Xiaoyu, T. Pang, L. Wenming, and X. Pan, "miR-370 targeted FoxM1 functions as a tumor suppressor in laryngeal squamous cell carcinoma (LSCC)," *Biomedicine & Pharmacotherapy*, vol. 68, no. 2, pp. 149–154, 2014.
- [24] X. Zhang, J. Zeng, M. Zhou et al., "The tumor suppressive role of miRNA-370 by targeting FoxM1 in acute myeloid leukemia," *Molecular Cancer*, vol. 11, no. 1, p. 56, 2012.
- [25] K. Felekis, E. Touvana, C. Stefanou, and C. Deltas, "MicroRNAs: a newly described class of encoded molecules that play a role in health and disease," *Hippokratia*, vol. 14, no. 4, pp. 236–240, 2010.
- [26] H. Wang, Y. Zhu, M. Zhao et al., "miRNA-29c suppresses lung cancer cell adhesion to extracellular matrix and metastasis by targeting integrin β 1 and matrix metalloproteinase2 (MMP2)," *PLoS One*, vol. 8, no. 8, p. e70192, 2013.
- [27] S. Wei and W. Ma, "miR-370 functions as oncogene in melanoma by direct targeting pyruvate dehydrogenase B," *Biomedicine & Pharmacotherapy*, vol. 90, pp. 278–286, 2017.

Retraction

Retracted: Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] K. Zhou, J. Xu, X. Yin, and J. Xia, "Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 7236245, 9 pages, 2020.

Research Article

Long Noncoding RNA HAGLROS Promotes Cell Invasion and Metastasis by Sponging miR-152 and Upregulating ROCK1 Expression in Osteosarcoma

Kaifeng Zhou, Jun Xu, Xiaofan Yin , and Jiangni Xia 

Department of Orthopaedics, Minhang Hospital, Fudan University, 170 Xin-Song Road, Shanghai 201199, China

Correspondence should be addressed to Xiaofan Yin; yin_xiaofan@fudan.edu.cn and Jiangni Xia; xiajiangni770523@163.com

Received 9 June 2020; Revised 21 August 2020; Accepted 28 August 2020; Published 9 September 2020

Academic Editor: Tao Huang

Copyright © 2020 Kaifeng Zhou 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.

Background. Long noncoding RNAs (lncRNAs) played a crucial role in a number of biological processes. lncRNA HAGLROS was demonstrated to facilitate cell proliferation and migration in various cancers. However, the functions and molecular mechanisms of HAGLROS in osteosarcoma remained to be elucidated. **Methods.** qRT-PCR assay was used to detect the relative expression of HAGLROS in osteosarcoma tissue samples and cells. CCK-8 and Transwell assays were performed to assess the effects of HAGLROS on OS cells proliferation and invasion. Luciferase reporter assay verified the interaction between ROCK1 and miR-152. **Results.** In our study, we found that the expression of HAGLROS increased osteosarcoma samples and cell lines compared with normal tissues and cells. HAGLROS knockdown inhibited certain functions of U2OS and SW1353 cells in vitro. Moreover, HAGLROS depletion inhibited tumor growth and metastasis in vivo. Mechanically, we found that HAGLROS sponged miR-152 to promote ROCK1 expression in U2OS and SW1353 cells. **Conclusion.** In summary, our study indicated that HAGLROS could promote osteosarcoma progression by sponging miR-152 to promote ROCK1 expression. The results showed HAGLROS/miR-152/ROCK1 axis might act as a novel therapeutic strategy for osteosarcoma.

1. Background

Osteosarcoma (OS) is a rare malignant tumor with low incidence, frequent onset in adolescents and the elderly, high metastasis, and poor prognosis [1]. In addition to genetic factors, the causes of osteosarcoma mainly includes certain chemical agents, high doses of radiation, and certain viruses [2]. Over the past decades, *considerable* progress has been made in the treatment of osteosarcoma, including the development of chemical drugs and the improvement of surgical techniques, which has greatly improved the survival time and life quality of OS patients. However, there is still lacking the understanding of the pathogenic mechanisms related to osteosarcoma.

The genetic information in DNA is transcribed, translated, and modified into proteins, which participates in regulating a series of physiological metabolic activities of cells. Interestingly, more than 95% of the human genomes could not be translated into proteins. Recent studies revealed that

these long noncoding RNAs had an important role in the regulation of physiological activities in human cancers. Long noncoding RNA (lncRNA) is composed of more than 200 nucleotides [3]. lncRNAs participate in regulatory processes in various forms, including binding to DNA, proteins, and RNA molecules or combining with the above substances to regulate the transcription and translation of genetic information [4]. lncRNAs can be used as miRNA recognition elements and regulatory elements and interact with miRNAs as part of the regulatory network to affect the activities of miRNAs [5]. For example, in the zebrafish model, the interaction between 7SL lncRNA and miR-125b was detected. miR-125b could downregulate the expression of 7SL RNA lncRNA in zebrafish, thereby affecting the regulation effect of 7SL lncRNA on intracellular physiological activities [6]. HAGLROS increases in multiple tumor cells and acts as an oncogene, which may promote tumor development by regulating the miR-100/ATG5 axis and PI3K/AKT/mTOR signaling [7].

In this study, we found that the proliferation of OS cells was inhibited after knockdown of HAGLROS. The metastasis of OS cells was also reduced after knockdown of HAGLROS. Furthermore, it was found that lncRNA HAGLROS could interact with miR-152 to affect the expression of downstream gene ROCK1.

2. Material and Methods

2.1. Patients and Tissue Samples. 10 paired OS tissues and match normal bone tissues were collected from the OS patients with surgical treatment from Minhang Hospital, Fudan University, from January 2014 to May 2017. All tissues were stored in liquid nitrogen and used for extraction of RNA. Informed consent was obtained from all patients. This project was approved by the Institute Research Ethics Committee at Minhang Hospital, Fudan University (Shanghai, China).

2.2. Cell Lines. MG-63, hFOB 1.19, SW1353, and U2OS were obtained from ATCC and cultured in RPMI-1640 (BI, Israel) containing 10% FBS (BI, Israel) with 5% CO₂ at 37°C.

2.3. Small Interfering RNA Synthesis and Transfection. Small interfering RNAs (siRNAs) specifically targeting HAGLROS were purchased from the GenePharma Company (Shanghai, China). The siRNA sequences were 5'-CCUAUUUACUG GCAGGAGUTT-3' for HAGLROS; 5'-UUCUCCGAACG UGUCACGUTT-3' for NC. Transfections were performed using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions.

2.4. Cell Counting Kit-8 Proliferation Assay. In order to detect the cell proliferation capacity, about 2000 SW1353 or U2OS cells with different treatments were seeded into a 96-well plate. After a certain time, 10 μ L CCK-8 (Dojindo Chemical Laboratory, Kumamoto, Japan) was added to each well and then incubated at 37°C for 2 hours, and the optical density (OD) of each well at 450 nm was detected by using a MB-580 machine (HEALES, Shenzhen, China) [8].

2.5. Transwell Assays. Briefly, the SW1353 and U2OS cells were starved for 12 hours in RPMI-1640 medium without FBS. Cells at 1×10^6 cells/well were added into the upper chamber of Transwell with (for invasion assay) or without (for migration assay) diluted Matrigel treatment (BD Biosciences, San Jose, CA, USA) and then in the bottom chamber were added complete medium. After incubation for 2 days, the migrating cells were treated with methanol and stained with 10 μ g/mL DAPI (Solarbio, Beijing, China). The migration or invasion cells were counted using an inverted fluorescence microscope.

2.6. Dual-Luciferase Reporter Assay. The pMIR-REPORT-HAGLROS-WT/MUT and pMIR-REPORT-ROCK1-3' UTR-WT/MUT (Sangon Biotech, Shanghai, China) were constructed. Furthermore, the Dual-Luciferase Reporter Assay System (E1910, Promega, WI, USA) was used to detect the luciferase activities of these reporters according to the manufacturer's protocol [9].

2.7. qRT-PCR Assay. We extracted RNA by using TRIzol reagent, as described by the manufacturer's protocol (Sangon Biotech, Shanghai, China). cDNAs were synthesized with the Primer-Script™ one-step RT-PCR kit (Takara, Otsu, Shiga, Japan). The PCR amplification was performed on an Applied Biosystems 7900HT (Biosystems, Foster City, CA, USA). The relative quantification values of lncRNA were determined by the $2^{-\Delta\Delta C_t}$ method with GAPDH as an internal reference.

2.8. Statistical Analysis. Statistical analysis was conducted with the GraphPad Prism 6.0 software (La Jolla, CA, USA). To analyze the relationship between gene expression and the clinical characteristics of the tumors, the chi-squared test, *t*-test, Fisher's exact test, or Mann-Whitney's *U*-test were used as appropriate. $p < 0.05$ was regarded as statistically significant.

3. Results

3.1. HAGLROS Was Upregulated in OS Samples. lncRNA HAGLROS was upregulated in OS. To investigate the level of HAGLROS expression in human OS, RT-PCR assays were performed in 10 OS and 10 normal samples. As displayed in Figure 1, we found the RNA levels of HAGLROS distinctly increased in OS, in comparison with normal samples (Figure 1(a)) ($p < 0.05$). Furthermore, by using qRT-PCR assay, we revealed that HAGLROS was upregulated in MG-63, SW1353, and U2OS cells compared to hFOB 1.19 cells (Figure 1(b)) ($p < 0.05$). In order to further confirm these findings, we analyzed HAGLROS expression in TCGA and found that HAGLROS was upregulated in OS samples compared to normal tissues (Figure 1(c)). Our findings indicated that HAGLROS might contribute to the development of OS.

3.2. Knockdown of HAGLROS Inhibited OS Cell Proliferation. In order to further explore the molecular roles of HAGLROS in OS, we conducted functional validation in OS. We observed the HAGLROS expression was reduced in SW1353 and U2OS cells treated with si-HAGLROS which was designed to knock down HAGLROS (Figures 2(a) and 2(b)) ($p < 0.01$). CCK-8 assay showed that the proliferation rate of HAGLROS knockdown group was significantly decreased compared with the negative control group in both SW1353 and U2OS cells (Figures 2(c) and 2(d)) ($p < 0.05$).

3.3. Knockdown of HAGLROS Inhibited OS Cell Migration and Invasion. In order to further investigate the effects of HAGLROS on the metastasis capacity of SW1353 and U2OS cells, the Transwell assay showed both the migration and invasion abilities were remarkably suppressed after knockdown HAGLROS in SW1353 and U2OS cells compared with negative controls (Figures 3(a) and 3(b)) ($p < 0.05$).

3.4. HAGLROS Served as a Sponge of miR-152 to Promote ROCK1. Next, we explored the location of this lncRNA in OS to understand its molecular mechanisms. As presented in Figure 4, we observed HAGLROS was mainly localized to the cytoplasm (Figure 4(a)), suggesting that HAGLROS may play its functions as a miRNA sponge in cytoplasm. In

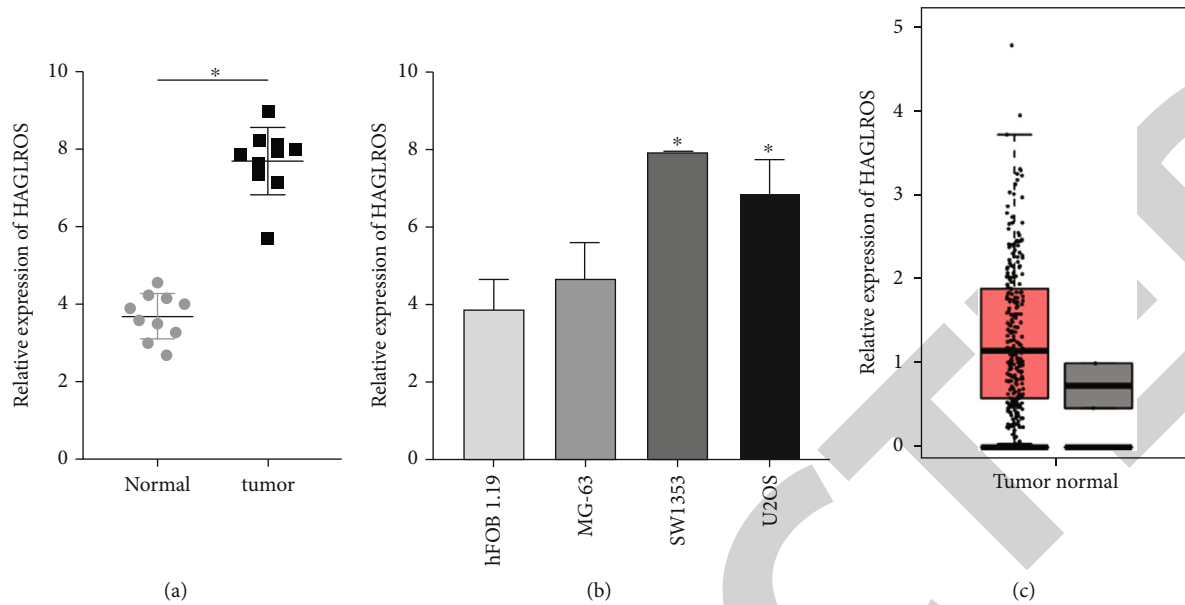


FIGURE 1: HAGLROS was upregulated in OS samples. HAGLROS expression in paired OS tissues and normal samples collected from 10 patients with OS. $*p < 0.05$. (a) HAGLROS expression in normal cells (hFOB 1.19) and human OS cells (MG-63, SW1353, and U2OS) was detected by qRT-PCR. $*p < 0.05$. (b) HAGLROS was upregulated in OS samples compared to normal tissues. $*p < 0.05$.

silico analysis indicated that HAGLROS could interact with miR-152 to upregulate ROCK1. Further validation showed miR-152 was downregulated in MG-63, SW1353 and U2OS cells compared to hFOB 1.19 cells (Figure 4(b)) ($p < 0.05$). siRNA-mediated HAGLROS knockdown contributed to miR-152 induction in SW1353 and U2OS cells (Figure 4(c)) ($p < 0.001$). The transfection efficiency after overexpression or knockdown of miR-152 was shown in Figures 4(d) and 4(e). Furthermore, we constructed HAGLROS luciferase vectors with wild-type or mutant miR-152 binding site to validate the direct binding between HAGLROS and miR-152. As expected, the results showed HEK 293T cells transfected with miR-152 and wild-type, but not the mutant HAGLROS constructions, had a significantly lower luciferase activity (Figure 4(f)) ($p < 0.05$).

Furthermore, we aimed to validate the effects of HAGLROS/miR-152 axis on ROCK1. Interestingly, ROCK1 was upregulated in OS cell lines (Figure 5(a)) ($p < 0.05$). Dual-luciferase reporter assay showed HEK 293T cells transfected with miR-152 and wild-type ROCK1 had significantly reduced luciferase activity, but not the mutant ROCK1 (Figure 5(b)) ($p < 0.05$). From the rescue assay, miR-152 inhibitor counteracted HAGLROS knockdown-mediated downregulation on ROCK1 expression in both SW1353 and U2OS cells (Figures 5(c) and 5(d)) ($p < 0.05$). The above experimental results could confirm that HAGLROS served as a sponge of miR-152 to promote ROCK1.

4. Discussion

With the rapid development of sequencing technology and increasing scientific investment, more and more lncRNAs have been discovered and studied. The constant discovered lncRNA sites have extended a vast space for the study of life

science and also widened the way for scientific and medical research. Studies have found that lncRNAs were involved in the organization, transcription, and posttranscriptional regulation of chromatin in cells [10, 11], which updates our understanding of the regulation of traditional eukaryotes' genomes and makes us aware of the complexity of eukaryotic life processes. Interestingly, with the in-depth research on lncRNA, its roles in the development of cancer have been increasingly discovered and understood. From the perspective of function, more and more lncRNAs have been found to be tumor suppressor or protooncogenes at the molecular level [12, 13], which greatly enriches our understanding of the complexity of cancer occurrence. The occurrence and development of cancer are the results of a series of cellular regulatory mechanisms that do not play their normal roles effectively, while lncRNA is deeply involved in these processes. To some extent, the stem cell characteristics of cancer cells are similar to those of adult stem cells. For example, lncRNA ANRIL can interact with polycomb repressive complexes 1 (PRC1) and polycomb repressive complexes 2 (PRC2) to inhibit the expression of INK4b-ARF-INK4a, resist the aging mechanism, and enhance the stem cell characteristics of cancer cells [14]. Tumor growth is a process in which cancer cells get rid of the mechanism of inhibition and clearance and continue to proliferate and grow and break through the barriers between tissues. Multiple lncRNAs were found to be abnormally expressed in tumor and normal tissues. For example, overexpression of lncRNA H19 was found to promote the development of gastric cancer and metastasis of cancer cells [15]. The high expression of HAGLROS in colorectal cancer is inversely related to the survival time of CRC patients. Downregulation of HAGLROS may induce apoptosis of CRC cells and inhibit autophagy by regulating the miR-100/ATG5 axis [16]. Similarly, the high expression

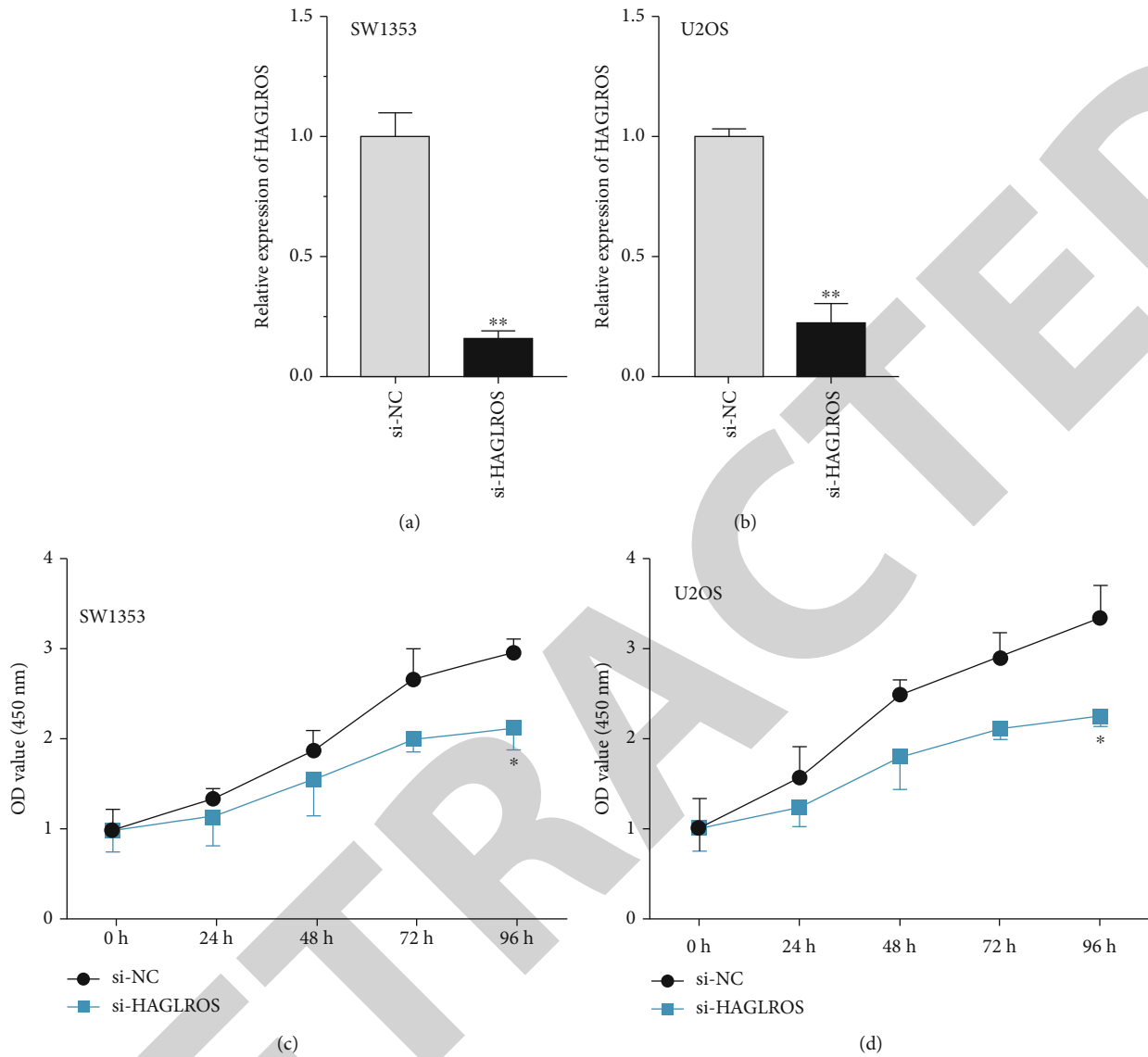


FIGURE 2: Knockdown of HAGLROS inhibited OS cell proliferation. (a, b) qRT-PCR analysis for HAGLROS expression in SW1353 and U2OS cell lines with si-NC or si-HAGLROS. $**p < 0.01$. (c, d) CCK-8 assay showed SW1353 and U2OS cells proliferation ability in different transfected groups. $*p < 0.05$.

of lncRNA PVT1 in NSCLC could promote the progression of NSCLC [17]. Here, the results indicated lncRNA HAGLROS could promote the proliferation of OS cells (SW1353 and U2OS) and enhance the invasion and migration of OS cells.

There are a large number of noncoding RNAs in human cells. These RNAs can be abundant in cells because of not the wrong transcription, but their key regulatory roles in cell metabolism. Among them, some single-stranded endogenous RNAs between 19 and 25 nucleotides in length are called microRNAs (miRNAs) [18]. These single-stranded miRNAs bind to target mRNAs to negatively regulate the expression of these genes [19]. With the deepening of research, the role of miRNAs in cell proliferation, differentiation, metastasis, apoptosis, and other metabolic activities has been revealed more. Of

course, more and more studies implied the functional importance of miRNA in cancers. For example, miR-15 and miR-16 which were discovered earlier negatively regulated B-cell lymphoma 2 (BCL2) at the posttranscriptional level, and their expression was negatively correlated with BCL2 in chronic lymphocytic leukemia (CLL). These microRNAs can be induced by the inhibition of BCL2 CLL cells apoptosis, thus giving play to the role of tumor suppressor genes [20]. Conversely, miRNAs can also promote tumor development and progression. miR-21 expression was upregulated in human glioblastoma tumors and other cancer samples as well as in the established cancer cell lines compared with normal tissues [21, 22]. Downregulation of miR-21 expression leads to increased apoptosis of tumor cells, suggesting that miR-21 may affect the gradual development and transformation of tumors into

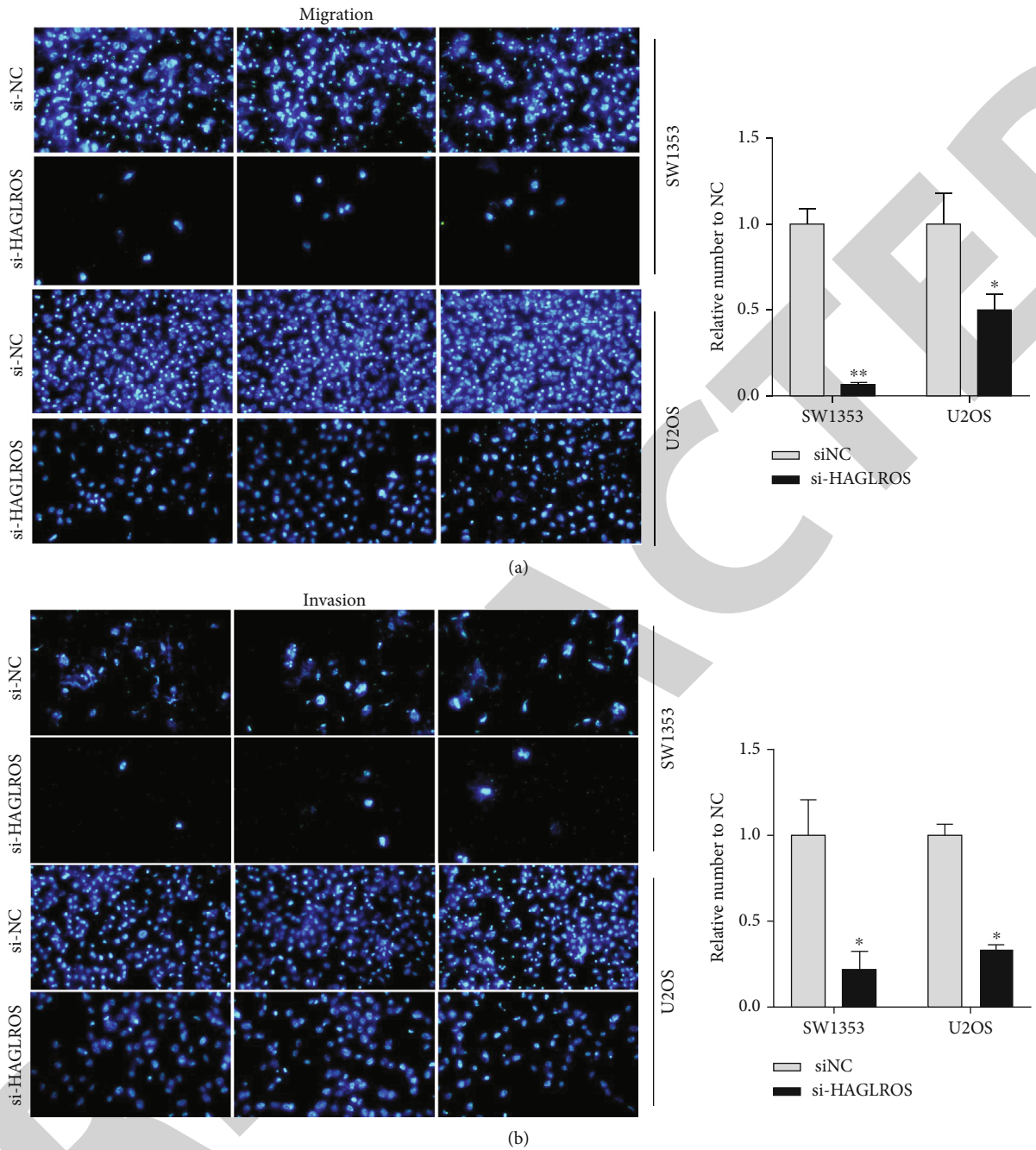


FIGURE 3: Knockdown of HAGLROS inhibited OS cell migration and invasion. (a, b) The effect of HAGLROS knockdown on the migration and invasion of SW1353 and U2OS cells was estimated through Transwell assay. * $p < 0.05$, ** $p < 0.01$.

malignant tumors by blocking the expression of relevant apoptotic genes [23]. While this study involved in miR-152 in the department of endometrial cancer cells can interfere with cell processes, which affect the cell cycle arrest and inhibit the growth of tumor cells in vitro [24].

The regulatory mechanism of lncRNA in cells also affects the regulatory effect of miRNA, and the antagonistic or synergistic effect between the two together affects the expression of downstream genes [5]. lncRNA HAGLROS can sponge miR-152 in the experiments of lung cancer cell lines in vitro, thereby reducing the anticancer regulation ability

and promoting the proliferation and metastasis of lung cancer cells [25]. These reports are consistent with our findings. In this study, lncRNA HAGLROS sponged miR-152 to reduce its downstream genes ROCK1 expression, leading to an increase in the ability of cell proliferation and ability.

miR-148/152 family is related to the regulation of metabolic activities and cell growth. The expressions of miR-148 and miR-152 are often coexpressed in cancer cells [26]. miR-152 was reported in human glioblastoma stem cells as a tumor suppressor [27]. ROCK1 protein is Rho relative coiled-coil protein kinase, one of the isomers with adjusting

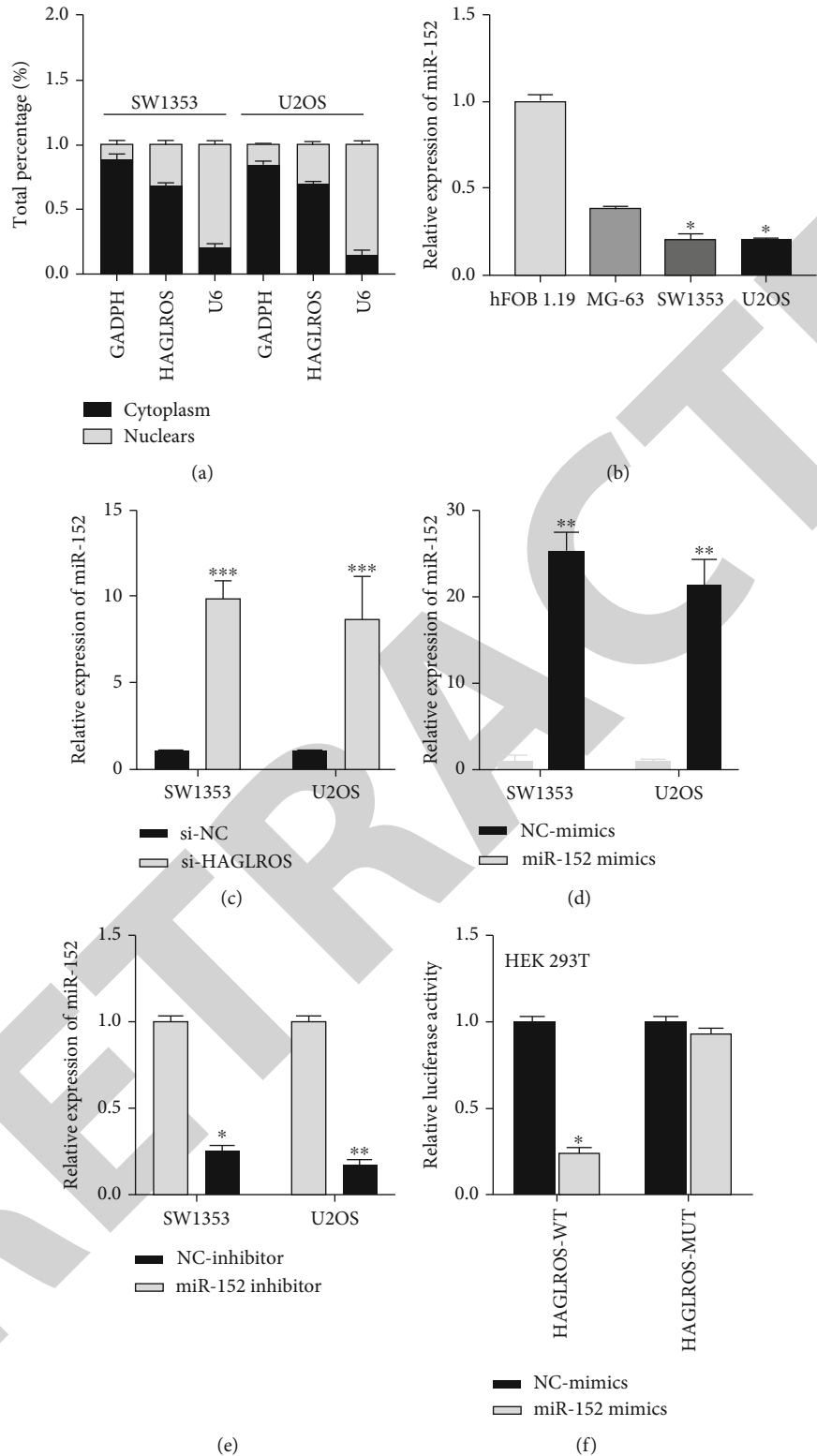


FIGURE 4: HAGLROS served as a sponge of miR-152. qRT-PCR analysis of the location of HAGLROS in SW1353 and U2OS cells. The relative expression of miR-152 in OS cells was less than half that in normal cells. $*p < 0.05$. OS cells with si-HAGLROS had higher expression of miR-152 than those with si-NC. $***p < 0.001$. (d, e) The transfection efficiency of miR-152 mimics (or inhibitor) was estimated by qRT-PCR. $*p < 0.05$, $**p < 0.01$. (f) The relationship between HAGLROS and miR-152 was confirmed with the use of luciferase reporter assay. $*p < 0.05$.

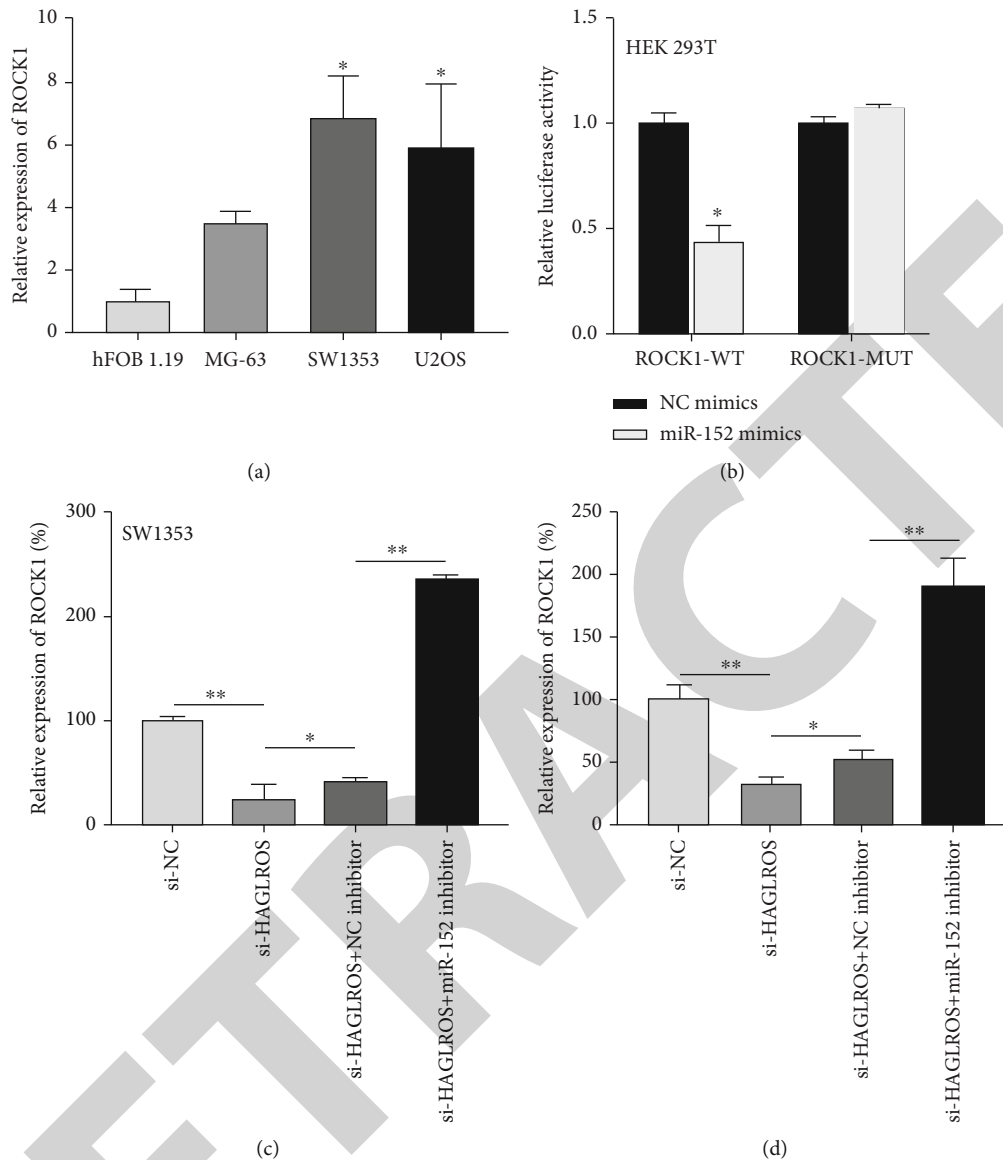


FIGURE 5: ROCK1 a target gene of miR-152 was modulated by HAGLROS. (a) Relative expression of ROCK1 in OS cell lines and normal cell lines was tested using qRT-PCR. * $p < 0.05$. (b) Luciferase reporter assay revealed the molecular incorporation of ROCK1 and miR-152. * $p < 0.05$. (c, d) qRT-PCR analysis was utilized to estimate ROCK1 expression in si-NC, si-HAGLROS, si-HAGLROS + NC-inhibitor, and si-HAGLROS + miR-152 inhibitor groups. * $p < 0.05$, ** $p < 0.01$.

the cytoskeleton reorganization and the action such as cell adhesion, and ROCK1 was reported to have the function to regulate cell movement and cell migration [28]. It is reported that miR-148a can directly target the ROCK1, thereby suppressing ROCK1 transcription and translation level and suppress the invasion and metastasis of tumor cells [29]. Our results showed that miR-152 could also suppress ROCK1 expression, thus inhibiting the proliferation and metastasis of OS cells.

5. Conclusions

In conclusion, the result showed that silencing the lncRNA HAGLROS by si-lncRNA HAGLROS significantly suppressed the proliferation and metastasis of OS cells. Mechan-

ical studies showed that lncRNA HAGLROS could sponge miR-152, thereby suppressing the inhibition of the downstream ROCK1 gene and promoting the proliferation, invasion, and metastasis of OS cells. Our study provided new potential biomarkers for OS.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethical Approval

The experimental protocol was established, according to the ethical guidelines of the Helsinki Declaration, and was

approved by the Institute Research Ethics Committee at Minhang Hospital, Fudan University.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

Kaifeng Zhou and Jun Xu were from the development of methodology; sample collection was done by Jun Xu; analysis and interpretation of data were done by Xiaofan Yin; and writing, review, and/or revision of the manuscript were done by Jiangni Xia and Xiaofan Yin. Kaifeng Zhou and Jun Xu contributed equally to this work.

Acknowledgments



This work is supported by the National Natural Science Foundation of China (Grant No. 81772433) and Natural Science Research Projects in Minhang District (Grant No. 2019MHZ086).

References

- [1] L. Mirabello, R. J. Troisi, and S. A. Savage, "Osteosarcoma incidence and survival rates from 1973 to 2004," *Cancer*, vol. 115, no. 7, pp. 1531–1543, 2009.
- [2] B. Fuchs and D. J. Pritchard, "Etiology of Osteosarcoma," *Clinical Orthopaedics and Related Research*, vol. 397, pp. 40–52, 2002.
- [3] R. Achawanantakun, J. Chen, Y. Sun, and Y. Zhang, "LncRNA-ID: long non-coding RNA IDentification using balanced random forests," *Bioinformatics*, vol. 31, no. 24, pp. 3897–3905, 2015.
- [4] L. W. Harries, "Long non-coding RNAs and human disease," *Biochemical Society Transactions*, vol. 40, no. 4, pp. 902–906, 2012.
- [5] J. Liz and M. Esteller, "lncRNAs and microRNAs with a role in cancer development," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1859, no. 1, pp. 169–176, 2016.
- [6] S. Jalali, D. Bhartiya, M. K. Lalwani, S. Sivasubbu, and V. Scaria, "Systematic transcriptome wide analysis of lncRNA-miRNA interactions," *Plos One*, vol. 8, no. 2, p. e53823, 2013.
- [7] G. Mu, Q. Liu, S. Wu, Y. Xia, and Q. Fang, "Long noncoding RNA HAGLROS promotes the process of mantle cell lymphoma by regulating miR-100/ATG5 axis and involving in PI3K/AKT/mTOR signal," *Artificial cells, nanomedicine, and biotechnology*, vol. 47, no. 1, pp. 3649–3656, 2019.
- [8] S. Chen, Z. Jin, L. Dai et al., "Aloperine induces apoptosis and inhibits invasion in MG-63 and U2OS human osteosarcoma cells," *Biomedicine & Pharmacotherapy*, vol. 97, pp. 45–52, 2018.
- [9] F. Alcaraz-Pérez, V. Mulero, and M. L. Cayuela, "Application of the dual-luciferase reporter assay to the analysis of promoter activity in Zebrafish embryos," *BMC Biotechnology*, vol. 8, no. 1, p. 81, 2008.
- [10] R.-Z. He, D.-X. Luo, and Y.-Y. Mo, "Emerging roles of lncRNAs in the post-transcriptional regulation in cancer," *Genes & Diseases*, vol. 6, no. 1, pp. 6–15, 2019.
- [11] T. C. Roberts, K. V. Morris, and M. S. Weinberg, "Perspectives on the mechanism of transcriptional regulation by long non-coding RNAs," *Epigenetics*, vol. 9, no. 1, pp. 13–20, 2014.
- [12] L. Qian, J. Huang, N. Zhou et al., "LncRNA loc285194 is a p53-regulated tumor suppressor," *Nucleic Acids Research*, vol. 9, 2013.
- [13] Y. Yan, Q. Fan, L. Wang, Y. Zhou, J. Li, and K. Zhou, "LncRNA Snhg1, a non-degradable sponge for miR-338, promotes expression of proto-oncogene CST3 in primary esophageal cancer cells," *Oncotarget*, vol. 8, no. 22, pp. 35750–35760, 2017.
- [14] F. Aguilo, S. D. Cecilia, and M. J. Walsh, "Long non-coding RNA ANRIL and polycomb in human cancers and cardiovascular disease," *Current Topics in Microbiology and Immunology*, vol. 394, pp. 29–39, 2015.
- [15] H. Li, B. Yu, J. Li, L. Su, and B. Liu, "Overexpression of lncRNA H19 enhances carcinogenesis and metastasis of gastric cancer," *Oncotarget*, vol. 5, no. 8, pp. 2318–2329, 2014.
- [16] Y. Zheng, K. Tan, and H. Huang, "Long noncoding RNA HAGLROS regulates apoptosis and autophagy in colorectal cancer cells via sponging miR-100 to target ATG5 expression," *Journal of Cellular Biochemistry*, vol. 120, pp. 3922–3933, 2018.
- [17] Y. R. Yang, S. Z. Zang, C. L. Zhong, Y. X. Li, S. S. Zhao, and X. J. Feng, "Increased expression of the lncRNA PVT1 promotes tumorigenesis in non-small cell lung cancer," *International Journal of Clinical and Experimental Pathology*, vol. 7, no. 10, pp. 6929–6935, 2014.
- [18] V. N. Kim, "MicroRNA biogenesis: coordinated cropping and dicing," *Nature reviews Molecular cell biology*, vol. 6, no. 5, pp. 376–385, 2005.
- [19] R. Perales, D. M. King, C. Aguirre-Chen, C. M. Hammell, and G. Ruvkun, "LIN-42, the *Caenorhabditis elegans* PERIOD homolog, negatively regulates microRNA transcription," *PLoS Genetics*, vol. 10, no. 7, article e1004486, 2014.
- [20] Y. Pekarsky and C. M. Croce, "Role of miR-15/16 in CLL," *Cell Death & Differentiation*, vol. 22, no. 1, pp. 6–11, 2015.
- [21] J. A. Chan, A. M. Krichevsky, and K. S. Kosik, "MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells," *Cancer Research*, vol. 65, no. 14, pp. 6029–6033, 2005.
- [22] C. Gong, Y. Yao, Y. Wang et al., "Up-regulation of miR-21 mediates resistance to trastuzumab therapy for breast cancer," *Journal of Biological Chemistry*, vol. 286, no. 21, pp. 19127–19137, 2011.
- [23] X. Zhou, J. Zhang, Q. Jia et al., "Reduction of miR-21 induces glioma cell apoptosis via activating caspase 9 and 3," *Oncology Reports*, vol. 24, 2010.
- [24] T. Tsuruta, K.-i. Kozaki, A. Uesugi et al., "miR-152 is a tumor suppressor microRNA that is silenced by DNA hypermethylation in endometrial cancer," *Cancer Research*, vol. 71, no. 20, pp. 6450–6462, 2011.
- [25] W. L. Wang, D. J. Yu, and M. Zhong, "LncRNA HAGLROS accelerates the progression of lung carcinoma via sponging microRNA-152," *European Review for Medical and Pharmacological Sciences*, vol. 23, pp. 6531–6538, 2019.
- [26] X. Liu, J. Li, F. Qin, and S. Dai, "miR-152 as a tumor suppressor microRNA: target recognition and regulation in cancer," *Oncology Letters*, vol. 11, no. 6, pp. 3911–3916, 2016.

Research Article

Comprehensive Analysis of Differentially Expressed circRNAs Reveals a Colorectal Cancer-Related ceRNA Network

Feng Que,^{1,2} Hua Wang,^{1,2} Yi Luo ,^{2,3} Li Cui,^{2,4} Lanfu Wei,^{2,5} Zhaohong Xi,^{2,5} Qiu Lin,^{1,2} Yongsheng Ge,^{1,2} and Wei Wang ^{1,2}

¹Department of Colorectal Surgery, Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, 210028 Nanjing, Jiangsu, China

²Jiangsu Province Academy of Traditional Chinese Medicine, 210028 Nanjing, Jiangsu, China

³Department of Oncology, Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, 210028 Nanjing, Jiangsu, China

⁴Key Laboratory of New Drug Delivery System of Chinese Materia Medica, Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, 210028 Nanjing, Jiangsu, China

⁵Department of Gastroenterology, Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, 210028 Nanjing, Jiangsu, China

Correspondence should be addressed to Wei Wang; wangweinj88@163.com

Received 23 June 2020; Accepted 25 July 2020; Published 1 September 2020

Guest Editor: Tao Huang

Copyright © 2020 Feng Que 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 morbidity and mortality of colorectal cancer (CRC) remained to be very high worldwide. Recently, circRNAs had been revealed to have a crucial role in cancer prognosis and progression. Numerous researches have shown that RNA sequencing technology and in silico method were widely used to identify pathogenic mechanisms and uncover promising targets for diagnosis and therapy. In this study, these methods were analyzed to obtain differentially expressed circRNAs (DECs). We identified upregulated 316 circRNAs and reduced 76 circRNAs in CRC samples, in comparison with those in normal tissues. In addition, a competitive endogenous network of circRNA-miRNA-mRNA was established to predict the mechanisms of circRNAs. Bioinformatics analysis revealed that these circRNAs participated in metabolism regulation and cell cycle progression. Of note, we observed the hub genes and miRNAs in this ceRNA network were associated with the survival time in CRC. We think this study could provide potential prognostic biomarkers and targets for CRC.

1. Introduction

The morbidity and mortality of colorectal cancer (CRC) remained to be very high worldwide [1, 2]. Over the past 10 years, great progress has been made in CRC prevention, diagnosis, and treatment [3]. Nevertheless, the prognosis of CRC is still poor [4]. It is therefore of importance to understand the mechanisms affecting CRC pathogenesis.

circRNAs are different from linear RNAs, which have a 5' cap and 3' tail structure [5]. circRNAs are characterized by forming a covalent closed-loop structure in the absence of 5'-3' polarity or polyadenylated tails [6]. More and more

researches have shown that circRNAs are a sort of affluent, miscellaneous, and conservative RNA molecules [7, 8]. Of note, up-and-coming evidences suggested that circRNAs modulated various biological processes, such as cell viability, differentiation, apoptosis, and angiogenesis [9, 10]. Presently, numerous researches have revealed that circRNAs are dysregulated in various carcinomas, implying that circRNAs displayed an important role in the occurrence and development of human carcinomas. With the development of high throughput RNA sequencing and bioinformatics methods, the advantage of circRNAs emerged gradually. Currently, some researches had shown that circRNAs modulated

alternative splicing and gene expression level through sponging microRNAs (miRNAs) [11, 12]. At the same time, disorders of circRNAs took part in carcinogenesis and cancer development of CRC [13, 14], liver cancer [15], and gastric cancer [16–19]. These findings indicated that circRNAs, as a new kind of endogenous noncoding RNA, have been a new hotspot in the cancer research and probably would exhibit important functions in the development of tumor.

miRNAs are a member of noncoding RNAs with approximate 22 nucleotides in length [20, 21]. miRNAs exhibit inhibitory function on target gene expression via miRNA response elements (MREs) in the 3'-UTR of transcripts [22]. Several researches had revealed that circRNAs sponged miRNAs by binding to corresponding MREs in many diseases, which has also been identified in the occurrence and development of cancer as previously described [23, 24]. Numerous evidences showed that circRNAs, exhibiting as miRNAs spongers, could modulate CRC growth, progress, and metastasis [25, 26]. For example, ciRS-7 sponged miR-7 and suppressed target gene expression in many tumors, including CRC [26]. Additionally, circHIPK3 probably displayed as miR-1207/miR-637/miR-7 sponge and retarded its antitumor function, thereby promoting CRC cell viability [27–29]. hsa_circ_0091074 modulated TAZ level by microRNA-1297 in breast cancer cell [30]. CircSMC3 modulated tumor genesis of gastric cancer via targeting miR-4720-3p/TJP1 axis [31]. In colon cancer, hsa_circ_0055625 deriving from the expression profile of circRNAs promoted colon cancer cell viability via sponging miR-106b-5p [32]. hsa_circ_0007843 was a sponger of miR-518c-5p and modulated colon cancer cells migration and invasion [33]. Nevertheless, further researches are still needed to investigate the probable mechanisms of tumorigenesis, might aid in the diagnosis and treatment of CRC, and may be of help for CRC diagnosis and treatment.

In our study, we systematically assessed circRNA expression in 3 paired CRC and normal tissues. We discovered that the expression profile of circRNAs was dramatically distinct between CRC and normal tissues. Meanwhile, we constructed a circRNA-miRNA regulatory network in CRC. Our data suggested that circRNAs were linked to CRC occurrence and development, thus supplying more long-range perspective indicators and new biomarkers for CRC.

2. Materials and Methods

2.1. Samples. This study got approval from The Ethics Review Board of Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine. All CRC patients signed written informed consent. A total of 3 paired CRC and normal were used in this study.

2.2. Construction and Sequencing of RNA Library. Whole RNA was harvested by TRIzol reagent (Invitrogen, CA, USA) as manually described and then subjected to detect quality and concentration of purified RNA by Bioanalyzer 2100 (Agilent, CA, USA) and Qubit 2.0, respectively.

The library of circRNA was established referring to the instruction of the NEBNext Ultra Small RNA Sample Library Preparation Kit of Illumina. RNase R was applied to digest linear RNA, and rRNA probe was used to remove rRNA accordingly. The first strand is synthesized using stochastic hexapolymers and templates of circRNA. Subsequently, the second strand of cDNA was completed. AMPure XP beads were applied to make purification of lncRNAs and circRNAs. T4 DNA polymerase and Klenow DNA polymerase could attain the goal of the blunt end of DNA. Poly(A) tail was added into 3' end of DNA and then sequenced. AMPure XP beads were applied to choose the size of fragments. USER enzyme was used to degrade the second strand comprising cDNA. The library of ncRNA and circRNA was obtained by PCR amplification. Finally, the libraries were subjected to paired-end sequencing with pair end 150 bp reading length on an Illumina HiSeq sequencer (Illumina, San Diego, CA, USA) according to a previous report [34].

Differentially expressed circRNAs (DEC) between cancer and normal tissue were identified with defined threshold values > 1.0 ($|\log_{2}FC| > 1$) and p values < 0.01 (p value < 0.01).

2.3. Construction of Competing Endogenous (ceRNA) Network. The CSCD database (<http://gb.whu.edu.cn/CSCD>) was applied to predict MREs, RNA-binding proteins (RBPs), and open reading frames (ORFs) [35]. The links between miRNAs and the circRNAs or mRNAs were forecasted by a ceRNA network. The interplay between mRNA and miRNA, of which sequences and interpretation were derived from miRBase [36], was forecasted by miRTarbase [37], TargetScan [38], and miRDB [39]. The ceRNA network comprising target genes of miRNA and circRNAs was established utilizing Cytoscape software (version 3.6.1) [40].

2.4. Bioinformatics Analysis. GO (<http://www.geneontology.org>) was applied to identify and annotate the sequences of homologous genes and proteins in a variety of organisms, which can help us clarify the particular role of specified genes.

The KEGG (<http://www.genome.jp/kegg/>) [41] database was used to predict the potential pathways regulated by candidate mRNAs or circRNAs. Adjusted $p < 0.05$ and $q < 0.05$ represented obvious annotations of enriched function of circRNAs.

2.5. Statistical Analysis. GraphPad Prism software version 6 (GraphPad Software, Inc.) was applied for data processing. The represented data obtained from at least three independent experiments in triplicates was shown as the mean \pm SD. Two-tailed paired Student's t -test was applied to compare two groups. The overall survival curve was generated using the Kaplan–Meier method and the log-rank test. Significant difference was shown as $p < 0.05$.

3. Results

3.1. Identification of DECs in CRC. RNA sequencing was applied to determine the expression profile of circRNAs in 3 CRC and normal tissues. Hierarchical clustering, as one of most extensive clustering analysis toolsets, was applied to

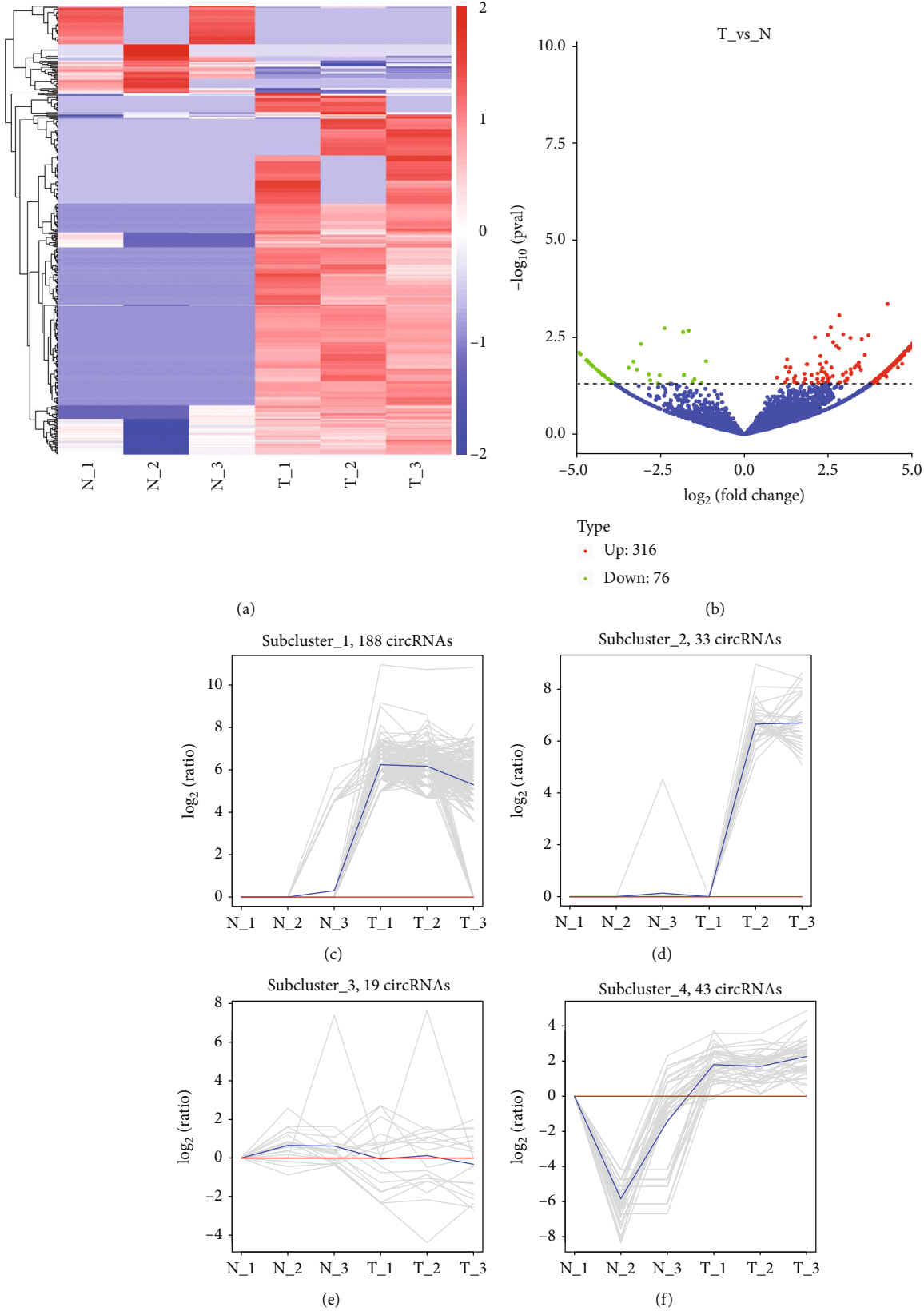


FIGURE 1: Continued.

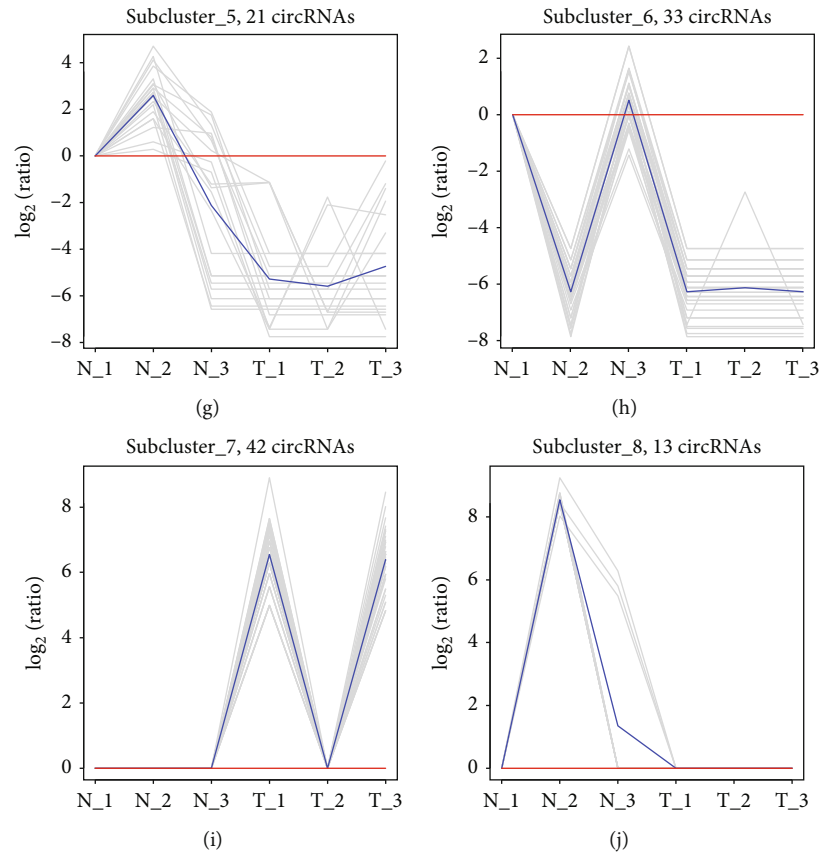


FIGURE 1: Identification of DECs in CRC. (a) Hierarchical clustering identified differently expressed circRNAs in CRC. (b) The volcano plot showed differently expressed circRNAs in CRC. (c) Subcluster 1 included 188 circRNAs. (d) Subcluster 2 included 33 circRNAs. (e) Subcluster 3 included 19 circRNAs. (f) Subcluster 4 included 43 circRNAs. (g) Subcluster 5 included 21 circRNA. (h) Subcluster 6 included 33 circRNAs. (i) Subcluster 7 included 42 circRNAs. (j) Subcluster 8 included 13 circRNAs.

assess gene expression data (Figure 1(a)). The difference between normal and CRC samples was visualized by the volcano plot. The vertical lines indicated approximate 2.0-fold, and the horizontal lines suggested a p value of 0.05. Red dots included in the indicated figures represented the differentially expressed circRNA, meaning significant difference (Figure 1(b)). In our results, compared with normal tissues, 316 circRNA expression level was increased and 76 circRNA expression level was reduced.

Of note, some of these circRNAs had been demonstrated to be related to cancer progression. For example, *hsa_circ_0084663* was validated to be related to sorafenib-resistant liver cancer [42]. Very interestingly, we found that *hsa_circ_0061776* and *hsa_circ_0006528* were also upregulated in colon samples by analyzing the supplementary table in an independent report [25]. Meanwhile, we found these differentially expressed circRNAs could be divided into 8 clusters. As shown in Figure 1, subcluster 1 included 188 circRNAs (Figure 1(c)), subcluster 2 included 33 circRNAs (Figure 1(d)), subcluster 3 included 19 circRNAs (Figure 1(e)), subcluster 4 included 43 circRNAs (Figure 1(f)), subcluster 5 included 21 circRNAs (Figure 1(g)), subcluster 6 included 33 circRNAs (Figure 1(h)), subcluster 7 included 42 circRNAs (Figure 1(i)), and subcluster 8 included 13 circRNAs (Figure 1(j)).

3.2. The Function Prediction of the Host Genes of circRNAs. Previous studies had showed that circRNAs may play its roles in human diseases via their host genes. Thus, the potential functions of the host genes were assessed by GO and KEGG. The host genes of DECs are enriched in GO terms including primarily contained organization of cellular component or biogenesis, cell cycle, miRNAs (mainly for gene silencing), mitotic cell cycle, cell cycle process, and posttranscriptional gene silencing by RNA. The host genes of DECs are enriched in CC terms, including cytosol, organelle, cytoplasm, membrane-less organelles, intracellular membrane-less organelles, and nucleus. The differentially expressed genes are enriched in molecular function term and protein binding (Figure 2(a)).

Through KEGG analysis, the function of the host genes of DECs participated in modulating signaling pathways, including adherent junction, VEGF signaling pathway, thyroid cancer, endometrial cancer, serotonergic synapse, leukocyte transendothelial migration, bacterial invasion of epithelial cells, non-small cell lung cancer, and long-term depression (Figure 2(b)).

3.3. Construction of a circRNA-miRNA-mRNA Network. Numerous evidences showed DECs could competitively sponge miRNAs and then suppressed downstream genes of

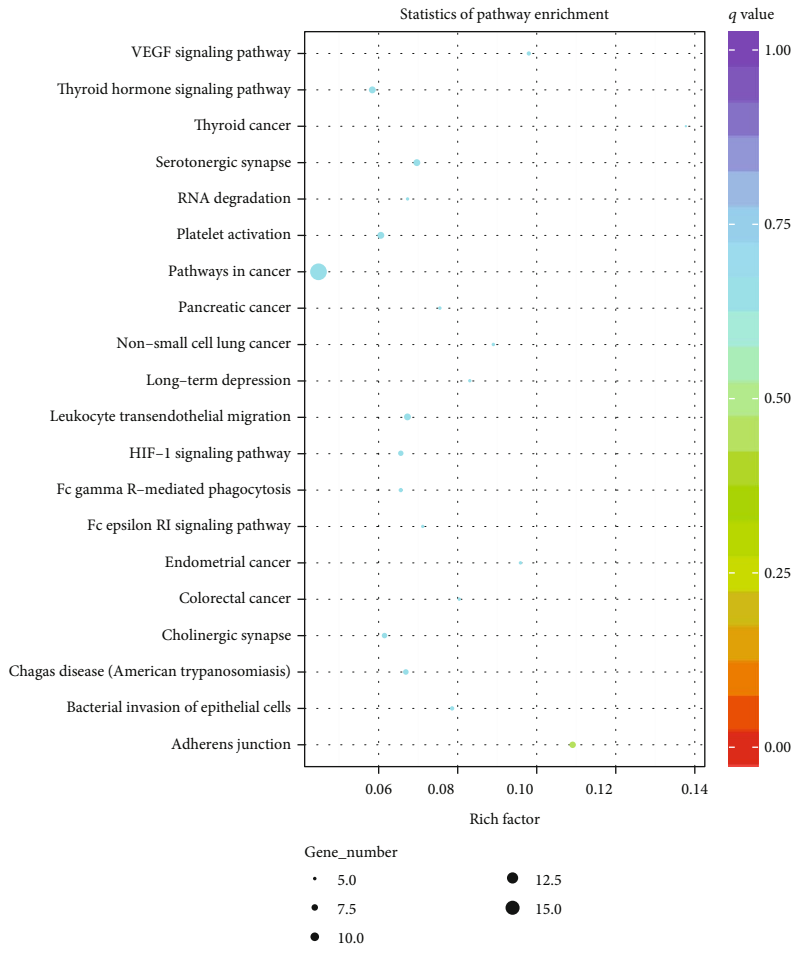
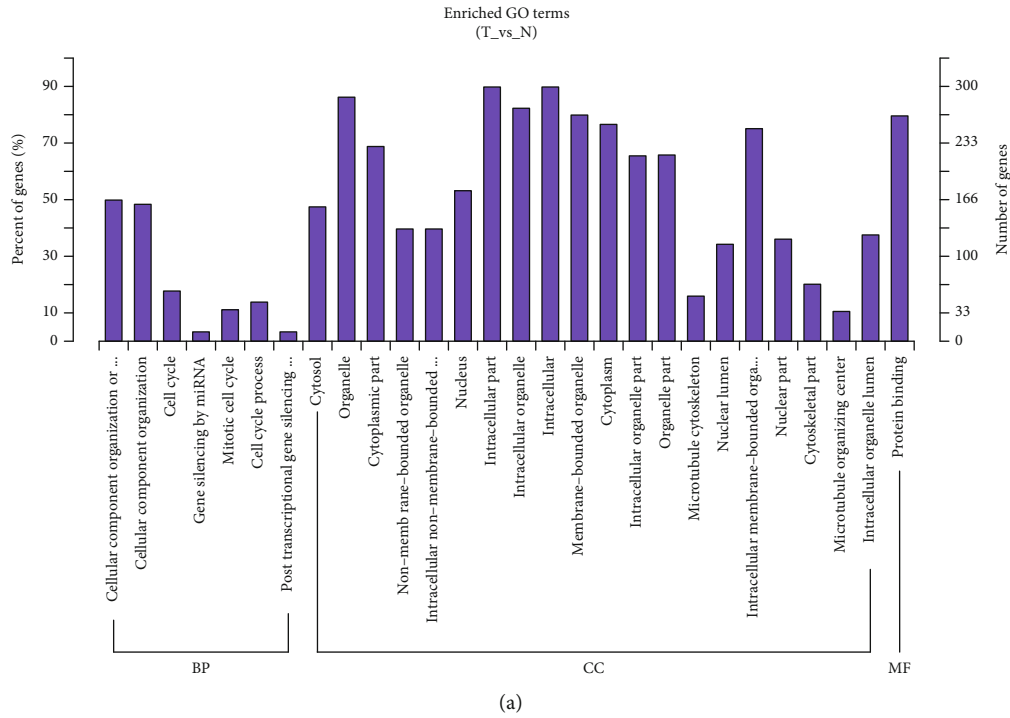


FIGURE 2: The function prediction of the host genes of circRNAs. (a) The GO analysis of the potential functions of the host genes of DECs. (b) The KEGG pathway analysis of the potential functions of the host genes of DECs.

miRNAs. In the present study, we used an integrated database, CSCD, to predict the interaction between miRNAs and circRNAs. Then, the specific mRNA-miRNA interactions were identified using miRTarBase, TargetScan, and miRDB. The network was constructed by Cytoscape. As presented in Figure 3, a total of 41 miRNAs, 166 circRNAs, and 2427 mRNAs were included in this ceRNA network (Supplementary table 1).

From the network, 6 circRNAs (hsa_circ_0004841, hsa_circ_0007523, hsa_circ_0008038, hsa_circ_0000799, hsa_circ_0002744, and hsa_circ_0005620) were found to be linked to more than 10 different miRNAs, which were identified as key circRNAs in CRC. Meanwhile, 14 mRNAs (including QKI, HECTD2, SYNCRIP, RUNX1, SEMA6D, VCL, MBNL1, HMGA2, ABCC5, ARRDC4, CPD, JPH1, MTDH, and TP53INP1) were identified as key regulators in this network, which were targeted by more than 25 miRNAs. In addition, hsa-miR-93-5p, hsa-miR-20a-5p, hsa-miR-17-5p, hsa-miR-106a-5p, hsa-let-7b-5p, hsa-miR-27a-3p, hsa-miR-15a-5p, hsa-miR-16-5p, hsa-let-7c-5p, hsa-miR-103a-3p, hsa-let-7d-5p, hsa-miR-107, hsa-let-7e-5p, hsa-miR-23a-3p, hsa-let-7a-5p, hsa-miR-30a-5p, hsa-miR-19a-3p, and hsa-miR-19b-3p were identified as the key miRNAs by targeting more than 500 genes.

3.4. GO and KEGG Pathway of ceRNA Network. Considering the above DECs, the functional role was still unclear, and the function of circRNA was predicted with their targeting mRNAs. GO and KEGG pathway analyses of differentially expressed mRNAs with significance were conducive to understand circRNAs.

The data showed that the target gene function of circRNAs was related to the modulation of metabolism, energy pathways, cell growth and/or maintenance, transport, cell communication, signal transduction, nucleobase regulation, nucleoside, nucleotide and nucleic acid metabolism, cell cycle, apoptosis, gene expression regulation, epigenetics, cell motility, enzyme activity negative regulation, DNA replication, lipid metabolism, chromosome segregation, and steroid metabolism (Figure 4(a)).

KEGG pathway analysis results revealed the target genes of circRNAs were related to the pathways, including syndecan-1-mediated signaling events, glypican pathway, nectin adhesion pathway, TRAIL signaling pathway, glypican 1 network, integrin family, ErbB receptor signaling network, VEGF and VEGFR signaling network, proteoglycan syndecan-mediated signaling events, LKB1 signaling events, mesenchymal-to-epithelial transition, and epithelial-to-mesenchymal transition (Figure 4(b)).

3.5. The Dysregulation of Hub Genes and miRNAs Was Correlated to the Survival Time in Patients with CRC. Then, to further explore the functions of these hub genes in the network in carcinogenesis and the development of CRC, we analyzed the correlation between the expression of hub mRNAs, or miRNAs and survival time in patients with CRC using GEPIA [43] and Kaplan–Meier plotter database. Unfortunately, both databases do not contain circRNAs. Thus, we did not analyze the correlation between the survival time

and the expression of circRNAs in CRC. As shown in Figure 5, the Kaplan–Meier curves indicated that higher expressions of QKI (Figure 5(a)), ABCC5 (Figure 5(b)), RUNX1 (Figure 5(c)), CALD1 (Figure 5(d)), and CLIP4 (Figure 5(e)) were dramatically linked to poorer overall survival. However, overexpressions of SYNCRIP (Figure 5(f)) and SEMA6D (Figure 5(g)) were related to longer survival time in patients with CRC.

Meanwhile, as presented in Figure 6, the Kaplan–Meier curves showed that hsa-miR-20a (Figure 6(a)), hsa-let-7b (Figure 6(b)), and has-miR-15 (Figure 6(c)) were related to longer survival time in patients with CRC. However, the higher expression level of hsa-let-7d (Figure 6(d)) would result in shorter overall survival time. All the results demonstrated that dysregulation of hub genes and miRNAs could be the potential targets for the prognosis of CRC.

4. Discussion

RNA sequencing technology has made it possible to extensively explore gene expression and promoted the study of susceptibility to disease, which is beneficial for disease treatment at the molecular level. Numerous researches have shown that RNA sequencing technology and in silico method were widely used to identify pathogenic mechanisms and uncover promising targets for diagnosis and therapy [44–46].

In this study, the RNA sequencing was applied and the bioinformatics data was analyzed to obtain differentially expressed circRNAs (DECs). Then, the probable DEC-sponged miRNAs was identified by Cancer-Specific CircRNA Database (CSCD). In addition, target mRNAs were predicted by bioinformatics analysis and a competitive endogenous network of circRNA-miRNA-mRNA was established. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases were applied to analyze candidate mRNAs and then assume the signaling pathways underlying in CRC. To identify DECs and explore the hidden mechanisms in this study might be helpful to develop novel treatment for CRC.

Here, we systematically analyzed and compared the expression profile of circRNAs in 3 CRC and normal tissues. The data suggested that circRNA expression profile was largely distinct in CRC tissues, compared to that in normal tissues. Notably, 289 of circRNAs expression presented ectopic in CRC after comparison with those in normal tissues. 76 of circRNA expression level were reduced, and 316 of circRNAs expression level were induced in CRC. We also observed this interesting phenomenon. We found the number of upregulated circRNAs is more than 3-fold compared to the number of downregulated circRNAs. Very interestingly, multiple previous studies also reported this phenomenon. For example, He et al. identified 94 downregulated circRNAs and 28 upregulated circRNAs in gastric cancer with GSE89143 [47] and identified 144 downregulated circRNAs and 52 upregulated circRNAs with GSE78092 [48]. Shi et al. observed 469 upregulated circRNAs and 275 downregulated circRNAs in ESCC [49]. Wen et al. found 109 circRNAs that were significantly upregulated and 56

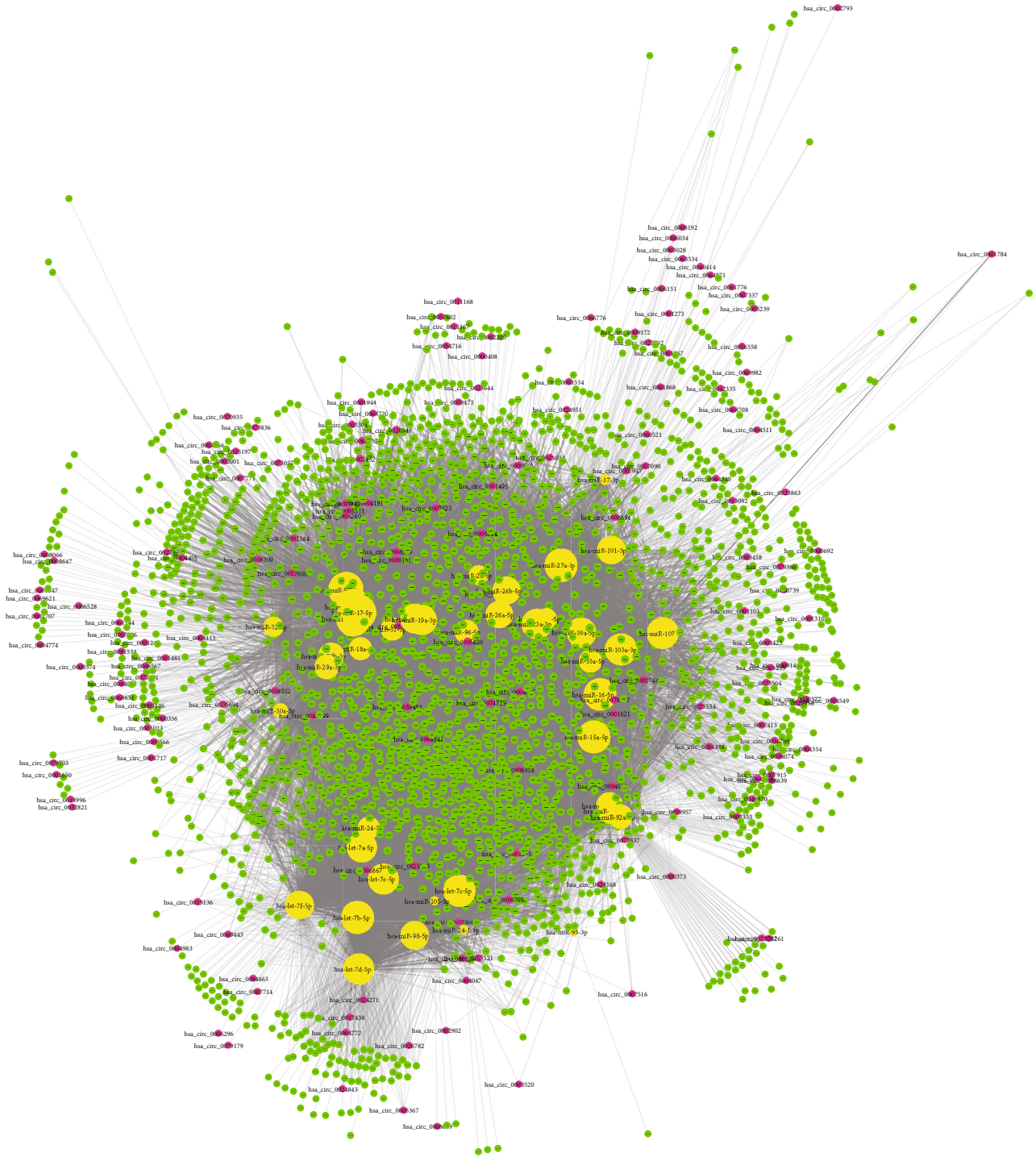


FIGURE 3: Construction of a circRNA-miRNA-mRNA network. A total of 41 miRNAs, 166 circRNAs, and 2427 mRNAs were included in this ceRNA network.

circRNAs that were downregulated among the rheumatoid arthritis patients by using RNA-seq method [50]. Our findings together with previous reports showed the circRNA expression pattern between disease and nondisease samples was not similar with mRNAs and lncRNAs, suggesting that posttranscriptional regulation may have a crucial role in modulating circRNA formation. The function of circRNAs in the host participated in modulating cell cycle, gene silenc-

ing by miRNA, mitotic cell cycle, cell cycle process, and post-transcriptional gene silencing by RNA.

Recently, the links between noncoding RNAs and cancer were greatly investigated. Newly generated researches showed that noncoding RNAs displayed importance in carcinogenesis and cancer development [51]. Some reports have shown that circRNAs displayed importance in the process of biology and the progression of diseases through sponging

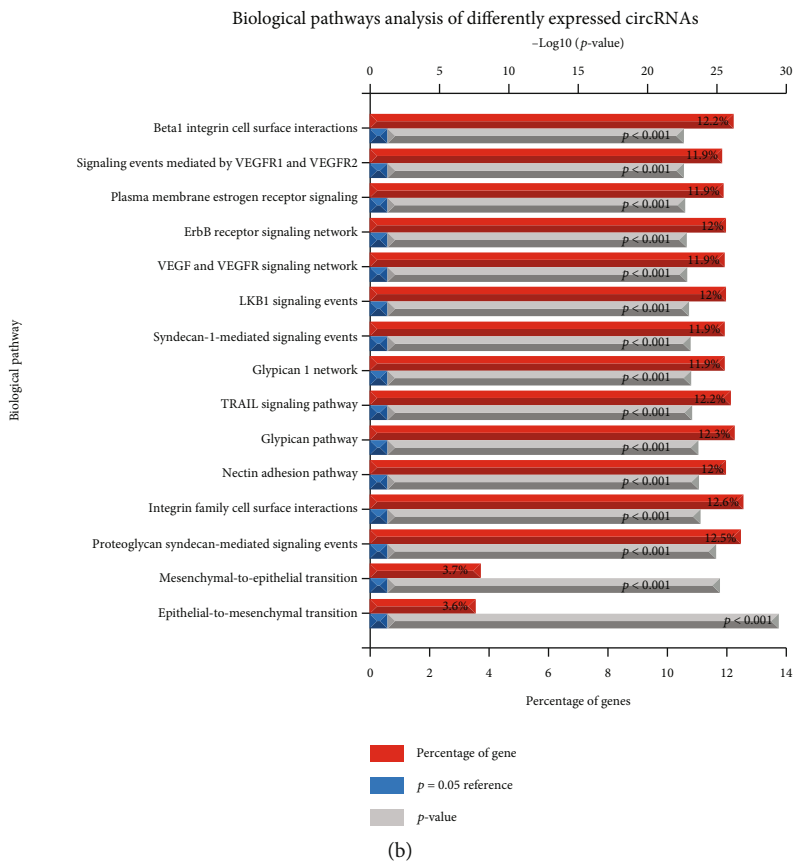
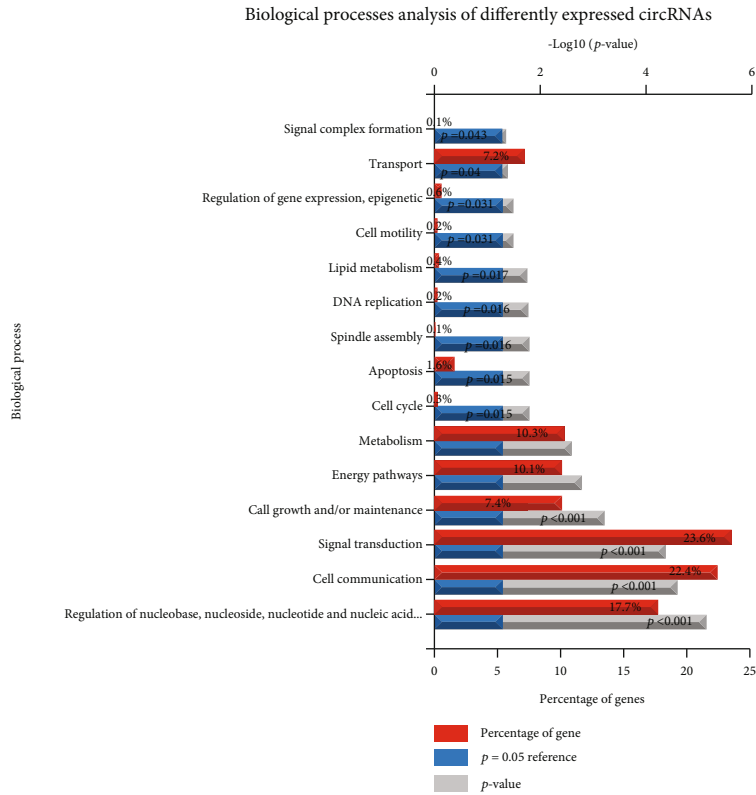


FIGURE 4: GO and KEGG pathway analysis of ceRNA networks. (a) The GO analysis of the potential functions of ceRNA networks. (b) The KEGG pathway analysis of the potential functions of ceRNA networks.

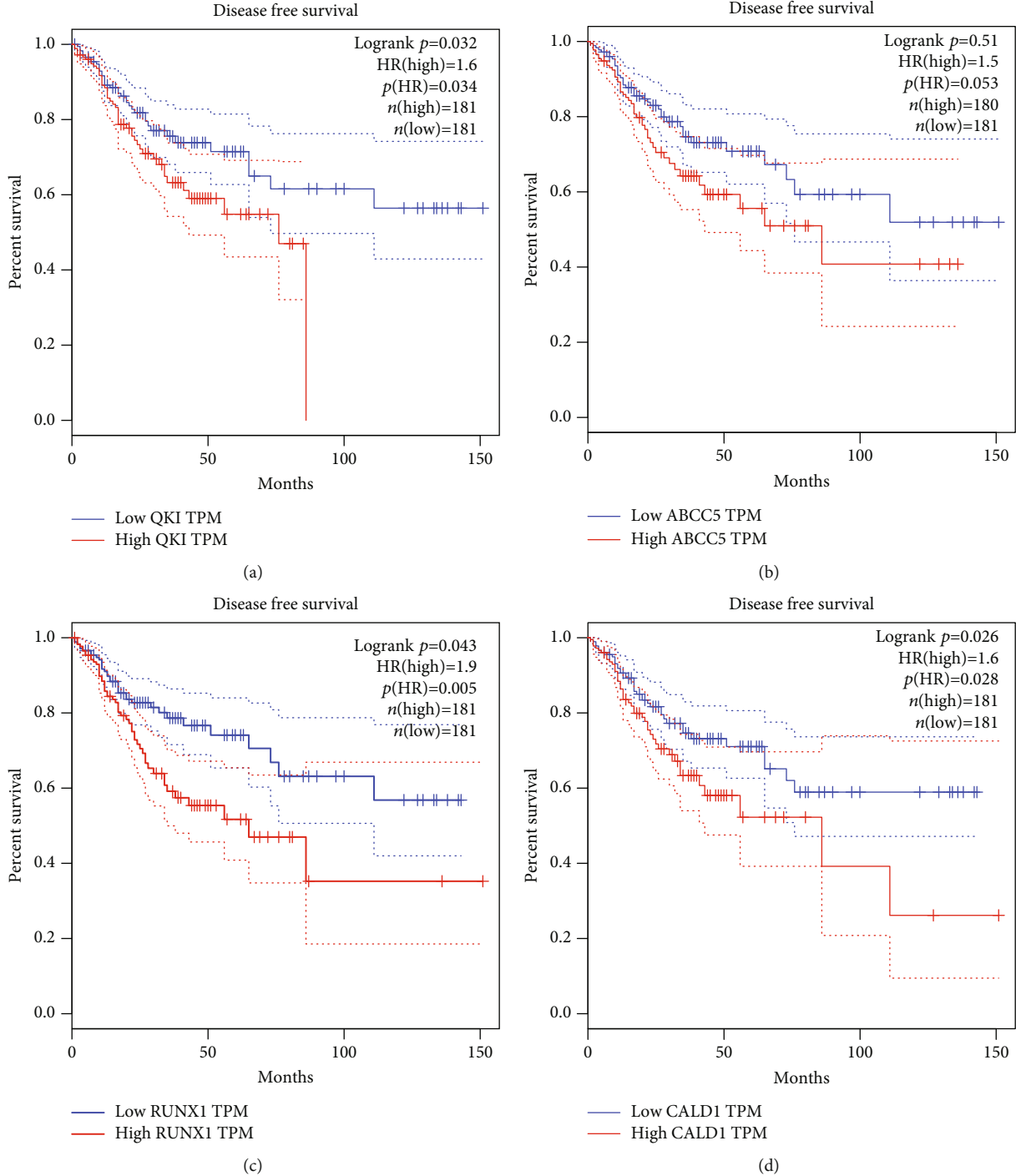


FIGURE 5: Continued.

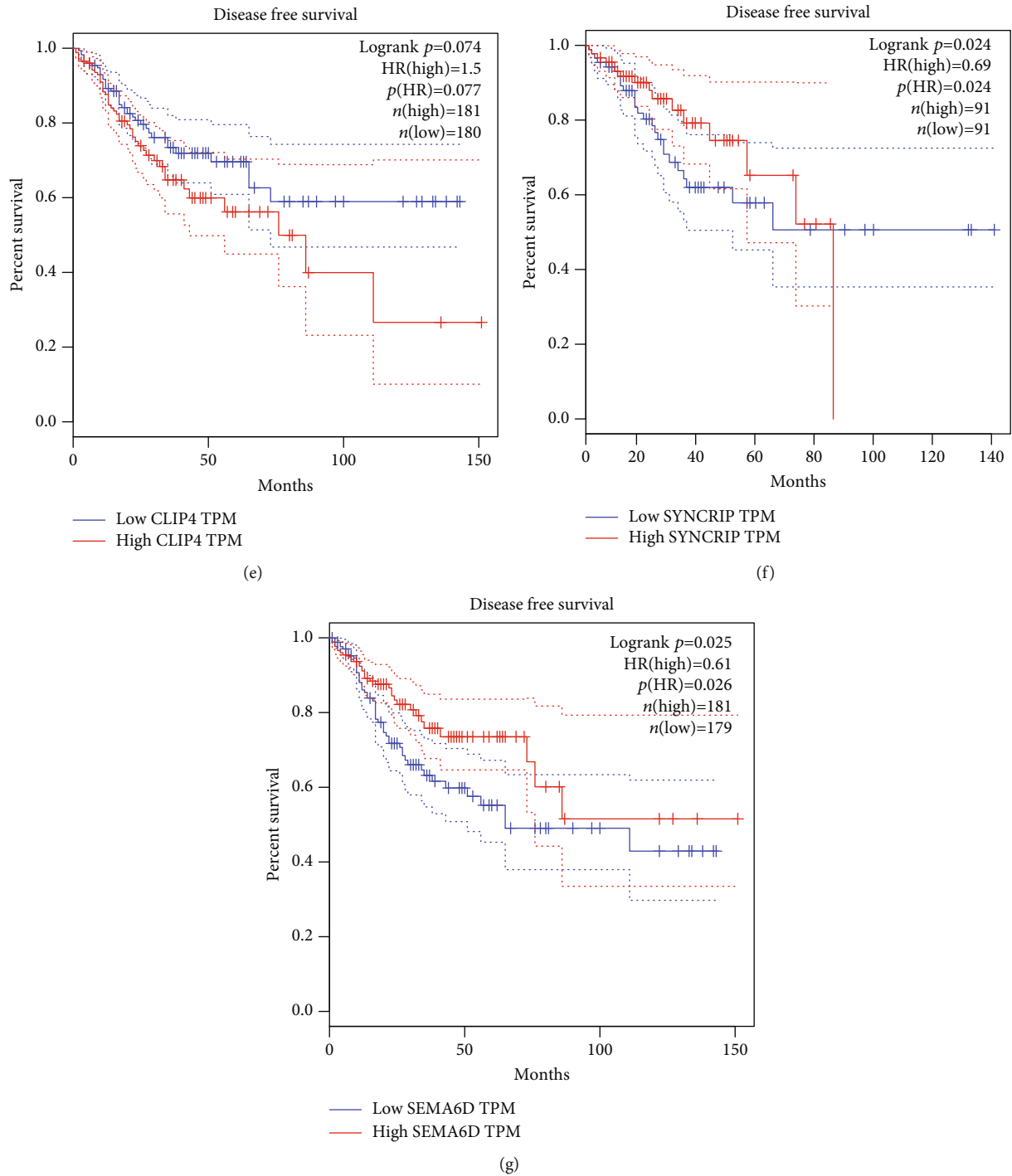


FIGURE 5: The dysregulation of hub genes was correlated to the survival time in patients with CRC. The Kaplan–Meier curves indicated that higher expression of QKI (a), ABCC5 (b), RUNX1 (c), CALD1 (d), and CLIP4 (e) and lower expression of SYNCRIP (f) and SEMA6D (g) were dramatically linked to poorer overall survival in patients with CRC.

miRNA [13, 14]. For example, circular RNA Circ100084, exhibiting as the sponge of miR-23a-5p, modulated the expression of IGF2 in hepatocarcinoma [52]. In this study, a total of 41 miRNAs, 166 circRNAs, and 2427 mRNAs were included in this ceRNA network. Bioinformatics analysis showed that the function of circRNA target genes had an association with metabolism regulation, energy

pathway regulation, cell growth and/or maintenance, transport, cell communication, signal transduction, nucleobase regulation, nucleoside, nucleotide and nucleic acid metabolism, cell cycle, apoptosis, gene expression regulation, epigenetics, cell motility, enzyme activity negative regulation, DNA replication, lipid metabolism, and chromosome segregation.

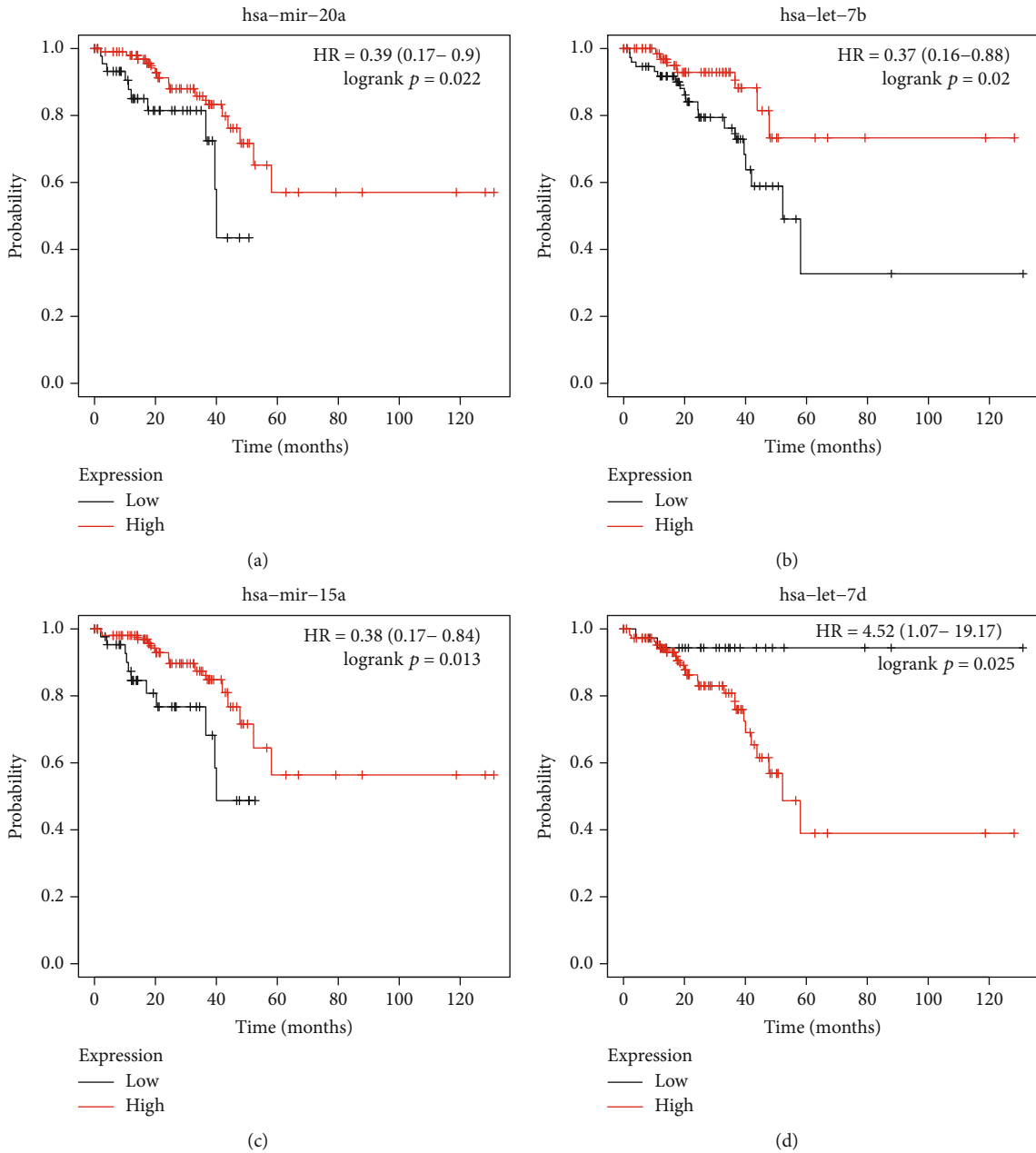


FIGURE 6: The dysregulation of hub miRNAs was correlated to the survival time in patients with CRC. The Kaplan–Meier curves indicated that higher expression of hsa-miR-20a (a), hsa-let-7b (b), and has-miR-15 (c) and lower expression of hsa-let-7d (d) were dramatically linked to longer overall survival in patients with CRC.

In the present study, 14 mRNAs (including QKI, HECTD2, SYNCRIP, RUNX1, SEMA6D, VCL, MBNL1, HMGA2, ABCC5, ARDC4, CPD, JPH1, MTDH, and TP53INP1) were identified as primary regulators in this network, which were targeted by more than 25 miRNAs. Several reports had illustrated the key characters of these mRNAs in human cancers as previously described. QKI protein has been revealed in human and was associated with many human diseases, such as cancer, and neurological diseases, such as human hereditary ataxia, various sclerosis, or schizophrenia. Currently, several findings suggested that knockout of QKI-5 genome would result in increased cell viability and

dedifferentiation in cancers, indicating that QKI-5 was an inhibitor of tumor in multiple types of many cancers [53]. QKI was identified as a key regulator of alternative splicing in cancers. Very interestingly, QKI was found to be related to the formation of circRNAs. Ablated QKI led to arrestment of circRNA expression-related EMT [54]. Further reports confirmed the active roles of QKI during the biogenesis of circRNAs [54]. Studies towards solid tumors have indicated RUNX1 possessing a context-dependent function displayed as an oncogene or a suppressor of tumor. These functions of Runx1 have been shown in breast, prostate, lung, and skin cancers, presenting a relationship between different subtypes

of cancers and stages of tumor progress. There are more and more evidences showing that Runx1 inhibited the invasiveness of most kinds of breast cancer, especially in the early stage of tumor development. In colon cancer, Systems Pharmacogenomics identified RUNX1 as an aspirin-responsive transcription factor. Vinculin (Vcl), a 117 kDa membrane-related protein, was expressed in global cells and functioned importantly in mechanotransduction. MBNL1, a RNA-binding protein, bind to 3'UTRs and facilitate mRNA decay. Hence, HMGA2 is considered to be an oncogene, which serves as a critical regulator of proliferation and survival. Besides, HMGA2 overexpression is linked to initial of metastasis and poorly prognostic status in a large number of cancers types [55]. ABCC5, also named by multidrug-resistance protein 5, has been shown to transport nucleosides and antifolates. The enhanced ABCC5 level was shown to be related to the occurrence of breast cancer, hepatocellular carcinoma, and pancreatic ductal adenocarcinoma. What is more, MTDH has been demonstrated to function vitally in tumor genesis, development, and resistance to chemotherapy. The abnormal expression and dysfunction of MTDH are related to the viability, survival, and metastasis of tumor cells. Apoptotic protein TP53INP1 (tumor protein 53-inducible nuclear protein 1) participated in the response from cellular stress.

Here, we identified 18 key miRNAs involved in regulating the activity of circRNAs. Among these miRNAs, higher expression of hsa-miR-20a, has-Let-7b, and hsa-miR-15 but lower expression of has-Let-7d were related to longer survival time in CRC. The miR-20a expression level was raised in CRC patients after a comparison with that in control. Several recent studies demonstrated that miR-20a retarded autophagy induced by hypoxia through targeting ATG5/FIP200 in CRC and modulated the sensitivity of CRC cells to NK cells by targeting MICA. Universal ablation of miR-15 (microRNA 15) and miR-16 in cell lines and tissues of numerous cancers revealed that the function of miR-15a/16-1 exhibited vitally in the progression of tumor. The prospect of miR-15a/16-1 was herein noteworthy in cancer therapy. In colon cancer, upregulation of miR-15a hindered cell viability and cycle. In normal cells, let-7 modulated cell viability, cycle, apoptosis, metabolism, and stemness. However, the let-7 microRNA level in CRC was shown to be decreased and was taken as a suppressor of tumor.

There are some limitations in our research. Firstly, the amount of samples is not sufficient. Secondly, our results merely from present toolsets and databases needed to be further improved. Thirdly, the parameter of prognosis regarding DEcircRNAs in CRC should be identified. More clinical samples and experiments would be supplemented in the following studies to consolidate our conclusions and evaluate the characters of these DEcircRNAs.

In conclusion, the present study identified 316 upregulated circRNAs and 76 downregulated circRNAs in CRC samples, in comparison with those in normal tissues. Bioinformatics analysis revealed that these circRNAs participated in metabolism regulation and cell cycle progression. Furthermore, we constructed differently expressed circRNAs to regulate ceRNA networks based on RNA-seq methods and

bioinformatics analysis. Finally, we thought our study could provide novel biomarkers and insights for CRC prognosis.

Data Availability

The datasets used during the present study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

Feng Que and Hua Wang are co-first authors and contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 81773947).

Supplementary Materials

Supplementary table 1: the detailed information about circRNA-miRNA-mRNA network. (*Supplementary Materials*)

References

- [1] S. Chakradhar, "Colorectal cancer: 5 big questions," *Nature*, vol. 521, no. 7551, p. S16, 2015.
- [2] H. Brody, "Colorectal cancer," *Nature*, vol. 521, no. 7551, p. S1, 2015.
- [3] N. Keum and E. Giovannucci, "Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies," *Nature Reviews. Gastroenterology & Hepatology*, vol. 16, no. 12, pp. 713–732, 2019.
- [4] C. De Divitiis, G. Nasti, M. Montano, R. Fisichella, R. V. Iaffaioli, and M. Berretta, "Prognostic and predictive response factors in colorectal cancer patients: between hope and reality," *World Journal of Gastroenterology*, vol. 20, no. 41, pp. 15049–15059, 2014.
- [5] E. Lasda and R. Parker, "Circular RNAs: diversity of form and function," *RNA*, vol. 20, no. 12, pp. 1829–1842, 2014.
- [6] L. L. Chen and L. Yang, "Regulation of circRNA biogenesis," *RNA Biology*, vol. 12, no. 4, pp. 381–388, 2015.
- [7] T. Shen, M. Han, G. Wei, and T. Ni, "An intriguing RNA species—perspectives of circularized RNA," *Protein & Cell*, vol. 6, no. 12, pp. 871–880, 2015.
- [8] Z. Zhang, T. Yang, and J. Xiao, "Circular RNAs: promising biomarkers for human diseases," *eBioMedicine*, vol. 34, pp. 267–274, 2018.
- [9] M. Su, Y. Xiao, J. Ma et al., "Circular RNAs in Cancer: emerging functions in hallmarks, stemness, resistance and roles as potential biomarkers," *Molecular Cancer*, vol. 18, no. 1, p. 90, 2019.
- [10] S. Qu, Z. Liu, X. Yang et al., "The emerging functions and roles of circular RNAs in cancer," *Cancer Letters*, vol. 414, pp. 301–309, 2018.

- [11] S. Greco, B. Cardinali, G. Falcone, and F. Martelli, "Circular RNAs in muscle function and disease," *International Journal of Molecular Sciences*, vol. 19, no. 11, p. 3454, 2018.
- [12] S. Xu, L. Zhou, M. Ponnusamy et al., "A comprehensive review of circRNA: from purification and identification to disease marker potential," *PeerJ*, vol. 6, article e5503, 2018.
- [13] W. Yuan, S. Peng, J. Wang et al., "Identification and characterization of circRNAs as competing endogenous RNAs for miRNA-mRNA in colorectal cancer," *PeerJ*, vol. 7, article e7602, 2019.
- [14] Y. Tian, Y. Xu, H. Wang et al., "Comprehensive analysis of microarray expression profiles of circRNAs and lncRNAs with associated co-expression networks in human colorectal cancer," *Functional & Integrative Genomics*, vol. 19, no. 2, pp. 311–327, 2019.
- [15] Q. Chen, Z. Chen, S. Cao et al., "Role of CircRNAs_100395 in proliferation and metastases of liver cancer," *Medical Science Monitor*, vol. 25, pp. 6181–6192, 2019.
- [16] X. Zhang, L. Zhang, L. Cui, M. Chen, D. Liu, and J. Tian, "Expression of circRNA circ_0026344 in gastric cancer and its clinical significance," *International Journal of Clinical and Experimental Pathology*, vol. 13, no. 5, pp. 1017–1023, 2020.
- [17] J. Liu, Z. Li, W. Teng, and X. Ye, "Identification of downregulated circRNAs from tissue and plasma of patients with gastric cancer and construction of a circRNA-miRNA-mRNA network," *Journal of Cellular Biochemistry*, 2020.
- [18] H. X. Ding, Q. Xu, B. G. Wang, Z. Lv, and Y. Yuan, "MetaDE-based analysis of circRNA expression profiles involved in gastric cancer," *Digestive Diseases and Sciences*, 2020.
- [19] R. Li, J. Jiang, H. Shi, H. Qian, X. Zhang, and W. Xu, "CircRNA: a rising star in gastric cancer," *Cellular and Molecular Life Sciences*, vol. 77, no. 9, pp. 1661–1680, 2020.
- [20] B. Pal and R. L. Anderson, "MiRNAs prognostic for basal and BRCA1 breast cancer," *Oncotarget*, vol. 9, no. 87, pp. 35717–35718, 2018.
- [21] D. A. Clump, C. R. Pickering, and H. D. Skinner, "Predicting outcome in head and neck Cancer: miRNAs with potentially big effects," *Clinical Cancer Research*, vol. 25, no. 5, pp. 1441–1442, 2019.
- [22] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng, "Overview of microRNA biogenesis, mechanisms of actions, and circulation," *Frontiers in Endocrinology*, vol. 9, p. 402, 2018.
- [23] X. Chen, R. Mao, W. Su et al., "Circular RNA circHIPK3 modulates autophagy via MIR124-3p-STAT3-PRKAA/AMPK α signaling in STK11 mutant lung cancer," *Autophagy*, vol. 16, no. 4, pp. 659–671, 2020.
- [24] L. Li, K. Wan, L. Xiong, S. Liang, F. Tou, and S. Guo, "CircRNA hsa_circ_0087862 acts as an oncogene in non-small cell lung cancer by targeting miR-1253/RAB3D axis," *Oncotargets and Therapy*, vol. 13, pp. 2873–2886, 2020.
- [25] X. Zheng, L. Chen, Y. Zhou et al., "A novel protein encoded by a circular RNA circPPP1R12A promotes tumor pathogenesis and metastasis of colon cancer via Hippo-YAP signaling," *Molecular Cancer*, vol. 18, no. 1, p. 47, 2019.
- [26] W. Weng, Q. Wei, S. Toden et al., "Circular RNA ciRS-7-a promising prognostic biomarker and a potential therapeutic target in colorectal cancer," *Clinical Cancer Research*, vol. 23, no. 14, pp. 3918–3928, 2017.
- [27] Y. Yan, M. Su, and B. Qin, "CircHIPK3 promotes colorectal cancer cells proliferation and metastasis via modulating of miR-1207-5p/FMNL2 signal," *Biochemical and Biophysical Research Communications*, vol. 524, no. 4, pp. 839–846, 2020.
- [28] Y. Zhang, C. Li, X. Liu et al., "circHIPK3 promotes oxaliplatin-resistance in colorectal cancer through autophagy by sponging miR-637," *eBioMedicine*, vol. 48, pp. 277–288, 2019.
- [29] K. Zeng, X. Chen, M. Xu et al., "CircHIPK3 promotes colorectal cancer growth and metastasis by sponging miR-7," *Cell Death & Disease*, vol. 9, no. 4, p. 417, 2018.
- [30] J. Hu, C. Ji, K. Hua et al., "Hsa_circ_0091074 regulates TAZ expression via microRNA-1297 in triple negative breast cancer cells," *International Journal of Oncology*, vol. 56, no. 5, pp. 1314–1326, 2020.
- [31] T. Xia, Z. Pan, and J. Zhang, "CircSMC3 regulates gastric cancer tumorigenesis by targeting miR-4720-3p/TJP1 axis," *Cancer Medicine*, vol. 9, no. 12, pp. 4299–4309, 2020.
- [32] S. Hao, L. Cong, R. Qu, R. Liu, G. Zhang, and Y. Li, "Emerging roles of circular RNAs in colorectal cancer," *Oncotargets and Therapy*, vol. 12, pp. 4765–4777, 2019.
- [33] J. H. He, Z. P. Han, J. G. Luo et al., "Hsa_Circ_0007843 acts as a miR-518c-5p sponge to regulate the migration and invasion of colon cancer SW480 cells," *Frontiers in Genetics*, vol. 11, p. 9, 2020.
- [34] X. N. Li, Z. J. Wang, C. X. Ye, B. C. Zhao, Z. L. Li, and Y. Yang, "RNA sequencing reveals the expression profiles of circRNA and indicates that circDDX17 acts as a tumor suppressor in colorectal cancer," *Journal of Experimental & Clinical Cancer Research*, vol. 37, no. 1, p. 325, 2018.
- [35] S. Xia, J. Feng, K. Chen et al., "CSCD: a database for cancer-specific circular RNAs," *Nucleic Acids Research*, vol. 46, no. D1, pp. D925–D929, 2018.
- [36] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Research*, vol. 34, no. 9, pp. D140–D144, 2006.
- [37] S. D. Hsu, F. M. Lin, W. Y. Wu et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D163–D169, 2011.
- [38] V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *Elife*, vol. 4, 2015.
- [39] X. Wang, "miRDB: a microRNA target prediction and functional annotation database with a wiki interface," *RNA*, vol. 14, no. 6, pp. 1012–1017, 2008.
- [40] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [41] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.
- [42] M. Y. Wu, Y. P. Tang, J. J. Liu, R. Liang, and X. L. Luo, "Global transcriptomic study of circRNAs expression profile in sorafenib resistant hepatocellular carcinoma cells," *Journal of Cancer*, vol. 11, no. 10, pp. 2993–3001, 2020.
- [43] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic Acids Research*, vol. 45, no. W1, pp. W98–W102, 2017.

- [44] H. Fei, S. Chen, and C. Xu, "RNA-sequencing and microarray data mining revealing: the aberrantly expressed mRNAs were related with a poor outcome in the triple negative breast cancer patients," *Annals of Translational Medicine*, vol. 8, no. 6, p. 363, 2020.
- [45] T. Machackova, K. Trachtova, V. Prochazka et al., "Tumor microRNAs identified by small RNA sequencing as potential response predictors in locally advanced rectal cancer patients treated with neoadjuvant chemoradiotherapy," *Cancer Genomics Proteomics*, vol. 17, no. 3, pp. 249–257, 2020.
- [46] R. T. Davis, K. Blake, D. Ma et al., "Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing," *Nature Cell Biology*, vol. 22, no. 3, pp. 310–320, 2020.
- [47] Y. Shao, J. Li, R. Lu et al., "Global circular RNA expression profile of human gastric cancer and its clinical significance," *Cancer Medicine*, vol. 6, no. 6, pp. 1173–1180, 2017.
- [48] Y. S. Huang, N. Jie, K. J. Zou, and Y. Weng, "Expression profile of circular RNAs in human gastric cancer tissues," *Molecular Medicine Reports*, vol. 16, no. 3, pp. 2469–2476, 2017.
- [49] P. Shi, J. Sun, B. He et al., "Profiles of differentially expressed circRNAs in esophageal and breast cancer," *Cancer Management and Research*, vol. 10, pp. 2207–2221, 2018.
- [50] J. Wen, J. Liu, P. Zhang et al., "RNA-seq reveals the circular RNA and miRNA expression profile of peripheral blood mononuclear cells in patients with rheumatoid arthritis," *Bio-science Reports*, vol. 40, no. 4, 2020.
- [51] F. Calore, F. Lovat, and M. Garofalo, "Non-coding RNAs and cancer," *International Journal of Molecular Sciences*, vol. 14, no. 8, pp. 17085–17110, 2013.
- [52] J. Yang, Y. Li, Z. Yu et al., "Circular RNA Circ100084 functions as sponge of miR-23a-5p to regulate IGF2 expression in hepatocellular carcinoma," *Molecular Medicine Reports*, vol. 21, no. 6, pp. 2395–2404, 2020.
- [53] Y. Feng and A. Bankston, "The star family member," *Advances in Experimental Medicine and Biology*, vol. 693, pp. 25–36, 2010.
- [54] S. J. Conn, K. A. Pillman, J. Toubia et al., "The RNA binding protein quaking regulates formation of circRNAs," *Cell*, vol. 160, no. 6, pp. 1125–1134, 2015.
- [55] X. Gao, M. Dai, Q. Li, Z. Wang, Y. Lu, and Z. Song, "HMGA2 regulates lung cancer proliferation and metastasis," *Thoracic Cancer*, vol. 8, no. 5, pp. 501–510, 2017.

Research Article

A Semantic Analysis and Community Detection-Based Artificial Intelligence Model for Core Herb Discovery from the Literature: Taking Chronic Glomerulonephritis Treatment as a Case Study

Yun Zhang,¹ Yongguo Liu ,¹ Jiajing Zhu,¹ Shuangqing Zhai,² Rongjiang Jin,³ and Chuanbiao Wen ⁴

¹Knowledge and Data Engineering Laboratory of Chinese Medicine, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

²School of Basic Medical Science, Beijing University of Chinese Medicine, Beijing 100029, China

³College of Health Preservation and Rehabilitation, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China

⁴College of Medical Information Engineering, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

Correspondence should be addressed to Yongguo Liu; liuyg_cn@163.com

Received 11 March 2020; Revised 14 July 2020; Accepted 14 August 2020; Published 1 September 2020

Academic Editor: Lin Lu

Copyright © 2020 Yun Zhang 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 Traditional Chinese Medicine (TCM) formula is the main treatment method of TCM. A formula often contains multiple herbs where core herbs play a critical therapeutic effect for treating diseases. It is of great significance to find out the core herbs in formulae for providing evidences and references for the clinical application of Chinese herbs and formulae. In this paper, we propose a core herb discovery model CHDSC based on semantic analysis and community detection to discover the core herbs for treating a certain disease from large-scale literature, which includes three stages: corpus construction, herb network establishment, and core herb discovery. In CHDSC, two artificial intelligence modules are used, where the Chinese word embedding algorithm ESSP2VEC is designed to analyse the semantics of herbs in Chinese literature based on the stroke, structure, and pinyin features of Chinese characters, and the label propagation-based algorithm LILPA is adopted to detect herb communities and core herbs in the herbal semantic network constructed from large-scale literature. To validate the proposed model, we choose chronic glomerulonephritis (CGN) as an example, search 1126 articles about how to treat CGN in TCM from the China National Knowledge Infrastructure (CNKI), and apply CHDSC to analyse the collected literature. Experimental results reveal that CHDSC discovers three major herb communities and eighteen core herbs for treating different CGN syndromes with high accuracy. The community size, degree, and closeness centrality distributions of the herb network are analysed to mine the laws of core herbs. As a result, we can observe that core herbs mainly exist in the communities with more than 25 herbs. The degree and closeness centrality of core herb nodes concentrate on the range of [15, 40] and [0.25, 0.45], respectively. Thus, semantic analysis and community detection are helpful for mining effective core herbs for treating a certain disease from large-scale literature.

1. Introduction

Artificial intelligence is the general term of the modern technology of computer science [1], including image recognition [2], network analysis [3], and natural language processing [4]. Artificial intelligence technologies have been utilized in various fields of medicine, for example, automatic disease diagnosis [5], pathogenic network analysis [6], and biological

text analysis [7] [8]. Meanwhile, Traditional Chinese Medicine (TCM) plays an important role and provides a unique theoretical and practical way to treat diseases for thousands of years in Chinese history. TCM has many treatments, such as acupuncture, medicinal wine, medicinal formula, and medicinal diet [9, 10]. Among them, medicinal formula, also called as the TCM formula, is the frequently used mode and is made up of several Chinese herbs. The TCM formula has

many characteristics, such as compatibility combination, efficacy, treatment mechanism, and medication taboo [9, 11]. Compatibility combination can reflect the rationality of herb combination in formulae and guide TCM doctors to make up formulae [12], which mainly contains the “Sovereign-Minister-Assistant-Courier” combination rule and herb pair combination rule [13] [14]. Among them, the “Sovereign-Minister-Assistant-Courier” combination rule, also known as the “Jun-Chen-Zuo-Shi” combination rule, is a major combination principle of TCM formulae [15]. According to this principle, the sovereign herb plays a major role for dealing with main symptoms and syndromes of diseases, the minister herb helps the sovereign herb to strengthen herbal efficacy, and the assistant and courier herbs provide supporting function to reconcile formulae (e.g., reducing side effects) [15, 16]. Thus, the herbs acting as the sovereign or minister play a key role in terms of treating diseases, while others play an assistant role [16, 17]. In this way, the herbs serving as the sovereign or minister are viewed as core herbs in TCM formulae [17–19]. In other words, a formula contains multiple herbs, and core herbs play a critical therapeutic effect for treating diseases. Many formulae are collected in books, medical records, and scientific literature; however, most of them do not record their core herbs [19], which is difficult for young doctors and learners to master the core concern of formulae and prescribe effective formulae for treating different diseases. Thus, discovering core herbs can help doctors and learners to understand the quintessence of formulae quickly and provide evidences and references for the clinical application of herbs and formulae [16, 18, 19]. Through discovered core herbs, doctors can optimize the herb combination of formulae and synergize herb efficacies to prescribe more effective formulae for treating diseases [15], [19].

In general, researchers mainly explored core herbs by manual analysis [20–22], data analysis [19, 23–27], and clinical and pharmacology experiments [16, 28]. The traditional way to discover core herbs is the manual analysis on TCM books. Researchers first collected the relative books about the TCM treatment of a certain disease and then explored the possible relations between herbs and this disease. Finally, they discovered core herbs according to frequent relations, which is suitable for small-scale researches [20–22]. Recently, researchers utilized data analysis methods, such as statistical approach, association rule, mutual information, and entropy clustering, to analyse the frequency of herbs and their co-occurrence relations in formulae for discovering herbal compatibility rules and core herbs from medical records [19, 23–27]. Data analysis approaches can deal with large-scale medical records; however, they need structured data. It is known that medical records contain personal information (e.g., name, age, and sex), diagnostic information (e.g., laboratory index, symptom, syndrome, and disease), and treatment information (e.g., western drug, Chinese herb, formula, and medical advice) [29] [30]. In order to discover core herbs for treating a certain disease, researchers must extract partial diagnostic and treatment information from large-scale records, which costs more time. Meanwhile, it is worth noting that existing core herb discovery models cannot understand the inner meanings and functions of herbs in

these records [19, 23–27]. For example, herb *liquorice root* (Gan Cao) has many attributes, such as usage, efficacy, and taboo; however, existing models only consider the characters of Chinese words as text, then they cannot capture the implicit characteristics of this herb. In clinical experiments, researchers evaluated the efficacy of different herb combinations of TCM formulae on subjects to find effective herbs as core herbs [28]. In pharmacology experiments, researchers designed evaluation indexes, such as the network recovery index, to measure the scores of different ingredients in TCM formulae to find high score ingredients and considered the herbs with these ingredients as core herbs [16]. The experimental ways focus on few classical formulae and can analyse herb components in clinical trial and microscopic analysis perspectives to achieve high accuracy. However, enumerating all potential herb combinations and ingredients in an experimental way maybe impossible.

Besides books and medical records, there is rich scientific literature containing medical knowledge about TCM formulae [31]. To our best knowledge, there are few researches about discovering core herbs from the scientific literature. We consider some reasons: (1) literature is unstructured text, where disease, formula, and herb information are unevenly distributed in full text and cannot be processed easily; (2) it is hard to analyse the semantics of herbs in the literature; and (3) there are no good ways to represent herb semantics. Minority researchers studied classical literature to mine treatment patterns [32, 33], but they also process them artificially to deal with problem (1). However, they also do not analyse the inner meanings and functions of herbs in the literature for problems (2) and (3). In order to mimic the human learning mode for relatively accurately comprehending the literature and improve the efficiency of literature analysis, we introduce semantic analysis and community detection to handle these problems for analysing large-scale literature and discover core herbs efficiently.

In this paper, we propose an artificial intelligence model CHDSC for discovering the core herb for treating a certain disease based on semantic analysis and community detection, whose framework is shown in Figure 1. CHDSC mainly contains two artificial intelligence modules, in which a semantic analysis module is a natural language processing algorithm for analysing the semantics of herbs in large-scale literature by a Chinese word embedding algorithm ESSP2VEC as described in Section 3.1, and the community detection module is a network analysis algorithm to discover herb communities in the herbal semantic network by a label propagation-based algorithm LILPA as described in Section 3.2. The herbal semantic network is constructed by the semantic similarity of herbs based on the results of the semantic analysis module. The semantics of herbs contain which disease can be treated and how is it treated, then the herbal semantic network can reflect the relations between herbs and disease. Herbs in each community have the same or similar efficacy for treating multiple syndromes of a certain disease. Further, we consider important herbs in each herb community as the core herbs for treating the syndromes characterized by the community.

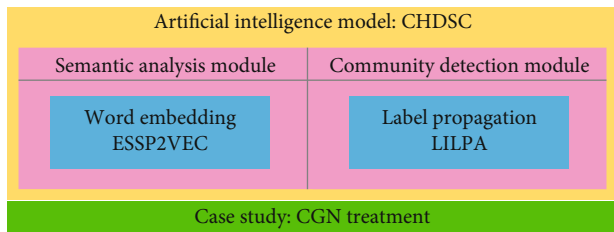


FIGURE 1: The framework of CHDSC. CHDSC consists of two modules, semantic analysis and community detection, in which the former is a Chinese word embedding algorithm ESSP2VEC to analyse the semantics of herbs and the latter is a label propagation-based algorithm LILPA to detect herb communities and core herbs. As a case study, we take CGN as an example and discover core herbs to treat different syndromes of CGN.

In order to validate the proposed model, we choose chronic glomerulonephritis (CGN) and discover the core herbs for treating this disease as a case study. Chronic Kidney Disease (CKD) is a class of kidney diseases with proteinuria, oedema, and haematuria as clinical symptoms [34]. The overall prevalence of CKD is 10.8%, and the number of patients in China is up to about 119.5 million [34]. CGN is a typical disease of CKD, which has different symptoms, such as oedema, haematuria, anaemia, albuminuria, and kidney function decrease and may lead to different degrees of renal dysfunction and chronic renal failure. CGN may damage heart function and the central nervous system and threaten life when it is severe [35]. In TCM, CGN is mainly recognized as the syndrome of qi deficiency of the spleen and kidney, the syndrome of deficiency of both qi and yin, the syndrome of yin deficiency of the liver and kidney, the syndrome of yang deficiency of the spleen and kidney, and the syndrome of liver depression and qi stagnation [36]. In addition, CGN also contains the syndrome of fluid-dampness, the syndrome of dampness-heat, the syndrome of blood stasis, and the syndrome of damp-turbidity [36]. It is shown that TCM treatment can improve and recover renal function and alleviate clinical symptoms [37]. Thus, discovering core herbs in TCM formulae for CGN treatment is helpful for improving the curative effect and precisely prescribing medicine. TCM doctors can utilize effective core herbs to form new formulae for treating different syndromes of patients with CGN.

In order to discover core herbs for treating different syndromes of CGN, we propose CHDSC with three stages: corpus construction, herb network establishment, and core herb discovery. The literature of CGN treatment in TCM is acquired from the China National Knowledge Infrastructure (CNKI). In the first stage, the CGN corpus is constructed by preprocessing the collected large-scale literature. In the second stage, a semantic analysis module based on word embedding is proposed by integrating the stroke, structure, and pinyin features of Chinese characters to analyse the semantics of herbs in literature, then the semantic similarity among herbs is measured, and a herbal semantic network is built according to semantic similarity. In the last stage, a community detection module based on label propagation is used to discover herb communities and core herbs in the herbal semantic network. We also analyse the community size,

degree, and closeness centrality distributions of the network to mine the rules of core herbs. Experimental results show that CHDSC uncovers three major herb communities where herbs in each community can be used for treating multiple syndromes of CGN, and discovers the core herbs for curing different syndromes of CGN with high accuracy. Core herbs mainly exist in the herb communities with more than 25 herbs. The degree and closeness centrality of core herb nodes in the herb network concentrate on the range of [15, 40] and [0.25, 0.45], respectively.

2. Related Work

In general, there are three type ways to discover core herbs: manual analysis, data analysis, and clinical and pharmacology experiments.

In manual analysis, researchers searched Chinese books about the TCM treatment of a specified disease, extracted corresponding formulae, and found core herbs by hand. Wang [20] investigated some classical books such as Shen-Nong-Ben-Cao-Jing and Huang-Di-Nei-Jing to discuss the methods for exploring the sovereign, minister, assistant, and courier herbs. Wang and Wang [21] analysed the compatibility and function of Zhi-Gan-Cao-Tang and found that *liquorice root* (Gan Cao) is its core herb with the efficacy of making up qi, blood, yin, and yang. Song and Niu [22] drew the rules on the determination of the sovereign herbs of Xie-Xin-Tang and analysed its sovereign herbs.

In data analysis, researchers discovered core herbs based on the frequency of herbs and their cooccurrence relations in datasets. Meanwhile, most studies focused on medical records. Zhou et al. [19] proposed an Effect Degree- (ED-) based algorithm to discover core herbs and compatibility rules with three steps: core herb discovery based on ED, network construction based on pointwise mutual information, and herb compatibility rule detection. They found 42 core herbs for treating consumptive lung disease. Zhan et al. [23] collected CGN treatment data in a Chinese biomedical literature database and mined the relationship among symptoms, syndromes, herbs, and formulae by the stratification algorithm based on keyword frequency, then they discovered that *milkvetch root* (Huang Qi), *danshen root* (Dan Shen), and *Indian bread* (Fu Ling) are core herbs. Ma et al. [24] extracted herbs, therapies, syndromes, and diseases in TCM formulae from medicine records and built a relation graph by NetDraw. The degree and closeness centrality were calculated to discover core herbs, then they found nine core herbs for treating gastric abscess. You et al. [25] established a formula database of bone marrow suppression treatment with a TCM kidney-tonifying method after radiotherapy and chemotherapy and applied cluster techniques and association rules to analyse medication rules. They found that *milkvetch root* (Huang Qi), *atractylodis macrocephalae rhizoma* (Bai Zhu), and *ligustri lucidi fructus* (Nv Zhen Zi) are frequently used herbs. Most data analysts also discovered the compatibility rules and treatment patterns of TCM formulae where core herbs are contained. Chen et al. [26] mined symptom-herb patterns with the triangular relationship of symptoms, syndromes, and herbs from medical records. They found

the main symptom-herb patterns on four real-world patient records (insomnia, diabetes, infertility, and Tourette syndrome). Chang et al. [27] investigated the treatment patterns among stroke patients by a nationwide population-based study using random samples of one million individuals from the national health insurance research database in Taiwan. They found that Bu-Yang-Huan-Wu-Tang and *danshen root* (Dan Shen) are commonly used.

In clinical and pharmacology experiments, researchers analysed effective herb combinations or ingredients of a given formula to explore core herbs for treating a certain disease. Yan et al. [28] proposed a study protocol to explore the core herbs for treating primary insomnia in TCM, in which they performed a triple-blind, randomized, and parallel-group clinical trial to analyse the formulae of prestigious TCM clinicians and used association rules to find effective core herbs. Wu et al. [16] identified the roles of “Sovereign-Minister-Assistant-Courier” of herbs in the Qi-Shen-Yi-Qi formula for treating myocardial ischemia by the network pharmacology approach. They integrated disease-associated genes and protein-protein interaction experiments to construct an organism disturbed network of myocardial ischemia and developed a network-based index, Network Recovery Index (NRI), to measure the therapeutic efficacy of the Qi-Shen-Yi-Qi formula. As a result, the whole formula gets the NRI score of 864.48 and outperforms a single herb. Additionally, *danshen root* (Dan Shen) and *milkvetch root* (Huang Qi) obtain the NRI scores of 734.31 and 680.27, respectively; thus, the two herbs are regarded as core herbs.

The above researches obtain good results for discovering core herbs; however, manual analysis and medical experiments need high cost for large-scale samples. Meanwhile, for data analysis, researchers need to process medical records manually to obtain structured data. Data analysis methods are based on cooccurrence relations and do not contain the inner meaning of herbs in medical records. On the other hand, there is large-scale literature containing the domain knowledge of formulae. In this paper, we focus on analysing literature and introduce semantic analysis and community detection to analyse the meanings of herbs in the literature to discover core herbs for treating a disease in TCM.

3. Artificial Intelligence Module

In this section, we introduce the semantic analysis and community detection modules used in CHDSC for core herb discovery. Firstly, we propose a Chinese word embedding algorithm ESSP2VEC to deal with large-scale literature and analyse the semantics of herbs based on the stroke, structure, and pinyin features of Chinese characters by predicting the contextual words of Chinese words. Secondly, we adopt a label importance-based label propagation algorithm LILPA to detect herb communities and core herbs, in which labels are propagated according to label importance based on node importance and node attraction. If the nodes own the same label when LILPA ends, they are allocated to the same community.

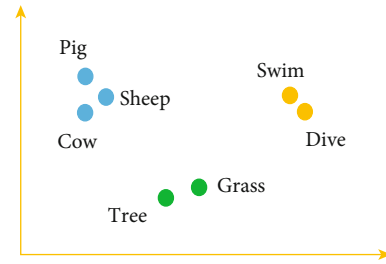


FIGURE 2: Example of semantic word vectors.

3.1. Word Embedding Algorithm. In order to analyse the semantics of herbs in large-scale literature, we propose the Chinese word embedding algorithm ESSP2VEC. We consider that it is a suitable model for learning the meanings of herbs from large-scale literature to handle problems (1) and (2). Word embedding is utilized to analyse word meanings based on the distributional hypothesis that similar words tend to appear in similar contexts; in other words, the semantics of words are included in their contextual words [38, 39]. Words are expressed as semantic word vectors, then we can consider that the meanings of words are contained in them [39]. In order to understand semantic word vectors intuitively, we take an example to visual some words in a two-dimensional surface by their semantic word vectors. As shown in Figure 2, semantic word vectors can contain some meanings of words and better distinguish different types of words, such as the animals (pig, sheep, and cow), the plants (tree and grass), and the actions (swim and dive). Thus, word embedding can capture the semantics of words to a certain degree. In collected large-scale literature, we can analyse the semantics of herbs and express them as semantic word vectors, which is a way to improve problem (3) to embody the semantics of herbs.

Further, herbs in the collected literature are recorded as Chinese words. Chinese words are made up of Chinese characters, which contain many semantically related internal features [40] [41]. Researchers proposed many Chinese word embedding algorithms for analysing the meanings of Chinese words by exploiting the character feature of Chinese words [42] and the internal features of Chinese characters, such as radical [43], component [44], stroke n -grams [39], structure [40], and pinyin [40]. Here, we introduce these features briefly.

- (i) *Character* (https://en.wikipedia.org/wiki/Chinese_characters): characters are logogram developed for the writing of Chinese, which makes up Chinese words [42].
- (ii) *Radical* ([https://en.wikipedia.org/wiki/Radical_\(Chinese_characters\)](https://en.wikipedia.org/wiki/Radical_(Chinese_characters))): radical is the first stroke or morphological component of Chinese characters, which is the catalogue of symbols that are classified according to the structure and meaning of Chinese characters in a dictionary [43].

- (iii) *Component*: component is a character-forming unit and has the function of assembling Chinese characters [44].
- (iv) *Stroke n -gram* ([https://en.wikipedia.org/wiki/Stroke_\(CJK_character\)](https://en.wikipedia.org/wiki/Stroke_(CJK_character))): stroke is the uninterrupted dots and lines of various shapes that compose Chinese characters, such as horizontal, vertical, left-falling, right-falling, and turning, which is the smallest constitutional unit of Chinese characters. Stroke n -gram is the combination of strokes according to stroke order (https://en.wikipedia.org/wiki/Stroke_order) [39].
- (v) *Structure*: structure is the azimuth relationship (13 patterns) among strokes, such as left-right and left-middle-right [40].
- (vi) *Pinyin* (<https://en.wikipedia.org/wiki/Pinyin>): pinyin is the romanization of Chinese characters, which consists of initials, finals, and tones [40].

However, existing researches do not consider these features together. Stroke n -grams include radical and component features and can capture partial semantics of the entire character [39] [40]. Meanwhile, the structure feature can capture the implication meanings of characters, and the pinyin feature can help us to understand the meanings of onomatopoeia and distinguish the characters which have the same stroke n -gram and structure [40]. Then, we can catch relatively comprehensive semantics of Chinese characters from the stroke n -gram, structure, and pinyin features. Thus, we propose ESSP2VEC to integrate the stroke n -gram, structure, and pinyin features of Chinese characters for analysing the semantics of Chinese words.

The architecture of ESSP2VEC is shown in Figure 3 with an explanatory example. In this example, we have a sentence “carry forward the spirit of laborious struggle vigorously,” where the target word is “laborious (<https://www.zdic.net/hans/%E8%89%B0%E8%8B%A6>),” which is made up of two Chinese characters, and its contextual words are “vigorously,” “carry forward,” “struggle,” and “spirit.” ESSP2VEC consists of input, feature extraction, feature encoding, ensemble feature, and output layers.

- (i) *Input layer*: input layer is used to receive the target word w_t .
- (ii) *Feature extraction layer*: this layer is used to decompose word w_t to independent characters and extract the stroke, structure, and pinyin of each character.
- (iii) *Feature encoding layer*: this layer is used to encode the stroke, structure, and pinyin features. We adopt the code defined in [40] to encode the stroke, structure, and pinyin features.
- (iv) *Ensemble feature layer*: this layer is designed to generate stroke n -gram (all combinations of stroke) by moving a slide window with different lengths on the stroke sequence as shown in Figure 3 and integrate stroke n -gram, structure, and pinyin features.

- (v) *Output layer*: output layer is designed as a *softmax* layer [45] to calculate the probability that the contextual words of word w_t are predicted based on the ensemble features of word w_t .

Similar to [39–45], we predict the contextual words based on the target word in ESSP2VEC. In particular, the target word is expressed as its ensemble features. Given corpus C represented as the sequence of words $w_1, \dots, w_t, \dots, w_{N_{\text{word}}}$ formally, where the word w_t is the target word and N_{word} is the number of words. The set of the contextual words of word w_t is represented as

$$C_t = \{w_{t+i}\}, (i \in [-c, 0) \cup (0, c]), \quad (1)$$

where c is the size of the contextual words and word w_c represented the element of C_t , $w_c \in C_t$, then the objective of ESSP2VEC is to maximize the log-likelihood in equation (2) where we hope to get the maximization of possibility $p(w_c | w_t)$, that is, word w_c can be predicted correctly with maximum possibility based on the target word w_t .

$$\mathcal{L} = \frac{1}{N_{\text{word}}} \sum_{t=1}^{N_{\text{word}}} \sum_{w_c \in C_t} \log p(w_c | w_t). \quad (2)$$

Then, the *softmax* function is used to model probability $p(w_c | w_t)$ of predicting word w_c given word w_t , which is defined as

$$p(w_c | w_t) = \frac{e^{s(w_t, w_c)}}{\sum_{j=1}^{N_{\text{word}}} e^{s(w_t, w_j)}}, \quad (3)$$

where $s(w_t, w_c)$ is a scoring function to map the pairs of word w_t and word w_c to a real number.

Chinese characters with similar stroke n -gram, structure, and pinyin may have similar semantics [40]. Thus, Chinese characters having similar ensemble features should have similar senses. Then, we define $s(w_t, w_c)$ as equation (4) to calculate their similarity based on the ensemble features of word w_t and its contextual word w_c , where $F(w_t)$ denotes the collection of the stroke n -grams of word w_t ; $v_{\text{stroke } n\text{-gram}}$, $v_{\text{structure}}$, and v_{pinyin} are the embeddings of stroke n -gram, structure, and pinyin features, respectively; and v_{w_c} is the initial semantic word vector of word w_c . By replacing w_c as w_j , we also can compute $s(w_t, w_j)$.

$$s(w_t, w_c) = \left(\left(\sum_{\text{stroke } n\text{-gram} \in F(w_t)} v_{\text{stroke } n\text{-gram}} \right) + v_{\text{structure}} + v_{\text{pinyin}} \right) \bullet v_{w_c}. \quad (4)$$

We optimize the objective function of equation (2) based on standard gradient methods [39]. After the training process, the semantic word vectors of contextual words are the output. Thus, we can obtain semantic word vectors $U = \{u_1, \dots, u_t, \dots, u_{N_{\text{word}}}\}$ of all words in the corpus, where u_t denotes the semantic word vector of word w_t and N_{word} is the number

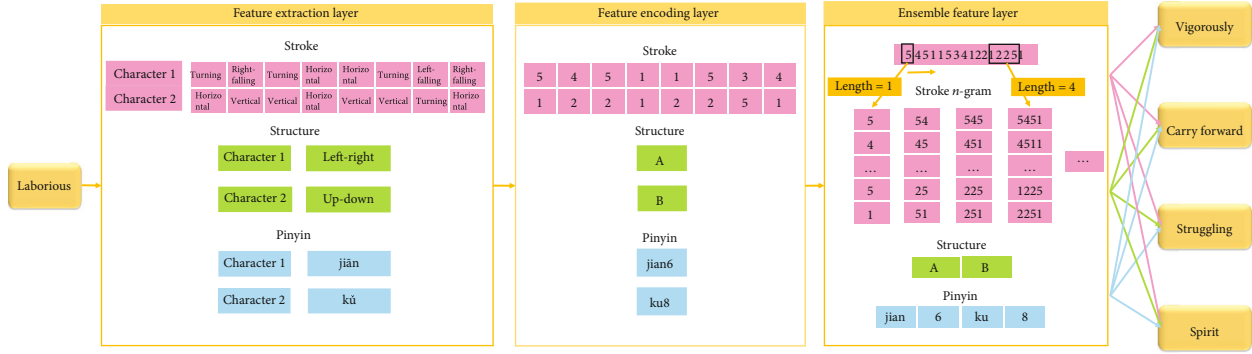


FIGURE 3: The architecture of ESSP2VEC. First, we decompose a Chinese word to characters and extract their stroke, structure, and pinyin features in the feature extraction layer. Second, the three features are encoded in the feature encoding layer. Third, we generate stroke n -gram and integrate stroke n -gram, structure, and pinyin features in the ensemble feature layer. Finally, the contextual words are predicted based on the ensemble features of the target word to learn the semantics of the target word.

of nonrepetitive words in the corpus. By training ESSP2VEC in the collected large-scale literature corpus, we can analyse the semantics of herbs and express them as semantic word vectors.

There also are word embedding algorithms designed for other languages. For example, Park et al. [46] proposed a Korean word embedding algorithm, which uses *jamo* feature of Korean characters to construct the *jamo* n -gram of Korean words and predict the contextual words of the target word based on its *jamo* n -gram. Korean characters can be decomposed into *jamos* in turn, which are the smallest lexicographic units representing the consonants and vowels [46]. The *jamo* feature is extracted to construct the *jamo* n -gram of Korean words, which is similar to the stroke n -gram of Chinese words. Then, the model predicts the contextual words of the target word based on its *jamo* n -gram to obtain final word embeddings. For English, Bojanowski et al. [47] proposed the FastText algorithm to capture the subword feature of English words to construct character n -gram and predict the contextual words of the target word based on its character n -gram. English words can be divided into 26 alphabets, which are the smallest component units of English words. Different character combinations can form different features, such as etyma, prefixes, and suffixes, which contain part semantics of words [47]. The subword feature is extracted to construct the character n -gram of English words, which is similar to the stroke n -gram of Chinese words. For example, the 3-grams of the word *where* are $\langle wh, whe, her, ere, re \rangle$ [47]. Then, FastText predicts the contextual words of the target word based on its n -gram to obtain final word embeddings.

The above three methods both generate the n -gram of one feature of the target word (i.e., the stroke n -gram of Chinese, the *jamo* n -gram of Korean, and the character n -gram of English) and predict the contexts of the target word based on its n -grams. For the proposed algorithm, ESSP2VEC not only constructs the stroke n -gram of Chinese words but also integrates the other two features (structure and pinyin) to analyse relatively comprehensive semantics of Chinese words. That is, ESSP2VEC considers both the morphological and phonetic features of Chinese words. Meanwhile, ESSP2VEC considers the similarity between the contextual

words and the internal features of the target word to conduct prediction.

3.2. Label Propagation-Based Algorithm. According to the theory of ESSP2VEC, we can analyse the semantics of herbs and obtain their semantic word vectors. However, how to use the semantic word vectors to find core herbs is a challenge. In order to discover core herbs for treating a certain disease by the semantic word vectors, we first compute the semantic similarity among herbs and construct a herbal semantic network, where herbs are considered as nodes and if the semantic similarity between two herbs is larger than the average value of all similarity among herbs (threshold value), edges are formed between the two herbs. Then, we adopt a label importance-based label propagation algorithm LILPA [48] to detect communities in the herbal semantic network, which can further improve problem (3). Herbs in a community may have the same or similar efficacy and can treat multiple syndromes of a certain disease. Finally, we identify important nodes in each community as core herbs for treating the syndromes of the disease. Here, we introduce LILPA briefly.

There are many real-world networks, such as social networks, collaboration networks, and herb networks, in which nodes represent objects and edges represent their relations [18]. Real-world networks often consist of subnetworks or communities with nodes more tightly linked with respect to the rest of the networks [3]. Community detection can be informally considered as a problem of finding such communities in networks, which aims at assigning community labels to nodes such that the nodes in the same community share higher similarity than the nodes in different communities [49] [50]. Communities in networks are the division of networks into the groups of nodes having dense intra-connections and sparse interconnections [51]. In other words, the connections among nodes in communities are dense, while the connections between communities are sparse. Thus, community detection focuses on discovering communities with dense connection nodes in networks. If the nodes own the same label when the algorithm ends, these nodes are allocated to the same community. Table 1

TABLE 1: Corresponding concepts.

Core herb discovery	LILPA
Herb	Node
The relations among herbs	Edge
Efficacy	Label
Herb group for treating multiple syndromes	Community
Core herbs for treating multiple syndromes	Nodes with a top- k degree in each community

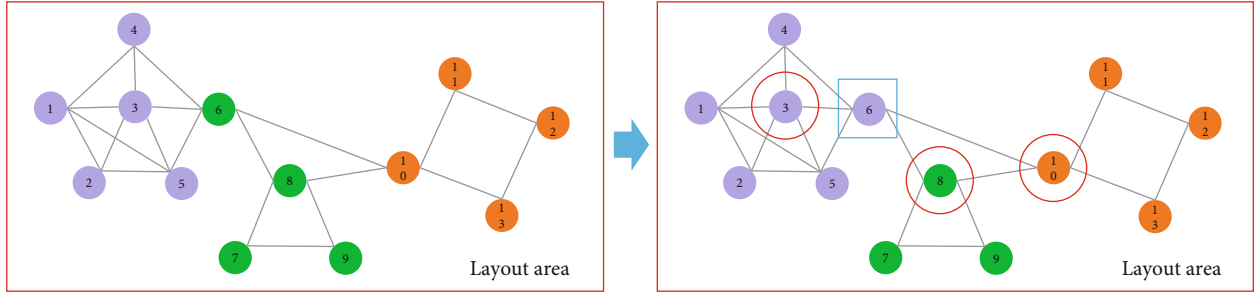


FIGURE 4: Example of label propagation for community detection.

shows the corresponding concepts between core herb discovery and LILPA.

An example is given in Figure 4 to explain the process of community detection based on label propagation. Node v_6 is chosen to update its labels firstly. Its neighbour nodes $v_3, v_4, v_5, v_8, v_{10}$ launch their own label with belonging coefficient to node v_6 (we assume that the belonging coefficient equals to 1). Then, node v_6 receives labels (purple, 1), (purple, 1), (purple, 1), (green, 1), and (orange, 1). By normalizing their belonging coefficients, we obtain node v_6 with labels (purple, 0.6), (green, 0.2), and (orange, 0.2). If the belonging coefficient is smaller than $1/R$ (we assume the filtering threshold $R = 2$), then the green and orange labels are filtered. Finally, the label of node v_6 is updated to the purple label, so node v_6 is assigned to purple community. Then, other nodes are chosen to update their labels. The above process is conducted continuously until the labels of nodes are kept unchanged. Finally, community detection is finished, and we can discover three communities in the example network. That is, nodes $v_1, v_2, v_3, v_4, v_5, v_6$ are assigned to a community; nodes v_7, v_8, v_9 are assigned to a community; and nodes $v_{10}, v_{11}, v_{12}, v_{13}$ are assigned to another community. We can find that intercommunal relations among communities are sparser than the connections within the communities. In each community, the nodes with a large degree (the number of neighbours) are considered as important nodes, such as v_3, v_8 , and v_{10} . In addition, nodes are drawn in a layout area.

Given an undirected and unweighted network $G = (V, E)$, where $V = \{v_1, \dots, v_i, \dots, v_{N_{\text{node}}}\}$ represents the set of nodes and $E = \{e_1, \dots, e_i, \dots, e_{M_{\text{edge}}}\}$ represents the set of edges. N_{node} and M_{edge} are the number of nodes and edges, respectively. The neighbour nodes of node v_i are expressed as $Z(v_i) = \{v_j | e_{v_i, v_j} \in E\}$, and its degree is expressed as k_{v_i} . The labels of node v_i are stored in $B(v_i) = \{(l_1^i, c_1^i), \dots, (l_j^i, c_j^i), \dots,$

$(l_H^i, c_H^i)\}$, where label l_j^i is the j th label with a belonging coefficient c_j^i of node v_i , $\sum_{j=1}^H c_j^i = 1$, and H is the number of labels in $B(v_i)$. The community characterized by label l is expressed as O^l . Nodes are drawn in a rectangle layout area with length L and width W . The position and displacement of node v_i in the layout are denoted as \vec{P}_{v_i} and \vec{D}_{v_i} , respectively.

In the above example, node v_6 is randomly chosen to update its labels. In order to fix the updating order of nodes to improve stability, node importance is defined to reflect the weight of nodes in networks as

$$I_{v_i} = C_{v_i} \times k_{v_i} + \sum_{v_j \in Z(v_i)} \frac{k_{v_j}}{\sum_{v_k \in Z(v_i)} k_{v_k}} \times C_{v_j} \times k_{v_j}, \quad (5)$$

where $C_{v_i} = (N_{\text{node}} - 1) / \sum_{v_j \in V} d_{v_i, v_j}$ is the closeness centrality of node v_i to measure its centrality in networks and d_{v_i, v_j} is the shortest distance between nodes v_i and v_j .

Communities are the clusters of nodes owning dense intraconnections and sparse external connections [49]. In order to increase the attraction among nodes to obtain dense internal connection, the node attraction between nodes v_i and v_j is defined as

$$F_{v_i, v_j}^A = \frac{x_{v_i, v_j}^2}{\sqrt{(W \times L) / N}}, \quad (6)$$

where x_{v_i, v_j} is the straight-line distance between nodes v_i and v_j in the layout area calculated by the positions of nodes in the layout area, which is different with d_{v_i, v_j} which is the shortest distance between nodes v_i and v_j calculated by the edge weight of networks.

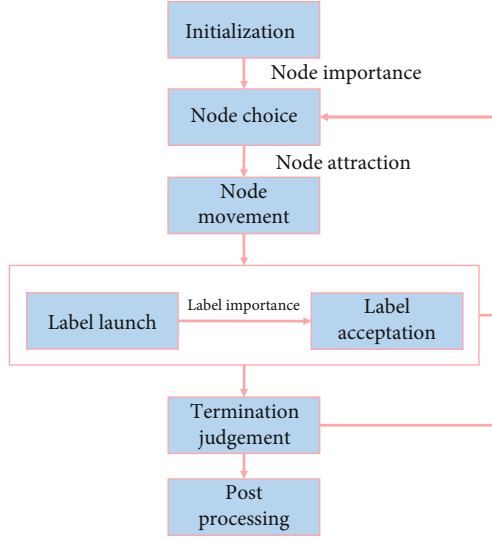


FIGURE 5: The process of LILPA. LILPA first initializes each node with a unique label, chooses nodes to update according to node importance, and moves nodes in the layout area according to node attraction. Then, the neighbour nodes of the updating node launch labels, and the updating node accepts labels according to label importance. The above steps except initialization are iteratively executed until all nodes are updated once. If LILPA reaches termination condition, then it goes to postprocessing, else, it returns to the step of node choice for the next iteration.

When label l with a belonging coefficient c is sent from node v_j to node v_i , the weight of this label is influenced by the node importance of sender, propagation distance (related to the node attraction among nodes), and its belonging coefficient [18]. Then, label importance is defined to measure the weight of labels of a node when they reach other nodes as

$$LP_{l,v_j \rightarrow v_i} = I_{v_i} \times c \times \sqrt{F_{v_i,v_j}^A}. \quad (7)$$

The processes of LILPA consist of initialization, node choice, node movement, label launch, label acceptance, termination judgement, and postprocess, as shown in Figure 5.

Step 1. Initialization. Nodes are allotted with labels (e.g., node's id) and random positions, then the node importance of all nodes is computed.

- (1) Set $S = V$, $B(v_i) = \{(l_1^i = i, c_1^i = 1)\}$, $\vec{P}_{v_i} = (x_i \in [-L/2, L/2], y_i \in [-W/2, W/2])$, and $\vec{D}_{v_i} = \vec{0}$ for $v_i \in V$, $r = 1$, and $t = 1$. Here, S represents the node set where nodes have not been updated
- (2) Node importance is calculated, then the nodes in S are ordered in ascending order of node importance.

Step 2. Node choice. Node v_i is chosen to update its labels, which satisfies $I_{v_i} = \min(I_{v_j} \mid v_j \in S)$, then set $B(v_i) = \emptyset$. Nodes with small importance can be influenced by nodes

with large importance easily [18], then the labels of nodes with small importance are preferentially updated.

Step 3. Node movement. Node v_i moves to a new position according to its displacement.

- (1) The displacement of node v_i is calculated by

$$\vec{D}_{v_i} = - \sum_{v_j \in Z(v_i)} \frac{\vec{P}_{v_i} - \vec{P}_{v_j}}{|\vec{P}_{v_i} - \vec{P}_{v_j}|} \times F_{v_i,v_j}^A + \sum_{v_j \in Z(v_i)} \frac{\vec{P}_{v_i} - \vec{P}_{v_j}}{|\vec{P}_{v_i} - \vec{P}_{v_j}|} \times F_{v_i,v_j}^R, \quad (8)$$

$$F_{v_i,v_j}^R = \frac{W \times L}{N \times x_{v_i,v_j}}$$

- (2) The position of the node is updated by

$$\vec{P}_{v_i} = \vec{P}_{v_i} + \frac{\vec{D}_{v_i}}{|\vec{D}_{v_i}|} \times \min\left(|\vec{D}_{v_i}|, \frac{\min(W, L)}{4}\right) \quad (9)$$

- (3) If node v_i is out of the layout area, then its position is restricted in the layout area by equations (10) and (11)

$$x_{v_i} = \min\left(\frac{L}{2}, \max\left(-\frac{L}{2}, x_{v_i}\right)\right), \quad (10)$$

$$y_{v_i} = \min\left(\frac{W}{2}, \max\left(-\frac{W}{2}, y_{v_i}\right)\right) \quad (11)$$

Step 4. Label launch. In this step, every node in the neighbouring nodes of node v_i sends its label with the maximal belonging coefficient to node v_i .

- (1) For node v_j in $Z(v_i)$, label l^j is chosen, which satisfies $c^{v_j} = \max(c^{v_k} \mid (l^k, c^{v_k}) \in B(v_j))$, then node v_j sends label l^j to node v_i
- (2) When label l^j reach node v_i , it is assigned with label importance calculated by equation (7), then $B(v_i) = B(v_i) \cup (l^j, LP_{l^j,v_j \rightarrow v_i})$
- (3) The label importance is added when the labels with the same id reach node v_i

Step 5. Label acceptance. This step is used to accept useful labels and filter the labels with small belonging coefficients.

- (1) By normalizing the label importance of labels in $B(v_i)$, $B(v_i) = \{(l_1^i, c_1^i), \dots, (l_j^i, c_j^i), \dots, (l_H^i, c_H^i)\}$, $c_j^i = LP_{l_j^i} / \sum_{k=1}^H LP_{l_k^i}$

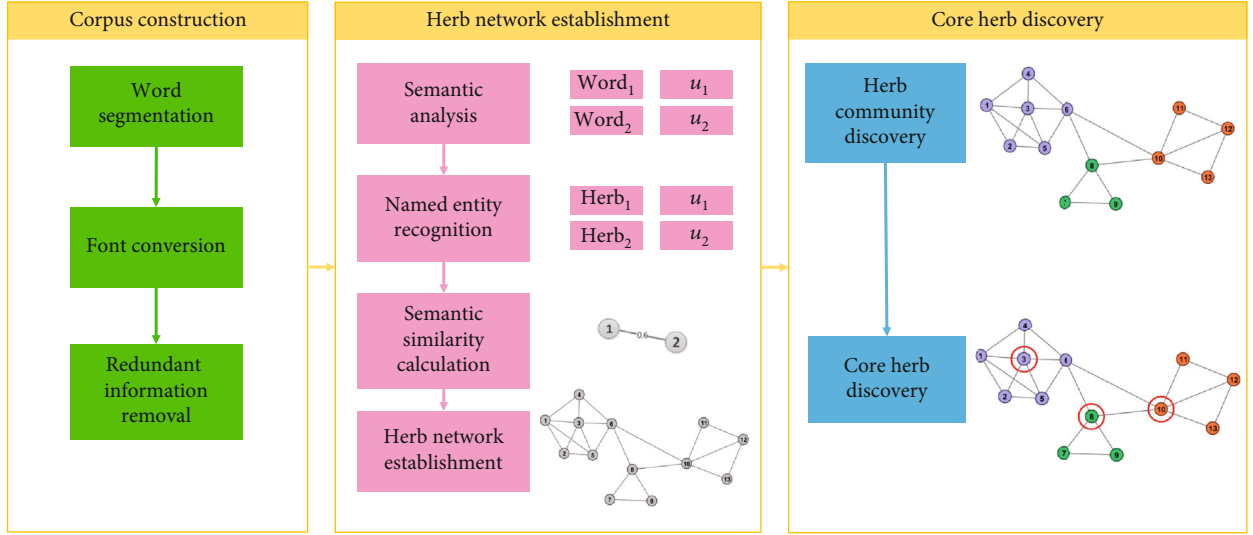


FIGURE 6: The process of CHDSC.

- (2) For label (l_j^i, c_j^i) in $B(v_i)$, if $c_j^i < 1/R$, then $B(v_i) = B(v_i) - (l_j^i, c_j^i)$. The updated $B(v_i)$ of node v_i is gained by normalizing again, then the updating of labels of node v_i is finished. Here, R is the filtering threshold
- (3) If $r = N_{\text{node}}$, then step 6 is executed, else the method sets $r = r + 1$, $S = S - \{v_i\}$ and returns to Step 2 to update other nodes.

Step 6. Termination judgement. The minimal number set m_t of nodes signed by each community identifier is computed. If $m_t = m_{t-1}$ or LILPA reaches the maximum number of iterations, LILPA goes to Step 7 for postprocessing, else sets $S = V$, $\vec{D}_{v_i} = \vec{0}$, $r = 1$, $t = t + 1$ and returns to Step 2 for the next iteration.

Step 7. Postprocessing. Nodes with label l are allocated to community O^l . If nodes have multiple labels, then they are assigned to multiple communities.

We apply LILPA to herbal semantic network and discover herb community set $O = \{O^1, \dots, O^i, \dots, O^k\}$, where k is the number of herb communities. In each community O^i , herbs have the same or similar efficacy for treating multiple syndromes of a certain disease. Then, we can discover core herbs for treating the syndromes by choosing nodes with large degree in community O^i .

4. The Proposed Model

In this paper, we aim to import herb knowledge implied in large-scale literature into core herb discovery. Thus, we propose CHDSC to analyse the semantics of herbs in literature based on semantic analysis module ESSP2VEC, calculate the semantic similarity among herbs to build herbal semantic network, and discover herb communities and core herbs in

the network based on community detection module LILPA. CHDSC includes three stages: corpus construction, herb network establishment, and core herb discovery, whose process is shown in Figure 6.

Before applying CHDSC to discover core herbs for treating a certain disease, we should choose a target disease; here, we denote the target disease as T . After discussing with TCM experts, we select keywords in Chinese about the TCM treatment of disease T to search scientific literature in CNKI.

4.1. Corpus Construction. In this stage, corpus C about the TCM treatment of disease T is built by preprocessing the collected literature, which is used to train ESSP2VEC for analysing the semantics of herbs in literature.

Step 1. Word segmentation. Different from English sentences that use space as the natural interval among words, Chinese sentences are made up of continuous words. In order to analyse the semantics of Chinese words in literature, in this paper, Chinese sentences of the full text of literature are divided into Chinese words.

Step 2. Font conversion. Since traditional Chinese characters may exist in the literature, we convert them into simplified Chinese characters to make uniform the process.

Step 3. Redundant information removal. This step is to remove messy code, punctuations, and English abstract to obtain the pure corpus C , whose number of words is N_{word} .

4.2. Herb Network Establishment. In this stage, herbal semantic network G is constructed by extracting the semantic word vectors of herbs and calculating their semantic similarity to reflect the relations between herbs and the target disease.

Step 1. Semantic analysis. Corpus C is input into ESSP2VEC to analyse the semantics of words in literature. Then, we obtain the semantic word vectors U .

```

Input: the collected literature, standard herb name dictionary  $D$ , the size of context windows  $c = 5$ , filtering threshold  $R$ ;
Output: core herb set  $D^{\text{core}}$ ;
Stage 1 Corpus construction
 $C_1 = \text{Word-segmentation}()$ ;
 $C_2 = \text{Font-conversion}(C_1)$ ;
 $C = \text{Redundant-information-removal}(C_2)$ ;
Stage 2 Herb network establishment
Step 1 Semantic analysis
 $U = \text{ESSP2VEC}(C, c)$ ;
Step 2 Name entity recognition
 $X = \emptyset, U_X = \emptyset$ ;
For each word  $w_t$  in  $C$ 
  If  $w_t \in D$ 
     $X = X \cup \{w_t\}$  and  $U_X = U_X \cup \{u_t\}$ ;
  End For
Step 3 Semantic similarity calculation
 $\forall w_i, w_j \in X, i \neq j$ 
  Calculate  $Q(w_i, w_j)$  by equation (12);
Step 4 Herb network establishment
 $V = X, E = \emptyset$ ;
For each herb  $w_i$  in  $X$ 
  If  $Q(w_i, w_j) \geq \sum_{j=1}^{|X|} Q(w_i, w_j) / |X|$ 
     $E = E \cup \{e_{w_i, w_j}\}$ ;
  End For
Stage 3 Core herb discovery
Step1 Herb community discovery
 $O = \text{LILPA}(G)$ ;
Step 2 Core herb discovery
For each community  $O^i$  in  $O$ 
   $D_i^{\text{core}} = \text{herbs represented by the nodes having top-8 degree in } O^i$ ;
   $D^{\text{core}} = D^{\text{core}} \cup \{D_i^{\text{core}}\}$ ;
End For
Return  $D^{\text{core}}$ ;

```

ALGORITHM 1: CHDSC.

Step 2. Name entity recognition. All Chinese words in the corpus including symptoms, syndromes, diseases, herbs, and other words are used to train word embedding because the semantics of herbs are contained in the contexts of words [38]. Then, the results contain the semantic word vectors of symptoms, syndromes, diseases, herbs, and other words. In this step, the semantic word vectors U_X of herbs is extracted from U by name entity recognition, where X represents the herbs existing in the collected literature. We construct a standard herbal name dictionary D according to the regulated herb name in *The Pharmacopoeia of the People's Republic of China* [52]. If herbs exist in the corpus and the standard herb thesaurus simultaneously, then we extract the herbs and their semantic word vectors.

Step 3. Semantic similarity calculation. Here, we adopt cosine similarity [53] to measure the semantic similarity among herbs, which is defined as

$$Q(w_i, w_j) = \frac{u_i \cdot u_j}{|u_i| |u_j|}. \quad (12)$$

If the semantic similarity among herbs is greater than the average value of all similarities among herbs, we consider that they own similar efficacy and can treat some syndromes of a disease.

Step 4. Herb network establishment. Herb semantic network is constructed by herbs and their semantic similarity. The herbs form nodes in the network, and if the similarity of two herbs is greater than the average value of all similarities among herbs, then an edge is formed between the two herbal nodes.

4.3. Core Herb Discovery. In this stage, core herb set $D^{\text{core}} = \{D_1^{\text{core}}, \dots, D_i^{\text{core}}, \dots, D_k^{\text{core}}\}$ is discovered in herb community $O = \{O^1, \dots, O^i, \dots, O^k\}$.

Step 1. Herb community discovery. Herbs in herb communities own the same or similar efficacy to treat multiple syndromes of a disease. Herb communities are revealed by LILPA.

TABLE 2: Description of training and evaluation dataset.

Function	Dataset	Reference	Task	Scale
Training	SogouCA	[40]	—	300 million words
	WA-1124	[42]	Word analogy	1124 instances
Evaluation	WS-240	[39]	Word similarity	240 instances
	WS-296	[39]	Word similarity	296 instances

Step 2. Core herb discovery. In each community, nodes are important if they have a large degree. We choose eight herbs with top-8 degree in each community as core herbs.

4.4. Complexity Analysis. The complexity of CHDSC is mainly in the two artificial intelligence modules. Here, we briefly analyse the time complexity of ESSP2VEC and LILPA.

4.4.1. The Time Complexity of ESSP2VEC. The contextual words are predicted based on each word taking time $O(cN_{\text{word}})$. We adopt an optimal strategy, negative sampling [45], which considers the target word and its contextual words as positive sample pairs and takes the target word and random words as negative sample pairs, whose number is N_{neg} . Then, the problem of predicting contextual words can be replaced as a set of independent binary classification tasks so as to independently predict the presence (or absence) of contextual words [39] [40]. Then, equation (3) can be rewritten as

$$p(w_c | w_t) = \frac{e^{s(w_t, w_c)}}{\sum_{j=1}^{c+N_{\text{neg}}} e^{s(w_t, w_j)}}. \quad (13)$$

Thus, the complexity of predicting the contextual words of a word can be reduced to $O(c(c + N_{\text{neg}}))$. Then, predicting the contextual words of all words costs time $O(c(c + N_{\text{neg}})N_{\text{word}})$. For ESSP2VEC, we represent each word as its stroke n -gram with structure and pinyin features and predict the contextual words based on the ensemble features of the target word, then the total complexity is $O((L_{\text{max}}(L_{\text{max}} + 1)/2)c(c + N_{\text{neg}})N_{\text{word}})$, where L_{max} is the maximum length of stroke n -gram. In general, $c, N_{\text{neg}}, L_{\text{max}} \ll N_{\text{word}}$, then the total complexity is near $O(h_1 N_{\text{word}})$, where h_1 is a constant.

4.4.2. The Time Complexity of LILPA. The time complexity of LILPA is estimated as follows.

- (1) *Initialization:* the shortest distances among nodes are calculated with time $O(N_{\text{node}} \log N_{\text{node}})$. The Quick-sort algorithm is adopted for sorting nodes by node importance with time $O(N_{\text{node}} \log N_{\text{node}})$. Thus, initialization costs time $O(N_{\text{node}} \log N_{\text{node}})$
- (2) *Node choice:* choosing a node to update its label costs constant time
- (3) *Node movement:* calculating the attractive and repulsive forces between node v_i and its neighbours and the displacement of node v_i takes time $O(|N(v_i)|)$

TABLE 3: Description of real-world networks.

Network	N_{node}	M_{edge}	N_{com}	$\langle k \rangle$	Dia	Reference
Karate	34	78	2	4.588	5	[55]
Dolphins	62	159	2	5.129	8	[56]
Football	115	615	12	10.661	4	[57]
Netscience	1589	2742	16	3.451	17	[58]
Power	4941	6594	—	2.669	46	[59]
PGP	10680	24316	—	4.554	24	[60]
Cond2003	31163	120029	—	7.703	16	[61]
Cond2005	40421	175693	—	8.693	18	[61]

N_{com} : the number of communities; $\langle k \rangle$: the average degree of networks; dia: the diameter of networks.

- (4) *Label launch:* the neighbours of node v_i cost the worst time $O(|N(v_i)| n_1 \log n_1)$ to send their labels, where n_1 is the maximum number of labels of the neighbours of node v_i . In general, $n_1 \ll N_{\text{node}}$, then label launch needs constant time
- (5) *Label acceptance:* accepting the labels of node v_i takes $O(n_2)$, where n_2 is the number of labels reaching node v_i . In general, $n_2 \ll N_{\text{node}}$, then label acceptance takes constant time
- (6) *Termination judgement and postprocessing:* the same as COPRA [54], the former costs time $O(\beta N_{\text{node}})$ and the latter needs time $O((\beta^3 + 1)N_{\text{node}} + \beta(N_{\text{node}} + M_{\text{edge}}))$

For the label update process of node v_i , Steps 2–5 need constant time. Thus, updating the labels of N_{node} nodes in one iteration needs time $O(N_{\text{node}})$. Thus, the time complexity of LILPA is $O(N_{\text{node}} \log N_{\text{node}} + (\beta^3 + 2\beta + t + 1)N_{\text{node}} + \beta M_{\text{edge}})$. In general, $\beta, t \ll N_{\text{node}}, M_{\text{edge}}$, then the total complexity is near $O(N_{\text{node}} \log N_{\text{node}} + h_2 N_{\text{node}} + h_3 M_{\text{edge}})$, where h_2 and h_3 are constants.

5. Experiment Setup

In this section, we first introduce datasets, evaluation criteria, and comparison algorithms, which are used to evaluate the performance of artificial intelligence modules. Then, we take a case study by choosing CGN as the target disease and apply CHDSC to discover the core herbs for treating multiple syndromes of CGN in TCM.

TABLE 4: Results of word analogy and word similarity tasks.

Algorithm	Word analogy (%)	Word similarity (%)		Average rank
	WA-1124	WS-240	WS-296	
CBOW	22.77 (7)	46.40 (8)	56.26 (7)	7.33
Skip-Gram	58.45 (3)	55.36 (2)	60.76 (4)	3.00
Glove	19.39 (8)	48.36 (7)	47.02 (8)	7.67
CWE	47.69 (6)	51.67 (5)	61.17 (3)	4.67
JWE	57.65 (4)	51.00 (6)	60.22 (6)	5.33
GWE	48.84 (5)	53.45 (4)	60.63 (5)	4.67
CW2VEC	63.17 (2)	54.85 (3)	61.41 (2)	2.33
ESSP2VEC	64.85 (1)	55.38 (1)	61.71 (1)	1.00

TABLE 5: Average value of NMI.

Algorithm	Karate	Dolphins	Football	Netscience	Average rank
COPRA	0.3596 (8)	0.5976 (6)	0.8836 (7)	0.3566 (5)	6.5000
SLPA	0.6915 (3)	0.6678 (3)	0.8862 (6)	0.3651 (2)	3.5000
DLPA ⁺	0.5489 (5)	0.4753 (8)	0.9044 (2)	0.3858 (1)	4.0000
WLPA	0.5016 (6)	0.6599 (4)	0.9013 (3)	0.3350 (8)	5.2500
LPA_NI	0.6598 (4)	0.6436 (5)	0.8823 (8)	0.3636 (3)	5.0000
NGLPA	0.4408 (7)	0.7108 (2)	0.8887 (5)	0.3471 (7)	5.2500
LPANNI	0.7782 (2)	0.5809 (7)	0.8997 (4)	0.3627 (4)	4.2500
LILPA	0.9855 (1)	0.8125 (1)	0.9079 (1)	0.3526 (6)	2.2500

5.1. Data Description. For evaluating the effectiveness of the semantic analysis module (i.e., word embedding algorithm ESSP2VEC), we employ a universal data SogouCA shown in Table 2, which contains 300 million words after preprocessing by the same operation of corpus construction to train ESSP2VEC to obtain word semantic vectors. Then, we use datasets (1) WA-1124, (2) WS-240, and (3) WS-296 to evaluate the proposed module on word analogy and word similarity tasks, respectively, as described in Section 5.2. For estimating the performance of community detection module (i.e., label propagation algorithm LILPA), we use eight real-world networks shown in Table 3. Each algorithm independently runs 50 times.

5.2. Evaluation Criteria. In order to evaluate the quality of semantic word vectors obtained by word embedding algorithms, we test them on word analogy and word similarity tasks.

- (i) Word analogy task is used to measure the model ability of exploring the semantic relations among words [42] [45]. Given three words w_1 , w_2 , and w_3 , the word embedding models judge word w_4 that correctly answers the question “ w_1 to w_2 is w_3 to what?” For example, there is a question “*Beijing* is to *China* as *Berlin* is to what?” such that the cosine similarity between vectors $(v_{w_2} - v_{w_1} + v_{w_3})$ and v_{w_4} is maximized. By correctly answering this question, such as *Germany*, the models are considered that they can capture semantic relationships among words. We

adopt the test data WA-1124 with 1124 instances for evaluating Chinese word semantic vectors [42]

- (ii) Word similarity task is designed to evaluate the model ability of capturing semantic relatedness and closeness among words [39] [40]. Word similarity is measured by the cosine similarity between the corresponding word vectors, then we calculate the Spearman correlation coefficient between the word similarity and the human similarity scores to estimate the quality of word vectors. We adopt two datasets WS-240 and WS-296 for evaluation [39]

In order to measure the quality of detected communities in networks, we use two criteria Normalized Mutual Information (NMI) and Overlap Modularity (OM). If the true communities of real-world networks are known, the two criteria are both adopted; otherwise, only OM is adopted.

- (i) NMI is used to compute the difference between the communities detected by algorithms and true community structures and varies between 0 and 1 [62]. The larger the value, the smaller the difference
- (ii) OM reflects the quality of divisions assessed by the relative density of edges within communities and between communities [63], which varies between 0 and 1. The larger the value, the better the quality

5.3. Comparison Algorithms. To evaluate the effectiveness of ESSP2VEC, we compare it with seven word embedding

TABLE 6: Average value of OM.

Algorithm	Karate	Dolphins	Football	Netscience	Power	PGP	Cond2003	Cond2005	Average rank
COPRA	0.2348 (8)	0.3741 (7)	0.5972 (6)	0.8784 (7)	0.1696 (8)	0.5117 (8)	0.6306 (5)	0.4256 (8)	7.1250
SLPA	0.3742 (5)	0.4757 (5)	0.6016 (3)	0.9043 (6)	0.6225 (6)	0.7641 (4)	0.6341 (2)	0.6019 (5)	4.5000
DLPA ⁺	0.4210 (2)	0.5166 (3)	0.5960 (7)	0.8456 (8)	0.5993 (7)	0.6761 (6)	0.4764 (8)	0.4371 (7)	6.0000
WLPA	0.3682 (6)	0.3695 (8)	0.5981 (5)	0.9279 (2)	0.7731 (2)	0.6231 (7)	0.5959 (6)	0.6117 (3)	4.8750
LPA_NI	0.4136 (4)	0.5055 (4)	0.5985 (4)	0.9140 (4)	0.7473 (4)	0.7861 (3)	0.6313 (3)	0.6111 (4)	3.7500
NGLPA	0.3314 (7)	0.5189 (2)	0.5848 (8)	0.9209 (3)	0.7631 (3)	<i>0.8092 (1)</i>	0.5907 (7)	0.4431 (6)	4.6250
LPANNI	0.4147 (3)	<i>0.5423 (1)</i>	<i>0.6090 (1)</i>	0.9070 (5)	0.6608 (5)	0.7575 (5)	0.6312 (4)	0.6175 (2)	3.6250
LILPA	<i>0.4213 (1)</i>	0.4003 (6)	0.6061 (2)	<i>0.9319 (1)</i>	<i>0.7817 (1)</i>	0.8001 (2)	<i>0.6852 (1)</i>	<i>0.6223 (1)</i>	<i>1.8750</i>

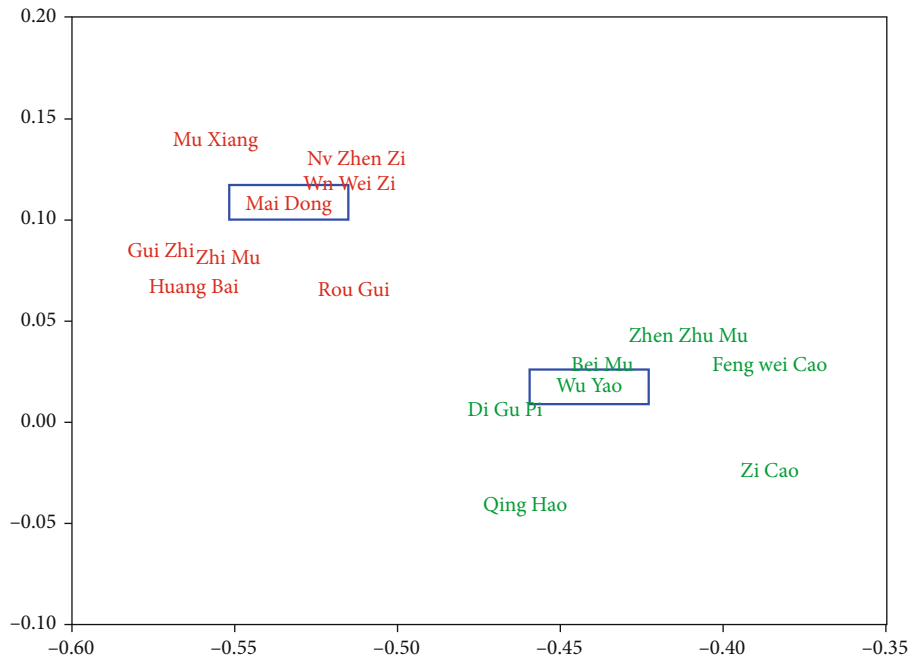


FIGURE 7: Example of semantic word vectors of herbs.

algorithms, including (1) three general word embedding algorithms CBOW [45], Skip-Gram [45], and GloVe [64], which can be used for any languages, and (2) four Chinese word embedding algorithms CWE [42], JWE [44], GWE [65], and CW2VEC [39], which are designed for the Chinese language and consider the radical, component, character, and stroke n -gram features, respectively. For baselines, we set the size of the contextual window equalling to ESSP2VEC.

To show that LILPA can find better communities, we compare it with seven label propagation-based community detection algorithms COPRA [54], SLPA [66], DLPA⁺ [67], WLPA [68], LPA_NI [69], NGLPA [70], and LPANNI [49]. In this paper, we use the given parameters for baselines if real-world networks are used in the original articles. Otherwise, we utilize the ways proposed in the original articles to gain the best solution.

5.4. Case Study. In order to further validate the effectiveness of core herb discovery model CHDSC, we choose CGN as

the target disease to conduct a case study. After discussing with TCM experts, we select keyword pairs in Chinese (1) “chronic glomerulonephritis” and “Chinese medicine” and (2) “chronic glomerulonephritis” and “Chinese native medicine,” to search the scientific literature in CNKI. Then, we apply CHDSC to analyse the collected literature to discover the core herbs for treating different syndromes of CGN.

6. Results and Discussion

The results for word analogy and word similarity tasks are shown in Table 4. The average values of NMI and OM for real-world networks are shown in Tables 5 and 6, respectively. We mark the optimal values in italics. The number in brackets is the rank of methods for each task or network, and the average rank of each algorithm is shown in the last column. Finally, we choose CGN as the target disease to conduct a case study.

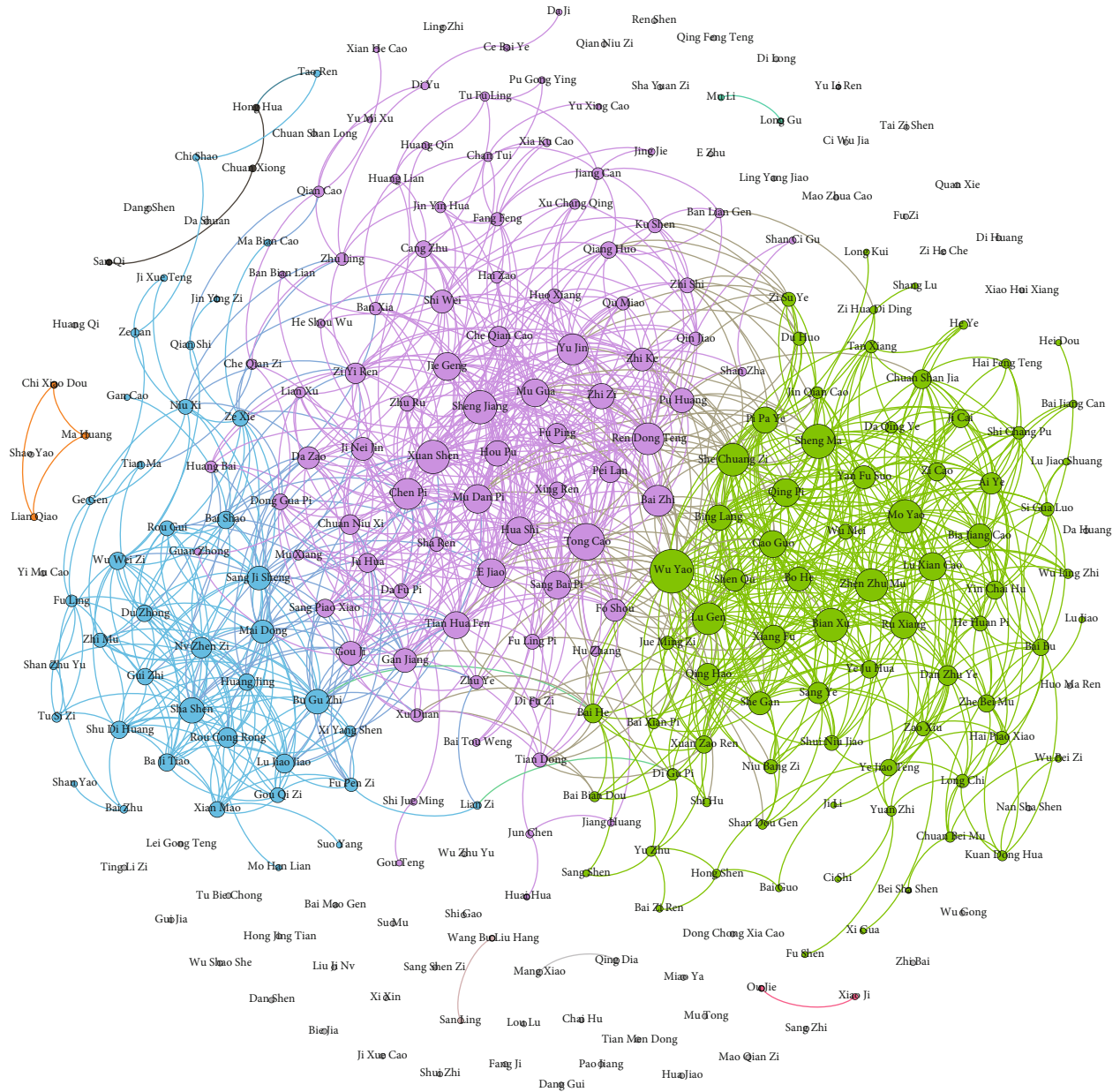
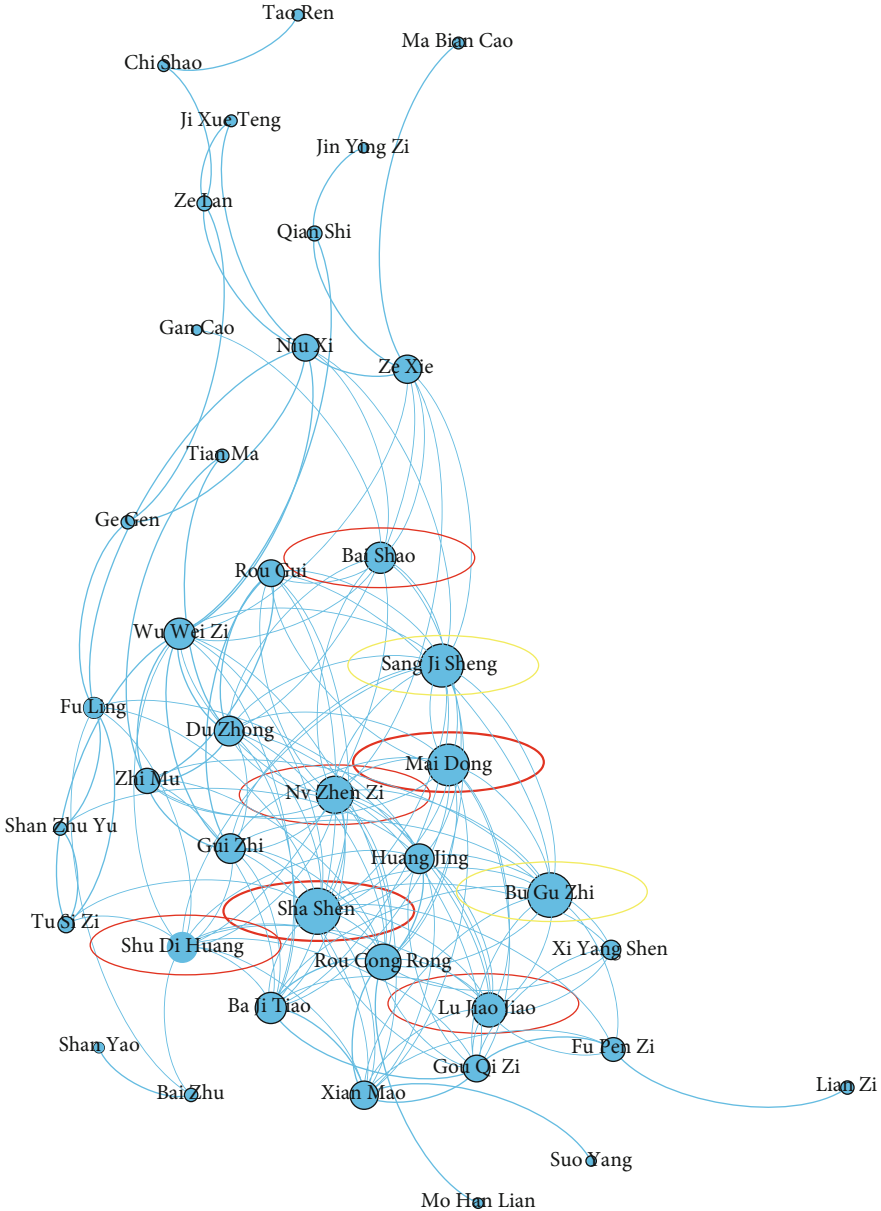


FIGURE 8: Results of herb communities.

6.1. *Results of Word Embedding Algorithm.* As shown in Table 4, we can find that ESSP2VEC obtains the best result in all tasks. For word analogy task, CBOW and Glove achieve about 20% accuracy, CWE and GWE obtain over 40% accuracy, Skip-Gram and JWE gain over 50% accuracy, and the accuracy of CW2VEC and ESSP2VEC is over 60%. In general, the proposed algorithm ESSP2VEC outperforms the best baseline CW2VEC. For word similarity task in terms of WS-240, CBOW and Glove gain over 40% accuracy and other algorithms achieve over 50% accuracy. ESSP2VEC outstrips the best baseline Skip-Gram. For word similarity in terms of WS-296, the accuracy of CBOW and Glove is under 60%; on the contrary, other algorithms obtain over 60% accuracy. ESSP2VEC outperforms the best baseline CW2VEC.

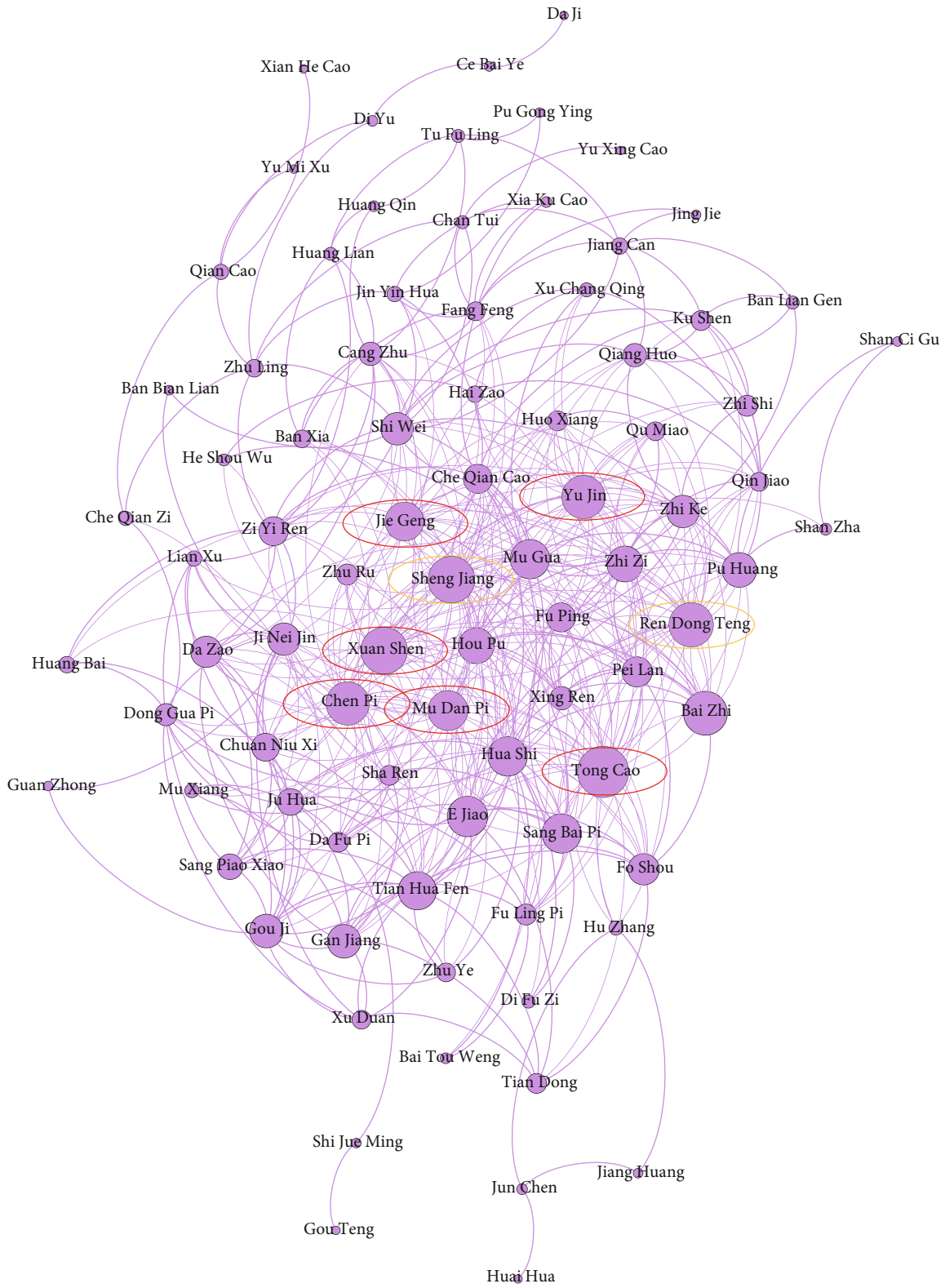
Thanks to the ideas of using the target word to predict its contexts and the effectiveness of integrating the stroke, structure, and pinyin features of Chinese characters, ESSP2VEC obtains the best average rank on word analogy and word similarity tasks. Comparing with state-of-the-art word embedding algorithms, we can consider that the proposed algorithm ESSP2VEC can obtain good accuracy and analyse the semantics of herbs in the literature.

6.2. *Results of Label Propagation-Based Algorithm.* As shown in Table 5, we can find that LILPA obtains the best NMI for the Karate, Dolphins, and Football networks and achieves the best average rank, which illustrates that LILPA can discover communities close to the true ones. In particular, LILPA



(a) Blue community

FIGURE 9: Continued.



(b) Purple community

FIGURE 9: Continued.

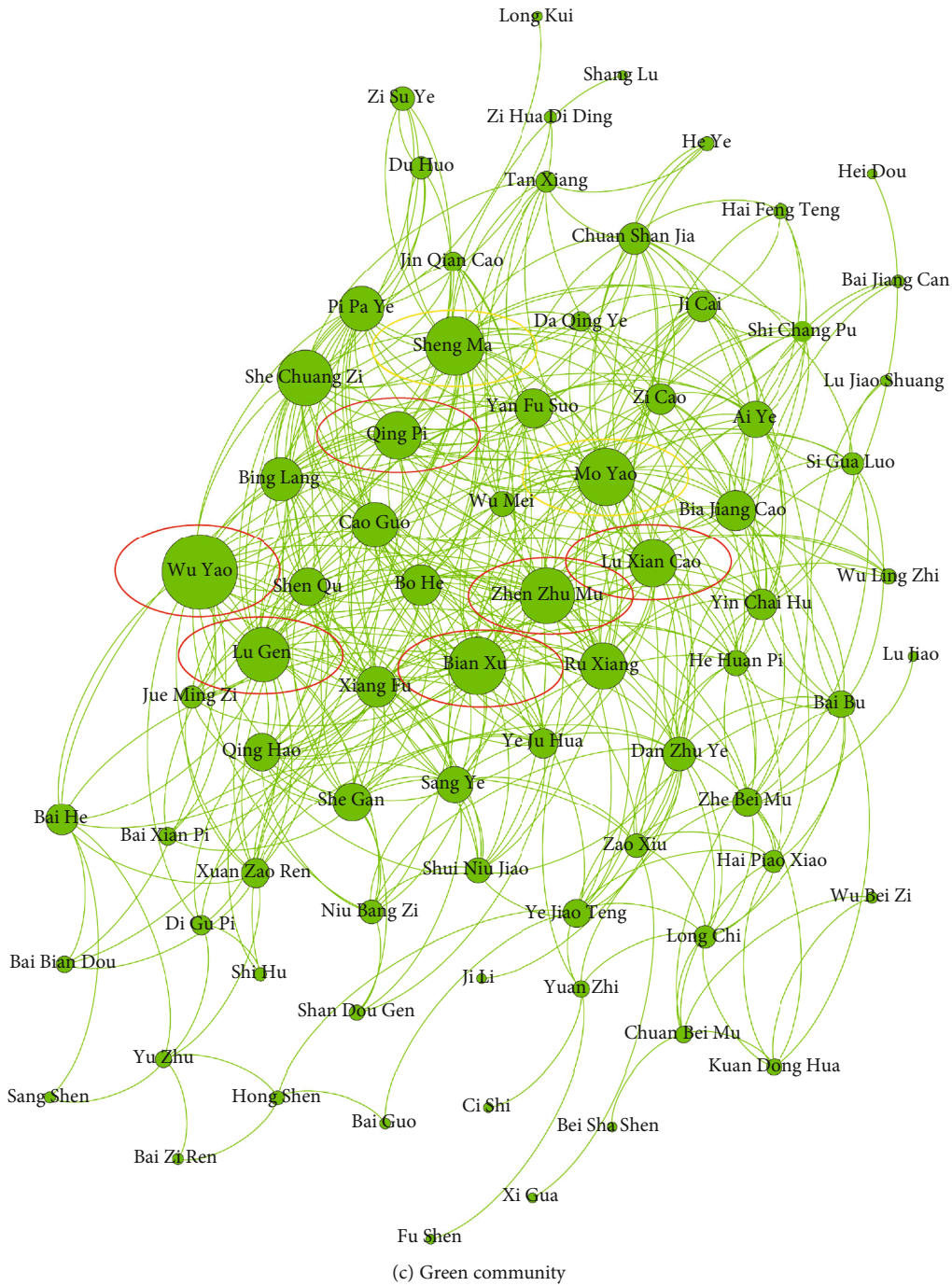


FIGURE 9: Results of core herbs in each community. Herbs in red circles are the core herbs identified correctly for treating multiple syndromes of CGN. Herbs in yellow circles are complementary herbs.

outperforms the best baseline LPANNI over 20.73% for the Karate network and outstrips the best baseline NGLPA over 10.17% for the Dolphins network. Although LILPA obtains poor NMI than some algorithms for the Netscience network, the difference with the optimal value is small. As shown in Table 6, LILPA gains the first rank in five networks and the second rank in two networks, then it achieves the best average rank. LILPA gets poor OM in the Dolphins network, while it obtains the best NMI in this network. With the

increase of network scale, LILPA keeps good performance. LILPA can find better communities in different scale networks than other baselines. In general, according to the average rank, LILPA outperforms baselines in terms of NMI and OM, which is profited by node importance, node attraction, and label importance. Compared with state-of-the-art label propagation-based algorithms, we can infer that LILPA can discover good communities and can detect high-quality herb communities in the herbal semantic network.

TABLE 7: Top-8 herbs in blue community.

Herb (Chinese pinyin)	Herb (English name)	Degree	Closeness centrality
<i>Sha Shen</i>	<i>Coastal glehnia root</i>	26	0.35
Bu Gu Zhi	Malaytea scurfpea fruit	25	0.33
Sang Ji Sheng	Chinese taxillus herb	23	0.35
<i>Mai Dong</i>	<i>Dwarf lilyturf tuber</i>	22	0.34
<i>Nv Zhen Zi</i>	<i>Glossy privet fruit</i>	20	0.30
<i>Lu Jiao Jiao</i>	<i>Deerhorn glue</i>	17	0.31
<i>Shu Di Huang</i>	<i>Prepared rehmannia root</i>	15	0.28
<i>Bai Shao</i>	<i>Debark peony root</i>	15	0.34

TABLE 8: Top-8 herbs in purple community.

Herb (Chinese pinyin)	Herb (English name)	Degree	Closeness centrality
<i>Tong Cao</i>	<i>Ricepaperplant pith</i>	39	0.40
<i>Xuan Shen</i>	<i>Figwort root</i>	35	0.39
Sheng Jiang	Fresh ginger	34	0.38
Ren Dong Teng	Honeysuckle stem	33	0.41
<i>Chen Pi</i>	<i>Dried tangerine peel</i>	32	0.39
<i>Yu Jin</i>	<i>Turmeric root tuber</i>	32	0.40
<i>Mu Dan Pi</i>	<i>Tree peony root bark</i>	29	0.41
<i>Jie Geng</i>	<i>Platycodon root</i>	28	0.37

6.3. *Results of the Application of CHDSC on CGN.* In this section, we choose CGN as the target disease T for the reason mentioned in Section 1. According to the above experiments, we can consider that CHDSC with ESSP2VEC and LILPA can discover core herbs accurately. Then, we apply CHDSC to discover core herbs for CGN treatment in TCM. After searching the literature in CNKI, we collect 449 samples of literature by keywords “chronic glomerulonephritis” and “Chinese medicine” and 677 samples of literature by keywords “chronic glomerulonephritis” and “Chinese native medicine.”

After corpus construction, we obtain CGN corpus containing 1126 samples of literature with 0.8 million words. All articles are related to the TCM treatment of CGN, so we expect semantic analysis can obtain high-quality semantic word vectors of herbs, since a pure in-domain corpus yields better performance than a mixed-domain corpus [71].

After herb network establishment, we obtain the semantic word vectors of 274 herbs and build a herbal semantic network with 274 nodes and 1293 edges. Some nodes have no edges with others because these herbs may have small similarity with other herbs. In order to understand the semantic word vectors intuitively, we choose two herbs *dwarf lilyturf tuber* (Mai Dong) and *combined spicebush root* (Wu Yao), discover the herbs owing large semantic similarity with one of them, and visualize these herbs in a two-dimensional surface. As shown in Figure 7, *dwarf lilyturf tuber* (Mai Dong) and some herbs are clustered together (denoted as O^1 with red color), and *combined spicebush root* (Wu Yao) and some herbs are also gathered together (denoted as O^2 with green color). Meanwhile, we can observe that groups

O^1 and O^2 have obvious interval, then we can infer that the semantic word vectors of *dwarf lilyturf tuber* and *combined spicebush root* can reflect their characteristics to find similar herbs. CHDSC can capture the semantics of herbs in the literature to a certain extent and generate effective semantic word vectors.

After core herb discovery, CHDSC discovers three large herb communities in herbal semantic network as shown in Figure 8. The herbs in the same community own similar efficacy and can treat multiple syndromes of CGN. According to the analysis of TCM experts, the herbs in the blue community have the efficacies of nourishing the liver and kidney and nourishing yin and blood, which can be mainly used for treating the syndrome of deficiency of both qi and yin and the syndrome of yin deficiency of the liver and kidney. The herbs in the purple community have the efficacies of removing dampness and diuresis, clearing heat and removing toxicity and dispelling wind evil and are often used for treating the syndrome of yang deficiency of the spleen and kidney. Meanwhile, they can be used to treat the syndromes of dampness-heat and fluid-dampness. The herbs in the green community have the efficacies of activating qi and eliminating dampness, clearing heat and removing toxicity, and resolving masses, which are used to treat the syndrome of liver depression and qi stagnation. According to the pathogenesis of CGN in TCM (intermingled deficiency and excess) and the TCM treatment points for CGN (supple deficiency and expel excess and strengthening vital qi to eliminate pathogenic factor) [36, 37], we find that the herbs in the blue community are mainly used for supplying deficiency and

TABLE 9: Top-8 herbs in green community.

Herb (Chinese pinyin)	Herb (English name)	Degree	Closeness centrality
<i>Wu Yao</i>	<i>Combined spicebush root</i>	48	0.42
Sheng Ma	Large-trifoliate bugbane rhizome	36	0.38
<i>Bian Xu</i>	<i>Common knotgrass herb</i>	35	0.32
Mo Yao	Myrrh	35	0.34
<i>Zhen Zhu Mu</i>	<i>Nacre</i>	34	0.34
<i>Lu Gen</i>	<i>Reed rhizome</i>	33	0.38
<i>Lu Xian Cao</i>	<i>Pyrola herb</i>	23	0.31
<i>Qing Pi</i>	<i>Immature tangerine peel</i>	23	0.38

the herbs in the purple and green communities are mainly used for expelling excess. Thus, the herbs in the blue community are necessary for treating CGN in TCM, and the ones in the purple and green communities are used to treat the secondary symptoms. CHDSC discovers herb communities where herbs can treat most primary syndromes of CGN; however, herbs in herb communities do not cover all syndromes of CGN, which may be because some syndromes are less recorded in the literature and the scale of the literature is limited.

The herbs represented by the nodes with the top-8 degree in each community are regarded as core herbs for treating multiple syndromes of CGN as shown in Figure 9 (their Chinese pinyin and English name are shown in Tables 7–9). According to the analysis of TCM experts, for the herbs with the top-8 degree, the herbs in red circles are the core herbs identified correctly for treating the CGN syndromes represented by corresponding herb communities and the herbs in yellow circles are complementary herbs (the core herbs identified correctly are indicated in italics in Tables 7–9). It is seen that CHDSC can discover core herbs for treating most syndromes of CGN with high accuracy from large-scale literature, which can give references for the clinical application of herbs. Thus, we can consider that CHDSC can automatically discover core herbs for treating a disease from large-scale literature. The herbs in red circles are core herbs and can be used to treat main symptoms of CGN; the herbs in yellow circles are used to play support efficacy according to the symptoms of patients because patients may suffer from other diseases and need to be treated at the same time.

In order to further explore the herbal semantic network, we analyse its community size, degree, and closeness centrality distributions to mine the rules of CGN core herbs.

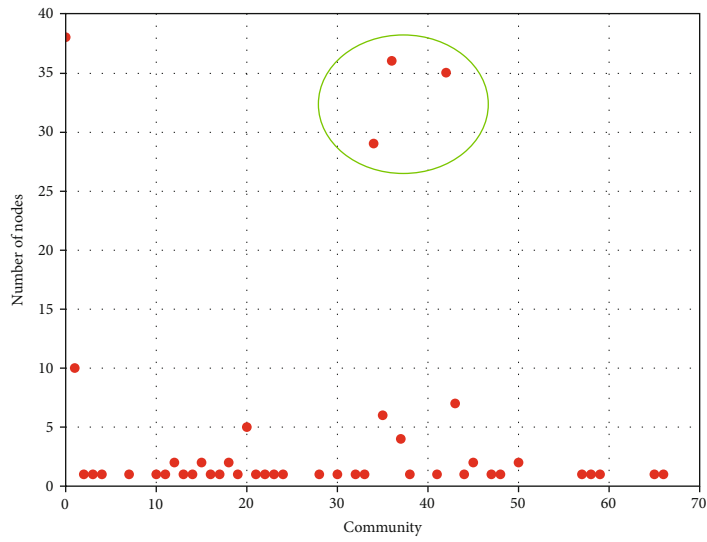
- (i) Community size distribution can reflect the community number of networks and the number of nodes in each community
- (ii) Degree distribution can measure the number of nodes with different degrees
- (iii) Closeness centrality distribution can reflect the number of nodes with different closeness. The closeness centrality of a node is a measure of centrality in

a network. The more central a node is, the closer it is to other nodes

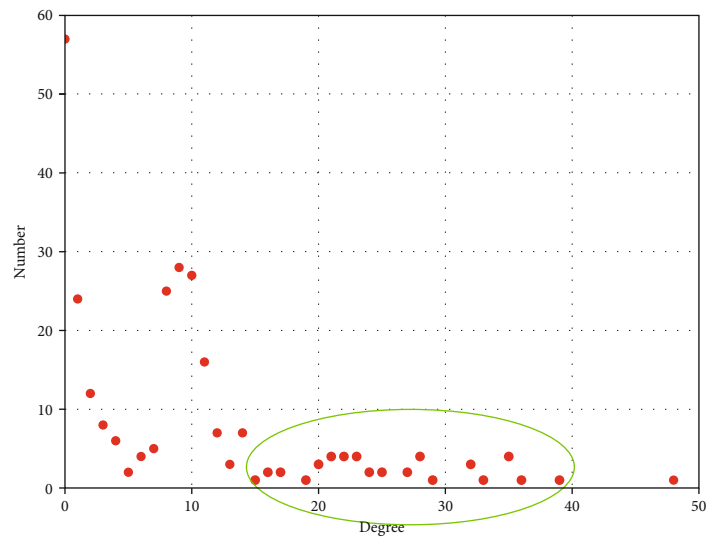
The results are shown in Figure 10, and the degree and closeness centrality values of core herb nodes are shown in Tables 7–9. As shown in Figure 10(a), there are three large communities, in which each community owns more than 25 nodes. Other communities are small and only have few nodes because the literature records complex symptoms and syndromes, and these herbs in small communities are used to treat other symptoms of patients. Thus, core herbs are discovered from the three communities for treating the main symptoms and syndromes of CGN in TCM. As shown in Figures 10(b) and 10(c) and Tables 7–9, the degree and closeness centrality of core herb nodes concentrate on the range of [15, 40] and [0.25, 0.45], respectively. It suggests that core herbs are represented by the important and central nodes in the herb network. Thus, if we construct a new herb network from new literature, then we can prejudge the core herbs for treating CGN according to their degree and closeness centrality, which can reduce cost and increase accuracy. For other diseases, we also can utilize the above rules to prejudge core herbs according to their degree and closeness centrality. So, these circled states can reflect the distribution rules of core herbs and are important for doctors and researchers to explore core herbs for CGN and other diseases.

7. Conclusions

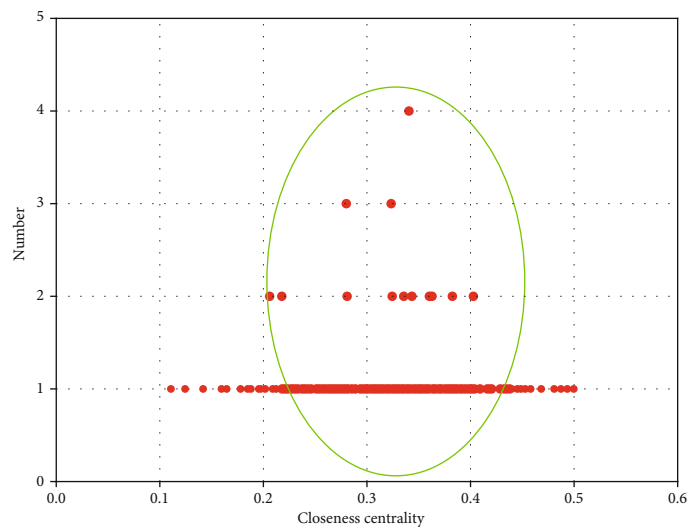
In this paper, we study the problem of core herb discovery in TCM and propose an artificial intelligence model CHDSC to discover core herbs for treating a certain disease from large-scale literature based on semantic analysis and community detection, in which word embedding algorithm ESS2VEC is designed to analyse the semantics of herbs in the literature, and label propagation-based algorithm LILPA is used to discover herb communities and core herbs. In the case study, CHDSC discovers three large herb communities where herbs can treat most syndromes of CGN and identifies core herbs for treating these syndromes with high accuracy. CHDSC can discover effective core herbs, which is helpful for the clinical application of herbs and formulae. In addition, the proposed model is



(a) Community size distribution



(b) Degree distribution



(c) Closeness centrality distribution

FIGURE 10: Results of network distributions.

introduced to discover core herbs for treating CGN as an example; it also can be applied to other diseases.

We also find that some syndromes cannot be covered by discovered core herbs and some core herbs with low degree (e.g., *asiatic cornelian cherry fruit* (Shan Zhu Yu) in blue community) are not discovered. These syndromes may be less recorded in the literature, and the collected literature may not contain the usage of these core herbs in most cases. Improving the semantic analysis and community detection modules is an important area of future research. For example, importing the “Sovereign-Minister-Assistant-Courier” combination rule in LILPA can combine TCM domain knowledge with community detection to guide label propagation and form a supervised way. The source and scale of literature have the influence on results, so enlarging the scale of the corpus and selecting authoritative literature can enhance the accuracy. In addition, for the Chinese word embedding algorithm ESSP2VEC, we can consider the syntax and Part of Speech (POS) [72] as features and predict the contextual words based on soft tree [73] to learn the semantics of Chinese words, which will also be the subject of future research.

Data Availability

The text data used to support the findings of this study have been deposited in <https://github.com/yunzhangwww/TCM-literature-corpus>

Conflicts of Interest

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

Acknowledgments

This work was supported by the National Key R&D Program of China (Nos. 2017YFC1703905 and 2018YFC1704105) and Sichuan Province Science and Technology Department (Nos. 2020YFS0372 and 2020YFS0302).

References

- [1] C. Wallis, “How artificial intelligence will change medicine,” *Nature*, vol. 576, no. 7787, article S48, 2019.
- [2] W. Zhang, X. He, and W. Lu, “Exploring discriminative representations for image emotion recognition with CNNs,” *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 515–523, 2020.
- [3] L. Yang, X. C. Cao, D. X. He, C. Wang, X. Wang, and W. X. Zhang, “Modularity based community detection with deep learning,” in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2252–2258, New York, NY, USA, July 2016.
- [4] M. Gimenez, J. Palanca, and V. Botti, “Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis,” *Neurocomputing*, vol. 378, pp. 315–323, 2020.
- [5] T. Paulraj, K. S. V. Chelliah, and S. Chinnasamy, “Lung computed axial tomography image segmentation using possibilistic fuzzy C-means approach for computer aided diagnosis system,” *International Journal of Imaging Systems and Technology*, vol. 29, no. 3, pp. 374–381, 2019.
- [6] Y. Zhang and Y. Zhao, “Pathogenic network analysis predicts candidate genes for cervical cancer,” *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 3186051, 8 pages, 2016.
- [7] A. Onan, “Biomedical text categorization based on ensemble pruning and optimized topic modelling,” *Computational and Mathematical Methods in Medicine*, vol. 2018, Article ID 2497471, 22 pages, 2018.
- [8] D. Kerstin and V. H. Frank, “Recent advances in extracting and processing rich semantics from medical texts,” *Artificial Intelligence in Medicine*, vol. 93, pp. 11–12, 2018.
- [9] J. Li, J. W. Lian, Y. X. Zhou et al., *Formula of Traditional Chinese Medicine*, China Press of Traditional Chinese Medicine, China, 9th edition, 2016.
- [10] M. Jiang, C. Lu, C. Zhang et al., “Syndrome differentiation in modern research of traditional Chinese medicine,” *Journal of Ethnopharmacology*, vol. 140, no. 3, pp. 634–642, 2012.
- [11] X. Gao, J. Shang, H. Liu, and B. R. Yu, “A meta-analysis of the clinical efficacy of TCM decoctions made from formulas in the Liu-Wei-Di-Huang-Wan categorized formulas in treating diabetic nephropathy proteinuria,” *Evidence-based Complementary and Alternative Medicine*, vol. 2018, Article ID 2427301, 10 pages, 2018.
- [12] X. J. Wang and B. L. Zhang, “Elucidation of compatibility principle and scientific value of Chinese medical formulae based on pharmacometabolomics,” *China Journal of Chinese Materia Medica*, vol. 35, no. 10, pp. 1346–1348, 2010.
- [13] Y. Zhao, “The ‘Jun-Chen-Zuo-Shi’ combination rule of TCM formula in the Ming dynasty,” *Journal of Traditional Chinese Medical Literature*, vol. 32, no. 3, pp. 23–25, 2014.
- [14] Y. Zhao, “The herb property combination rule of TCM formula in the Ming dynasty,” *Journal of Traditional Chinese Medical Literature*, vol. 32, no. 1, pp. 32–34, 2014.
- [15] Y. J. Bai, *Design, synthesis, and biological characterization of herb molecules based on ‘Jun-Chen-Zuo-Shi’ strategy*, Northwest University, China, 2014.
- [16] L. H. Wu, Y. Wang, Z. Li, B. Zhang, Y. Y. Cheng, and X. H. Fan, “Identifying roles of ‘Jun-Chen-Zuo-Shi’ component herbs of QiShenYiQi formula in treating acute myocardial ischemia by network pharmacology,” *Chinese Medicine*, vol. 9, no. 1, p. 24, 2014.
- [17] K. Li, *The multidimensional data analysis and judgment study of major herbs in formula*, Chengdu University of Traditional Chinese Medicine, China, 2007.
- [18] Y. Zhang, Y. Liu, J. J. Zhu, C. Yang, W. Yang, and S. Zhai, “NALPA: a node ability based label propagation algorithm for community detection,” *IEEE Access*, vol. 8, pp. 46642–46664, 2020.
- [19] W. Zhou, F. Wang, C. J. Wang, and J. Y. Xie, “Mining core herbs and their combination rules using effect degree,” *Journal of Frontiers of Computer Science & Technology*, vol. 7, no. 11, pp. 994–1001, 2013.
- [20] J. J. Wang, “Discussion on concept of ‘the monarch and his subjects, assistant and envoy’ and applied principle from documents,” *Chinese Journal of Basic Medicine in Traditional Chinese Medicine*, vol. 10, no. 5, pp. 63–65, 2004.

- [21] Y. T. Wang and L. Wang, "The explore of 'Jun' herbs in Zhi-Gan-Cao-Tang," *Global Traditional Chinese Medicine*, vol. 8, no. 8, pp. 955-956, 2015.
- [22] X. L. Song and X. Niu, "The explore of 'Jun' herbs in 'Ban-Xia-Sheng-Jiang-Gan-Cao' of Xie-Xin-Tang," *Chinese Journal of Experimental Traditional Medical Formulae*, vol. 13, no. 9, pp. 66-68, 2007.
- [23] J. P. Zhan, G. Zheng, M. Jiang et al., "Exploring association rules of traditional Chinese medicine syndrome-symptom-formula-herb in chronic glomerulonephritis by a novel text mining approach," *Chinese Journal of Experimental Traditional Medical Formulae*, vol. 19, no. 3, pp. 315-320, 2013.
- [24] Y. K. Ma, D. Z. Zhang, A. Wulamu, Y. Xie, H. Zang, and J. Zhang, "The core drugs analysis based on social network analysis about traditional Chinese medicine records semantic relation," *Procedia Computer Science*, vol. 31, pp. 328-335, 2014.
- [25] X. You, Y. K. Xu, J. Huang et al., "A data mining-based analysis of medication rules in treating bone marrow suppression by kidney-tonifying method," *Evidence-Based Complementary and Alternative Medicine*, vol. 2019, Article ID 1907848, 9 pages, 2019.
- [26] J. P. Chen, J. Poon, S. K. Poon, L. Xu, and D. M. Y. Sze, "Mining symptom-herb patterns from patient records using tripartite graph," *Evidence-Based Complementary and Alternative Medicine*, vol. 2015, Article ID 435085, 14 pages, 2015.
- [27] C. C. Chang, Y. C. Lee, C. C. Lin et al., "Characteristics of traditional Chinese medicine usage in patients with stroke in Taiwan: a nationwide population-based study," *Journal of Ethnopharmacology*, vol. 186, pp. 311-321, 2016.
- [28] S. Y. Yan, R. S. Zhang, X. Z. Zhou, P. Li, L. Y. He, and B. Y. Liu, "Exploring effective core drug patterns in primary insomnia treatment with Chinese herbal medicine: study protocol for a randomized controlled trial," *Trials*, vol. 14, no. 1, p. 61, 2013.
- [29] V. Žitkus, R. Butkienė, R. Butleris, R. Maskeliūnas, R. Damaševičius, and M. Woźniak, "Minimalistic approach to coreference resolution in Lithuanian medical records," *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 9079840, 14 pages, 2019.
- [30] H. Liang, B. Tsui, H. Ni et al., "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature Medicine*, vol. 25, no. 3, pp. 433-438, 2019.
- [31] X. Zhou, B. Liu, Z. Wu, and Y. Feng, "Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks," *Artificial Intelligence in Medicine*, vol. 41, no. 2, pp. 87-104, 2007.
- [32] S. Y. Yu, J. Yang, M. X. Yang et al., "Application of acupoints and meridians for the treatment of primary dysmenorrhea: a data mining-based literature study," *Evidence-Based Complementary and Alternative Medicine*, vol. 2015, Article ID 752194, 8 pages, 2015.
- [33] M. J. Choi, B. T. Choi, H. K. Shin, B. C. Shin, Y. K. Han, and J. U. Baek, "Establishment of a comprehensive list of candidate antiaging medicinal herb used in Korean medicine by text mining of the classical Korean medical literature, "Dongeuibogam," and preliminary evaluation of the antiaging effects of these herbs," *Evidence-Based Complementary and Alternative Medicine*, vol. 2015, Article ID 873185, 29 pages, 2015.
- [34] L. X. Zhang, F. Wang, L. Wang et al., "Prevalence of chronic kidney disease in China: a cross-sectional survey," *Lancet*, vol. 379, no. 9818, pp. 815-822, 2012.
- [35] H. Y. Wang, *Nephrology*, People's Medical Publishing House, China, 3th edition, 2012.
- [36] B. H. Liu and Y. Xu, "The diagnose, syndrome differentiation and therapeutic effect evaluation of chronic glomerulonephritis (trial scheme)," *Shanghai Journal of Traditional Chinese Medicine*, vol. 40, no. 6, pp. 8-9, 2006.
- [37] J. T. Liu, Y. Jin, and F. F. Li, "Research process of traditional Chinese medicine of chronic glomerulonephritis," *Chinese Archives of Traditional Chinese Medicine*, vol. 31, no. 10, pp. 2127-2129, 2013.
- [38] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146-147, 1954.
- [39] S. S. Cao, W. Lu, J. Zhou, and X. L. Li, "Cw2vec: learning Chinese word embeddings with stroke n-gram information," in *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 5053-5061, New Orleans, LA, 2018.
- [40] Y. Zhang, Y. Liu, J. Zhu et al., "Learning Chinese word embeddings from stroke, structure and pinyin of characters," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1011-1020, Beijing, China, November 2019.
- [41] Y. X. Meng, W. Wu, F. Wang et al., "Glyce: glyph-vectors for Chinese character representations," in *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2742-2753, Vancouver, Canada, December 2019.
- [42] X. X. Chen, X. Lei, Z. Y. Liu, M. S. Sun, and H. B. Luan, "Joint learning of character and word embeddings," in *Proceedings of the 24th International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 1236-1242, Buenos Aires, Argentina, July 2015.
- [43] Y. M. Sun, L. Lin, N. Yang, Z. Z. Ji, and X. L. Wang, "Radical-enhanced Chinese character embedding," in *Neural Information Processing. ICONIP 2014*, C. K. Loo, K. S. Yap, K. W. Wong, A. Teoh, and K. Huang, Eds., vol. 8835 of Lecture Notes in Computer Science, pp. 279-286, Springer, Cham, 2014.
- [44] J. X. Yu, X. Jian, H. Xin, and Y. Q. Song, "Joint embeddings of Chinese words, characters, and fine-grained subcharacter components," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 286-291, Copenhagen, Denmark, September 2017.
- [45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013, <https://arxiv.org/abs/1301.3781>.
- [46] S. Park, J. Byun, S. Baek, Y. Cho, and A. Oh, "Subword-level word vector representations for Korean," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2429-2438, Melbourne, Australia, July 2018.
- [47] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, no. 1, pp. 135-146, 2017.
- [48] Y. Zhang, Y. G. Liu, Q. Q. Li, R. J. Jin, and C. B. Wen, "LILPA: a label importance based label propagation algorithm for community detection with application to core drug discovery," *Neurocomputing*, vol. 413, pp. 107-133, 2020.
- [49] M. L. Lu, Z. L. Zhang, Z. H. Qu, and Y. Kang, "LPANNI: overlapping community detection using label propagation in large-scale complex networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 9, pp. 1736-1749, 2019.

- [50] Y. Zhang, Y. G. Liu, J. T. Li et al., "WOCDA: a whale optimization based community detection algorithm," *Physica A*, vol. 539, article 122937, 2020.
- [51] J. R. Zhu, B. L. Chen, and Y. F. Zeng, "Community detection based on modularity and k-plexes," *Information Sciences*, vol. 513, pp. 127–142, 2020.
- [52] Pharmacopoeia Commission of the Ministry of Health of the People's Republic of China, *The Pharmacopoeia of the People's Republic of China*, China Medical Science Press, China, 2010.
- [53] M. Abdel-Basset, M. Mohamed, M. Elhoseny, L. H. Son, F. Chiclana, and A. E.-N. H. Zaied, "Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar disorder diseases," *Artificial Intelligence in Medicine*, vol. 101, article 101735, 2019.
- [54] S. Gregory, "Finding overlapping communities in networks by label propagation," *New Journal of Physics*, vol. 12, no. 10, article 103018, 2010.
- [55] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.
- [56] D. Lusseau, "The emergent properties of a dolphin social network," *Royal Society of London Series B*, vol. 270, Supplement 2, 2003.
- [57] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.
- [58] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, article 036104, 2006.
- [59] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [60] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical Review E*, vol. 70, no. 5, article 056122, 2004.
- [61] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.
- [62] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, article 033015, 2009.
- [63] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics Theory and Experiment*, vol. 2009, article 03024, no. 3, p. P03024, 2009.
- [64] J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.
- [65] S. Tzu-Ray and H. Lee, "Learning Chinese word representations from glyphs of characters," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 264–273, Copenhagen, Denmark, September 2017.
- [66] J. R. Xie, B. K. Szymanski, and X. M. Liu, "SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 344–349, Vancouver, Canada, December 2011.
- [67] K. Liu, J. B. Huang, H. L. Sun, M. J. Wan, Y. T. Qi, and H. Li, "Label propagation based evolutionary clustering for detecting overlapping and non-overlapping communities in dynamic networks," *Knowledge-Based Systems*, vol. 89, pp. 487–496, 2015.
- [68] C. Tong, J. W. Niu, J. M. Wen, Z. Y. Xie, and F. Peng, "Weighted label propagation algorithm for overlapping community detection," in *2015 IEEE International Conference on Communications (ICC)*, pp. 1–6, London, UK, June 2015.
- [69] X. K. Zhang, J. Ren, C. Song, J. Jia, and Q. Zhang, "Label propagation algorithm for community detection based on node importance and label influence," *Physics Letters A*, vol. 381, no. 33, pp. 2691–2698, 2017.
- [70] M. Shen and Z. Ma, "A novel node gravitation-based label propagation algorithm for community detection," *International Journal of Modern Physics C*, vol. 30, no. 6, article 1950049, 2019.
- [71] S. W. Lai, K. Liu, S. Z. He, and J. Zhao, "How to generate a good word embedding," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 5–14, 2016.
- [72] M. Woźniak, D. Polap, R. Damasevicius, and W. Wei, "Design of computational intelligence-based language interface for human-machine secure interaction," *Journal of Universal Computer Science*, vol. 24, no. 4, pp. 537–553, 2018.
- [73] M. Woźniak and D. Polap, "Soft trees with neural components as image-processing technique for archeological excavations," *Personal and Ubiquitous Computing*, vol. 24, pp. 1–13, 2020.

Retraction

Retracted: mir-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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. Zhu, Z. Shen, D. Man, H. Ruan, and S. Huang, "miR-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 8209504, 10 pages, 2020.

Research Article

miR-152-3p Affects the Progression of Colon Cancer via the KLF4/IFITM3 Axis

Xiaoyi Zhu,¹ Zhan Shen,¹ Da Man,¹ Hang Ruan,¹ and Sha Huang^{1,2}

¹Department of Colorectal Surgery, Shulan (Hangzhou) Hospital, Hangzhou 310000, China

²Department of Plastic Surgery, Shulan (Hangzhou) Hospital, Hangzhou 310000, China

Correspondence should be addressed to Sha Huang; huangsha1020@163.com

Received 27 May 2020; Accepted 20 August 2020; Published 1 September 2020

Academic Editor: Tao Huang

Copyright © 2020 Xiaoyi Zhu 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.

Objective. The purpose of this study was to investigate the relationship between miR-152-3p and the KLF4/IFITM3 axis, thereby revealing the mechanism underlying colon cancer occurrence and development, consequently providing a promising target for colon cancer treatment. **Methods.** Bioinformatics methods were implemented to analyze the differential expression of miRNAs and mRNAs in colon cancer, confirm the target miRNA, and predict the downstream targeted mRNAs. qRT-PCR and Western blot were performed to detect the expression of miR-152-3p, KLF4, and IFITM3. CCK-8 and colony formation assays were conducted for the assessment of cell proliferation, and flow cytometry was carried out for the detection of cell apoptosis. Finally, dual-luciferase reporter gene assay was employed to verify the targeting relationship between miR-152-3p and KLF4. **Results.** miR-152-3p was highly expressed in colon cancer cells, whereas KLF4 was poorly expressed. Dual-luciferase assay verified that miR-152-3p targeted to bind to KLF4 and suppressed its expression. Moreover, silencing miR-152-3p or overexpressing KLF4 was found to downregulate IFITM3, thereby inhibiting cell proliferation and potentiating cell apoptosis. In rescue experiments, we found that miR-152-3p deficiency decreased the expression of IFITM3 and weakened cancer cell proliferation, and such effects were restored when miR-152-3p and KLF4 were silenced simultaneously. **Conclusion.** In sum, we discovered that miR-152-3p can affect the pathogenesis of colon cancer via the KLF4/IFITM3 axis.

1. Introduction

Colon cancer is the fourth common malignant tumor worldwide and the fifth cause of cancer-related deaths, with confirmed cases around 1,096,601 in 2018 and deaths up to 551,269 [1]. Due to the change in people's lifestyle and diet, the morbidity of colon cancer annually increases, and the age of people suffering from this cancer tends to be lower, leading to a high rank (3th) among gastrointestinal malignancies in China [2]. Therapies currently for colon cancer mainly include surgical resection, chemotherapy, radiotherapy, and targeted therapy-based comprehensive treatment [3], but the efficacy in advanced patients remains poor with clinical symptoms partially relieved [4]. Statistically, for the patients with no metastasis, with local metastasis, and with distant metastasis, their 5-year survival rate was 90%, 70%, and 10%, respectively [5]. Thus, exploring biomarkers for early

diagnosis of colon cancer and therapeutic targets is beneficial for better treatment and prognosis.

KLF4 (Kruppel-like factor 4), a zinc finger transcription factor, is involved in the regulation of thymocyte as well as the colonic goblet cell proliferation, differentiation, apoptosis, and metabolism [6, 7]. Studies have proved that KLF4 is differentially expressed in various human cancers, such as prostate cancer [8], liver cancer [9], and breast cancer [10]. In addition, KLF4 has been considered as a molecular target that can be used in cancer treatment. For example, FBXO32 can promote the degradation of the KLF4 proteasome to suppress the occurrence of breast cancer [11]. In hepatocellular carcinoma, KLF4 can function on cell growth and migration via the CDH3/GSK-3 β axis [12], while in colon cancer KLF4 is always reported acting as an oncogene. Decreased KLF4 was observed to inhibit NDRG2 signal-dependent cell proliferation in colorectal cancer, which provides a theoretical

basis for early diagnosis and treatment 13. Moreover, KLF4 can enhance the sensitivity of colon cancer cells to cisplatin through altering the expression of HMGB1 (high-mobility group box 1) and hTERT (human telomerase reverse transcriptase) 14. Another study also found that IFITM3 is a direct transcription target of KLF4 in colon cancer, and decreased KLF4 leads to the upregulation of IFITM3, thereby promoting progression and metastasis 15.

In the present study, we predicted that there were targeted binding sites of miR-152-3p on KLF4 3'UTR and found that miR-152-3p was highly expressed in colon cancer, which has never been reported. In view of this, we made further efforts and confirmed that miR-152-3p targeted KLF4 to mediate the expression of IFITM3 and affect the development of colon cancer.

2. Materials and Methods

2.1. Bioinformatics Analysis. Expression profiles of mRNAs and miRNAs associated with colon cancer were obtained from TCGA database (<https://portal.gdc.cancer.gov/>). “edgeR” package was employed to perform differential analysis ($|\log FC| > 2$, $p_{adj} < 0.05$) to find the differentially expressed miRNAs (DEmiRNAs), which were subjected to survival analysis combined with the matched clinical information. Three databases miRDB (<http://mirdb.org/>), miRTarBase (<http://mirtarbase.mbc.nctu.edu.tw/php/index.php>), and TargetScan (http://www.targetscan.org/vert_71/) were applied to predict the targets of the target miRNA. Venn diagram was plotted to find the candidate targeted mRNAs.

2.2. Cell Culture. Human normal colon cell line CCD-18Co (BNCC337724), human embryonic kidney cell line HEK-293T (BNCC338274), and colon cancer cell lines HT29 (ATCC HTB-38), HCT116 (BNCC337692), and SW480 (BNCC100604) were all purchased from the American Type Culture Collection. All cells were grown in Dulbecco's Modified Eagle's Medium (DMEM; Gibco, Thermo Fisher Scientific, Inc., Waltham, MA, USA) containing 10% fetal bovine serum (FBS; HyClone; GE Healthcare Life Sciences, Logan, UT, USA) and 100 U/mL streptomycin/penicillin (Gibco; Thermo Fisher Scientific, Inc.) and maintained in 5% CO₂ at 37°C. Mediums were replaced every 2 or 3 days.

2.3. Cell Transfection. miR-152-3p mimic, miR-152-3p inhibitor, si-KLF4, and their negative controls (mimic NC, inhibitor NC, and si-NC) were purchased from GenePharma (Shanghai, China). Overexpression plasmids targeting KLF4 and IFITM3 (oe-KLF4 and pre-IFITM3) and their negative controls (oe-NC and pre-NC) were ordered from Miaoling Biotechnology (Wuhan, China). For transfection, Lipofectamine 2000 (Thermo Fisher Scientific, Inc.) was used following the manufacturer's instructions. Transfected cells were maintained in DMEM containing 5% CO₂ at 37°C for subsequent experiments. All cells grew in complete mediums for at least 24h and were washed in PBS (pH 7.4) before transfection.

2.4. qRT-PCR. Total RNA was extracted using TRizol (Invitrogen), treated by DNase I (TaKaRa) for the removal of

the genomic DNA, and then reversely transcribed into cDNA by reverse transcriptase M-MLV (TaKaRa). Applied Biosystems 7300 Real-Time PCR System (Applied Biosystems, USA) was employed to test the KLF4 mRNA expression, and the result was expressed in $2^{-\Delta\Delta Ct}$. Primer sequences were as follows: KLF4-F: 5'-ATGGCTGTCAGCGACGCGCTGC-3', KLF4-R: 5'-TTAAAAATGCCTCTTCATGTGTAAGCG-3'; GAPDH-F: 5'-GCACCGTCAAGCTGAGAAC-3', GAPDH-R: 5'-TGGTGAAGACGCCAGTGG-3'.

RNAiso technology (TaKaRa, Dalian, China) was applied to isolate the total RNA from cancer cells and solid tumors. The expression of miR-152-3p was analyzed by a TaqMan RT kit (Applied Biosystems) and TaqMan MicroRNA kit (Applied Biosystems) under the Applied Biosystems 7300 Real-Time PCR System (Applied Biosystems, USA), and $2^{-\Delta\Delta Ct}$ was performed to normalize the miR-152-3p expression. The primers were designed as follows: miR-152-3p-F: 5'-ACACTCCAGCTGGGTCAGTGCATGACAG-3', miR-152-3p-R: 5'-CTCAACTGGTGTCTGGAGTCCGGCAATTCAGTTGAGCCAAGTT-3'; U6-F: 5'-GCTTCGGCA GCACATATACTAAAAT-3', U6-R: 5'-CGCTTCAGAAT TTGCGTGCAT-3'.

2.5. Western Blot. After 48 h of transfection, cells were washed in cold PBS for three times and then lysed on ice with whole cell lysate for 10 min, with the concentration of the product sequentially determined using the BCA protein assay kit (Thermo Fisher Scientific, Waltham, MA, USA). Subsequently, 30 μ g of the total proteins was separated by polyacrylamide gel electrophoresis (PAGE) and then transferred onto polyvinylidene fluoride (PVDF) membranes (Amersham, USA). After being blocked in 5% skim milk at room temperature for 1 h, the membranes were incubated overnight at 4°C with primary antibodies, followed by horseradish peroxidase- (HRP-) labeled secondary antibody goat anti-rabbit IgG H&L (ab6721, 1:2000, Abcam, Cambridge, UK) at room temperature for 1 h. Primary antibodies included KLF4 rabbit polyclonal antibody (ab215036, 1:1000, Abcam, Cambridge, UK), IFITM3 rabbit polyclonal antibody (ab109429, 1:1000, Abcam, Cambridge, UK), and GAPDH rabbit polyclonal antibody (ab9485, 1:2500, Abcam, Cambridge, UK). PBST (PBS buffer containing 0.1% Tween-20) was used to wash the membranes after each reaction. An optical luminometer (GE, USA) was employed to visualize the protein bands.

2.6. Dual-Luciferase Reporter Gene Assay. Amplified wild-type (WT) and mutant (MUT) KLF4 3'UTR were inserted into the pMIR reporter vectors (Ambion; Thermo Fisher Scientific, Inc.). Constructs WT-KLF4 and MUT-KLF4 were cotransfected with the miR-152-3p mimic/miR-152-3p inhibitor or their negative controls into HEK-293T cells (BNCC338274), respectively. The Renilla luciferase vector pRL-TK (TaKaRa, Dalian, China) was taken as the internal reference. The luciferase activity was assayed by a luciferase reporter kit (Promega, Madison, WI).

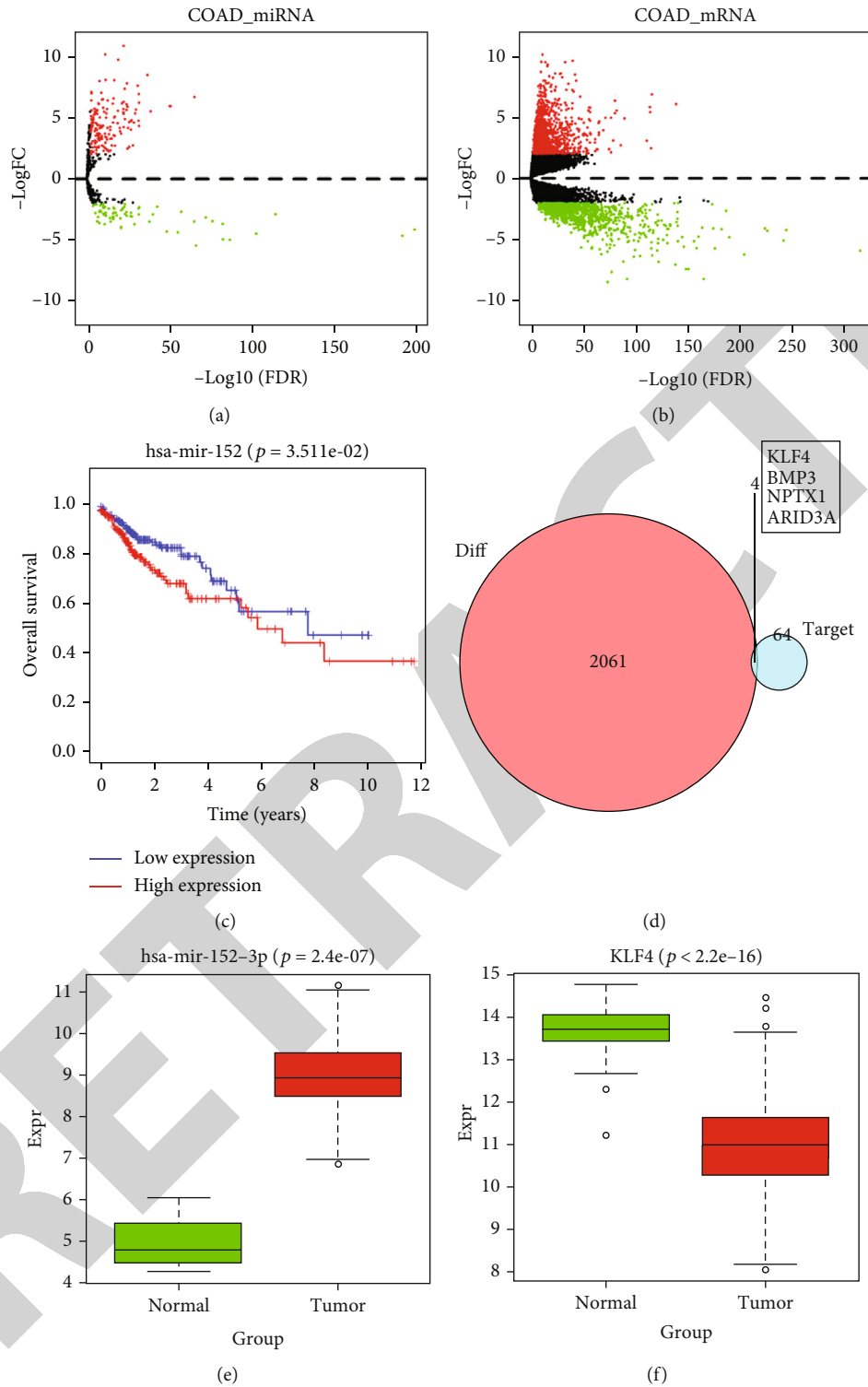


FIGURE 1: Bioinformatics analysis. Volcano plots were made to find the (a) DEMiRNAs and (b) DEMRNAs in TCGA-COAD dataset. (c) Survival analysis was performed, and it was found that miR-152-3p was of remarkable survival significance. (d) Venn diagram was plotted to find the candidate target genes from the predicted mRNAs of miR-152-3p and the DEMRNAs in TCGA-COAD dataset. (e) miR-152-3p was shown to be significantly upregulated in the colon cancer tissue in TCGA-COAD dataset. (f) KLF4 was verified to be noticeably lowly expressed in cancer tissue in TCGA-COAD dataset.

2.7. CCK-8. 96-well plates were used for cell culture at a density of 1×10^4 cells/well. The specific procedures proceeded as previously described 16. According to the protocols of

the CCK-8 kit (Dojindo, Japan), cell viability was examined at 0, 24, 48, and 72 h, respectively. The absorbance at 450 nm in wavelength of each well was read.

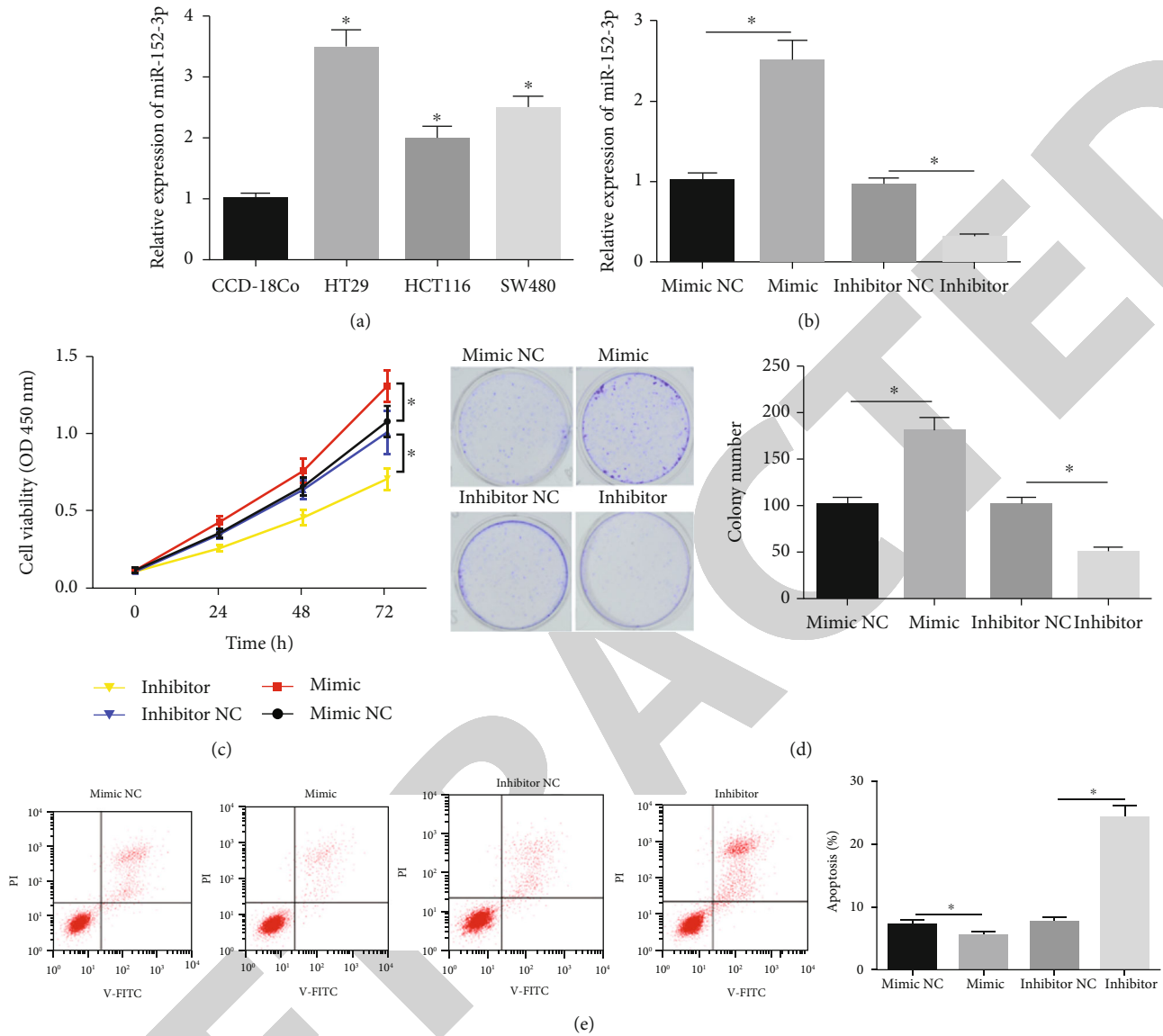


FIGURE 2: miR-152-3p is upregulated in colon cancer cells and affects cancer cell proliferation and apoptosis. miR-152-3p was significantly increased in (a) colon cancer cell lines HT29, HCT116, and SW480 relative to the normal cell line CCD-18Co. (b) qRT-PCR, (c) CCK-8, (d) colony formation assay, and (e) flow cytometry were conducted to assess the expression of miR-152-3p, cell viability, colony-forming ability, and apoptosis in each treatment group (* means $p < 0.05$).

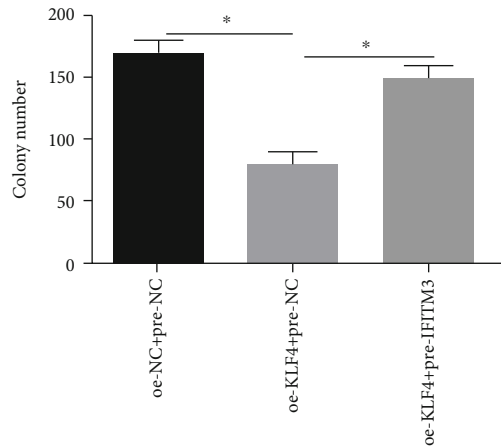
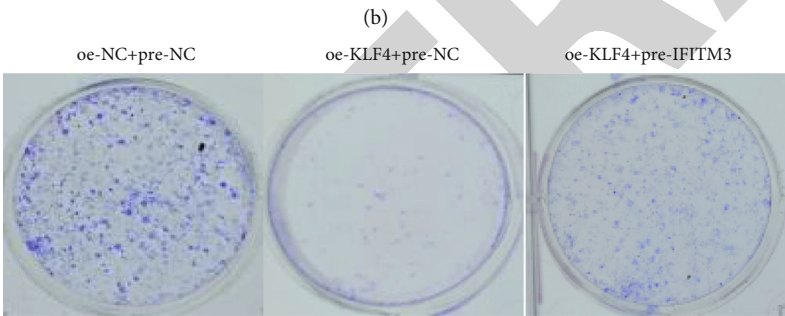
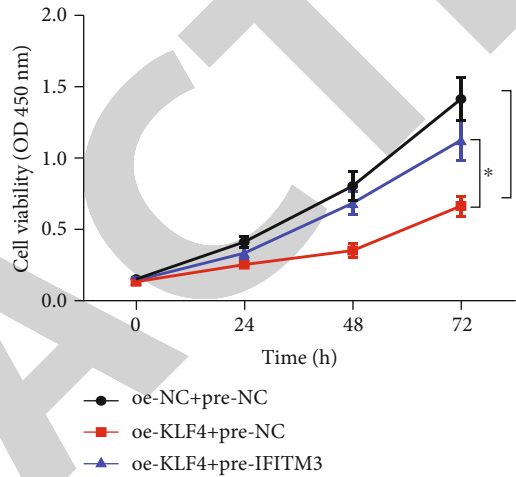
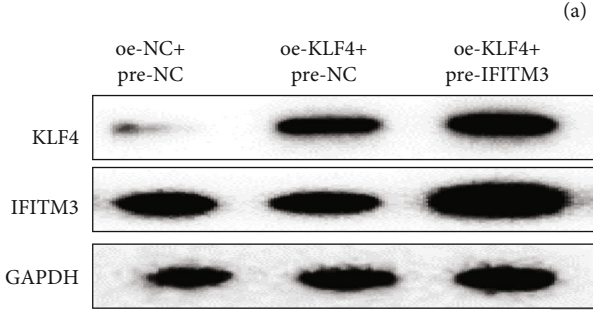
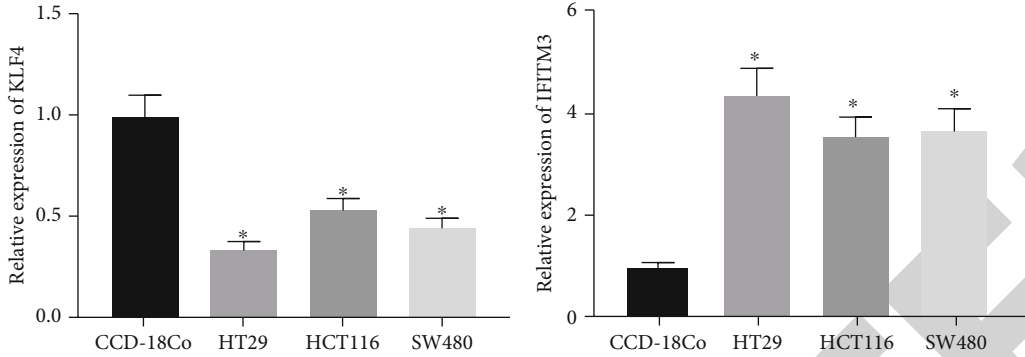
2.8. Colony Formation Assay. Transfected cells were seeded into 6-well plates (1×10^3 cells/well) and cultured in DMEM supplemented with 10% FBS at 37°C . After two weeks, the cells were washed in PBS, fixed with 10% formalin, and stained by 0.1% crystal violet (Sigma, USA). Cell colonies that were visible to the naked eyes were counted at the end.

2.9. Flow Cytometry. Annexin V and Propidium Iodide (PI) fluorescein staining kits (Bender MedSystems, Austria) were applied in this experiment. 5×10^5 cells were suspended in $500 \mu\text{L}$ (1x) binding buffer (10 mM HEPES pH 7.4, 140 mM NaCl, and 2.5 mM CaCl_2) for preparation. Then, the cell suspension was incubated with Annexin V (1:20) for 5 min, followed by PI for another 15 min. Cell apoptosis was analyzed by flow cytometry, and cell apoptotic rate was calculated.

2.10. Statistical Analysis. All data were processed by the SPSS 22.0 software (SPSS Inc., Chicago, IL, USA) and GraphPad Prism 6.0 software (GraphPad Prism 6.0; San Diego, CA, USA). Mean \pm standard deviation (SD) was used to express the measurement data, while t test and one-way ANOVA were performed to analyze the differences between two groups and among multiple groups. All results were representative of at least three independent experiments. $p < 0.05$ was considered statistically significant.

3. Results

3.1. Bioinformatics Analysis. In all, 239 DEmiRNAs (Figure 1(a)) and 2065 DEMRNAs (Figure 1(b)) were obtained by differential analysis. Among the DEmiRNAs, 4 miRNAs (miR-145, miR-152, miR-193b, and miR-216a) with survival



(d) FIGURE 3: Continued.

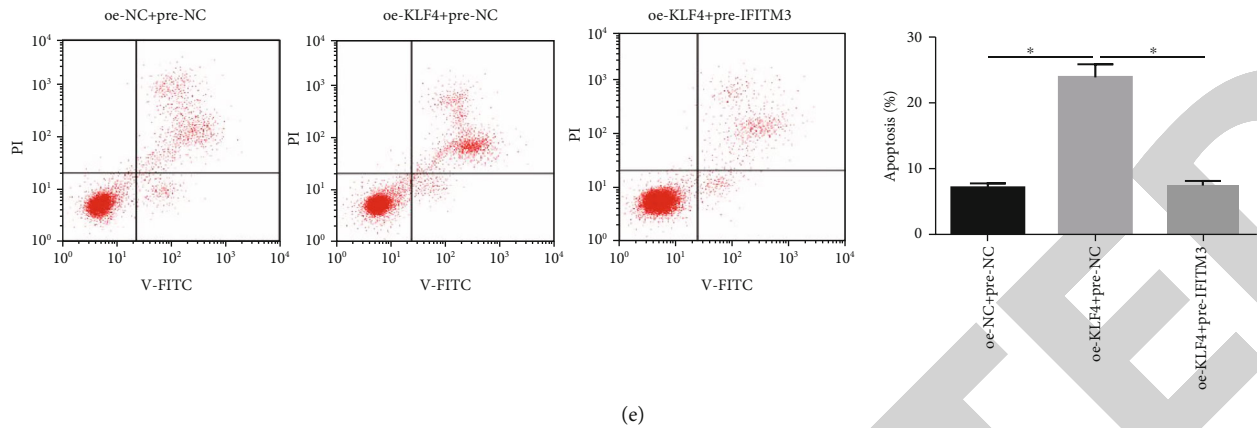


FIGURE 3: The overexpression of KLF4 mediates the IFITM3 expression to regulate colon cancer cell proliferation and apoptosis. (a) The expression of KLF4 and IFITM3 in the normal cell line CCD-18Co and colon cell lines HT29, HCT116, and SW480 was detected by qRT-PCR. (b) Western blot was carried out for the protein examination of KLF4 and IFITM3 in cells transfected with oe-NC+pre-NC, oe-KLF4+pre-NC, and oe-KLF4+pre-IFITM3. (c) CCK-8, (d) colony formation assay, and (e) flow cytometry were performed to determine cell viability, colony-forming ability, and cell apoptosis (* means $p < 0.05$).

significance were screened, of which miR-152 exhibited a significant correlation with the survival of colon cancer (Figure 1(c)) and was remarkably elevated in the colon cancer tissue (Figure 1(e)). Relevant literature reported that miR-152 is involved in the ceRNA network 17. Thus, miRDB, miRTarBase, and TargetScan databases were used to predict the targets for miR-152, and eventually, 68 targeted mRNAs were obtained. Thereafter, 4 candidate mRNAs, including KLF4, BMP3, NPTX1, and ARID3A (Figure 1(d)), were identified from the intersection of these predicted mRNAs and downregulated DEMRNAs in TCGA. KLF4, which was observed to be significantly decreased in the cancer tissue in TCGA-COAD dataset, was selected for follow-up analysis (Figure 1(f)).

3.2. miR-152-3p Is Highly Expressed in Colon Cancer Cells and Affects the Proliferation and Apoptosis of Cancer Cells In Vitro. In order to further investigate the role of miR-152-3p in colon cancer, qRT-PCR was primarily conducted to examine the expression of miR-152-3p in tumor cells (Figure 2(a)). It turned out that the miR-152-3p expression was significantly elevated in the cancer cell lines HT29, HCT116, and SW480 relative to that in the normal cell line CCD-18Co. Hence, the HT29 cell line where miR-152-3p was most highly expressed was selected for follow-up analysis.

miR-152-3p mimic, miR-152-3p inhibitor, and their negative controls were transfected into HT29 cells, respectively. Transfection efficiency was detected by qRT-PCR, and it was found that miR-152-3p was remarkably overexpressed or silenced in cells transfected with miR-152-3p mimic or inhibitor (Figure 2(b)). Thereafter, CCK-8 and colony formation assays were performed for the examination of cell proliferation. As shown in Figures 2(c) and 2(d), cells transfected with miR-152-3p mimic had an increased cell viability and stronger colony-forming ability, whereas cells with low miR-152-3p expression were accompanied with a reduced cell proliferation. Moreover, as revealed by flow cytometry,

miR-152-3p silencing greatly increased cell apoptosis (Figure 2(e)).

3.3. Overexpression of KLF4 Mediates IFITM3 to Suppress the Proliferation of Colon Cancer Cells. KLF4 as a potential target of miR-152-3p is worthy of further exploration. Published literature reported that KLF4 can negatively mediate IFITM3 and plays a crucial role in the pathogenesis of colon cancer 15. Therefore, we firstly carried out qRT-PCR to detect the expression of these two genes in colon cancer cell lines. Compared with normal cells, KLF4 was significantly downregulated in colon cancer cells, while IFITM3 was upregulated (Figure 3(a)). Subsequently, Western blot was used to assess the protein levels of KLF4 and IFITM3 in cells transfected with oe-NC+pre-NC, oe-KLF4+pre-NC, and oe-KLF4+pre-IFITM3, finding that IFITM3 was greatly decreased in the oe-KLF4+pre-NC group by comparison with the oe-NC+pre-NC group (Figure 3(b)).

Moreover, the KLF4 overexpression was found to reduce cell viability and colony-forming ability as well as promote cell apoptosis, yet such effects were reversed upon the cooverexpression of KLF4 and IFITM3 (Figures 3(c)–3(e)). Collectively, it elucidated that KLF4 could function on the progression of colon cancer by regulating the expression of IFITM3, and its inhibitory effect on cancer cells could be suppressed with the upregulation of IFITM3.

3.4. KLF4 Is a Direct Target of miR-152-3p. As mentioned above, we predicted that miR-152-3p might target to regulate KLF4 (Figure 4(a)). To deeply explore the relationship between miR-152-3p and KLF4, dual-luciferase reporter gene assay was conducted, demonstrating that the markedly decreased luciferase activity happened in cells transfected with miR-152-3p mimic and WT-KLF4, while the highest activity occurred in cells with the miR-152-3p inhibitor and WT-KLF4 (Figure 4(b)). Western blot showed that the miR-152-3p overexpression reduced the KLF4 protein level but elevated the IFITM3 protein expression (Figure 4(c)).

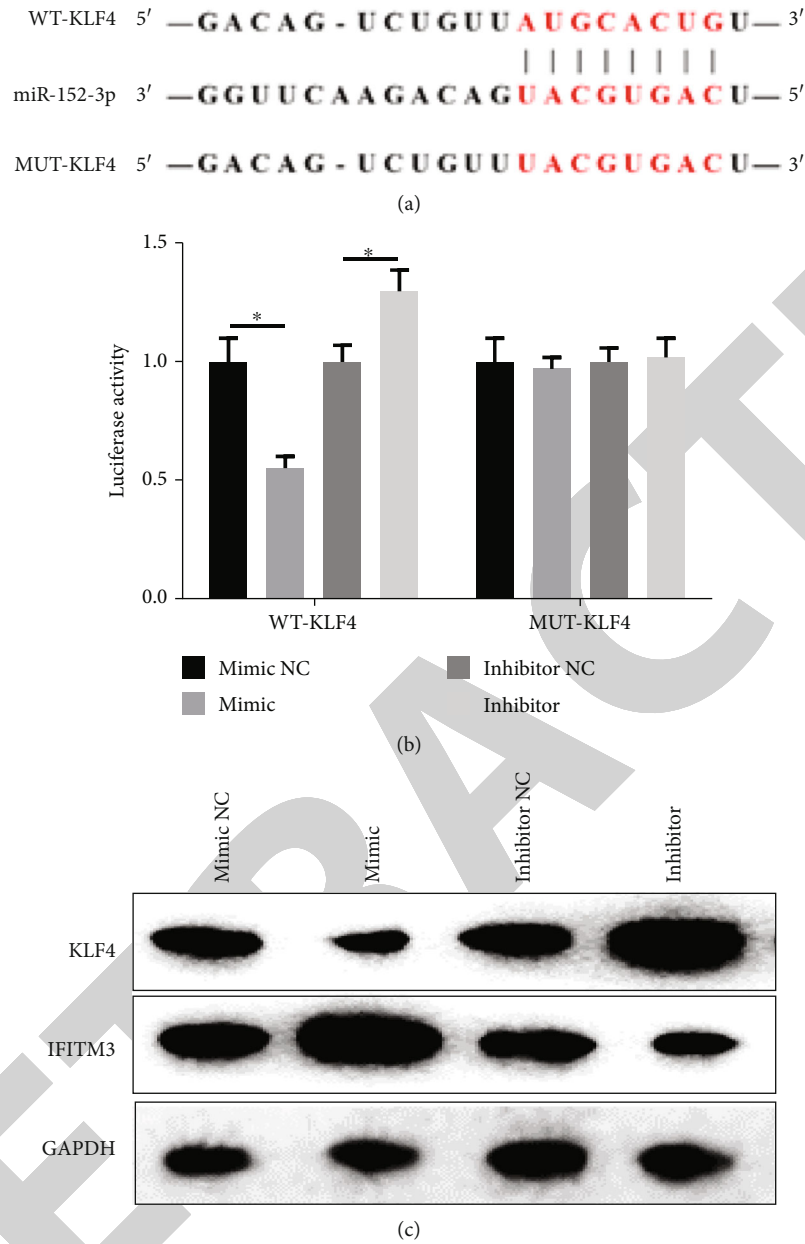


FIGURE 4: miR-152-3p targets KLF4 and decreases its expression. (a) Binding sites of miR-152-3p on KLF4 3'UTR were predicted by the bioinformatics method. (b) Dual-luciferase assay was done for the validation of the targeting relationship between miR-152-3p and KLF4. (c) Western blot was conducted to test the protein expression of KLF4 and IFITM3 in each treatment group (* means $p < 0.05$).

3.5. miR-152-3p Affects Colon Cancer Cell Growth via Regulating the KLF4/IFITM3 Axis. To validate the regulation of miR-152-3p on KLF4/IFITM3 and clarify the role of such regulation in colon cancer, rescue experiment was carried out. Inhibitor NC+si-NC, miR-152-3p inhibitor+si-NC, and miR-152-3p inhibitor+si-KLF4 were used to transfect cells. As plotted in Figure 5(a), KLF4 was greatly upregulated when miR-152-3p was inhibited, whereas IFITM3 was significantly downregulated. Notably, the IFITM3 protein level was recovered near to the level in the inhibitor NC+si-NC group when miR-152-3p and KLF4 were silenced concurrently. In addition, CCK-8 and colony formation assays revealed that the low miR-152-3p expression repressed cell viability and

colony-forming ability, but such inhibitory effect was reversed after KLF4 was silenced (Figures 5(b) and 5(c)). Moreover, in agreement with the results concluded above, increased KLF4 indicated a high apoptosis rate, as shown in Figure 5(d).

4. Discussion

In our study, we first confirmed that miR-152-3p was highly expressed in colon cancer cells, and silencing miR-152-3p could inhibit cell proliferation and growth. Also, we found that KLF4 was a direct target of miR-152-3p by conducting bioinformatics methods and dual-luciferase reporter gene assay. Finally, we used various experiments and elucidated

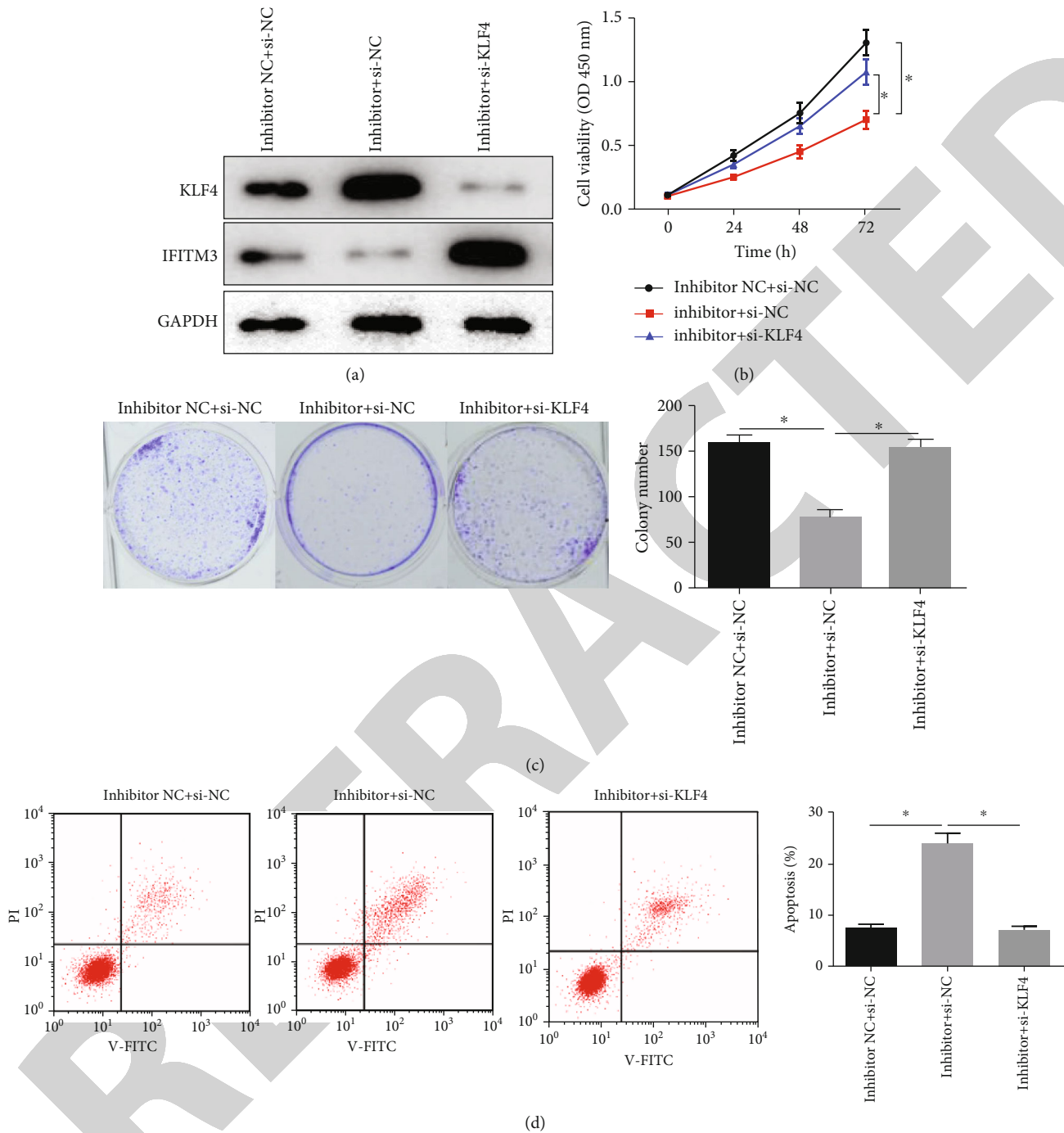


FIGURE 5: miR-152-3p affects colon cancer cell proliferation and apoptosis via the KLF4/IFITM3 axis. (a) Western blot was carried out to determine the protein expression of KLF4 and IFITM3 in cells transfected with inhibitor NC+si-NC, miR-152-3p inhibitor+si-NC, and miR-152-3p inhibitor+si-KLF4. (b) CCK-8, (c) colony formation assay, and (d) flow cytometry were conducted to assay cell viability, colony-forming ability, and apoptosis in each treatment group (* means $p < 0.05$).

a miR-152-3p-dependent mechanism by which miR-152-3p affected colon cancer cell proliferation and growth via the KLF4/IFITM3 axis.

MicroRNAs (miRNAs) can promote the degradation and inhibit the translation of mRNAs by interacting with 3' UTR of the target genes in eukaryotic cells, thereby participating in gene regulation in the posttranscriptional level 18. Being able to be expressed in various malignancies 19, miRNAs can not

only serve as protooncogenes 20, 21 but also act as tumor suppressor genes affecting tumorigenesis and cancer cell differentiation 22. In diverse cancer tissues and cells, miR-152-3p exhibits significantly different expression levels. For instance, in breast cancer, the upregulation of miR-152-3p exerts its antitumor role by negatively regulating PIK3CA to suppress the activation of AKT and RPS6, thus inhibiting the HCC1806 cell proliferation 23. However, in glioma 24

and leukemia 25, the expression of miR-152-3p was reported to be elevated in cancer tissue and cells, which is significantly higher than that in normal tissue and cells. Interestingly, there is a study revealing that miR-152-3p is differentially expressed in different T_stages (T1-T4) of colon cancer, with increased expression in T2-T4 and relatively higher expression in T2 and T4 26. In this study, bioinformatics analysis and qRT-PCR revealed that miR-152-3p was markedly elevated in colon cancer tissue and cells, and silencing miR-152-3p led to decreased cell proliferation but increased cell apoptosis.

KLF4 is a transcription factor belonging to the Kruppel-like family and mediates some basic biological progresses, such as cell proliferation, differentiation, and migration 27. KLF4 is highly expressed in colon cells, especially in well-differentiated cells 28. In our research, KLF4 was verified to be significantly downregulated in colon cancer cells. Additionally, when KLF4 was overexpressed, cancer cell proliferation was greatly repressed, indicating the antitumor role of KLF4 overexpression in colon cancer. Some literatures have reported underlying mechanisms of KLF4 in cancers. For example, miR-10b mediates colorectal cancer cell metastasis and proliferation via targeting KLF4 29, which is consistent with our study. Furthermore, high KLF4 expression plays an inhibitory role in colon cancer development through suppressing IFITM3, which is an interferon-inducible gene overexpressed in colorectal cancer 15. In the present study, we conducted Western blot to assay the protein expression of IFITM3 and also found that IFITM3 was significantly downregulated with the presence of KLF4 overexpression.

Furthermore, targeted binding sites of miR-152-3p on KLF4 were predicted through bioinformatics methods, and miR-152-3p was verified to suppress KLF4 via qRT-PCR and Western blot. Dual-luciferase reporter gene assay indicated that KLF4 was a direct target of miR-152-3p, and that was in agreement with the report on glioma 24. In the cell level, we found that silencing miR-152-3p was observed to potentiate the synthesis of the KLF4 protein, resulting in the inhibition of cell proliferation and the promotion of apoptosis. Moreover, the trend of the expression of miR-152-3p and IFITM3 was demonstrated to be consistent. Rescue experiments elucidated that the expression of IFITM3 was remarkably restored when miR-152-3p and KLF4 were simultaneously silenced, indicating that miR-152-3p might affect the IFITM3 expression by targeting KLF4, thus regulating colon cancer cell proliferation and growth.

In conclusion, our study explored a miR-152-3p-dependent mechanism by which miR-152-3p affects colon cancer progression via the KLF4/IFITM3 axis, which was never studied before. Our findings provide potent reference for the molecular targeted therapy towards colon cancer.

Data Availability

The data used to support the findings of this study are included within the article. The data and materials in the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

All authors contributed to the data analysis and drafting and revising of the article, gave the final approval of the version to be published, and agreed to be accountable for all aspects of the work.

References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] R. Zheng, H. Zeng, S. Zhang, and W. Chen, "Estimates of cancer incidence and mortality in China, 2013," *Chinese Journal of Cancer*, vol. 36, no. 1, p. 66, 2017.
- [3] A. B. Benson, A. P. Venook, L. Cederquist et al., "Colon Cancer, Version 1.2017, NCCN Clinical Practice Guidelines in Oncology," *Journal of the National Comprehensive Cancer Network*, vol. 15, no. 3, pp. 370–398, 2017.
- [4] F. Lordick and J. Mossner, "Aktuelle standards in der diagnostik und therapie des kolonkarzinoms," *Deutsche Medizinische Wochenschrift*, vol. 142, no. 7, pp. 487–490, 2017.
- [5] D. Liska, L. Stocchi, G. Karagkounis et al., "Incidence, patterns, and predictors of locoregional recurrence in colon cancer," *Annals of Surgical Oncology*, vol. 24, no. 4, pp. 1093–1099, 2017.
- [6] J. P. Katz, N. Perreault, B. G. Goldstein et al., "The zinc-finger transcription factor Klf 4 is required for terminal differentiation of goblet cells in the colon," *Development*, vol. 129, no. 11, pp. 2619–2628, 2002.
- [7] Y. Li, J. McClintick, L. Zhong, H. J. Edenberg, M. C. Yoder, and R. J. Chan, "Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf 4," *Blood*, vol. 105, no. 2, pp. 635–637, 2005.
- [8] L. Z. Wei, Y. Q. Wang, Y. L. Chang et al., "Imbalance of a KLF4-mi R-7 auto-regulatory feedback loop promotes prostate cancer cell growth by impairing micro RNA processing," *American Journal of Cancer Research*, vol. 8, pp. 226–244, 2018.
- [9] M.-T. Sung, H.-T. Hsu, C.-C. Lee et al., "Krüppel-like factor 4 modulates the migration and invasion of hepatoma cells by suppressing TIMP-1 and TIMP-2," *Oncology Reports*, vol. 34, no. 1, pp. 439–446, 2015.
- [10] F. Yu, J. Li, H. Chen et al., "Krüppel-like factor 4 (KLF4) is required for maintenance of breast cancer stem cells and for cell migration and invasion," *Oncogene*, vol. 30, no. 18, pp. 2161–2172, 2011.
- [11] H. Zhou, Y. Liu, R. Zhu et al., "FBXO32 suppresses breast cancer tumorigenesis through targeting KLF4 to proteasomal degradation," *Oncogene*, vol. 36, no. 23, pp. 3312–3321, 2017.
- [12] L. Li, S. Yu, Q. Wu, N. Dou, Y. Li, and Y. Gao, "KLF4-mediated CDH3 upregulation suppresses human hepatoma cell growth and migration via GSK-3 β signaling," *International Journal of Biological Sciences*, vol. 15, no. 5, pp. 953–961, 2019.

Research Article

Construction of circRNA-Associated ceRNA Network Reveals Novel Biomarkers for Esophageal Cancer

Yunhao Sun,^{1,2} Limin Qiu,² Jinjin Chen,³ Yao Wang,² Jun Qian,² Lirong Huang,² and Haitao Ma¹ 

¹Department of Thoracic Surgery, The First Affiliated Hospital of Soochow University, China

²Department of Thoracic Surgery, Yancheng City No.1 People's Hospital, China

³Oncology Department, Yancheng City No.1 People's Hospital, China

Correspondence should be addressed to Haitao Ma; haitao_ma110@163.com

Received 7 June 2020; Accepted 27 July 2020; Published 28 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Yunhao Sun 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.

Objective. Esophageal cancer (ESCC) is reported to be the eighth most common malignant tumors worldwide with high mortality. However, the functions of majority circRNAs in ESCC requires to be further explored. **Methods.** This study identified differently expressed circRNAs in 3 paired ESCC using RNA-sequencing method. The interactions among circRNAs, miRNAs, and mRNAs were predicted using bioinformatics analysis. **Results.** In this study, using RNA-sequencing method and integrated bioinformatics analysis, 418 overexpressed circRNAs and 637 reduced circRNAs in ESCC sample were identified. Based on the mechanism that circRNAs could play as ceRNAs to modulate targets expression, circRNA-miRNA and circRNA-miRNA-mRNA networks were constructed in this study. Based on the network analysis, 7 circRNAs, including circ_0002255, circ_0000530, circ_0001904, circ_0001005, circ_0000513, circ_0000075, and circ_0001121, were identified as key circRNAs in ESCC. We found that circ_0002255 was related to the regulation of substrate adhesion-dependent cell spreading. circ_0001121 was involved in regulating nucleocytoplasmic transport. circ_0000513 played a key role in regulating Adherens junction, B cell receptor signaling pathway. Meanwhile, we observed circ_0000075 was involved in regulating zinc II ion transport, transition metal ion homeostasis, and angiogenesis. **Conclusion.** We thought this study could provide novel biomarkers for the prognosis of ESCC.

1. Introduction

In recent years, the functional importance of noncoding RNAs (ncRNAs) in the tumorigenesis and the development of cancers have been found. CircRNAs are a type of special endogenous RNA molecules [1]. With the development of high-throughput RNA sequencing, circRNAs were found to be present in human cells. Emerging reports have revealed the important roles of circRNAs in multiple human diseases, such as malignant tumors [1–3]. The findings indicated that circRNAs were abnormally expressed and involved in regulating cancer proliferation and therapy resistance through various mechanisms, such as sponging miRNAs or proteins, and regulating RNA splicing and transcription [3–5].

Esophageal cancer (ESCC) is reported to be the eighth most common malignant tumors worldwide with high mor-

tality [6, 7]. Previous studies showed more than 455800 patients were diagnosed with ESCC, and almost 400200 patients died from this disease [8]. Despite novel methods, such as radiotherapy and chemotherapy, were used in the ESCC treatment, the five-year survival rate of ESCC patients is as low as about 25% due to distant metastasis and therapy resistance [9, 10]. It is therefore of great importance to explore an effective treatment to prevent ESCC progression.

A number of reports have indicated that circRNAs were related to the development of ESCC. A report by Chen et al. showed circLARP4 suppressed ESCC progression via sponging miR-1323 and modulating PI3K signaling [11]. Another study by Pan et al. found that hsa_circ_0006948 modulated miR-490-3p/HMGA2 axis, thus regulating tumorigenesis and EMT processes in ESCC [12]. Moreover, the special expression pattern of circRNAs in ESCC was

validated as potential biomarkers for the prognosis of this disease. For example, hsa_circRNA_100873 upregulation was correlated to lymphatic metastasis of ESCC [13], and Circ-SLC7A5 was validated as a potential prognostic circulating biomarker for detection of ESCC, which was correlated to advanced stage and worse prognosis [14]. Despite a few studies revealed the functions of circRNAs in ESCC [15, 16], the functions of majority circRNAs require to be further explored.

Recently, the progress in RNA-sequencing method had expanded the understanding of the molecular mechanism of cancers. A series of novel mRNAs and noncoding RNAs were revealed to be related to the tumorigenesis. For example, Li et al. revealed that circDDX17 was downregulated in colorectal cancer with RNA sequencing and suppressed tumor development [17]. Huang et al. reported abundant mRNA, circRNA, and lncRNA in blood could act as diagnostic markers for cancers by using extracellular vesicles long RNA sequencing [18]. Yu et al. found hsa_circ_0001445 was identified to be downregulated by RNA-sequencing and suppress liver cancer metastasis [19]. Also, using RNA-sequencing method could provide novel biomarkers for ESCC.

This study identified differently expressed circRNAs in ESCC using RNA-sequencing method. The interactions among circRNAs, miRNAs, and mRNAs were predicted using bioinformatics analysis. We thought this study was able to provide novel biomarkers for ESCC.

2. Materials and Methods

2.1. Tissue Specimens. Three paired ESCC tissues and adjacent normal tissues were collected from patients who received radical gastrectomy at the Department of Thoracic Surgery, The First Affiliated Hospital of Soochow University, from 2019 to 2020. All specimens were collected under the guidance of the HIPAA protocol and supervised by the ethics committee. TNM stage classification complied with the TNM classification system of the International Union Against Cancer (7th edition). These patients were diagnosed with ESCC with average age: 62.7.

2.2. RNA-seq Analysis. The total RNA was isolated with RNAsiso Plus (TaKaRa Japan). The Ribo-Zero rRNA Removal Kit (Illumina, San Diego, CA, USA) and the CircRNA Enrichment Kit (Cloud-seq, USA) were used to remove the rRNA and enrich the circRNAs. The RNA-seq libraries were constructed by using TruSeq Stranded Total RNA Library Prep Kit (Illumina, San Diego, CA, USA). The libraries were denatured as single-stranded DNA molecules, captured on Illumina flow cells, amplified in situ as clusters, and finally sequenced for 150 cycles on Illumina HiSeq™ 4000 Sequencer (Illumina, San Diego, CA, USA). All these assays were conducted according to the manufacturer's instructions. The raw data were listed as a supplementary table 1.

2.3. Identification and Quantification of Human circRNAs. For each sample, the cleaned RNA-seq reads were first

mapped to the human reference genome (GRCh37/hg19, UCSC Genome Browser [20]) by TopHat2 [21]. Then, the unmapped reads of each sample in the TopHat2 results were used to identify the circRNAs by UROBORUS pipeline [22].

2.4. Differential Expression Analysis. Differentially expressed circRNAs between ESCC and normal samples were determined using the “limma” package (3.38.3) in R (5.3.2) [23, 24]. A paired Student's *t*-test was used to identify any significant differences in circRNA expression between tumor and tumor-adjacent normal tissues. The thresholds of fold-change >2 were set to screen the significantly DESCCs.

2.5. Functional Analysis. Bioinformatics analysis was conducted using the DAVID online database (<https://david.ncifcrf.gov/home.jsp>) [25]. The results were visualized by the imageGP online software (<http://www.ehbio.com/ImageGP/index.php/Home/Index/index.html>).

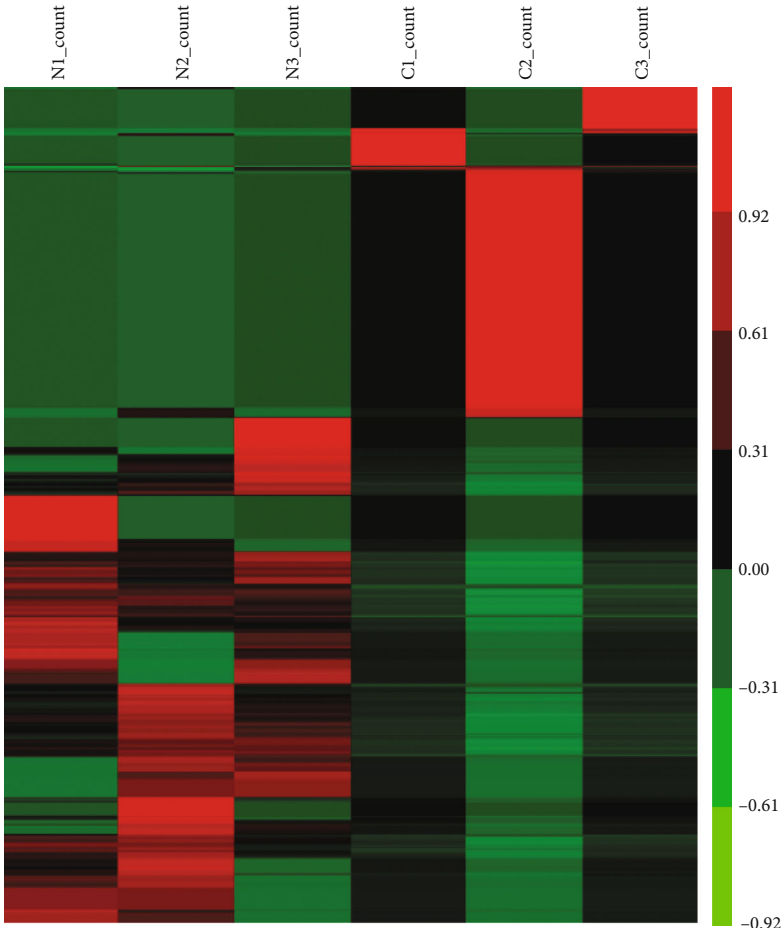
2.6. Correlation Analysis of circRNAs and mRNAs in ESCC. An Agilent circRNA and mRNA expression profile microarray was used to screen the differentially expressed circRNA and mRNA gene expression. The regulation of the mRNA target expression of circRNAs was evaluated to investigate whether circRNAs could act as “miRNA sponges.” CircRNA-miRNA interaction analysis was conducted by Cytoscape 3.2.1 software (Cytoscape Consortium). The size of each node represents the number of putative miRNAs that were functionally connected to each circRNA.

3. Result

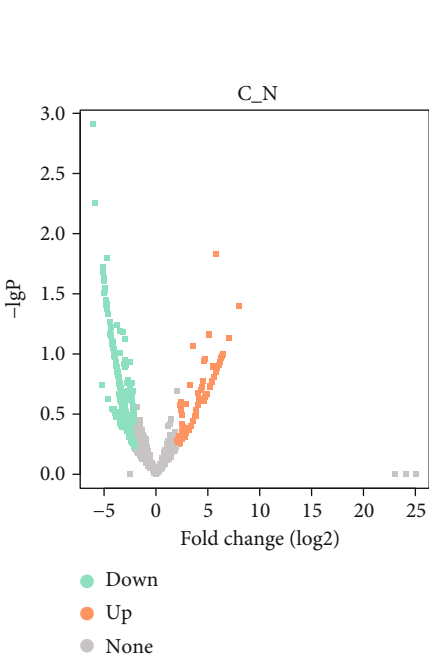
3.1. Identification and Validation of Differentially Expressed circRNAs in ESCC. By analyzing the expression pattern of circRNAs between ESCC tumors and normal tissues, 1055 circRNAs were identified to be differently expressed in ESCC tissues with fold change ≥ 2 (Figures 1(a) and 1(b)). Among these circRNAs, 418 circRNAs were overexpressed, and 637 circRNAs were reduced in ESCC sample compared to normal tissues (Figure 1(c)). Heatmap and volcano plot analysis demonstrated these significant differentially expressed circRNAs (Figures 1(a) and 1(b)).

3.2. Enrichment Analysis of circRNAs' Parental Genes. Furthermore, we perform GO analysis to explore the potential functional roles of circRNAs' parental genes. Our results showed that the top 10 biological processes related to parental genes of differently expressed circRNA included cellular component organization, biosynthetic process, macromolecule biosynthetic process, primary metabolic process, RNA metabolic process, and transcription from RNA polymerase II promoter (Figure 2(a)). Meanwhile, the top 10 molecular functions and cellular components related to these circRNAs' parental genes were shown in Figures 2(b) and 2(c).

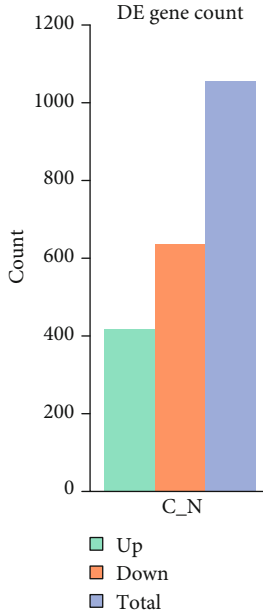
The KEGG analysis revealed that the pathways related to parental genes of differently expressed circRNAs included ErbB signaling pathway, focal adhesion, and lysine degradation (Figure 2(d)).



(a)

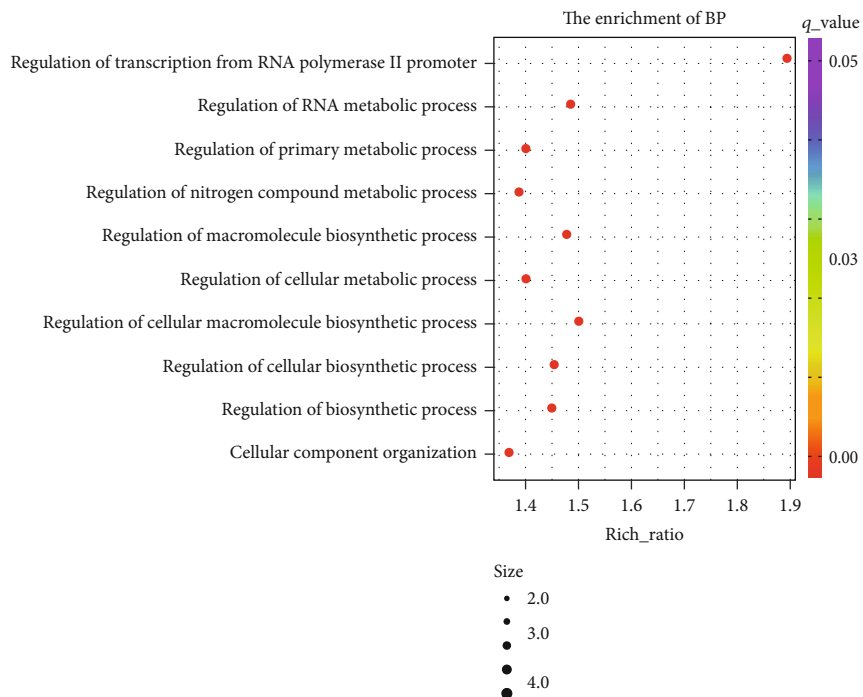


(b)

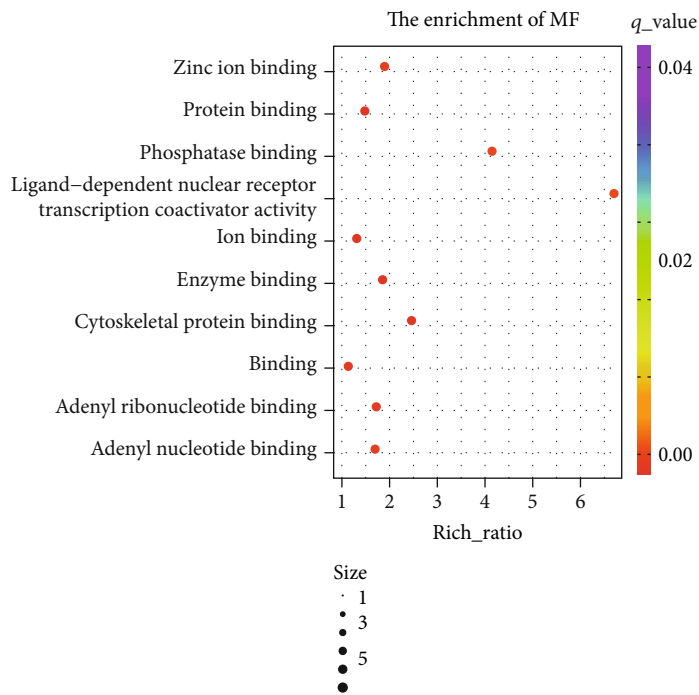


(c)

FIGURE 1: Analysis of differentially expressed circRNAs in ESCC by RNA-sequencing. (a) Heatmap analysis of differentially expressed circRNAs between ESCC and normal groups. (b) The volcano plot analysis of differentially expressed circRNAs between ESCC and normal groups. (c) The summarization of upregulated and downregulated circRNAs between ESCC and normal groups.



(a)



(b)

FIGURE 2: Continued.

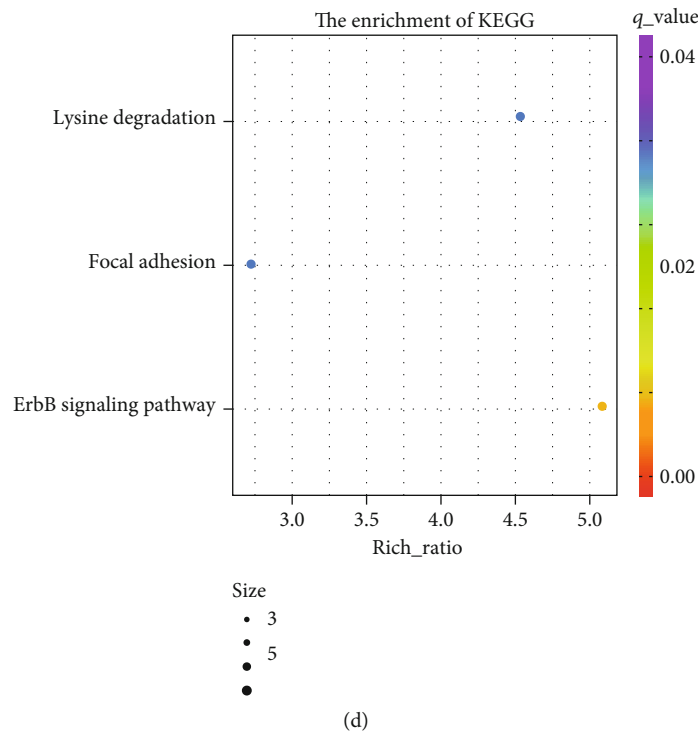
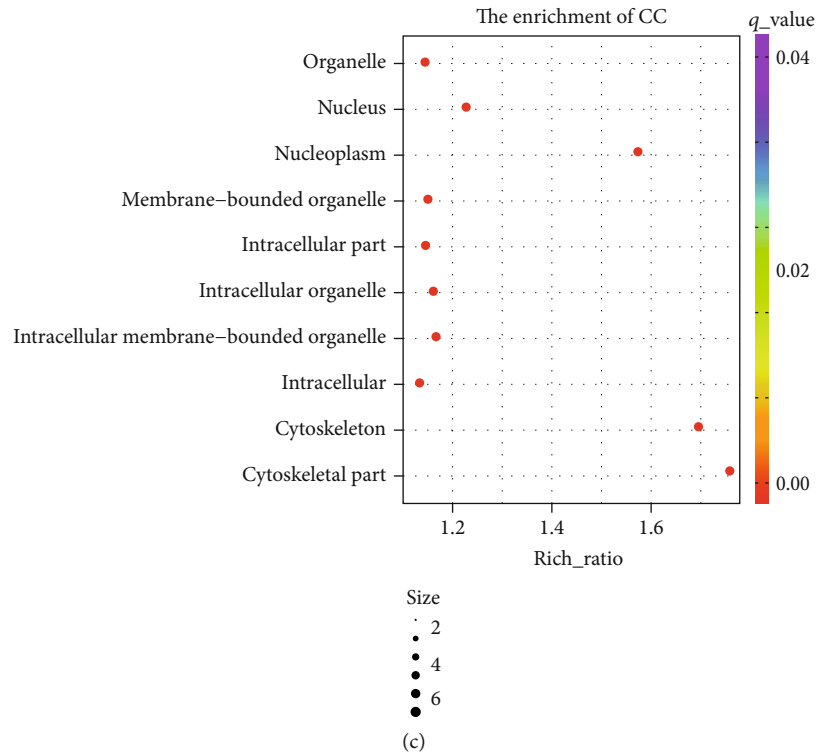


FIGURE 2: In silico analysis of circRNAs' parental genes. (a) Enrichment of the top 10 BP of circRNAs' parental genes. (b) Enrichment of the top 10 MF of circRNAs' parental genes. (c) Enrichment of the top 10 CC of circRNAs' parental genes. (d) Enrichment of the top 10 pathways of circRNAs' parental genes. The size: the number of genes. MF: molecular functions; CC: cellular components; BP: biological processes.

3.3. Construction of circRNA-miRNA-mRNA Network. A number of studies showed circRNAs act as sponges of miRNA to suppress their activities. Therefore, we constructed a circRNA-miRNA interaction network using bioinformatics methods. The interaction between circRNA and

miRNAs was predicted using circinteractome database (<https://circinteractome.nia.nih.gov/>) [26].

Next, we constructed a circRNA-miRNA-mRNA network in ESCC. The miRNA-mRNA pairs were identified using Starbase [27] and TARGETSCAN [28] database. A

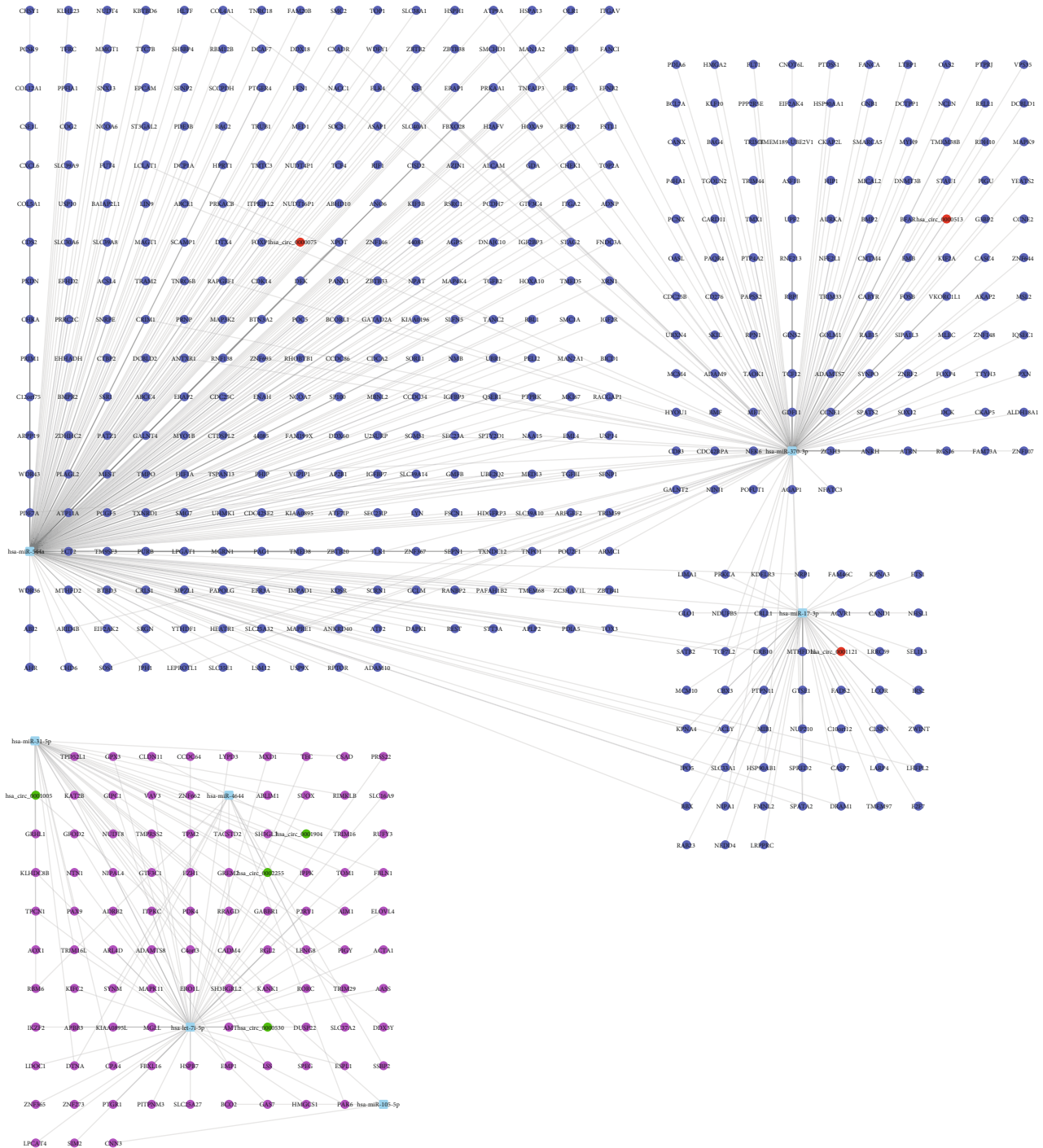


FIGURE 3: Construction of circRNA associated ceRNA network. Green circle: upregulated circRNAs; red circle: downregulated circRNAs; blue circle: miRNAs; purple circle: upregulated mRNAs; deep blue circle: downregulated mRNAs.

total of 8975 mRNAs were identified as potential circRNA-miRNA targets. Then, we extracted differently expressed mRNAs in ESCC using GEPIA database [29]. Finally, ESCC specific circRNA associated ceRNA network was constructed with Cytoscape 3.6.1 software [30], which included 7 circRNAs (circ_0002255, circ_0000530, circ_0001904, circ_0001005, circ_0000513, circ_0000075, circ_0001121), 7 miRNAs (hsa-miR-31-5p, hsa-let-7i-5p, hsa-miR-4644,

hsa-miR-105-5p, hsa-miR-370-3p, hsa-miR-544a, hsa-miR-17-3p), and 548 mRNAs (Figures 3(a) and 3(b)).

3.4. Enrichment Analysis of Key circRNAs in This Network. Next, we conducted the bioinformatics analysis of Key circRNAs in this network using Clue-GO plugin [31] in Cytoscape 3.6.1 software. The results revealed that hsa_circ_0002255 was related to the regulation of substrate

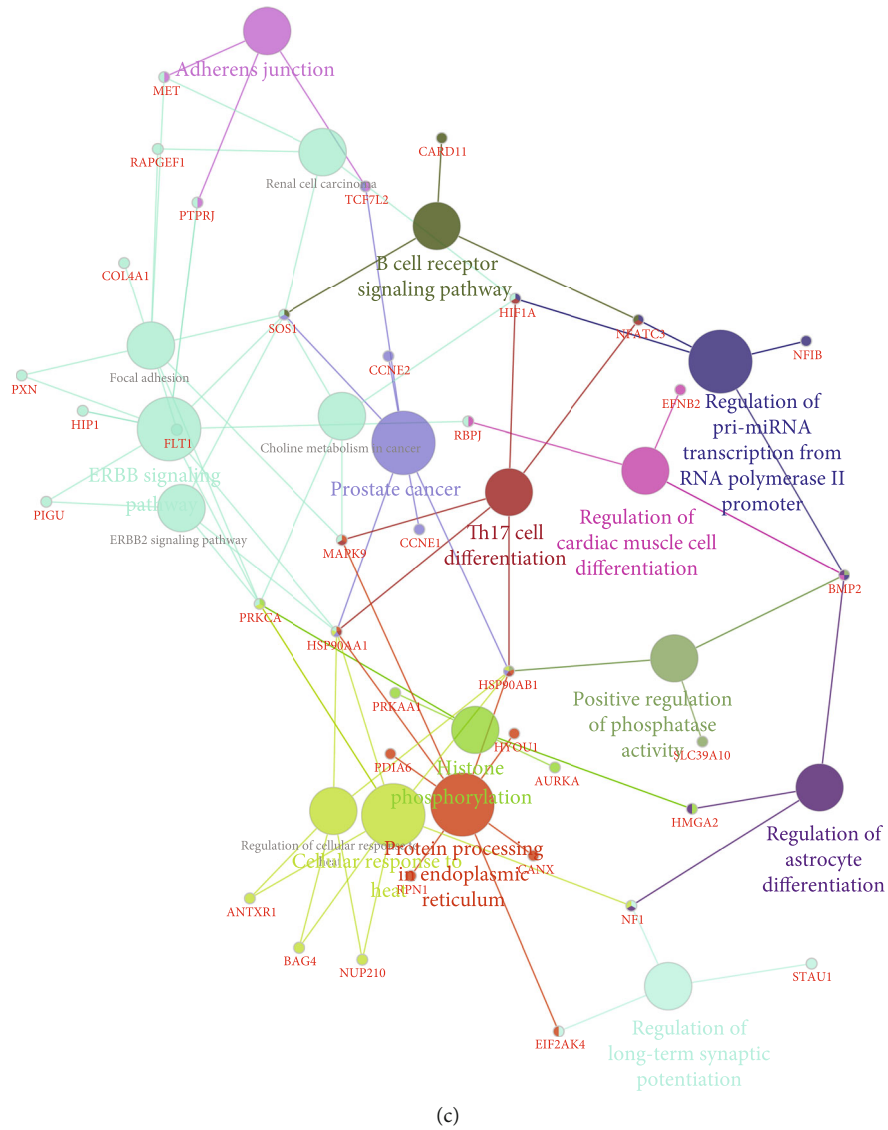
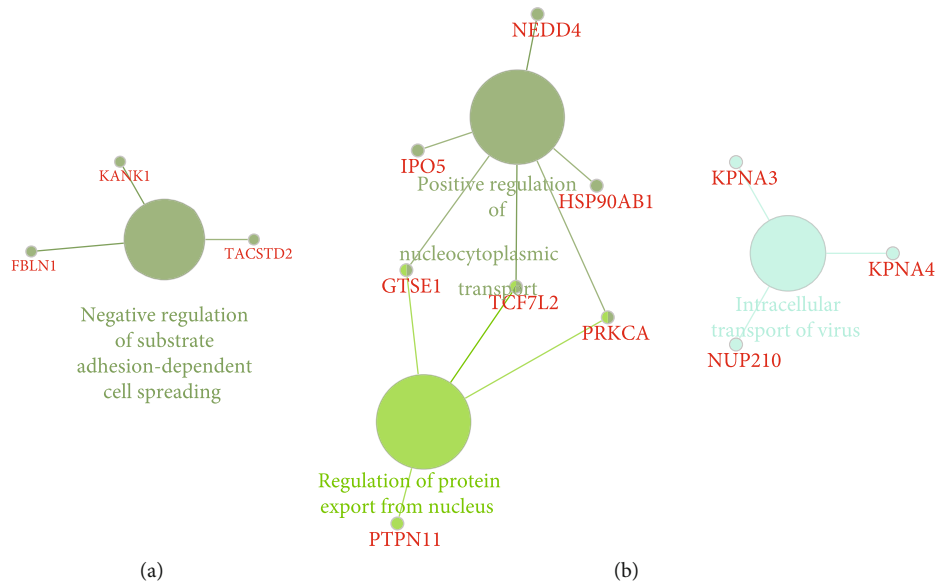


FIGURE 4: Continued.

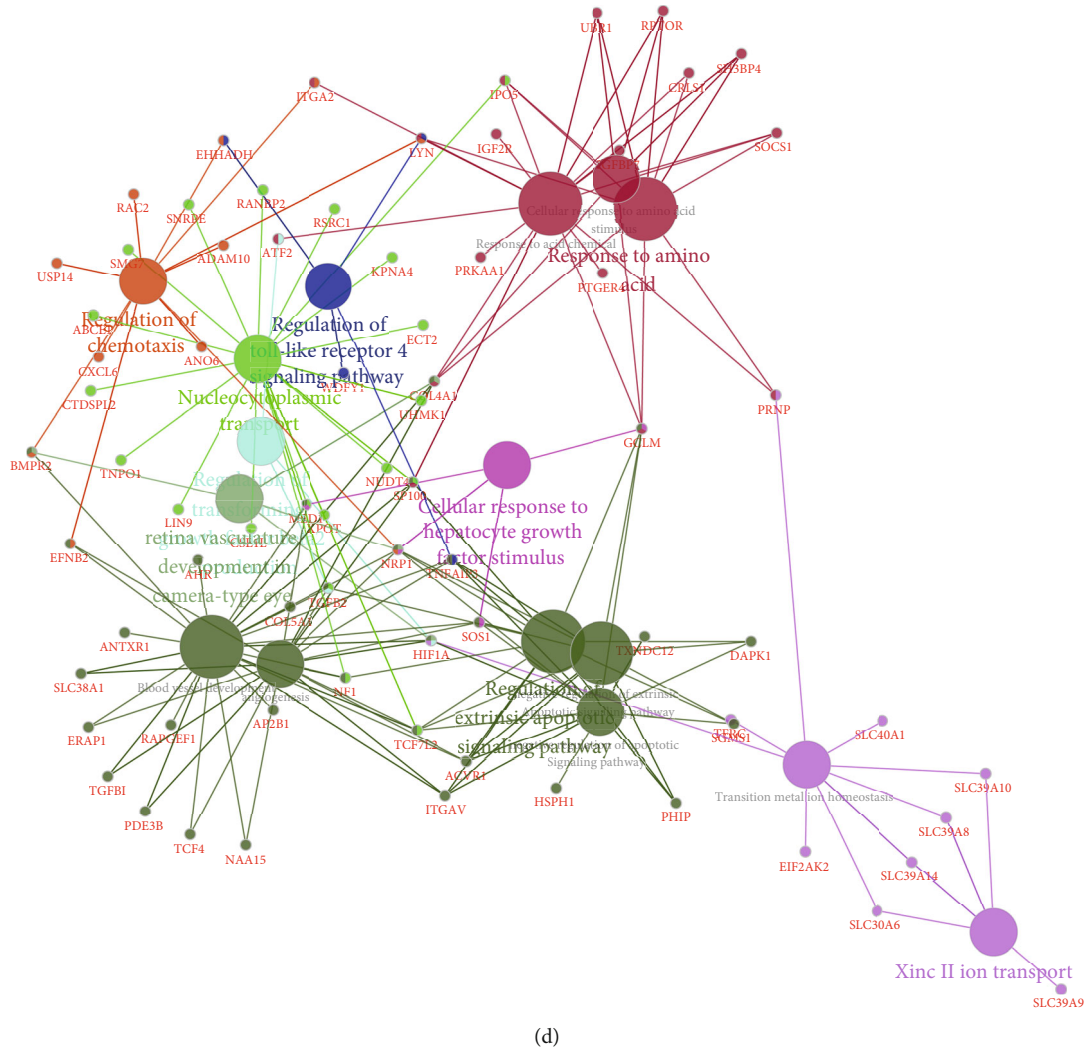


FIGURE 4: Enrichment analysis of key circRNAs. (a) Enrichment of hsa_circ_0002255 in esophageal cancer. (b) Enrichment of hsa_circ_0001121 in esophageal cancer. (c) Enrichment of hsa_circ_0000513 in esophageal cancer. (d) Enrichment of hsa_circ_0000075 in esophageal cancer. The circle: biological pathways; the dots: genes.

adhesion-dependent cell spreading (Figure 4(a)). hsa_circ_0001121 was involved in regulating nucleocytoplasmic transport and protein export from nucleus (Figure 4(b)).

Moreover, we identified hsa_circ_0000513 played a key role in regulating Adherens junction, B cell receptor pathway, ERBB signaling, pri-miRNA transcription, regulation of phosphatase activity, histone phosphorylation, and protein processing in endoplasmic reticulum (Figure 4(c)). Among these pathways, we specially indicated that ERBB signaling was potentially regulated by this circRNA via PTPRJ, SOS1, HIP1, PXN, and PIGU.

Meanwhile, we observed hsa_circ_0000075 was involved in regulating zinc II ion transport, transition metal ion homeostasis, angiogenesis, blood vessel development, extrinsic apoptotic signaling pathway, response to amino acid, nucleocytoplasmic transport, response to acid chemical, toll-like receptor 4, cellular response to hepatocyte growth factor stimulus, chemotaxis transforming, and growth factor beta2 production (Figure 4(d)).

3.5. *The Dysregulation of Key miRNAs Was Related to the Survival Time in ESCC.* Next, we predicted the prognostic value of key miRNAs in ESCC with TCGA data. The results showed that higher expression level of hsa-let-7i (Figure 5(a)), hsa-mir-4644 (Figure 5(b)), hsa-mir-17 (Figure 5(c)), hsa-mir-544a (Figure 5(d)), hsa-mir-105 (Figure 5(e)) were associated with shorter overall survival time in ESCC patients.

4. Discussion

Recently, the circRNAs have been reported to be related to ESCC. CircRNA dysregulation was related to prognosis and tumor proliferation regulation of ESCC. For example, Zhang et al. revealed 2,046 circRNAs were frequently altered in ESCC tissues [32]. Su et al. identified 57 induced circRNAs and 17 reduced circRNAs in radioresistant ESCC cells compared to normal ESCC cells [33]. Also, the special functions of several circRNAs had been clearly demonstrated. For

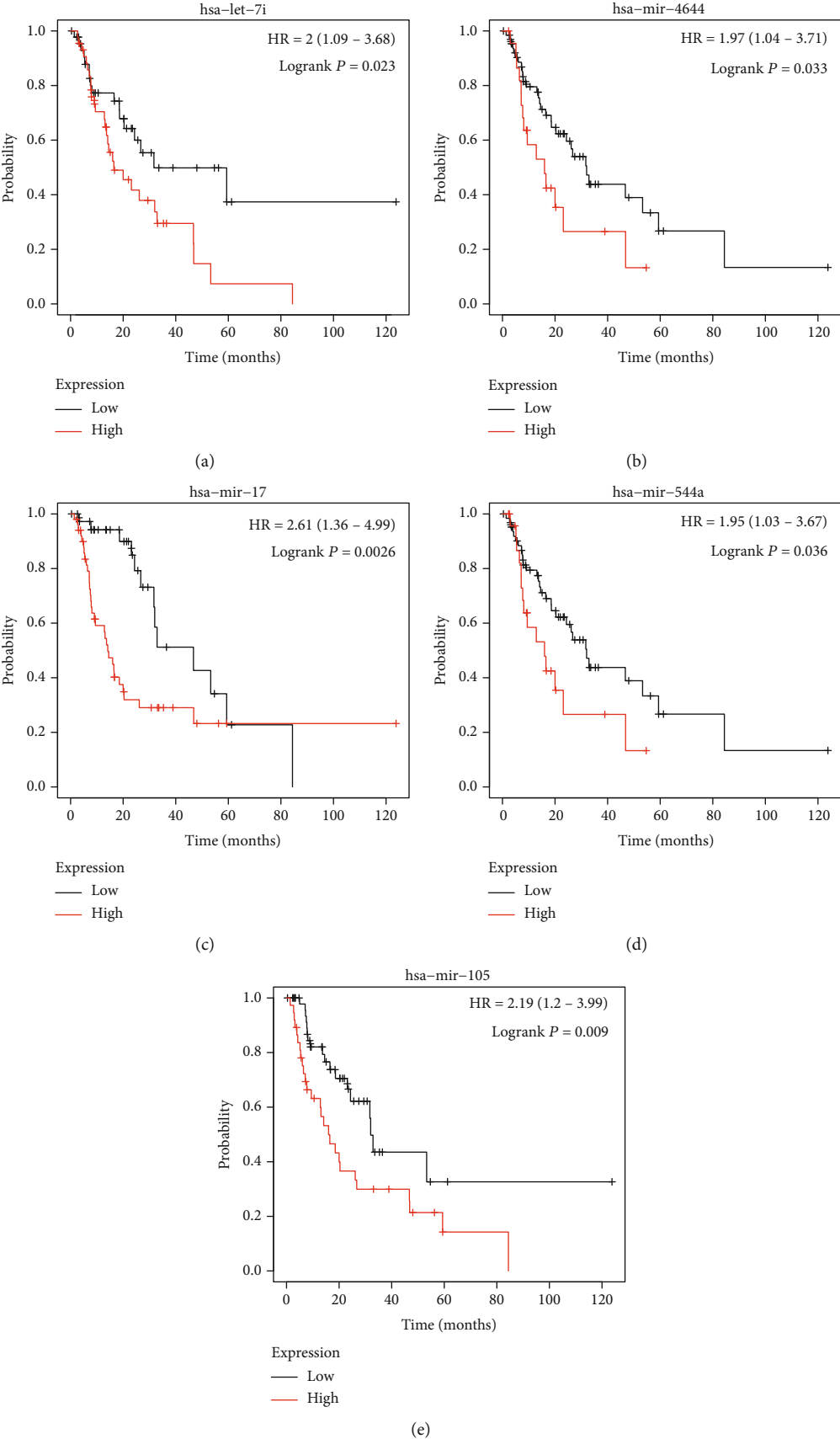


FIGURE 5: The dysregulation of key miRNAs was related to the survival time in ESCC (a-e) higher expression level of hsa-let-7i (a), hsa-mir-4644 (b), hsa-mir-17 (c), hsa-mir-544a (d), hsa-mir-105 (e) were associated with shorter overall survival time in ESCC patients.

example, CiRS-7 promotes growth and metastasis of ESCC via regulation of miR-7/HOXB13 [34]. However, these studies just revealed a limited amount of circRNAs in ESCC. According to circBase database, more than 50000 circRNAs existed in human cells [35]. Therefore, this was still an urgent need to identify differently expressed circRNAs in ESCC to expand our understanding of the mechanism related to ESCC development. In this study, using RNA-sequencing method and integrated bioinformatics analysis, 418 overexpressed circRNAs and 637 reduced circRNAs in ESCC sample were identified. Based on the mechanism that circRNAs could play as ceRNAs to modulate targets expression [36, 37], circRNA-miRNA and circRNA-miRNA-mRNA networks were constructed in this study. Based on the network analysis, 7 circRNAs, including circ_0002255, circ_0000530, circ_0001904, circ_0001005, circ_0000513, circ_0000075, and circ_0001121, were identified as key circRNAs in ESCC. We found that circ_0002255 was related to the regulation of substrate adhesion-dependent cell spreading. circ_0001121 was involved in regulating nucleocytoplasmic transport. circ_0000513 played a key role in regulating Adherens junction, B cell receptor signaling pathway. Meanwhile, we observed circ_0000075 was involved in regulating zinc II ion transport, transition metal ion homeostasis, and angiogenesis.

CircRNAs have been shown to function as regulators of parental gene transcription and alternative splicing and miRNA sponges. Exon-intron circular RNAs (EIciRNAs) hold U1 snRNP through interaction with U1 snRNA, and then, the EIciRNA-U1 snRNP complexes further interact with Pol II transcription complex at the promoters of parental genes to enhance gene transcription and expression [38, 39]. Zhang et al. [39, 40] found that circEIF3J and circPAIP2 with higher expression levels can complement U1 and interact with U1 small ribonucleoprotein to promote the transcription of EIF3J and PAIP2 genes in cis. Intronic circRNAs (CiRNAs) also positively regulate Pol II transcription. For example, ci-ankrd52, generated from gene ANKRD52, is capable of accumulating to its transcription sites and regulates elongation Pol II machinery acting as a positive regulator for transcription [39]. Moreover, circRNAs could act as ceRNAs to affect parental gene expression. For example, circ-VANGL1 as a competing endogenous RNA modulates VANGL1 expression via miR-605-3p [41]. Thus, prediction of the molecular functions related to circRNAs' parental genes could provide more clues to understand the potential functions of circRNAs. The present study showed the pathways related to parental genes of differently expressed circRNAs included ErbB signaling pathway, focal adhesion, and lysine degradation.

Recently, circRNA-mediated ceRNA pathways played a crucial role in cancer initiation and development. For example, circRNA-UCK2 suppressed prostate cancer viability and metastasis through sponging miRNA-767-5p [42]. circFOXO3 was found to promote prostate cancer and glioma progression through sponging miR-29a-3p [43] and miR-138-5p [44]. CircPTPRA suppressed bladder cancer via sponging miR-636 [45]. Also, several cancer-related ceRNA networks were identified. Song et al. constructed a colorectal

cancer-related ceRNA network, which includes 13 circRNAs, 62 miRNAs, and 301 mRNAs [46]. In this study, we for the first time built a miRNA-mRNA network in ESCC, containing 33 circRNAs and 158 miRNAs. hsa_circ_0001904, hsa-miR-1273g-3p, hsa-miR-6089, hsa-miR-6873-3p, hsa-miR-8485, and hsa-miR-939-5p were identified as key regulators in ESCC. miR-1273g was found to suppress colorectal cancer proliferation via activation of AMPK signaling [47]. hsa-miR-6089 played a crucial role in regulating inflammation through regulating TLR4 [48, 49]. miR-939 had been revealed to be a key regulator in human cancers, including lung cancer [50], colorectal cancer [51], tongue squamous cell carcinoma [52], epithelial ovarian cancer [53, 54], and gastric cancer [55]. Overexpression of this miRNA enhanced lung cancer progression [50]. In gastric cancer, knockdown of miR-939 modulated metastasis and chemoresistance via dysregulation of SLC34A2 and Raf/MEK/ERK pathway [55].

Also, we built an ESCC-related circRNA-miRNA-mRNA network, including 7 circRNAs, 7 miRNAs, and 548 mRNAs. Very interestingly, bioinformatics analysis showed that hsa_circ_0000513 played a key role in regulating ERBB signaling pathway, regulation of pri-miRNA transcription, and histone phosphorylation in endoplasmic reticulum. ERBB signaling pathway was reported to be activated in ESCC [56, 57]. For example, inhibitors of ERBB signaling were found to suppress ESCC cell migration. miRNAs played an important role in ESCC via affecting cell growth, migration, and autophagy. Very interestingly, we showed hsa_circ_0000513 may affect miRNA functions through modulating their transcription. A recent study showed hsa_circ_0000075 participated in the AF pathogenesis via TGF-beta signaling pathway. However, the roles of hsa_circ_0000075 in ESCC remained unclear. The present study showed that hsa_circ_0000075 was involved in regulating angiogenesis, blood vessel development, and regulation of extrinsic apoptotic signaling pathway.

Despite this study identified differently expressed circRNAs and predicted their functions in ESCC with bioinformatics method, several limitations should be noted. Firstly, the molecular functions and mechanisms of these circRNAs should be further confirmed using experimental assays. Secondly, the prognostic value of key circRNAs should be further explored. The correlation between circRNAs expression and tumor stage, survival time should be further evaluated with collected clinical samples. Finally, the raw data of the Ribo-zero library-based RNA-seq data should be further analyzed to confirm circRNA-mRNA interaction in the future study.

In our study, we identified 418 overexpressed circRNAs and 637 downregulated circRNAs in ESCC and conducted bioinformatics to reveal the potential mechanisms and molecular functions of these circRNAs in ESCC. We thought this study could provide novel biomarkers for the prognosis of ESCC.

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 competing interests.

Authors' Contributions

Yunhao Sun and Limin Qiu contributed equally to this work.

Supplementary Materials

Supplementary Table 1: the different expressed circRNAs in ESCC. (*Supplementary Materials*)

References

- [1] S. Memczak, M. Jens, A. Elefsinioti et al., "Circular RNAs are a large class of animal RNAs with regulatory potency," *Nature*, vol. 495, no. 7441, pp. 333–338, 2013.
- [2] W. R. Jeck, J. A. Sorrentino, K. Wang et al., "Circular RNAs are abundant, conserved, and associated with ALU repeats," *RNA*, vol. 19, no. 2, pp. 141–157, 2013.
- [3] L. S. Kristensen, M. S. Andersen, L. V. W. Stagsted, K. K. Ebbesen, T. B. Hansen, and J. Kjems, "The biogenesis, biology and characterization of circular RNAs," *Nature Reviews. Genetics*, vol. 20, no. 11, pp. 675–691, 2019.
- [4] C. Nicot, "RNA-Seq reveal the circular RNAs landscape of lung cancer," *Molecular Cancer*, vol. 18, no. 1, p. 183, 2019.
- [5] J. N. Vo, M. Cieslik, Y. Zhang et al., "The landscape of circular RNA in cancer," *Cell*, vol. 176, no. 4, pp. 869–881.e13, 2019.
- [6] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *International Journal of Cancer*, vol. 127, no. 12, pp. 2893–2917, 2010.
- [7] G. Abbas and M. Krasna, "Overview of esophageal cancer," *Annals of Cardiothoracic Surgery*, vol. 6, no. 2, pp. 131–136, 2017.
- [8] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: a Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.
- [9] A. Rustgi and H. B. El-Serag, "Esophageal carcinoma," *The New England Journal of Medicine*, vol. 372, no. 15, pp. 1472–1473, 2015.
- [10] A. K. Rustgi and H. B. El-Serag, "Esophageal carcinoma," *The New England Journal of Medicine*, vol. 371, no. 26, pp. 2499–2509, 2014.
- [11] Z. Chen, N. Yao, H. Gu et al., "Circular RNA_LARP4 sponges miR-1323 and hampers progression of esophageal squamous cell carcinoma through modulating PTEN/PI3K/AKT pathway," *Digestive Diseases and Sciences*, vol. 371, no. 8, pp. 1–12, 2020.
- [12] Z. Pan, J. Lin, D. Wu et al., "Hsa_circ_0006948 enhances cancer progression and epithelial-mesenchymal transition through the miR-490-3p/HMGA2 axis in esophageal squamous cell carcinoma," *Aging*, vol. 11, no. 24, pp. 11937–11954, 2019.
- [13] B. Zheng, Z. Wu, S. Xue et al., "hsa_circRNA_100873 upregulation is associated with increased lymphatic metastasis of esophageal squamous cell carcinoma," *Oncology Letters*, vol. 18, no. 6, pp. 6836–6844, 2019.
- [14] Q. Wang, H. Liu, Z. Liu et al., "Circ-SLC7A5, a potential prognostic circulating biomarker for detection of ESCC," *Cancer Genetics*, vol. 240, pp. 33–39, 2020.
- [15] Z. F. Xie, H. T. Li, S. H. Xie, and M. Ma, "Circular RNA hsa_circ_0006168 contributes to cell proliferation, migration and invasion in esophageal cancer by regulating miR-384/RBBP7 axis via activation of S6K/S6 pathway," *European Review for Medical and Pharmacological Sciences*, vol. 24, no. 1, pp. 151–163, 2020.
- [16] Z. Xu, X. Tie, N. Li, Z. Yi, F. Shen, and Y. Zhang, "Circular RNA hsa_circ_0000654 promotes esophageal squamous cell carcinoma progression by regulating the miR-149-5p/IL-6/STAT3 pathway," *IUBMB Life*, vol. 72, no. 3, 2020.
- [17] X. N. Li, Z. J. Wang, C. X. Ye, B. C. Zhao, Z. L. Li, and Y. Yang, "RNA sequencing reveals the expression profiles of circRNA and indicates that circDDX17 acts as a tumor suppressor in colorectal cancer," *Journal of Experimental & Clinical Cancer Research*, vol. 37, no. 1, p. 325, 2018.
- [18] Y. Li, J. Zhao, S. Yu et al., "Extracellular vesicles long RNA sequencing reveals abundant mRNA, circRNA, and lncRNA in human blood as potential biomarkers for cancer diagnosis," *Clinical Chemistry*, vol. 65, no. 6, pp. 798–808, 2019.
- [19] X. Zhang, H. Zhou, W. Jing et al., "The circular RNA hsa_circ_0001445 regulates the proliferation and migration of hepatocellular carcinoma and may serve as a diagnostic biomarker," *Disease Markers*, vol. 2018, Article ID 3073467, 9 pages, 2018.
- [20] M. Haussler, A. S. Zweig, C. Tyner et al., "The UCSC genome browser database: 2019 update," *Nucleic Acids Research*, vol. 47, no. D1, pp. D853–D858, 2019.
- [21] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, p. R36, 2013.
- [22] X. Song, N. Zhang, P. Han et al., "Circular RNA profile in gliomas revealed by identification tool UROBORUS," *Nucleic Acids Research*, vol. 44, no. 9, article e87, 2016.
- [23] M. E. Ritchie, B. Phipson, Y. H. Di Wu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, article e47, 2015.
- [24] I. Diboun, L. Wernisch, C. A. Orengo, and M. Koltzenburg, "Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma," *BMC Genomics*, vol. 7, no. 1, p. 252, 2006.
- [25] G. Dennis Jr., B. T. Sherman, D. A. Hosack et al., "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 5, p. P3, 2003.
- [26] D. B. Dudekula, A. C. Panda, I. Grammatikakis, S. De, K. Abdelmohsen, and M. Gorospe, "CircInteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs," *RNA Biology*, vol. 13, pp. 34–42, 2015.
- [27] J. H. Li, S. Liu, H. Zhou, L. H. Qu, and J. H. Yang, "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data," *Nucleic Acids Research*, vol. 42, pp. D92–D97, 2013.
- [28] V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *Elife*, vol. 4, 2015.
- [29] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling

- and interactive analyses," *Nucleic Acids Research*, vol. 45, no. W1, pp. W98–W102, 2017.
- [30] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [31] G. Bindea, B. Mlecnik, H. Hackl et al., "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, 2009.
- [32] J. Song, Y. Lu, W. Sun, M. Han, Y. Zhang, and J. Zhang, "Changing expression profiles of lncRNAs, circRNAs and mRNAs in esophageal squamous carcinoma," *Oncology Letters*, vol. 18, no. 5, pp. 5363–5373, 2019.
- [33] H. Su, F. Lin, X. Deng et al., "Profiling and bioinformatics analyses reveal differential circular RNA expression in radioresistant esophageal cancer cells," *Journal of Translational Medicine*, vol. 14, no. 1, p. 225, 2016.
- [34] R. C. Li, S. Ke, F. K. Meng et al., "CiRS-7 promotes growth and metastasis of esophageal squamous cell carcinoma via regulation of miR-7/HOXB13," *Cell Death & Disease*, vol. 9, no. 8, p. 838, 2018.
- [35] P. Glazar, "circBase: a database for circular RNAs," *RNA*, vol. 20, no. 11, pp. 1666–1670, 2014.
- [36] D. W. Thomson and M. E. Dinger, "Endogenous microRNA sponges: evidence and controversy," *Nature Reviews. Genetics*, vol. 17, no. 5, pp. 272–283, 2016.
- [37] Y. Tay, J. Rinn, and P. P. Pandolfi, "The multilayered complexity of ceRNA crosstalk and competition," *Nature*, vol. 505, no. 7483, pp. 344–352, 2014.
- [38] A. C. Panda, S. De, I. Grammatikakis et al., "High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs," *Nucleic Acids Research*, vol. 45, no. 12, article e116, 2017.
- [39] Y. Zhang, X. O. Zhang, T. Chen et al., "Circular intronic long noncoding RNAs," *Molecular Cell*, vol. 51, no. 6, pp. 792–806, 2013.
- [40] S. Qu, X. Yang, X. Li et al., "Circular RNA: a new star of non-coding RNAs," *Cancer Letters*, vol. 365, no. 2, pp. 141–148, 2015.
- [41] Z. Zeng, W. Zhou, L. Duan et al., "Circular RNA circ-VANGL1 as a competing endogenous RNA contributes to bladder cancer progression by regulating miR-605-3p/VANGL1 pathway," *Journal of Cellular Physiology*, vol. 234, no. 4, pp. 3887–3896, 2019.
- [42] Z. Xiang, C. Xu, G. Wu, B. Liu, and D. Wu, "CircRNA-UCK2 increased TET1 inhibits proliferation and invasion of prostate cancer cells via sponge MiRNA-767-5p," *Open Medicine (Wars)*, vol. 14, no. 1, pp. 833–842, 2019.
- [43] Z. Kong, X. Wan, Y. Lu et al., "Circular RNA circFOXO3 promotes prostate cancer progression through sponging miR-29a-3p," *Journal of Cellular and Molecular Medicine*, vol. 24, no. 1, pp. 799–813, 2020.
- [44] S. Zhang, K. Liao, Z. Miao et al., "CircFOXO3 promotes glioblastoma progression by acting as a competing endogenous RNA for NFAT5," *Neuro-Oncology*, vol. 21, no. 10, pp. 1284–1296, 2019.
- [45] Q. He, L. Huang, D. Yan et al., "CircPTPRA acts as a tumor suppressor in bladder cancer by sponging miR-636 and upregulating KLF9," *Aging*, vol. 11, no. 23, pp. 11314–11328, 2019.
- [46] W. Song and T. Fu, "Circular RNA-associated competing endogenous RNA network and prognostic nomogram for patients with colorectal cancer," *Frontiers in Oncology*, vol. 9, p. 1181, 2019.
- [47] F. Wu, F. Liu, L. Dong et al., "miR-1273g silences MAGEA3/6 to inhibit human colorectal cancer cell growth via activation of AMPK signaling," *Cancer Letters*, vol. 435, pp. 1–9, 2018.
- [48] S. Yan, P. Wang, J. Wang et al., "Long non-coding RNA HIX003209 promotes inflammation by sponging miR-6089 via TLR4/NF- κ B signaling pathway in rheumatoid arthritis," *Frontiers in Immunology*, vol. 10, p. 2218, 2019.
- [49] D. Xu, M. Song, C. Chai et al., "Exosome-encapsulated miR-6089 regulates inflammatory response via targeting TLR4," *Journal of Cellular Physiology*, vol. 234, no. 2, pp. 1502–1511, 2019.
- [50] X. Han, C. Du, Y. Chen et al., "Overexpression of miR-939-3p predicts poor prognosis and promotes progression in lung cancer," *Cancer Biomarkers*, vol. 25, no. 4, pp. 325–332, 2019.
- [51] Y. Zhang, X. Liu, Q. Li, and Y. Zhang, "lncRNA LINC00460 promoted colorectal cancer cells metastasis via miR-939-5p sponging," *Cancer Management and Research*, vol. 11, pp. 1779–1789, 2019.
- [52] Y. Chen, Y. Guo, and W. Yan, "lncRNA RP5-916L7.2 correlates with advanced tumor stage, and promotes cells proliferation while inhibits cells apoptosis through targeting miR-328 and miR-939 in tongue squamous cell carcinoma," *Clinical Biochemistry*, vol. 67, pp. 24–32, 2019.
- [53] M. Tang, L. Jiang, Y. Lin et al., "Platelet microparticle-mediated transfer of miR-939 to epithelial ovarian cancer cells promotes epithelial to mesenchymal transition," *Oncotarget*, vol. 8, no. 57, pp. 97464–97475, 2017.
- [54] X. Ying, Q. Li-ya, Z. Feng, W. Yin, and L. Ji-hong, "MiR-939 promotes the proliferation of human ovarian cancer cells by repressing APC2 expression," *Biomedicine & Pharmacotherapy*, vol. 71, pp. 64–69, 2015.
- [55] J. X. Zhang, Y. Xu, Y. Gao et al., "Decreased expression of miR-939 contributes to chemoresistance and metastasis of gastric cancer via dysregulation of SLC34A2 and Raf/MEK/ERK pathway," *Molecular Cancer*, vol. 16, no. 1, p. 18, 2017.
- [56] L. Zhang, J. Ma, Y. Han et al., "Targeted therapy in esophageal cancer," *Expert Review of Gastroenterology & Hepatology*, vol. 10, no. 5, pp. 595–604, 2016.
- [57] P. Gaur, M. P. Kim, and B. J. Dunkin, "Esophageal cancer: recent advances in screening, targeted therapy, and management," *Journal of Carcinogenesis*, vol. 13, no. 1, p. 11, 2014.

Research Article

Discovery of Prognostic Signature Genes for Overall Survival Prediction in Gastric Cancer

Changyuan Meng,^{1,2} Shusen Xia ,^{1,2} Yi He,^{1,2} Xiaolong Tang,^{1,2} Guangjun Zhang,^{1,2} and Tong Zhou ^{1,2}

¹The Second Department of Gastrointestinal Surgery, The Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan, China

²Institute of Hepatobiliary, Pancreatic and Intestinal Disease, North Sichuan Medical College, Nanchong, Sichuan, China

Correspondence should be addressed to Tong Zhou; zhigai92002426745@163.com

Received 28 May 2020; Revised 4 July 2020; Accepted 7 July 2020; Published 25 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Changyuan Meng 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.

Background. Gastric cancer (GC) is one of the most common malignant tumors in the digestive system with high mortality globally. However, the biomarkers that accurately predict the prognosis are still lacking. Therefore, it is important to screen for novel prognostic markers and therapeutic targets. **Methods.** We conducted differential expression analysis and survival analysis to screen out the prognostic genes. A stepwise method was employed to select a subset of genes in the multivariable Cox model. Overrepresentation enrichment analysis (ORA) was used to search for the pathways associated with poor prognosis. **Results.** In this study, we designed a seven-gene-signature-based Cox model to stratify the GC samples into high-risk and low-risk groups. The survival analysis revealed that the high-risk and low-risk groups exhibited significantly different prognostic outcomes in both the training and validation datasets. Specifically, *CGB5*, *IGFBP1*, *OLFML2B*, *RAI14*, *SERPINE1*, *IQSEC2*, and *MPND* were selected by the multivariable Cox model. Functionally, PI3K-Akt signaling pathway and platelet-derived growth factor receptor (PDGFR) were found to be hyperactive in the high-risk group. The multivariable Cox regression analysis revealed that the risk stratification based on the seven-gene-signature-based Cox model was independent of other prognostic factors such as TNM stages, age, and gender. **Conclusion.** In conclusion, we aimed at developing a model to predict the prognosis of gastric cancer. The predictive model could not only effectively predict the risk of GC but also be beneficial to the development of therapeutic strategies.

1. Introduction

Gastric cancer (GC) is the fifth most common malignancies worldwide in 2018, accounting for 5.7% of total new cases and 8.2% of cancer-related deaths [1]. Most GC cases are from developing countries, and increased prevalence in the younger population is observed [2]. The major risk factor for GC is *Helicobacter pylori* infection, and its eradication is considered as the most critical for the prevention of GC [3]. Meanwhile, GC often exhibits a high metastasis rate, and most GC patients are not effectively diagnosed at early stages, where surgical resection could become unavailable, which leads to the generally poor prognoses of GC patients [4]. Therefore, there is an urgent need to focus on accurately

identifying markers of prognostic value, in order to provide personalized treatment strategies and to improve the survival of GC patients.

Thanks to the development in sequencing technologies, the utilization of gene expression data makes it possible to explore the molecular background of GC. GC is considered a heterogeneous disease, and so far, several classifications of molecular subtypes of GC have been established. The genomic studies reveal that mutations in *CDH1*, *ERBB4*, *MET*, and *CD44* are closely associated with poor prognosis in gastric cancer [5, 6]. A recent research has reported 4 molecular subtypes that can be identified using immunohistochemical analysis, the Pentaplex assay and certain gene expression (*VIM*, *ZEB1*, *MDM2*, and *CDKN1A*), which

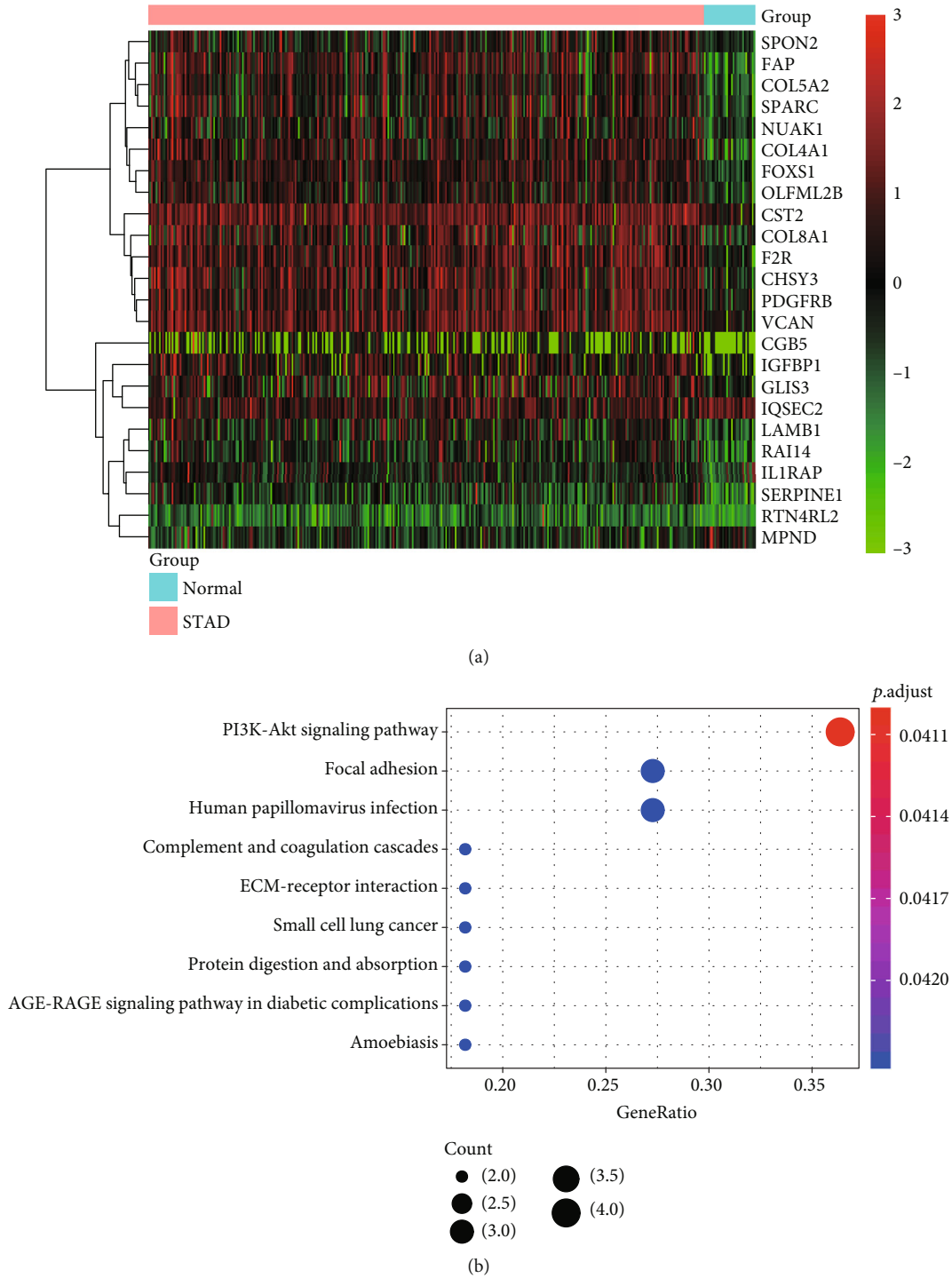


FIGURE 1: The expression patterns and functionalities of prognostic genes in GC. (a) The expression patterns of the 24 prognostic genes selected by differential expression analysis and univariable Cox regression analysis. The expression levels were scaled to -3 to 3. (b) The pathways enriched by the 24 prognostic genes. The node color and size represent the statistical significance and the number of genes included in the pathway.

are the mesenchymal-like type, Microsatellite-unstable type, tumor protein 53- (TP53-) active and TP53-inactive types, each of them characterized by distinctive prognosis and recurrence patterns [7]. A 19-gene signature was developed

to distinguish grades and stages of GC, with an overall accuracy at 79.6%, but among those detected genes, only *CLDN7*, *CLDN1*, and *DPT* exhibited significantly varied expression when compared with normal tissues [8]. Notably, another

TABLE 1: The hazard ratio and statistical significance of the seven signature genes in univariate and multivariate analyses.

Genes	Univariate analysis HR (95% CI)	p value	Multivariate analysis HR (95% CI)	p value
<i>CGB5</i>	1.79 (1.28-2.50)	6.02E-04	1.78 (1.27-2.50)	8.99E-04
<i>IGFBP1</i>	1.66 (1.19-2.30)	2.63E-03	1.33 (0.95-1.87)	1.02E-01
<i>OLFML2B</i>	1.77 (1.27-2.48)	7.60E-04	1.64 (1.13-2.38)	9.36E-03
<i>RAI14</i>	1.82 (1.30-2.54)	4.19E-04	1.35 (0.93-1.96)	1.16E-01
<i>SERPINE1</i>	1.95 (1.39-2.72)	9.32E-05	1.62 (1.14-2.30)	6.66E-03
<i>IQSEC2</i>	0.71 (0.51-0.99)	4.35E-02	0.77 (0.55-1.07)	1.24E-01
<i>MPND</i>	0.65 (0.47-0.91)	1.14E-02	0.73 (0.51-1.03)	7.67E-02

study has presented a prognostic scoring system developed with 53 gene signatures for GC, including well-reported cancer hallmark genes like *FGFR4*, *CEP55*, and *MCM2* [9]. However, the identification of biomarkers with high prognostic efficacy and the establishment of prognostic scoring with fewer but more effective markers are still essential. In the present study, we aimed at identifying a combination of prognostic genes to predict the risk of GC and stratify the samples, which might be beneficial to the development of therapeutic strategies.

2. Materials and Methods

2.1. Data Acquisition. The gene expression data from the Cancer Genome Atlas (TCGA) project [10] were collected from the UCSC Xena database [11]. We only retained 350 gastric cancer and 32 normal tissues with detailed clinical information. The independent validation dataset was collected from Gene Expression Omnibus [12] (GEO) with accession GSE84433. The TCGA dataset was normalized by log-transforming the FPKM (Fragment Per Kilobase Per Million Reads) +1. The microarray gene expression data of GSE84433 was normalized following a previous study [13]. The former dataset was used for selecting genes for model training, and the latter was used to validate the model performance.

2.2. Selection of Prognostic Genes in Gastric Cancer. To select the prognostic genes in gastric cancer, we first conducted differential expression analysis between the gastric cancer and adjacent normal tissues. Wilcoxon rank-sum test and fold change were employed to identify the upregulated and downregulated genes in gastric cancer. The adjusted p value of 0.05 and fold change of 2 were chosen as the thresholds for the differentially expressed genes (DEGs). Furthermore, a univariate Cox regression analysis was conducted to identify those overall survival-associated genes from the DEGs ($p < 0.05$). The optimal combination of prognostic genes was selected by a stepwise method with the R language *step* function. The gene sets with minimal Akaike information criterion (AIC) values were selected as the predictors in the multivariable Cox model.

2.3. Overrepresentation Enrichment Analysis (ORA). The ORA was employed to identify the pathways enriched by a

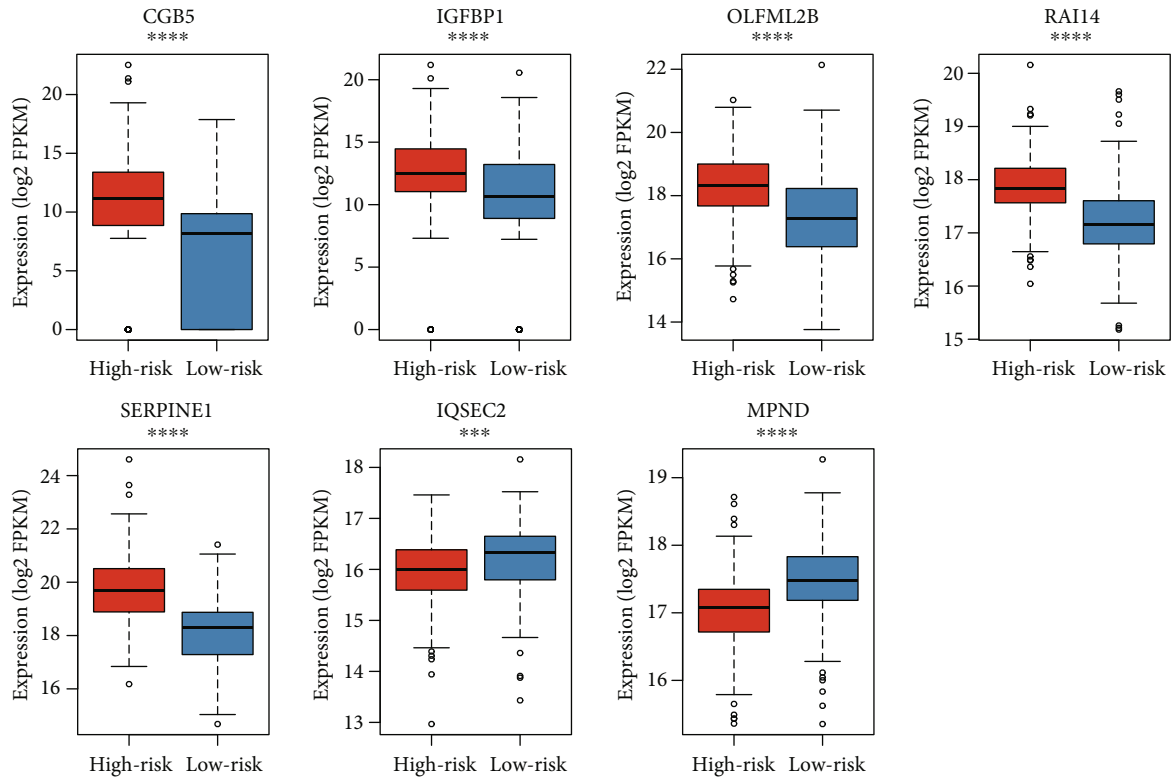
given gene set. The Fisher's exact test was used to test the statistical significance of each pathway. The analysis and visualization was implemented in the R package *clusterProfiler* [14].

2.4. Discovery of Drug-Target. The upregulated genes in the gastric cancer samples with worse prognosis were used to identify the potential therapeutic targets. The drug-target data was curated by R *maftools* package [15] *drugInteractions*, which searched for the drugs based on the genes.

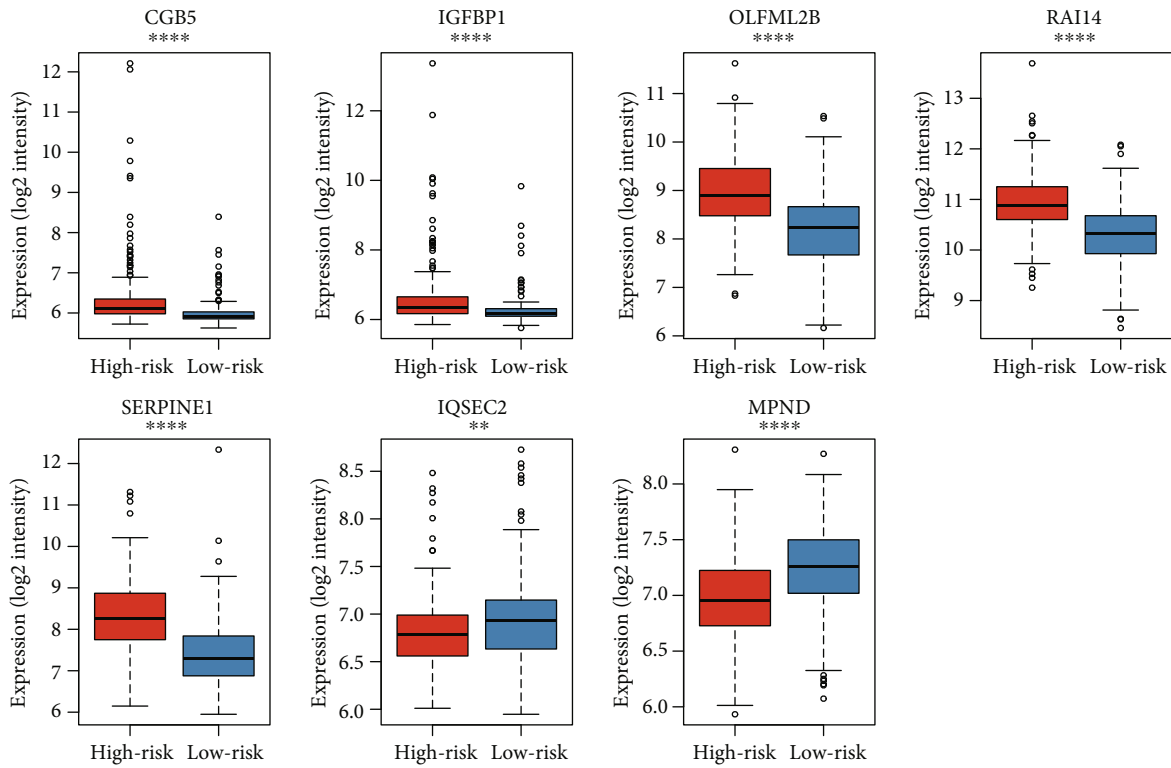
2.5. Survival Analysis. The Cox proportional hazard regression analysis was employed to identify genes associated with the overall survival of gastric cancer. The genes were binarized based on the median of expression levels. The samples were stratified into high-risk and low-risk groups based on the median of risk scores estimated by the Cox model.

3. Results

3.1. Identification of Prognostic Genes in Gastric Cancer. To identify the prognostic genes in gastric cancer, we first collected gene expression data of 350 gastric cancer and 32 normal tissues from the Cancer Genome Atlas (TCGA) project. Subsequently, we conducted a differential expression analysis of the gene expression data by comparing the tumor with the normal tissues. Moreover, we also conducted Cox regression analysis to identify the upregulated and downregulated genes that were associated with overall survival (OS) of the gastric cancer (adjusted p value < 0.05 and fold change > 1). Specifically, we identified a total of 24 prognostic genes in gastric cancer including 22 upregulated and 2 downregulated genes (Supplementary Table S1, Figure 1(a), adjusted p value < 0.05). To reveal the functionality of these genes, we conducted overrepresentation enrichment analysis (ORA) of the 24 prognostic genes and found that these genes were enriched in cancer-related pathways, such as PI3K-Akt signaling pathway, focal adhesion, complement and coagulation cascades, and ECM-receptor interaction. These results indicated that these prognostic genes could not only act as predictors for OS prediction but also be used for interpreting the reason of the worse prognosis in gastric cancer.



(a)



(b)

FIGURE 2: The gene expression levels of the seven gene signatures in the two risk groups. The differential expression levels of the seven prognostic genes between the high-risk and low-risk groups in TCGA (a) and GSE84433 (b) datasets, which were referred to as training and validation datasets, respectively. The red and blue boxes represent the high-risk and low-risk groups. (* < 0.05, ** < 0.01, *** < 0.001, and **** < 0.0001).

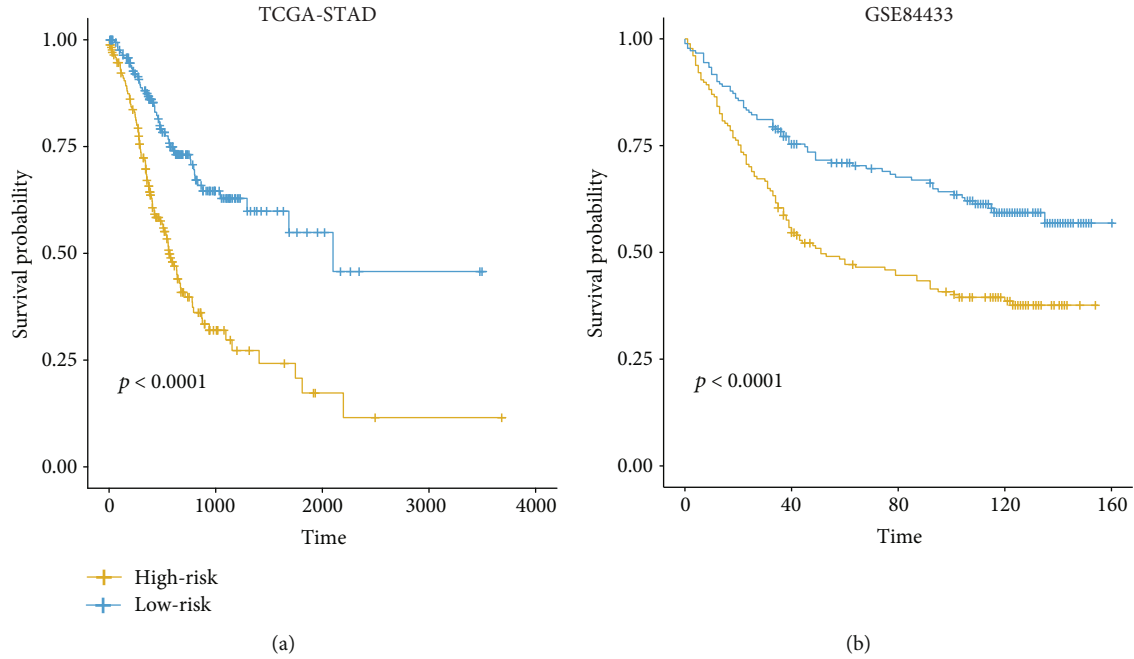


FIGURE 3: The Kaplan-Meier (KM) curves of the two risk groups in the training and validation datasets. The difference of the probabilities of the overall survival in the training (a) and validation (b) datasets. The log-rank test was used to test the differences between the high-risk and low-risk groups. The yellow and blue lines represent the high-risk and low-risk groups.

3.2. Construction and In Silico Validation of Multivariable Cox Model for OS Prediction. With the 24 prognostic genes, a stepwise method was employed to identify a subset of genes in the multivariate analysis. Specifically, *CGB5*, *IGFBP1*, *OLFML2B*, *RAI14*, *SERPINE1*, *IQSEC2*, and *MPND* were selected by the multivariable Cox model (Table 1). The samples in TCGA and the validation cohorts were then stratified into high-risk and low-risk groups by the median of the risk scores. The seven signature genes were observed to be remarkably differentially expressed between the two groups in both TCGA (Figure 2(a)) and the validation cohorts (Figure 2(b)). The log-rank test revealed that the high-risk group had a significantly worse prognosis than the low-risk group (Figure 3(a)). Moreover, the two groups in the validation cohort were also observed to have significantly different prognostic outcomes in the independent dataset (Figure 3(b)). Furthermore, we compared the seven-gene-signature with others by Cui et al. [8] and Wang et al. [9], and our proposed gene signatures exhibited higher performance than the others (Supplementary Table S2). These results suggested that the seven-gene-signature-based Cox model was capable of predicting the overall survival of gastric cancer.

3.3. The Risk Stratification Is an Independent Prognostic Factor in Gastric Cancer. To demonstrate the independence of the risk stratification, we built a multivariable Cox model on the risk stratification with TNM stage, age, and gender as cofactors. Consistently, the risk stratification still maintained higher statistical significance than the TNM stage in the multivariable Cox model (Table 2). Moreover, the

TABLE 2: The multivariate Cox analysis of the risk stratification, TNM stage, age, and gender.

Factors	HR (95% CI)	<i>p</i> value
Risk stratification		
High-risk	1 (reference)	
Low-risk	0.40 (0.28-0.58)	1.41E-06
TNM stage		
I	1 (reference)	
II	1.44 (0.75-2.80)	2.73E-01
III	1.99 (1.07-3.69)	3.02E-02
IV	3.82 (1.86-7.84)	2.58E-04
Age	1.03 (1.01-1.04)	5.81E-03
Gender		
Female	1 (reference)	
Male	1.04 (0.72-1.51)	8.31E-01

older age was an unfavorable factor in gastric cancer. Consistently, we found that high-risk group had a shorter overall survival than the low-risk group in both samples with early stage (I-II) and those with advanced stage (III-IV) (Figures 4(a) and 4(b)). These results indicated that the risk stratification is an independent prognostic factor in gastric cancer.

3.4. The Biomarkers and Pathways Associated with OS in Gastric Cancer. To further interpret the underlying

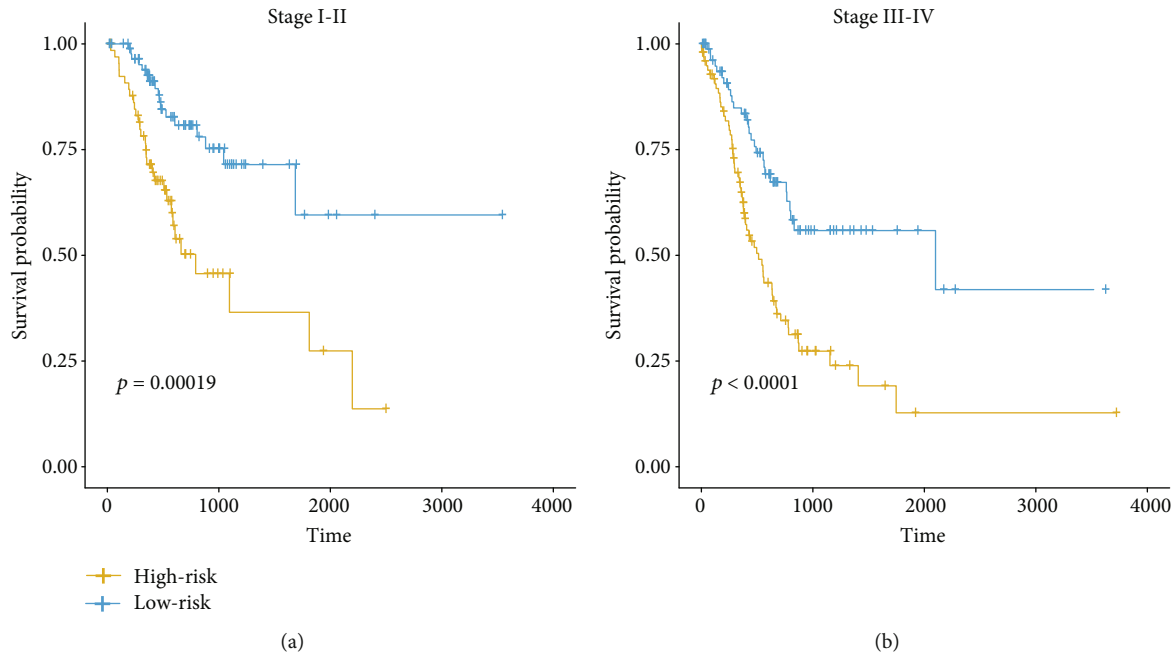


FIGURE 4: The differential prognostic outcomes in the early-stage and advanced GC. The early-stage and advanced GC were defined by those samples with TNM stage I-II, and III-IV, respectively. The KM curves of the early-stage and advanced GC were displayed in (a) and (b). Log-rank test was used to test the difference.

mechanism and key molecules resulting in poor outcome in gastric cancer, we compared the gene expression profiles of the high-risk group with those of the low-risk group. ORA analysis of these upregulated genes in high-risk group revealed that PI3K-Akt signaling pathway and tumor microenvironment-related pathways such as focal adhesion, ECM-receptor interaction, and complement and coagulation cascades might play key roles in the high-risk group of gastric cancer (Figure 5(a)). Notably, two receptors of growth factor in PI3K-Akt signaling, PDGFRA and PDGFRB, were significantly upregulated in the high-risk group of both TCGA and validation cohorts (Figure 5(b)). Moreover, drugs including Nilotinib, Crenolanib, Dasatinib, Benzotatate, Carboplatin, Sunitinib, Regorafenib, Paclitaxel, Ponatinib, Gefitinib, and Imatinib were found to target the two receptors, suggesting that the high-risk group might be treated by these PDGFR inhibitors.

4. Discussion

Gastric cancer (GC) is one of the most common malignant tumors in the digestive system. Here, we designed a seven-gene-signature-based Cox model to stratify the GC samples into high-risk and low-risk groups. The survival analysis revealed that the high-risk and low-risk groups exhibited significantly different prognostic outcomes in both the training and validation datasets, suggesting that the seven-gene-signature-based Cox model was capable of predicting the overall survival of gastric cancer.

Specifically, *CGB5*, *IGFBP1*, *OLFML2B*, *RAI14*, *SERPINE1*, *IQSEC2*, and *MPND* were selected by the multivariable Cox model. *CGB5* is one of the key hCG β encoding

genes, which acts as a proangiogenic factor in some tumors [16, 17], suggesting that *CGB5* might also promote angiogenesis in gastric cancer. *IGFBP1* is involved in the insulin signaling pathway [18], which also participates in the regulation of the PI3K-Akt signaling pathway [19–21]. In accordance with this, the PI3K-Akt signaling pathway was found to be hyperactive in the high-risk group. Notably, the platelet-derived growth factor receptor [22, 23], *PDGFRA* and *PDGFRB*, was significantly upregulated in the high-risk group, further demonstrating that the PDGF/PDGFR and PI3K-Akt signaling pathway were responsible for the worse prognostic outcome and might be the potential therapeutic targets in gastric cancer. Among the drugs inhibiting the activity of PDGFR, Crenolanib [24] and Regorafenib [25] have been found to act as potential targeted therapies in gastric cancer. The remaining prognostic genes such as *OLFML2B*, *RAI14*, *SERPINE1*, and *MPND* were also reported to be dysregulated and associated with poor prognosis in gastric cancer [26–29].

The further evaluation of the risk stratification revealed that it is an independent prognostic factor in gastric cancer. With the TNM stage, age, and gender as cofactors, the risk stratification still maintained statistical significance in the multivariable Cox model, indicating that the risk stratification, combined with TNM stage, age, and gender, had the potential to be applied in OS prediction of gastric cancer.

In summary, we aimed at developing a combination of prognostic gene signatures and building a robust model for GC risk prediction. The predictive model could not only effectively predict the risk of GC but also be beneficial to the development of therapeutic strategies.

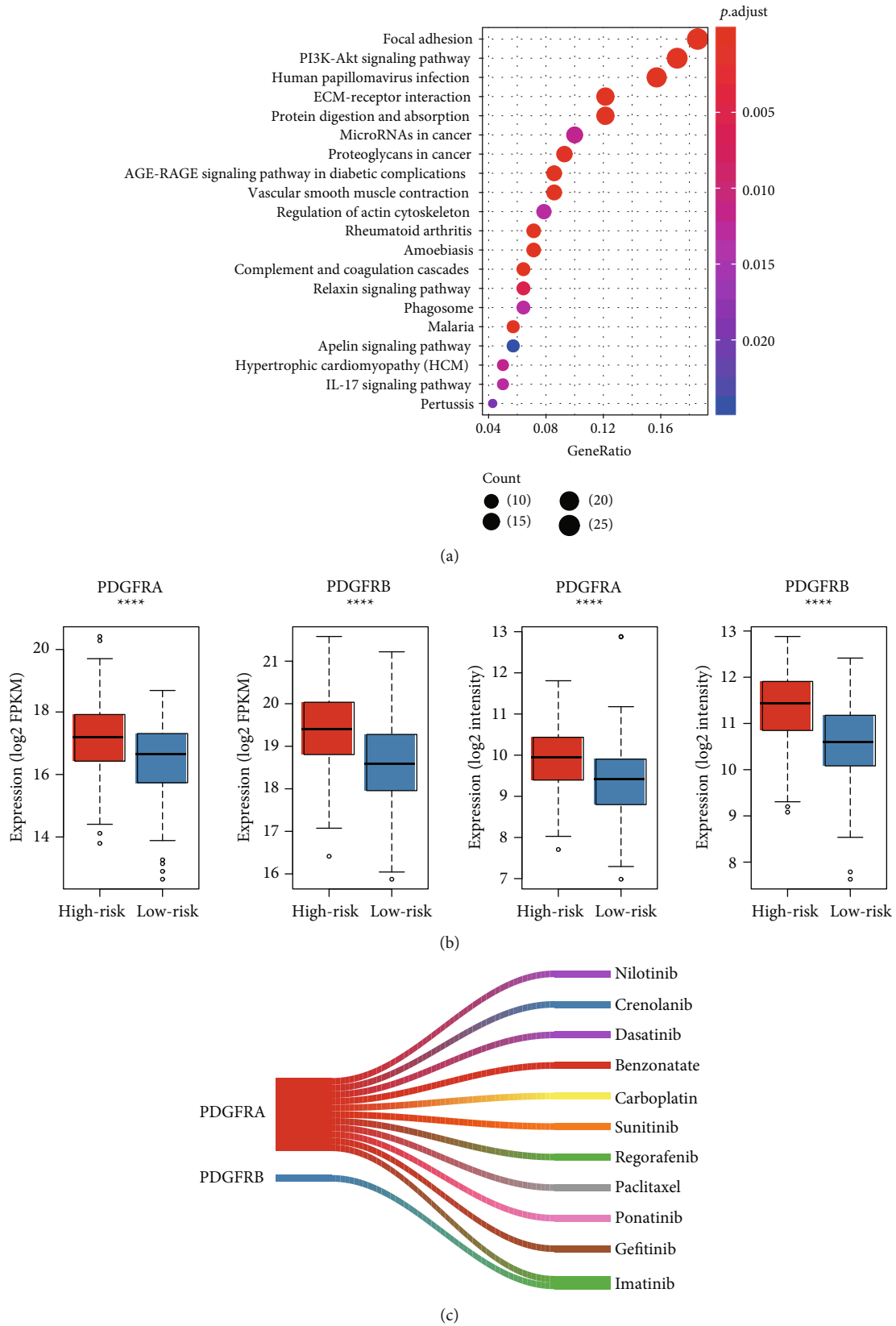


FIGURE 5: The critical biomarkers and pathways in the high-risk group of GC. (a) The pathways enriched by the upregulated genes in the high-risk group of GC. (b) The differential expression levels of *PDGFRA* and *PDGFRB* between the high-risk and low-risk groups. The left two panels represent the data in the TCGA cohort, and the right two represent the GSE84433 cohort. (c) The drugs that potentially inhibit the *PDGFRA* or *PDGFRB*. (* < 0.05, ** < 0.01, *** < 0.001, and **** < 0.0001).

Data Availability

TCGA data were collected from UCSC Xena database and the independent validation dataset was collected from Gene Expression Omnibus with accession GSE84433.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Sichuan Youth Science and Technology Foundation (2017JQ0039), the Scientific and Technological Cooperation Project of Nanchong City (18SXHZ0577, 18SXHZ0575), the Key Scientific Project of Sichuan Health and Health Committee (19ZD005), the Key Scientific Project of The Affiliated Hospital of North Sichuan Medical College (19ZD004), and the Scientific and Technological Cooperation Project of Nanchong City (18SXHZ0548).

Supplementary Materials

Description of the two supplementary tables. Supplementary Table 1: 22 upregulated and 2 downregulated genes in gastric cancer. Supplementary Table 2: other studies involved in our proposed signatures. (*Supplementary Materials*)

References

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] C. M. den Hoed and E. J. Kuipers, "Gastric cancer: how can we reduce the incidence of this disease?," *Current Gastroenterology Reports*, vol. 18, no. 7, 2016.
- [3] L. H. Eusebi, A. Telese, G. Marasco, F. Bazzoli, and R. M. Zagari, "Gastric cancer prevention strategies: a global perspective," *Journal of Gastroenterology and Hepatology*, 2020.
- [4] Z. Song, Y. Wu, J. Yang, D. Yang, and X. Fang, "Progress in the treatment of advanced gastric cancer," *Tumour Biology*, vol. 39, no. 7, 2017.
- [5] G. Corso, J. Carvalho, D. Marrelli et al., "Somatic mutations and deletions of the E-cadherin gene predict poor survival of patients with gastric cancer," *Journal of Clinical Oncology*, vol. 31, no. 7, pp. 868–875, 2013.
- [6] J. Shi, D. Yao, W. Liu et al., "Frequent gene amplification predicts poor prognosis in gastric cancer," *International Journal of Molecular Sciences*, vol. 13, no. 4, pp. 4714–4726, 2012.
- [7] R. Cristescu, J. Lee, M. Nebozhyn et al., "Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes," *Nature Medicine*, vol. 21, no. 5, pp. 449–456, 2015.
- [8] J. Cui, F. Li, G. Wang, X. Fang, J. D. Puett, and Y. Xu, "Gene-expression signatures can distinguish gastric cancer grades and stages," *PLoS One*, vol. 6, no. 3, article e17819, 2011.
- [9] P. Wang, Y. Wang, B. Hang, X. Zou, and J. H. Mao, "A novel gene expression-based prognostic scoring system to predict survival in gastric cancer," *Oncotarget*, vol. 7, no. 34, pp. 55343–55351, 2016.
- [10] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of gastric adenocarcinoma," *Nature*, vol. 513, no. 7517, pp. 202–209, 2014.
- [11] M. Goldman, B. Craft, M. Hastie et al., "The UCSC Xena Platform for cancer genomics data visualization and interpretation," 2018.
- [12] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D991–D995, 2013.
- [13] S. J. Yoon, J. Park, Y. Shin et al., "Deconvolution of diffuse gastric cancer and the suppression of CD34 on the BALB/c nude mice model," *BMC Cancer*, vol. 20, no. 1, 2020.
- [14] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS: A Journal of Integrative Biology*, vol. 16, no. 5, pp. 284–287, 2012.
- [15] A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler, "Maftools: efficient and comprehensive analysis of somatic variants in cancer," *Genome Research*, vol. 28, no. 11, pp. 1747–1756, 2018.
- [16] A. Schanz, M. Lukosz, A. P. Hess, D. M. Baston-Bust, J. S. Krussel, and C. Heiss, "hCG stimulates angiogenic signals in lymphatic endothelial and circulating angiogenic cells," *Journal of Reproductive Immunology*, vol. 110, pp. 102–108, 2015.
- [17] S. Brouillet, P. Hoffmann, S. Chauvet et al., "Revisiting the role of hCG: new regulation of the angiogenic factor EG-VEGF and its receptors," *Cellular and Molecular Life Sciences*, vol. 69, no. 9, pp. 1537–1550, 2012.
- [18] D. van der Kaay, C. Deal, S. de Kort et al., "Insulin-like growth factor-binding protein-1: serum levels, promoter polymorphism, and associations with components of the metabolic syndrome in short subjects born small for gestational age," *The Journal of Clinical Endocrinology and Metabolism*, vol. 94, no. 4, pp. 1386–1392, 2009.
- [19] I. Nepstad, K. J. Hatfield, I. S. Gronningsaeter et al., "Effects of insulin and pathway inhibitors on the PI3K-Akt-mTOR phosphorylation profile in acute myeloid leukemia cells," *Signal Transduction and Targeted Therapy*, vol. 4, no. 1, 2019.
- [20] C. Godoy-Parejo, C. Deng, W. Liu, and G. Chen, "Insulin stimulates PI3K/AKT and cell adhesion to promote the survival of individualized human embryonic stem cells," *Stem Cells*, vol. 37, no. 8, pp. 1030–1041, 2019.
- [21] A. Molinaro, B. Becattini, A. Mazzoli et al., "Insulin-driven PI3K-AKT signaling in the hepatocyte is mediated by redundant PI3K α and PI3K β activities and is promoted by RAS," *Cell Metabolism*, vol. 29, no. 6, pp. 1400–1409.e5, 2019.
- [22] G. Wang, B. Shi, Y. Fu et al., "Hypomethylated gene *NRPI* is co-expressed with *PDGFRB* and associated with poor overall survival in gastric cancer patients," *Biomedicine & Pharmacotherapy*, vol. 111, pp. 1334–1341, 2019.
- [23] F. Huang, M. Wang, T. Yang et al., "Gastric cancer-derived MSC-secreted PDGF-DD promotes gastric cancer progression," *Journal of Cancer Research and Clinical Oncology*, vol. 140, no. 11, pp. 1835–1848, 2014.
- [24] Y. Hayashi, M. R. Bardsley, Y. Toyomasu et al., "Platelet-Derived Growth Factor Receptor- α Regulates Proliferation of Gastrointestinal Stromal Tumor Cells With Mutations in *KIT* by Stabilizing ETV1," *Gastroenterology*, vol. 149, no. 2, pp. 420–432.e16, 2015.

- [25] S. Fukuoka, H. Hara, N. Takahashi et al., “Regorafenib plus nivolumab in patients with advanced gastric or colorectal cancer: an open-label, dose-escalation, and dose-expansion phase Ib trial (REGONIVO, EPOC1603),” *Journal of Clinical Oncology*, vol. 38, no. 18, pp. 2053–2061, 2020.
- [26] B. Xu, Z. Bai, J. Yin, and Z. Zhang, “Global transcriptomic analysis identifies *SERPINE1* as a prognostic biomarker associated with epithelial-to-mesenchymal transition in gastric cancer,” *PeerJ*, vol. 7, article e7091, 2019.
- [27] J. Liu, Z. Liu, X. Zhang, T. Gong, and D. Yao, “Bioinformatic exploration of OLFML2B overexpression in gastric cancer base on multiple analyzing tools,” *BMC Cancer*, vol. 19, no. 1, 2019.
- [28] C. Chen, A. Maimaiti, X. Zhang et al., “Knockdown of RAI14 suppresses the progression of gastric cancer,” *OncoTargets and Therapy*, vol. 11, pp. 6693–6703, 2018.
- [29] J. Zhang, J. Y. Huang, Y. N. Chen et al., “Whole genome and transcriptome sequencing of matched primary and peritoneal metastatic gastric carcinoma,” *Scientific Reports*, vol. 5, no. 1, article 13750, 2015.

Retraction

Retracted: lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p

Computational and Mathematical Methods in Medicine

Received 19 September 2023; Accepted 19 September 2023; Published 20 September 2023

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] B. Zhu, J. Liu, Y. Zhao, and J. Yan, "lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 3128053, 10 pages, 2020.

Research Article

lncRNA-SNHG14 Promotes Atherosclerosis by Regulating ROR α Expression through Sponge miR-19a-3p

Baoliang Zhu,¹ Jing Liu,² Ying Zhao,³ and Jing Yan¹ 

¹Department of Physiology, Jining Medical College, Jining, Shandong, China

²Department of Pharmacy, Jining Medical College, Jining, Shandong, China

³Department of Biochemistry, Jining Medical College, Jining, Shandong, China

Correspondence should be addressed to Jing Yan; yanjing102@mail.jnmc.edu.cn

Received 8 June 2020; Accepted 2 July 2020; Published 25 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Baoliang Zhu 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.

Coronary heart disease (CHD) is the most common cardiovascular disease with high prevalence, disability, and mortality. The balance between proliferation and apoptosis of vascular smooth muscle cells (VSMCs) plays a key role in the initiation of atherosclerosis. In this study, we found a significant decrease in the expression of lncRNA-SNHG14 in atherosclerotic plaque tissues of ApoE^{-/-} mice. Overexpression of lncRNA-SNHG14 can inhibit VSMC proliferation while promoting apoptosis. There is a potential reciprocal regulatory relationship between lncRNASNHG14 and miR-19a-3p, which inhibit each other's expression in vascular smooth muscle cells. In addition, the luciferase reporter gene analysis results showed that there was a direct interaction between miR-19a-3p and the 3' UTR of ROR α . The results of qRT-PCR showed that the level of ROR α mRNA was significantly increased in the aortas treated with miR-19a-3p and SNHG14 compared with that treated with miR-19a-3p alone. In conclusion, we demonstrated that lncRNA-SNHG14 regulates the apoptosis/proliferation balance of VSMCs in atherosclerosis.

1. Introduction

Coronary atherosclerotic heart disease (CHD), referred to as coronary heart disease, is mainly due to the occurrence of atherosclerosis in the coronary artery, which makes the lumen narrow or obstructed [1, 2]. This heart disease is associated with coronary spasm, leading to myocardial ischemia, hypoxia, or necrosis [3]. CHD is the most common cardiovascular disease, which has become the number one killer of human health in the 21st century due to its high morbidity, high disability, as well as high mortality [4]. With the acceleration of the aging social process in China, the incidence of this disease has been increasing year by year. The latest data of WHO in 2017 show that as many as 17.7 million people succumb to cardiovascular disease (CVD) every year, making up about 30% of the total global deaths, among which CHD ranks first [5]. However, the causes and mechanisms of CHD are not completely clear, and there are still many shortcomings in the prevention and treatment of this disease, which cannot fundamentally curb the increasing trend of

the incidence and mortality of CHD. In order to explore its pathogenesis, it is of great significance to carry out experimental studies, especially intervention studies on animal models.

Pathological changes of atherosclerosis (AS) are characterized by lipid deposition in the intima and subintima of large and middle arteries [6]. In addition, smooth muscle cell (SMC) migration to the intima, proliferation and matrix proliferation, and inflammatory cell infiltration are also involved. These processes lead to intimal thickening and the formation of atherosclerotic lesions or fibrolipid plaque lesions [7, 8]. The research and debate on the mechanism of AS has lasted for more than 160 years, forming a variety of theories and factions, such as lipid infiltration theory, injury-response theory, inflammatory response theory, macrophage receptor deletion theory, SMC-causing mutation theory, platelet aggregation, and thrombosis theory [9]. However, neither doctrine alone can comprehensively explain the occurrence nor development of AS. Vascular smooth muscle cells (VSMC) is an important cellular

component of the vascular wall. During development and maturation, VSMC is responsible for vasoconstriction and relaxation and responds to the stimulation of hemodynamic and environmental signals to regulate blood pressure and control vascular homeostasis in the body [10–12]. VSMC has strong plasticity. Normal VSMCs have no significant activity of proliferation, migration, and secretion of extracellular matrix, which is called contractile/differentiated VSMCs. However, VSMCs exhibit significant proliferative and migratory activities when they are immature, when physiological conditions change (such as long-term exercise, pregnancy), or when they are under pathological conditions (such as hypertension), and synthesize a large amount of extracellular matrix, which is called secretory/proliferative VSMC at this time [13–15]. The proliferation and differentiation of VSMC is a key regulatory process that affects the maturation and development of the vascular system. When vascular intima is damaged, VSMC overproliferation, migration, and synthesis of a large number of cellular matrix can induce cardiovascular diseases such as vascular restenosis, hypertension, and atherosclerosis [16, 17].

Long noncoding RNAs (lncRNAs) refer to bioactive RNAs with a length of >200 bases that cannot be translated into proteins and display mRNA-like features such as 5' capping, splicing, and polyadenylation. Several studies have confirmed that lncRNAs become powerful bioregulators by regulating a series of cellular processes in the nucleus or cytoplasm [18, 19]. Recent evidence has emerged that a variety of lncRNAs are involved in the regulation of AS and inflammatory response. For example, Wu's team [20] found that the expression of lincRNA-p21 was downregulated in the ApoE knockout mouse AS model. By interfering with the gene expression *in vitro*, it was confirmed that lincRNA-p21 inhibited the proliferation of VSMC and monocyte macrophages and induced apoptosis. Hu et al. [21] found that RP5-833A20.1 is an lncRNA regulating the NFIA gene, which may reduce the expression of NFIA by inducing the expression of mi-38R2-5p. Overexpression of NFIA increased HDL, decreased LDL and VLDL, increased reverse transport of CHOL, and inhibited AS formation.

Recent studies have shown that lncRNA-SNHG14 is upregulated in gliomas participates in tumor proliferation and migration as an oncogene [22]. The expression of miR-19a-3p was downregulated in gastric cancer, acting as a tumor suppressor by regulating the expression of different genes [23]. Recent studies have shown that lncRNA-SNHG14 promotes microglial activation in cerebral infarction by regulating miR145-5p/PLA2G4a [24]. However, the expression of lncRNA-SNHG14 in atherosclerosis remains unclear, and the relationship between the two and the clinicopathological features of patients have not yet been published. Herein, we aimed at further revealing the role of lncRNA in cardiovascular diseases and its possible molecular mechanism by studying the role of lncRNA-SNHG14 in the pathological process of atherosclerosis.

2. Methods

2.1. Bioinformatics Analysis. The prediction module of the DIANA LncBase2 tool (<https://omictools.com/diana->

[lncbase-tool](https://omictools.com/diana-lncbase-tool/)) was used to predict lncRNA-SNHG14-miR-19a-3p interaction. Target relationships between miR-19a-3p and ROR α were predicted using miRanda and target Scan.

2.2. Mouse Studies. Clean C57BL/6J mice were purchased from Jining Medical College Laboratory Animal Center, and ApoE knockout (ApoE^{-/-}) mice were purchased from Jining Medical College Laboratory Animal Center. The mice were raised in the SPF grade mouse feeding room of Jining Medical College Laboratory Animal Center. Animal husbandry meets relevant management requirements, and all animal operations meet the ethical requirements of laboratory animals. Genotypes were homozygous ApoE knockout mice detected by RT-PCR using the genomic DNA of rat tail tissue as a template. Then, 10 male suckling mice were fed in two cages after the end of lactation and fed with 60% high-fat diet for one month to induce atherosclerosis. This study was approved by the Ethics Committee on Animal Experiments of XX Hospital.

2.3. Cell Culture and Transfection. Human primary aortic smooth muscle cells (HA-VSMC) are adherent cells with large cell morphology, spindle shape, and slow growth. In order to make it grow better, SmGM smooth muscle cell growth medium consisting of smooth muscle cell basal medium (Lonza, USA) was used in this study. TM-2 Bullet Kit™ Cell culture medium and kit, the mixed cell culture medium was prepared according to requirements. After adding various growth factors, fetal bovine serum was added to make the serum concentration reach 5%, and finally, penicillin streptomycin was added to prevent cell contamination. RAW264.7 mouse macrophages were also adherent growth cells with small cell morphology and polygonal or round shape, which were routinely cultured in DMEM medium containing 8% fetal bovine serum. The cells were placed in an incubator with 5% CO₂ at 37°C. After the cells grew to logarithmic phase, experiments were carried out.

To induce overexpression of lncRNASNHG14 in VSMCs, pcDNA3.1-lncRNA-SNHG14 vectors were transfected into the cells. To enhance the miR-19a-3p level, miR-19a-3p mimics were transfected. siRNA of lncRNA-SNHG14 and miR-19a-3p was used to knockdown the expression of lncRNASNHG14 and miR-19a-3p in VSMCs. The pcDNA3.1 empty vectors and scramble control sequences were used as negative transfection controls. One day before transfection, HA-VSMC/RAW264.7 cells were passaged and seeded on cell culture plates at a certain density. Ensure that the cells can grow to 75%-85% fusion within 24 hours, ready for transfection. Add 100 nM siRNA to 100 ml of Opti-MEM and mix gently. Mix the Lipofectamine reagent with 100 μ l serum-free DMEM or Opti-MEM, dilute the 4 μ l Lipofectamine RNAiMAX reagent, mix gently, and stand for 5 minutes at room temperature. The diluted siRNA and reagents were mixed and left for 20 minutes, and then the complex was added to the cell plate for subsequent experiments. If the cell line is sensitive, remove the complex and replace the medium after incubation for 4-6 hours to prevent cell death.

Both small interference sequence of specific targeting SNHG14 (named si-SNHG14) and negative control sequence (named si-NC) were designed and synthesized by Shanghai Gma Biotechnology Co., LTD. The specific sequences were si-SNHG14, forward: 5'-GCUGAUUUUUAGGCACUA TT-3' and reverse 5'-UAGUGCCUUAAAUAUCAGCTT-3'. Si-NC forward: 5'-UUCUCCGAACGUGUCACGUTT-3' and reverse 5'-AACGUGACACGUUCGGAGAATT-3'.

2.4. CCK-8 Assay. CCK-8 solution (CCK-8; Dojindo) was added to each well with 10 microliters. A cell-free pore with the appropriate amount of cell culture medium was set up; drugs and CCK-8 solution were as a blank control. HA-VSMC (1.5 h) and RAW264.7 (1 h) were incubated in the cell incubator and absorbance was measured at 450 nm.

2.5. Flow Cytometry Assay for Apoptosis. The cells were collected and stained with cimin V-FITC and propidium iodide after transfection for 48 h. With Guava_easyCyte flow cytometry instrument testing process cells apoptosis rate, built-in software was used for analysis.

2.6. Real-Time Quantitative PCR (qRT-PCR). Total RNA was extracted from cells by TRIzol™ reagent. In order to detect the mRNA levels of lncRNA-SNHG14 and miR-19a-3p, reverse transcription PCR was carried out with the Prime-Script RT Master Mix kit. Next, SYBR premix EX Taq II was used to amplify and quantify the cDNA according to 2 microliters of 5 * RT Buffer, 0.5 microliters Enzyme mix, 0.5 microliters Primer mix, 2 microliters RNA, 5 uL Rnase free water. Then, set the Takara reverse transcription instrument at 37°C for 5 min, 95°C for 5 min, 4°C for reverse transcription PCR. The qRT-PCR instrument runs the PCR program: preheating at 95°C for 2 min, 8 cycles at 95°C for 30 s, 60°C for 4 s, and 72°C for 30 s, followed by 40 cycles at 95°C for 30 s, 56°C for 45 s, 72°C for 30 s, heating, and unlinking the chain to detect the fluorescence intensity. The sequence of lncRNA-SNHG14 is lncRNA-SNHG14 F: 5'-GGGTGTTTACGTAGACCAGAACC-3', R: 5'-CTTCCA AAAGCCTTCTGCCTTAG-3'. The primer sequences of GAPDH are F: 5'- CCAAAATCAGATGGGGCAATG CTGG-3', R: 5'- TGATGGCATGGACTGTGGTCATTCA-3'; the primer sequences of miR-19a-3p are F: 5'- CGCT GTGCAAATCTATGCAA-3', R: 5'-CGGCCAGTGTTCAGACTAC-3'. The RORα primer sequences are F: 5'-GCTT CGGCAGCACATATACTAAAAT-3', R: 5'-CGCTTCACG AATTTGCGTGCAT-3'.

2.7. Western Blot. The VSMC was cleaved with a RIPA buffer, and the total protein was collected. The protein was isolated by SDS-PAGE and transferred to the NC membrane using a conventional protocol. The membrane was sealed with 5% skim milk at room temperature for 2 hours and incubated with RORα (ab60134, Abcam, 1:1000) and β-actin (ab8226, Abcam, 1:1000) primary antibody at 4°C for 12 hours. Wash the membrane and incubate it with HRP secondary antibody for 2 h at room temperature. The blotting was visualized

using an electrochemical luminescent Western blot kit. The strip strength was quantified using ImageJ software.

2.8. Dual-Luciferase Reporter Assay. Add 70 microliter PLB (Obio Technology) into the hole to be tested, shake, and lyse for 15 min in the dark. Then, add 100 microliter LARI and 20 microliter cell lysate into the white 96-well plate together, and measure the fluorescence value of firefly after mixing. Later, add 100 uL Stop&Glo into the sample immediately, the fluorescence value measured again after mixing was the internal reference kidney fluorescence value. Finally, the intensity of fluorescence was expressed by the ratio of the firefly fluorescence value to the kidney fluorescence value.

2.9. Statistical Analysis. All data were expressed as mean standard deviation (mean SD). Unpaired Student *t*-test was used when comparing two groups of data, and one-way ANOVA was used to compare more than two groups of data. Statistical differences were considered when *P* < 0.05. Data statistics and plotting were performed using SPSS17.0 and GraphPad Prism 5.0.

3. Results

3.1. Expression of lncRNA-SNHG14 in Atherosclerotic Plaques. Atherosclerosis was successfully induced in ApoE^{-/-} mice fed a 60% high-fat diet for one month. Total RNA from atherosclerotic plaque tissues of male ApoE^{-/-} mice and aortic vascular tissues of normal C57BL/6J mice was extracted and quantified by qPCR. We found that compared with wild-type C57BL/6J mice, the expression of lncRNA-SNHG14 in atherosclerotic plaque tissue of ApoE^{-/-} mice was significantly reduced, as shown in Figure 1(a). After transfection of siRNA against murine and human lncRNA-SNHG14 in RAW264.7 cells and HA-VSMC cells, respectively, the expression of lncRNA-SNHG14 was successfully knocked down in both cells by qPCR detection, as shown in Figure 1(b).

3.2. Detection of the Role of lncRNA-SNHG14 in Cell Proliferation and Apoptosis. To further clarify the function of lncRNA-SNHG14, we transfected siRNA against murine and human lncRNA-SNHG14 in RAW264.7 cells and HA-VSMC cells. After cell counting detection, we found that the number of both cells increased significantly 48 hours after transfection compared with the siRNA transfection group with unrelated sequence after lncRNA-SNHG14 silencing (Figure 2(a)). After the transfection of siRNA targeting murine and human lncRNA-SNHG14 in RAW264.7 cells and HA-VSMC cells, CCK-8 cell proliferation assay indicated that the proliferation level of the two cells was significantly increased 48 hours after transfection compared with the siRNA control group after lncRNA-SNHG14 silencing (Figure 2(b)).

In addition, we also transfected siRNA against murine and human lncRNA-SNHG14 in RAW264.7 cells and HA-VSMC cells, respectively, and detected apoptosis by flow cytometry. We found that after lncRNA-SNHG14 silencing, the apoptotic level of the two cells was significantly lower than that of the siRNA control group, indicating that the

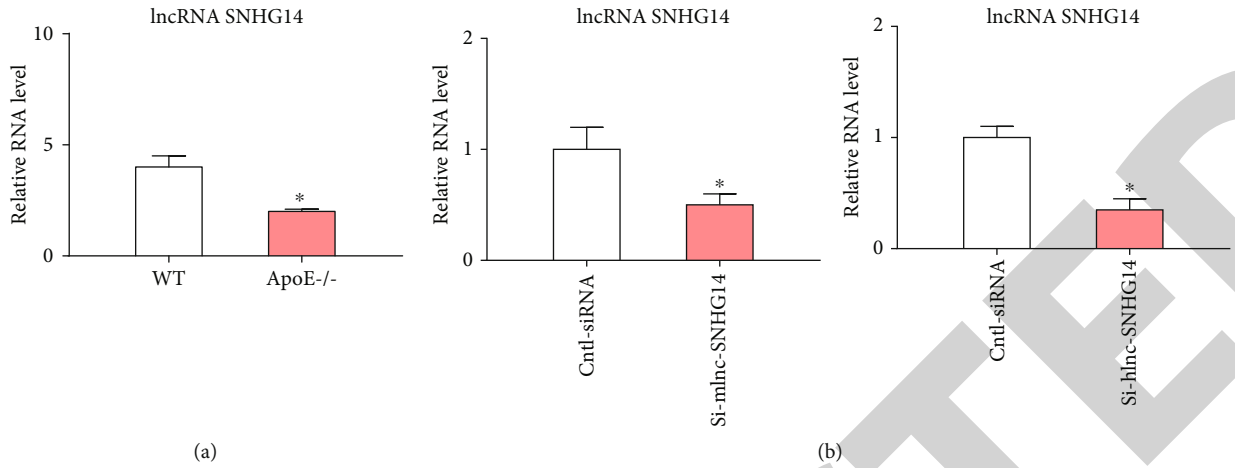


FIGURE 1: Expression of lncRNA-SNHG14 in atherosclerotic plaques. (a) Expression of lncRNA-SNHG14 in mouse atherosclerotic plaques. WT: 5 normal C57BL/6 mice; ApoE-/-: 5 ApoE-/- knockout mice, * $P < 0.05$. (b) Detection of lncRNA-SNHG14 expression in RAW264.7/HA-VSMC cells. Cntl-siRNA was the irrelevant sequence siRNA control transfection group, si-mlncRNA-SNHG14 was the siRNA transfection group for mouse lncRNA-SNHG14, si-hlncRNA-SNHG14 was the siRNA transfection group for human lncRNA-SNHG14, * $P < 0.05$.

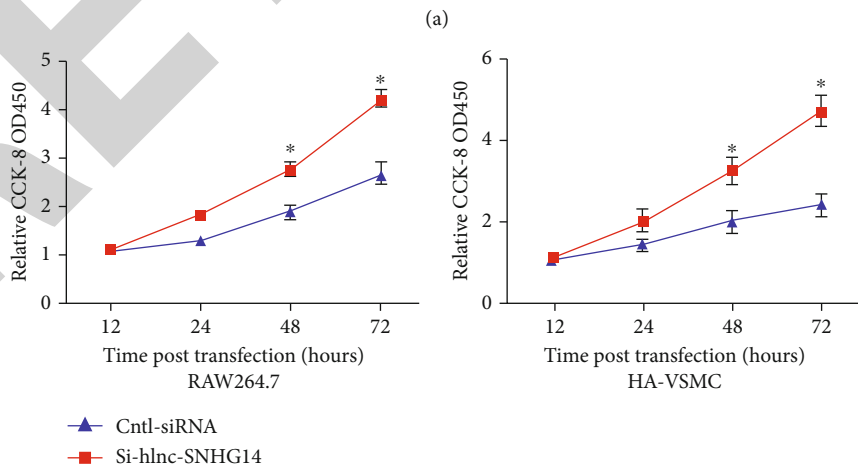
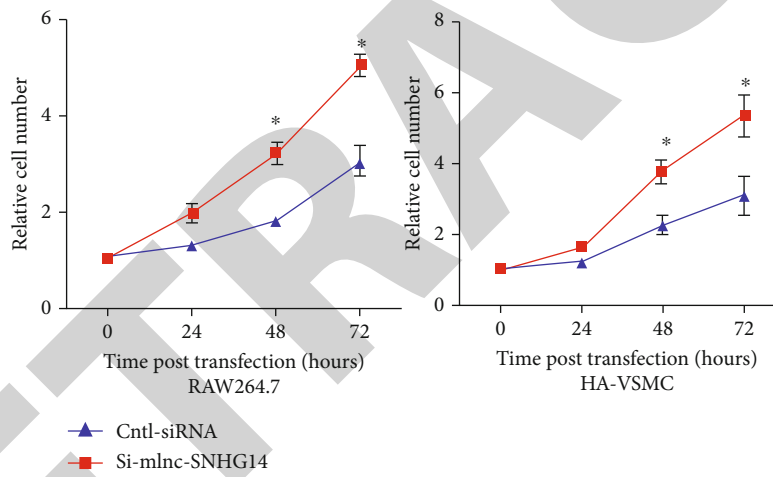


FIGURE 2: Continued.

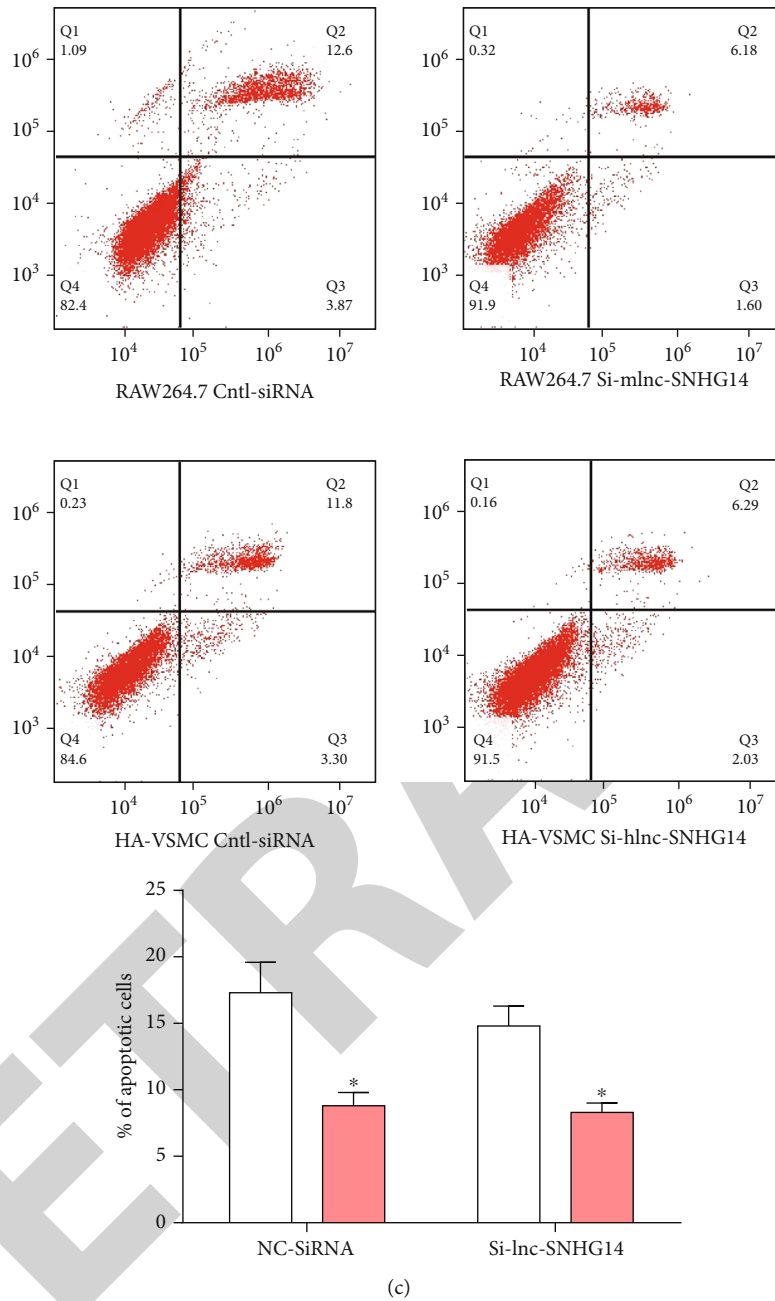


FIGURE 2: Detection of the role of lncRNA-SNHG14 in cell proliferation and apoptosis. (a) Cell counts of RAW264.7/HA-VSMC cells at different time points after lncRNA-SNHG14 silencing. Left: RAW264.7 cells; Right: HA-VSMC cells. The ordinate represents the relative cell number; the harvesting time of 0, 24, 48, 72 cells after transfection, in hours. (Cntl-siRNA was the irrelevant sequence siRNA control transfection group, si-mlncRNA-SNHG14 was the siRNA transfection group for mouse lncRNA-SNHG14, si-hlncRNA-SNHG14 was the siRNA transfection group for human lncRNA-SNHG14, * $P < 0.05$). (b) Cell proliferation detection of RAW264.7/HA-VSMC cells at different time points after lncRNA-SNHG14 silencing. Left: RAW264.7 cells; Right: HA-VSMC cells. The ordinate represents the relative absorbance value; the harvesting time of 0, 24, 48, 72 cells after transfection, in hours. (Ctl-siRNA was the irrelevant sequence siRNA control transfection group, si-mlncRNA-SNHG14 was the siRNA transfection group for mouse lncRNA-SNHG14, si-hlncRNA-SNHG14 was the siRNA transfection group for human lncRNA-SNHG14, * $P < 0.05$). (c) Detection of apoptosis in RAW264.7/HA-VSMC cells after lncRNA-SNHG14 silencing. Flow cytometry two-dimensional dot plot and Flow cytometry two-dimensional dot plot data statistical graph. The ordinate of the two-dimensional dot plot of flow cytometry data is the percentage of apoptotic cells. Cntl-siRNA was an irrelevant sequence siRNA control transfection group, and si-lncRNA-SNHG14 was a siRNA transfection group for mouse or human lncRNA-SNHG14, * $P < 0.05$.

silencing of lncRNA-SNHG14 inhibited the apoptosis of the two cells (Figure 2(c)).

3.3. Interaction between lncRNA-SNHG14 and miR-19a-3p.

To verify whether there is a regulatory relationship between lncRNA-SNHG14 and miR-19a-3p, we performed targeted binding prediction between lncRNA SNHG14 and miR-19a-3p (Figure 3(a)). In addition, we transfected lncRNA-SNHG14 overexpression plasmid or infected with miR-19a-3p lentivirus on vascular smooth muscle cells. The results showed that the upregulation of lncRNA-SNHG14 could significantly inhibit the expression of miR-19a-3p. When the expression of miR-19a-3p was upregulated, the expression of lncRNA-SNHG14 was also inhibited. This suggests that lncRNA-SNHG14 and miR-19a-3p inhibit each other's expression in vascular smooth muscle cells, and there is a potential mutual regulatory relationship between the two (Figure 3(b)).

We constructed dual-luciferase reporter plasmids containing the sequences of wild-type and mutant (in which all three potential binding targets were mutated) of the potential targeting binding sites of miR-19a-3p and lncRNA-SNHG14 to verify their binding *in vivo*. Luciferase assay found that the relative fluorescence activity of wild-type luciferase plasmid decreased significantly compared to the mutant group (Figure 3(c)). Subsequently, we validated the binding *in vitro* of lncRNA-SNHG14 to miR-19a-3p by RNA pull-down assay. It was found that the biotin-labeled lncRNA-SNHG14 sense strand could pull out more miR-19a-3p than lncRNA-SNHG14 antisense strand. The binding and adsorption of lncRNA-SNHG14 on miR-19a-3p were further illustrated (Figure 3(d)).

3.4. SNHG14/miR-19a-3p Promotes VSMC Proliferation and Inhibits Its Apoptosis by Targeting ROR α .

Through bioinformatics prediction analysis, there is a binding site of miR-19a-3p in the 3'UTR of ROR α , and it is speculated that ROR α may be a downstream target of SNHG14 to play a regulatory role through miR-19a-3p (Figure 4(a)). The results of luciferase reporter gene analysis showed that there was a direct interaction between miR-19a-3p and the 3'UTR of ROR α (Figure 4(b)). Furthermore, Western Blot assay showed that after overexpression of miR-19a-3p, the expression of ROR α decreased compared with the control group ($P < 0.05$), confirming that ROR α can be used as a regulatory target of miR-19a-3p (Figure 4(c)).

We injected the SNHG14 overexpression plasmid with adenovirus pMIR-19a-3p and the corresponding control tail vein into ApoE $^{-/-}$ mice and aortic tissues, respectively. The results of qRT-PCR showed that the level of ROR α mRNA was significantly increased in the aortas treated with miR-19a-3p and SNHG14 compared with that treated with miR-19a-3p alone (Figure 5(a)). Western blot also showed that the expression of ROR α protein in the aorta treated with SNHG14 and miR-19a-3p was higher than that in the tissues treated with only miR-19a-3p, but lower than that in the tissues treated with only SNHG14. The result suggests that SNHG14 can reverse the inhibitory effect of miR-19a-3p on

the expression of ROR α in the liver and aorta of ApoE $^{-/-}$ mice (Figure 5(b)).

4. Discussion

lncRNA-SNHG14, also known as UBE3A-ATS, is located on human chromosome 15q11.2. Knockout of the SNHG14 gene significantly inhibits the survival, migration, invasion, and promotes apoptosis of gastric cancer SGC-7901 cells [25]. Studies in renal cancer have found that lncRNA-SNHG14 is upregulated and can be used as a ceRNA to promote the migration and invasion of clear cell renal cancer [26]. Qi and other studies have shown that lncRNA-SNHG14 promotes microglial activation by regulating miR145-5P/PLA2G4a and participates in the occurrence of cerebral infarction [24]. The study confirmed that the expression level of lncRNA-SNHG14 in the serum of acute cerebral infarction was upregulated, and with the aggravation of acute cerebral infarction, the expression level of lncRNA-SNHG14 gradually increased, suggesting that lncRNA-SNHG14 participates in the occurrence and progression of acute cerebral infarction. However, to our knowledge, there is no published evidence that lncRNA-SNHG14 is associated with dysfunction of VSMC in atherosclerosis. We examined the expression of lncRNA-SNHG14 in patients and analyzed its function *in vitro*. As expected, downregulated lncRNA-SNHG14 was observed in the AS mouse model, and lncRNA-SNHG14 could inhibit VSMC proliferation but induce apoptosis. These results suggest that lncRNA-SNHG14 plays a role in atherosclerosis and may be used as a potential target for therapy. The mechanism by which lncRNA-SNHG14 regulates proliferation/apoptosis deserves further study. However, the molecular functions performed by lncRNAs and the corresponding mechanisms are essentially complex. To simplify the problem, we only tried a popular theory called ceRNA theory to partially explain how lncRNA-SNHG14 performs its function. In brief, we established a model of lncRNA-microRNA-mRNA regulation based on the concept of ceRNA, and designed experiments to confirm whether certain mechanisms are suitable for this model. In addition, we performed a bioinformatics analysis to understand the potential interaction between lncRNA-SNHG14 and microRNA. Using the DIANA LncBase2 tool, we speculated that >200 microRNAs might interact with lncRNA-SNHG14. Among these candidate microRNAs, we chose miR-19a because it is fully studied in the regulation of VSMC function.

miRNAs are a class of endogenous noncoding small RNAs with a length of 21-25 nucleotides, which can cleave or repress target gene mRNAs by binding to the 3'-non-coding region of target mRNAs. miR-19a is located in chromosome region 13q31.3 and has been confirmed to be associated with malignant transformation of metastatic breast and colon cancer [27, 28]. Studies have shown that in hyperlipidemia, mildly oxidized LDL can stimulate HIF-1 α expression in vascular endothelial cells, and endothelial HIF-1 α can trigger miR-19a-mediated CXCL1 expression and monocyte adhesion to promote atherosclerosis progression [29]. miR-19a is an important member of the

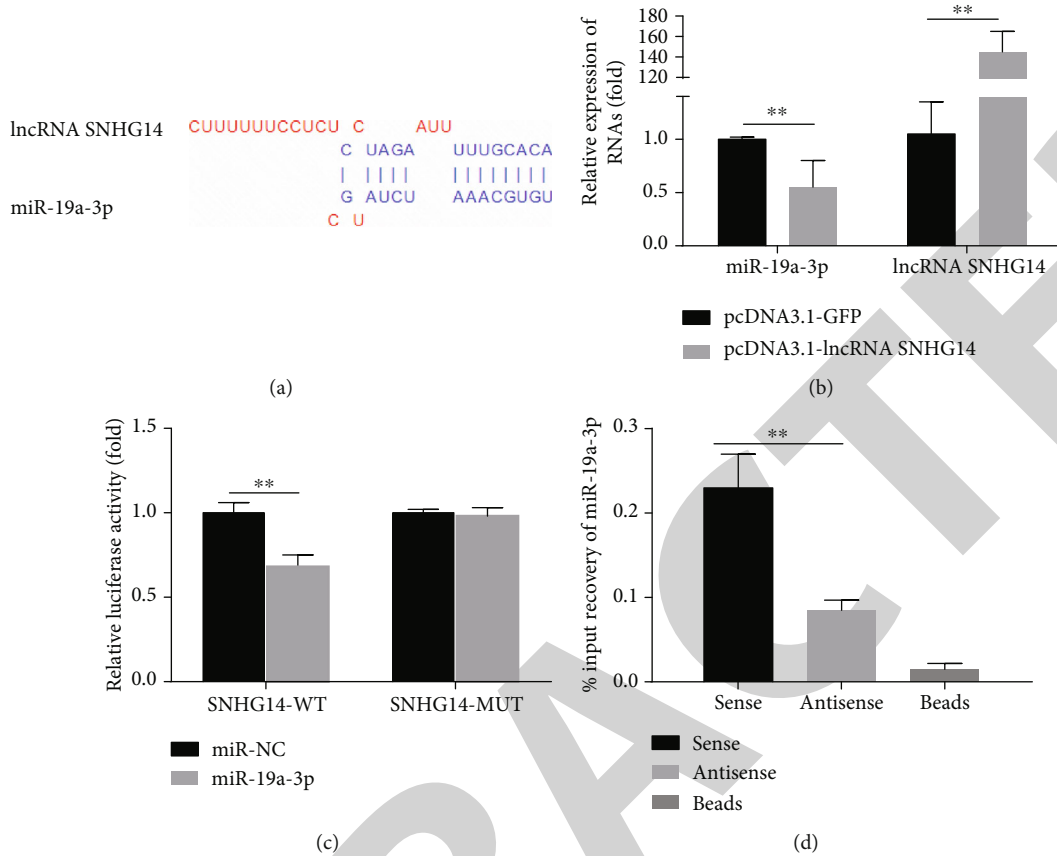


FIGURE 3: Interaction between lncRNA-SNHG14 and miR-19a-3p. (a) Targeted binding between lncRNA SNHG14 and miR-19a-3p. (b) Left: the expression result of miR-19a-3p when overexpressing lncRNA SNHG14; Right: the expression result of lncRNA SNHG14 when overexpressing miR-19a-3p; * $P < 0.05$, ** $P < 0.01$. (c) Analysis results of fluorescence data after cotransfection of mimic with dual luciferase plasmid; (d) RNA pull-down validates the binding of miR-19a-3p to lncRNA SNHG14; * $P < 0.05$, ** $P < 0.01$.

polycistronic gene cluster of miR-17-92. Current studies have found that the gene cluster in which miR-19a resides can be activated during atherosclerosis, promoting vascular inflammation and foam cell formation [30]. miR-19a was also found to be upregulated in endothelial vascular cells under hypoxia-inducible factors and shear stress, increasing the proliferation and antiapoptotic ability of endothelial vascular cells [29]. Animal models of atherosclerosis provide more evidence for the role of miR-19a in vivo. ApoE^{-/-} mice fed a high-fat diet were treated with an antagonist of miR-19a, which is consistent with our findings. Histological analysis of thoracic and abdominal aorta specimens revealed that atherosclerotic plaques and lipid content were significantly reduced in mice [31]. The above evidence suggests that inhibition of miR-19a can alleviate the inflammatory response and slow down the development of atherosclerosis. At present, it has been reported that miR-19a is elevated in serum and atherosclerotic lesions of patients with coronary artery disease, which may be one of the initiating factors of atherosclerosis [32].

Nuclear receptors are a class of ligand-dependent transcription factor superfamily. Retinoic acid-related orphan receptors (RORs) are named for their similarity in gene sequence to retinoic acid receptors (RARs) and retinoid X receptors (RXRs). RORs include three subfamilies:

ROR α , ROR β , and ROR γ . Because the endogenous ligands of RORs receptors are unknown, they are called orphan nuclear receptors. Through the study of ROR α molecular structure, action characteristics, and ROR α gene mutation deficient (ROR α sg/sg) mice, we found that ROR α plays an important role in the regulation of lipid metabolism. Compared with wild-type mice, the serum triglyceride (TG), total cholesterol (TC), and high-density lipoprotein (HDL) of ROR α sg/sg mice were lower under normal dietary conditions [33, 34], which indicated that ROR α was closely related to lipid metabolism. Abnormal expression of ROR α can cause metabolic disorders, leading to an increased prevalence of various metabolic-related diseases, including obesity, type 2 diabetes, atherosclerosis, and so on. When atherosclerosis occurs, ROR α expression is significantly reduced in smooth muscle cells and endothelial cells. Vascular smooth muscle cell apoptosis and extracellular matrix homeostasis are key links in regulating atherosclerotic plaque stability. When unstable plaques are formed, the phagocytic clearance of apoptotic cells by macrophages is defective or inadequate, leading to decreased clearance of apoptotic cells. The increase of apoptotic cells and the decrease of macrophage's funeral effect on apoptotic cells make a large accumulation of apoptotic cells in plaque, which will promote the further release of

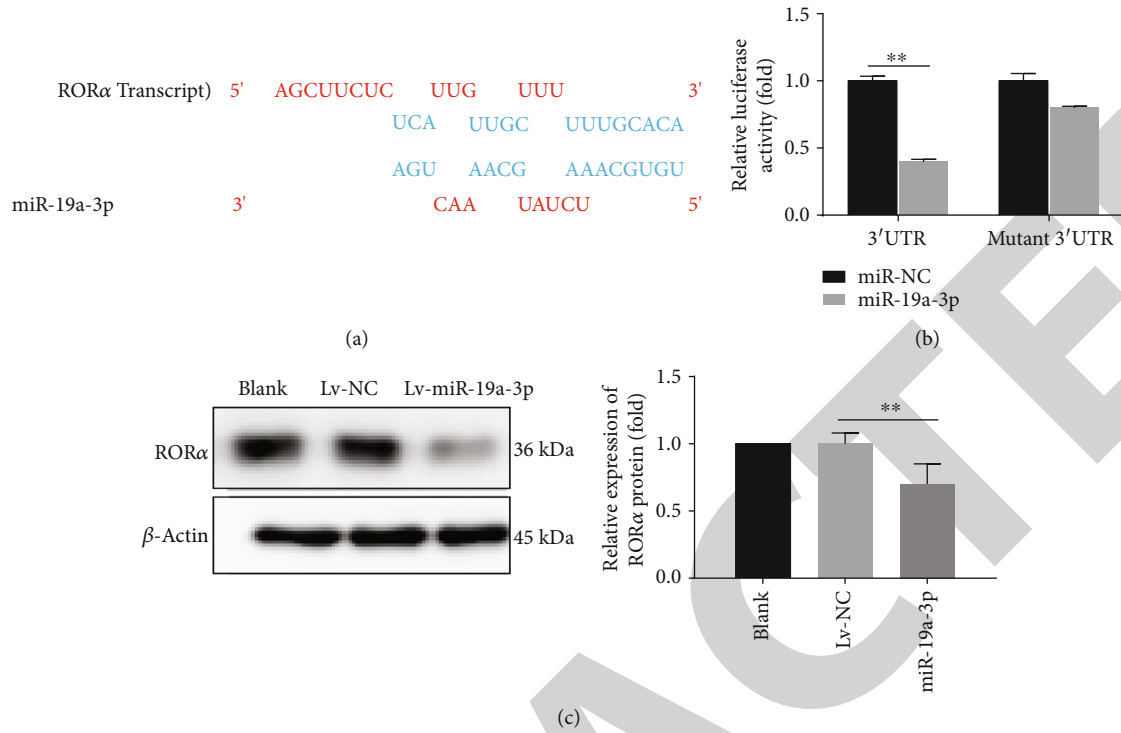


FIGURE 4: SNHG14/miR-19a-3p promotes VSMC proliferation and inhibits its apoptosis by targeting RORα. (a) Patterns of construction of wild-type and mutant dual-luciferase plasmids; the red part is the mutated base site; (b) Analysis results of fluorescence value data after cotransfection of mimic and dual-luciferase plasmids; ** $P < 0.1$. (c) Relative expression of RORalpha protein in cells; * $P < 0.05$, ** $P < 0.01$.

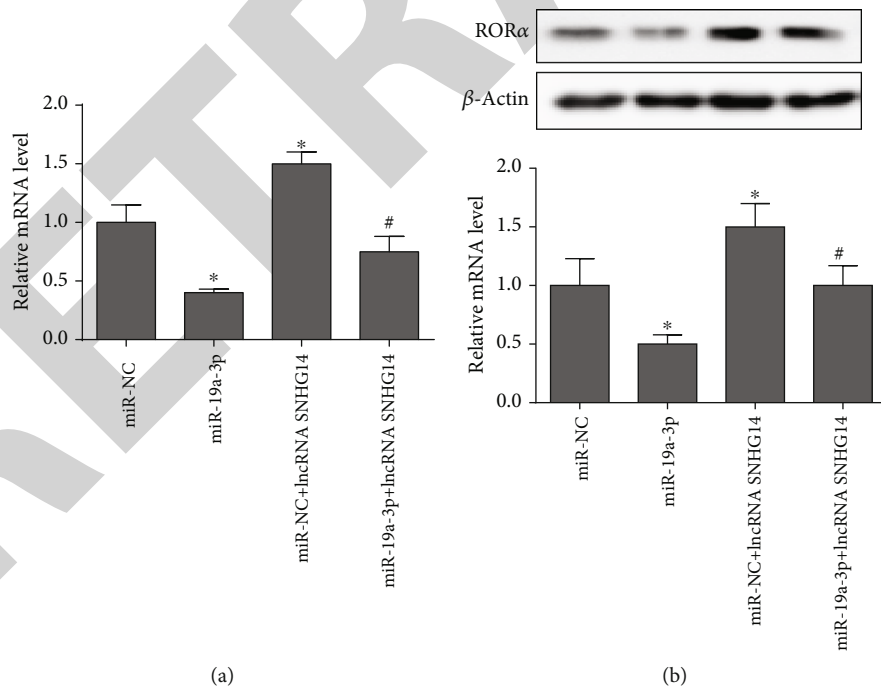


FIGURE 5: Overexpression of SNHG14 reverses the inhibition of RORalpha mRNA (a) and protein (b) levels in ApoE^{-/-} mouse aorta by miR-19a-3p. Note: * $P < 0.05$ vs. miR-NC group; # $P < 0.05$ vs. miR-19a-3p group.

inflammatory factors and matrix-degrading enzymes, trigger the expansion of necrotic core, further thinning and rupture of plaque fibrous cap, and promote the occurrence of acute coronary events.

In this study, we first used ApoE knockout mice to detect the expression of SNHG14 in atherosclerotic plaques by qRT-PCR after atherosclerosis was induced by high-fat diet feeding. We found that the expression of SNHG14 in

atherosclerotic plaque tissues of ApoE^{-/-} mice was significantly decreased compared with wild-type C57BL/6J mice. Thus, the possible correlation between SNHG14 and AS was preliminarily verified. Next, we used mouse macrophage cell line RAW264.7 and human primary cultured aortic vascular smooth muscle cell HA-VSMC. The cell model of SNHG14 gene silencing was prepared by siRNA transfection, and then the levels of cell proliferation and apoptosis were detected in vitro. Results after silencing SNHG14 in RAW264.7 cells and HA-VSMC cells, the proliferation level of the above two cells increased while the apoptotic level was inhibited. This finding indicates that SNHG14 can affect the proliferation and apoptosis of the above two AS-related cells at the in vitro level. In addition, the upregulation of lncRNA-SNHG14 can cause changes in the microRNA 19a-3p/ROR α axis. Combined with the results of bioinformatics analysis and dual-luciferase reporter gene experiment, we believe that lncRNA-SNHG14 has the potential of sponge adsorption of microRNA-19a-3p.

It has been reported that both microRNA-19a-3p and ROR α are involved in regulating the function of VSMCs in atherosclerosis, thus indirectly confirming the possible target relationship between microRNA-19a-3p and ROR α . These results suggest that there may be a regulatory mode of lncRNA-SNHG14/microRNA-19a-3p/ROR α in the dysfunction of atherosclerotic VSMCs.

5. Conclusion

In conclusion, we believe that lncRNA-SNHG14 is an important regulator of vascular smooth muscle cell proliferation and apoptosis in the process of atherosclerosis. lncRNA-SNHG14 can exert this function by regulating the microRNA-19a-3p/ROR α axis as ceRNA. Restoring the expression of lncRNA-SNHG14 in vascular smooth muscle cells may be a potential target for atherosclerotic treatment.

Data Availability

The related data can be provided if any researchers required.

Conflicts of Interest

The authors declare no financial conflicts of interest.

Authors' Contributions

JY designed the project. BLZ and JL carried out all experiments. JL and YZ performed the statistical analysis. JBZ prepared the manuscript. JY contributed to revising the manuscript. All authors have seen and approved the final manuscript.

References

- [1] C. Collet, D. Capodanno, Y. Onuma et al., "Left main coronary artery disease: pathophysiology, diagnosis, and treatment," *Nature Reviews Cardiology*, vol. 15, no. 6, pp. 321–331, 2018.
- [2] J. M. Lee, K. H. Choi, B. K. Koo et al., "Prognostic implications of plaque characteristics and stenosis severity in patients with coronary artery disease," *Journal of the American College of Cardiology*, vol. 73, no. 19, pp. 2413–2424, 2019.
- [3] M. Ishii, K. Kaikita, K. Sato et al., "Acetylcholine-provoked coronary spasm at site of significant organic stenosis predicts poor prognosis in patients with coronary vasospastic angina," *Journal of the American College of Cardiology*, vol. 66, no. 10, pp. 1105–1115, 2015.
- [4] G. A. Karpouzas, J. Malpeso, T. Y. Choi, D. Li, S. Munoz, and M. J. Budoff, "Prevalence, extent and composition of coronary plaque in patients with rheumatoid arthritis without symptoms or prior diagnosis of coronary artery disease," *Annals of the Rheumatic Diseases*, vol. 73, no. 10, pp. 1797–1804, 2014.
- [5] WHO, *World Heart Day 2017 - Scale up prevention of heart attack and stroke*. [DB/OL]. (2017-09-24)[2018-02-01]http://www.who.int/cardiovascular_diseases/world-heart-day-2017/en/.
- [6] P. Raggi, J. Genest, J. T. Giles et al., "Role of inflammation in the pathogenesis of atherosclerosis and therapeutic interventions," *Atherosclerosis*, vol. 276, pp. 98–108, 2018.
- [7] M. Cattaneo, R. Wyttenbach, R. Corti, D. Staub, and A. Gallino, "The growing field of imaging of atherosclerosis in peripheral arteries," *Angiology*, vol. 70, no. 1, pp. 20–34, 2018.
- [8] A. N. Orekhov, E. Andreeva, I. A. Mikhailova, and D. Gordon, "Cell proliferation in normal and atherosclerotic human aorta: proliferative splash in lipid-rich lesions," *Atherosclerosis*, vol. 139, no. 1, article S0021915098000446, pp. 41–48, 1998.
- [9] X. Hou and H. Chen, "Proposed antithrombotic strategy for acute ischemic stroke with large-artery atherosclerosis: focus on patients with high-risk transient ischemic attack and mild-to-moderate stroke," *Annals of Translational Medicine*, vol. 8, no. 1, p. 16, 2020.
- [10] S. Wang, Z. Cheng, and X. Chen, "Promotion of PTEN on apoptosis through PI3K/Akt signal in vascular smooth muscle cells of mice model of coronary heart disease," *Journal of Cellular Biochemistry*, vol. 120, no. 9, pp. 14636–14644, 2019.
- [11] P. Keul, A. Polzin, K. Kaiser et al., "Potent anti-inflammatory properties of HDL in vascular smooth muscle cells mediated by HDL-S1P and their impairment in coronary artery disease due to lower HDL-S1P: a new aspect of HDL dysfunction and its therapy," *FASEB Journal*, vol. 33, no. 1, pp. 1482–1495, 2018.
- [12] H.-B. Wu, Z.-W. Wang, F. Shi et al., "Av β 3 single-stranded DNA aptamer attenuates vascular smooth muscle cell proliferation and migration via Ras-PI3K/MAPK pathway," *Cardiovascular Therapeutics*, vol. 2020, Article ID 6869856, 12 pages, 2020.
- [13] U. Pohl, "Connexins: key players in the control of vascular plasticity and function," *Physiological Reviews*, vol. 100, no. 2, pp. 525–572, 2020.
- [14] A. F. Kolb, L. Petrie, C. D. Mayer, L. Pirie, and S. J. Duthie, "Folate deficiency promotes differentiation of vascular smooth muscle cells without affecting the methylation status of regulated genes," *The Biochemical Journal*, vol. 476, no. 19, pp. 2769–2795, 2019.
- [15] M. Liu and D. Gomez, "Smooth muscle cell phenotypic diversity," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 39, no. 9, pp. 1715–1723, 2019.
- [16] N. Li, A. P. Rickel, H. J. Sanyour, and Z. Hong, "Vessel graft fabricated by the on-site differentiation of human mesenchymal stem cells towards vascular cells on vascular extracellular

Research Article

Identification of Key mRNAs and lncRNAs in Neonatal Sepsis by Gene Expression Profiling

Lin Bu,^{1,2} Zi-wen Wang,² Shu-qun Hu ,² Wen-jing Zhao,² Xiao-juan Geng,² Ting Zhou,¹ Luo Zhuo ,¹ Xiao-bing Chen,¹ Yan Sun,¹ Yan-li Wang,¹ and Xiao-min Li ¹

¹Department of Emergency Medicine, First People's Hospital of Lianyungang, Hospital of the Clinical Medical School of Nanjing Medical University, Lianyungang 222002, China

²Department of Intensive Care Unit, Affiliated Hospital of Xuzhou Medical University, Xuzhou 221000, China

Correspondence should be addressed to Xiao-min Li; dechun663930@163.com

Received 2 February 2020; Revised 2 April 2020; Accepted 8 April 2020; Published 25 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Lin Bu 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.

Neonatal sepsis is one of the most prevalent causes of death of the neonates. However, the mechanisms underlying neonatal sepsis remained unclear. The present study identified a total of 1128 upregulated mRNAs and 1008 downregulated mRNAs, 28 upregulated lncRNAs, and 61 downregulated lncRNAs in neonatal sepsis. Then, we constructed PPI networks to identify key regulators in neonatal sepsis, including ITGAM, ITGAX, TLR4, ITGB2, SRC, ELANE, RPLP0, RPS28, RPL26, and RPL27. lncRNA coexpression analysis showed HS.294603, LOC391811, C12ORF47, LOC729021, HS.546375, HNRPA1L-2, LOC158345, and HS.495041 played important roles in the progression of neonatal sepsis. Bioinformatics analysis showed DEGs were involved in the regulation cellular extravasation, acute inflammatory response, macrophage activation of NF-kappa B signaling pathway, TNF signaling pathway, HIF-1 signaling pathway, Toll-like receptor signaling pathway, and ribosome, RNA transport, and spliceosome. lncRNAs were involved in regulating ribosome, T cell receptor signaling pathway, RNA degradation, insulin resistance, ribosome biogenesis in eukaryotes, and hematopoietic cell lineage. We thought this study provided useful information for identifying novel therapeutic markers for neonatal sepsis.

1. Introduction

Neonatal sepsis was a severe systematic infectious disease in neonates induced by bacteria, fungi, and viruses [1]. Neonatal sepsis is one of the most prevalent causes of death of neonates [2]. Adult sepsis has been studied in depth, but many abundant studies stated that the neonatal immune response to sepsis is different from adults; comparable research on neonatal vascular endothelium is not enough. Neonatal endothelial cells expressing lower amounts of adhesion molecules show a reduced capacity to reactive oxygen species [3]. In the past decades, emerging studies showed activation of lymphocytes, neutrophils, and mononuclear macrophages played crucial roles in the progression of neonatal sepsis. A few genes were identified to be associated with neonatal sepsis. For example, TLR2 and TLR4 were associ-

ated with the recognition of the bacteria in neonates [4]. PIK3CA, TGFBR2, CDKN1B, KRAS, E2F3, TRAF6, and CHUK were reported to be key regulators in neonatal sepsis [5]. However, the detailed mechanisms underlying these processes remained elusive.

Long noncoding RNAs (lncRNAs) were a class of ncRNAs longer than 200 bps. Emerging studies showed lncRNAs were important regulators in multiple human diseases such as diabetes, cancers, and neonatal sepsis. lncRNAs regulate target expression in different levels, including transcriptional and posttranscriptional levels. lncRNAs could bind to RNA, protein, and DNA molecules in cells. Very few reports are aimed at elucidating the functions and roles of lncRNAs in neonatal sepsis. Until now, only one report showed lncRNA SNHG16 reverses the effects of miR-15a/16 on the LPS-induced inflammatory pathway in neonatal sepsis [6]. Exploring the

roles of lncRNAs in neonatal sepsis could provide novel clues for us to understand the mechanisms underlying this disease progression.

The previous study is aimed at identifying differently expressed mRNAs and lncRNAs in neonatal sepsis by analyzing GSE25504 [7]. Protein-protein interaction network and coexpression network analysis were used to identify key mRNAs and lncRNAs. Bioinformatics analysis was also conducted to predict the potential roles of these genes in neonatal sepsis. This study could provide novel clues to understand the mechanisms of underlying neonatal sepsis progression.

2. Materials and Methods

2.1. Microarray Data. Three gene expression profile GSE25504 [8] was downloaded from the GEO database. GSE25504, which was based on the GPL6947 platform, was submitted by Dickinson et al. The GSE25504 dataset contained 38 negative blood culture result samples and 25 positive blood culture result samples. The analysis for differential gene expression between tumor and normal tissue was performed using GeneSpring software version 11.5 (Agilent Technologies, Inc., Santa Clara, CA, USA). Student's *t*-test was used to identify DEGs with an alteration of ≥ 2 -fold. $p < 0.05$ was considered to indicate a statistically significant difference. We applied Limma package to identify DEGs with R software [9].

2.2. Coexpression Network Construction and Analysis. In this study, Pearson's correlation coefficient of differently expressed gene- (DEG-) lncRNA pairs was calculated according to the expression value of them. The coexpressed DEG-lncRNA pairs with the absolute value of Pearson's correlation coefficient ≥ 0.8 were selected, and the coexpression network was established by using Cytoscape software.

2.3. Pathway Enrichment Analysis. Pathway analysis was used to find the significant pathways according to Kyoto Encyclopedia of Genes and Genomes (KEGG). Fisher's exact test was adopted to select the significant pathways, and the threshold of significance was defined by FDR and *p* value. Significant pathways were extracted according to the thresholds of $p < 0.05$ and intersection gene count > 1 .

2.4. Integration of the Protein-Protein Interaction (PPI) Network. The Search Tool for the Retrieval of Interacting Genes version 10.0 (STRING: <http://string-db.org>) [10] was used for the exploration of potential DEG interactions at the protein level. The PPI networks of DEGs by STRING were derived from validated experiments. A PPI score of > 0.4 was considered significant. The PPI networks were visualized using Cytoscape software [11] (<http://www.cytoscape.org>). $p < 0.05$ was considered to indicate a statistically significant difference.

3. Results

3.1. Identification of Differently Expressed mRNAs and lncRNAs in Neonatal Sepsis by Analyzing Whole Blood Expression Profiling. The present study is aimed at identify-

ing differently expressed mRNAs and lncRNAs in neonatal sepsis by analyzing whole blood mRNA expression profiling, GSE25504, from the NCBI GEO dataset (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25504>). A total of 38 negative blood culture result samples and 25 positive blood culture result samples were included in this dataset. As shown in Figures 1(a) and 1(b), 1128 upregulated mRNAs and 1008 downregulated mRNAs with \log_2 fold change (FC) ≥ 1.0 and false discovery rate (FDR) ≤ 0.01 were identified as differently expressed genes (DEGs). Meanwhile, this study identified 28 upregulated lncRNAs and 61 downregulated lncRNAs in positive samples compared to negative samples as differently expressed lncRNAs (DELncs).

3.2. PPI Network Analysis of DEGs in Neonatal Sepsis. The above analysis revealed multiple differently expressed genes in neonatal sepsis. However, the interactions of these DEGs remained largely unclear. To obtain the interactions among the 22 upregulated mRNAs and 863 downregulated mRNAs in the neonatal sepsis, the present study constructed and presented PPI networks using the STRING database and Cytoscape software. The combined score > 0.4 was used as the cut-off criterion. Following the construction of PPI network, a MCODE plug-in analysis was performed to identify hub networks (degree cut-off ≥ 2 and the nodes with edges ≥ 2 -core) in the PPI network using Cytoscape software (Figure 2). As shown in Figure 2, upregulated hub network 1 included 71 nodes and 1187 edges, upregulated hub network 2 included 66 nodes and 611 edges, and upregulated hub network 3 included 62 nodes and 529 edges. As shown in Figure 3, downregulated hub network 1 included 94 nodes and 4048 edges, downregulated hub network 2 included 30 nodes and 247 edges, and downregulated hub network 3 included 26 nodes and 199 edges. Blue nodes indicate upregulated genes, and pink nodes indicate downregulated genes in the neonatal sepsis.

Also, we identified several key regulators in these PPI networks. The key regulators in upregulated PPI networks included ITGAM (degree = 131), ITGAX (degree = 101), TLR4 (degree = 100), ITGB2 (degree = 92), SRC (degree = 87), and ELANE (degree = 81). The key regulators in downregulated PPI networks included RPLP0 (degree = 128), RPS28 (degree = 128), RPL26 (degree = 124), RPL27 (degree = 123), NSA2 (degree = 122), RPS15 (degree = 120), RPS10 (degree = 117), RPS13 (degree = 117), RPS20 (degree = 117), RPL36 (degree = 110), FAU (degree = 108), NHP2L1 (degree = 106), RPL23 (degree = 106), RPS25 (degree = 105), RPL9 (degree = 101), RPL30 (degree = 100), and RPL35A (degree = 100).

3.3. Bioinformatics Analysis of DEGs in Neonatal Sepsis. Furthermore, we explored the potential functions of DEGs in neonatal sepsis. We next performed bioinformatics analysis of upregulated and downregulated hub PPI networks in thyroid cancer using Cytoscape's ClueGo plug-in. Only significant biological processes and pathways ($p \leq 0.05$) were shown. Our results (Figure 4) showed upregulated hub network 1 was involved in regulation of myeloid cell apoptotic process, cellular extravasation, acute inflammatory response,

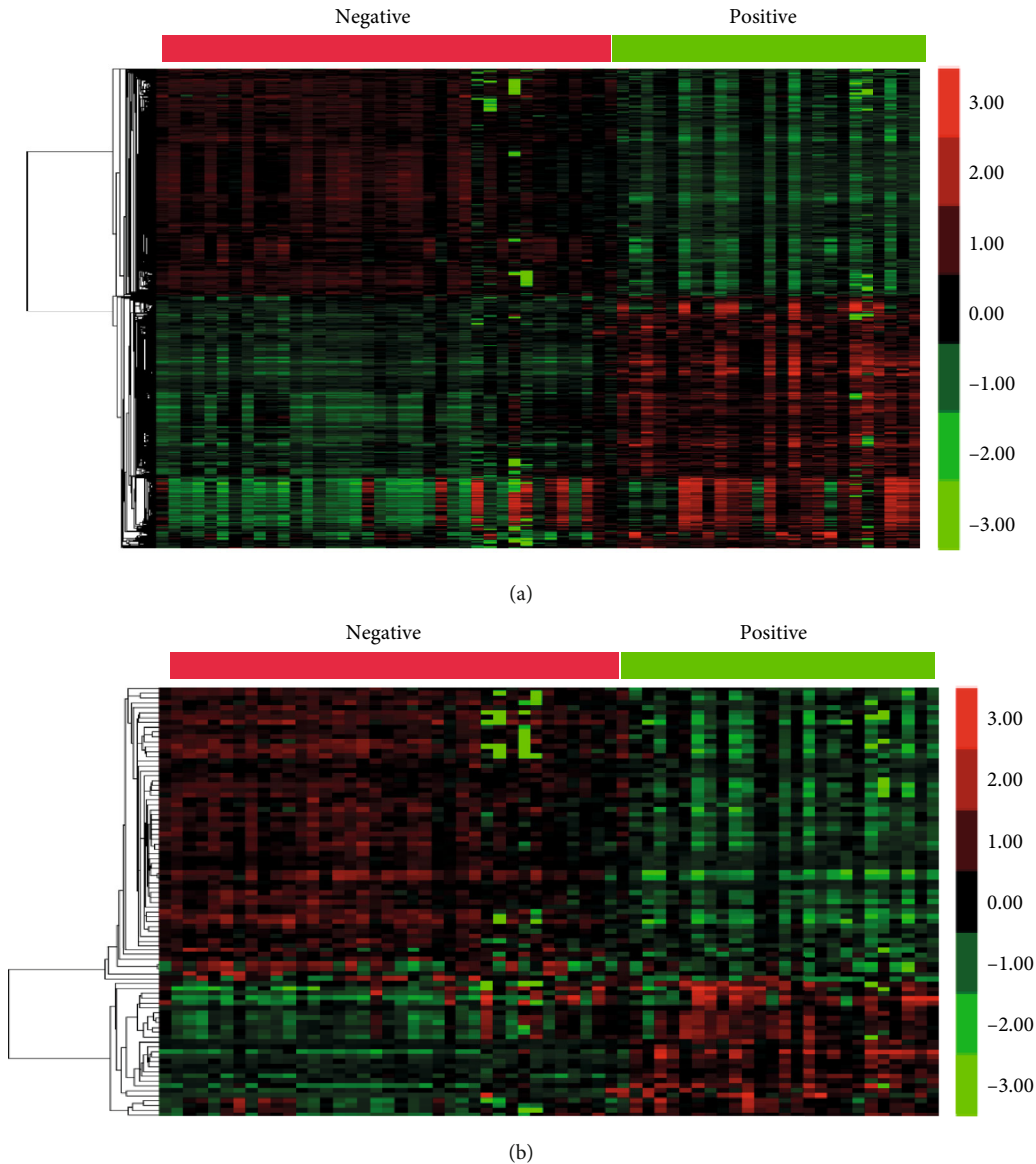


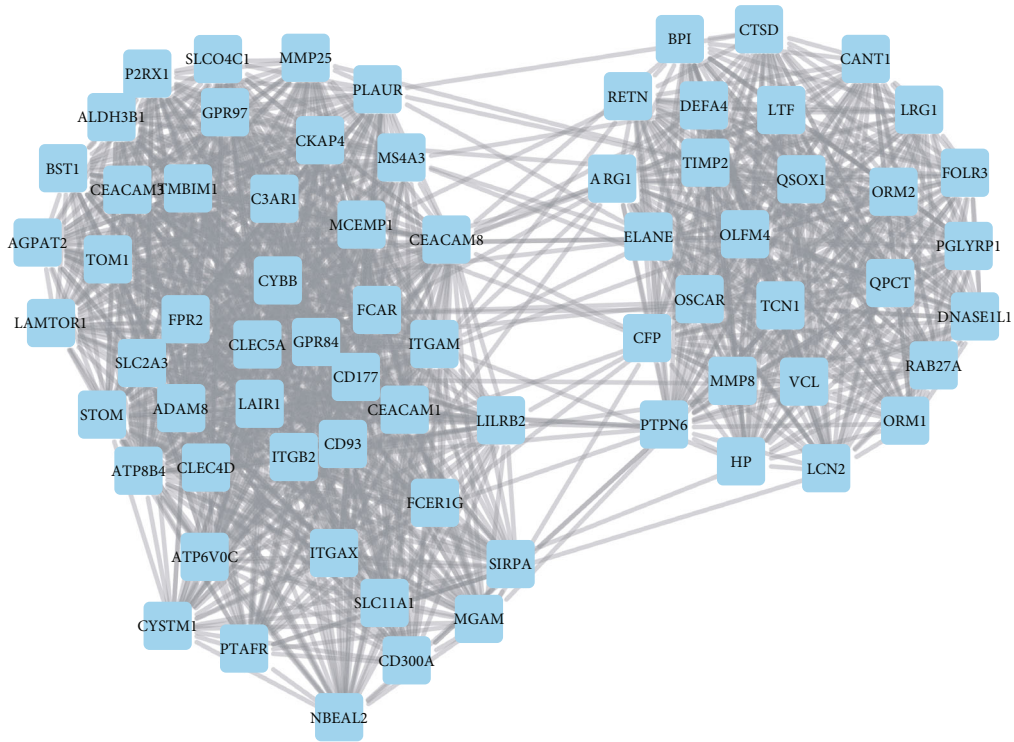
FIGURE 1: Identification of differently expressed mRNAs and lncRNAs in neonatal sepsis by analyzing whole blood mRNA expression profiling. (a) Hierarchical clustering analysis showed differential mRNA expression in negative blood culture result samples and positive blood culture result samples. (b) Hierarchical clustering analysis showed differential lncRNA expression in negative blood culture result samples and positive blood culture result samples.

neutrophil degranulation, macrophage activation, antimicrobial humoral response, and collagen metabolic process. Upregulated hub network 2 was involved in regulating NF-kappa B signaling pathway, TNF signaling pathway, HIF-1 signaling pathway, Toll-like receptor signaling pathway, tuberculosis, legionellosis, and complement and coagulation cascades. Upregulated hub network 3 was involved in regulating ubiquitin-mediated proteolysis, Toll-like receptor signaling pathway, chemokine signaling pathway, and circadian entrainment.

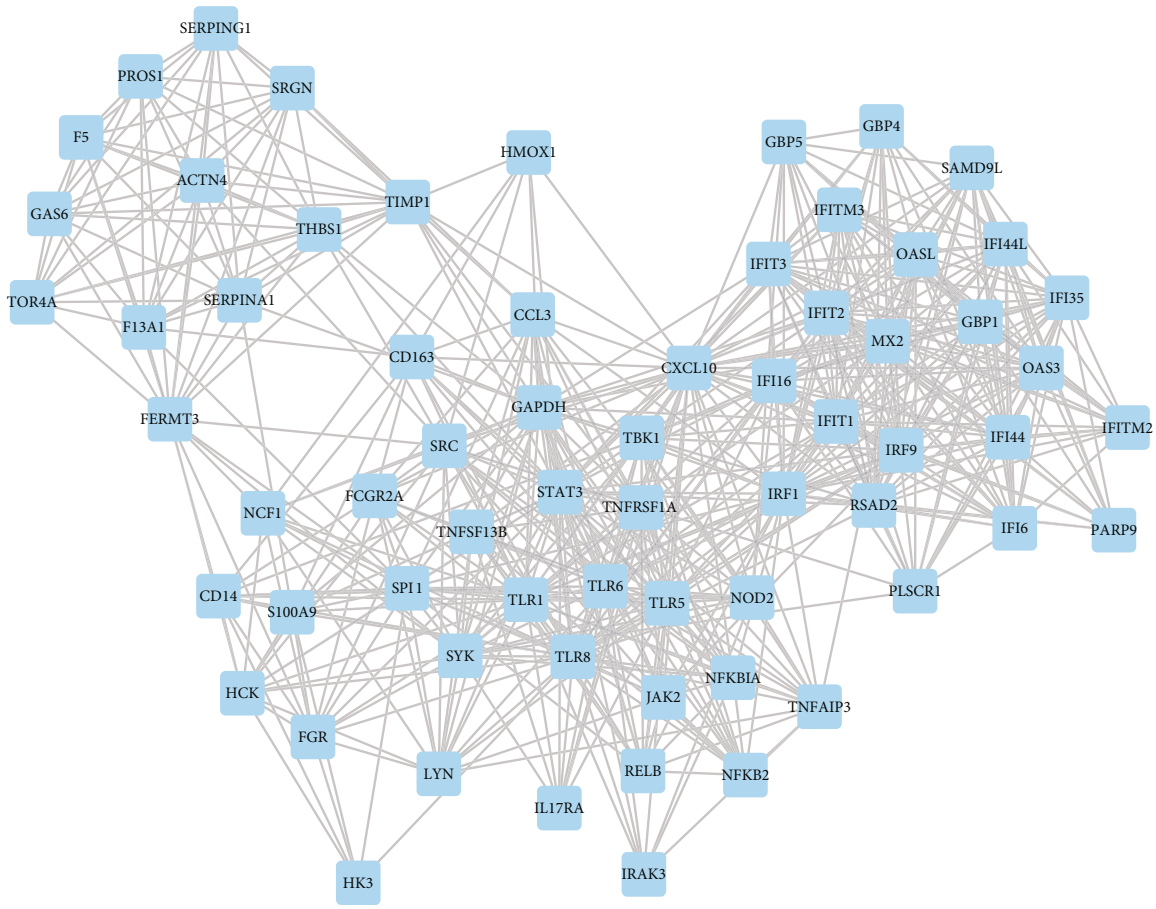
Meanwhile, our results (Figure 5) showed downregulated hub network 1 was involved in regulating ribosome, and RNA transport. Downregulated hub network 2 was involved in regulating Parkinson's disease and oxidative phosphoryla-

tion. Downregulated hub network 3 was involved in regulating spliceosome.

3.4. Coexpression Network Analysis of DElncs in Neonatal Sepsis. We next explored the interactions between mRNAs and lncRNAs. We performed Pearson's correlation calculation of lncRNA-mRNA pair in neonatal sepsis. Based on the correlation analysis results, we constructed an mRNA-lncRNA coexpression network, including 62 lncRNAs, 726 mRNAs, and 2041 interactions between lncRNAs and mRNAs (p value < 0.05 and absolute value of correlation coefficient > 0.85). Eight lncRNAs are significantly associated with more than 100 genes, suggesting their key roles in this network (Figure 6), including HS.294603 (degree =



(a)



(b)

FIGURE 2: Continued.

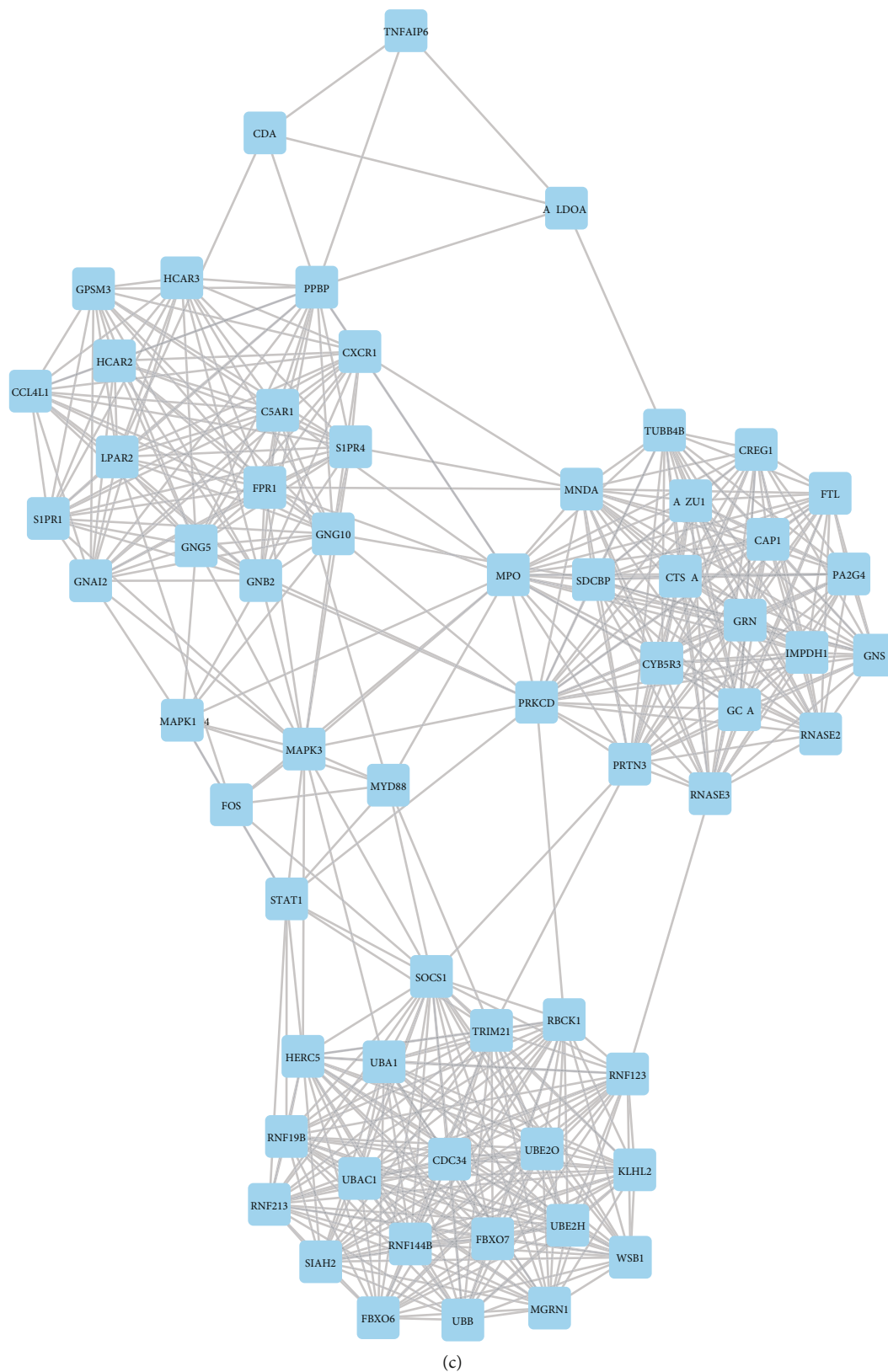


FIGURE 2: Construction of upregulated PPI networks in neonatal sepsis. PPI network analysis showed upregulated hub PPI network 1 (a), hub PPI network 1 (b), and hub PPI network 3 (c).

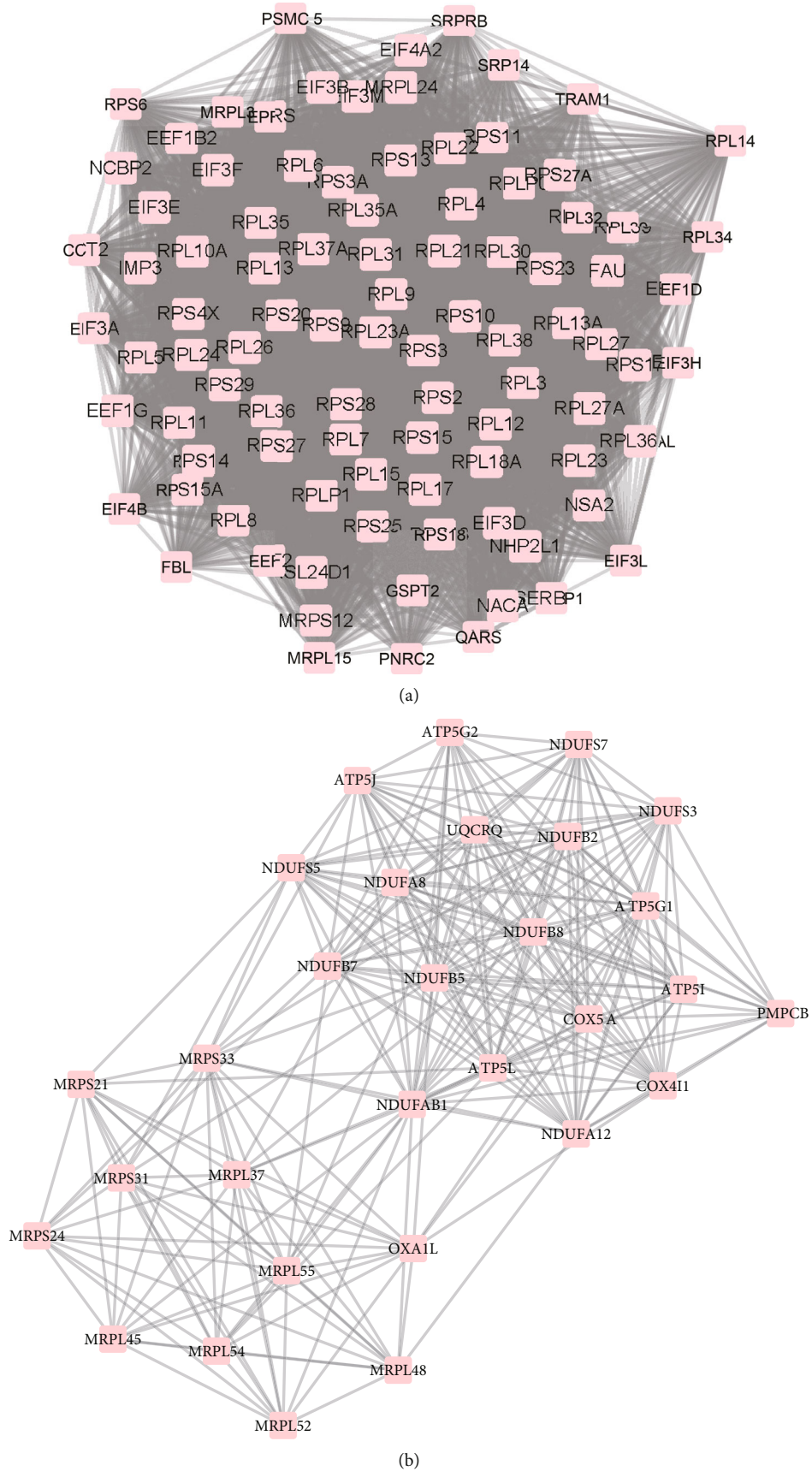


FIGURE 3: Continued.

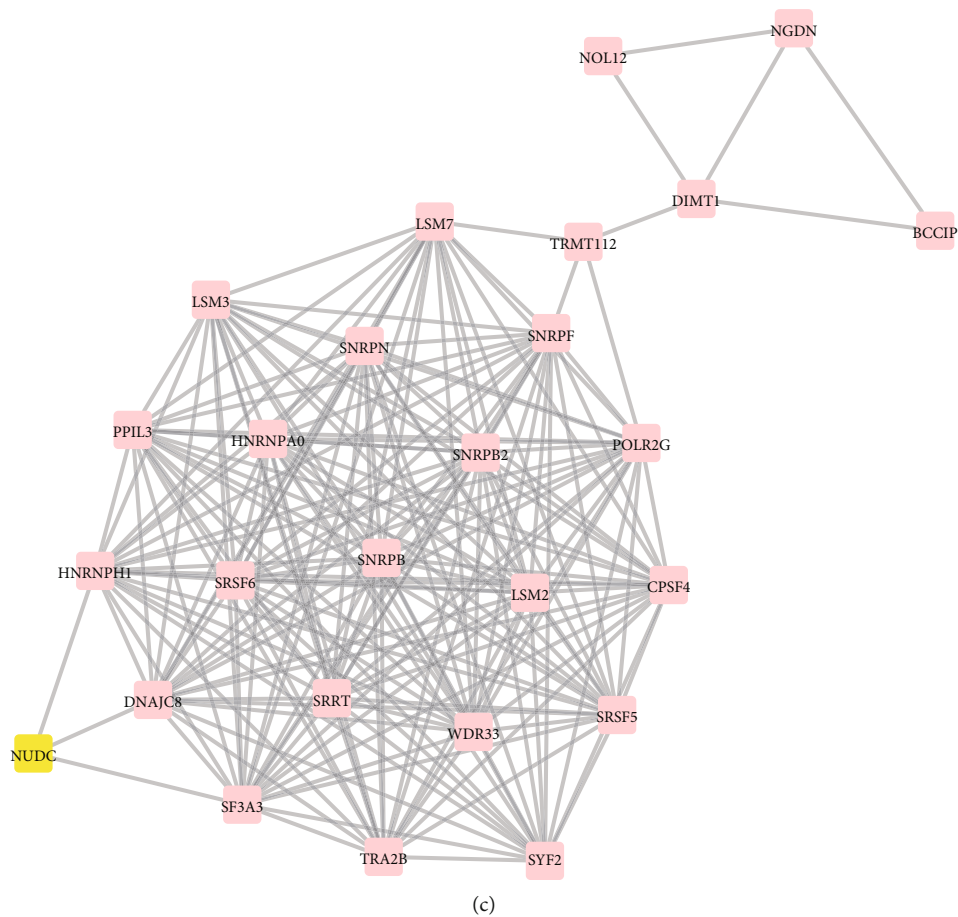


FIGURE 3: Construction of downregulated PPI networks in neonatal sepsis. PPI network analysis showed downregulated hub PPI network 1 (a), hub PPI network 1 (b), and hub PPI network 3 (c).

195), LOC391811 (degree = 179), C12ORF47 (degree = 168), LOC729021 (degree = 155), HS.546375 (degree = 151), HNRPA1L-2 (degree = 127), LOC158345 (degree = 100), and HS.495041 (degree = 100).

3.5. Bioinformatics Analysis of DElncs in Neonatal Sepsis. Bioinformatics analysis for DElncs in neonatal sepsis was also conducted. GO analysis (Figure 7) showed that differentially expressed lncRNAs were associated with translation, cytoplasmic translation, rRNA processing, ribosomal large subunit biogenesis, regulation of translational initiation, tRNA processing, response to peptidoglycan, positive regulation of natural killer cell mediated cytotoxicity, T cell receptor signaling pathway, and negative regulation of apoptotic process. KEGG pathway analysis indicated these lncRNAs were associated with ribosome, T cell receptor signaling pathway, RNA degradation, insulin resistance, ribosome biogenesis in eukaryotes, and hematopoietic cell lineage.

4. Discussion

Neonatal sepsis is the most common cause of death of new born children with few certainly reported biomarkers. Infections remained to be the main risk factor that causes the neonatal death. In the past decades, only few reports indicated

the potential mechanisms underlying the progression of neonatal sepsis. For example, Medzhitov et al. reported that TLR2 and TLR4 were associated with the recognition of bacteria in neonates [12]. Meng et al. identified core regulators involved in the regulation of neonatal sepsis using bioinformatics analysis [13]. Wynn et al. used gene microarray to identify whole genome gene expression change in very low birth weight with neonatal sepsis [14]. The present study is aimed at identifying differently expressed mRNAs and lncRNAs in neonatal sepsis by analyzing GSE25504. A total of 1128 upregulated mRNAs, 1008 downregulated mRNAs, 28 upregulated lncRNAs, and 61 downregulated lncRNAs were identified. Of note, several DEGs identified by this study had also been reported to be associated with neonatal sepsis. For example, IL1R2 and SOCS3 were reported to drive the neonatal innate immune response to sepsis [15]. In order to elucidate the interactions among these DEGs, we constructed upregulated and downregulated genes regulating PPI networks in neonatal sepsis.

Several key genes were identified in neonatal sepsis, including ITGAM, ITGAX, TLR4, ITGB2, SRC, ELANE, RPLP0, RPS28, RPL26, and RPL27. ITGAM (CD11b) was reported as an early diagnostic marker of neonatal sepsis. TLR4 had been reported to be a key regulator in neonatal sepsis [16]. TLR4 was associated with the recognition of the

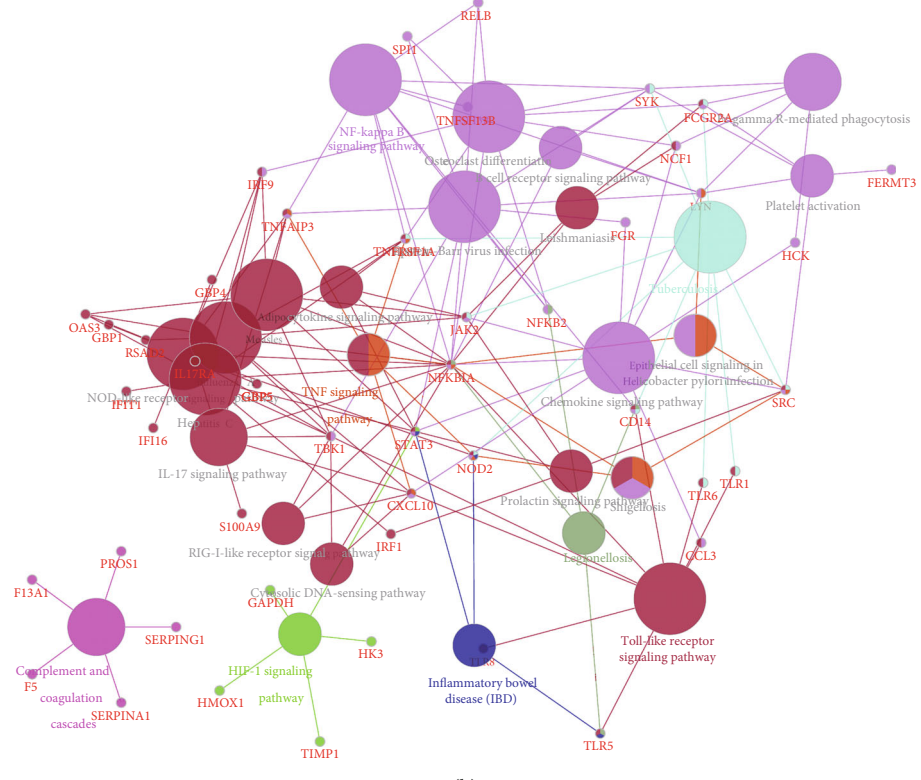
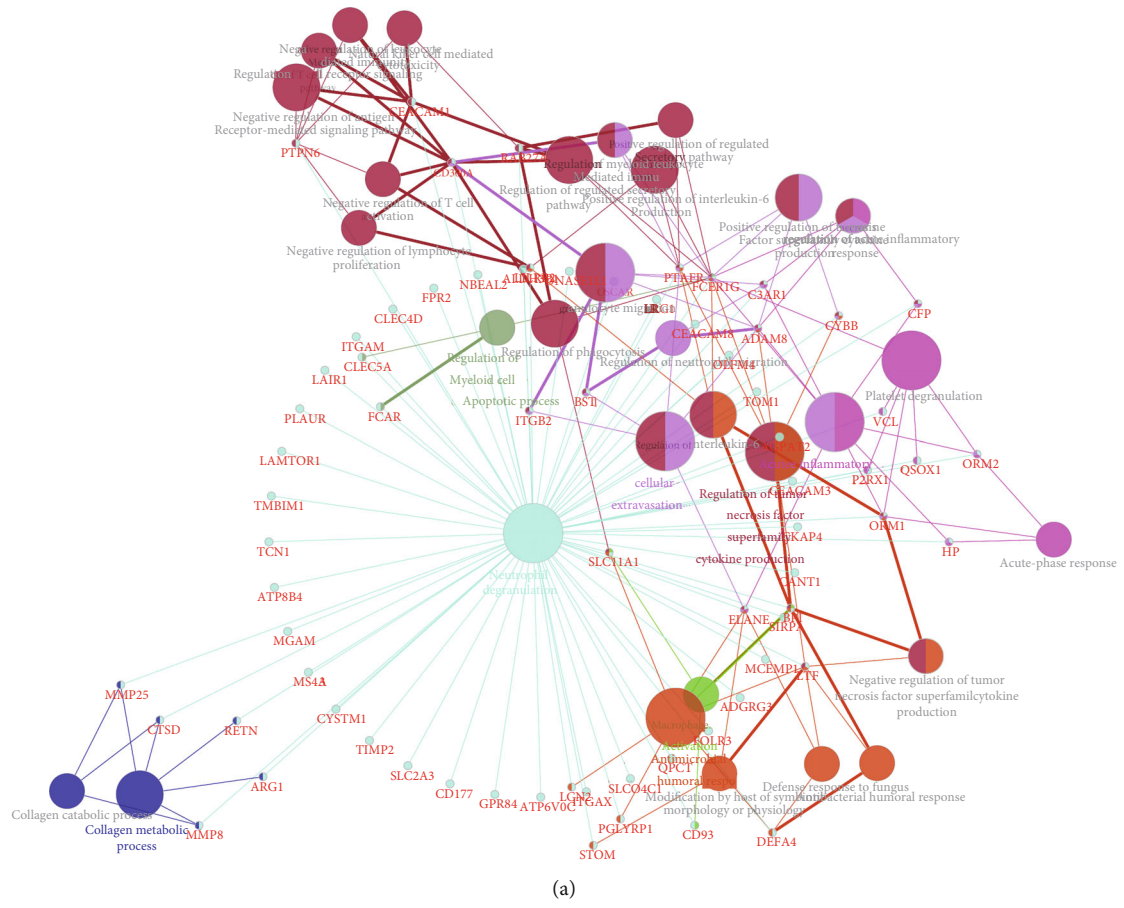


FIGURE 4: Continued.

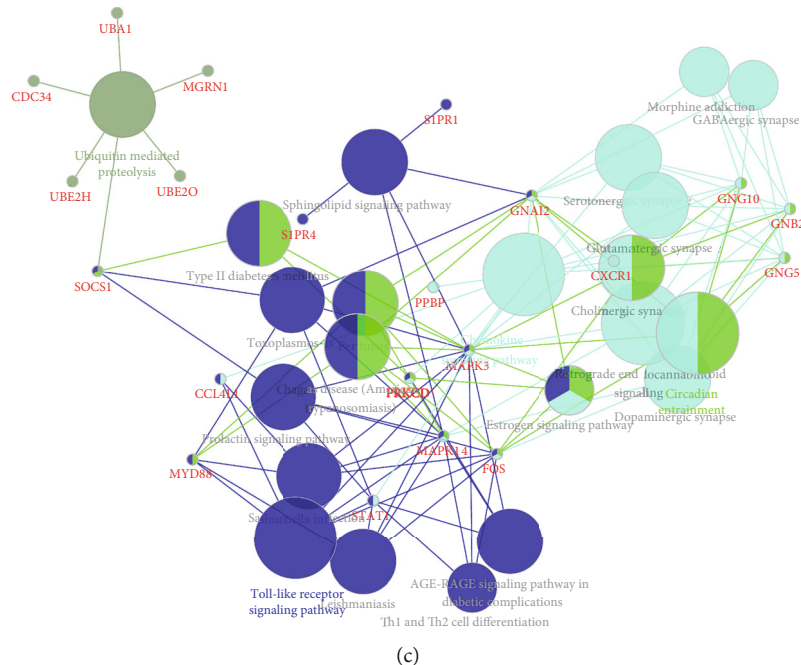


FIGURE 4: Bioinformatics analysis of hub upregulated PPI networks in neonatal sepsis. Bioinformatics analysis of up-regulated hub PPI network 1 (a), hub PPI network 1 (b), and hub PPI network 3 (c).

bacteria in neonates [17]. The single-nucleotide polymorphisms (SNPs) in TLR4 were regarded as genetic modulators of infection in neonatal sepsis [18]. Bioinformatics analysis revealed these DEGs were significantly associated with multiple biological processes, including myeloid cell apoptotic process, cellular extravasation, acute inflammatory response, neutrophil degranulation, macrophage activation, antimicrobial humoral response, collagen metabolic process, NF-kappa B signaling pathway, TNF signaling pathway, HIF-1 signaling pathway, Toll-like receptor signaling pathway, tuberculosis, legionellosis, complement and coagulation cascades, chemokine signaling pathway, circadian entrainment, ribosome, RNA transport, and spliceosome. This signaling had been demonstrated to play crucial roles in neonatal sepsis. For example, altered neonatal Toll-like receptor (TLR) function is hypothesized to contribute to the heightened susceptibility to infection and perpetuated inflammation in term and preterm neonates, clinically evident in neonatal sepsis and increased rates of inflammatory disorders [19].

Emerging studies had demonstrated noncoding RNAs, such as lncRNAs and miRNAs, were involved in regulating the progression of human diseases. In neonatal sepsis, multiple miRNAs were reported. For example, microRNA-300/NAMPT regulates inflammatory responses through activation of the AMPK/mTOR signaling pathway in neonatal sepsis [20]. miR-15a/16 are upregulated in the serum of neonatal sepsis patients and inhibit the LPS-induced inflammatory pathway [21]. lncRNAs were a type of ncRNAs longer than 200 bps. Emerging evidences showed lncRNAs played important roles in human diseases, such as diabetes, multiple cancers, and neurodegenerative diseases. A recent study showed lncRNA SNHG16 reverses the effects of miR-15a/16

on the LPS-induced inflammatory pathway in neonatal sepsis [6]. However, the molecular functions of lncRNAs in neonatal sepsis remained unclear. This study identified 28 upregulated lncRNAs and 61 downregulated lncRNAs in neonatal sepsis. Coexpression analysis were used to identify key lncRNAs, including HS.294603, LOC391811, C12ORF47, LOC729021, HS.546375, HNRPA1L-2, LOC158345, and HS.495041. Bioinformatics analysis showed these lncRNAs were involved in regulating ribosome, T cell receptor signaling pathway, RNA degradation, insulin resistance, ribosome biogenesis in eukaryotes, and hematopoietic cell lineage.

In this study, there also existed some limitations. Firstly, more samples were needed considering the small sample size in the present study. Secondly, further experimental validation would be required for future verification. Moreover, specific functions of those dysregulated circRNAs had not been further excavated in this study. Therefore, the further researches with a larger samples group should be performed and more experimental validation and much deeper analysis were still needed in the near future.

5. Conclusion

In conclusion, the present study identified a total of 1128 upregulated mRNAs, 1008 downregulated mRNAs, 28 upregulated lncRNAs, and 61 downregulated lncRNAs in neonatal sepsis. Then, we constructed PPI networks to identify key regulators in neonatal sepsis, including ITGAM, ITGAX, TLR4, ITGB2, SRC, ELANE, RPLP0, RPS28, RPL26, and RPL27. lncRNA coexpression analysis showed HS.294603, LOC391811, C12ORF47, LOC729021, HS.546375, HNRPA1L-2, LOC158345, and HS.495041 played important roles in

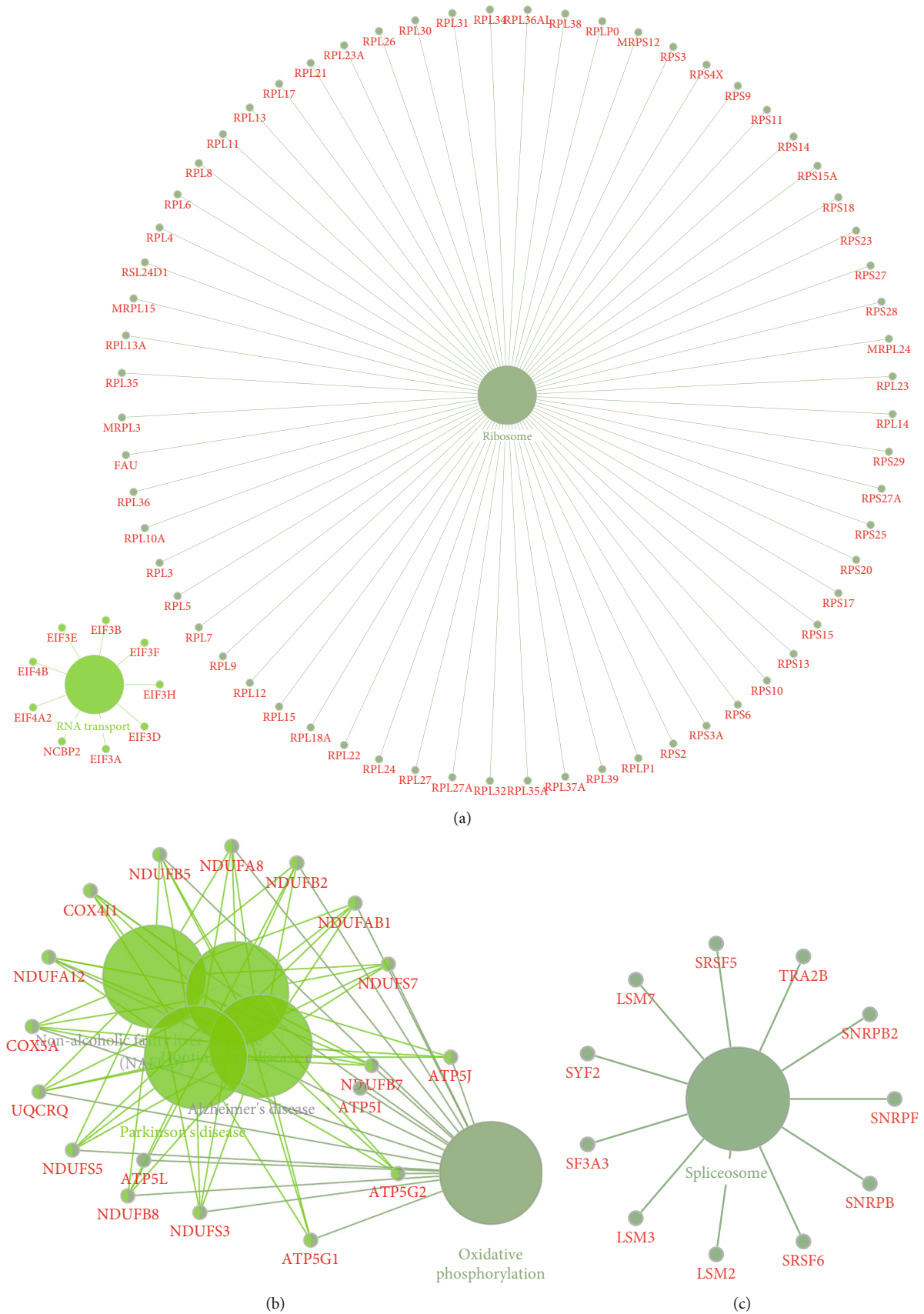


FIGURE 5: Bioinformatics analysis of hub downregulated PPI networks in neonatal sepsis. Bioinformatics analysis of down-regulated hub PPI network 1 (a), hub PPI network 1 (b), and hub PPI network 3 (c).

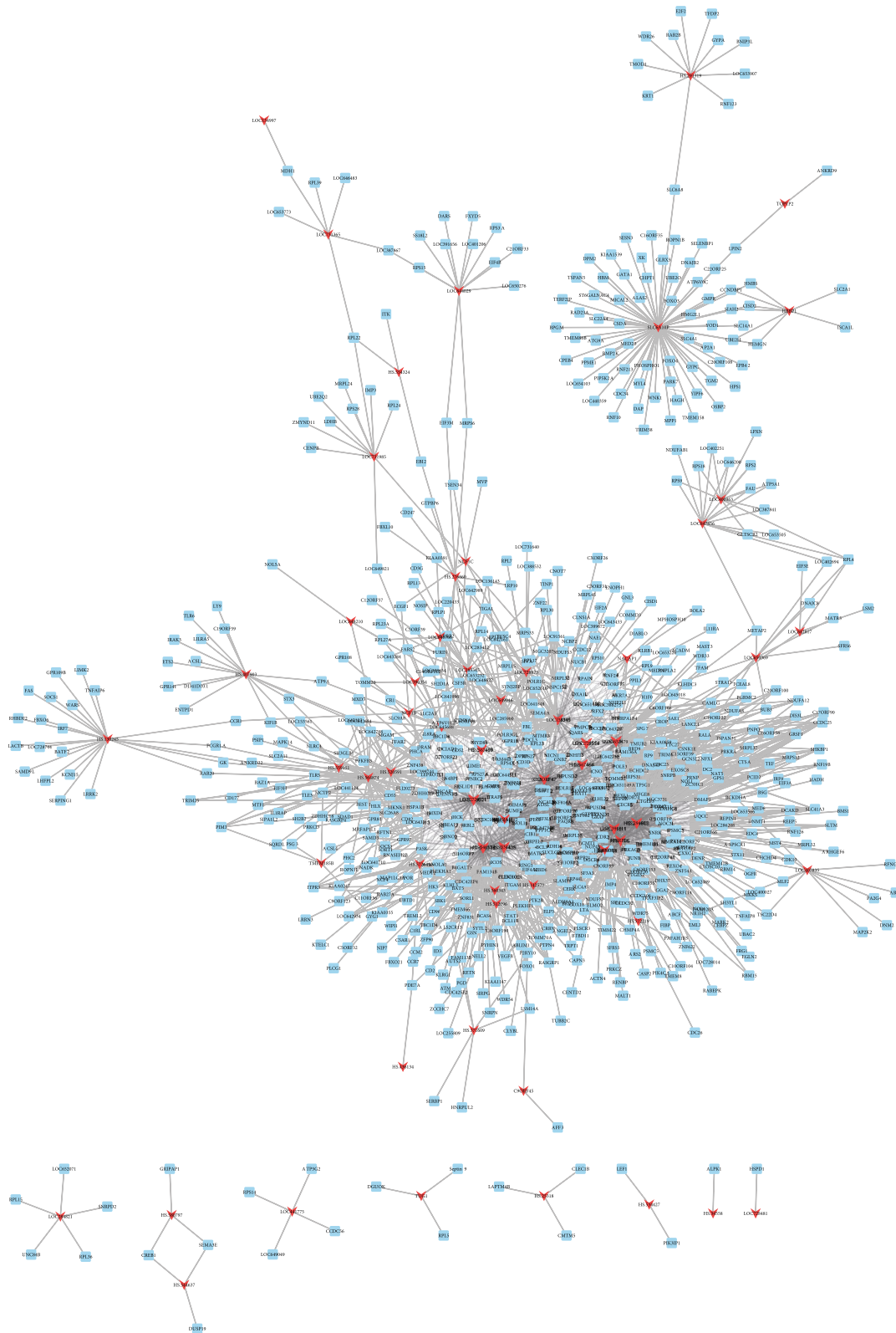


FIGURE 6: Construction of differently expressed lncRNAs regulating coexpression networks in neonatal sepsis.

the progression of neonatal sepsis. Bioinformatics analysis showed DEGs were involved in the regulation cellular extravasation, acute inflammatory response, macrophage activation

of NF-kappa B signaling pathway, TNF signaling pathway, HIF-1 signaling pathway, Toll-like receptor signaling pathway, and ribosome, RNA transport, and spliceosome.

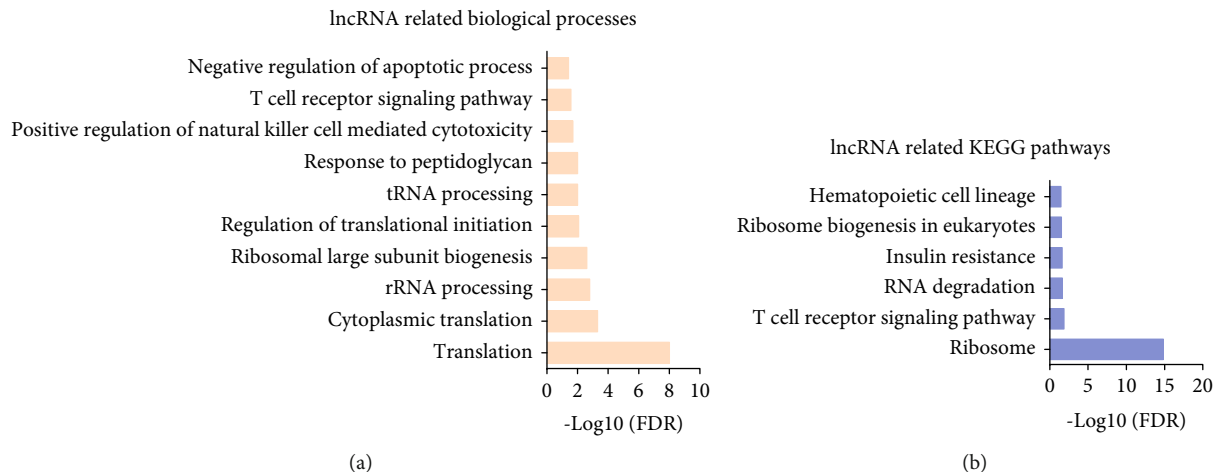


FIGURE 7: Bioinformatics analysis of differently expressed lncRNAs regulating coexpression networks in neonatal sepsis. (a, b) GO and KEGG analysis of differently expressed lncRNAs in neonatal sepsis.

lncRNAs were involved in regulating ribosome, T cell receptor signaling pathway, RNA degradation, insulin resistance, ribosome biogenesis in eukaryotes, and hematopoietic cell lineage. We thought this study provided useful information for identifying novel therapeutic markers for neonatal sepsis.

Data Availability

All the data were reserved and can be accessed in GSE25504 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25504>).

Conflicts of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

Lin Bu and Xiao-min Li guaranteed the integrity of the entire study and reviewed the manuscript. Xiao-min Li is assigned to the study concepts and definition of intellectual content. Lin Bu is responsible for the study design, literature research, manuscript preparation, and manuscript editing. Lin Bu and Shu-qun Hu performed the data acquisition, data analysis, and statistical analysis. Lin Bu, Wen-jing Zhao, Xiao-juan Geng, and Yan-li Wang participated in the clinical studies. Lin Bu, Ting Zhou, Luo Zhuo, Xiao-bing Chen, and Yan Sun participated in the experimental studies.

References

- [1] N. Chauhan, S. Tiwari, and U. Jain, "Potential biomarkers for effective screening of neonatal sepsis infections: An overview," *Microbial Pathogenesis*, vol. 107, pp. 234–242, 2017.
- [2] S. ULLAH, K. RAHMAN, and M. HEDAYATI, "Hyperbilirubinemia in neonates: types, causes, clinical examinations, preventive measures and treatments: a narrative review article," *Iranian Journal of Public Health*, vol. 45, no. 5, pp. 558–568, 2016.
- [3] C. Pietrasanta, L. Pugni, A. Ronchi et al., "Vascular Endothelium in Neonatal Sepsis: Basic Mechanisms and Translational Opportunities," *Frontiers in pediatrics*, vol. 7, 2019.
- [4] B. Schaub, A. Bellou, F. K. Gibbons et al., "TLR2 and TLR4 stimulation differentially induce cytokine secretion in human neonatal, adult, and murine mononuclear cells," *Journal of Interferon and Cytokine Research*, vol. 24, no. 9, pp. 543–552, 2004.
- [5] Y. X. Meng, X. H. Cai, and L. P. Wang, "Potential Genes and Pathways of Neonatal Sepsis Based on Functional Gene Set Enrichment Analyses," *Computational and Mathematical Methods in Medicine*, vol. 2018, 10 pages, 2018.
- [6] X. Wang, X. Wang, X. Liu et al., "miR-15a/16 are upregulated in the serum of neonatal sepsis patients and inhibit the LPS-induced inflammatory pathway," *International Journal of Clinical and Experimental Medicine*, vol. 8, no. 4, pp. 5683–5690, 2015.
- [7] C. L. Smith, P. Dickinson, T. Forster et al., "Identification of a human neonatal immune-metabolic network associated with bacterial infection," *Nature Communications*, vol. 5, no. 1, 2014.
- [8] P. Dickinson, C. L. Smith, T. Forster et al., "Whole blood gene expression profiling of neonates with confirmed bacterial sepsis," *Genomics data*, vol. 3, pp. 41–48, 2015.
- [9] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, Springer, 2005.
- [10] D. Szklarczyk, A. L. Gable, D. Lyon et al., "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, 2019.
- [11] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: software for visualization and analysis of biological networks," *Methods in Molecular Biology*, vol. 696, pp. 291–303, 2011.
- [12] S. A. Spector, M. Qin, J. Lujan-Zilbermann et al., "Genetic variants in toll-like receptor 2 (TLR2), TLR4, TLR9, and FCγ receptor II are associated with antibody response to quadrivalent meningococcal conjugate vaccine in HIV-infected youth," *Clinical and Vaccine Immunology*, vol. 20, no. 6, pp. 900–906, 2013.

- [13] K. Meng, X. Hu, X. Peng, and Z. Zhang, "Incidence of venous thromboembolism during pregnancy and the puerperium: a systematic review and meta-analysis," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 28, no. 3, pp. 245–253, 2015.
- [14] J. L. Wynn, S. O. Guthrie, H. R. Wong et al., "Postnatal Age Is a Critical Determinant of the Neonatal Host Response to Sepsis," *Molecular Medicine*, vol. 21, no. 1, pp. 496–504, 2015.
- [15] H. Jiang, C. van de Ven, L. Baxi, P. Satwani, and M. S. Cairo, "Differential gene expression signatures of adult peripheral blood vs cord blood monocyte-derived immature and mature dendritic cells," *Experimental Hematology*, vol. 37, no. 10, pp. 1201–1215, 2009.
- [16] J. P. Zhang, C. Chen, and Y. Yang, "Changes and clinical significance of Toll-like receptor 2 and 4 expression in neonatal infections," *Chinese journal of pediatrics*, vol. 45, no. 2, p. 130, 2007.
- [17] M. Triantafilou, F. G. J. Gamper, P. M. Lepper et al., "Lipopolysaccharides from atherosclerosis-associated bacteria antagonize TLR4, induce formation of TLR2/1/CD36 complexes in lipid rafts and trigger TLR2-induced inflammatory responses in human vascular endothelial cells," *Cellular Microbiology*, vol. 9, no. 8, pp. 2030–2039, 2007.
- [18] A. Gast, J. L. Bermejo, R. Claus et al., "Association of inherited variation in Toll-like receptor genes with malignant melanoma susceptibility and survival," *PLoS One*, vol. 6, no. 9, p. e24370, 2011.
- [19] L. Stridh, A. Mottahedin, M. E. Johansson et al., "Toll-like receptor-3 activation increases the vulnerability of the neonatal brain to hypoxia-ischemia," *The Journal of Neuroscience*, vol. 33, no. 29, pp. 12041–12051, 2013.
- [20] Y. Li, J. Ke, C. Peng, F. Wu, and Y. Song, "microRNA-300/NAMPT regulates inflammatory responses through activation of AMPK/mTOR signaling pathway in neonatal sepsis," *Biomedicine & Pharmacotherapy*, vol. 108, pp. 271–279, 2018.
- [21] W. Wang, C. Lou, J. Gao, X. Zhang, and Y. Du, "LncRNA SNHG16 reverses the effects of miR-15a/16 on LPS-induced inflammatory pathway," *Biomedicine & Pharmacotherapy*, vol. 106, pp. 1661–1667, 2018.

Research Article

Comparison of the Therapeutic Effects of Tension Band with Cannulated Screw and Tension Band with Kirschner Wire on Patella Fracture

Chengwu Liu ¹, Haitao Ren ², Chunyan Wan ³, and Jianlin Ma ⁴

¹Department of Orthopedics, Qingdao Chengyang People's Hospital, Qingdao, China

²Department of Microsurgery, Qingdao Chengyang People's Hospital, Qingdao, China

³Department of Surgery, Qingdao Chengyang People's Hospital, Qingdao, China

⁴Department of Spinal Arthrology, Qingdao Chengyang People's Hospital, Qingdao, China

Correspondence should be addressed to Jianlin Ma; mjlcyrmyy@163.com

Received 22 May 2020; Revised 1 July 2020; Accepted 4 July 2020; Published 25 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Chengwu Liu 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.

Background. Patella fracture accounts for 1% of bone injury, of which anatomical reduction is of great significance to the recovery. Tension band with cannulated screw and Kirschner wire is commonly used methods for the treatment of displaced patella fracture. However, there is still some controversy on the clinical efficacy of the two treatment methods. **Objective.** This study aimed at comparing the therapeutic effects of the cannulated screw and Kirschner wire tension bands on patella fracture and at providing more data basis for clinical selection of treatment methods for patella fracture. **Methods.** Altogether, 146 patients with displaced patella fracture admitted to our hospital from March 2016 to February 2018 were selected and divided into two groups according to the different treatment methods. Among them, 71 patients received tension band with a cannulated screw (TBWCS group) and 75 patients received tension band with Kirschner wire (TBWKW group). Two groups of patients were compared in terms of surgical treatment effect after one year of treatment, complications within six months after the operation and operation-related indexes. The pain visual analogue scale (VAS) score, knee flexion degree, Lysholm score, and Bostman score were recorded at 1, 3, 6, and 12 months after operation, and the activity of daily living scale (ADL) score was evaluated at the last follow-up. **Results.** During the operation of patella fracture patients, the intraoperative blood loss, hospitalization time, and knee flexion loss of patients in TBWCS group were less than those in the TBWKW group ($P < 0.05$), the starting time of postoperative functional exercise was earlier than that of patients in TBWKW group ($P < 0.05$), and the incidence rate of secondary operation was lower than that of patients in the TBWKW group ($P < 0.05$), but there was no statistical difference in the operation time, incision length, and postoperative fracture gap between the two groups. The results of curative effect analysis showed that the knee flexion, Lysholm score, and Bostman score of patients treated with tension band with cannulated screw were higher than those treated with Kirschner wire ($P < 0.05$), and VAS score was lower. Tension band with cannulated screw had a better curative effect on patella fracture ($P < 0.05$), lower complication rate ($P < 0.05$), and higher quality of life of patients ($P < 0.05$). **Conclusion.** Tension band with cannulated screw has a good curative effect on patella fracture, low incidence of complications, early start of postoperative functional exercise, and high quality of life.

1. Introduction

Patella is the largest sesamoid bone of the human body, which has the function of transmitting muscle strength during knee extension [1]. Patella fracture accounts for 1% of bone injury [2]. If not treated properly, complications such as traumatic knee arthritis and knee function limitation

may occur [3, 4]. Anatomical reduction is of great significance to the recovery of the patella function.

Intraoperative fixation is a common method for the treatment of displaced patella fracture, which can better recover joint function and quadriceps femoris function, and prevent osteoarthritis [5, 6]. Traditionally, the Kirschner wire tension band method is widely used, which is superior to other

surgical methods in terms of tension resistance and can recover patella function to the maximum extent [7]. However, foreign body reaction caused by Kirschner wire indwelling affects skin and tissues, such as inflammatory granulation tissue [8]. In addition, Kirschner wire loosening, bending, and the like may also be caused during knee exercise, resulting in treatment failure [9]. About 10%-20% of the patients suffered from displacement between fracture fragments, and 5% of the patients underwent two operations [10]. In view of this, medical workers made an improvement plan, namely, cannulated screw method, using a cannulated screw instead of Kirschner wire, cannulated screw can be embedded into bone tissue, causing less subcutaneous foreign body sensation, and can be used for annular binding fixation of more severely crushed fracture blocks without lacunae [11, 12]. However, a tension band with a cannulated screw will irritate the skin, and the metal tail will sometimes sting and damage the bone [13]. Therefore, there is still some controversy on the clinical efficacy of the two treatment methods.

This study compared the therapeutic effects of cannulated screw and Kirschner wire with tension band on patella fracture, providing more data basis for clinical selection of treatment methods for patella fracture.

2. Data and Methods

2.1. Research Participants. This study applied prospective analysis. Altogether, 146 patients with displaced patella fracture admitted to our hospital from March 2016 to February 2018 were selected and divided into two groups according to the different treatment methods. Among them, 71 patients received tension band with a cannulated screw (TBWCS group), and 75 patients received tension band with Kirschner wire (TBWKW group).

2.2. Inclusion and Exclusion Criteria. Inclusion criteria: patients were aged 30-75 years old; patients underwent anteroposterior and lateral X-ray films; patients were diagnosed as acute closed patella fracture with separation and displacement by X-ray; patients had complete medical records and follow-up data; patients signed informed consent forms and could cooperate with medical staff to complete relevant diagnosis and treatment work. Exclusion criteria: patients with pathological, open, and comminuted fractures; patients with Rockwood classification of patella fractures in I, VI, and VII; patients with suppurative infection of joints, old fractures, severe acetabular destruction, or obvious degeneration; patients combined with fractures of other parts; patients with cardiopulmonary dysfunction and severe diabetes (fasting blood glucose greater than 10 mmol/L); patients with cognitive impairment could not cooperate with follow-up.

2.3. Treatment Method. Patients in the TBWCS group were treated with a tension band with a cannulated screw, while patients in the TBWKW group were treated with a tension band with Kirschner wire.

Kirschner wire tension band: the patient underwent general anesthesia and took the supine position. A midline longitudinal incision was taken on the patella, the full-thickness

skin flap was raised, the fracture position was exposed, and then thorough debridement was carried out. Under the perspective of the c-arm X-ray machine, Kirschner wire was pinned to replace and fix the fracture fragments, and continuous traction was carried out to maintain the correction stability. Kirschner wire with a diameter of 2 mm was used to fix it, and the joint cavity was cleaned. After the fixation, the 8-figure fixation outside the wire, the deep tissue was buckled and embedded, the joint cavity was cleaned, the drainage tube was retained, the wound was sutured, and the pressure dressing was carried out after the operation.

Tension band with cannulated screw: the patient underwent general anesthesia and took the supine position. A midline longitudinal incision was taken on the patella, the full-thickness skin flap was raised, the fracture position was exposed, then, thorough debridement was carried out. A sharp reduction clamp was carried out to temporarily fix the fracture fragments. After the finger proved that the patella joint surface was flat, two Kirschner wires with a diameter of 1.6 mm was used to longitudinally pass through the fractured patella, and the Kirschner wires were used as parallel as possible and located in the anterior 1/3 of the patella. Under the guidance of Kirschner wire, a 4.5 mm semithreaded stainless steel cannulated compression screw was screwed in. After the cannulated screw was embedded into the patella bone, the needle was pulled out. Steel wire was inserted from the screw, fixed in the shape of 8-figure, buckled, and embedded into deep tissue. The joint cavity was then cleaned, the drainage tube was retained, and the wound was sutured.

2.4. Follow-Up Arrangements. The patients in this study were followed up for 12 months. The patients were followed up, and the pain visual analogue scale (VAS), knee flexion, Lysholm score, and Bostman score were recorded at 1, 3, 6, and 12 months after operation, and the activity of daily living scale (ADL) score was evaluated at the last follow-up. ADL Score includes tips for assessing a resident's need for assistance with activities of daily living (ADLs). VAS score was 0-10, and the high score was closely related to the severity of the pain; Lysholm score was 0-100, and the high score was closely related to the better joint function; Bostman score was 0-30, and the high score was closely related to the better recovery of knee joint function; ADL score was 0-100, and the high score was closely related to the better quality of life of patients.

2.5. Observation Index. Two groups of patients were compared in terms of surgical treatment effect after one year of treatment, complication occurrence within 6 months after the operation and operation-related indexes, including operation time, incision length, fracture gap after the operation (Measurement was conducted according to CT image), intraoperative blood loss, hospitalization time (Measurement was conducted according to CT image), angle of knee flexion loss of affected limb, the start time of postoperative functional exercise, and incidence rate of secondary operation.

2.6. Efficacy Evaluation Criteria. (1) Markedly effective: patients had no pain in the knee joint after operation; knee

TABLE 1: Comparison of general data of two groups of patients (mean \pm SD; n , %).

	TBWCS ($n = 71$)	TBWKW ($n = 75$)	χ^2/t	P
Gender			0.392	0.531
Male	33 (46.48)	31 (41.33)		
Female	38 (53.52)	44 (58.67)		
Age (years)	57.23 \pm 8.67	60.74 \pm 14.82	1.734	0.085
Body mass index (kg/cm ²)	22.49 \pm 1.83	22.36 \pm 1.86	0.671	0.425
Affected side			3.441	0.064
Left	44 (61.97)	35 (46.67)		
Right	27 (38.03)	40 (53.33)		
Cause of fracture			1.705	0.426
Fall injury	41 (57.75)	40 (53.33)		
Traffic accident	19 (26.76)	17 (22.67)		
Other	11 (15.49)	18 (24.00)		
AO/OTA			0.913	0.633
Transverse, middle (45-C1.1)	41 (57.75)	49 (65.33)		
Transverse, proximal (45-C1.2)	11 (15.49)	9 (12.00)		
Transverse, distal (45-C1.3)	19 (26.76)	17 (22.67)		
Underlying diseases			0.087	0.957
Hypertension	10 (14.08)	9 (12.00)		
Coronary heart disease	14 (19.72)	15 (20.00)		
Diabetes	5 (7.04)	5 (6.67)		
ASA classification			0.185	0.912
I	24 (33.80)	23 (30.67)		
II	42 (59.15)	46 (61.33)		
III	5 (7.04)	6 (8.00)		
Displaced distance of fracture fragment (mm)	4.3 \pm 1.8	4.7 \pm 1.7	1.381	0.169
Injury time before operation (days)	2.8 \pm 1.2	3.3 \pm 1.4	1.849	0.067

joint movement was normal; imaging examination showed that bone healing was satisfactory, there was no traumatic arthritis, bursitis, and other complications. (2) Effective: the patient's knee joint movement was slightly limited after operation, and the bone healing was satisfactory by imaging examination. (3) Ineffective: none of the above curative effects have been achieved, delayed healing, or malunion of fracture occurred, and knee joint movement was limited. Total effective rate = markedly effective rate + effective rate.

2.7. Statistical Analysis. SPSS 19.0 (Asia Analytics Formerly SPSS China) was used to analyze the data. Measurement data were expressed by %, and the comparison of rates used χ^2 test. The counting data were expressed by Mean \pm standard deviation (mean \pm SD). K-S was used to test whether the data conform to the normal distribution, the Wilcoxon test was used for nonnormal distribution data between the two groups. The comparison between the two groups for normal distribution data adopted t -test. The comparison between the two groups adopted t -test, the comparison at different time points adopted repeated measurement analysis of variance, and the back testing adopted the LSD test. $P < 0.05$ indicates that the difference is statistically significant.

TABLE 2: Clinical efficacy (N , %).

	TBWCS ($n = 71$)	TBWKW ($n = 75$)	χ^2	P
Markedly effective	54 (76.06)	49 (65.33)	2.018	0.155
Effective	17 (23.94)	20 (26.67)	0.143	0.705
Ineffective	0 (0.00)	6 (8.00)	Fisher	0.028
Total efficiency	71 (100.00)	69 (92.00)	Fisher	0.028

3. Results

3.1. General Data. There were 71 patients in the TBWCS group, including 33 male and 38 female patients, with an age of (57.23 \pm 8.67) years, and 75 patients in the TBWKW group, including 31 male and 44 female patients, with an age of (60.74 \pm 14.82) years. There was no statistical difference in gender ratio and age between the two groups. There was no significant difference between the two groups in other data, such as body mass index and fracture causes ($P > 0.05$), see Table 1 for details.

3.2. Clinical Efficacy. One year after treatment, the markedly effective rate, effective rate, and ineffective rate of patients in the TBWCS group were 76.06%, 23.94%, and 0.00%,

TABLE 3: Operation related indicators (mean \pm SD).

	TBWCS ($n = 71$)	TBWKW ($n = 75$)	χ^2/t	P
Operation time (min)	57.37 \pm 9.83	54.87 \pm 9.58	1.556	0.122
Incision length (cm)	6.24 \pm 1.02	5.94 \pm 1.09	1.715	0.089
Intraoperative blood loss (mL)	155.38 \pm 28.37	187.47 \pm 23.27	7.490	<0.001
Hospitalization time (days)	13.58 \pm 3.53	14.57 \pm 1.12	2.309	0.022
Start time of postoperative functional exercise (days)	32.13 \pm 5.88	46.12 \pm 9.02	11.035	<0.001
Fracture healing time (weeks)	10.7 \pm 1.7	13.2 \pm 3.1	5.995	<0.001
Angle of limb loss (angles)	10.8 \pm 5.6	19.3 \pm 7.2	7.932	<0.001
Fracture gap after operation			0.669	0.716
0 mm	52 (73.24)	58 (77.33)		
≤ 2 mm	14 (19.72)	11 (14.67)		
≥ 3 mm	5 (7.04)	6 (8.00)		
Secondary operation	1 (1.41)	11 (14.67)	6.832	0.009

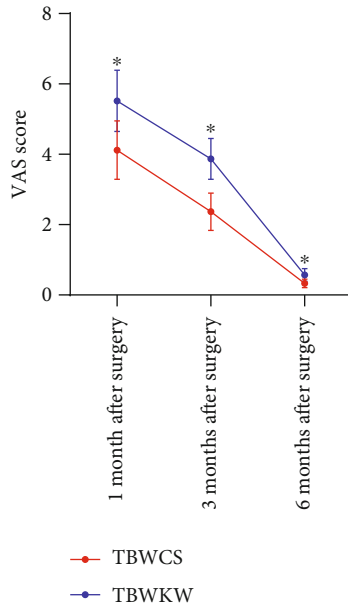


FIGURE 1: Difference of VAS score between the two groups within 6 months after the operation. Repeated measurement analysis of variance showed that the VAS scores of the two groups of patients decreased gradually with time. The VAS scores of patients in the TBWCS group ($n = 71$) were lower than those in the TBWKW group ($n = 75$) at 1, 3, and 6 months after operation. * indicates compared with the TBWCS group, $P < 0.05$.

respectively, while those of patients in the TBWKW group were 65.33%, 26.67%, and 8.00%, respectively. There was no statistical difference in the markedly effective rate and effective rate between the two groups ($P > 0.05$), but the ineffective rate of patients in the TBWCS group was significantly lower than that in the TBWKW group ($P < 0.05$). (Table 2).

3.3. Operation Related Indicators. There was no statistical difference between the two groups in terms of operation time, incision length, and fracture gap after operation. The intraoperative blood loss, hospitalization time, and angle of knee flexion loss in patients with TBWCS were less than those in

patients with TBWKW ($P < 0.05$). The starting time of the postoperative functional exercise of the TBWCS group was earlier than those of the TBWKW group, and the incidence of secondary operation was lower than those in patients with TBWKW ($P < 0.05$) (Table 3).

3.4. Pain Score. We evaluated the changes of VAS scores of the two groups of patients within 1 year after the operation, but we did not continue to follow up on the VAS scores because those of the two groups of patients were less than 1 point 6 months after the operation. The follow-up results showed that the VAS scores of the two groups decreased gradually with the time ($P < 0.05$). The VAS scores of the patients in the TBWCS group were lower than those in the TBWKW group at 1, 3, and 6 months after operation ($P < 0.05$) (Figure 1).

3.5. Knee Flexion Degree. At 1, 3, 6, and 12 months after operation, knee flexion degree of the two groups of patients gradually increased with the time ($P < 0.05$), and the knee flexion of patients in TBWCS group was higher than that of patients in TBWKW ($P < 0.05$) (Figure 2).

3.6. Lysholm Score. At 1, 3, 6, and 12 months after operation, the Lysholm scores of the two groups of patients gradually increased with the time ($P < 0.05$). The Lysholm scores of the patients in the TBWCS group were higher than those in the TBWKW group ($P < 0.05$) (Figure 3).

3.7. Bostman Score. At 1, 3, 6, and 12 months after the operation, the Bostman score of the two groups of patients gradually increased with the time ($P < 0.05$), while the Bostman score of the TBWCS group was higher than that of the TBWKW group ($P < 0.05$) (Figure 4).

3.8. Comparison of Complications between Two Groups of Patients. There was no significant difference in the incidence of knee joint movement limitation, traumatic arthritis, bursitis, displaced internal fixation, reduction loss, and delayed fracture healing between the two groups, but the total

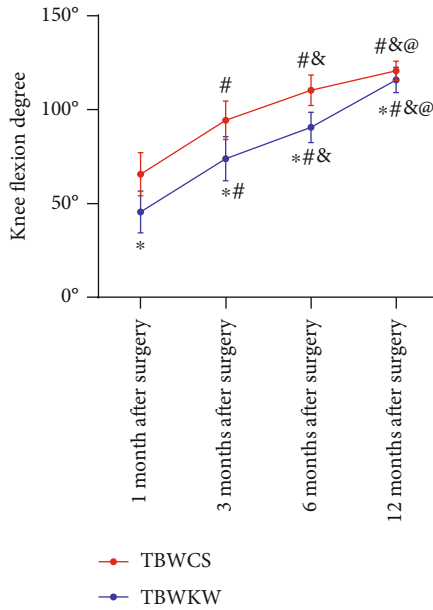


FIGURE 2: Difference of knee flexion between the two groups within 12 months after the operation. Repeated measurement analysis of variance showed that the knee flexion of the two groups of patients gradually increased with time. The knee flexion of patients in the TBWCS group ($n = 71$) was higher than that of patients in the TBWKW group ($n = 75$). * indicates compared with the TBWCS group, $P < 0.05$. # indicates compared with 1 month after treatment in the same group, $P < 0.05$; & indicates compared with 3 months after treatment in the same group, $P < 0.05$; @ indicates compared with 4 months after treatment in the same group, $P < 0.05$.

incidence of complications in the TBWCS group was lower than that in the TBWKW group ($P < 0.05$) (Table 4).

3.9. *Quality of Life Assessment.* There was no statistical difference in ADL scores between the two groups before the operation. Twelve months after operation, the ADL scores of patients in the TBWCS group were higher than those in the TBWKW group ($P < 0.05$) (Figure 5).

4. Discussion

Surgical is a common method for the treatment of displaced patella fracture, of which tension band fixation is the current treatment standard [14]. However, due to the complexity of patella fracture, there is no consistent conclusion on the best clinical treatment scheme at present. This study compared the therapeutic effects of cannulated screw tension band and Kirschner wire tension band on patella fracture and found that the cannulated screw tension band had more advantages in treatment of patella fracture, with a fast recovery of patients and low incidence of complications.

Cannulated screw tension band is an improved technique based on the Kirschner wire tension band. Theoretically, the cannulated screw tension band has the advantages of stable fixation and implant protection [15]. In a biomechanical analysis on the treatment of patella fracture with wire tension band by Lee et al. [16], a cannulated screw tension band has higher load-carrying capacity and rigidity, and can absorb

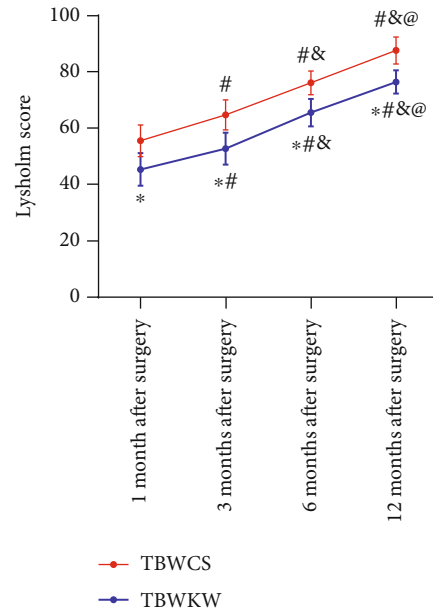


FIGURE 3: Difference of Lysholm score between the two groups within 12 months after the operation. Repeated measurement analysis of variance showed that the Lysholm scores of the two groups of patients gradually increased with time. The Lysholm scores of patients in the TBWCS group ($n = 71$) were higher than those in the TBWKW group ($n = 75$). * indicates compared with the TBWCS group, $P < 0.05$. # indicates compared with 1 month after treatment in the same group, $P < 0.05$; & indicates compared with 3 months after treatment in the same group, $P < 0.05$; @ indicates compared with 4 months after treatment in the same group, $P < 0.05$.

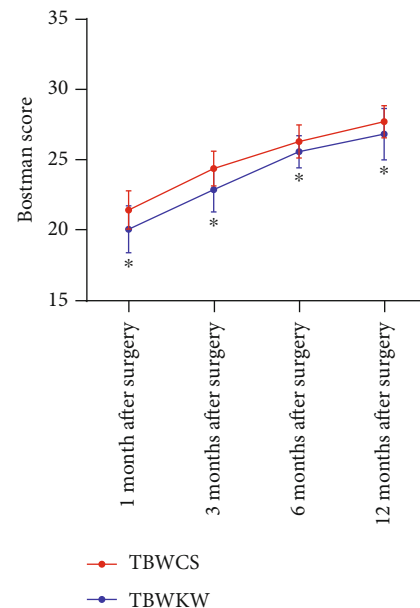


FIGURE 4: Bostman score difference between the two groups within 12 months after the operation. Repeated measurement analysis of variance showed that the Bostman score of the two groups of patients gradually increased with time. The Bostman score of patients in the TBWCS group ($n = 71$) was higher than that of patients in the TBWKW group ($n = 75$). * indicates compared with the TBWCS group, $P < 0.05$.

TABLE 4: Comparison of complications of two groups of patients within 6 months after operation (N , %).

	TBWCS ($n = 71$)	TBWKW ($n = 75$)	χ^2	P
Limited knee joint movement	2 (2.82)	5 (6.67)	0.491	0.484
Traumatic arthritis	0 (0.00)	2 (2.67)	Fisher	0.497
Bursitis	0 (0.00)	1 (1.33)	Fisher	0.999
Displaced internal fixation	1 (1.41)	2 (2.67)	0.002	0.962
Reduction loss	2 (2.82)	4 (5.33)	0.122	0.728
Delayed fracture healing	1 (1.41)	3 (4.00)	0.204	0.652
Total	6 (8.45)	17 (22.67)	4.534	0.033

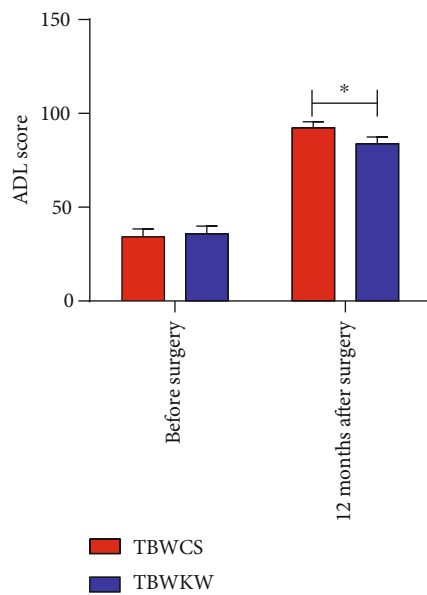


FIGURE 5: Quality of life assessment of two groups of patients. Repeated measurement analysis of variance showed that ADL scores of patients in the TBWCS group ($n = 71$) were higher than those in the TBWKW group ($n = 75$) 12 months after the operation. * indicates $P < 0.05$.

higher energy. However, Wang et al. [17] showed that there was no significant difference between cannulated screw and Kirschner wire tension band in improving Iowa knee joint score of patella fracture patients. Lin et al. [18] also found that after 12 months of treatment, there was no significant difference between cannulated screws and Kirschner wire tension band in improvement of VAS score, knee joint mobility, flexion, and extension of patella fracture patients. Hoshino et al. [19] also reported that the failure rate of patella fracture fixation with a cannulated screw tension band was higher than that with the Kirschner wire tension band. The results of this study show that during the operation of patella fracture patients, the intraoperative blood loss, hospitalization time, and knee flexion loss of patients in TBWCS group were less than those in the TBWKW group, the starting time of postoperative functional exercise was earlier than that of patients in the TBWKW group, and the incidence rate of secondary operation was lower than that of patients in the

TBWKW group, but there was no statistical difference in the operation time, incision length, and postoperative fracture gap between the two groups. The results of curative effect analysis showed that the knee flexion, Lysholm score, and Bostman score of patients treated with tension band with cannulated screw were higher than those treated with Kirschner wire, and VAS score was lower. Tension band with cannulated screw had a better curative effect on patella fracture, lower complication rate, and higher quality of life of patients. A meta-analysis report showed that there was no difference in the success rate of operation, operation time, fracture healing time, and the number of infections between the cannulated screw and Kirschner wire tension band in treating patella fracture, but cannulated screw tension band was superior to Kirschner wire tension band in reducing the incidence of complications [20]. In the internal fixation of patella fracture, biodegradable implants were not as effective as metal implants in the treatment of displaced patella fracture, but implant stimulation was the main reason for the second operation, and the removal rate of symptomatic implants after treatment with cannulated screw tension band was low (8%) [21]. Tan et al. [22] also reported in the study that the cannulated screw tension band had a better curative effect in the treatment of patella fracture compared with the Kirschner wire tension band, and it reduced the occurrence of pain caused by implants and implant loosening. These studies all supported our conclusion.

However, some problems need to be paid attention to when using a cannulated screw tension band to treat patella fracture. The cannulated screw placed in the operation needs to be of appropriate size and can be completely embedded into bone. The head and tail of the screw do not penetrate through the upper and lower ends of the patella, thus ensuring the action of the tension band of steel wire and reducing the friction loss between the screw and steel wire, steel wire, and patella.

To sum up, the cannulated screw tension band has a better curative effect on patella fracture, low incidence of complications, early start of postoperative functional exercise, and higher quality of life.

Data Availability

All the data can be provided if other researchers need.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] J. L. Prudhon, J. H. Caton, T. Aslanian, and R. Verdier, "How is patella height modified after total knee arthroplasty?," *International Orthopaedics*, vol. 42, no. 2, pp. 311–316, 2018.
- [2] P. T. Simonian, T. L. Simonian, and L. E. Simonian, "Percutaneous tension-band suture technique for distal patella fracture fixation," *MOJ Orthopedics & Rheumatology*, vol. 8, no. 3, pp. 315–318, 2017.
- [3] N. A. Bonazza, G. S. Lewis, E. Z. Lukosius, E. P. Roush, K. P. Black, and A. Dhawan, "Effect of transosseous tunnels on

- patella fracture risk after medial patellofemoral ligament reconstruction: a cadaveric study,” *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 34, no. 2, pp. 513–518, 2018.
- [4] G. K. H. Shea, K. Hoi-Ting So, K. W. Tam, D. K. H. Yee, C. Fang, and F. Leung, “Comparing 3 different techniques of patella fracture fixation and their complications,” *Geriatric Orthopaedic Surgery & Rehabilitation*, vol. 10, 2019.
- [5] Y. Sun, K. Sheng, Q. Li, D. Wang, and D. Zhou, “Management of comminuted patellar fracture fixation using modified cerclage wiring,” *Journal of Orthopaedic Surgery and Research*, vol. 14, no. 1, p. 324, 2019.
- [6] D. S. S. Borkar, D. S. S. Konde, D. S. Thosar, and D. R. Patil, “Minimally invasive technique of tension band wiring in patella fractures,” *International Journal of Orthopaedics Sciences*, vol. 4, no. 2k, pp. 729–731, 2018.
- [7] R. Kansay, A. Malhotra, D. Singh, S. Gupta, A. Soni, and G. Chander, “Migrated K wire into popliteal fossa from tension band wiring of patella: a case report,” *International Journal of Scientific Research*, vol. 8, no. 1, 2019.
- [8] I. Zderic, K. Stoffel, C. Sommer, D. Höntzsch, and B. Gueorguiev, “Biomechanical evaluation of the tension band wiring principle. A comparison between two different techniques for transverse patella fracture fixation,” *Injury*, vol. 48, no. 8, pp. 1749–1757, 2017.
- [9] W. Ju, D. Sun, and B. Qi, “Novel method of Kirschner wire bending for treatment of displaced patella fractures,” *International Journal of Clinical and Experimental Medicine*, vol. 11, no. 3, pp. 1955–1958, 2018.
- [10] M. Ling, S. Zhan, D. Jiang, H. Hu, and C. Zhang, “Where should Kirschner wires be placed when fixing patella fracture with modified tension-band wiring? A finite element analysis,” *Journal of Orthopaedic Surgery and Research*, vol. 14, no. 1, p. 14, 2019.
- [11] D. Franks, J. Shatrov, M. Symes, D. G. Little, and T. L. Cheng, “Cannulated screw versus Kirschner-wire fixation for Milch II lateral condyle fractures in a paediatric sawbone model: a biomechanical comparison,” *Journal of Children’s Orthopaedics*, vol. 12, no. 1, pp. 29–35, 2018.
- [12] H. Y. Choi, S. J. Hyun, K. J. Kim, T. A. Jahng, and H. J. Kim, “Freehand S2 alar-iliac screw placement using K-wire and cannulated screw: technical case series,” *Journal of Korean Neurosurgical Society*, vol. 61, no. 1, pp. 75–80, 2018.
- [13] M. Nienhaus, I. Zderic, D. Wahl, B. Gueorguiev, and P. M. Rommens, “A locked intraosseous nail for transverse patellar fractures: a biomechanical comparison with tension band wiring through cannulated screws,” *JBJS*, vol. 100, no. 12, article e83, 2018.
- [14] B. Matthews, K. Hazratwala, and S. Barroso-Rosa, “Comminuted patella fracture in elderly patients: a systematic review and case report,” *Geriatric Orthopaedic Surgery & Rehabilitation*, vol. 8, no. 3, pp. 135–144, 2017.
- [15] R. K. Alluri, J. R. Hill, P. Navo, A. Ghiassi, M. Stevanovic, and A. Mostofi, “Washer and post augmentation of 90/90 wiring for proximal interphalangeal joint arthrodesis: a biomechanical study,” *The Journal of Hand Surgery*, vol. 43, no. 12, pp. 1137.e1–1137.e10, 2018.
- [16] K. H. Lee, Y. Lee, Y. H. Lee, B. W. Cho, M. B. Kim, and G. H. Baek, “Biomechanical comparison of three tension band wiring techniques for transverse fracture of patella: Kirschner wires, cannulated screws, and ring pins,” *Journal of Orthopaedic Surgery*, vol. 27, no. 3, 2019.
- [17] C. Wang, L. Tan, B. C. Qi et al., “A retrospective comparison of the modified tension band technique and the parallel titanium cannulated lag screw technique in transverse patella fracture,” *Chinese Journal of Traumatology*, vol. 17, no. 4, pp. 208–213, 2014.
- [18] T. Lin, J. Liu, B. Xiao, D. Fu, and S. Yang, “Comparison of the outcomes of cannulated screws vs. modified tension band wiring fixation techniques in the management of mildly displaced patellar fractures,” *BMC Musculoskeletal Disorders*, vol. 16, no. 1, p. 282, 2015.
- [19] C. M. Hoshino, W. Tran, J. V. Tiberi III et al., “Complications following tension-band fixation of patellar fractures with cannulated screws compared with Kirschner wires,” *JBJS*, vol. 95, no. 7, pp. 653–659, 2013.
- [20] Y. Zhang, Z. Xu, W. Zhong, F. Liu, and J. Tang, “Efficacy of K-wire tension band fixation compared with other alternatives for patella fractures: a meta-analysis,” *Journal of Orthopaedic Surgery and Research*, vol. 13, no. 1, p. 226, 2018.
- [21] G. Busel, B. Barrick, D. Auston et al., “Patella fractures treated with cannulated lag screws and fiberwire® have a high union rate and low rate of implant removal,” *Injury*, vol. 51, no. 2, pp. 473–477, 2020.
- [22] H. Tan, P. Dai, and Y. Yuan, “Clinical results of treatment using a modified K-wire tension band versus a cannulated screw tension band in transverse patella fractures,” *Medicine*, vol. 95, no. 40, article e4992, 2016.

Research Article

Application of Deep Learning for Early Screening of Colorectal Precancerous Lesions under White Light Endoscopy

Junbo Gao ¹, Yuanhao Guo ¹, Yingxue Sun ¹ and Guoqiang Qu ²

¹Information Engineering College, Shanghai Maritime University, Shanghai 201306, China

²Department of Gastroenterology, Eastern Hospital, Shanghai Sixth People Hospital, Shanghai 201306, China

Correspondence should be addressed to Junbo Gao; jbgao@shmtu.edu.cn

Received 29 May 2020; Accepted 30 June 2020; Published 25 August 2020

Guest Editor: Lin Lu

Copyright © 2020 Junbo Gao 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.

Background and Objective. Colorectal cancer (CRC) is a common gastrointestinal tumour with high morbidity and mortality. Endoscopic examination is an effective method for early detection of digestive system tumours. However, due to various reasons, missed diagnoses and misdiagnoses are common occurrences. Our goal is to use deep learning methods to establish colorectal lesion detection, positioning, and classification models based on white light endoscopic images and to design a computer-aided diagnosis (CAD) system to help physicians reduce the rate of missed diagnosis and improve the accuracy of the detection rate. **Methods.** We collected and sorted out the white light endoscopic images of some patients undergoing colonoscopy. The convolutional neural network model is used to detect whether the image contains lesions: CRC, colorectal adenoma (CRA), and colorectal polyps. The accuracy, sensitivity, and specificity rates are used as indicators to evaluate the model. Then, the instance segmentation model is used to locate and classify the lesions on the images containing lesions, and mAP (mean average precision), AP_{50} , and AP_{75} are used to evaluate the performance of an instance segmentation model. **Results.** In the process of detecting whether the image contains lesions, we compared ResNet50 with the other four models, that is, AlexNet, VGG19, ResNet18, and GoogLeNet. The result is that ResNet50 performs better than several other models. It scored an accuracy of 93.0%, a sensitivity of 94.3%, and a specificity of 90.6%. In the process of localization and classification of the lesion in images containing lesions by Mask R-CNN, its mAP, AP_{50} , and AP_{75} were 0.676, 0.903, and 0.833, respectively. **Conclusion.** We developed and compared five models for the detection of lesions in white light endoscopic images. ResNet50 showed the optimal performance, and Mask R-CNN model could be used to locate and classify lesions in images containing lesions.

1. Introduction

Colorectal cancer (CRC) is a common malignancy of the digestive system. According to the latest data, the morbidity and mortality of CRC rank among the top four in cancer [1]. With the improvement of Chinese people's living standards, the incidence of CRC increases year by year. The most common precancer disease of CRC is colorectal adenoma (CRA) [2]. At present, total colonoscopy is still the best screening method for colorectal polyps, CRA, and CRC [3]. Early detection of CRA and endoscopic resection of adenoma under colonoscopy can reduce or avoid the occurrence of CRC, thereby reducing the mortality rate of CRC [4, 5]. Early diagnosis of digestive system tumours has always been a hot spot for the medical community to conquer. However, it is

difficult to detect early precancerous lesions of the digestive system because they generally involve a small range and are shallow in depth, and the morphological manifestations under endoscopy are not obvious [6]. Moreover, the evaluation results of endoscopy often depend on the subjective experience of the operating physician, which is highly subjective and requires a high level of clinical skills and work experience of the physician. The low-qualified or fatigued physician is more likely to misdiagnose the lesion [7].

At present, the application of artificial intelligence (AI) in the medical field has shown an exciting dawn, and its exploration in the field of digestive endoscopy has also achieved some preliminary results [8]. In the study conducted by [9], two different shape description features were compared to distinguish whether there are polyps in the colorectal region.

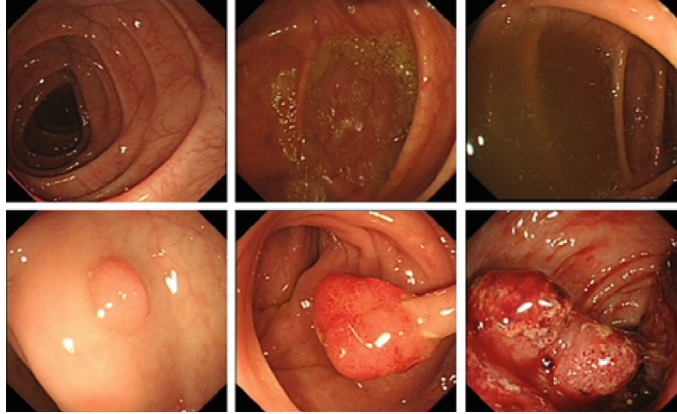


FIGURE 1: Colorectal polyps in different stages of neoplasia and different grades of bowel cleanliness under white light endoscopy.

The algorithm was tested on 300 images, among which 150 images contain polyps and 150 images of normal mucosa, with an accuracy of 86%. In the study [10], the author used the fitting ellipse method and multiscale Gaussian texture geometric features, and the true positive of the algorithm was 64.8%. Reference [11] combined the advantages of wavelet transform and local uniform binary mode to characterize image features and used support vector machine (SVM) as the classifier. Their data set contained a total of 1200 images (600 polyp images and 600 nonpolyp tissue images), and the algorithm accuracy reached 91.6%.

With the emergence of deep learning algorithm, machine learning has gradually gotten rid of the limitation of low efficiency and imprecision in manually extracting data features, which has brought revolutionary progress to the research and development of artificial intelligence. Zhang et al. [12] used white light endoscopic images to distinguish polyps from adenoma, with an accuracy rate of 85.9%. Patino-Barrientos et al. [13] used the VGG16 model to perform Kudo's Classification for Colon Polyps on white light endoscopic images. The accuracy rate was 83%, and the F1 score was 0.83. Ruikai et al. [14] used the YOLOv3 algorithm to detect and locate polyps on white light endoscopes. The accuracy of the results was 88.6%, and the recall rate was 71.6%.

In summary of the above studies, it can be concluded that the application of convolutional neural network of deep learning in the detection, location, and classification of colorectal polyps is feasible and has achieved good results. Moreover, due to the late development of endoscopic medical technology in China, the overall technical level lags behind that of developed countries, so most of the endoscopic techniques used in hospitals in China are still dominated by ordinary white light endoscopes. In the daily diagnosis, the endoscopist must thoroughly examine each image of each patient; the process is very cumbersome. Therefore, the development of a computer-aided diagnosis system based on white light endoscopy can greatly reduce the burden on medical personnel and is of great significance.

The contribution of our study is to establish a model of detection, localization, and classification of colorectal lesions based on white light endoscopy. Compared with other models based on the research of white light endoscopy, the

models of this study have improved to some extent in some evaluation indicators of experimental results. In the process of detecting whether white light endoscopic pictures contain lesions, we compared five convolutional neural network models. As a result, the ResNet50 model showed higher detection performance; it scored an accuracy of 93.0%, a sensitivity of 94.3%, and a specificity of 90.6%. In the process of locating and classifying lesions, the Mask R-CNN model was used to segment the images, and a satisfactory result was obtained; its mAP, AP_{50} , and AP_{75} were 0.676, 0.903, and 0.833, respectively.

2. Materials and Method

2.1. Data Set. In this study, images of patients undergoing colorectal examination under white light endoscopy (as shown in Figure 1) were used.

These images were derived from the Digestive Endoscopy Centre, East Hospital, Shanghai Sixth People's Hospital, China. The time span is from June 2015 to September 2019. We created a database containing 3413 WLE images of dimensions $420 \times 389 \times 3$ (RGB), of which 1709 of them contained lesions (CRC, CRA, and polyps) and 1704 of them are normal colorectal mucosa. Images with clear surface and boundary under the white light endoscope and complete film were selected, and corresponding microscope pathology was recorded, which was marked and classified by trained endoscopy physicians and gastroenterologists. Images of normal mucosa have varying degrees of cleanliness and air bubbles. All images containing lesions have been checked to ensure the accuracy of the data set. We randomly divide the image into a training set, a validation set, and a test set (respectively, 70%, 15%, and 15% of the full data set), and limit the balance between the validation set and the test set.

2.2. Study Design. In this study, we will follow the following process (Figure 2) to detect, locate, and classify white light endoscopic colorectal lesions.

First of all, we will use the convolutional neural network (CNN) [15] model to distinguish whether the white light endoscopic images contain lesions (CRC, CRA, and polyps) and whether output images also contain lesions. Next, an

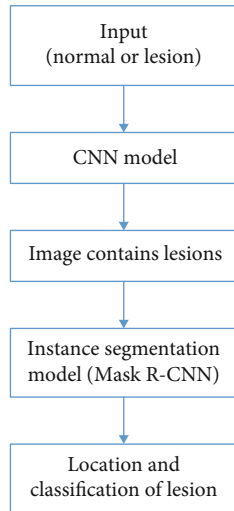


FIGURE 2: Colorectal lesion detection localization and classification process.

instance segmentation [16] model Mask R-CNN is used on the image containing a lesion to locate the position of the lesion, and a prediction of the corresponding category of the lesion is given.

2.3. CNN Architecture. Convolutional neural networks (CNN) are a kind of Feedforward Neural Network that contains convolutional computation and has a deep structure and is one of the representative algorithms of deep learning [17, 18]. LeCun and his collaborators constructed the convolutional neural network LeNet-5 and achieved success in the recognition of handwritten digits [19]. LeNet-5 and its subsequent variants define the basic structure of modern convolutional neural networks. The alternating convolutional layer and pooling layer in its construction are considered to be able to extract the higher-order features of the input image.

We evaluated five network architectures: AlexNet [20], GoogLeNet [21], ResNet50 [22], ResNet18 [23], and VGG19 [24]. These networks all use a hierarchical structure, with the output of the previous layer as the input of the next layer, continuously extracting and building the higher-order features of the input picture. Because our white light endoscopy image data set is not far enough to support from scratch or train the network, we will use transfer learning to use the above-mentioned pretrained CNN convolutional layer as a feature extractor. These pretrained networks use a large amount of image data (including 1000 categories) for training, and on ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) good results have been achieved, so these networks have the ability to classify various images. According to the use of the feature extraction scheme and transfer learning to solve the new classification problem, the last layer or the last three layers of CNN must be fine-tuned. In this study, we changed the final classification layer to output two categories, namely, normal images and images containing lesions.

For the binary classification, the success of classification of WLE images using CNN is measured by accuracy,

TABLE 1: Performance of different networks on test dataset.

Network	Accuracy (%)	Sensitivity (%)	Specificity (%)
AlexNet	85.5	78.9	92.2
VGG19	89.5	85.9	93.0
ResNet18	87.9	84.4	91.4
ResNet50	93.0	90.6	95.6
GoogLeNet	87.9	82.8	93.0

sensitivity, and specificity, which are widely employed by colleagues to assess the performance of classification. Here is their definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of positives} + \text{number of negatives}}, \quad (1)$$

$$\text{Sensitivity} = \frac{\text{Number of correct positive predictions}}{\text{Number of positives}}, \quad (2)$$

$$\text{Specificity} = \frac{\text{Number of correct negative predictions}}{\text{Number of negatives}}. \quad (3)$$

In this study, we observed that pretrained ResNet50 usually performs better than the other four networks (Table 1). Therefore, we will introduce this network architecture in detail. ResNet50's network consists of 50 layers, including 17 layers (16 convolutional layers and 1 fully connected layer) of learnable weights. Each convolutional layer contains 64 to 2048 kernels of size 1×1 and 3×3 . In order to enhance the robustness of the internal deformation of the class and avoid overfitting, the convolution kernel with a size of 1×1 before the shortcut uses the Rectified Linear Unit (ReLU). ReLU, as the activation function of the neural network, defines the non-linear output of the neuron after linear transformation; here is the definition:

$$f(x) = \max(0, w^T x + b). \quad (4)$$

In order to improve network performance, we have further optimized the network architecture. And the stochastic gradient descent with momentum (SGDM) is used as the optimization algorithm. The learning rate was initially set to $1e-4$ and was adaptively modified during the training process until the verification criteria were met. The maximum epoch size of the training process is 50. In order to standardize the model and reduce overfitting, image data enhancement is used in the model training process, including rotation, cropping, and mirror conversion. These data enhancement operations do not affect the content or size of the image. Finally, in order to overcome the generalization gap while taking into account the limited GPU memory, a mini batch size of 16 was chosen.

2.4. Mask R-CNN Network Architecture. Mask R-CNN was extended from Faster R-CNN [25]. Faster R-CNN is a popular target detection framework, which was extended to the instance segmentation framework by Mask R-CNN. Mask R-CNN [26] is a two-stage framework. The first stage scans

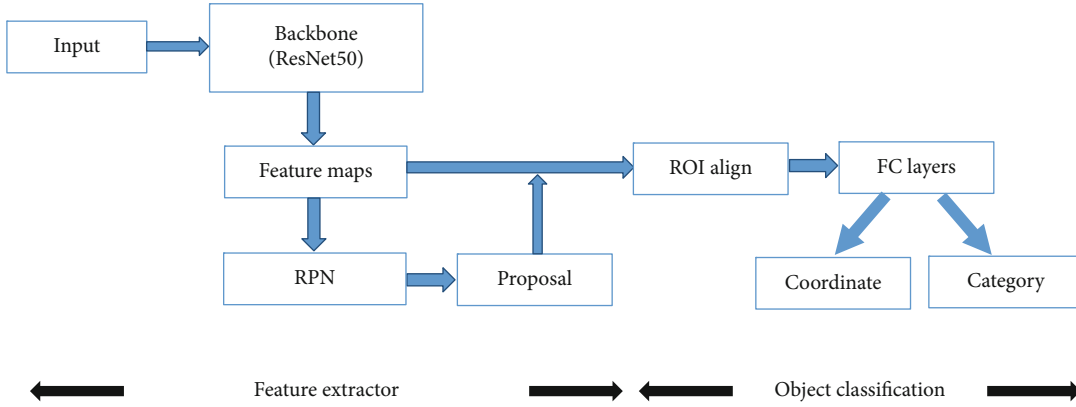


FIGURE 3: Mask R-CNN model training process.

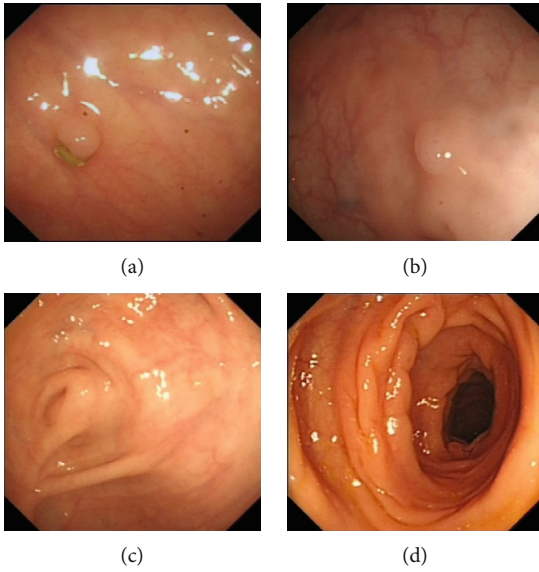


FIGURE 4: Misclassification examples: (a, b) false negatives; (c, d) false positives.

the images and relies on the Region Proposal Network (RPN) algorithm [27] to generate proposals (region of interest, or ROI), and the second stage classifies the proposals and generates bounding boxes and masks. Mask R-CNN can effectively detect the target in the image, add a parallel branch for border box recognition, and generate a high-quality segmentation mask for each instance. So it was called Mask R-CNN.

Mask R-CNN was trained in the following steps (Figure 3):

Step 1. Enter an image you want to process and then carry out the corresponding pre-processing operation or the pre-processed picture.

Step 2. Input it into a pretrained neural network (ResNe50, etc.) to obtain the corresponding feature map.

Step 3. Set a predetermined ROI for each point in the feature map, so as to obtain multiple candidate ROI.

Step 4. Send these candidate ROIs into the RPN network for binary classification (foreground or background) and BB regression and filter out some candidate ROIs.

Step 5. Perform ROI align operation on the remaining ROIs, that is, match the original image with the pixel of the feature map first and then match the feature map with the fixed feature.

Step 6. Perform operations on each ROI in Fully Convolutional Networks (FCN) for classification, bounding-box regression, and mask generation.

To develop our Mask R-CNN, we selected 1709 images containing lesions (CRC, CRA, and polyps) as the image database. Next, manually label these images to mark the location and correspondence of the lesions in the picture category, which is a tedious labelling task, and finally, we created an image database in MSCOCO format for lesion location and classification. Each image contains a bounding box around the large intestine lesion in the format of $[x, y, \text{width}, \text{height}]$, which specifies the lesion's position and size in the upper left corner of the image. We further divided the images into 70% for training, 15% for validation, and 15% (256 images) for testing the Mask R-CNN network.

In the Mask R-CNN developed in this study, ResNet50 was used as the backbone network. The minimum batch is set to 1, so that each iteration processes multiple image areas from one training image. Each image is controlled by two different parameters, positive training samples and negative training samples. These two values are set to overlap with ground truth boxes by a factor of $[0.6-1.0]$ and $[0-0.3]$, respectively. Considering the bounding box as a true positive box containing lesions, we chose a threshold of 0.7 for the IoU measure, which is a good threshold for calculating the "intersection" of various bounding boxes.

IoU (Intersection-over-Union) represents the overlap rate between the generated candidate bound and the ground truth bound, that is, the ratio of their intersection and union.

$$\text{IoU} = \frac{\text{area}(C) \cap \text{area}(G)}{\text{area}(C) \cup \text{area}(G)}. \quad (5)$$

TABLE 2: Performance of Mask R-CNN model on test dataset.

Network	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AR	AR ₁₀	AR ₁₀₀	AR _S	AR _M	AR _L
Mask R-CNN	67.6	90.3	83.3	100	65.1	64.8	75.4	78.2	78.2	100	79.9	76.5

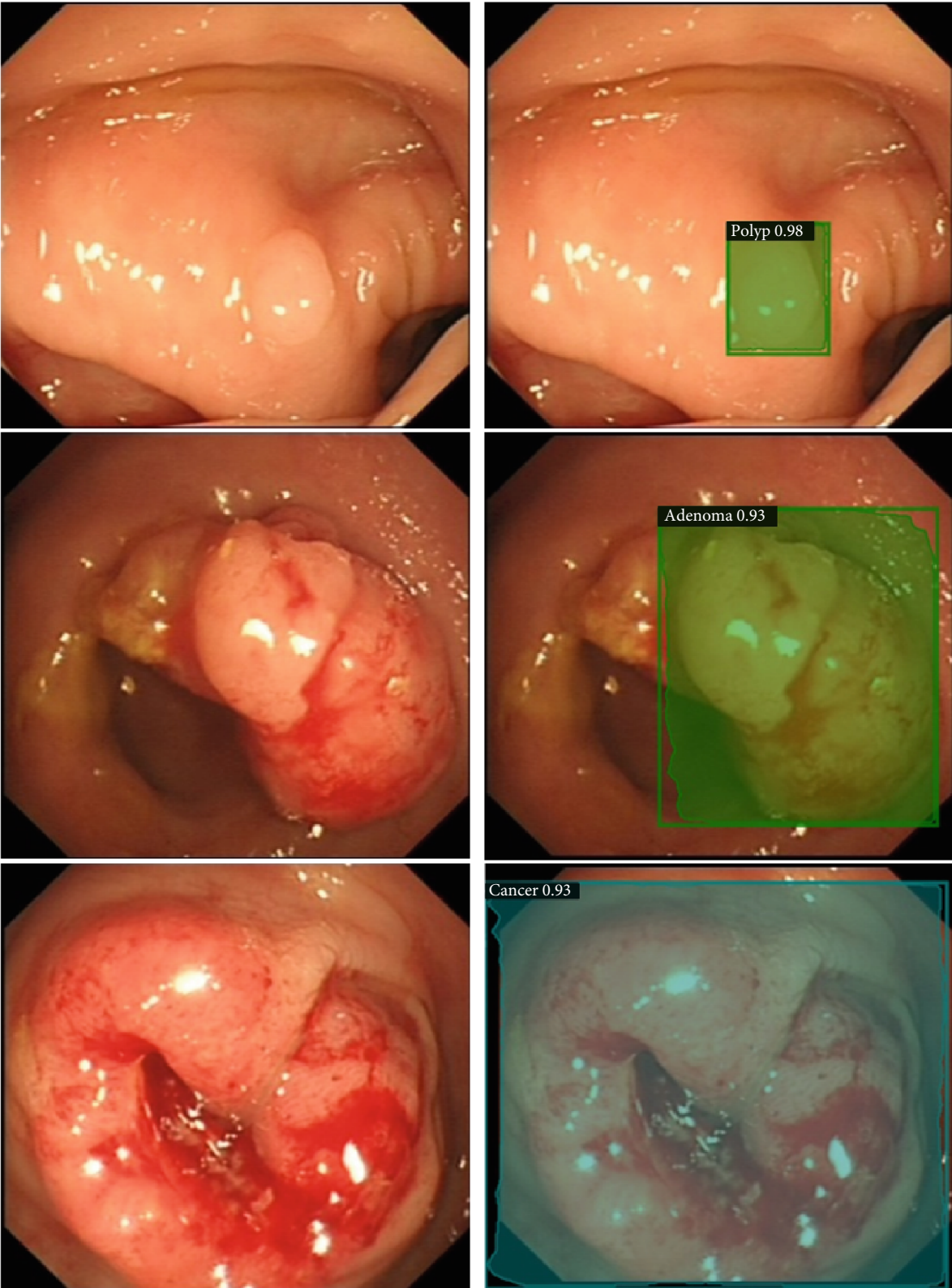


FIGURE 5: Colorectal lesion localization and classification using Mask R-CNN.

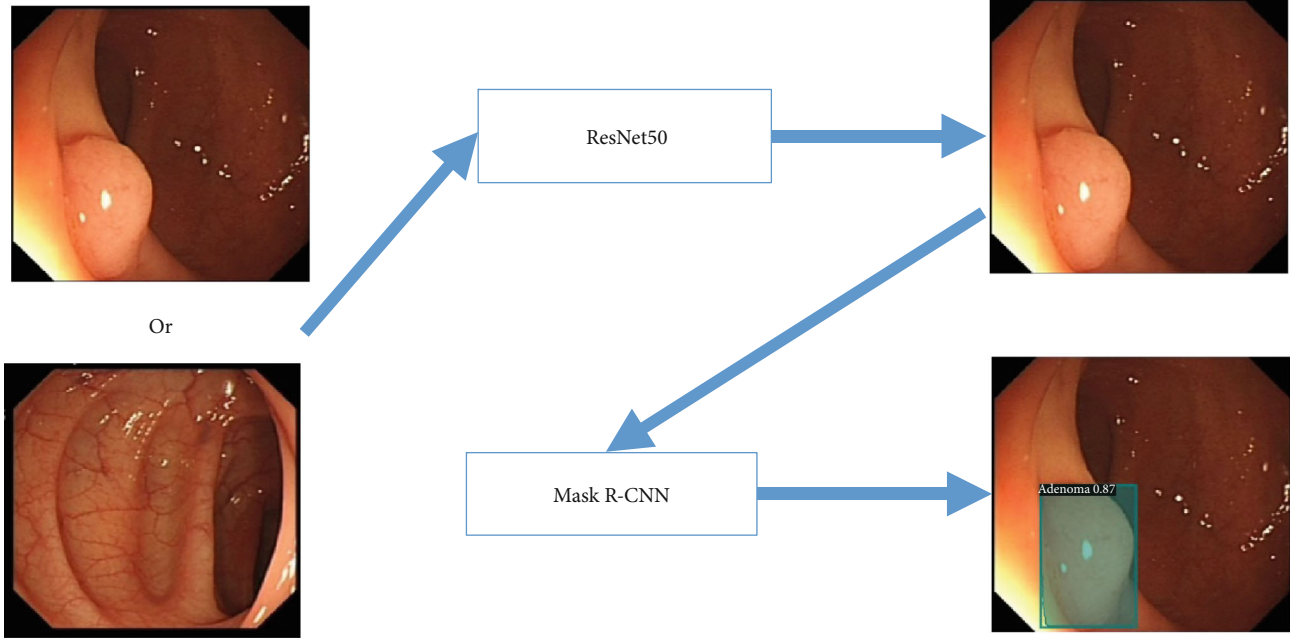


FIGURE 6: Our computer-aided diagnosis system workflow.

In this study, we used some outcome indicators that evaluated the MSCOCO data set to evaluate our model, such as mAP (mean average precision), AP_{50} , and AP_{75} , as the main evaluation criteria for the results.

For the binary classification, the sample can be divided into four cases of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Precision predicts the correct value in the case of predicting positive samples; recall predicts the correct value in instances with positive labels.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (7)$$

$$\text{mAP} = \int_0^1 P(R) dR. \quad (8)$$

In formula (8), P represents precision and R represents recall. mAP is the average AP value of multiple verification set individuals. So the higher the mAP score is, the higher the confidence in the test results will be and the more likely polyps will be contained in the boundary box.

3. Results and Discussion

3.1. Performance of CNN Model. Table 1 lists the performance of different networks after transfer learning in the process of detecting whether white light endoscopic pictures contain lesions, mainly reflected in the accuracy, sensitivity, and specificity of the network. We found that the modified ResNet50 performed significantly better than the other four networks.

The reason why high sensitivity is very important is because the consequences of false negatives (missed polyps) are much more serious than false positives (misdiagnosed as polyps). Figure 4 shows some examples of misclassified images. Missed polyps (false negatives) and normal mucosa are misidentified as polyps (false positives).

We have observed a situation that easily leads to missed diagnosis of polyps, that is, the size of polyps is small. Usually in endoscopic detection, some smaller polyps are also easily missed [28], but these polyps are less likely to form tumours at advanced stages, and our models are often correct when detecting large polyps, so to a certain extent, the consequences of missed diagnosis of small polyps are reduced.

3.2. Performance of Mask R-CNN Model. Since the data set we marked is in MSCOCO format, a series of result metrics of the MSCOCO data set will be used to evaluate our data set. Here is their definition:

mean average precision (AP):

AP: AP at IoU = 0.50 : 0.05 : 0.95 (primary challenge metric),

AP_{50} : AP at IoU = 0.50,

AP_{75} : AP at IoU = 0.75,

AP_S : AP for small objects: area < 32²,

AP_M : AP for medium objects: 32² < area < 96²,

AP_L : AP for large objects: area > 96².

Average recall (AR):

AR: AR given 1 detection per image,

AR_{10} : AR given 10 detections per image,

AR_{100} : AR given 100 detections per image,

AR_S : AR for small objects: area < 32²,

AR_M : AR for medium objects: 32² < area < 96²,

AR_L : AR for large objects: area > 96².

The results are shown in Table 2.

An example of the Mask R-CNN we developed to locate and classify lesions is presented in Figure 5.

4. Conclusions

Due to the late development of endoscopic medical technology in China, the overall technical level lags behind that of developed countries; ordinary white light endoscopy is still the endoscopic technique used in most hospitals in China. The endoscopist needs to examine each image of each patient carefully. The process is cumbersome. Therefore, the development of a computer-aided diagnosis system based on a white light endoscope is of great significance. Compared with other models based on white light endoscopic research, the model developed in this study has improved the effect and can assist the endoscopist in daily diagnosis, which will greatly reduce the daily burden of medical staff.

In this study, we used white light endoscopic images as a screening tool for colorectal lesions. Through comprehensive evaluation of supervised machine learning algorithms, we find that different algorithms have different prediction performances on image data. By comparing the predictive performance of the classifier, we found that ResNet50 is a good model.

In addition, we also annotated images containing lesions, constructed a data set in MSCOCO format, and used the instance segmentation algorithm Mask R-CNN to perform experiments on this data set. Through comparison and analysis of some result indicators, we found that location and classification of colorectal lesions have achieved a good result.

We hope to develop a computer-aided diagnosis system; the process is shown in Figure 6.

Through the combination of the two models, a white light endoscopic image is inputted, the ResNet50 model is used to determine whether it contains a lesion, and the image containing the lesion is then input into the Mask R-CNN model to locate and classify the lesion.

Future work includes exploring architectures such as capsule networks and attention model, which may be difficult to implement but can provide more specific interpretability. Our goal is to develop a more accurate, real-time colorectal lesion detection, localization, and classification model and deploy it on a system where white light endoscopy works to help physicians better perform colonoscopy for patients.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by the Shanghai University of Medicine and Health Sciences Seed Foundation (SFP-18-22-14-006) in China.

References

- [1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *Ca A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018.
- [2] G. Soltani, A. Poursheikhani, M. Yassi, A. Hayatbakhsh, M. Kerachian, and M. A. Kerachian, "Obesity, diabetes and the risk of colorectal adenoma and cancer," *BMC Endocrine Disorders*, vol. 19, no. 1, p. 113, 2019.
- [3] X. Wu, X. He, S. Li, X. Xu, X. Chen, and H. Zhu, "Long non-coding RNA ucoo2kmd.1 regulates CD44-dependent cell growth by competing for miR-211-3p in colorectal cancer," *PLoS ONE*, vol. 11, no. 3, 2016.
- [4] M. Løberg, M. Kalager, Ø. Holme, G. Hoff, H.-O. Adami, and M. Bretthauer, "Long-term colorectal-cancer mortality after adenoma removal," *New England Journal of Medicine*, vol. 371, no. 9, pp. 799–807, 2014.
- [5] D. A. Corley, C. D. Jensen, A. R. Marks et al., "Adenoma detection rate and risk of colorectal cancer and death," *New England Journal of Medicine*, vol. 370, no. 14, pp. 1298–1306, 2014.
- [6] X. Jie, L. I. Peng, and Z. Shutian, "The importance of endoscopic diagnosis and treatment in early gastrointestinal cancer," *Chinese Journal of Colorectal Diseases*, vol. 3, no. 6, 2014.
- [7] T. D. Lange, P. Halvorsen, and M. Riegler, "Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy," *World Journal of Gastroenterology*, vol. 24, no. 45, pp. 5057–5062, 2018.
- [8] K. Togashi, "Applications of artificial intelligence to endoscopy practice: the view from Japan Digestive Disease Week 2018," *Digestive Endoscopy*, vol. 31, no. 3, pp. 270–272, 2019.
- [9] B. Li, Y. Fan, M. Q. H. Meng, and L. Qi, "Intestinal polyp recognition in capsule endoscopy images using color and shape features," in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1490–1494, IEEE, 2009.
- [10] J. Jia, S. Sun, T. Terrence, and P. Wang, "Accurate and efficient polyp detection in wireless capsule endoscopy images," *US Patent*, vol. 14, no. 471, 2014.
- [11] B. Li and Q. H. Meng, "Automatic polyp detection for wireless capsule endoscopy images," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10952–10958, 2012.
- [12] R. Zhang, Y. Zheng, T. W. C. Mak et al., "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain," *IEEE Journal of Biomedical & Health Informatics*, vol. 21, no. 1, pp. 41–47, 2017.
- [13] S. Patino-Barrientos, D. Sierra-Sosa, B. Garcia-Zapirain, C. Castillo-Olea, and A. Elmaghraby, "Kudo's classification for colon polyps assessment using a deep learning approach," *Applied Sciences*, vol. 10, no. 2, p. 501, 2020.
- [14] R. Zhang, Y. Zheng, C. C. Poon, D. Shen, and J. Y. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognition*, vol. 83, pp. 209–219, 2018.
- [15] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object

- detection,” *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, 2020.
- [16] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, “Object instance segmentation and fine-grained localization using hypercolumns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 627–639, 2017.
- [17] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [18] J. Gu, Z. Wang, J. Kuen et al., *Recent advances in convolutional neural networks*, Computer Science, 2015.
- [19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [20] Y. Fu and C. Aldrich, “Froth image analysis by use of transfer learning and convolutional neural networks,” *Minerals Engineering*, vol. 115, pp. 68–78, 2018.
- [21] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, 2015.
- [22] N. Ali, E. Quansah, K. Köhler et al., “Automatic label-free detection of breast cancer using nonlinear multimodal imaging and the convolutional neural network ResNet50,” *Translational Biophotonics*, vol. 1, no. 1-2, 2019.
- [23] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, “Learning structured sparsity in deep neural networks,” in *Advances in neural information processing systems*, pp. 2074–2082, 2016.
- [24] D. Xia, P. Chen, B. Wang, J. Zhang, and C. Xie, “Insect detection and classification based on an improved convolutional neural network,” *Sensors*, vol. 18, no. 12, p. 4169, 2018.
- [25] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [26] Y. Liu, J. Liu, H. Pu, Y. Liu, and S. Song, “Instance segmentation of outdoor sports ground from high spatial resolution remote sensing imagery using the improved mask R-CNN,” *International Journal of Geoences*, vol. 10, no. 10, pp. 884–905, 2019.
- [27] Y. P. Chen, Y. Li, and G. Wang, “An enhanced region proposal network for object detection using deep learning method,” *Plos One*, vol. 13, no. 9, 2018.
- [28] A. Murino, C. Hassan, and A. Repici, “The diminutive colon polyp,” *Current Opinion in Gastroenterology*, vol. 32, no. 1, pp. 38–43, 2016.

Research Article

Functional Modular Network Identifies the Key Genes of Preoperative Inhalation Anesthesia and Intravenous Anesthesia in Off-Pump Coronary Artery Bypass Grafting

Hongfei Zhao,¹ Weitian Wang,¹ Liping Liu,² Junlong Wang,¹ and Quanzhang Yan ¹

¹Department of Anesthesiology, Weifang People's Hospital, Weifang 261000, China

²Department of Neurology, Weifang People's Hospital, Weifang 261000, China

Correspondence should be addressed to Quanzhang Yan; yannan555766@163.com

Received 27 April 2020; Revised 12 June 2020; Accepted 15 June 2020; Published 17 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Hongfei Zhao 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.

Off-pump coronary artery bypass grafting (OPCABG) is an effective strategy for revascularization. Preoperative anesthesia appears critical due to surgical instability and the risk of organ damage. This study, based on a functional module network, analysed the effects of preoperative inhalation anesthesia and intravenous anesthesia on OPCABG and performed a pivot analysis of its potential drug regulators. We obtained microarray data of sevoflurane anesthesia and propofol anesthesia from the GEO database and analysed the difference between the two groups of data, resulting in 5701 and 3210 differential genes to construct the expression matrix. WGCNA analysis showed that sevoflurane anesthesia clustered into 7 functional disorder modules, including *PDCD6IP*, *WDR3*, and other core genes; propofol anesthesia clustered to form two functional disorder modules, including *KCNB2* and *LHX2*, two core genes. Enrichment analysis of the functions and pathways of interest suggests that both anesthesia-related module genes tend to function as pathways associated with ion and transmembrane transport. The underlying mechanism may be that targeted regulation of transmembrane-associated biological processes and ion pathways in the core genes of each module affect the surgical process. Pivot analysis of potential drug regulators revealed 229 potential drugs for sevoflurane anesthesia surgery, among which zinc regulates three functional disorder modules via *AHSG*, *F12*, etc., and 67 potential drugs for propofol anesthesia surgery, among which are propofol, methadone, and buprenorphine, regulate two functional disorder modules through four genes, *CYP2C8*, *OPRM1*, *CYP2C18*, and *CYP2C19*. This study provides guidance on clinical use or treatment by comparing the effects of two anesthetics on surgery and its potential drugs.

1. Introduction

Currently, off-pump coronary artery bypass grafting (OPCABG) is an innovative technique in cardiac surgery. In recent years, an aging population, increased risk of surgery, and improved technology contribute to the resurgence of OPCABG [1]. OPCABG can reduce postoperative complications such as systemic inflammatory response, myocardial damage, kidney damage, and brain damage [2]. Currently, it is the best choice for modern cardiac surgery. Mortality and stroke rate of the elderly after surgery are extremely low, indicating that surgery is the safe management option for coronary artery disease in this population [3]. However, during the operation, the patient needs to be anesthetized. Thus,

choosing which kind of anesthesia is very important. Common methods are inhalation anesthesia, intravenous anesthesia, sevoflurane anesthesia, and propofol anesthesia. Compared with propofol-based total intravenous anesthesia (TIVA), sevoflurane anesthesia reduced cardiac biomarker release and hospital stay. It also could reduce mortality compared with CFG [4]. This may be due to the fact that sevoflurane can better protect the heart muscle during cardiac surgery [5]. The induction characteristics of sevoflurane anesthesia in congenital heart disease are similar. Sevoflurane induced good tolerance technology, suitable for children with congenital heart disease [6]. Although sevoflurane anesthesia has numerous benefits, when sevoflurane is paused to use after surgery, rhythmic heart separation returns to sinus

rhythm. Sometimes, it might cause atrioventricular conduction disturbances, leading to rhythmic arrhythmias [7]. Another common anesthetic is propofol. It can be quickly induced and rapidly eliminated with short duration of action, smooth recovery of anesthesia, and few side effects. So it has been widely used. It is more effective and harmless than hypnotic drug [8]. Propofol is also considered as the best anesthetic alternative in experiments comparing the recovery periods of the two anesthesia regimens [9]. Propofol is also reported to be approved for continuous intravenous sedation. Surgical clinical studies have revealed that the combination of propofol and opioids is a reasonable anesthetic option [10]. OPCABG surgery is associated with lymphopenia. Propofol anesthesia with protective effects is superior to sevoflurane maintenance anesthesia [11]. When it comes to the difference between the two anesthesia methods, the researchers have many different opinions. Of course, some people thought that no difference exists in myocardial protection after sevoflurane anesthesia or propofol anesthesia in OPCABG surgery [12]. Although propofol-induced anesthesia can attenuate the feedback pathway of cardiac baroreflex and the feedforward pathway can be immune to anesthesia [13], we still have no conclusion about the use of the two anesthetic methods. To further investigate the difference between the two anesthetic methods, we conducted a study on molecular mechanisms of its regulation.

Here, we analysed the effects of postoperative inhalation anesthesia and intravenous anesthesia on off-pump coronary artery bypass grafting based on a functional modular network, to explore the underlying molecular mechanisms.

2. Materials and Methods

2.1. Differential Expression Analysis. We collected an expression microarray dataset for inhaled anesthesia and intravenous anesthesia prior to coronary artery bypass grafting from the NCBI Gene Expression Omnibus (GEO) database [14], numbered GSE4386, and performed variance analysis in the collected disease samples (containing interleukin 23; no interleukin 23) using the R language limma package [15]. With threshold $P < 0.05$, significantly differentially expressed genes were obtained. In the end, a total of 6699 differential genes were obtained, including 2212 common genes, 3489 specific differential genes for sevoflurane anesthesia, and 998 differentially expressed genes for propofol anesthesia. We used two sets of differential genes to construct an expression profile matrix for nonexternal coronary artery bypass grafting.

2.2. Coexpression Analysis. WGCNA [16] on two sets of differential expression profiles, respectively, clusters similar or identical genes to form a module, also known as a functional disorder module, in order to explore synergistically express relationships of differential genes in the two groups. Since the two sets of data in this study are subject to scale-free networks, correlation coefficients can be used for cluster analysis. Firstly, the correlation coefficient between the genes is taken to the N^{th} power by the correlation coefficient weighting, and the Person Coefficient between the genes is obtained.

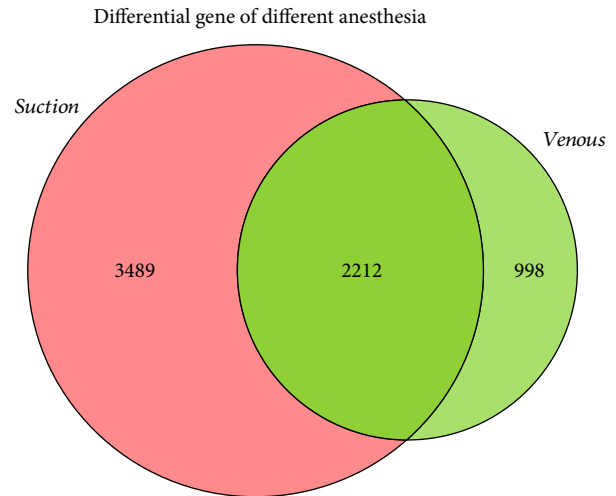


FIGURE 1: Venn diagram: Red represents the differential gene of sevoflurane anesthesia surgery, green represents the differential gene of propofol anesthesia, and the middle overlap is the difference gene shared by both.

Then, the results of the Person Coefficient are clustered to obtain the clustering tree. Various branches of the cluster tree means various functional barrier modules while diverse colours represent diverse modules, and the genes of the same branch have strong correlations. There are numerous regulatory genes in each module, and we have extracted these genes with relatively large regulatory powers as key genes leading to the dysfunction of functional modules. Seven key genes were obtained for sevoflurane anesthesia-related modules, namely PDCD6IP, DNAH10, WDR3, PROP1, ASCL2, LRRC2-AS1, and SDC3. In addition, the two key genes of the module related to propofol anesthesia are KCNB2 and LHX2, respectively. Therefore, we believe that these core genes are involved in the molecular regulation of anesthesia for nonexternal coronary artery bypass grafting.

2.3. Analysis of Functional and Pathway Enrichment. Exploring the function and signalling pathways involving genes is often an effective means of studying the molecular mechanisms of disease. We performed the GO function and KEGG pathway enrichment analysis for the differentially expressed genes of seven functional disorders related to sevoflurane anesthesia and two functional disorder modules related to propofol anesthesia. The enrichment analysis of the GO function and the KEGG pathway uses the clusterProfiler package in the R language [17, 18], with threshold $pvalueCutoff = 0.05$. Through the perspective of data, we filter out the functions and paths of interest that interact with multiple modules and draw bubble maps based on the count values that act between the modules. This study screened 15 interesting functions and pathways, and used ggplot2 in R language to draw bubble maps for display. The size of the bubble in the display represents the count value of the function and the path. The larger the value of the count, the more obvious the potential effect, while the colour refers to its LogFC value.

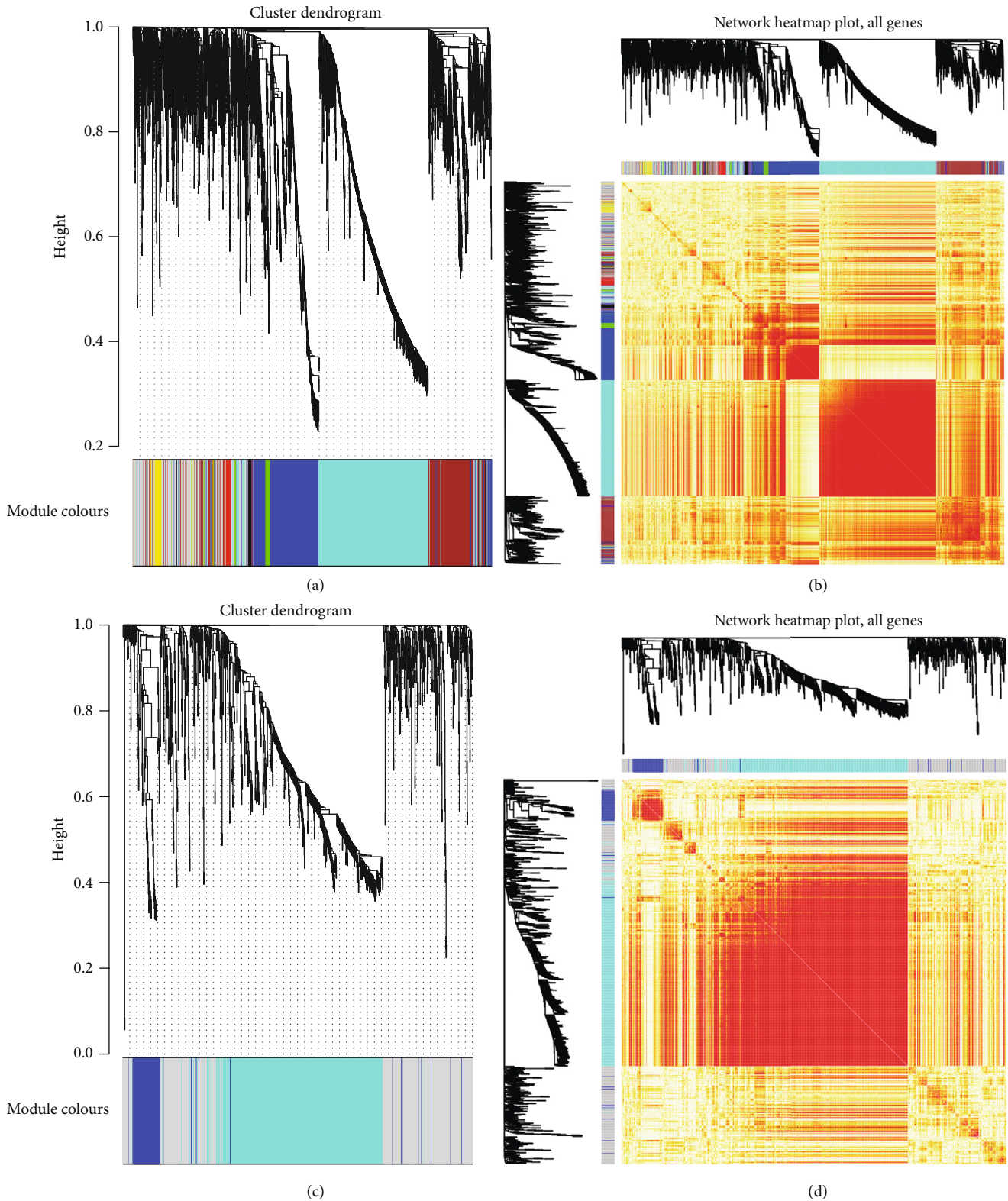


FIGURE 2: Synergistic expression of preoperative anesthesia OPCAB. (a) Synergistic expression of sevoflurane anesthesia for OPCABG, 7 coexpression groups obtained by clustering were identified as modules, and 7 colours represented 7 coexpression modules. (b) Here are heatmaps of all genes expression in the sample, whose expression behaviour is clustered into 7 coexpression modules. (c) Coexpression of propofol anesthesia for OPCABG, two coexpression groups obtained by clustering were identified as modules, and two colours represented two coexpression modules. (d) Propofol anesthetizes the expression heatmap of all genes in the sample, and its expression behaviour is clustered into two coexpression modules.

2.4. Identification of Drug Regulation of Modular Genes. To further enhance the impact of the two anesthetics on surgery, we performed a pharmacomodulator predictive analysis of the functional modules. Comparing the effects of two different anesthesia methods on the operation and the mechanism of action leaves guidance for clinical application. A pivotal analysis of the effects of the drug supplements given during anesthesia on the surgery will provide a better understanding of the effects of the drug on anesthesia. We calculated the enrichment target significance in each module based on the hypergeometric test ($pvalueCutoff = 0.01$, $LogFC = 0.5$) and obtained the drug related to the module. The data was input into Cytoscape for a module-drug network diagram. 229 drugs with potential effects on sevoflurane anesthesia were obtained. We selected related drugs and modules for network regulation. In addition, there are 67 drugs with potential effects on the operation of propofol anesthesia. We have a network diagram of potential interactions between drugs in the module for the interaction between these 67 modules and drugs.

2.5. Potential Role of Drug Target Genes. After analysing the interaction between the module and the drug, we do not know about the regulated target gene, so we need to analyse its target gene. This study was based on the pivot analysis of drugs and modules and review of the target gene for drug action in the DrugBank database, with threshold $P < 0.05$. We interpreted drugs with potential effect on sevoflurane anesthesia and propofol anesthesia as well as the relationship list of their target genes to Cytoscape and got regulatory network diagram of Module_Drug_TargetGene. When constructing the Module_Drug_TargetGene regulatory network map for the interaction of sevoflurane anesthesia surgery, we selected the zinc-related interaction relationship mapping.

3. Results

3.1. Time Series Expression of Dysregulated Molecules for Postoperative Anesthesia for Off-Pump Coronary Artery Bypass Grafting. First, we constructed the gene expression profiles of OPCABG under two anesthesia and analysed the differences in order to further understand the effect of anesthesia on off-pump coronary artery bypass grafting. Resulting in 5701 differential genes from sevoflurane anesthesia CABG surgery and 3210 differential genes from propofol anesthesia CABG surgery. The Venn map of the two groups of differential genes revealed 6699 differential genes comprising 2212 shared genes, 3489 specific for sevoflurane anesthesia, and 998 specific for propofol anesthesia (Figure 1).

3.2. Identification of Functional Module Networks. To further explain the effect of two anesthesia methods on off-pump coronary artery bypass grafting, based on the WGCNA analysis, we constructed 7 functional disorder modules using sevoflurane anesthesia-specific differential genes (Figures 2(a) and 2(b)). Two functional disorder modules were constructed by specific differential genes of propofol anesthesia (Figures 2(c) and 2(d)). In addition, we identified the module's hub genes by regulation of genes within

TABLE 1: Key genes of sevoflurane anesthesia related modules.

Colour	Hub genes	Module
Black	PDCD6IP	m6
Blue	DNAH10	m1
Brown	WDR3	m2
Green	PROP1	m5
Red	ASCL2	m7
Turquoise	LRRC2-AS1	m3
Yellow	SDC3	m4

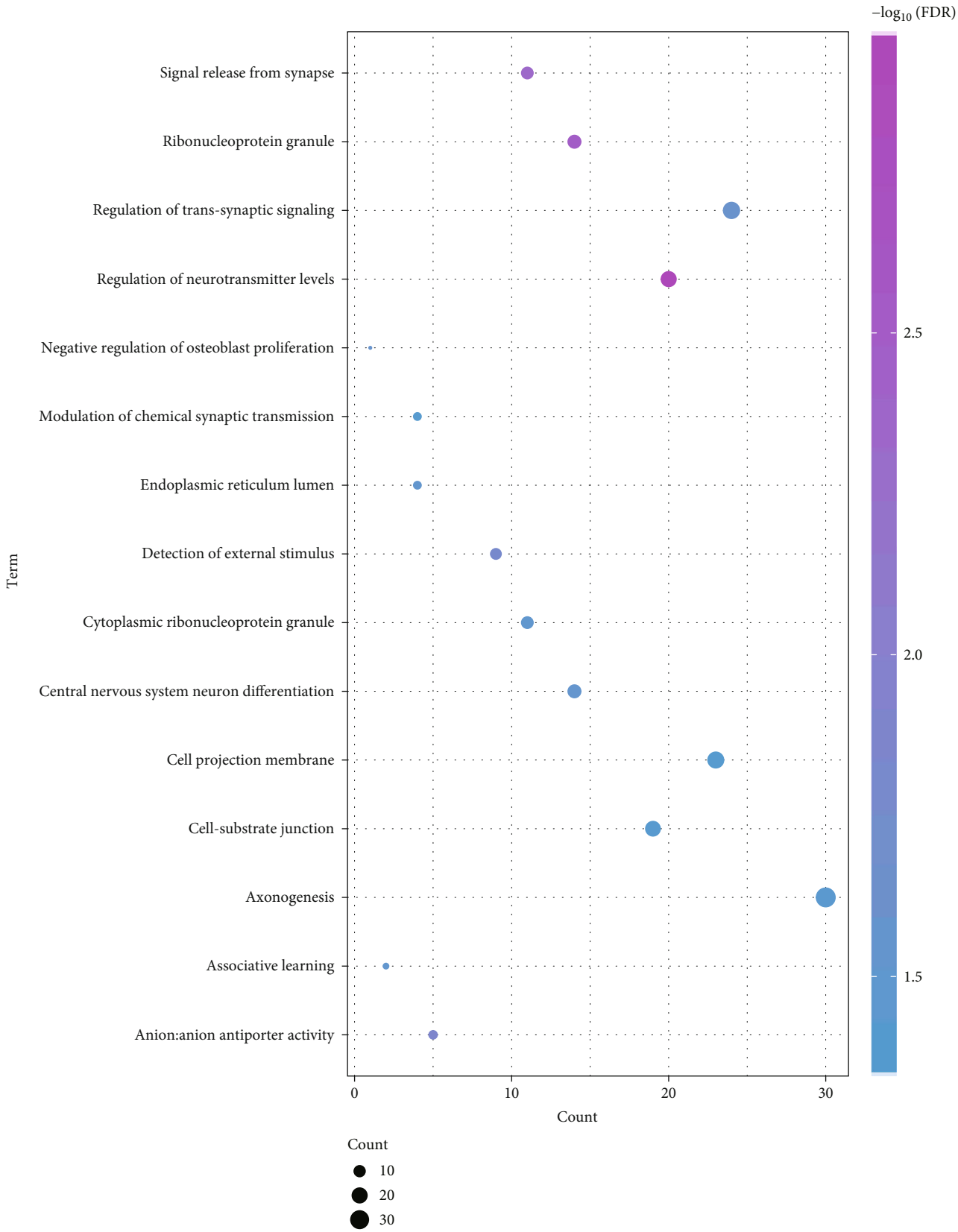
TABLE 2: Key genes of propofol anesthesia related modules.

Colour	Hub genes	Module
Blue	KCNB2	m2
Turquoise	LHX2	m1

the functional disorder module (Tables 1 and 2). Then, we analysed the various effect of its key genes in two anesthesia situations that participated in different functions and pathways on OPCABG.

3.3. Functions and Pathways Involved in the Gene of Interest. In order to further understand the biological characteristics of the functional impairment module, we performed the GO function and KEGG pathway enrichment analysis for the two functional module networks. The sevoflurane anesthesia-related module gene was involved in 371 cell composition entries, 680 molecular function terms, 2567 biological processes, and 128 signal pathways (Figures 3(a) and 3(b)). In light of functional analysis, we observed that related functional modules favour various biological process-related functions, including axonogenesis, anion antiporter activity, regulation of neurotransmitter levels, and PI3K-Akt signalling pathway. Besides, propofol anesthesia-related module genes involved 84 cell component entries, 179 molecular functional terms, 648 biological processes, and 15 signalling pathways (Schedule 2-2, Figures 3(c) and 3(d)). We observed that the relevant functional modules are mainly involved in ion channel and transmembrane transport, such as metal ion transmembrane transporter activity, channel activity, passive transmembrane transporter activity, and neuroactive ligand-receptor interaction. These signalling pathways have been shown to be associated with the development and progression of OPCABG under anesthesia.

3.4. Drugs with Potential Effects on OPCABG under Anesthesia. A pharmacomodulator predictive analysis of the functional modular genes was carried out for the impact of the two anesthetics on surgery. According to the number of regulatory modules with P value < 0.05 , 229 drugs with potential effects on sevoflurane anesthesia were obtained (Schedule 3-1, Figure 4(a)), in which zinc significantly participated in the regulation of three modules. In addition, 67 drugs with potential effects on propofol anesthesia surgery (Schedule 3-2, Figure 4(b)), among which are propofol,



(a) FIGURE 3: Continued.

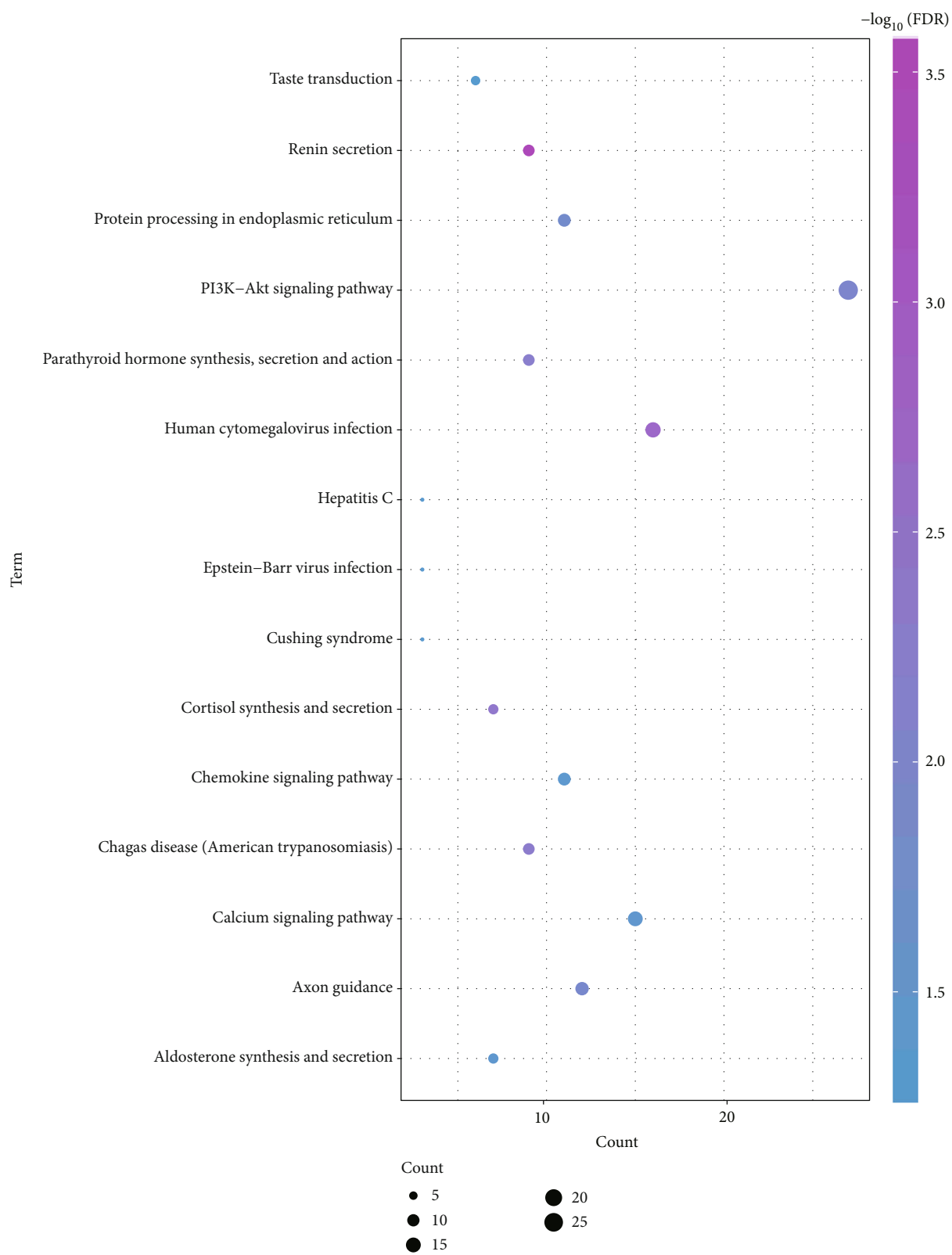
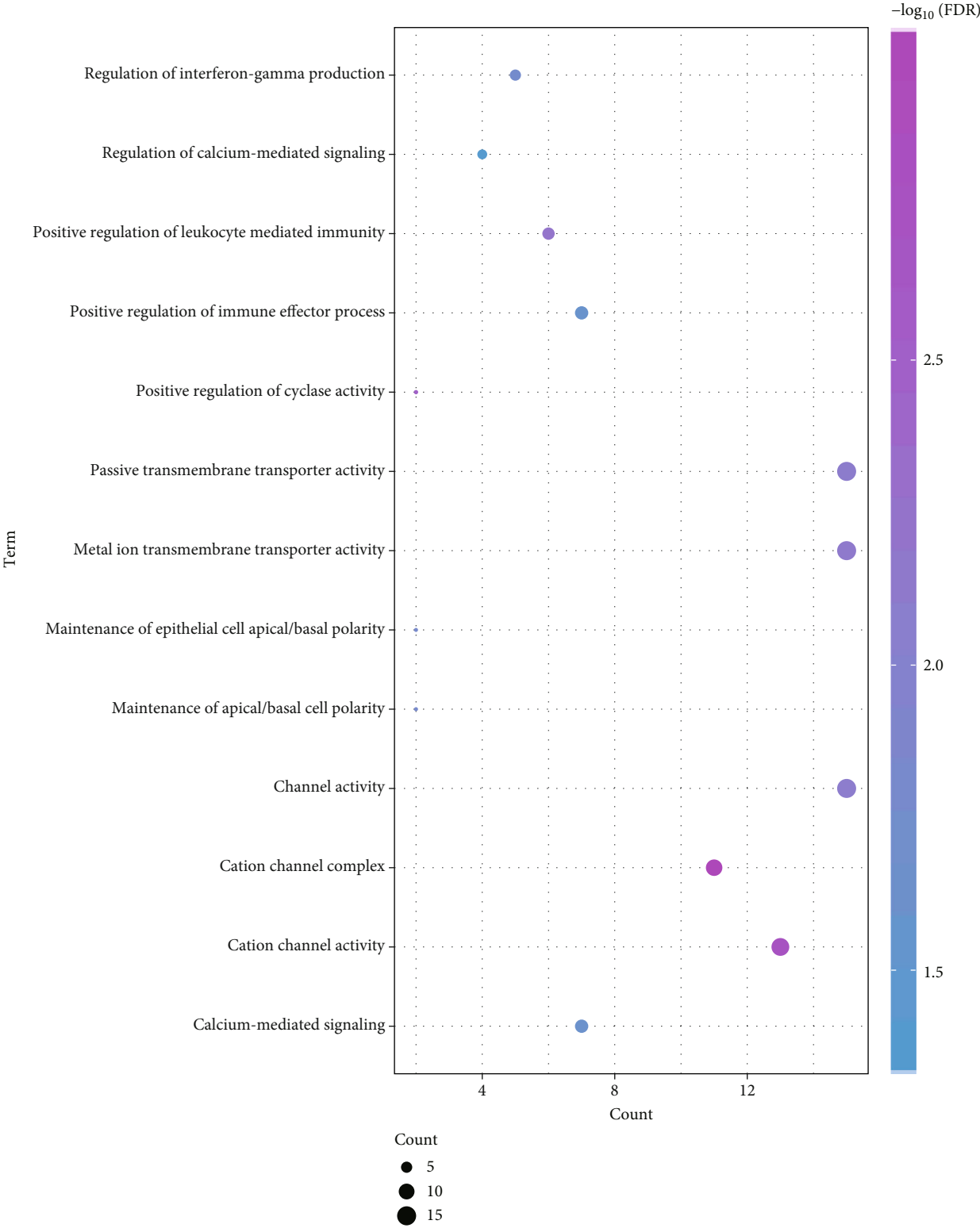


FIGURE 3: Continued.



(c)

FIGURE 3: Continued.

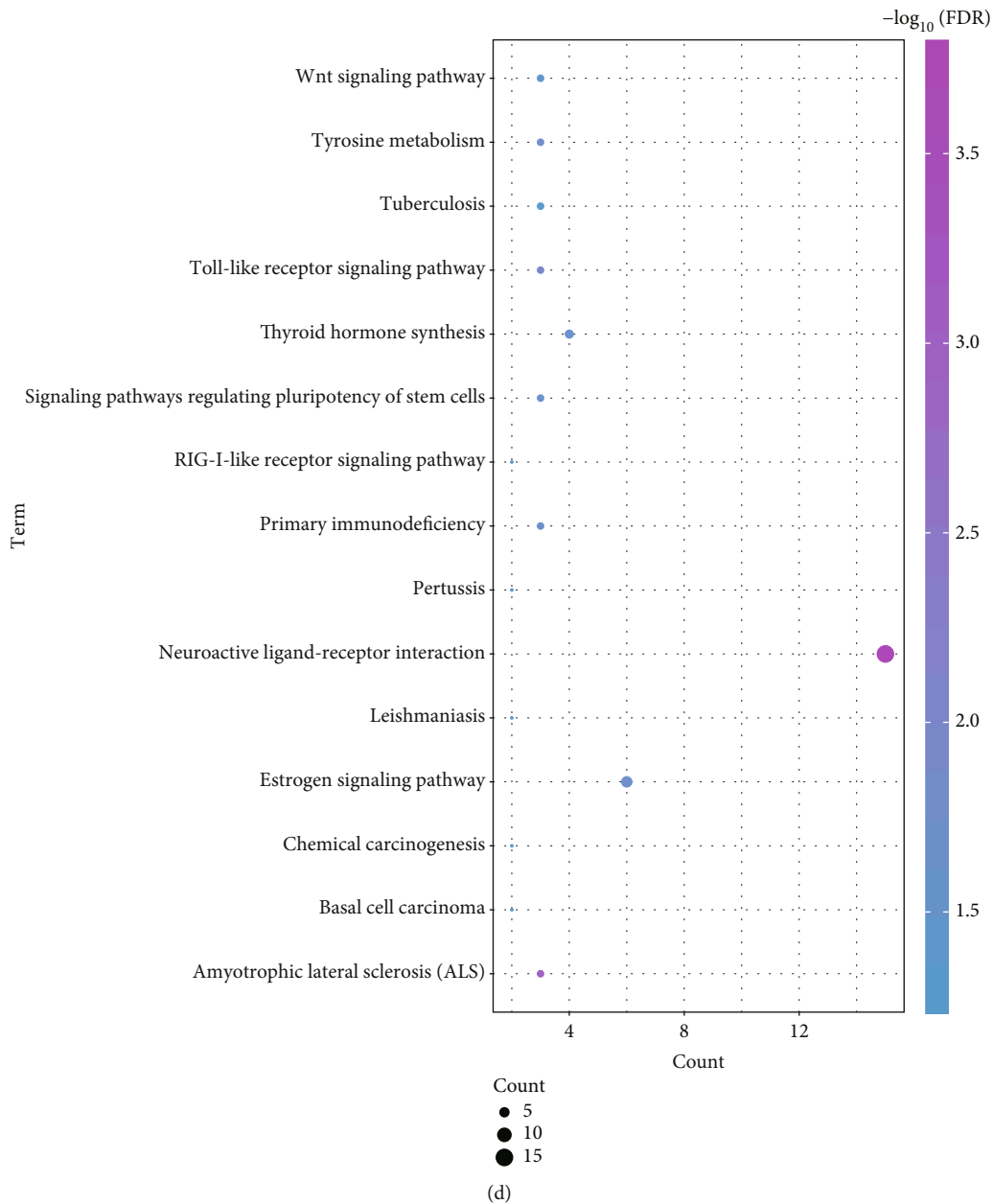


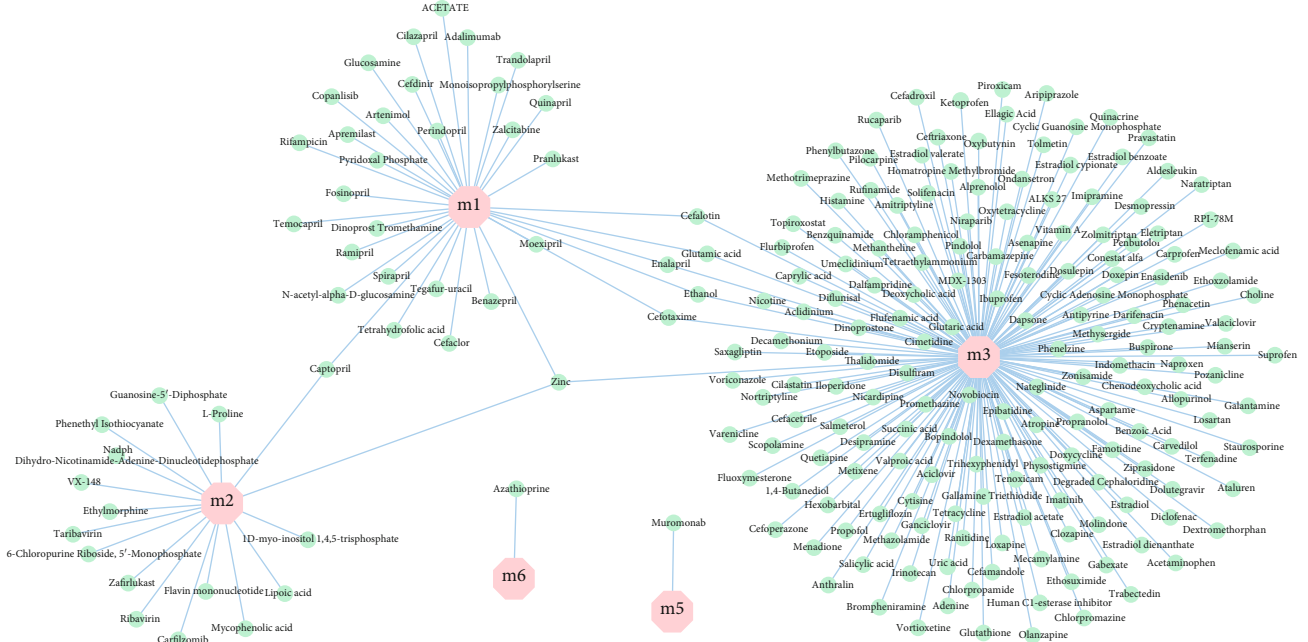
FIGURE 3: Functional and pathway enrichment analysis excerpts of preoperative anesthesia for OPCABG. (a) Sevoflurane anesthesia for excision analysis of GO gene functional enrichment of OPCAB. From blue to purple, the enrichment increases significantly. The larger the circle, the greater the proportion of the module gene in the GO function entry gene. (b) Sevoflurane anesthesia for excision analysis of the KEGG gene functional enrichment of OPCABG. From blue to purple, the enrichment increases significantly. The larger the circle, the greater the proportion of the module gene in KEGG function entry gene. (c) Propofol anesthesia for excision analysis of the GO gene functional enrichment of OPCABG. From blue to purple, the enrichment increases significantly. The larger the circle, the greater the proportion of the module gene in GO function entry gene. (d) Propofol anesthesia for OPCABG modular gene KEGG pathway enrichment analysis excerpt. From blue to purple, the enrichment increases significantly. The larger the circle, the greater the proportion of the module gene in KEGG function entry gene.

methadone, and buprenorphine, significantly involved in the regulation of the two modules. Most drugs can have certain auxiliary or obstructive effects on anesthesia with certain impact on the surgical procedure.

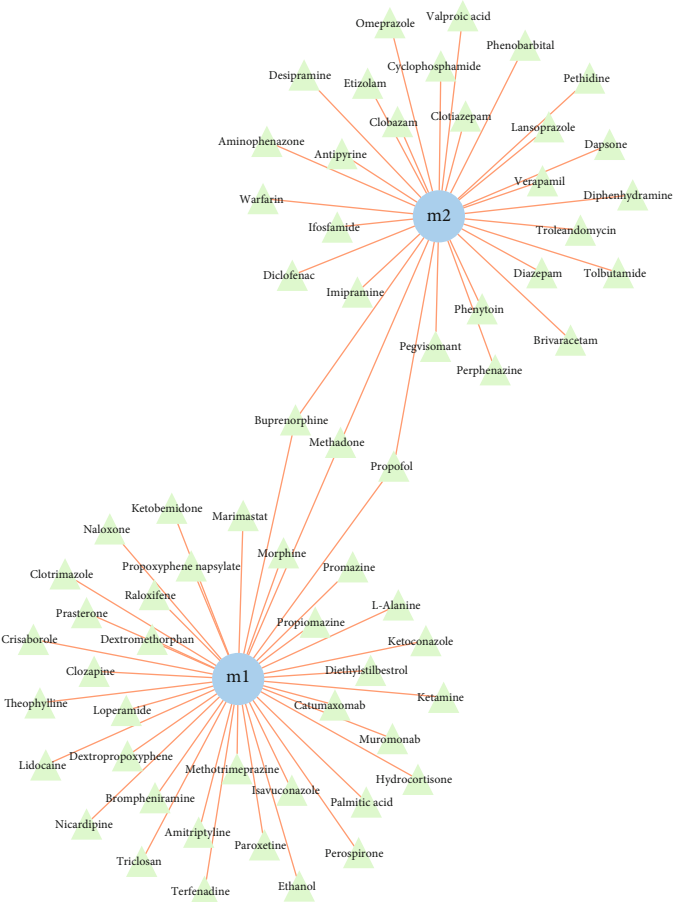
3.5. Potential Role of Drug Target Genes. Based on the pivot analysis of the drug, the target gene for drug action was

traced back through the DrugBank database with threshold $P < 0.05$. Module_Drug_TargetGene regulatory relationship table for its target genes and drugs with potential effects on sevoflurane anesthesia was obtained, and zinc-related interactions were selected to construct a regulatory network map (Figure 5(a)). The figure indicates that zinc affects the progression of OPCABG through AHSG, F12, and other genes

(1S,6R,9AS,11R,11BR)-9A,11B-DIMETHYL-1-[(METHYLOXY)METHYL]-3,6,9-TRIOXO-1,6,6B,7,8,9,9A,10,11,11B-DECAHYDRO-3H-FURO[4,3,2-DE]INDENO[4,5-H]2]BENZOPYRAN-11-YL ACETATE

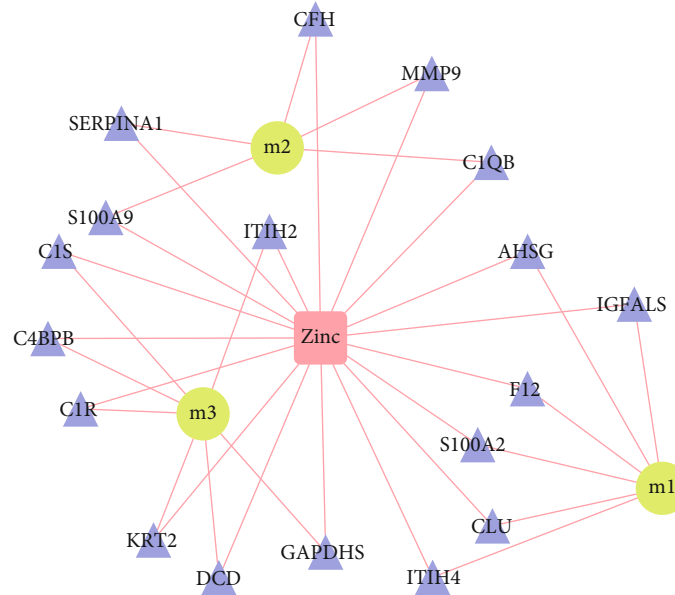


(a)

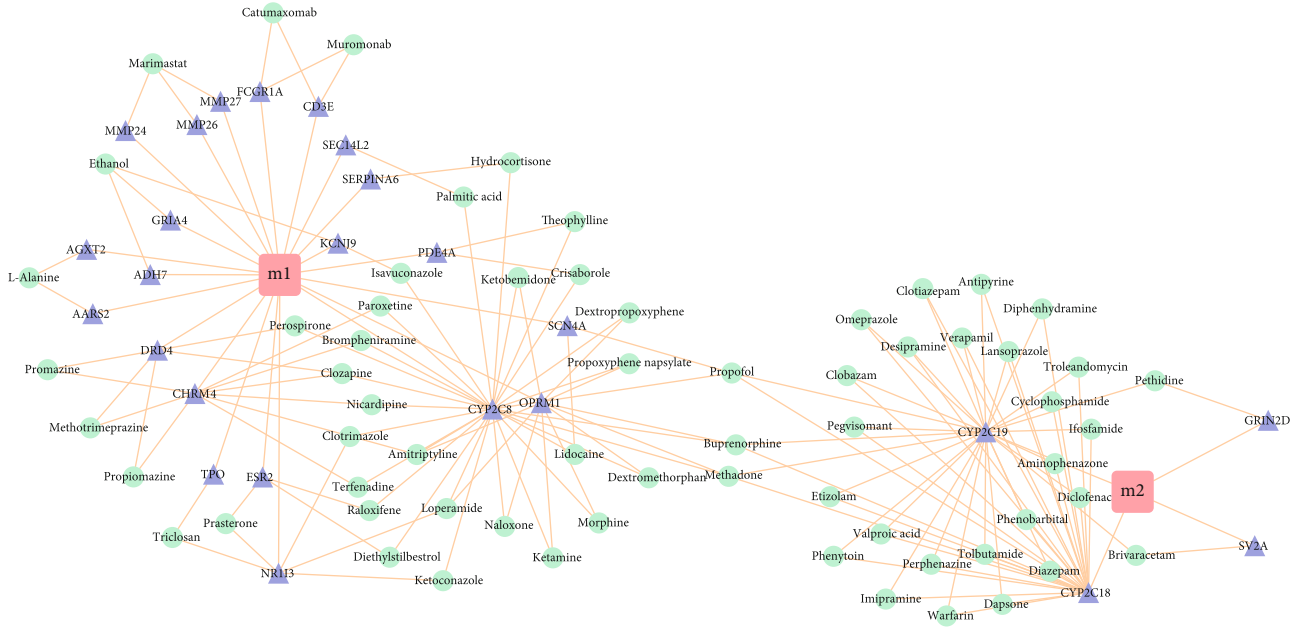


(b)

FIGURE 4: Regulation of drugs on dysfunction modules. (a) Potential module_drug regulatory relationship for preoperative sevoflurane anesthesia, with pink octagons representing modules and blue-green circles representing potential drugs. (b) Potential module_drug regulation of propofol anesthesia before surgery, with blue circles representing modules and green triangles representing potential drugs.



(a)



(b)

FIGURE 5: Regulation of drug target genes on dysfunction modules. (a) Potential Module_Drug_TargetGene regulation map for preoperative sevoflurane anesthesia, with green circles representing modules, pink quadrilateral for drugs, and purple triangles for target genes. (b) Potential Module_Drug_TargetGene regulation map of propofol anesthesia before surgery, pink quadrilateral represents the module, blue-green circles represent drugs, and purple triangles represent target genes.

regulating functional disorder modules. In addition, the Module_Drug_TargetGene regulatory network for target genes and drugs with potential effect on propofol anesthesia surgery was obtained (Schedule 4-1, Figure 5(b)).

4. Discussion

Coronary vascular disease has become a problem facing the world while common treatment is coronary artery bypass surgery (CABG). In the United States alone, about 500,000

patients need CABG surgery every year. OPCABG is a form of CABG surgery. According to statistics, due to its hemodynamic abnormalities during surgery [19], only about 20% of CABG surgeries are now performed under nonextracorporeal circulation. OPCABG has exhibited some advantages, especially in reducing postoperative complications such as systemic inflammation and myocardial and cerebral damage [20, 21]. Acute kidney injury (AKI) is also one of the common postoperative complications of OPCABG, which may be related to chlorine free radical IVF [22]. We need to

anesthetize patients before OPCABG surgery, and anesthesia may also cause unexpected hypothermia, which may be another complication of perioperative cardiovascular [23]. Therefore, it is very important to choose a suitable anesthesia method.

In this study, the typical anesthesia methods were selected: sevoflurane anesthesia and propofol anesthesia. The effects of two anesthesia methods on OPCABG were studied, based on a functional module network. We obtained microarray data from sevoflurane anesthesia (inhalation anesthesia) and propofol anesthesia (intravenous anesthesia) from the GEO database, differentially analysed the two groups of data, respectively, and obtained 5701 and 3210 differentially expressed genes, respectively. We believe that these differential genes are dysfunctional molecules. We then performed WGCNA analysis on the two groups of differential genes, and seven functional barrier modules were obtained by sevoflurane anesthesia clustering, including PDCD6IP, WDR3, and other core genes; propofol anesthesia clustering obtained two functional disorders module including two core genes, KCNB2 and LHX2. After obtaining the dysfunction module, the mechanism of action cannot be fully explained, so further enrichment analysis of the functions and pathways of interest is needed. The GO enrichment analysis of the sevoflurane anesthesia-related module gene found that it mainly focused on biological processes such as anion antiporter activity, and KEGG enrichment analysis found that it was mainly related to the PI3K-Akt signalling pathway. The GO enrichment analysis of the gene related to propofol anesthesia showed that it mainly focused on biological processes such as channel activity. The KEGG enrichment analysis found that it was mainly relative with the neuroactive ligand-receptor interaction plasma pathway. In addition, through the coexpression network, some scholars have found that anesthetics may protect the heart by activating the complement and coagulation system [24]. There is still some controversy about the impact of two anesthesia methods on surgery. On the one hand, clinical data have shown that sevoflurane can reduce death within 180 to 365 days after surgery with positive inotropic and vasoconstrictor support with minimal impact on cardiac index [25]. It can also be labelled with the sensitive biomarkers miR-499 and miR-208b [26]. On the other hand, studies have shown that 30% of hernia in propofol anesthesia improves hemodynamic stability by reducing the patient's norepinephrine requirement [27]. Of course, some new anesthesia methods are now available, such as remifentanyl target-controlled infusion, constant-rate infusion, chest epidural anesthesia (EA), and postoperative epidural infusion (EI) [28, 29]. During the operation, drugs are often used, so we analysed the potential drugs of the two anesthesia methods to provide guidance for clinical use. The pivot analysis of potential drug regulators in this study showed that there were 229 potential drugs in the operation of sevoflurane anesthesia and zinc regulated three functional disorder modules through core genes such as AHSG and F12; pivot analysis of potential drug regulators during propofol anesthesia surgery showed 67 potential agents, including propofol, methadone, and buprenorphine which regulate two functional barrier modules via the four genes

CYP2C8, OPRM1, CYP2C18, and CYP2C19. This study provides guidance on clinical use or treatment by comparing the effects of two anesthetics on surgery and its potential drugs.

Data Availability

The data associated with this manuscript can be accessed from GSE4386.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] R. Cartier, "Current trends and technique in OPCAB surgery," *Journal of Cardiac Surgery*, vol. 18, no. 1, pp. 32–46, 2003.
- [2] K. Nakazato and A. Sakamoto, "Opcab," *Masui*, vol. 63, no. 5, pp. 506–512, 2014.
- [3] V. Dhurandhar, A. Saxena, R. Parikh et al., "Comparison of the safety and efficacy of on-pump (ONCAB) versus off-pump (OPCAB) coronary artery bypass graft surgery in the elderly: a review of the ANZSCTS database," *Heart, Lung & Circulation*, vol. 24, no. 12, pp. 1225–1232, 2015.
- [4] V. V. Likhvantsev, G. Landoni, D. I. Levikov, O. A. Grebenchikov, Y. V. Skripkin, and R. A. Cherpakov, "Sevoflurane versus total intravenous anesthesia for isolated coronary artery bypass surgery with cardiopulmonary bypass: a randomized trial," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 30, no. 5, pp. 1221–1227, 2016.
- [5] R. Xu, R. Lu, H. Jiang et al., "Meta-analysis of protective effect of sevoflurane on myocardium during cardiac surgery," *European Review for Medical and Pharmacological Sciences*, vol. 18, no. 7, pp. 1058–1066, 2014.
- [6] P. Zeyneloglu, A. Donmez, and M. Sener, "Sevoflurane induction in cyanotic and acyanotic children with congenital heart disease," *Advances in Therapy*, vol. 25, no. 1, pp. 1–8, 2008.
- [7] J. Mizuno, S. Morita, K. Kamiya, M. Honda, K. Momoeda, and K. Hanaoka, "Isorhythmic dissociation during sevoflurane anesthesia," *Masui*, vol. 58, no. 5, pp. 645–648, 2009.
- [8] L. Pasin, G. Landoni, L. Cabrini et al., "Propofol and survival: a meta-analysis of randomized clinical trials," *Acta Anaesthesiologica Scandinavica*, vol. 59, no. 1, pp. 17–24, 2015.
- [9] Y. C. Tsai, L. Y. Wang, and L. S. Yeh, "Clinical comparison of recovery from total intravenous anesthesia with propofol and inhalation anesthesia with isoflurane in dogs," *The Journal of Veterinary Medical Science*, vol. 69, no. 11, pp. 1179–1182, 2007.
- [10] N. R. Searle and P. Sahab, "Propofol in patients with cardiac disease," *Canadian Journal of Anaesthesia*, vol. 40, no. 8, pp. 730–747, 1993.
- [11] L. Jia, R. Dong, F. Zhang et al., "Propofol provides more effective protection for circulating lymphocytes than sevoflurane in patients undergoing off-pump coronary artery bypass graft surgery," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 29, no. 5, pp. 1172–1179, 2015.
- [12] S. Suryaprakash, M. Chakravarthy, G. Muniraju et al., "Myocardial protection during off pump coronary artery bypass surgery: a comparison of inhalational anesthesia with sevoflurane or desflurane and total intravenous anesthesia," *Annals of Cardiac Anaesthesia*, vol. 16, no. 1, pp. 4–8, 2013.

- [13] G. Dorantes-Mendez, F. Aletti, N. Toschi et al., "Effects of propofol anesthesia induction on the relationship between arterial blood pressure and heart rate," *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2012, pp. 2835–2838, 2012.
- [14] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D991–D995, 2013.
- [15] M. E. Ritchie, B. Phipson, D. I. Wu et al., "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, p. e47, 2015.
- [16] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.
- [17] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS*, vol. 16, no. 5, pp. 284–287, 2012.
- [18] C. Gu, X. Shi, Z. Huang et al., "A comprehensive study of construction and analysis of competitive endogenous RNA networks in lung adenocarcinoma," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1868, no. 8, p. 140444, 2020.
- [19] R. Battu, A. Prasad, and M. Kanchi, "Perioperative optic neuropathy in patients undergoing off-pump coronary artery bypass graft surgery," *Annals of Cardiac Anaesthesia*, vol. 17, no. 2, pp. 92–97, 2014.
- [20] J. Huffmyer and J. Raphael, "The current status of off-pump coronary bypass surgery," *Current Opinion in Anaesthesiology*, vol. 24, no. 1, pp. 64–69, 2011.
- [21] T. M. Hemmerling, G. Romano, N. Terrasini, and N. Noiseux, "Anesthesia for off-pump coronary artery bypass surgery," *Annals of Cardiac Anaesthesia*, vol. 16, no. 1, pp. 28–39, 2013.
- [22] K. Bhaskaran, G. Arumugam, and P. V. Vinay Kumar, "A prospective, randomized, comparison study on effect of perioperative use of chloride liberal intravenous fluids versus chloride restricted intravenous fluids on postoperative acute kidney injury in patients undergoing off-pump coronary artery bypass grafting surgeries," *Annals of Cardiac Anaesthesia*, vol. 21, no. 4, pp. 413–418, 2018.
- [23] Y. J. Cho, S. Y. Lee, T. K. Kim, D. M. Hong, and Y. Jeon, "Effect of prewarming during induction of anesthesia on microvascular reactivity in patients undergoing off-pump coronary artery bypass surgery: a randomized clinical trial," *PLoS One*, vol. 11, no. 7, article e0159772, 2016.
- [24] X. Bu, B. Wang, Y. Wang et al., "Pathway-related modules involved in the application of sevoflurane or propofol in off-pump coronary artery bypass graft surgery," *Experimental and Therapeutic Medicine*, vol. 14, no. 1, pp. 97–106, 2017.
- [25] J. E. Pereira, A. Agarwal, H. Goma et al., "Inhalation versus intravenous anaesthesia for adults undergoing on-pump or off-pump coronary artery bypass grafting: A systematic review and meta-analysis of randomized controlled trials," *Journal of Clinical Anesthesia*, vol. 40, pp. 127–138, 2017.
- [26] X. Liu, X. Liu, R. Wang et al., "Circulating microRNAs indicate cardioprotection by sevoflurane inhalation in patients undergoing off-pump coronary artery bypass surgery," *Experimental and Therapeutic Medicine*, vol. 11, no. 6, pp. 2270–2276, 2016.
- [27] S. Devroe, G. Dewinter, M. Van de Velde et al., "Xenon as an adjuvant to propofol anesthesia in patients undergoing off-pump coronary artery bypass graft surgery: a pragmatic randomized controlled clinical trial," *Anesthesia and Analgesia*, vol. 125, no. 4, pp. 1118–1128, 2017.
- [28] A. Shu, L. Zhan, H. Fang et al., "Evaluation of remifentanyl anesthesia for off-pump coronary artery bypass grafting surgery using heart rate variability," *Experimental and Therapeutic Medicine*, vol. 6, no. 1, pp. 253–259, 2013.
- [29] M. Y. Kirov, A. V. Eremeev, A. A. Smetkin, and L. J. Bjertnaes, "Epidural anesthesia and postoperative analgesia with ropivacaine and fentanyl in off-pump coronary artery bypass grafting: a randomized, controlled study," *BMC Anesthesiology*, vol. 11, no. 1, p. 17, 2011.

Research Article

Integrated Genome-Wide Methylation and Expression Analyses Reveal Key Regulators in Osteosarcoma

Fei Wang,¹ Guoqing Qin,² Junzhi Liu,³ Xiunan Wang,⁴ and Baoguo Ye ⁵

¹Department of Orthopedics, China-Japan Union Hospital Jilin University, Changchun, Jilin, China

²Department of Orthopedics, Jilin Disabled Persons' Rehabilitation Center, Jilin Chunguang Rehabilitation Hospital, Changchun, Jilin, China

³Quality Control Department, China-Japan Union Hospital Jilin University, Changchun, Jilin, China

⁴Department of Orthopedics, The 964th Hospital of the PLA Joint Logistics Support Force, No. 4799 Xi'an Road, Lvyuan District, Changchun City, Jilin Province, China

⁵Department of Anesthesiology, China-Japan Union Hospital Jilin University, Changchun, Jilin, China

Correspondence should be addressed to Baoguo Ye; yebaoguo_cb@126.com

Received 7 June 2020; Accepted 23 July 2020; Published 13 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Fei Wang 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.

Osteosarcoma (OS) is one of the most common types of primary bone tumors in early adolescence with unsatisfied prognosis. Aberrant DNA methylation had been demonstrated to be related to tumorigenesis and progression of multiple cancers and could serve as the potential biomarkers for the prognosis of human cancers. In conclusion, this study identified 18 downregulated hypomethylation genes and 52 upregulated hypomethylation genes in OS by integrating the analysis the GSE97529 and GSE42572 datasets. Bioinformatics analysis revealed that OS-specific methylated genes were involved in regulating multiple biological processes, including chemical synaptic transmission, transcription, response to drug, and regulating immune response. KEGG pathway analysis showed that OS-specific methylated genes were associated with the regulation of Hippo, cAMP calcium, MAPK, and Wnt signaling pathways. By analyzing R2 datasets, this study showed that the dysregulation of these OS-specific methylated genes was associated with the metastasis-free survival time in patients with OS, including CBLN4, ANKMY1, BZW1, KRTCAP3, GZMB, KRTDAP, LY9, PFKFB2, PTPN22, and CLDN7. This study provided a better understanding of the molecular mechanisms underlying the progression and OS and novel biomarkers for the prognosis of OS.

1. Introduction

Osteosarcoma (OS) is one of the most common types of primary bone tumors in early adolescence, which was characterized by an aggressive osteolytic or osteoblastic appearance with a periosteal reaction [1]. Chemotherapy and surgery are the most important treatments for patients with OS [2, 3]. The survival rate of primary OS patients after treatments remains at 60–70% [4]. However, the prognosis of patients with progressive or recurrent OS was less than 20% [5]. In the past decades, emerging studies reported that multiple factors are associated with the tumorigenesis and progression of OS, including germline genetic variants [6], dysregulation of oncogenes or tumor suppressors [7], and the abnormal epi-

genetics change [8, 9]. A few proteins had been revealed to be related to the progression of OS. For example, GFRA1 was reported to promote autophagy and cisplatin-induced chemoresistance in OS [10]. The isoform 1 of TMIGD3 suppressed OS progression though downregulating NF- κ B [11]. Understanding the mechanisms related to OS development could provide new targets for OS.

DNA methylation could affect the gene expression though suppressing transcription [12]. Aberrant DNA methylation had been demonstrated to be involved in regulating tumorigenesis and progression of multiple cancers [13, 14]. In OS, DNA methylation-mediated suppression of miR-449c could promote cell cycle though inhibiting c-Myc in OS [15]. Hypomethylation of IRX1 was found to promote

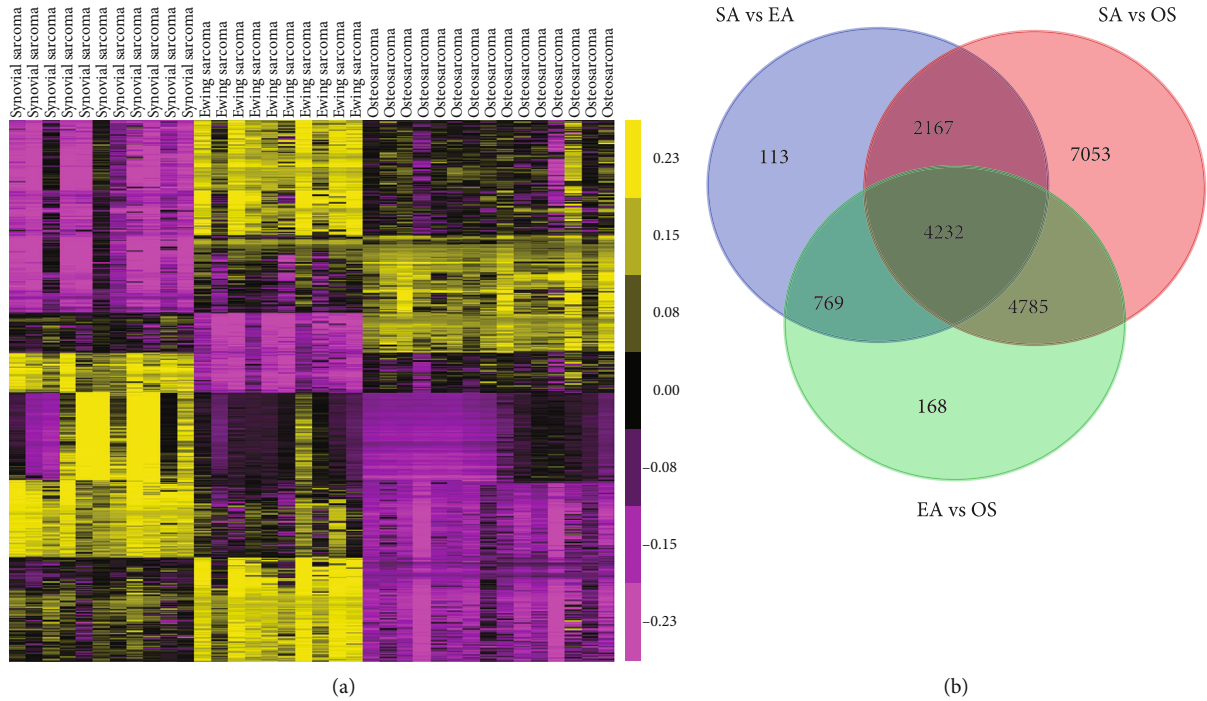


FIGURE 1: OS-specific methylated genes were identified by using the public dataset GSE97529. (a) DNA methylation status of 482,421 CpG sites in 10 Ewing's sarcoma, 11 synovial sarcoma, and 15 OS samples were included in this dataset.

OS metastasis by activating CXCL14/NF- κ B signaling [16]. Very interestingly, recent studies showed that aberrant DNA methylation was associated with the prognosis of OS. For example, the DNA methylation level of WNT6 was negatively correlated to the prognosis of children with osteosarcoma [17]. The hypermethylation of ESR1 was correlated to the worse overall survival of OS [18]. These results suggested that the DNA methylation status could be potential diagnostic and therapeutic targets for OS.

The present study analyzed the GSE97529 [19] dataset to identify OS-specific methylated genes. In silico analyses were performed to explore the functions of OS-specific methylated genes. Next, the GSE42572 dataset was used to validate the expression levels of OS-specific methylated genes [20]. Of note, we found that these OS-specific methylated genes were correlated to the prognosis of patients with OS. By these methods, it is hopeful that novel aberrant methylation genes and pathways will be screened in the OS and an understanding of the underlying molecular mechanisms will be enhanced.

2. Materials and Methods

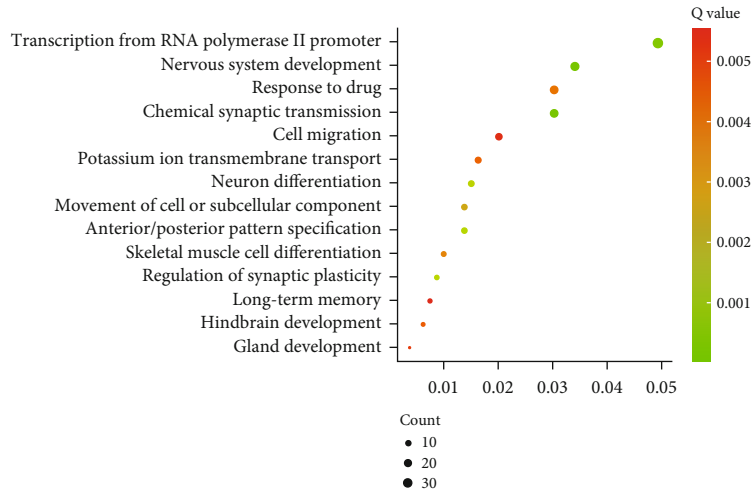
2.1. Microarray Data. The present study is aimed at identifying dysregulated OS-specific methylated genes in OS by analyzing public databases with bioinformatics analysis. Thus, we screened the GEO databases. The candidate databases were selected according to 3 standards: (1) the candidate database should contain clinical OS samples, (2) the number of clinical samples should be more than 10 cases, and (3) the candidate database was not noncoding RNA datasets. Finally, only the SE97529 and GSE42572 datasets were selected for

further analysis. We have included this information in Materials and Methods. The GSE97529 dataset was used to identify OS-specific methylated genes, which was downloaded from the NCBI GEO database (GSE97529). A total of 10 Ewing's sarcoma, 11 synovial sarcoma, and 15 OS samples were included in this dataset. The GSE42572 dataset was analyzed to identify differentially expressed genes in OS compared to normal samples, which was also downloaded from the NCBI GEO database (GSE42572). Differentially expressed genes (DEGs) and differentially methylated genes (DMGs) were identified by applying GEO2R. $P < 0.05$ and $|\text{fold change}| \geq 2$ is set as the cutoff criterion.

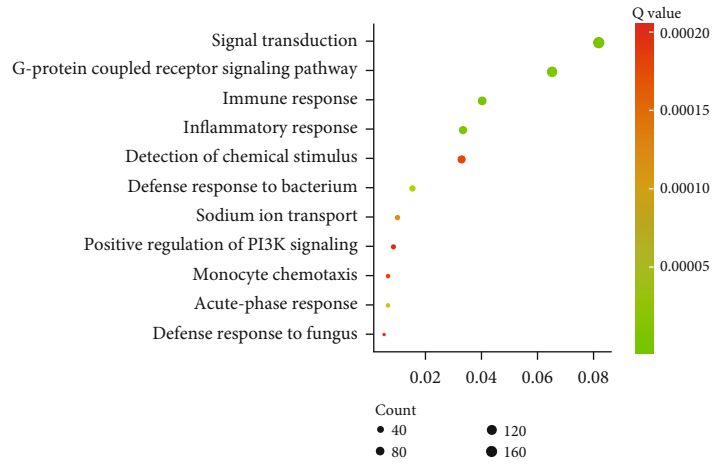
2.2. Functional and Pathway Enrichment Analyses. The DAVID system was used to predict the potential biological processes and KEGG pathways involved in target genes in this study [21]. $P < 0.05$ was set as the cutoff criterion.

2.3. Protein-Protein Interaction (PPI) Network Analysis. In the present study, PPI networks were used to reveal the interactions among differentially expressed OS-specific methylated genes using the STRING database (<https://string-db.org/>). PPI was visualized using Cytoscape [22].

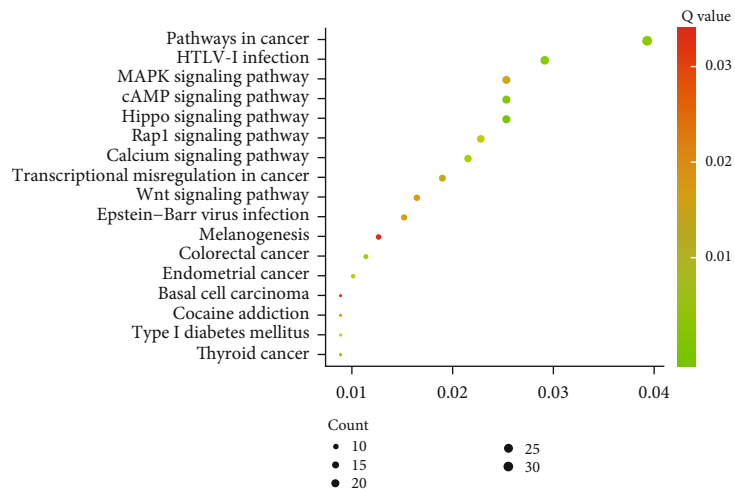
2.4. Survival Analysis. Survival analysis was performed using the OS microarray dataset (mixed osteosarcoma (mesenchymal)-Kuijjer-127-vst-ilmnhwg6v2) from the R2: Genomics Analysis and Visualization Platform (<http://r2.amc.nl>). The median expression of targets was selected as the cutoff to divide all OS samples into the high or low group.



(a)



(b)



(c)

FIGURE 2: Continued.

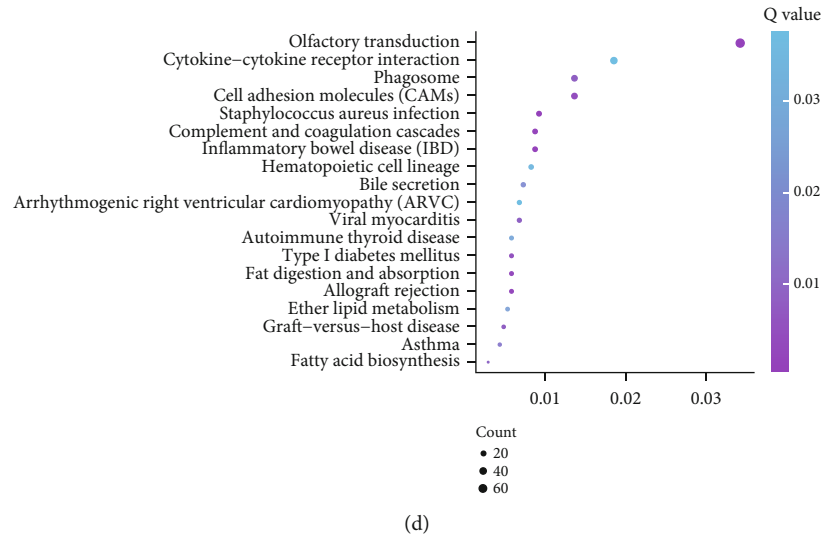


FIGURE 2: Bioinformatics analysis of hypermethylation genes and hypomethylation genes. (a) GO analysis of OS-specific hypermethylation genes. (b) GO analysis of OS-specific hypomethylation genes. (c) KEGG pathway analysis of OS-specific hypermethylation genes. (d) KEGG pathway analysis of OS-specific hypomethylation genes. The gene ratio was present in the X-axis.

3. Results

3.1. Identification of OS-Specific Methylated Genes. The public dataset GSE97529 was used to identify OS-specific methylated genes. DNA methylation status of 482,421 CpG sites in 10 Ewing's sarcoma, 11 synovial sarcoma, and 15 OS samples were included in this dataset (Figure 1(a)). Totally, we identified 3125 OS-specific methylated genes, including 875 hypermethylation genes and 2250 hypomethylation genes in OS samples compared to Ewing's sarcoma or synovial sarcoma samples (Figure 1(a)).

3.2. GO and KEGG Pathway Enrichment Analyses. GO analysis showed that hypermethylation genes were significantly associated with biological processes (BP) of the nervous system development, chemical synaptic transmission, transcription from RNA polymerase II promoter, anterior/posterior pattern specification, regulation of synaptic plasticity, neuron differentiation, movement of cell or subcellular component, skeletal muscle cell differentiation, response to drug, potassium ion transmembrane transport, hindbrain development, gland development, and cell migration (Figure 2(a)). Hypomethylation genes were significantly related to immune response, signal transduction, inflammatory response, acute-phase response, sodium ion transport, monocyte chemotaxis, detection of chemical stimulus, defense response to fungus, positive regulation of PI3K pathway, cell chemotaxis, chemotaxis, neutrophil chemotaxis, innate immune response, ion transmembrane transport, and cell adhesion (Figure 2(b)).

KEGG pathway analysis showed that significant pathways of hypermethylation genes in OS included the Hippo pathway, cAMP signaling, thyroid cancer, pathways in cancer, calcium signaling, endometrial cancer, Rap1 signaling pathway, transcriptional misregulation in cancer, MAPK signaling pathway, Epstein-Barr virus infection, Wnt signaling pathway, cocaine addiction, and basal cell carcinoma

(Figure 2(c)). And hypomethylation genes in OS were associated with Staphylococcus aureus infection, olfactory transduction, inflammatory bowel disease (IBD), complement and coagulation cascades, allograft rejection, fat digestion and absorption, graft-versus-host disease, phagosome, viral myocarditis, and fatty acid biosynthesis (Figure 2(d)).

3.3. OS-Specific Methylated Genes Were Differentially Expressed in OS. Subsequently, an independent public dataset, GSE42572, was used to identify differentially expressed genes in OS. As shown in Figure 3(a), we identified 614 upregulated genes and 696 downregulated genes in OS compared to healthy control samples (Figure 3(a)). Among DEGs, a total of 18 downregulated hypomethylation genes were screened out from overlapping 875 hypermethylation and 690 downregulated genes, while 52 upregulated hypomethylation genes were screened out from overlapping 2250 hypomethylation and 614 downregulated genes (Figure 3(b)). The 70 differentially expressed OS-specific methylated genes were presented by heat map (Figure 3(c)).

3.4. Construction of PPI Network to Identify Hub Differentially Expressed OS-Specific Methylated Genes. Furthermore, we constructed a PPI network to identify a hub differentially expressed OS-specific methylated gene using the STRING database. As presented in Figure 4, a total of 29 nodes and 30 edges were included in this network. The hub genes included NPSR1, PTAFR, LPAR5, PTGER3, NPY5R, KCNK3, KRTDAP, HCN4, KRT38, KCNIP2, KCNJ5, and KRTCAP3 (Figure 4).

3.5. The Survival Time Analysis of Differentially Expressed OS-Specific Methylated Genes. The above analysis was conducted with the GSE97529 and GSE42572 datasets. Unfortunately, the clinical information about metastasis-free survival time was not included in both databases. Thus, we analyzed an independent database, R2 dataset (<http://r2.amc.nl>), to

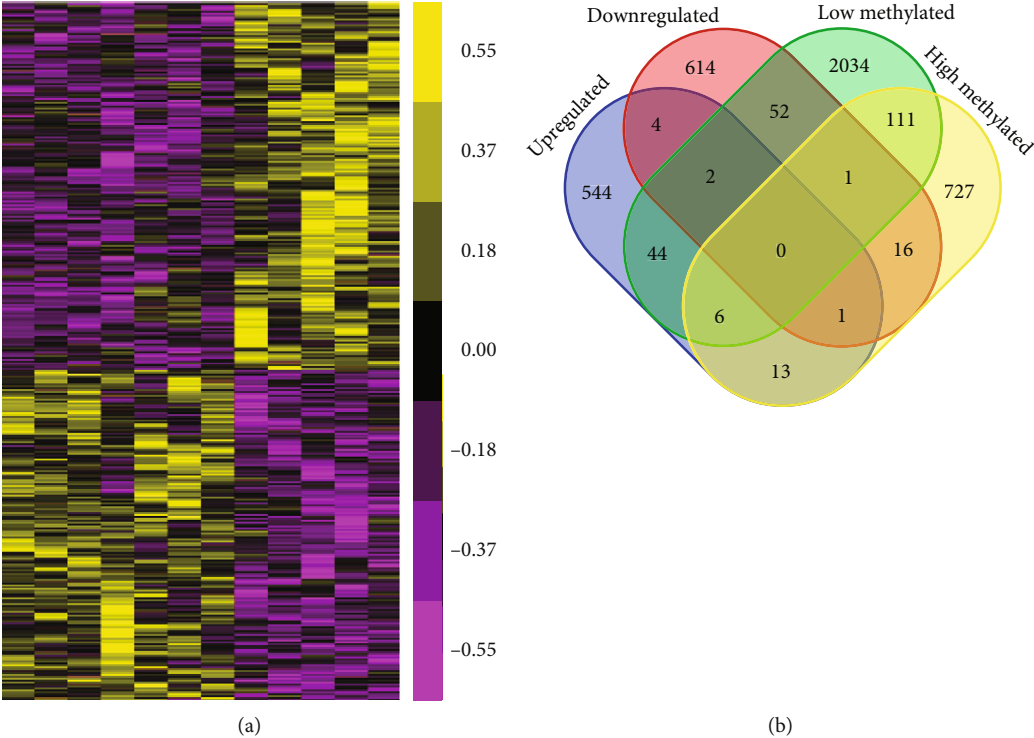


FIGURE 3: Continued.

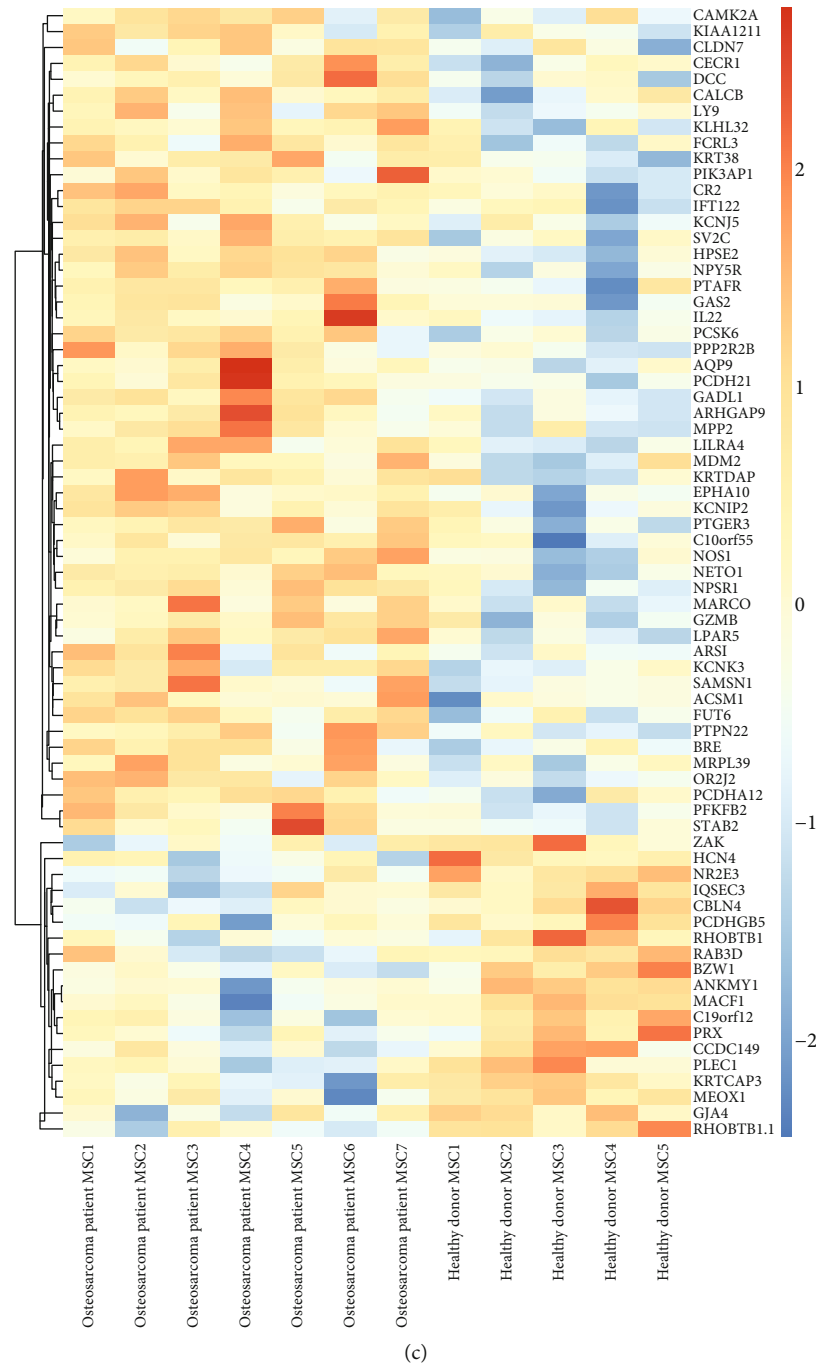


FIGURE 3: GSE42572 was analyzed to identify differentially expressed genes in OS compared to normal samples. (a) 614 induced genes and 696 reduced genes in OS compared to healthy control samples. (b) Among DEGs, a total of 18 downregulated hypomethylation genes and 52 upregulated hypomethylation genes were screened out. (c) The 70 differentially expressed OS-specific methylated genes were presented by heat map.

further evaluate the prognostic value of OS-specific methylated genes. The median expression of candidates in all OS samples was selected.

As the cutoff is used to divide OS samples into the high and low groups, it was shown that higher expression of CBLN4 ($P < 0.05$) was associated with longer metastasis-free survival time in patients with OS, as well as ANKMY1 ($P < 0.05$), BZW1 ($P < 0.05$), and KRTCAP3 ($P < 0.001$). However, higher expression of GZMB ($P < 0.05$), KRTDAP

($P < 0.05$), LY9 ($P < 0.05$), PFKFB2 ($P < 0.05$), PTPN22 ($P < 0.05$), and CLDN7 ($P < 0.05$) was associated with shorter metastasis-free survival time in patients with OS (Figure 5).

4. Discussion

The mechanisms underlying OS progression remained largely unclear. It has been widely accepted that DNA

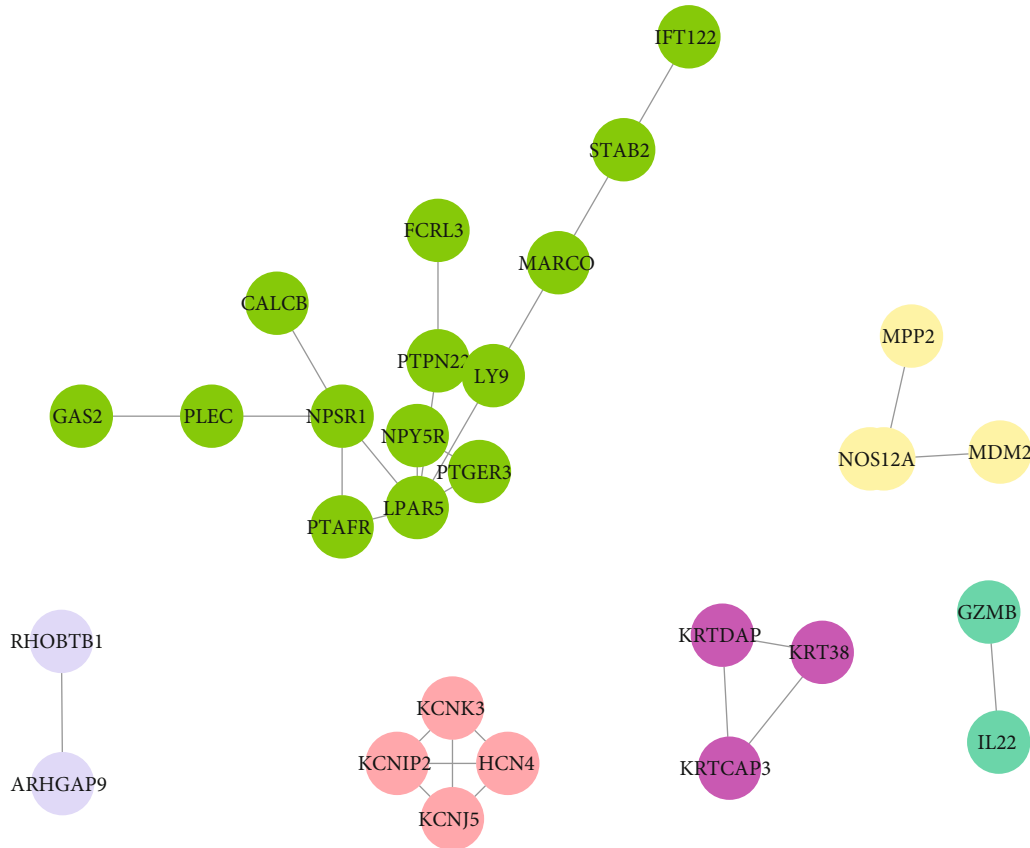


FIGURE 4: PPI network analysis. PPI network was used to identify a hub differentially expressed OS-specific methylated gene using the STRING database.

methylation was involved in regulating the tumorigenesis and development through modulating gene expression. DNA methylation has been shown to play an important role in gene regulation and implicated in various types of cancer. Emerging studies revealed that the cancer-specific CpG hypermethylation could turn off the expression of tumor suppressors; however, cancer-specific CpG hypomethylation could activate the expression of oncogenes [23]. Identification of aberrantly methylated genes in OS would be helpful to identify new diagnostic and therapeutic biomarkers for OS. The present study identified OS-specific methylated genes from Ewing's sarcoma or synovial sarcoma samples. Bioinformatics analysis revealed that OS-specific methylated genes were involved in regulating multiple biological processes, including chemical synaptic transmission, transcription, response to drug, and regulating immune response. Further validation indicated that OS-specific methylated genes were dysregulated in OS samples and correlated to the prognosis of patients with OS.

OS, together with Ewing's Sarcoma (EWS) and synovial sarcoma (SS), was the most common pediatric sarcomas [24]. These types of sarcomas occur in similar anatomical locations; however, the treatments for these sarcomas differed depending on the tumor type. The accurate diagnosis of OS remained to be a big challenge. Emerging studies demonstrated that aberrant DNA methylation was associated with the prognosis of human cancers, including OS. For

example, DNA methylation level of WNT6 and ESR1 was related to the prognosis of OS. The present study is aimed at identifying OS-specific methylated genes. A total of 3125 OS-specific methylated genes were identified, including 875 hypermethylation genes and 2250 hypomethylation genes in OS samples compared to Ewing's sarcoma or synovial sarcoma samples. Furthermore, GO and KEGG pathway analyses were further used to predict the potential roles of OS-specific methylated genes. Of note, our predictions showed that these methylated genes were associated with the Hippo signaling and Wnt signaling. Hippo pathway aberrations had been demonstrated in OS by multiple studies and involved in regulating primary tumor growth, angiogenesis, epithelial to mesenchymal transition, and metastatic dissemination [25]. The Hippo signaling played an important role controlling cancer cell proliferation and apoptosis [26]. Multiple studies indicated YAP was overexpressed in OS samples, and knockdown of YAP significantly inhibits OS cell growth and invasion [27]. Sox2, as a YAP upstream regulator, was reported to be required for tumor development and cancer cell proliferation in OS [28]. This study provided a potential mechanism to elucidate how the Hippo signaling activated in OS. Many studies support an aberrant activation of the canonical Wnt signaling pathway in osteosarcoma cells. For example, two recent studies described a high β -catenin level in osteosarcoma tissues compared to adjacent healthy tissues associated with poor prognosis and lung metastatic

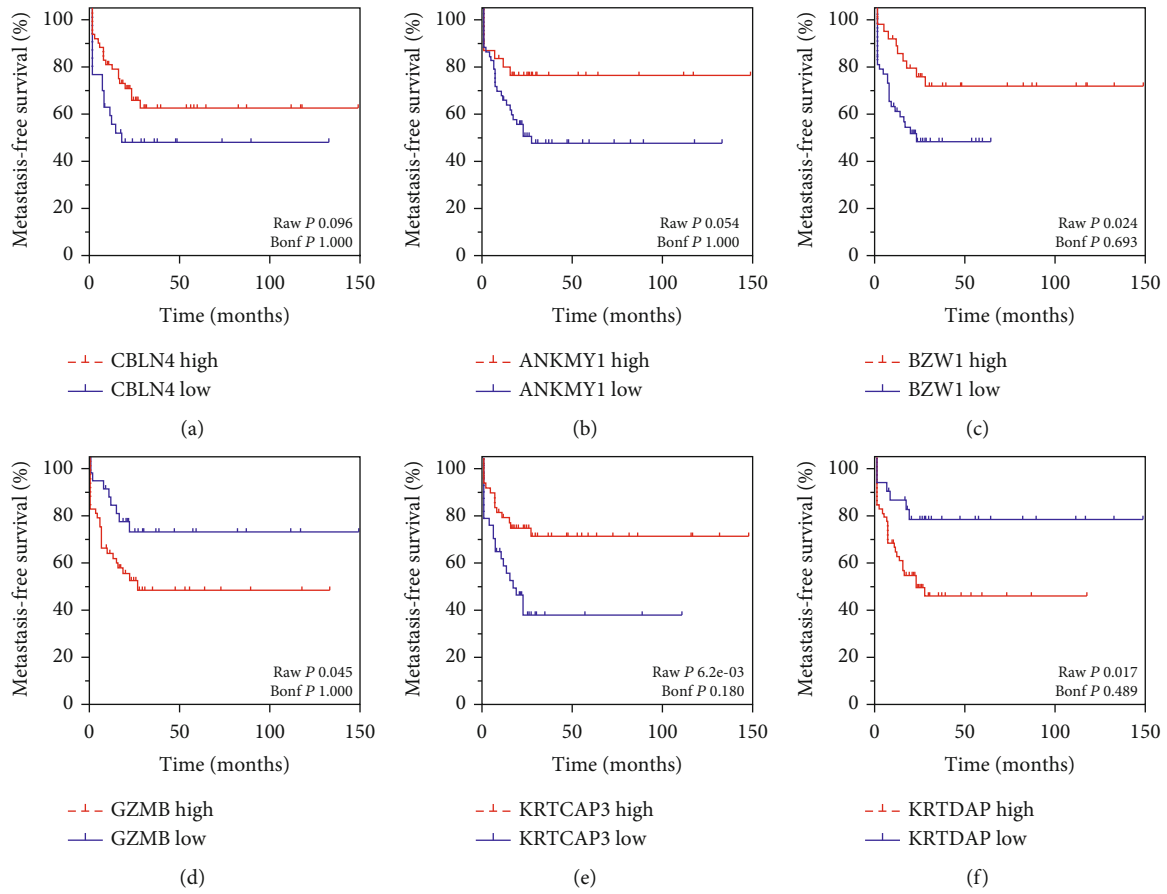


FIGURE 5: Continued.

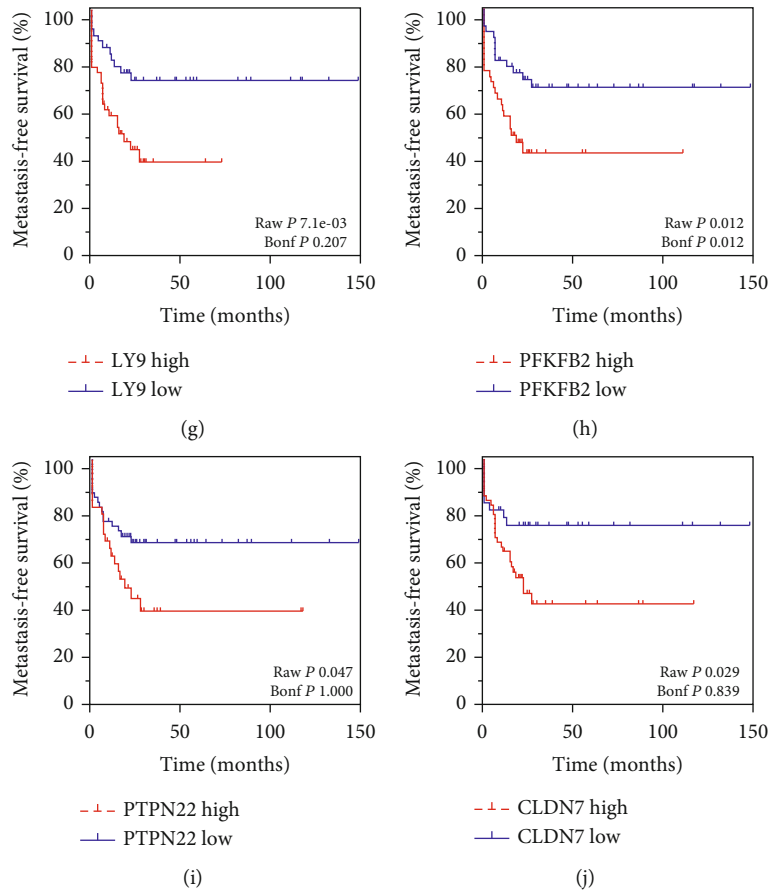


FIGURE 5: The prognostic values of differentially expressed OS-specific methylated genes were calculated by using the R2: Genomics Analysis and Visualization Platform. (a–j) Higher expression of CBLN4 (a) was associated with longer metastasis-free survival time in patients with OS, as well as ANKMY1 (b), BZW1 (c), and KRTCAP3 (e). However, higher expression of GZMB (d), KRTDAP (f), LY9 (g), PFKFB2 (h), PTPN22 (i), and CLDN7 (j) was associated with shorter metastasis-free survival time in patients with OS.

dissemination. Wnt signaling pathway played a crucial role in tumorigenicity and metastasis via regulation of the immune system, bone remodeling, angiogenesis, hypoxia response, and EMT [29].

Of note, this study showed that OS-specific methylated genes were significantly differentially expressed in OS samples. A total of 18 downregulated hypomethylation genes and 52 upregulated hypomethylation genes were identified in this study. PPI network analysis was constructed to reveal the relation among these genes. Totally, 29 nodes and 30 edges were included in this network. By analyzing R2 datasets, we found the dysregulation of these OS-specific methylated genes were associated with the metastasis-free survival time in patients with OS, including CBLN4, ANKMY1, BZW1, KRTCAP3, GZMB, KRTDAP, LY9, PFKFB2, PTPN22, and CLDN7. Among these regulators, BZW1 is a transcription factor related to the regulation of cell cycle and proliferation [30]. LY9 was a member of SLAM family of immunomodulatory receptors [31] and interacted with the adaptor molecule signaling lymphocyte activation molecule-associated proteins. A previous study showed LY9 was related to the cancer progression and correlated to overall survival of the patients with breast cancer. PFKFB2 is an enzyme involved in regulating the Warburg effect (also

termed as glycolysis) [32]. PFKFB2 had been found to have a key role in regulating tumor growth and survival in multiple cancer types, including gastric cancer, gliomas, and osteosarcoma [32–37].

Several limitations were also existed in this study. First, our studies revealed several hub OS-specific methylated genes. However, the roles of these genes remained to be unclear. The gain or loss of function assays should be performed to further explore their roles in OS. Next, the expression levels and methylation levels of hub OS-specific methylated genes in OS samples should be confirmed using clinical samples. Third, the direct interaction among these hub genes has not been confirmed using experimental assays.

5. Conclusion

In conclusion, this study identified 18 downregulated hypomethylation genes and 52 upregulated hypomethylation genes in OS and a series biological processes and pathways regulated by aberrantly methylated genes. PPI network analysis revealed the interactions among these genes. Moreover, the present study showed that the dysregulation of OS-specific methylated genes was correlated with the metastasis-free time in patients with OS, including CBLN4,

ANKMY1, BZW1, KRTCAP3, GZMB, KRTPAP, LY9, PFKFB2, PTPN22, and CLDN7. This study provided a better understanding of the molecular mechanisms underlying the progression and OS and novel biomarkers for the prognosis of OS.

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.

Authors' Contributions

Fei Wang and Guoqing Qin are co-first authors.

Acknowledgments

This work was supported by the Special fund for the key laboratory of Science and Technology Department of Jilin Province (20190201282JC).

References

- [1] E. Simpson and H. L. Brown, "Understanding osteosarcomas," *JAAPA*, vol. 31, no. 8, pp. 15–19, 2018.
- [2] G. Bacci, P. Picci, P. Ruggieri et al., "Primary chemotherapy and delayed surgery (neoadjuvant chemotherapy) for osteosarcoma of the extremities. The Istituto Rizzoli experience in 127 patients treated preoperatively with intravenous methotrexate (high versus moderate doses) and intraarterial cisplatin," *Cancer*, vol. 65, no. 11, pp. 2539–2553, 1990.
- [3] A. Luetke, P. A. Meyers, I. Lewis, and H. Juergens, "Osteosarcoma treatment - where do we stand? A state of the art review," *Cancer Treatment Reviews*, vol. 40, no. 4, pp. 523–532, 2014.
- [4] P. Picci, "Osteosarcoma (osteogenic sarcoma)," *Orphanet Journal of Rare Diseases*, vol. 2, no. 1, p. 6, 2007.
- [5] W. Wang, H. Shen, G. Cao, and J. Huang, "Long non-coding RNA Xist predicts poor prognosis and promotes malignant phenotypes in osteosarcoma," *Oncology Letters*, vol. 17, no. 1, pp. 256–262, 2019.
- [6] R. E. Windsor, S. J. Strauss, C. Kallis, N. E. Wood, and J. S. Whelan, "Germline genetic polymorphisms may influence chemotherapy response and disease outcome in osteosarcoma: a pilot study," *Cancer*, vol. 118, no. 7, pp. 1856–1867, 2012.
- [7] W. Liu, G. Xu, H. Liu, and T. Li, "MicroRNA-490-3p regulates cell proliferation and apoptosis by targeting Hmga2 in osteosarcoma," *FEBS Lett*, vol. 589, no. 20PartB, pp. 3148–3153, 2015.
- [8] J. Cui, W. Wang, Z. Li, Z. Zhang, B. Wu, and L. Zeng, "Epigenetic changes in osteosarcoma," *Bulletin du Cancer*, vol. 98, no. 7, pp. E62–E68, 2011.
- [9] K. Rao-Bindal and E. S. Kleinerman, "Epigenetic regulation of apoptosis and cell cycle in osteosarcoma," *Sarcoma*, vol. 2011, Article ID 679457, 5 pages, 2011.
- [10] M. Kim, J. Y. Jung, S. Choi et al., "Gfra1 promotes cisplatin-induced chemoresistance in osteosarcoma by inducing autophagy," *Autophagy*, vol. 13, no. 1, pp. 149–168, 2017.
- [11] S. V. Iyer, A. Ranjan, H. K. Elias et al., "Genome-wide RNAi screening identifies TMIGD3 isoform1 as a suppressor of NF- κ B and osteosarcoma progression," *Nature Communications*, vol. 7, no. 1, p. 13561, 2016.
- [12] F. Neri, S. Rapelli, A. Krepelova et al., "Intragenic DNA methylation prevents spurious transcription initiation," *Nature*, vol. 543, no. 7643, pp. 72–77, 2017.
- [13] N. Nishida, T. Nagasaka, T. Nishimura, I. Ikai, C. R. Boland, and A. Goel, "Aberrant methylation of multiple tumor suppressor genes in aging liver, chronic hepatitis, and hepatocellular carcinoma," *Hepatology*, vol. 47, no. 3, pp. 908–918, 2008.
- [14] C. G. Ekmekci, M. I. Gutierrez, A. K. Siraj, U. Ozbek, and K. Bhatia, "Aberrant methylation of multiple tumor suppressor genes in acute myeloid leukemia," *American Journal of Hematology*, vol. 77, no. 3, pp. 233–240, 2004.
- [15] Q. Li, H. Li, X. Zhao et al., "DNA methylation mediated down-regulation of Mir-449c controls osteosarcoma cell cycle progression by directly targeting oncogene C-Myc," *International Journal of Biological Sciences*, vol. 13, no. 8, pp. 1038–1050, 2017.
- [16] J. Lu and J. Wang, "Irx1 hypomethylation in osteosarcoma metastasis," *Oncotarget*, vol. 6, no. 19, pp. 16802–16803, 2015.
- [17] L. Li, C. Xu, P. Liu, and J. Huang, "Correlation study of DNA methylation of Wnt6 gene with osteosarcoma in children," *Oncology Letters*, vol. 14, no. 1, pp. 271–275, 2017.
- [18] V. Sonaglio, A. C. de Carvalho, S. R. Toledo et al., "Aberrant DNA methylation of Esr1 and P14arf genes could be useful as prognostic indicators in osteosarcoma," *Oncotargets and Therapy*, vol. 6, pp. 713–723, 2013.
- [19] S. P. Wu, B. T. Cooper, F. Bu et al., "DNA methylation-based classifier for accurate molecular diagnosis of bone sarcomas," *JCO Precision Oncology*, vol. 2017, 2017.
- [20] E. P. Buddingh, S. E. N. Ruslan, C. M. A. Reijnders et al., "Mesenchymal stromal cells of osteosarcoma patients do not show evidence of neoplastic changes during long-term culture," *Clinical Sarcoma Research*, vol. 5, no. 1, p. 16, 2015.
- [21] G. Dennis Jr., B. T. Sherman, D. A. Hosack et al., "David: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 5, p. 3, 2003.
- [22] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [23] J. C. Spainhour, H. S. Lim, S. V. Yi, and P. Qiu, "Correlation patterns between DNA methylation and gene expression in The Cancer Genome Atlas," *Cancer Informatics*, vol. 18, p. 117693511982877, 2019.
- [24] A. Ferrari, U. Dirksen, and S. Bielack, "Sarcomas of soft tissue and bone," *Progress in Tumor Research*, vol. 43, pp. 128–141, 2016.
- [25] S. Morice, G. Danieau, F. Redini, B. Brounais-Le-Royer, and F. Verrecchia, "Hippo/Yap signaling pathway: a promising therapeutic target in bone paediatric cancers?," *Cancers (Basel)*, vol. 12, no. 3, p. 645, 2020.
- [26] S. L. Teoh and S. Das, "The emerging role of the hippo pathway in lung cancers: clinical implications," *Current Drug Targets*, vol. 18, no. 16, pp. 1880–1892, 2017.
- [27] H. Wang, Y. C. Du, X. J. Zhou, H. Liu, and S. C. Tang, "The dual functions of yap-1 to promote and inhibit cell growth in human malignancy," *Cancer Metastasis Reviews*, vol. 33, no. 1, pp. 173–181, 2014.

- [28] Y. A. Chen, C. Y. Lu, T. Y. Cheng, S. H. Pan, H. F. Chen, and N. S. Chang, "Ww domain-containing proteins yap and Taz in the hippo pathway as key regulators in stemness maintenance, tissue homeostasis, and tumorigenesis," *Frontiers in Oncology*, vol. 9, p. 60, 2019.
- [29] P. McQueen, S. Ghaffar, Y. Guo, E. M. Rubin, X. Zi, and B. H. Hoang, "The Wnt signaling pathway: implications for therapy in osteosarcoma," *Expert Review of Anticancer Therapy*, vol. 11, no. 8, pp. 1223–1232, 2014.
- [30] S. Li, Z. Chai, Y. Li et al., "Bzw1, a novel proliferation regulator that promotes growth of salivary mucoepidermoid carcinoma," *Cancer Letters*, vol. 284, no. 1, pp. 86–94, 2009.
- [31] A. Angulo, M. Cuenca, P. Martinez-Vicente, and P. Engel, "Viral Cd229 (Ly9) homologs as new manipulators of host immunity," *Journal of Leukocyte Biology*, vol. 105, no. 5, pp. 947–954, 2019.
- [32] S. C. Ozcan, A. Sarioglu, T. H. Altunok et al., "Pfkfb2 regulates glycolysis and proliferation in pancreatic cancer cells," *Molecular and Cellular Biochemistry*, vol. 470, no. 1-2, pp. 115–129, 2020.
- [33] Q. Cheng and L. Wang, "LncRNA XIST serves as a ceRNA to regulate the expression of ASF1a, BRWD1M, and PFKFB2 in kidney transplant acute kidney injury via sponging hsa-miR-212-3p and hsa-miR-122-5p," *Cell Cycle*, vol. 19, no. 3, pp. 290–299, 2020.
- [34] H. Liu, K. Chen, L. Wang et al., "miR-613 inhibits Warburg effect in gastric cancer by targeting PFKFB2," *Biochemical and Biophysical Research Communications*, vol. 515, no. 1, pp. 37–43, 2019.
- [35] M. Camargo Barros-Filho, L. Barreto Menezes de Lima, M. Bisarro dos Reis et al., "Pfkfb2 promoter hypomethylation as recurrence predictive marker in well-differentiated thyroid carcinomas," *International Journal of Molecular Sciences*, vol. 20, no. 6, p. 1334, 2019.
- [36] Z. He, C. You, and D. Zhao, "Long non-coding RNA UCA1/miR-182/Pfkfb2 axis modulates glioblastoma-associated stromal cells-mediated glycolysis and invasion of glioma cells," *Biochemical and Biophysical Research Communications*, vol. 500, no. 3, pp. 569–576, 2018.
- [37] A. Sreedhar, P. Petruska, S. Miriyala, M. Panchatcharam, and Y. Zhao, "Ucp2 overexpression enhanced glycolysis via activation of PFKFB2 during skin cell transformation," *Oncotarget*, vol. 8, no. 56, pp. 95504–95515, 2017.

Research Article

A Comparative Analysis of Visual Encoding Models Based on Classification and Segmentation Task-Driven CNNs

Ziya Yu , Chi Zhang , Linyuan Wang , Li Tong , and Bin Yan 

PLA Strategy Support Force Information Engineering University, Zhengzhou 450001, China

Correspondence should be addressed to Bin Yan; ybspace@hotmail.com

Received 14 March 2020; Revised 31 May 2020; Accepted 6 June 2020; Published 1 August 2020

Guest Editor: Lin Lu

Copyright © 2020 Ziya Yu 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.

Nowadays, visual encoding models use convolution neural networks (CNNs) with outstanding performance in computer vision to simulate the process of human information processing. However, the prediction performances of encoding models will have differences based on different networks driven by different tasks. Here, the impact of network tasks on encoding models is studied. Using functional magnetic resonance imaging (fMRI) data, the features of natural visual stimulation are extracted using a segmentation network (FCN32s) and a classification network (VGG16) with different visual tasks but similar network structure. Then, using three sets of features, i.e., segmentation, classification, and fused features, the regularized orthogonal matching pursuit (ROMP) method is used to establish the linear mapping from features to voxel responses. The analysis results indicate that encoding models based on networks performing different tasks can effectively but differently predict stimulus-induced responses measured by fMRI. The prediction accuracy of the encoding model based on VGG is found to be significantly better than that of the model based on FCN in most voxels but similar to that of fused features. The comparative analysis demonstrates that the CNN performing the classification task is more similar to human visual processing than that performing the segmentation task.

1. Introduction

Complex neural circuits in the human brain allow us to easily understand the external visual world. However, the mechanisms of how visual areas encode visual stimuli have not yet been elucidated. Therefore, the development of a visual encoding model to predict the voxel response induced by any input stimulus, that is, simulating the complex nonlinear relationship between visual input and evoked voxel responses, has attracted wide attention [1, 2]. It can explain how the brain processes visual information through neural circuits [3]. In visual research based on functional magnetic resonance imaging (fMRI), linearized encoding has been widely applied to these models. It consists of a nonlinear mapping from visual stimuli to features and a linear mapping from features to voxel responses [4]. Nonlinear mapping is critical to visual encoding that can be implemented by various feature extractors such as Gabor wavelet pyramid (GWP) [5], histogram of oriented gradient (HOG) [6], local binary patterns (LBP) [7], scale-invariant feature transform

(SIFT) [8], and convolution neural networks (CNNs). On the other hand, linear mapping generally uses linear regression models with specific regularization.

In recent years, CNNs have been widely used in visual encoding models. CNNs, proposed based on early discoveries of the network structure and the visual system [9], can be used in a variety of computer vision tasks such as image classification [10], target recognition [11], and semantic segmentation [12]. Studies have shown that a deep network is comparable to the human visual system, which can automatically learn effective features from large data for specific tasks and predict voxel responses measured by fMRI in a multilevel manner [13]. Agrawal et al. [14] first proposed a CNN to predict human brain activity based on low-level visual input (pixels). Güçlü and van Gerven [15, 16] illustrated the similarity between a CNN and the mechanism of visual processing in both the ventral visual pathway, which is responsible for object recognition, and the dorsal visual pathway, which is responsible for motion perception. These studies demonstrated that a CNN is similar to a visual

pathway from a low level to a high level. Wen et al. [17] established an encoding model based on the deep residual network (DRN), which has been shown to perform better than the shallow AlexNet for video stimuli. Their study showed that improvements in prediction accuracy are due to the better feature expression of the deep network with a residual structure. Therefore, in computer vision, the choice of a network to obtain suitable feature transformations is critical, which directly influences the encoding performance [18].

In 2016, Yamins and DiCarlo [19] proposed a particularly important challenge, which was whether a model optimized for tasks other than classification can better explain neural data. In particular, task-driven deep networks performing different computer vision tasks can extract different features from the same image stimuli, resulting in variations in the performance of encoding models. Currently, studies on encoding models based on deep networks are limited to the visual classification task, which is different from the complexity and diversity of the human visual system.

Here, we explore the impact of network tasks on the performance of encoding models by building models based on the features extracted from a segmentation network, features extracted from a classification network, and the fusion of the two features. We use the largest dataset in the published dataset, BOLD5000 [20], to train and test the encoding model. We calculate the Pearson Correlation Coefficient between the predicted and experimental fMRI responses to compare the prediction performances of the three encoding models. Using the results, we describe the impact of changes in network tasks on the visual encoding model. We then discuss the advantages and disadvantages of simulating the human visual processing.

In this study, our main contributions are as follows: (1) we analyze the drawbacks of current encoding methods based on the complexity and diversity of the human visual system, (2) we propose to employ different task-driven networks to construct encoding models, and (3) we analyze the impact of different task-driven networks on the performance of encoding models and provide a possible direction for subsequent research on visual encoding.

2. Materials and Methods

2.1. Experimental Data. We used the public fMRI dataset, BOLD5000 [20], which can be downloaded from <https://bold5000.github.io/download.html>. Details of the visual stimuli and fMRI protocols of the dataset have been discussed elsewhere [20]. Hence, we only briefly summarize the details of the dataset in this subsection.

The dataset comprised fMRI data collected from four subjects, with three having a full set of data. Hence, we only used the data of three subjects. A full set of data included 16 MRI scan sessions, with 15 functional sessions and a session for the acquisition of high-resolution anatomical and diffusion data. Each functional session lasted 1.5 hours, consisting of 8 sessions with 9 image runs and an additional functional localizer run and 7 sessions with 10 image runs.

The stimuli included 5254 images, 4916 of which were unique. The images were obtained from three computer

vision datasets: Scene UNderstanding (SUN) [21], Common Objects in Context (COCO) [22], and ImageNet [23]. They were downsampled to 375×375 pixels and subtended a visual angle of approximately 4.6 degrees. The stimuli were presented using an event-related design. Each run comprised 37 stimuli, with approximately 2 from repeated images. Each image was presented for 1 second followed by a fixation cross for 9 seconds. At the beginning and end of each run, a fixation cross was displayed for 6 seconds and 12 seconds, respectively. fMRI data were acquired using a 3 T Siemens Verio MR scanner at the Carnegie Mellon University campus with a 32-channel phased array head coil. The repetition time (TR) was 2000 ms, the echo time (TE) was 30 ms, the field of view was 212 mm, and the slice thickness was 2 mm.

The data we used covered five visual areas in the human visual cortex, i.e., early visual area (EarlyVis), the lateral occipital complex (LOC), the occipital place area (OPA), the parahippocampal place area (PPA), and the retrosplenial complex (RSC). Note that different visual areas perform different visual functions. EarlyVis in this dataset goes beyond the typical V1 and V2 areas. Human visual cortex V1 is mainly responsible for the detection of local features and provides this information to the middle or even higher visual areas [24, 25]. V2 has a slightly complex modulation for positioning, spatial frequency color, and moderate modulation for complex shape [26, 27]. The other four areas belong to advanced visual areas, which perform more complex visual tasks such as perceiving the boundaries of a scene [28], processing shape [29], encoding and recognizing an environmental scene [30], and dealing with scenarios [31].

2.2. Overview of the Proposed Method. In general, linearized encoding adopts a two-step strategy, requiring two computational models to encode voxels. The first one is feature transformation, which is a nonlinear mapping from input space to feature space using feature extractors. The other is a linear regression model, which is a linear mapping from feature space to voxel space. The parameters of the feature transformation model are typically fixed and do not need further training. On the other hand, the linear weights of the linear regression model need to be trained. In this paper, we constructed CNN-based visual encoding models that use the classification network VGG and the segmentation network FCN to extract features of the input stimuli. Figure 1 shows the overall process.

5254 natural images were randomly divided into a group of 4754 images and a group of 500 images. Two groups of the images and their corresponding voxel responses were considered, with one group used as the training set and the other as the test set. We employed pretrained VGG16 and FCN32s to accomplish feature transformation and then used ROMP to construct a linear regression model. We mapped the features extracted by the two networks and the fused features to the voxel responses of visual areas to learn the weight coefficients. Hence, we attained three encoding models based on different CNN features. The encoding models were then tested on the test set to obtain the prediction accuracy for each voxel. Here, we defined the prediction accuracy as the Pearson Correlation Coefficient between the observed and

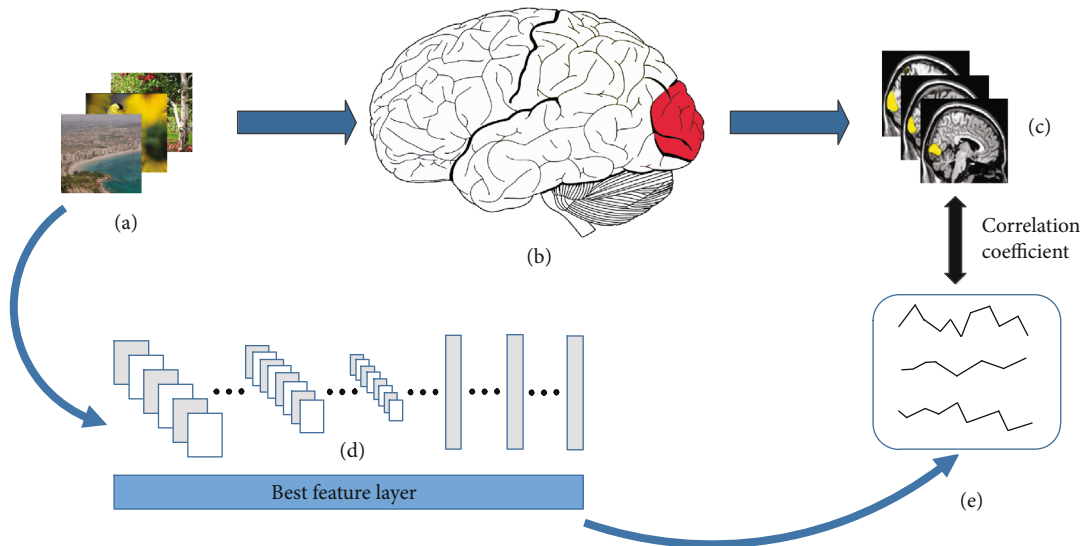


FIGURE 1: Main process of visual encoding. (a) Natural image stimuli; (b) visual processing of human brain; (c) real fMRI responses obtained by an MRI scanner; (d) CNN features of natural images extracted by pretrained CNN; (e) predicted voxel responses. When subjects saw the visual stimuli, the corresponding brain signals would be generated in the visual areas of the brain, and the fMRI responses were obtained through the MRI scanner. Using the pretrained network to extract the features of natural images, the CNN features of each layer were linearly mapped to voxel space, and the feature layer with the best prediction performance was selected as the best encoding feature layer to obtain predicted voxel responses. Then, the correlation coefficient between predicted responses and real responses was calculated to evaluate the prediction performance of the encoding model.

predicted responses across the test set. A high correlation coefficient corresponds to a high prediction accuracy of the encoding model, which means that the features and voxel responses are more linearly related.

2.3. Extracting Hierarchical Visual Features Based on VGG16.

To extract the features of natural images using a classification network, we employed the pretrained model of VGG16 based on the open-source deep learning framework of PyTorch [32]. VGG16, which is a classification model proposed by Oxford University in 2014 [33], comprises of 16 hidden layers (13 convolutional layers and 3 fully connected layers). Each artificial neuron in the convolutional layer corresponds to a feature detector, called a feature map, which represents the characteristics of the input stimuli. Each convolutional layer has 64, 64, 128, 128, 256, 256, 256, 256, 512, 512, 512, 512, 512, 512, 512, and 1000 (class of the dataset) kernels. In order to make the gradient descent and reverse propagation more effective, the activation function called the Rectified Linear Unit (ReLU) [34] is used by the artificial neurons at layers 1–15. The pooling layer that reduces redundancy can also be interpreted as a form of the nonlinear downsampling operation. The most common form of pooling layer is maximum pooling and average pooling. In the VGG16 architecture, layers 2, 4, 7, 10, and 13 adopt maximum pooling, while layers 14 and 15 adopt a nonlinear transformation to eliminate regularization. The architecture of VGG16 is shown in Table 1.

2.4. Extracting Hierarchical Visual Features Based on FCN32s.

To extract the features of natural images using a segmentation network, we employed pretrained FCN32s for semantic segmentation [11]. In this structure, the FCN

converts fully connected layers into convolutional layers. The output image of the last layer is sampled 32 times to obtain the image with the same size as the original input image. FCN32s initializes the network with the structural parameters of VGG16, discards the final classification layer, and converts all fully connected layers into convolutional layers; hence, it is called a fully convolutional network. We used a 1×1 convolution with a channel size of 21 to predict the score of each location (including background) of the Pascal class. Then, the deconvolution layer was added to the output at the pixel level to sample the output upwards.

FCN32 comprises 16 convolutional layers with each having 64, 64, 128, 128, 256, 256, 256, 256, 512, 512, 512, 512, 512, 4096, 4096, and 21 (class of the dataset) kernels. In the architecture of FCN32s, ReLU is used in layers 1–16. Layers 2, 4, 7, 10, and 13 adopt maximum pooling, while layers 14 and 15 adopt dropout regularization to realize nonlinear transformation. The architecture of FCN32s is shown in Table 1.

The FCN32s architecture we employed released in 2017 is available at <https://github.com/meetshah1995/pytorch-semseg>. We trained the FCN32s on 2913 high-resolution images from the Pascal-VOC 2012 dataset for semantic segmentation using PyTorch [32]. Each input image was represented as three RGB color channels and filtered through the convolutional layers. The stride of the convolutional layers was 3 pixels at layers 1–13, 7 pixels at layer 14, 1 pixel at layer 15, and 21 pixels at layer 16. In the training process, we adopted momentum and weight attenuation for random gradient descent. The learning rate was initialized to 0.01, and the final intersection over union (IOU) was 0.59. We trained FCN32s on the segmentation dataset to obtain a pretrained network for feature extraction of the encoding model.

TABLE 1: The layer index and corresponding layer names of VGG16 and FCN32s.

Index	1	2	3	4	5	6	7	8
Layer name of VGG16	conv1	conv2 mpool	conv3	conv4 mpool	conv5	conv6	conv7 mpool	conv8
Layer name of FCN32s	conv1	conv2 mpool	conv3	conv4 mpool	conv5	conv6	conv7 mpool	conv8
Index	9	10	11	12	13	14	15	16
Layer name of VGG16	conv9	conv10 mpool	conv11	conv12	conv13 mpool	fc1	fc2	fc3
Layer name of FCN32s	conv9	conv10 mpool	conv11	conv12	conv13 mpool	conv14	conv15	conv16

2.5. *Training the Mapping from the Features to Voxel Responses Based on Sparse Representation.* Corresponding to each layer of the CNN features, a linear model can be constructed to map CNN features into voxel responses of the visual areas. For responses of one voxel to all training samples, the model can be expressed by

$$y = Xw + \varepsilon. \quad (1)$$

Here, y is the measured voxel responses represented by an $m - by - 1$ matrix, where m is the number of training samples; X is the CNN features of images represented by an $m - by - (n + 1)$ matrix, where n is the dimension of features and the last column is the constant vector; w is the weight coefficient to be solved represented by an $(n + 1) - by - 1$ matrix; and ε is the noise term.

However, the number of training samples m is significantly smaller than the number of voxels n in visual areas. Hence, Equation (1) is an ill-posed equation without a unique solution. In addition, in several studies [35, 36], the visual cortex uses sparse coding for the expression of stimuli, which means that a specific stimulus can only activate a few specific visual neurons. Hence, sparse representation can be used as an effective tool to encode information related to natural images. Considering a sparse coefficient w , Equation (1) is converted into a traditional sparse representation problem, which is typically considered as an NP-Hard problem, defined as follows:

$$\min_w \|w\|_0 \text{ subject to } Xw = y. \quad (2)$$

To approximate the solution of Equation (2), we used the greedy algorithm [37], which follows the heuristic of making the locally optimal choice at each stage with the intent of finding a global optimum, which is quite fast by computing the support of the sparse signal iteratively [38]. Considering that the encoding model must be estimated for each voxel, the method we need to employ should be fast enough and simple to reduce the time cost. Therefore, we used the greedy algorithm to investigate the sparseness of the encoding model, in particular, the ROMP algorithm.

ROMP is an iterative fitting technique that reduces the difference between model fit and data [39, 40]. The specific calculation process is shown in Algorithm 1.

The features of each layer in FCN32 and VGG16 on the training set were mapped to the voxel space by ROMP, and the weight coefficients were obtained. Here, the coefficient of the final nonzero term was 100. Then, the predicted voxel

responses for the test set were obtained through the weight coefficients of each layer. We compared the correlation between the predicted responses and the measured voxel responses. Based on the correlation coefficient, the highest prediction accuracy was selected for each voxel; that is, the feature layer with the highest correlation was taken as the best feature layer for each voxel. The linear mapping from the best feature layer to the voxel response was added to obtain the voxel-wise encoding model.

2.6. *Combined Encoding Model Based on the Fusion of Features.* To fit the diversity of the mechanism of human vision, we fused some image features extracted from the classification and segmentation networks and established an encoding model based on the fused features. Firstly, we employed the ROMP algorithm to construct a linear mapping from the voxel responses to all the image features extracted from the FCN32s and VGG16 on the training set and obtained the predicted image features on the test set. For the specific one-dimensional feature on a certain layer of CNN, the model can be expressed by

$$y_1 = X_1 w_1 + \varepsilon_1. \quad (3)$$

Here, y_1 is the CNN features of images represented by an $p - by - 1$ matrix, where p is the number of training samples; X_1 is the measured voxel responses represented by an $p - by - (q + 1)$ matrix, where q is the number of voxels and the last column is the constant vector; w_1 is the weight coefficient to be solved represented by a $(q + 1) - by - 1$ matrix; and ε_1 is the noise term.

To reduce the influence of ineffective features, we calculated correlation coefficients between the predicted and real image features. According to the ranking of correlation coefficients from largest to smallest, the corresponding image features of the first 10% dimension (including the part of image features extracted by FCN32s and VGG16) were selected at each layer.

After feature selection, visual encoding was carried out according to the method mentioned in Training the Mapping from the Features to Voxel Responses Based on Sparse Representation. The features of selected dimensions were extracted from the training set and linearly mapped to the voxel responses by ROMP. Then, the predicted voxel responses based on different feature layers (including image features extracted from FCN32s and VGG16) were obtained by using the calculated weights. For each voxel, the feature with the highest prediction accuracy was selected as the best feature layer, and the voxel-wise visual encoding model was established.

ROMP algorithm.

Input: observation matrix \mathbf{X} (specific features of a certain layer of CNN), observation vector y (voxel responses), sparsity parameter p ;

Output: weight vector w ;

Process:

1. Initialization
Initialize the atomic support set $A = \emptyset$, residual $r_0 = y$, and repeat the following steps s times;
2. Atomic selection
Select the column index of the top n maximum or all non-zero values (the number of non-zero coordinates is less than p) in $u = abs[X^T r]$, and form an atomic support set \mathbf{J} ;
3. Regularization
Find a subset in the set \mathbf{J} so that any two inner product u_i and u_j satisfy $|u_i| \leq 2|u_j|$, and select the subset \mathbf{J}_0 with the maximum energy $\sum_j |u_j|^2, j \in \mathbf{J}_0$ among the subsets that satisfy the condition;
4. Update atomic support set and residual
 $A \leftarrow A \cup \mathbf{J}_0$. Update the residual: $\hat{w} = \arg \min_z \|y - \mathbf{X}|_A z\|_2$; $r = y - \mathbf{X}\hat{w}$, and return to the second step.

ALGORITHM 1

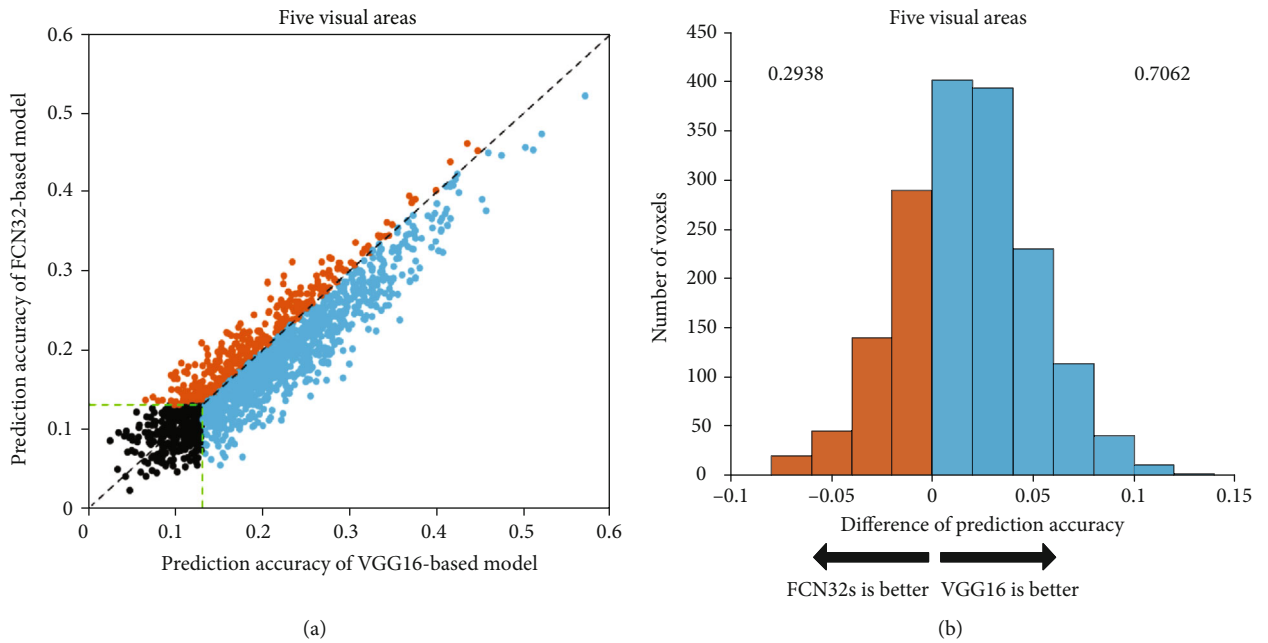


FIGURE 2: Comparison of the prediction accuracy between FCN32-based and VGG16-based encoding models in five visual areas of subject 1. (a) Prediction accuracies. The abscissa and ordinate represent the prediction accuracy of the FCN32-based encoding model and the VGG16-based encoding model, respectively. The orange dots represent the voxels that can be better predicted by the FCN32-based model than the VGG16-based model. The blue dots represent the opposite. And the black dots represent voxels with prediction accuracy less than 0.13. The green dashed lines indicate that the prediction accuracy is 0.13. (b) Distribution of the difference in prediction accuracies. The blue color denotes that the prediction accuracy is higher for the VGG16-based model. The orange color denotes that the prediction accuracy is higher for the FCN32-based model. The numbers on each side indicate the fraction of voxels with higher prediction accuracy under the model.

2.7. *Quantitative Standards.* We define the prediction accuracy for a voxel as a Pearson Correlation Coefficient between the measured and the predicted responses across all 500 images in the test set:

$$r = \frac{\text{cov}(v_p, v_m)}{\sqrt{\text{var}[v_p][v_m]}}. \quad (4)$$

In Equation (4), v_p represents the predicted voxel responses, v_m represents the measured voxel responses in the

test set, and r represents the correlation coefficient between them, i.e., the encoding accuracy.

To examine whether each voxel's prediction accuracy value significantly deviated from the null hypotheses, we randomly shuffled the pairing between measured and predicted responses across 500 images in the test set 1000 times and in each randomized sample recalculated the voxel's prediction. This calculation constructed a null hypothesis distribution for each voxel. For all voxels, the prediction accuracy value above 0.13 was significant ($p < 0.001$) relative to its null hypothesis distribution.

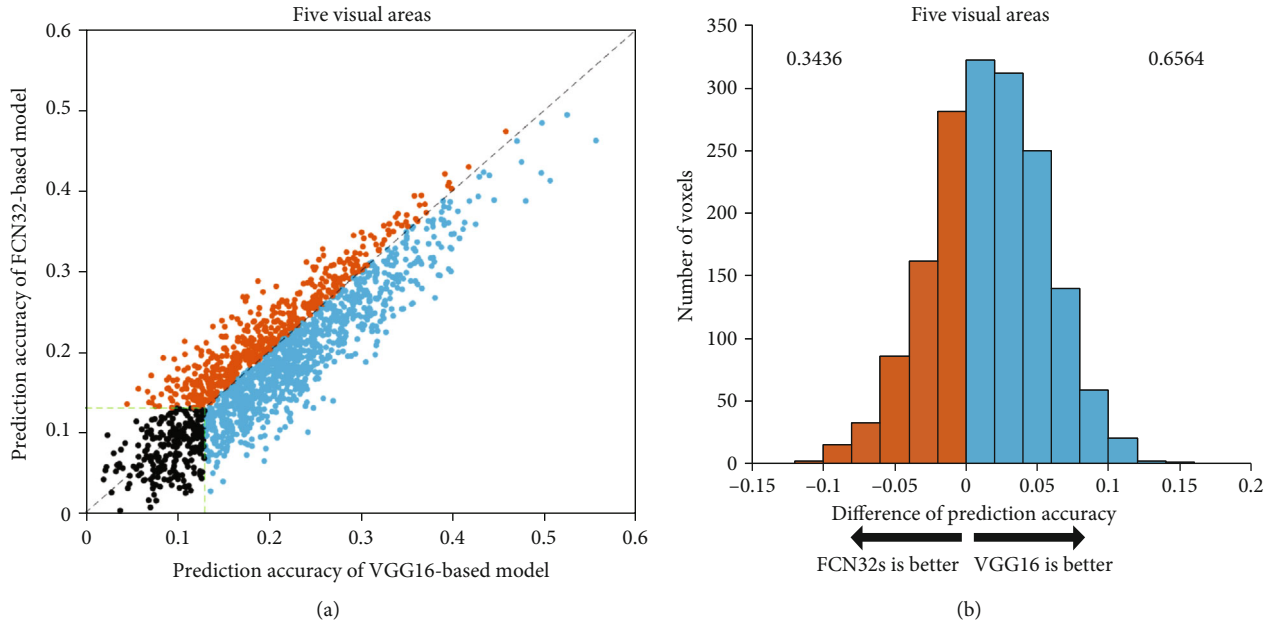


FIGURE 3: Comparison of prediction accuracy between FCN32-based and VGG16-based models for subject 2. Refer to Figure 2 for a detailed description of the plot elements.

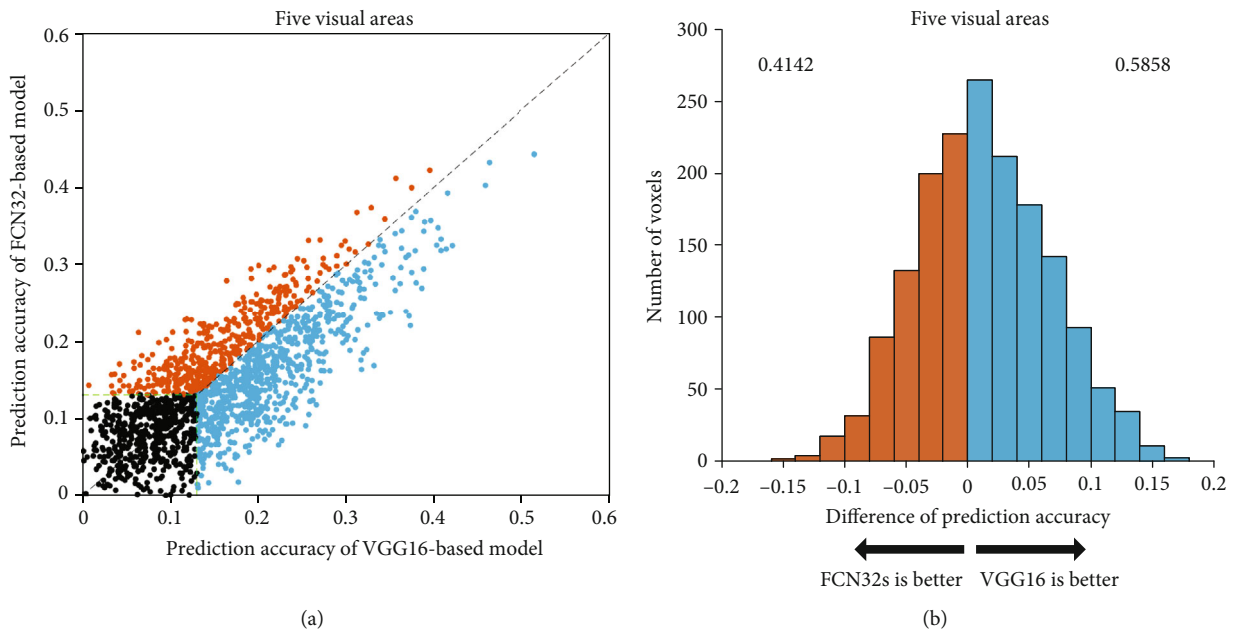


FIGURE 4: Comparison of prediction accuracy between VGG16-based and FCN32-based models for subject 3. Refer to Figure 2 for a detailed description of the plot elements.

To examine the significance of a model advantage, that is, the number of voxels that can be predicted by the model that is significantly more than that of the other, we randomly permuted (with a probability of 50%) the prediction accuracy of each voxel of the two models being compared and then calculated the advantage of each model (the percentage of voxels with the highest prediction accuracy). In this paper, we repeated such permutations 1000 times, and null hypothesis distribution was obtained. From the null hypothesis distribution, it is concluded that for any two models, the model

which can accurately predict more than 53% of voxel responses is significantly better than the other model ($p < 0.05$).

3. Results

3.1. Comparison of Prediction Accuracy

3.1.1. Comparison of VGG16-Based and FCN32-Based Encoding Models. To evaluate the encoding capabilities of different networks based on different training tasks, we

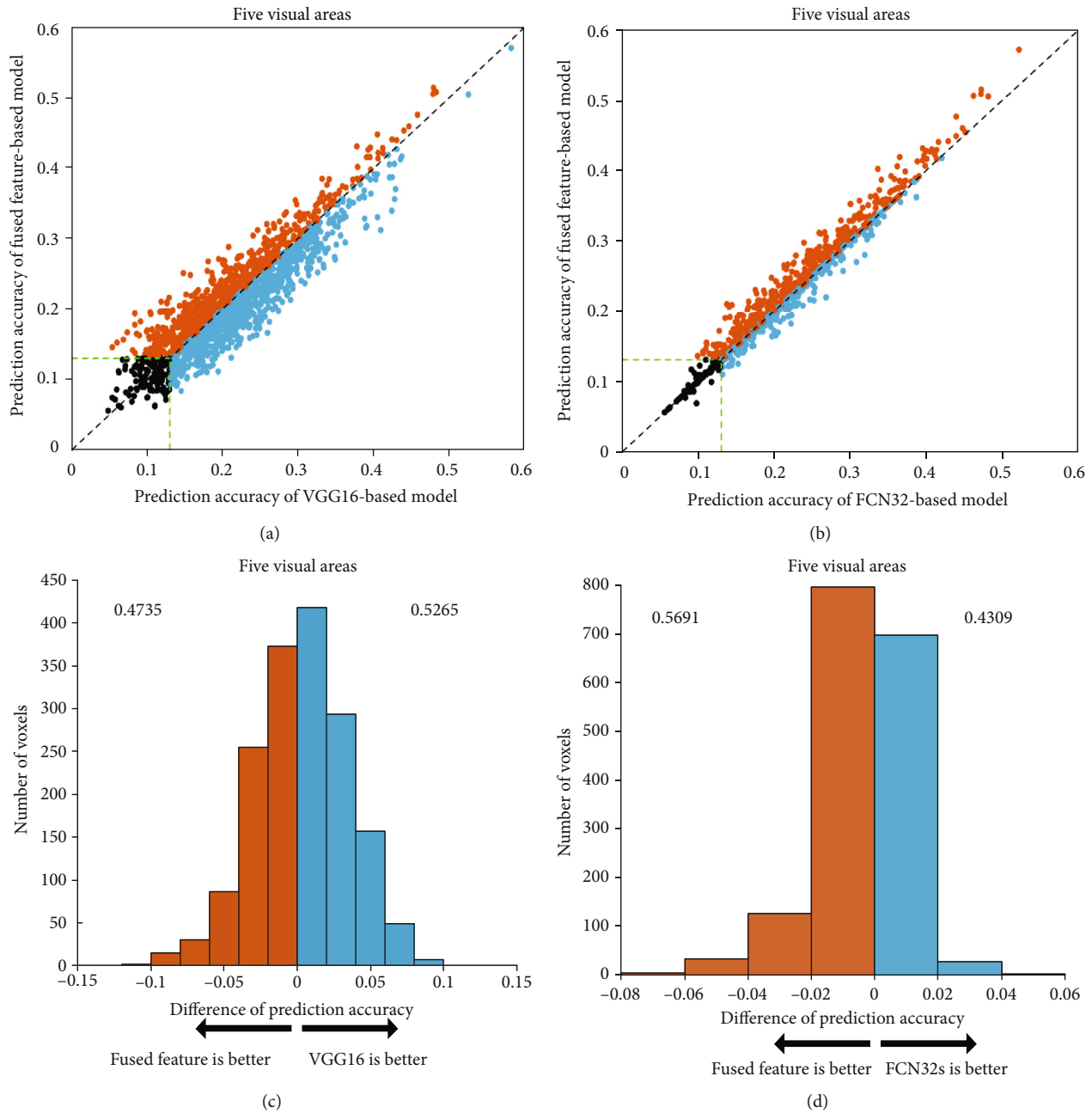


FIGURE 5: Comparison of the prediction accuracy between the fused feature-based encoding model and (a) VGG16-based or (b) FCN32-based encoding models in five visual areas. The ordinate represents the prediction accuracy of the fused feature-based encoding model, and the abscissas represent the prediction accuracy of the VGG16-based or FCN32-based encoding models. The orange dots represent the voxels that can be better predicted by the fused feature-based model than the VGG16-based or FCN32-based models. The blue dots represent the opposite. The green dashed lines and the black dots represent the same meanings as Figure 2. (c) Distribution of the difference between fused features and VGG16 or (d) FCN32-based model in prediction accuracies. The blue color denotes that the prediction accuracy is higher for the VGG16-based model or FCN32-based model. The orange color denotes that the prediction accuracy is higher for the fused feature-based model. The numbers on each side indicate the fraction of voxels with higher prediction accuracy under the model.

calculated the prediction accuracy of voxels in five ROIs based on two encoding models: classification network and segmentation network. We used a scatter plot to compare the accuracy of the two models and analyze their performances. Each plot represents a single voxel from the five ROIs. The ordinate of each point represents the highest

encoding accuracy of the FCN32s model, while the abscissa represents the highest encoding accuracy of the VGG16 model. Here, the correlation threshold for significance prediction is 0.13 ($p < 0.001$). The results show that the prediction accuracy of the encoding model based on VGG16 is better than that of the encoding model based on FCN32s.

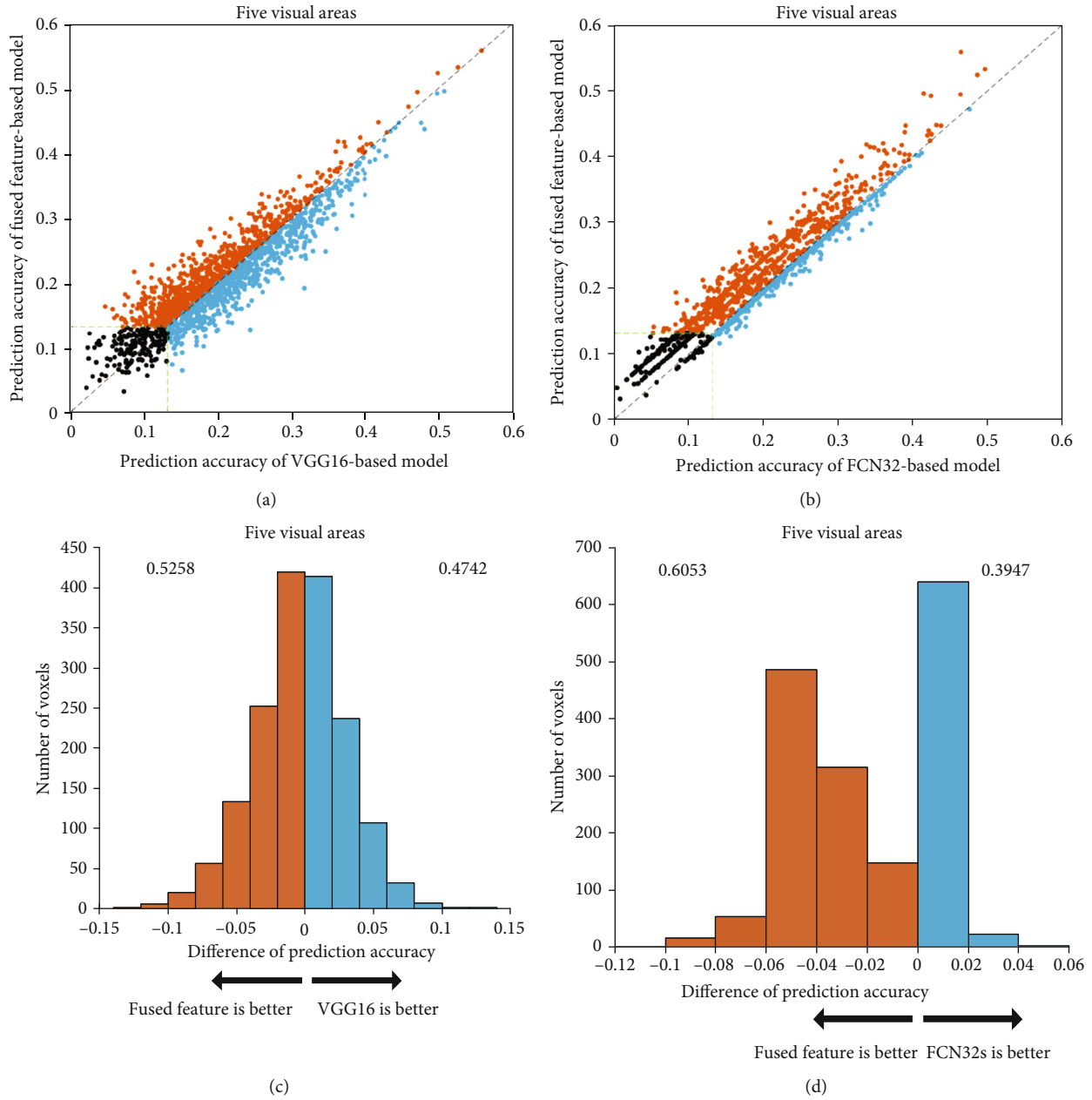


FIGURE 6: Comparison of prediction accuracy between fused feature-based and VGG16-based or FCN32-based models for subject 2. Refer to Figure 5 for a detailed description of the plot elements.

Figure 2 show the results for subject 1, while the results for subjects 2 and 3 are presented in Figures 3 and 4.

The VGG16-based model has significant advantages over the FCN32-based model in the five visual areas ($p < 0.05$). The results show that the encoding performance of the network based on classification features is significantly better than that of the network based on segmentation features, which indicates that different network tasks can affect the performance of the encoding model. However, some voxels have better prediction accuracy in the FCN32-based model than in the VGG16-based model, which indicates that there are still inconsistencies between the classification or segmentation networks and the visual encoding mechanism of the human brain.

3.1.2. Comparison between VGG16-Based, FCN32-Based, and Fused Feature-Based Encoding Models. To explore the relationship between segmentation features and classification features in visual encoding, i.e., the intersection and union of classification and segmentation tasks in the human visual system, we compared the prediction performance of the encoding model based on fused features with that of the VGG16-based and FCN32-based encoding models. The results shown in Figure 5 are used to compare the accuracy of the three models and analyze their performances.

Consistent with the results of subject 2 and subject 3 in Figures 6 and 7, the prediction performance of the fused feature-based encoding model is significantly better than that

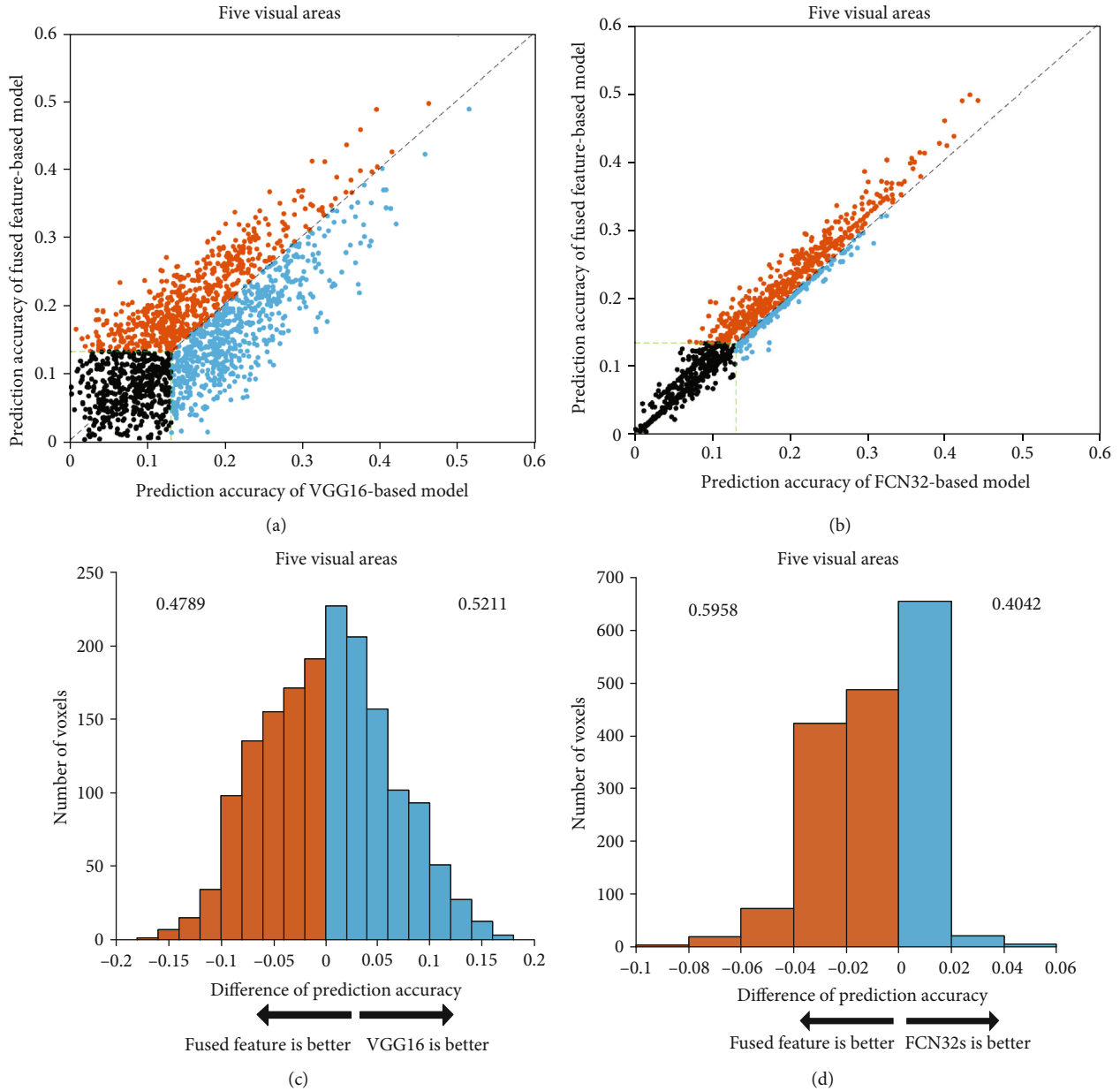


FIGURE 7: Comparison of prediction accuracy between fused feature-based and VGG16-based or FCN32-based models for subject 3. Refer to Figure 5 for a detailed description of the plot elements.

of the FCN32-based encoding model ($p < 0.05$), while it is slightly different from that of the VGG16-based encoding model. To a certain extent, this indicates that the fused features can significantly improve the prediction performance of the encoding model compared with the segmentation features but have little effect compared with the classification features. In other words, in the process of the human visual system perceiving external stimuli, the classification task performed by the visual areas covers most of the segmentation task; that is, in the process of completing the classification of external objects, the segmentation of objects is basically completed, which means that people can recognize the category, size, and location of objects almost at the same time when they see a picture.

3.2. Relationship between Feature Quantity and Prediction Accuracy. We compared and analyzed the influence of the number of features on the encoding performance for subject 1, as shown in Figure 8. Results for subject 2 and subject 3 are shown in Figures 9 and 10. The results show that too few or too many features can negatively impact the performance of the encoding model. In particular, a small number of features lead to the lack of effective information, while a high number of features lead to redundancy of effective information.

3.3. Contribution of Each CNN Layer to Prediction Performance. To further compare the encoding differences of different networks based on different training tasks as feature models and verify the hierarchical similarity between

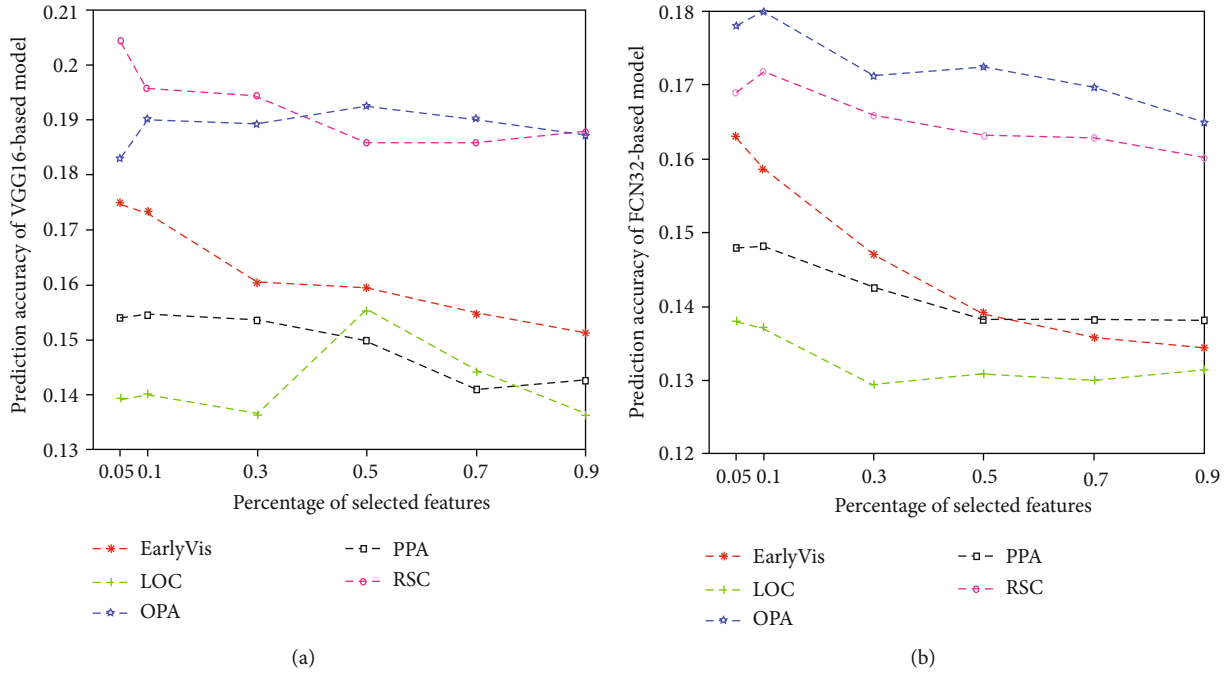


FIGURE 8: Relationship between the percentage of selected features and prediction accuracy of two encoding models in five visual areas: (a) VGG-based model and (b) FCN32-based model. The abscissa represents the percentage of selected features, and the ordinate represents the prediction accuracy of the models. The lines represent the results for five different visual areas.

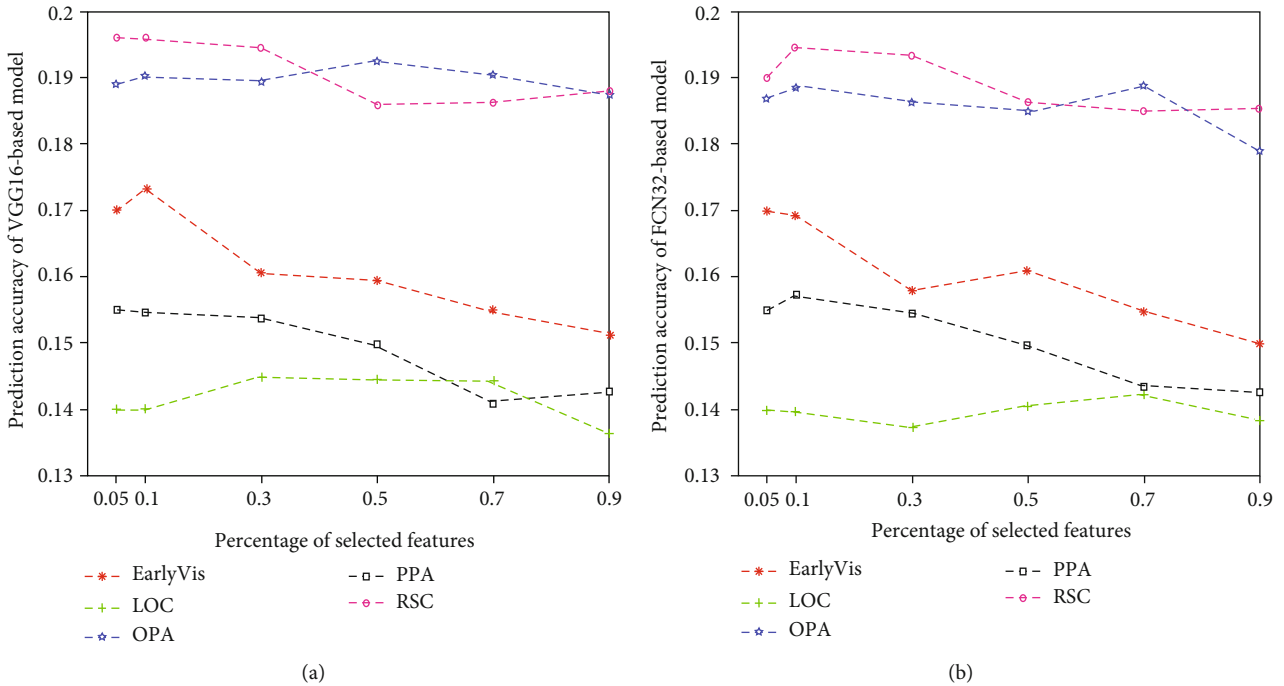


FIGURE 9: Relationship between the percentage of selected features and prediction accuracy of two encoding models (VGG16-based and FCN32-based models) in five visual areas of subject 2. Refer to Figure 8 for a detailed description of the plot elements.

CNNs and the human visual system, we analyzed the best encoding feature layer of the two CNNs. In detail, for voxels of different ROIs, we counted which layer of the CNN the best encoding layer came from. Figure 11 shows the contribution of each layer of the two feature models to voxel

responses in different visual areas. And the results of subject 2 and subject 3 are shown in Figures 12 and 13. From the figures, it is clear that voxel responses of the primary visual area can be better predicted by features in lower-level layers irrespective of the network's task. For the other

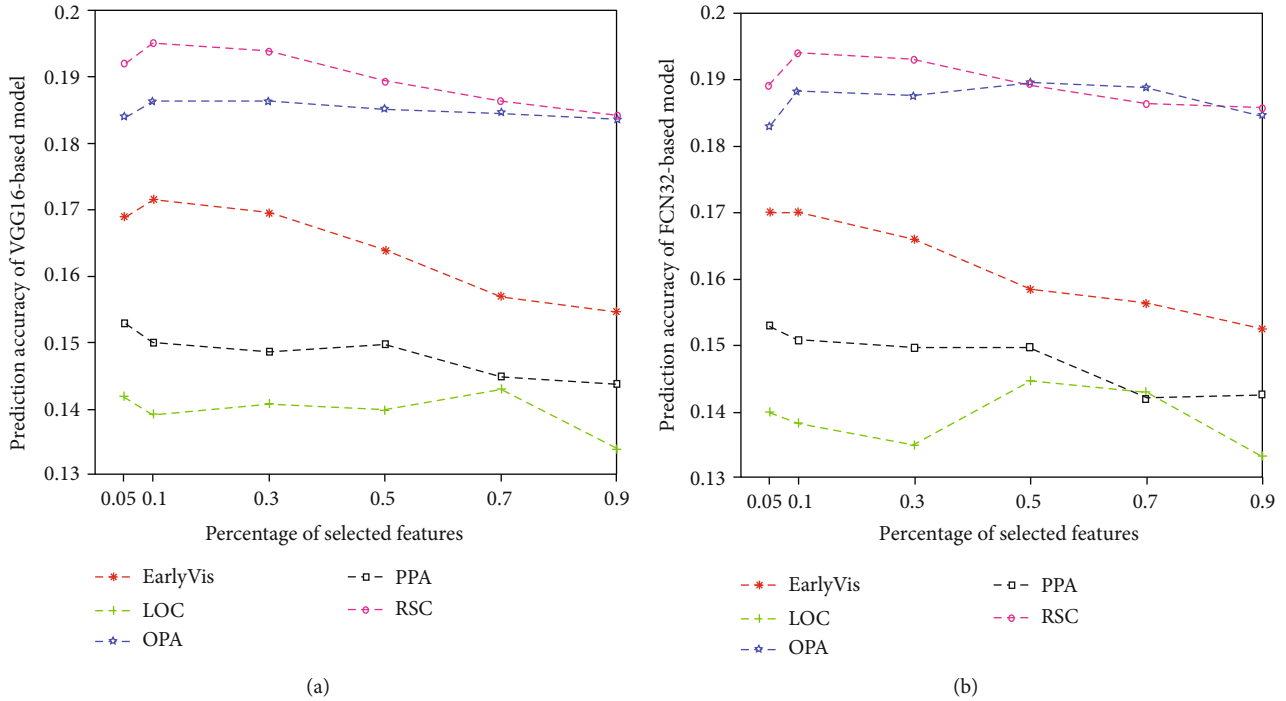


FIGURE 10: Relationship between the percentage of selected features and prediction accuracy of two encoding models (VGG16-based and FCN32-based models) in five visual areas of subject 3. Refer to Figure 8 for a detailed description of the plot elements.

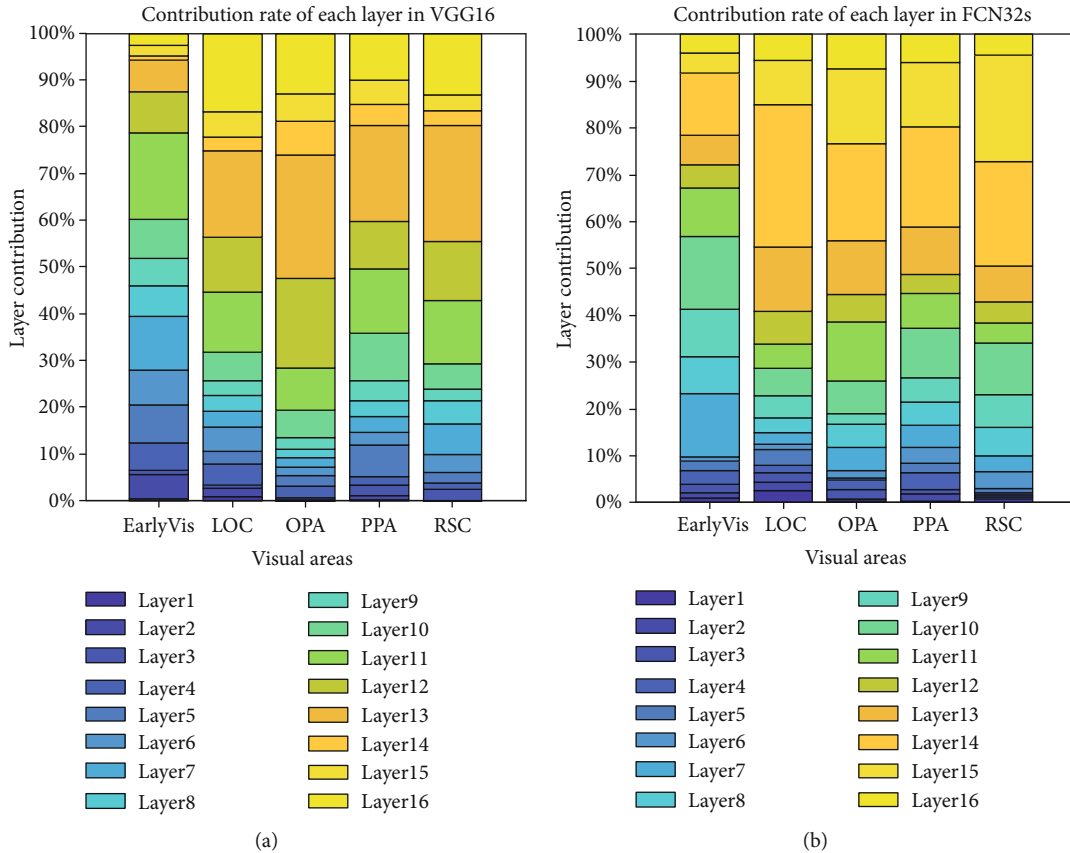


FIGURE 11: Contribution to the prediction accuracy of each layer in (a) VGG16 and (b) FCN32 networks. The ordinate represents the contribution of each layer, and the abscissa represents the five visual areas. The color bar from deep to shallow indicates the network layer from a low level to a high level.

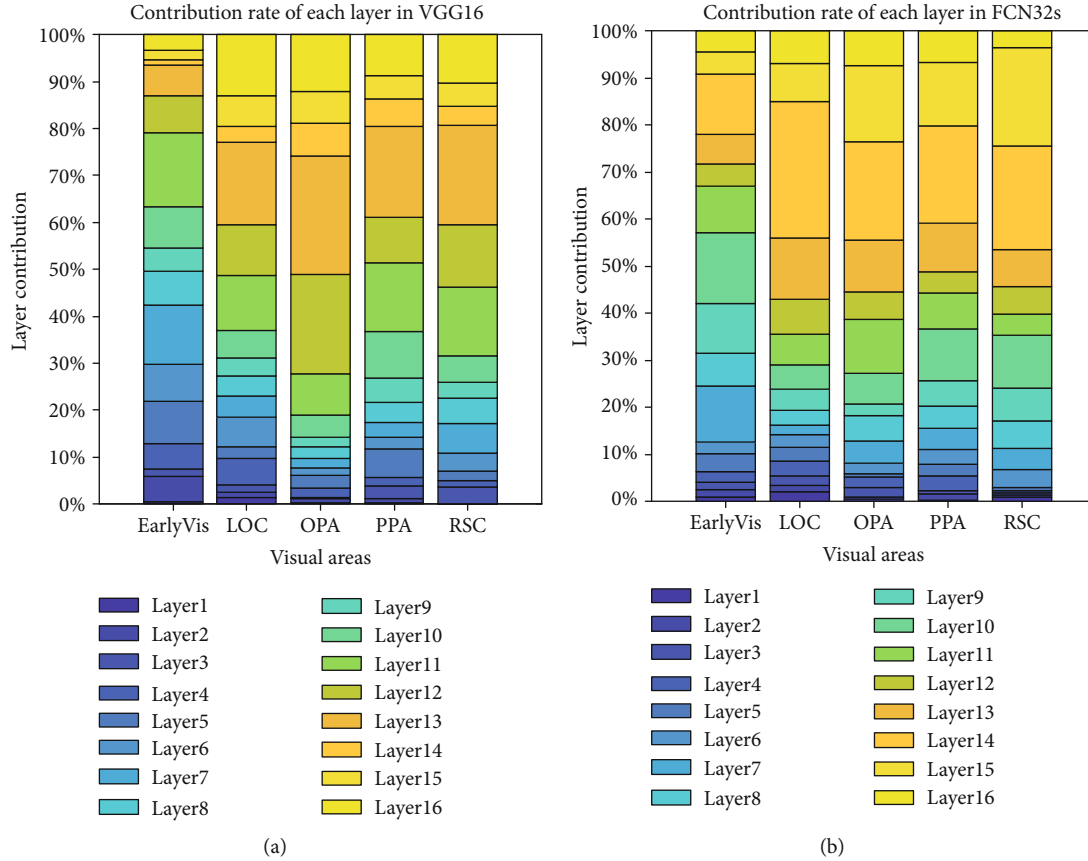


FIGURE 12: Contribution to the prediction accuracy of each layer in (a) VGG16 and (b) FCN32s networks for subject 2. Refer to Figure 11 for a detailed description of the plot elements.

four high-level visual areas, features in higher-level layers can better predict voxel responses (Mann-Kendall method, $p < 0.05$).

This study verifies the hierarchical similarity between CNNs and the human visual system. It also confirms that the human visual system and CNNs similarly process visual information in a hierarchical manner [15, 16]. Specifically, in the visual information processing pathway of the human brain, primary visual areas process relatively simple information, such as edges and shapes, and advanced visual areas process more complex visual features such as semantics and color. This is similar in CNNs where lower layers deal with simpler features and deeper layers deal with more complex features.

4. Discussions

4.1. Encoding Model Based on the Classification Network (VGG16) Has Better Prediction Performance. From the obtained prediction accuracies, we found that the encoding model based on the classification network is superior to that based on the segmentation network. Meanwhile, the prediction performance of the encoding model based on fused features is significantly better than that of the model based on segmentation and is almost the same as that of the model based on classification.

Our results show that different networks based on different computer vision tasks can affect the performance of the encoding models. We can also infer, to some extent, that the visual classification task can better fit human visual information processing than the visual segmentation task, with the human brain already completing the segmentation of objects in the process of completing the visual classification task. This is consistent with the discovery of David H. Hubel and Torsten Wiesel, 1981 Nobel Prize winners, that the information processing of the visual system is hierarchical in visual areas and the working process of the brain is iterative and abstract [41]. Upon obtaining the original information by the retina, visual area V1 firstly processes features related to edges and directions. Then, visual area V2 processes features related to contours and shapes. Finally, higher visual areas perform more refined classifications through more high-level abstractions iteratively. Hence, the human visual system already implements most of the segmentation tasks during information processing to realize the classification of external stimuli. This process is embedded in our brain and happens almost instantaneously.

From the point of view of natural evolution, primitive humans only need to identify whether an object in the field of vision is threatening them to avoid risk. This means that the object only needs to be categorized without the need for a specific segmentation.

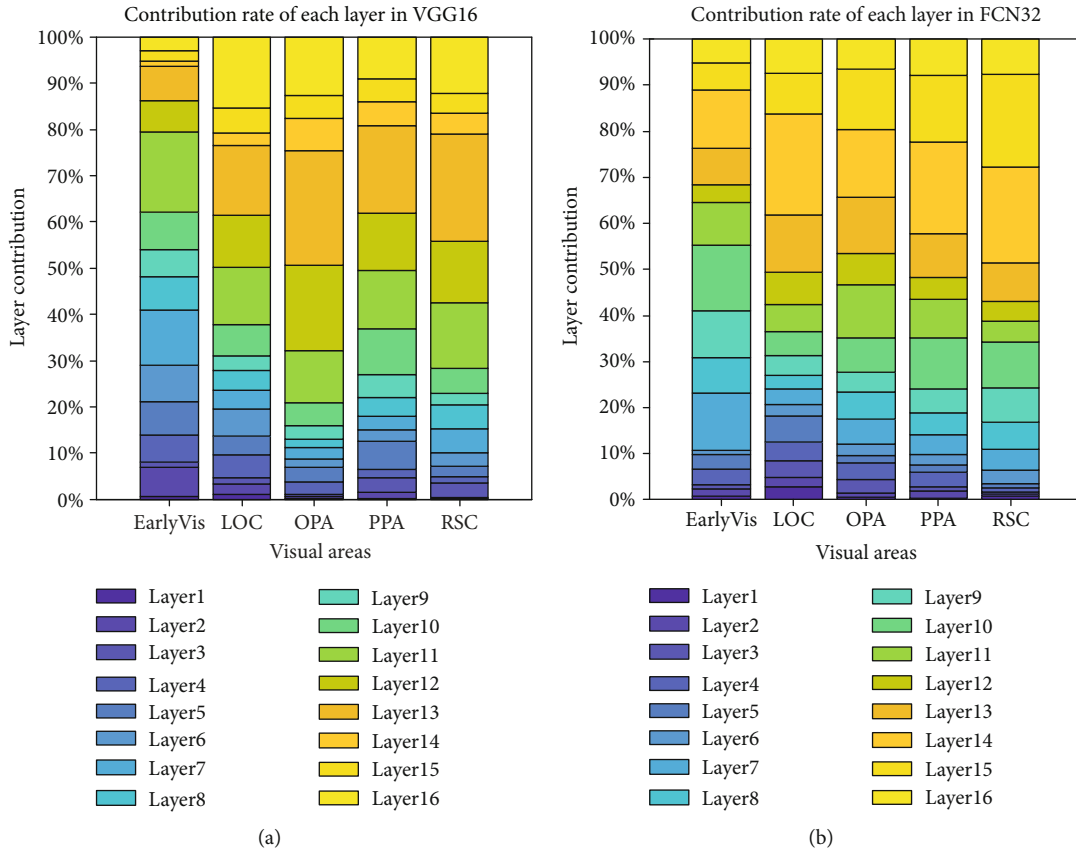


FIGURE 13: Contribution to the prediction accuracy of each layer in (a) VGG16 and (b) FCN32 networks for subject 3. Refer to Figure 11 for a detailed description of the plot elements.

From the experimental point of view, the subjects performed a task to judge the likes and dislikes of the input stimuli, which does not involve specific visual segmentation. This may have limitations that could affect the encoding performance. However, we can deduce that in the case of humans performing default visual tasks, the visual system gives priority to the classification of objects. On the other hand, when performing specific visual tasks, such as visual attention tasks, humans may give more priority to object segmentation.

4.2. Relationship between Classification and Segmentation Tasks in the Human Visual System. The visual encoding model based on classification and segmentation task-driven networks has advantages in predicting voxel responses, which indicates that the human visual system cannot be completely simulated by a certain task-driven network and performs various and complex visual tasks during visual information processing. We found that the prediction performance of the encoding model based on classification features is significantly better than that of the model based on segmentation features; hence, the CNN performing the classification task is more similar to the human visual system. The encoding model based on fused features and that based on classification features have almost the same performance, which indicates that the classification task is similar to most of the segmentation task. In other words, during visual processing, the human

brain completes most of the visual segmentation when the visual stimuli are classified.

4.3. The Prediction Accuracies of the Three Models Are Not High. From the perspective of encoding efficiency, Güçlü and van Gerven [16] employed a motion recognition network to predict the voxel responses in the dorsal pathway. In addition, a recent study investigated the impact of different computer vision tasks on deep networks performing visual encoding [19]. This demonstrates that research on encoding efficiency is beginning to gain attraction in the field.

In this study, we used the BOLD5000 dataset, which is the largest publicly published dataset. However, the obtained prediction accuracies of the three encoding models are not particularly high, which may be related to the diversity of stimuli in the dataset and absence of restrictions of the subjects' sights in the experiments. It should be emphasized that subjects only judged whether they liked or disliked the input images during the experiment. Hence, this limitation in the task may have an impact on the encoding performance. Moreover, it is unknown whether the performance of the encoding model based on the segmentation network would be improved if the subjects performed a corresponding visual segmentation task. This needs to be addressed in future work, highlighting its importance and relevance.

5. Conclusions

In conclusion, we explored the impact of different networks based on different tasks on encoding models. We found that the performance of the encoding model based on fused features is significantly better than that of the model based on segmentation and is almost the same as that of the model based on classification. This demonstrates that the CNN performing the classification task is more similar to the human visual system, and most of the segmentation of the visual system for the stimuli is completed with the process of object classification. However, we also found that the encoding model based on segmentation had better prediction performance on some voxels, which further illustrates the complexity and diversity of the human visual mechanism. In the future, we will consider more types of networks that perform different computer vision tasks, such as target detection and object recognition, which are aimed at not only improving the prediction performance but also better realizing the mechanism of human vision. Here, we demonstrated a valuable way of developing a computational neuroscience model from the perspective of computer vision.

Data Availability

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

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The work was supported by the National Key Research and Development Plan of China under grant 2017YFB1002502 and the National Natural Science Foundation of China (No.61701089).

References

- [1] M. Chen, J. Han, X. Hu, X. Jiang, L. Guo, and T. Liu, "Survey of encoding and decoding of visual stimulus via fMRI: an image analysis perspective," *Brain imaging and behavior*, vol. 8, no. 1, pp. 7–23, 2014.
- [2] N. Kriegeskorte, "Deep neural networks: a new framework for modeling biological vision and brain information processing," *Annual Review of Vision Science*, vol. 1, no. 1, pp. 417–446, 2015.
- [3] L. Paninski, J. Pillow, and J. Lewi, "Statistical models for neural encoding, decoding, and optimal stimulus design," in *Progress in Brain Research*, vol. 165, no. 6, pp. 493–507, 2007.
- [4] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *Neuroimage*, vol. 56, no. 2, pp. 400–410, 2011.
- [5] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008.
- [6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886–893, San Diego, CA, USA, June 2005.
- [7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.
- [8] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [9] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Image net classification with deep convolutional neural networks," *International Conference on Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [11] J. Ba, V. Mnih, and K. Kavukcuoglu, *Multiple object recognition with visual attention*, Computer Science, 2014.
- [12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 39no. 4, pp. 640–651, Boston, MA, USA, June 2015.
- [13] J. D. Cohen, N. Daw, B. Engelhardt et al., "Computational approaches to fMRI analysis," *Nature Neuroscience*, vol. 20, no. 3, pp. 304–313, 2017.
- [14] P. Agrawal, D. Stansbury, J. Malik, and J. L. Gallant, "Pixels to voxels: modeling visual representation in the human brain," 2014, <https://arxiv.org/abs/1407.5104>.
- [15] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *The Journal of Neuroscience*, vol. 35, no. 27, pp. 10005–10014, 2015.
- [16] U. Güçlü and M. A. van Gerven, "Increasingly complex representations of natural movies across the dorsal stream are shared between subjects," *Neuro Image*, vol. 145, Part B, pp. 329–336, 2017.
- [17] H. Wen, J. Shi, W. Chen, and Z. Liu, "Deep residual network predicts cortical representation and organization of visual features for rapid categorization," *Scientific Reports*, vol. 8, no. 1, p. 3752, 2018.
- [18] M. Schrimpf, J. Kubilius, H. Hong et al., *Brain-score: which artificial neural network for object recognition is most brain-like?*, no. article 407007, 2018bioRxiv, 2018.
- [19] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature neuroscience*, vol. 19, no. 3, pp. 356–365, 2016.
- [20] N. Chang, J. A. Pyles, A. Gupta, M. J. Tarr, and E. M. Aminoff, "BOLD 5000: a public fMRI dataset of 5000 images," 2018, <https://arxiv.org/abs/1809.01281>.
- [21] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010.
- [22] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014* 740–755.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, and F.-F. Li, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, Florida, USA, June 2009.

- [24] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.
- [25] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?," *Neural computation*, vol. 17, no. 8, pp. 1665–1699, 2005.
- [26] A. Anzai, X. Peng, and D. C. Van Essen, "Neurons in monkey visual area V2 encode combinations of orientations," *Nature neuroscience*, vol. 10, no. 10, pp. 1313–1321, 2007.
- [27] B. D. Willmore, R. J. Prenger, and J. L. Gallant, "Neural representation of natural images in visual area V2," *The Journal of Neuroscience*, vol. 30, no. 6, pp. 2102–2114, 2010.
- [28] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher, "The lateral occipital complex and its role in object recognition," *Vision Research*, vol. 41, no. 10-11, pp. 1409–1422, 2001.
- [29] M. X. Lowe, J. Rajsic, J. P. Gallivan, S. Ferber, and J. S. Cant, "Neural representation of geometry and surface properties in object and scene perception," *NeuroImage*, vol. 157, pp. 586–597, 2017.
- [30] R. Epstein, A. Harris, D. Stanley, and N. Kanwisher, "The parahippocampal place area: recognition, navigation, or encoding?," *Neuron*, vol. 23, no. 1, pp. 115–125, 1999.
- [31] S. D. Vann, J. P. Aggleton, and E. A. Maguire, "What does the retrosplenial cortex do?," *Nature Reviews Neuroscience*, vol. 10, no. 11, pp. 792–802, 2009.
- [32] N. Ketkar, "Introduction to pytorch," in *Deep learning with python*, pp. 195–208, Springer, 2017.
- [33] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, Computer Science, 2014.
- [34] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.
- [35] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.
- [36] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage*, vol. 19, no. 2, pp. 261–270, 2003.
- [37] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *Siam Journal on Applied Mathematics*, vol. 49, no. 3, pp. 906–931, 1989.
- [38] D. Needell, J. Tropp, and R. Vershynin, "Greedy signal recovery review," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, October 2008.
- [39] D. Needell and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 310–316, 2010.
- [40] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Computational Mathematics*, vol. 9, no. 3, pp. 317–334, 2009.
- [41] D. H. Hubel and T. N. Wiesel, "Brain Mechanisms of Vision," *Scientific American*, vol. 241, no. 3, pp. 150–163, 1979.

Research Article

A Simple Method to Train the AI Diagnosis Model of Pulmonary Nodules

Zhehao He , Wang Lv, and Jian Hu 

Department of Thoracic Surgery, The First Affiliated Hospital, College of Medicine, Zhejiang University, China

Correspondence should be addressed to Jian Hu; dr_hujian@zju.edu.cn

Received 28 May 2020; Accepted 29 June 2020; Published 1 August 2020

Guest Editor: Tao Huang

Copyright © 2020 Zhehao 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.

Background. The differential diagnosis of subcentimetre lung nodules with a diameter of less than 1 cm has always been one of the problems of imaging doctors and thoracic surgeons. We plan to create a deep learning model for the diagnosis of pulmonary nodules in a simple method. **Methods.** Image data and pathological diagnosis of patients come from the First Affiliated Hospital of Zhejiang University School of Medicine from October 1, 2016, to October 1, 2019. After data preprocessing and data augmentation, the training set is used to train the model. The test set is used to evaluate the trained model. At the same time, the clinician will also diagnose the test set. **Results.** A total of 2,295 images of 496 lung nodules and their corresponding pathological diagnosis were selected as a training set and test set. After data augmentation, the number of training set images reached 12,510 images, including 6,648 malignant nodular images and 5,862 benign nodular images. The area under the *P-R* curve of the trained model is 0.836 in the classification of malignant and benign nodules. The area under the ROC curve of the trained model is 0.896 (95% CI: 78.96%~100.18%), which is higher than that of three doctors. However, the *P* value is not less than 0.05. **Conclusion.** With the help of an automatic machine learning system, clinicians can create a deep learning pulmonary nodule pathology classification model without the help of deep learning experts. The diagnostic efficiency of this model is not inferior to that of the clinician.

1. Introduction

Malignant tumours are a type of malady that seriously threatens human life and health. In China, although the 5-year survival rate of malignant tumours is increasing year by year [1], the morbidity and mortality still increase every year [2]. Among them, lung cancer ranks first in the incidence of malignant tumours in China [1]. The results of the study show that screening low-dose spiral CT for people at high risk of lung cancer can significantly reduce lung cancer mortality [3]. However, the ensuing problem is that the detection rate of pulmonary nodules is increased. The differential diagnosis of subcentimetre lung nodules with a diameter of less than 1 cm has always been one of the problems of imaging doctors and thoracic surgeons [4].

In recent years, research and application of artificial intelligence based on deep learning are in full swing. In the field of medicine, the use of deep learning techniques for the diagnosis of imaging [5] and pathological [6] images is emerging.

However, deep learning is a subject with a high threshold, and such research often requires the in-depth participation of deep learning engineers. In order to further reduce the threshold of deep learning, people of insight have proposed the concept of automatic machine learning (AutoML) [7]. AutoML can completely automate the creation of the entire deep learning process, reducing the knowledge of researchers in various fields using deep learning for research work.

This study intends to use Microsoft's Custom Vision [8] AutoML system to train the model by learning the thin-layer CT imaging data of the lung nodules and the corresponding pathological diagnosis. Use the test data set to test the diagnostic model and compare the diagnosis of the clinician. Use the results to evaluate the effectiveness of the model.

2. Materials and Method

2.1. Training Set and Test Set. Retrieve the pathological diagnosis database of surgical specimens from the Department of

Pathology, the First Affiliated Hospital of Zhejiang University School of Medicine, from October 1, 2016, to October 1, 2019. In the database, screen out the pathological diagnosis with a higher ranking in the pathological results of pulmonary nodules. According to the patient data selected by the above diagnosis, the CT images of the patient in the hospital imaging system are retrieved one by one according to the patient's medical record number. The inclusion criteria include the following:

- (a) Must be CT images of lungs within 30 days before surgery
- (b) The CT image of the lungs should be a high-resolution horizontal sequence CT image (layer thickness 1.0~1.25 mm)
- (c) There is no limit to the size of the lung nodule, but they need to be spherical or quasispherical, the boundaries can be recognized, and the surroundings are surrounded by inflatable lung tissue, without atelectasis
- (d) There is only one lesion in the same lung lobe, or there are multiple lesions, but they are all removed, and the pathological results after surgery are the same

Download the patient's high-resolution CT image sequence (DICOM format) from the imaging system, and record the pathological diagnosis corresponding to the nodule. Randomly select 90% of all nodules as the training data set and 10% as the test data set.

2.2. Data Preprocessing. Convert DICOM format images to bitmap images. The conversion scheme is as follows: in DICOM format, each pixel records the CT value whose unit is the Hounsfield unit. The range of the CT value is from -1000 to 1000. We specified for each CT value the only colour corresponding to it. Through this conversion, we get a colour CT bitmap image (Figure 1).

Select the images in the sequence that contain a lung nodule with the diameter of the lung nodule in the image not less than 80% of the largest diameter of the nodule. Crop the selected bitmap image to obtain an approximately square rectangular image containing the nodule image. The side length of the cropped image should be between 2 and 3 times the diameter of the nodule. Moreover, the nodule pattern is located approximately in the middle.

2.3. Data Augmentation. Perform the following operations on the training set image: rotate 90 degrees, 180 degrees, and 270 degrees clockwise, flip horizontally, and flip vertically. The above means make the training data set data increased by six times.

2.4. Training a Deep Learning Diagnostic Model. Visit <https://www.customvision.ai>, register a new account, and log in. Create a new training project, and select "Classification" for the "Project Type" option, "Multiclass (Single tag per image)" for the "Classification Types" option, and "General" for the

"Domains" option. Upload all the training data set images, and add labels to the images according to the pathology type, and then start training. Wait for a moment, and record the training result data after the training is completed.

2.5. Evaluate the Trained Model with Test Data Set Images. Upload the test data set images on the test page to test the trained model. Since each nodule contains multiple test images, upload and test each image, record the percentage of each diagnosis possibility for each image, and average the multiple images. The diagnosis with the highest percentage is the final predicted diagnosis.

Invite three thoracic surgeons. View the lung nodules in the CT images corresponding to the test data set one by one, and diagnose according to the pathological grouping of the training data set. Make statistics after comparing the actual pathological results.

3. Results

Finally, a total of 2,295 images of 496 lung nodules and their corresponding pathological diagnosis were selected as a training set and test set. After data augmentation, the number of training set images expanded to 6 times before and eventually reached 12,510 images, including 6,648 malignant nodular images and 5,862 benign nodular images (Table 1).

The model trained using the training data set without data augmentation has a training result with a 50% probability threshold, the accuracy rate is 69.7%, the recall rate is 67.0%, and the area under the curve is 0.738. The training results of the model trained with the data augmentation training data set are as follows: at a 50% probability threshold, the accuracy rate is 78.8%, the recall rate is 76.2%, and the area under the curve is 0.836. After data augmentation, the area under the curve of the model is more excellent than before (Figure 2).

Use the model trained with enhanced data to make diagnostic predictions on the test data set. For benign and malignant classification, the model trained after data augmentation can reach a sensitivity of 88.24%, a specificity of 90.91%, and an overall accuracy rate of 90.0%. For pathological classification, the classification accuracy rate is 78%. For this test data set, three clinicians judged that the average sensitivity of benign and malignant classification is 86.27%, the average specificity is 65.66%, the average overall accuracy rate is 72.67%, and the average pathological accuracy rate is 48.67% (Table 2).

For the model trained after data augmentation and the three doctors, ROC curves are constructed for the diagnosis of benign and malignant nodules, which are used to judge their diagnostic value for the test data set. The area under the curve (AUC) corresponding to the model was 0.896 (95% CI: 78.96%~100.18%), and the area under the curve values corresponding to the three doctors were 0.759 (95% CI: 62.17%~89.70%), 0.775 (95% CI: 63.97%~90.93%), and 0.745 (95% CI: 60.12%~88.90%). The results mean that the model has a high value for the diagnosis of benign and malignant nodules in test data sets, and the corresponding optimal cutoff value is 0.791 (at this time, the sensitivity is 88.2% and the specificity is 90.9%). Moreover, the area under the curve

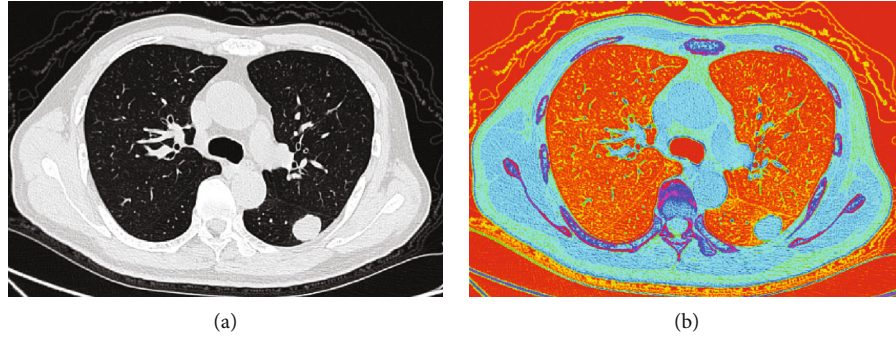


FIGURE 1: (a) A grayscale image of lung CT in a lung window and (b) a colour image after conversion.

TABLE 1: Training set and test set.

Type		Nodules			Images		
Benign or malignant	Pathology	Training set	Test set	All	Training set	Test set	All
Malignant	AAH/AIS/MIA	131	6	137	400	18	418
	IAC	72	8	80	460	36	496
	Metastatic cancer	54	3	57	248	10	258
	All	257	17	274	1108	64	1172
Benign	Chronic inflammation/granuloma	91	16	107	556	92	648
	Intrapulmonary lymph nodes	42	11	53	119	28	147
	Hemangioma	12	1	13	77	2	79
	Hamartoma	44	5	49	225	24	249
	All	189	33	222	977	146	1123
All		446	50	496	2085	210	2295

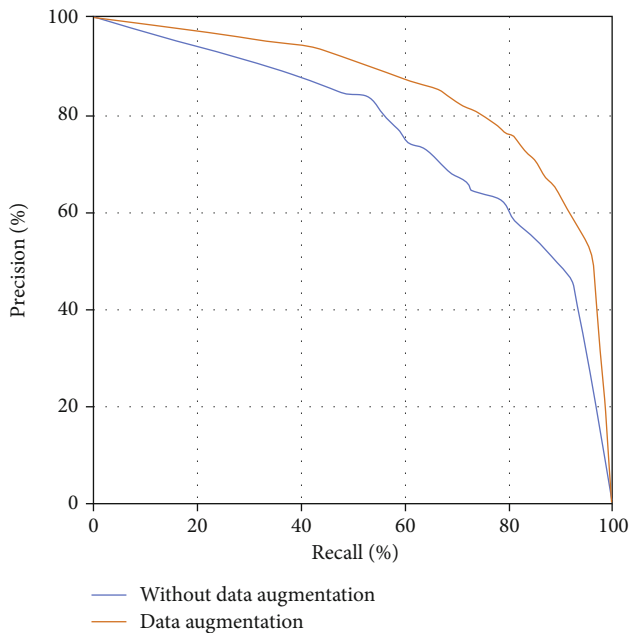


FIGURE 2: The P - R curve of the trained model with and without data augmentation.

is greater than that of the three doctors. However, the P value is not less than 0.05 (Table 3).

4. Discussion

Since deep learning has shown high accuracy in many computer vision tasks, in recent years, the research field of lung nodule detection and classification based on deep neural networks has rapidly heated up [9]. However, deep learning is a profession with a high threshold. Such research must rely on the participation of experienced deep learning engineers. In this study, the authors did not deeply study deep learning algorithms and specific operating practices. Only after a rough understanding of deep learning principles, an automatic deep learning system was used to create a deep learning diagnostic model. In this study, less than 500 cases of pulmonary nodules were collected as training data sets. Although the amount of data is not large, the final diagnostic model is still satisfactory and can be equivalent to the diagnosis of human doctors.

In previous studies [10], professional deep learning frameworks were often used to directly read lung nodule data in DICOM format to train models. However, in this study, Custom Vision can only read image data for training. To this end, we must convert DICOM format images into image format data for model training.

TABLE 2: Diagnosis results of the trained model and the doctors on the test data set.

	Sensitivity (%)	Specificity (%)	Accuracy rate (%)	Pathological accuracy rate (%)
Trained model (data augmentation)	88.24	90.91	90	78
Doctor A	88.24	63.64	72	46
Doctor B	88.24	66.67	74	48
Doctor C	82.35	66.67	72	52
Doctor average	86.27	65.66	72.67	48.67

TABLE 3: AUC and ROC curve best cutoff of the trained model and the doctors.

	AUC	Optimal cutoff	Sensitivity (%)	Specificity (%)	<i>P</i> (compared to the trained model)
Trained model (data augmentation)	0.896	0.791	88.2	90.9	
Doctor A	0.759	0.519	88.2	63.6	0.1212
Doctor B	0.775	0.549	88.2	66.7	0.1673
Doctor C	0.745	0.490	82.4	66.7	0.0963

In the CT image of lungs in DICOM format, the data of each pixel is between -1000 and 1000. In other words, the CT machine can recognize 2000 different density differences in the human body. The CT values of human organs are mostly concentrated in a relatively narrow range. In order to facilitate display and doctor reading, DICOM format images will be displayed as grayscale images through different window width and window level values. The doctor can very sensitively perceive the difference in the CT value within the window width by reading the CT image with the naked eye. The disadvantage is that the CT value outside the window width will eventually be displayed as completely white or completely black. This image conversion will lose data.

In order to avoid losing data, we have created an image conversion method. CT can identify 2000 different gray levels in the human body. In the computer, taking the 24-bit colour bitmap as an example, the number of colours that can be displayed is 16,777,216. Therefore, each different CT value in the DICOM format can be given a corresponding colour, so that all the information in the DICOM format image can be completely retained. In the colour image after conversion, human eyes cannot recognize the slight difference between some colours. However, for computer processing, it has entirely different colours.

Lung adenocarcinoma is the most common type of pathology in non-small-cell lung cancer, and it accounts for about 50% of all lung cancer patients [11]. With the changes in the epidemiology of lung cancer, the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS), and the European Respiratory Society (ERS) formed a joint working group in 2011 to announce a new classification method for lung adenocarcinoma [12]: atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma cancer (IAC). It is generally believed that AAH, AIS, MIA, and IAC are different stages of early non-small-cell lung cancer during the progression of the disease [13]. Statistical analysis of the prognosis of different types of lung adenocarcinoma revealed that AAH, AIS,

and MIA have an excellent prognosis [14], and their 5-year survival rate can reach almost 100%. The 5-year survival rate of invasive adenocarcinoma is significantly lower than that of the previous three types.

In the choice of surgical procedures, for the types of AAH, AIS, and MIA, recent studies [15] have been more inclined to perform sublobar resection (pulmonary wedge resection, segmentectomy, and combined subsection resection). The survival rate and local recurrence rate are not significantly different from those of lobectomy. Some scholars [16] even believe that because of the types of AAH, AIS, and MIA, the possibility of lymph node metastasis is extremely low. Stereotactic body radiotherapy (SBRT) treatment of these types of lesions can achieve similar treatment effect to traditional surgery. After the lesion reaches the level of invasive adenocarcinoma, lobectomy is more recommended.

In the process of rapid intraoperative pathological diagnosis, due to the influence of factors such as the material limitation, it is sometimes difficult for pathologists to distinguish between AAH, AIS, and MIA [17]. The three types of lesions have an excellent prognosis, and the clinical significance of surgical guidance is almost the same. Therefore, these three types are combined into a group as a low-risk group, and invasive adenocarcinoma is considered to belong to a high-risk group. Therefore, the classification model of deep learning can be more focused on identifying whether the lesion is invasive adenocarcinoma, which is of great significance for the formulation of surgical procedures and the prediction of disease prognosis.

In this study, the number of benign diseases is relatively small. For example, there are only 13 cases of pulmonary sclerosing hemangioma. If such a small number of cases are directly input into the deep learning engine for learning, it is bound to fail to obtain good results. Therefore, various forms of data augmentation are necessary. For image data, pure data augmentation methods generally include geometric transformation. In this study, the flip and rotation operations in geometric transformation are used. Rotating and flipping the image of a lung nodule do not affect the essence

and characteristics of the image. This operation method is simple, but the effect is pronounced. The model uses the enhanced image for training, which increases the area under the curve by about 0.1 compared to the previous one.

The two most basic indicators in the fields of deep learning related to information retrieval, classification, recognition, translation, etc. are the recall rate and precision rate. Recall rate = true positive/(true positive + false negative), and precision rate = true positive/(true positive + false positive). Therefore, the recall rate is the sensitivity in medical diagnosis, but the precision rate is not specific. Nonetheless, the relationship between the precision rate and recall rate is similar to the relationship between sensitivity and specificity: precision and recall affect each other, and the ideal situation is, of course, both high precision and recall. But under normal circumstances, precision rate is inversely proportional to recall rate.

Therefore, similar to the ROC curve formed by the correlation between sensitivity and specificity, the relationship between the precision rate and recall rate can also build a P - R curve, where the recall rate value is used as the x -axis and the precision rate value is used as the y -axis to indicate the different relationship between precision and recall. The average precision rate represents the average value of the precision rate during the change of the recall rate from 0 to 1, that is, the integration of the precision rate during the shift in the recall rate from 0 to 1, which is equivalent to the area under the PR . The area surrounded by the x - and y -axes (area under the P - R curve). In this way, the comparison between multiple models becomes intuitive. You only need to place the P - R curves of various models in the same coordinate system and compare the area under the curve.

By analyzing the diagnosis results, the AUC value corresponding to the ROC curve of the neural network is 0.896, indicating that the model is of higher value for the diagnosis of benign and malignant nodules in test data sets. The model can achieve 90% accuracy for benign and malignant classification and 78% accuracy for pathological classification. Moreover, the AUC value is higher than that of the three doctors. However, the P value is not less than 0.05, indicating that the model's diagnostic efficiency of benign and malignant classification is similar to that of the clinician.

When clinicians diagnose lung nodules, the sensitivity is not much different from that of the diagnostic model, but the specificity is significantly lower than that of the diagnostic model. The possible reason is that as a clinician when diagnosing pulmonary nodules, they tend to increase sensitivity, increase the detection rate of potentially malignant tumours, and reduce the rate of missed diagnosis. As a result, the false positives are high and the specificity is reduced.

5. Conclusion

This study shows that with the help of an automatic machine learning system, clinicians can create a deep learning pulmonary nodule pathology classification model without the help of deep learning experts. The diagnostic efficiency of this model is not inferior to that of the clinician, but the deep learning algorithm model will not replace the status of clini-

cians and radiologists. On the contrary, it can effectively help clinicians and radiologists in clinical work.

Data Availability

The data and materials in the current study are available from the corresponding author on reasonable request.

Disclosure

This work is part of the author's doctoral dissertation research work, which was performed by Zhehao He, and the supervisor is Jian Hu.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would particularly like to thank Rainbow Chen for the excellent technical support. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0113500); the Key Project of Zhejiang Province Science and Technology Plan, China (No. 2020C03058); the Key Discipline of Zhejiang Province Traditional Chinese Medicine (Combined Chinese and Western Medicine) (2017-XK-A33); the Zhejiang Lung Cancer Center (JBZX-202007); and the Natural Science Foundation of Zhejiang Province (LQ20H160050).

References

- [1] R. S. Zheng, K. X. Sun, S. W. Zhang et al., "Report of cancer epidemiology in China, 2015," *Chinese Journal of Oncology*, vol. 41, no. 1, pp. 19–28, 2019.
- [2] R. S. Zheng, X. Y. Gu, X. T. Li et al., "Analysis on the trend of cancer incidence and age change in cancer registry areas of China, 2000 to 2014," *Chinese Journal of Preventive Medicine*, vol. 52, no. 6, pp. 593–600, 2018.
- [3] V. A. Moyer, "Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement," *Annals of Internal Medicine*, vol. 160, no. 5, pp. 330–338, 2014.
- [4] M. J. Schuchert, A. Kilic, A. Pennathur et al., "Oncologic outcomes after surgical resection of subcentimeter non-small cell lung cancer," *The Annals of Thoracic Surgery*, vol. 91, no. 6, pp. 1681–1688, 2011.
- [5] S. Bhatia, Y. Sinha, and L. Goel, *Lung Cancer Detection: A Deep Learning Approach [M]//Soft Computing for Problem Solving*, Springer, Singapore, 2019.
- [6] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module," *PLoS One*, vol. 14, no. 3, article e0214587, 2019.
- [7] M. A. Zöllner and M. F. Huber, "Survey on automated machine learning," 2019, <https://arxiv.org/abs/1904.12054>.
- [8] M. Salvaris, D. Dean, and W. H. Tok, *Cognitive Services and Custom Vision*, Deep Learning with Azure, Apress, Berkeley, CA, 2018.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

- [10] C. Zhang, X. Sun, K. Dang et al., "Toward an expert level of lung cancer detection and classification using a deep convolutional neural network," *The Oncologist*, vol. 24, no. 9, pp. 1159–1165, 2019.
- [11] R. Zhang, Y. Zhang, F. Wen, W. Kai, and S. Zhao, "Analysis of pathological types and clinical epidemiology of 6,058 patients with lung cancer," *Chinese Journal of Lung Cancer*, vol. 19, no. 3, pp. 129–135, 2016.
- [12] P. E. Van Schil, A. D. L. Sihoe, and W. D. Travis, "Pathologic classification of adenocarcinoma of lung," *Journal of Surgical Oncology*, vol. 108, no. 5, pp. 320–326, 2013.
- [13] Y. Yatabe, A. C. Borczuk, and C. A. Powell, "Do all lung adenocarcinomas follow a stepwise progression?," *Lung Cancer*, vol. 74, no. 1, pp. 7–11, 2011.
- [14] J. Gu, C. Lu, J. Guo et al., "Prognostic significance of the IASLC/ATS/ERS classification in Chinese patients—a single institution retrospective study of 292 lung adenocarcinoma," *Journal of Surgical Oncology*, vol. 107, no. 5, pp. 474–480, 2013.
- [15] C. Cao, D. Chandrakumar, S. Gupta, T. D. Yan, and D. H. Tian, "Could less be more?—A systematic review and meta-analysis of sublobar resections versus lobectomy for non-small cell lung cancer according to patient selection," *Lung Cancer*, vol. 89, no. 2, pp. 121–132, 2015.
- [16] W. Weder, D. Moghanaki, B. Stiles, S. Siva, and G. Rocco, "The great debate flashes: surgery versus stereotactic body radiotherapy as the primary treatment of early-stage lung cancer," *European Journal of Cardio-Thoracic Surgery*, vol. 53, no. 2, pp. 295–305, 2018.
- [17] S. Liu, R. Wang, Y. Zhang et al., "Precise diagnosis of intraoperative frozen section is an effective method to guide resection strategy for peripheral small-sized lung adenocarcinoma," *Journal of Clinical Oncology*, vol. 34, no. 4, pp. 307–313, 2016.

Research Article

Comparison of Common Methods for Precision Volume Measurement of Hematoma

Minhong Chen,¹ Zhong Li¹, Jianping Ding,² Xingqi Lu,² Yinan Cheng,³ and Jiayun Lin¹

¹College of Science, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China 310018

²The Affiliated Hospital of Hangzhou Normal University, Hangzhou, Zhejiang, China 310015

³College of Science, Southern University of Science and Technology, Shenzhen, Guangdong, China 518055

Correspondence should be addressed to Zhong Li; lizhong@zstu.edu.cn and Jiayun Lin; lin_linjy@126.com

Received 21 March 2020; Accepted 18 May 2020; Published 17 July 2020

Guest Editor: Lei Chen

Copyright © 2020 Minhong 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.

Purpose. Our aim is to conduct analysis and comparison of some methods commonly used to measure the volume of hematoma, for example, slice method, voxelization method, and 3D-Slicer software method (projection method). **Method.** In order to validate the accuracy of the slice method, voxelization method, and 3D-Slicer method, these three methods were first applied to measure two known volumetric models, respectively. Then, a total of 198 patients diagnosed with spontaneous intracerebral hemorrhage (ICH) were recruited. The patients were split into 3 different groups based on the hematoma size: group 1: volume < 10 ml ($n = 89$), group 2: volume between 10 and 20 ml ($n = 59$), and group 3: volume > 20 ml ($n = 50$). And the shape of the hematoma was classed into regular (round to ellipsoid) with smooth margins ($n = 76$), irregular with frayed margins ($n = 85$), and multilobular ($n = 37$). The slice method, voxelization method, and 3D-Slicer method were adopted to measure the volume of hematoma, respectively, considering the nonclosed models and the models which may contain inaccurate normal information during CT scan. Moreover, the results were compared with the 3D-Slicer method for closed models. **Results.** There was a significant estimation error ($P < 0.05$) using these three methods to calculate the volume of the closed hematoma model. The estimated hematoma volume was calculated to be $14.2086743 \pm 0.900559087$ ml, $14.2119130 \pm 0.900851812$ ml, and $14.2123825 \pm 0.900835916$ ml using slice method 1, slice method 2, and the voxelization method, respectively, compared to $14.212656 \pm 0.900992371$ ml using the 3D-Slicer method. The mean estimation error was -0.00398172 ml, -0.00074303 ml, and -0.00027354 ml caused by slice method 1, slice method 2, and voxelization method, respectively. There was a significant estimation error ($P < 0.05$), applying these three methods to calculate the volume of the nonclosed hematoma model. The estimated hematoma volume was calculated to be $14.1928246 \pm 0.902210314$ ml using the 3D-Slicer method. The mean estimation error was calculated to be -0.00402121 ml, -0.00078237 ml, -0.00031288 ml, and -0.01983136 ml using slice method 1, slice method 2, voxelization method, and 3D-Slicer method, respectively. **Conclusions.** The 3D-Slicer software method is considered as a stable and capable method of high precision for the calculation of a closed hematoma model with correct normal direction, while it would be inappropriate for the nonclosed model nor the model with incorrect normal direction. The slice method and voxelization method can be the supplement and improvement of the 3D-Slicer software method, for the purpose of achieving precision medicine.

1. Introduction

Intracerebral hemorrhage (ICH) has been identified as a significant cause of death and disability around the world [1]. The increasing incidence of cerebral hemorrhage can cause progression of the disease. In addition, the amount of cerebral hemorrhage, or the cerebral hematoma volume, can be taken as a major indicator of early mortality at the time of

admission. It is also among the most effective indicators of the degree of neurological recovery within 90 days of the onset of the disease [2–6].

The diversity of hematoma shapes is one of the primary causes of errors in applying volume assessment methods. In practice, there will be brain lesions with inconspicuous lesions, irregular borders, discontinuities, and high noise. The shape of hypertensive cerebral hemorrhage can be

categorized into kidney shape, round shape, oval shape, fusi-form shape, and irregular shape, as shown in Figure 1. The diversity of hematoma shapes (Figure 1) makes it necessary to apply volumetric calculation methods that ensure both accuracy and robustness. Therefore, in order to facilitate the accurate diagnosis and treatment of disease, choosing an accurate, simple, and noninvasive approach to the measurement of intracranial hematoma volume is definitely conducive to the selection of treatment options, evaluation of clinical outcomes, and prediction of disease progression.

There are various methods to measure the volume of hematoma, and they are mainly classed into four categories, including the mathematical formula method, tool measurement method, CT machine measurement method, and software method. Among them, the Tada formula method is one of most commonly used formula methods. The formula is $V = 1/2 \times A \times B \times C$, where A indicates the long diameter, B represents the broad diameter, and C denotes the number of hematoma layers. The Tada formula has been extensively applied to assess the volume of intracerebral hematoma. Since the Tada formula in theory is derived from the ellipsoid volume formula, when the shape of an intracranial hematoma shows similarity to an ellipsoid, which has a regular shape, a hematoma such as a “ball” shape can be calculated using this method. However, when the shape of an intracranial hematoma is distant from the ellipsoid, that is, irregular hematoma or lobular hematoma, the Tada formula performs poorly [7]. In order to address this drawback, some improved ball volume formulas [8] were proposed based on the Tada formula. In spite of this, the accuracy of calculation for the volume formula remains associated with the shape of the hematoma. The more irregular the hematoma morphology, the more significant the error in the calculation results.

As computer technology progresses at a fast pace, the hematoma model can be measured and analyzed using different software methods. The 3D-Slicer method is one of the software methods purposed to measure the volume of a hematoma. It provides a free open source software platform for biomedical research to be conducted (<http://www.slicer.org>). With regard to the measurement principle, it is similar to the computer-aided volume analysis. The software is capable of identifying hematoma pixels based on CT data in cerebral hemorrhage images and reconstructing blood clots in a three-dimensional manner. Besides, it is free from restriction by hematoma morphology and bleeding sites. The 3D-Slicer method could ensure both accuracy and simplicity for hematoma assessment [9], which makes it gradually accepted as an effective measurement method [10–12]. In addition, the 3D-Slicer software method has been demonstrated to be faster and less user-intensive compared to manual delineation, which makes it suitable as a standard method. Xu et al. [7] analyzed the accuracy of the Tada formula by comparing with the 3D-Slicer software method, which led to the conclusion that hematoma assessment with software 3D-Slicer is a low-cost, accurate, and effective technique for the measurement of ICH volume. However, the stability of the 3D-Slicer software method has not yet been included in discussion. As for measurement of ICH volume, some other

methods can be analyzed and applied as well, such as the slice method and voxelization method.

In this paper, our aim is to improve the accuracy of hematoma assessment. The stability of the 3D-Slicer method was analyzed, and a comparison was performed between the 3D-Slicer method and two other methods. It was found out that, when the three-dimensional hematoma model is non-closed or the surface normal of the hematoma model is incorrect, the 3D-Slicer method will give rise to some errors, which can be rectified by two other methods, the slice method and the voxelization method.

2. Commonly Used Methods

2.1. 3D-Slicer Method (Projection Method). Jointly developed by Harvard University Brigham and Women’s Hospital and the Massachusetts Institute of Technology, 3D-Slicer software represents a free open source software platform for biomedical research. Hematoma is reconstructed using the original DICOM format data in 3D-Slicer software according to CT scanning, which ensures an accurate measurement for hematoma. Besides, the triangular mesh model is used for the volume measurement of hematoma by the 3D-Slicer method, slice method, and voxelization method.

2.1.1. Operation. Run 3D-Slicer software (3D-Slicer 4.6.2, Harvard University, USA), import the CT data of the patient in DICOM format, adjust the size of image, and proceed as follows: run Editor → Threshold → Apply. The CT threshold range is manually set, while the software automatically identifies and marks the pixels that constitute the hematoma. If necessary, editing is continued to completely separate the hematoma from the surrounding normal brain tissue. Run MakeModel → Models. Then, the three-dimensional shape of the hematoma and the volume of the hematoma can be determined, as shown in Figure 2.

2.1.2. Principle. 3D-Slicer software, as developed for the processing of image visualization and image analysis, is premised on VTK, ITK, Teem, QT, and other open source software [9, 13]. The principle of volume measurement is similar to the computer-aided volume analysis. The hematoma is segmented using the GrowCut method [9]. The hematoma volume is calculated following the three-dimensional reconstruction of hematoma. This method is simple, accurate, and resistant to the impact made by the shape and location of hematoma [14].

Its volume calculation is performed by referencing the volume calculation formula in the open source software VTK, where the major class for the calculation of volume and area in VTK is `vtkMassProperties` [15]. The principle of this method is premised on the triangulation projection, which means that the model volume refers to the algebraic sum of the convex polyhedral volume enclosed by all triangular patches and the projection plane.

It is assumed that the coordinates of each triangle vertex are $P_0(x_0, y_0, z_0)$, $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, the length of the triangle edge are a , b , and c , the normal of the triangular patch is $u(u_x, u_y, u_z)$, and the center of gravity of the

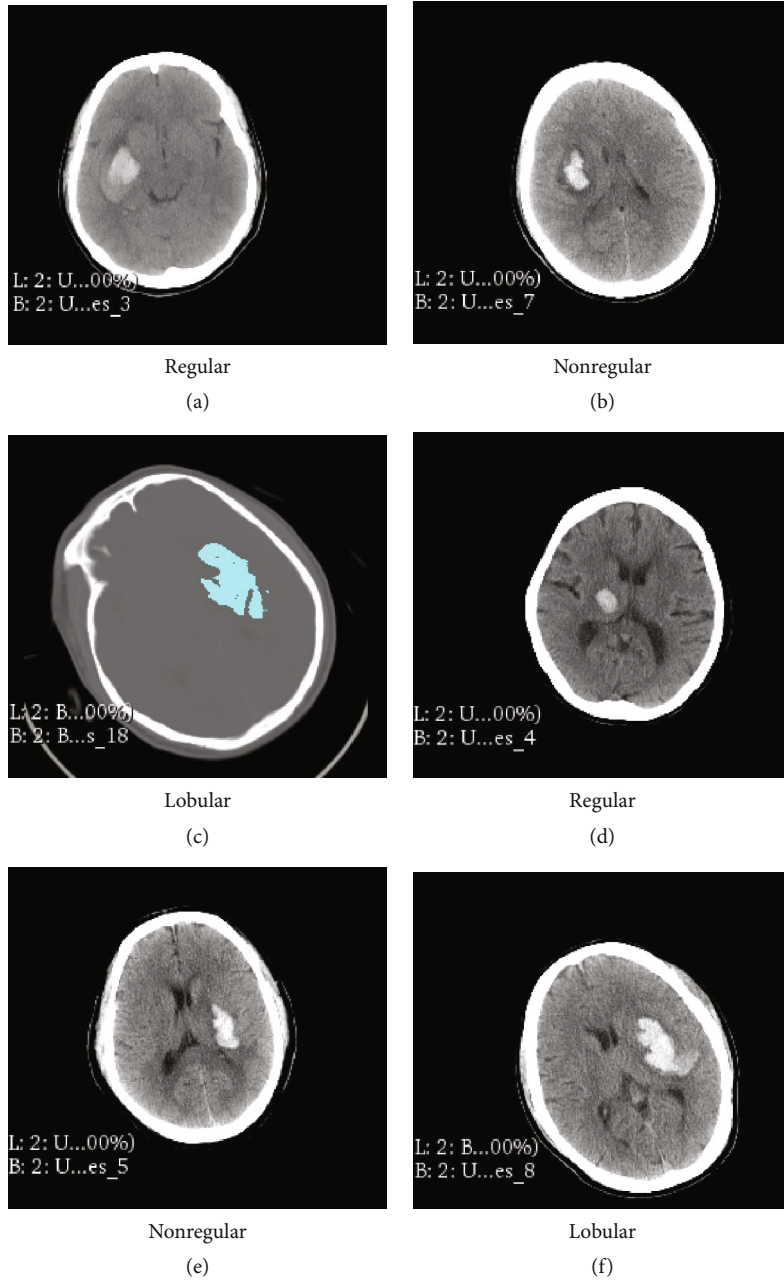


FIGURE 1: Hematoma shape classification.

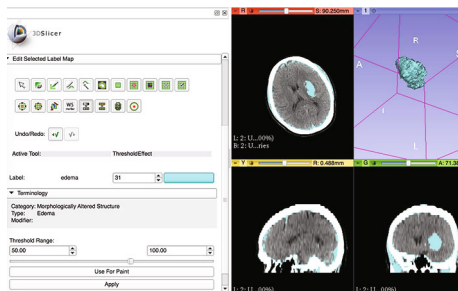


FIGURE 2: The hematoma model was reconstructed using 3D-Slicer software.

triangular patch is $avg(x, y, z)$. Then, the projection volume is expressed as

$$\begin{aligned} V_x &= area \cdot u_x \cdot avg_x, \\ V_y &= area \cdot u_y \cdot avg_y, \\ V_z &= area \cdot u_z \cdot avg_z, \end{aligned} \tag{1}$$

where $area = \sqrt{s \cdot (s - a) \cdot (s - b) \cdot (s - c)}$ means the triangular area and $s = (a + b + c)/2$. Therefore, the calculation formula for model volume is written as

TABLE 1: Volumes of pyramid and cubic models by voxelization, slice, and 3D-Slicer methods.

Model	Facets	Segments	Voxel unit	Slice method		Voxelization method	3D-Slicer method
				Method 1	Method 2		
Pyramid (closed)	2048	7	0.46875	6000.00	6000.18	6000.18	6000.00
		8	0.23438	6000.00	6000.04	6000.04	
Pyramid (unclosed)	1792	7	0.46875	6000.00	6000.18	6000.18	4625.18
		8	0.23438	6000.00	6000.04	6000.04	
Cubic (closed)	2304	7	0.15625	1000.00	1000.00	1000.00	1000.00
		8	0.07813	1000.00	1000.00	1000.00	
Cubic (unclosed)	2088	7	0.15625	1000.00	1000.00	1000.00	943.12
		8	0.07813	1000.00	1000.00	1000.00	

TABLE 2: Volumes of hematoma models by voxelization, slice, and 3D-Slicer methods.

Model	Facets	Segments	Voxel unit	Slice method		Voxelization method	3D-Slicer method
				Method 1	Method 2		
Hematoma 1 (closed)	13436	7	0.427	6131.83	6135.47	6135.47	6134.42
Hematoma 1 (unclosed)	13433	7	0.427	6131.83	6135.46	6135.46	6085.78
Hematoma 2 (closed)	2424	7	0.212	1056.25	1056.54	1056.54	1056.18
Hematoma 2 (unclosed)	2421	7	0.212	1056.25	1056.54	1056.54	929.30
Hematoma 3 (closed)	12582	7	0.599	12104.49	12107.78	12107.78	112107.68
Hematoma 3 (unclosed)	12579	7	0.599	12104.48	12107.77	12107.77	11872.43

$$V = |k_x \cdot V_x + k_y \cdot V_y + k_z \cdot V_z|, \quad (2)$$

where V_x , V_y , and V_z denote the sums of the projection volumes for the triangular patches, while k_x , k_y , and k_z represent the weights of each projection direction.

The measurement by 3D-Slicer software provides an accurate and simple method for the hematoma volume based on CT data. As shown in experiment, intracranial hematoma clearance (only about 2.71 ml left in average) is performed in combination with 3D-Slicer software, which achieves a 93.8% clearance rate [16]. However, it is discovered that the hematoma model required for the 3D-Slicer software method must be the closed triangular mesh model, and accurate normal information of the model surface needs to be known in advance. In some cases, the hematoma model may be non-closed or with incorrect normal information before the volume measurement. For example, when the boundary of a tumor surrounds that of the hematoma data, there is a possibility that the hematoma model is not closed. When the Marching Cube algorithm is applied to reconstruct the three-dimensional hematoma model, it will also give rise to the situation where the surface normal is inaccurate. Therefore, measuring the volume with the 3D-Slicer method in these cases will result in a significant error.

2.2. Slice Method. This method firstly performs layering on the three-dimensional hematoma model, then calculates the area of the corresponding section, and estimates the model volume based on the distance between adjacent planes. The

idea of slicing is to measure the volume of hematoma by the sum of quantitative measurement between consecutive sections; that is, the hematoma volume calculation formula is obtained as $V = \sum S_i \times h$, where S_i indicates the area of each CT slice and h denotes the thickness of the CT slice. The volume is determined based on the accumulation, which means that the three-dimensionally reconstructed hematoma is sliced, the adjacent section is supposed to form a round table, and the volume of all the sliced round tables is added as the total volume of the model. Different formulas can be obtained to calculate the volume of hematoma by applying different methods to calculate the volume of the round table V_i . For example, see the following.

Slice method 1: formula for each sliced round table volume is expressed as

$$V_i = \left(S_{i1} + S_{i2} + \sqrt{S_{i1}S_{i2}} \right) \frac{\text{step}}{3}, \quad (3)$$

where S_{i1} represents the upper floor area, S_{i2} indicates the lower floor area, and step refers to the interval between two slices.

Slice method 2: formula for each sliced round table volume is shown as follows:

$$V_i = (S_{i1} + S_{i2}) \frac{\text{step}}{2}, \quad (4)$$

where S_{i1} indicates the upper floor area, S_{i2} refers to the lower floor area, and step denotes the interval between two slices.

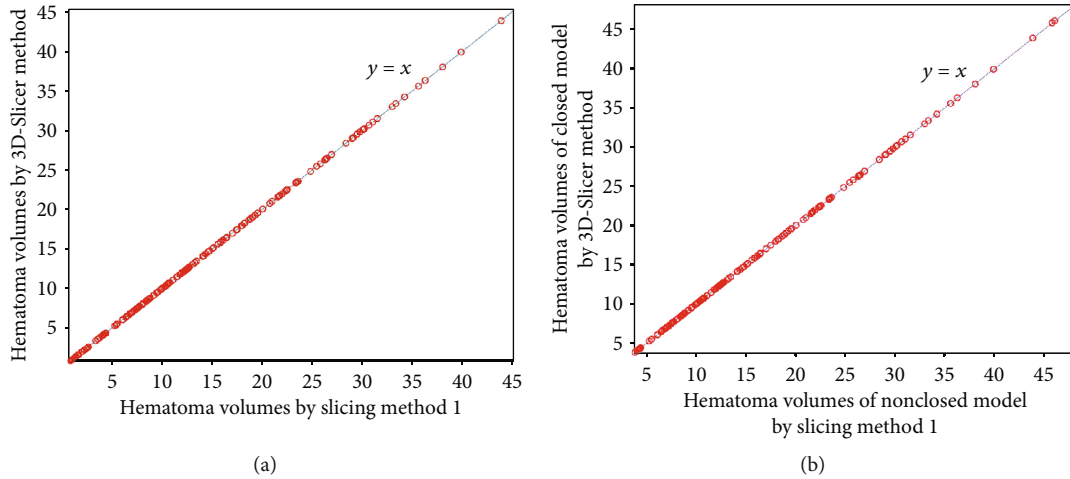


FIGURE 3: Comparisons of slice method 1 for measuring closed and nonclosed hematoma with the 3D-Slicer method.

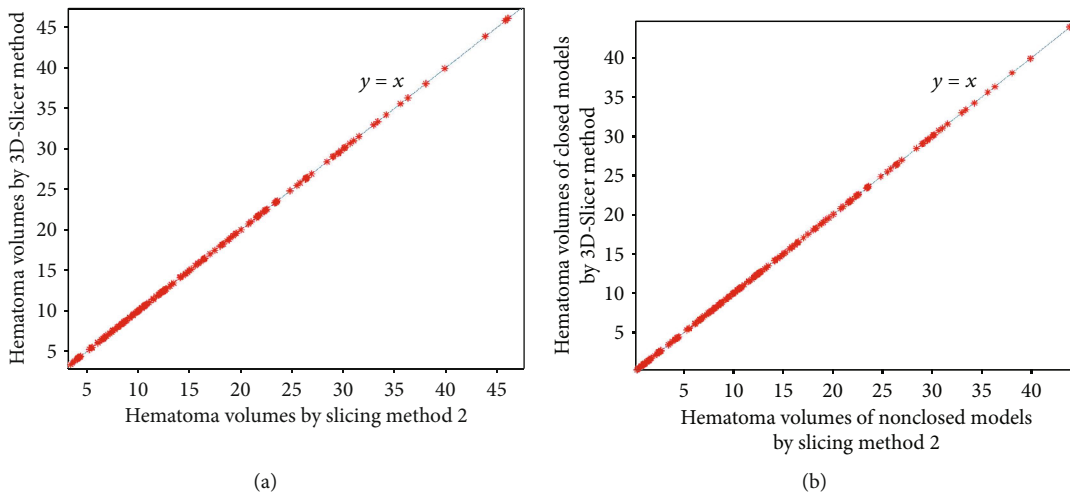


FIGURE 4: Comparisons of slice method 2 for measuring closed and nonclosed hematoma with the 3D-Slicer method.

From the aforementioned volume calculation formulas, it can be known that the calculation of the hematoma volume is related to the slice thickness (slice interval). A small thickness can improve accuracy, but this incurs more computation costs. Conversely, a large thickness reduces computation costs, but this causes accuracy to be compromised. Therefore, how to identify the appropriate slice thickness (interval) is a major problem facing the use of the slice method.

2.3. Voxelization Method. Voxelization provides a modeling method that approximates the geometric shape of a three-dimensional model by using spatial voxel units. These spatial voxels show similarity to pixels in a two-dimensional image and can be regarded as the expansion from a two-dimensional square area to a three-dimensional cube unit.

The realization of the voxelization method for the volume measurement involves two aspects. The octree operation is firstly implemented, and then, the calculation of boundary voxel volume is optimized. The major details are as follows:

- (1) *Implementation of the Octree Operation.* (a) The bounding box of the models is computed. (b) The octree is subdivided, the voxel with no intersection with the model mesh as a leaf voxel is marked, and the nonleaf voxel is subdivided again. (c) All leaf voxels are determined as either inside or outside the model. (d) The volume is defined as the sum of the volume of all inside leaf voxels and boundary voxels (i.e., the lowest nonleaf voxels).
- (2) *Optimization of the Boundary Voxel Volume.* The volume of the boundary voxel (the lowest nonleaf voxel) can be calculated using the slice method.

According to the voxelization method, spatial voxel units are required to approximate the three-dimensional model, and the computational complexity is higher compared to the 3D-Slicer software method (projection method) and the slice method. The computational accuracy of the voxelization method is determined by the size of the voxel unit and the

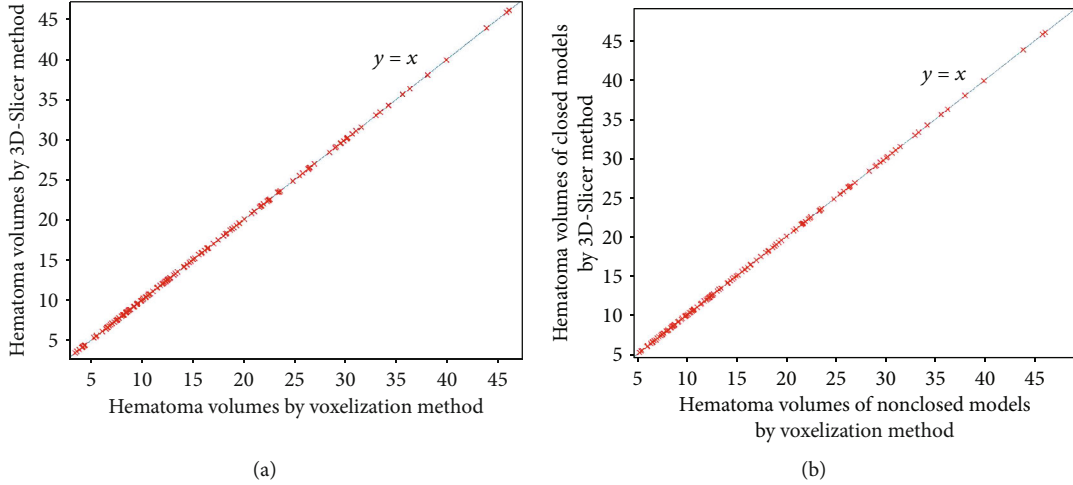


FIGURE 5: Comparisons of the voxelization method for measuring closed and nonclosed hematoma with the 3D-Slicer method.

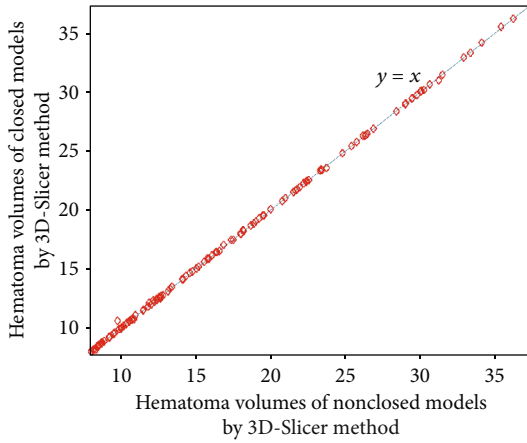


FIGURE 6: Comparisons of the 3D-Slicer method for measuring closed and nonclosed hematoma.

volume calculation of the boundary voxel. When the model volume is unknown, the results obtained by the voxelization method can be taken as the reference to compare the accuracy between the slice method and the 3D-Slicer software method.

2.4. Comparison of Three Measurement Methods

2.4.1. Standard Models with Known Volume. Firstly, a comparison is performed between the 3D-Slicer method, the slice method, and the voxelization method for the volume measurement of standard models, the volumes of which are known. For the slice method and voxelization method, we firstly calculate the volume with a large interval and a large voxel unit, respectively, and then reduce the interval and the voxel unit by a certain value until the calculated volume is relatively stable.

(1) Volume Measurement for Quadrangular Pyramid Model. It is assumed that the length of a pyramid is $l = 30$, the width is $w = 30$, and the height is $h = 20$. Then, the volume of a pyramid is calculated to be 6000 using the quadrangular

TABLE 3: Mean errors of closed hematoma (grouping by size) by slice and voxelization methods compared with the 3D-Slicer method.

	n	Slice method 1	Slice method 2	Voxelization method
First group	89	0.00125517	0.00220450	0.00082404
Second group	59	0.00405220	0.00067305	0.00067305
Third group	50	0.08751800	0.00175580	0.00175580

pyramid volume formula. The model is triangulated to construct a nonclosed 3D model and a closed 3D model with different triangular facets, respectively. Then, the 3D-Slicer method, the slice method, and the voxelization method are applied to measure the quadrilateral pyramid models, respectively. The results are indicated in Table 1.

(2) Volume Measurement for Cube. Suppose the length of the cube is $a = 10$, it can be known intuitively that the volume is 1000. Similarly, the model is triangulated to obtain a nonclosed 3D model and a closed 3D model with different triangular facets. Then, the 3D-Slicer method, the slice method, and the voxelization method are employed to measure the volumes, respectively. The results are presented in Table 1 as well.

2.4.2. Nonstandard Models with Unknown Volume. Two 3D models of hematomas stemming from the patients were first reconstructed using 3D-Slicer software. Besides, the 3D-Slicer method, the slice method, and the voxelization method are applied to measure the volumes of the nonclosed 3D model and the closed 3D model, respectively. The results are shown in Table 2.

2.4.3. Discussion

- (1) As for the closed cube model, the 3D-Slicer method is capable of ensuring accuracy. In comparison with the 3D-Slicer method, the results obtained by the slice

TABLE 4: Mean errors of nonclosed hematoma (grouping by size) by slice, voxelization methods, and 3D-Slicer method, compared with the 3D-Slicer method for closed models.

	n	Slice method 1	Slice method 2	Voxelization method	3D-Slicer method for nonclosed models
First group	89	0.00128517	0.00025000	0.00079449	0.01426753
Second group	59	0.00405797	0.00067915	0.00036763	0.04990034
Third group	50	0.00888480	0.00185180	0.00131520	0.00574640

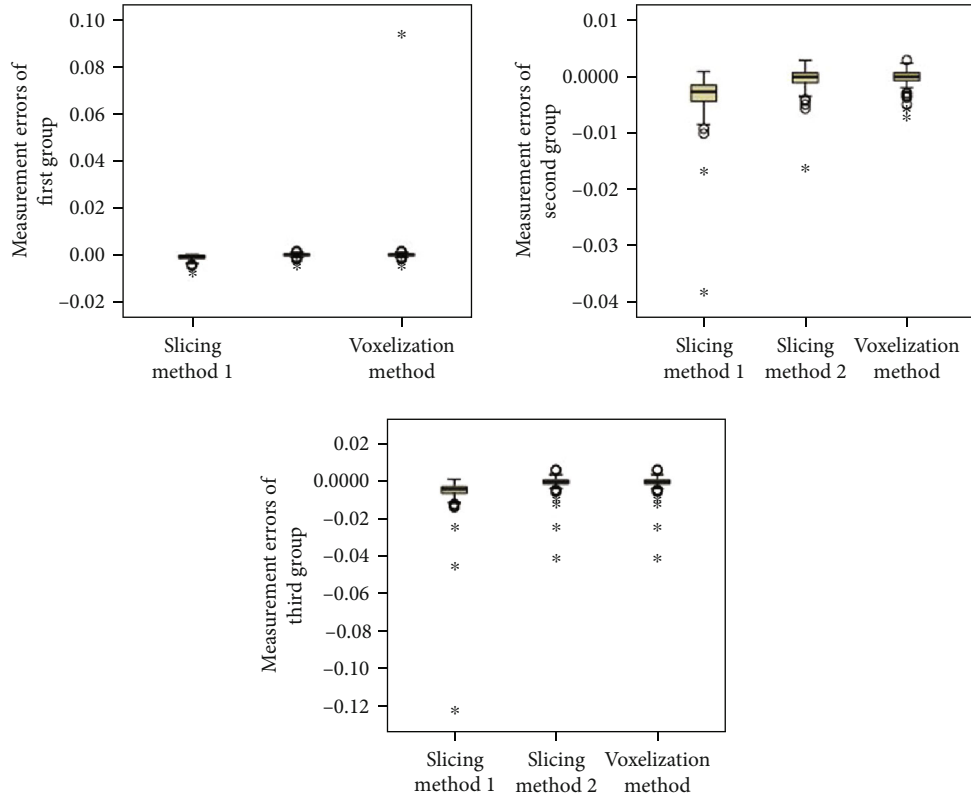


FIGURE 7: Distribution of measurement errors of closed hematoma models by slice method 1, slice method 2, voxelization method, and 3D-Slicer method (grouping by size).

method and the voxelization method show their accuracy and minor errors

- (2) The slice method and the voxelization method are consistent for either closed or nonclosed models. In addition, the results as obtained by the slice method and the voxelization method show similarity to the 3D-Slicer method applied for the closed model. However, the 3D-Slicer software method could result in a significant estimation error for the nonclosed model
- (3) The voxelization method and the slice method exhibit a low level of sensitivity to the number of triangular facets, and the volumes are identical for the model with different facets. The 3D-Slicer method shows sensitivity to the closeness of the model, and a small reduction of the facets will lead to a large error

3. Hematoma Volume Measurement and Comparison Analysis

For all patients, the 3D hematoma models were reconstructed using the 3D-Slicer software. Then, volume measurements were performed using the 3D-Slicer method, the slice method, and the voxelization method, respectively.

3.1. Materials and Methods

3.1.1. Patients. In this study, the patients admitted to the Affiliated Hospital of Hangzhou Normal University between December 2017 and January 2018 with diagnosis of spontaneous ICH were recruited. A total of 198 consecutive patients were recruited, including 132 male patients and 66 female patients, with the average age of 56.2 ± 28.8 . The patients with multiple sites of ICH were excluded from this study.

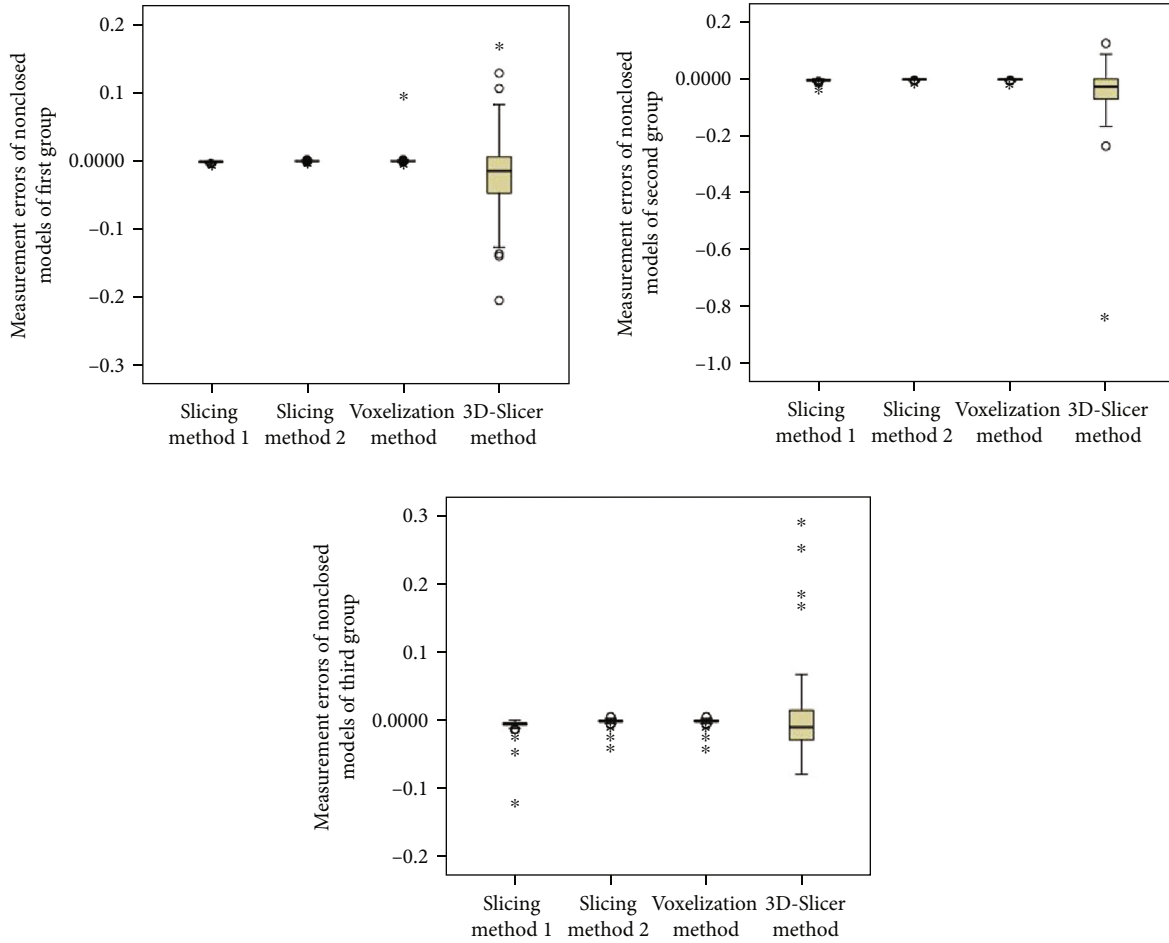


FIGURE 8: Distribution of measurement errors of nonclosed hematoma models by slice method 1, slice method 2, voxelization method, and 3D-Slicer method (grouping by size).

TABLE 5: Mean errors of closed hematoma (grouping by shape) by slice and voxelization methods compared with the 3D-Slicer method.

	<i>n</i>	Slice method 1	Slice method 2	Voxelization method
Regular group	76	0.00188145	0.00019487	0.00019487
Irregular group	85	0.00601259	0.00155753	0.00046388
Lobular group	37	0.00363027	0.00000216	0.00000216

All cases were included to the standard that the onset to head CT examination time is less than 24 hours.

3.1.2. *Imaging.* A total of 198 brain computed tomographic image data sets were acquired according to the hospital PACS system with the digital imaging standard in medicine format.

3.1.3. *Patient Groups.* The patients were split into 3 different groups depending on the hematoma size. Group 1 was comprised of 89 patients with volume < 10 ml, group 2 consisted

of 59 patients with volume ranging from 10 to 20 ml, and group 3 was made up of 50 patients with volume > 20 ml. Based on the maximal slice, the shape of the hematoma was classed into regular (round to ellipsoid) with smooth margins (76 cases), irregular with frayed margins (85 cases), and multilobular (37 cases).

3.1.4. *Statistical Analysis.* All of the statistical analyses were conducted with SPSS Statistics 21 (IBM Corporation, America). Moreover, GraphPad Prism was applied to draw charts. The relationship between the hematoma volume and the measurement method was analyzed by applying the simple linear correlation. Subsequent to the confirmation of distribution, the data were indicated as the mean ± SD, and unpaired *t*-test or 1-way ANOVA was conducted for comparison between different methods and groups, while the LSD method was applied to compare the two groups. A value of *P* < 0.05 was treated as statistically significant.

3.2. *Results.* We set the volumes of the closed models measured by the 3D-Slicer method as the standard values. The slice method (slice methods 1 and 2), voxelization method, and 3D-Slicer software method were compared using the closed hematoma and nonclosed hematoma models. For

TABLE 6: Mean errors of nonclosed hematoma (grouping by shape) by slice, voxelization methods, and 3D-Slicer method, compared with the 3D-Slicer method for closed models.

	n	Slice method 1	Slice method 2	Voxelization method	3D-Slicer method for nonclosed models
Regular group	76	0.00193513	0.00024842	0.00013171	0.01995789
Irregular group	85	0.00605059	0.00159624	0.00109388	0.01590706
Lobular group	37	0.00364405	0.00000946	0.00002892	0.02858676

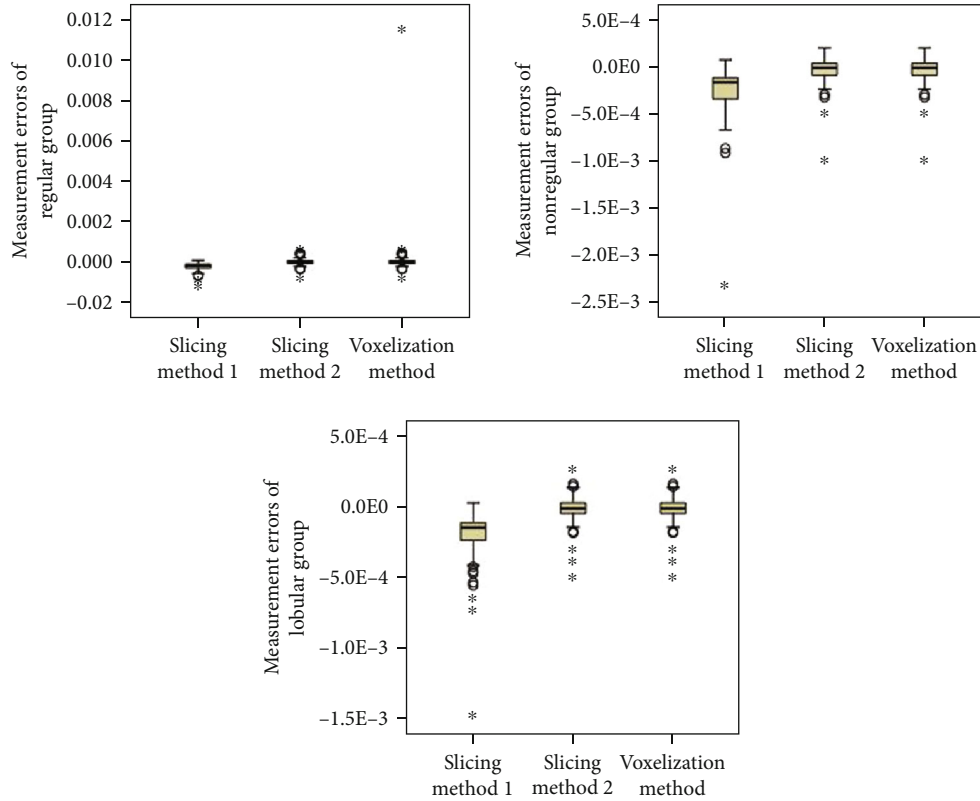


FIGURE 9: Distribution of measurement errors of closed hematoma models by slice method 1, slice method 2, voxelization method, and 3D-Slicer method (grouping by shape).

different methods, a simple correlation analysis was conducted under different models.

Figures 3–6 show the comparison results obtained by slice method 1, slice method 2, voxelization method, and 3D-Slicer method for the closed and nonclosed hematoma models. The results displayed in (a) are those for closed hematoma models. The scatter plots shown in Figures 3–6 have demonstrated that the results obtained from slice method 1, slice method 2, and voxelization methods are linearly related to those from the 3D-Slicer method. Moreover, their correlation is close to one. As revealed by the linear correlation analysis carried out by SPSS, the correlation coefficients between the slice methods 1 and 2, the voxelization method, and the 3D-Slicer method for the closed hematoma model were $r = 1$. There were statistically significant differences ($t = -5.627, P < 0.01$) observed for the results between the slice method, the voxelization method, and the 3D-Slicer method. From the results in Figures 3–5, we can see that the figures in (b) are similar to the results in the figures in (a).

That means the slice methods 1 and 2 and voxelization method are stable when the hematoma model was nonclosed, and the measurement results conform to those of the 3D-Slicer method when the hematoma model is closed. However, large errors will be caused by applying the 3D-Slicer method to the nonclosed hematoma model.

3.3. Analysis. When the patients are split into groups based on hematoma size, the statistical analyses are shown in Tables 3 and 4 and Figures 7 and 8. We can see that the mean errors of the results obtained by using the slice methods 1 and 2 and the voxelization method for closed and nonclosed hematoma measurements are broadly the same. The mean error of the voxelization method is less significant compared to the mean error of the slice method. The 3D-Slicer software method measures the nonclosed hematoma model with a significantly higher error than the slice method and the voxelization method. Specifically, for the first group, the error caused by the 3D-Slicer measurement for the nonclosed

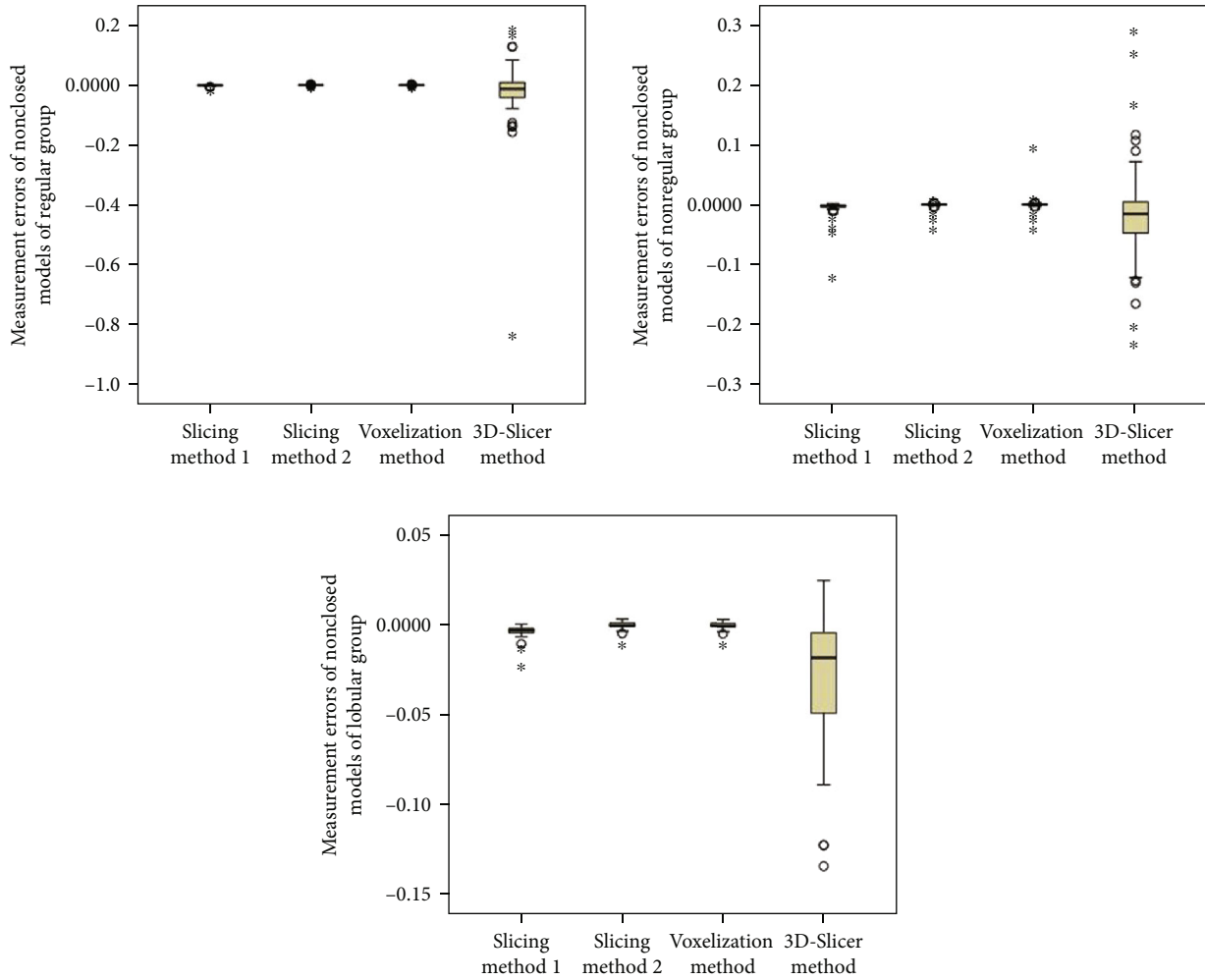


FIGURE 10: Distribution of measurement errors of nonclosed hematoma models by slice method 1, slice method 2, voxelization method, and 3D-Slicer method (grouping by shape).

model is 18 times that of the voxelization method. For the second group, the error by the 3D-Slicer measurement for the nonclosed model exceeds 100 times that of the voxelization method.

When the hematoma is classed by shape, the statistical analyses are shown in Tables 5 and 6 and Figures 9 and 10. It can be found out that slice methods 1 and 2 and voxelization methods are unaffected by the shape of the hematoma, and the measurement results obtained by these methods do not cause significant errors due to the irregular shape of hematoma. The mean errors as measured by slice method 1 and voxelization method show the same order of magnitude for the regular group and the irregular group. Besides, the voxelization method measures the hematoma of the lobulated group with less error, with its order of magnitude reaching 10^{-6} . Compared with the voxelization method, the 3D-Slicer method measures the nonclosed hematoma with significant errors. The error of the regular group, irregular group, and lobular group measured by the 3D-Slicer method is shown to be 151 times, 15 times, and nearly 1000 times that of the voxelization method, respectively.

4. Discussion

The accurate measurement of hematoma volume is of clinical significance as hematoma volume has been commonly used to correlate with treatment strategy, functional outcome, and mortality. It is inevitable for an inaccurately assessed hematoma volume to exert influence on the initial treatment decisions, thus leading to an undesirable outcome. Meanwhile, hematoma volume plays a crucial role in the prognosis of patients. The measurement of hematoma volume after cerebral hemorrhage can be taken as a potential indicator for prediction, which is of great significance to the clinical development of a sensible treatment. There are various forms of cerebral hemorrhage, especially for the presence of irregular hematoma, which makes it necessary to find an accurate method to determine the size of the volume based on different hematoma morphologies.

At present, the widely used methods to measure the volume of hematoma include the 3D-Slicer method and the Tada formula method. The Tada formula method is considered to be a rough calculation of hematoma due to its

inaccuracy in measuring irregular hematoma [7]. Moreover, the 3D-Slicer method is unaffected by the shape and location of the hematoma. Due to its real-time efficiency and low requirements on the operator, the 3D-Slicer method has been widely applied to measure the volume of hematoma.

For precision medicine, the 3D-Slicer method and other popular approaches to the volume measurement of hematoma were studied in this paper. The 3D-Slicer method caused a significant error in the measurement of the non-closed hematoma model or the model with wrong surface normal information. Nevertheless, the slice method and voxelization method were unaffected by closeness of the hematoma model nor the model with wrong normal information. Therefore, they can be treated as effective supplement methods of the 3D-Slicer method to measure the volume of hematoma. The drawbacks shown by slice and voxelization methods are the slice interval and the division of voxel units which affect both efficiency and accuracy. If there are significant errors between the 3D-Slicer method and the slice method (or the voxelization method), the voxelization method (or slice method) can be applied to validate the accuracy of these methods of measurement.

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 by the National Natural Science Foundation of China under Grant Nos. 11671009 and 11801513 and Zhejiang Provincial Natural Science Foundation of China under Grant Nos. LZ19A010002 and LQ18A010008.

References

- [1] S. Sacco, C. Marini, D. Toni, L. Olivieri, and A. Carolei, "Incidence and 10-year survival of intracerebral hemorrhage in a population-based registry," *Stroke*, vol. 40, no. 2, pp. 394–399, 2009.
- [2] J. P. Broderick, T. G. Brott, J. E. Duldner, T. Tomsick, and G. Huster, "Volume of intracerebral hemorrhage. A powerful and easy-to-use predictor of 30-day mortality," *Stroke*, vol. 24, no. 7, pp. 987–993, 1993.
- [3] M. T. C. Poon, A. F. Fonville, and R. al-Shahi Salman, "Long-term prognosis after intracerebral haemorrhage: systematic review and meta-analysis," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 85, no. 6, pp. 660–667, 2014.
- [4] S. Morris, R. M. Hunter, A. I. G. Ramsay et al., "Impact of centralising acute stroke services in English metropolitan areas on mortality and length of hospital stay: difference-in-differences analysis," *BMJ*, vol. 349, no. aug04 4, article g4757, 2014.
- [5] J. C. Hemphill III, S. M. Greenberg, C. S. Anderson et al., "Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 46, no. 7, pp. 2032–2060, 2015.
- [6] F. Milletari, "TOMAAAT: volumetric medical image analysis as a cloud service," 2018, <http://arxiv.org/abs/1803.06784>.
- [7] X. Xu, X. Chen, J. Zhang et al., "Comparison of the Tada formula with software slicer precise and low-cost method for volume assessment of intracerebral hematoma," *Stroke*, vol. 45, no. 11, pp. 3433–3435, 2014.
- [8] X. Lu and W. Lu, "Measurement of intracranial hematoma using the improved cubature formula," *Chinese Journal of Forensic Medicine*, vol. 26, no. 3, pp. 177–180, 2010.
- [9] A. Fedorov, R. Beichel, J. Kalpathy-Cramer et al., "3D Slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.
- [10] Z. Ma, X. Chen, Y. Huang et al., "MR diffusion-weighted imaging-based subcutaneous tumour volumetry in a xenografted nude mouse model using 3D Slicer: an accurate and repeatable method," *Scientific Reports*, vol. 5, no. 1, p. 15653, 2015.
- [11] G. Z. Cheng, R. San Jose Estepar, E. Folch, J. Onieva, S. Gangadharan, and A. Majid, "Three-dimensional printing and 3D slicer: powerful tools in understanding and treating structural lung disease," *Chest*, vol. 149, no. 5, pp. 1136–1142, 2016.
- [12] D. M. T. Devakumar, D. Devakumar, B. Sasidharan, S. R. Bowen, D. K. Heck, and E. J. J. Samuel, "Hybrid positron emission tomography segmentation of heterogeneous lung tumors using 3D Slicer: improved GrowCut algorithm with threshold initialization," *Journal of Medical Imaging*, vol. 4, no. 1, article 011009, 2017.
- [13] J. Egger, T. Kapur, A. Fedorov et al., "GBM volumetry using the 3D Slicer medical image computing platform," *Scientific Reports*, vol. 3, no. 1, pp. 1–7, 2013.
- [14] D. Petrovic, D. Mihailovic, S. Petrovic et al., "Asymptomatic flow of Rosai-Dorfman disease," *Vojnosanitetski Pregled*, vol. 71, no. 8, pp. 780–783, 2014.
- [15] <https://www.vtk.org/gitweb?p=VTK.git>.
- [16] G. Q. Xie, W. Shi, S. J. Chen et al., "Application of 3D-slicer in neuroendoscopic surgery for hypertensive intracerebral hemorrhage (in Chinese)," *Chinese Journal of Minimally Invasive Neurosurgery*, vol. 22, no. 3, pp. 109–111, 2017.

Research Article

CT-TEE Image Registration for Surgical Navigation of Congenital Heart Disease Based on a Cycle Adversarial Network

Yunfei Lu,¹ Bing Li,² Ningtao Liu,¹ Jia-Wei Chen ,¹ Li Xiao,³ Shuiping Gou ,¹
Linlin Chen,¹ Meiping Huang ,⁴ and Jian Zhuang⁵

¹Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China

²Department of Infectious Diseases, Ankang Central Hospital, Ankang 725000, China

³Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

⁴Catheterization Lab, Guangdong Cardiovascular Institute, Guangdong Provincial Key Laboratory of South China, Structural Heart Disease, Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510000, China

⁵Department of Cardiac Surgery, Structural Heart Disease, Guangdong General Hospital, Guangzhou 510000, China

Correspondence should be addressed to Jia-Wei Chen; jawaechan@gmail.com and Shuiping Gou; shpgou@mail.xidian.edu.cn

Yunfei Lu and Bing Li contributed equally to this work.

Received 25 March 2020; Revised 23 May 2020; Accepted 27 May 2020; Published 2 July 2020

Guest Editor: Lin Lu

Copyright © 2020 Yunfei Lu 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.

Transesophageal echocardiography (TEE) has become an essential tool in interventional cardiologist's daily toolbox which allows a continuous visualization of the movement of the visceral organ without trauma and the observation of the heartbeat in real time, due to the sensor's location at the esophagus directly behind the heart and it becomes useful for navigation during the surgery. However, TEE images provide very limited data on clear anatomically cardiac structures. Instead, computed tomography (CT) images can provide anatomical information of cardiac structures, which can be used as guidance to interpret TEE images. In this paper, we will focus on how to transfer the anatomical information from CT images to TEE images via registration, which is quite challenging but significant to physicians and clinicians due to the extreme morphological deformation and different appearance between CT and TEE images of the same person. In this paper, we proposed a learning-based method to register cardiac CT images to TEE images. In the proposed method, to reduce the deformation between two images, we introduce the Cycle Generative Adversarial Network (CycleGAN) into our method simulating TEE-like images from CT images to reduce their appearance gap. Then, we perform nongrid registration to align TEE-like images with TEE images. The experimental results on both children' and adults' CT and TEE images show that our proposed method outperforms other compared methods. It is quite noted that reducing the appearance gap between CT and TEE images can benefit physicians and clinicians to get the anatomical information of ROIs in TEE images during the cardiac surgical operation.

1. Introduction

Congenital heart disease accounts for 28% of all congenital malformations. In China, the incidence of congenital heart disease is 0.4-1% among the infants approximately 150000-200000 newborns annually. Transesophageal echocardiogra-

phy (TEE) is an imaging of congenital heart disease. The sensor is usually placed at the esophagus directly behind the heart and allows a continuous visualization of the movement of the visceral organ without trauma and the observation of the heartbeat in real time. Therefore, it has become an essential tool for most interventional cardiologists in

navigation during the surgery. However, TEE images usually provide very limited data on clear anatomically cardiac structures, which makes TEE images difficult to interpret. While high-resolution computed tomography (CT) images can provide anatomical information of cardiac structures, a CT scanner is not flexible to move and cannot be used in surgery. Despite the limitations of a CT scanner, the sufficient anatomical information of organs in CT images can benefit us to transfer the anatomical information to TEE images by the registrations between CT and TEE images.

Recently, similar ideas have been proposed to transfer the anatomical information of prostate from magnetic resonance (MR) image to CT images via registration methods. Cao et al. [1] developed a bidirection registration method from MR images to CT images via simulating CT and MR images with a structural random forest. Fundamentally, the similarity measurement is a core issue in registration. Recently, Cao et al. [2] proposed a similarity measurement for a registration-based convolution neural network. Fan et al. [3] proposed a registration method-based adversarial similarity network.

However, the registration between CT and TEE images is more challenging. For one, the quality of TEE images is much lower than other modalities, especially for children's cardiac TEE images. For others, the appearance patterns of CT and TEE images are quite different, even from the same subject. This big gap on appearance between two images makes most typical nongrid registration methods [4] be inefficient. It is quite significant in surgical operation to reduce the appearance gaps between these two modalities of medical images.

To reduce the appearance gap cross-image modalities, a generative adversarial network (GAN) [5] has been proposed to generate an image following a distribution. Nie et al. [6] introduced the GAN in medical image synthesis. Tanner et al. [7] proposed a GAN for MR-CT deformable registration. Yan et al. [8] proposed an end-to-end adversarial network for MR and TRUS image fusion. However, most GANs need paired data to train and it is difficult to collect sufficient number of paired medical image data. Wolterink et al. [9] proposed a GAN to synthesize CT image from an MR image which is trained by unpaired image data. On the other hand, Zhu et al. [10] proposed the Cycle Generative Adversarial Network (CycleGAN) for the image translation from different domains. Compared with the GAN, it can be efficiently trained by the unpaired image data, which are more easily collected. This advantage can benefit to the cross-domain medical image registration. In particular, for the cross-modal medical images with big appearance and morphological gaps, CycleGAN can be introduced to them.

Inspired by this, in this paper, we proposed a learning-based registration method to align CT and TEE images of the same subject. In the proposed method, we introduce a cycle adversarial network to generate TEE-like images from the corresponding CT images and CT-like images from the corresponding TEE images, with which reduces the appearance gap between two modalities. After that, the nongrid registration methods are applied to align TEE-like images with TEE images and CT images with CT-like images, respectively, to obtain two deformation fields. The final registration results

can be obtained by averaging these two deformation fields. It is quite significant to reduce the appearance gap between CT and TEE images which can benefit physicians and clinicians to get the anatomical information of ROIs in TEE images during the cardiac surgical operation.

The rest of this paper can be organized as follows. The proposed method will be introduced in Section 2, and the experimental results will be presented in Section 3. Finally, the conclusion will be made in Section 4.

2. Method

2.1. Datasets and Preprocessing. In this study, we collect a dataset of paired CT images and TEE images of 12 subjects with congenital heart disease. They include 2 adults and 10 teenagers. All the CT images are scanned by the SIEMENS CT VA1 DUMMY scanner, and their sizes of the XoY planes are 512×512 . Their spacing variously ranges from $0.28 \times 0.28 \text{ mm}^2$ to $0.45 \times 0.45 \text{ mm}^2$, while the TEE images are scanned by Philips iE33 Medical Imaging System with the size of the XoY plane being 800×600 . Their spacing variously ranges from $0.15 \times 0.15 \text{ mm}^2$ to $0.28 \times 0.28 \text{ mm}^2$. Due to the difficulty of data acquisition, TEE images for all patients only include 22 standard sections of the heart. Before training the CycleGAN, we resample all the images into the same spacing $0.45 \times 0.45 \text{ mm}^2$ and crop all the image into the same size.

2.2. CT-TEE Image Generation by CycleGAN. The CycleGAN is a weakly supervised GAN that is trained by unpaired samples achieving the cross-domain image translation with different distributions. In this section, we introduce it to CT-TEE image generation to reduce the appearance and morphological gap between them. The architecture of CT-TEE image generation with CycleGAN can be illustrated in Figure 1.

As shown in Figure 1, the architecture contains two directions of image generation: from a TEE image to generate a CT-like image and from a CT image to generate a TEE-like image. First, starting a TEE image as illustrated the top-left "BEGIN" in the figure, a CT-like image can be generated by the TEE-CT generator and then, a cyclic TEE image will be generated by the CT-TEE generator. The generated cyclic TEE image is forwarded to a discriminator giving a measure how it is like the real TEE image.

On the other thread, starting a CT image as illustrated the bottom-right "BEGIN" in the figure, a TEE-like image can be generated by the CT-TEE generator and then, a cyclic CT image will be generated by the TEE-CT generator. Similarly, the generator cyclic CT image is forwarded to a discriminator to measure how it is like the real CT image. The generator here is designed as the one in [11].

The discriminator here can be illustrated in Figure 2. It is shown that the discriminator includes two parts. The one is used to determine whether the generated image is a real or fake, and the other is used to classify whether the generated image is a TEE or a CT image. The discriminative loss will be used to update the parameters of the whole network. It can be computed as follows:

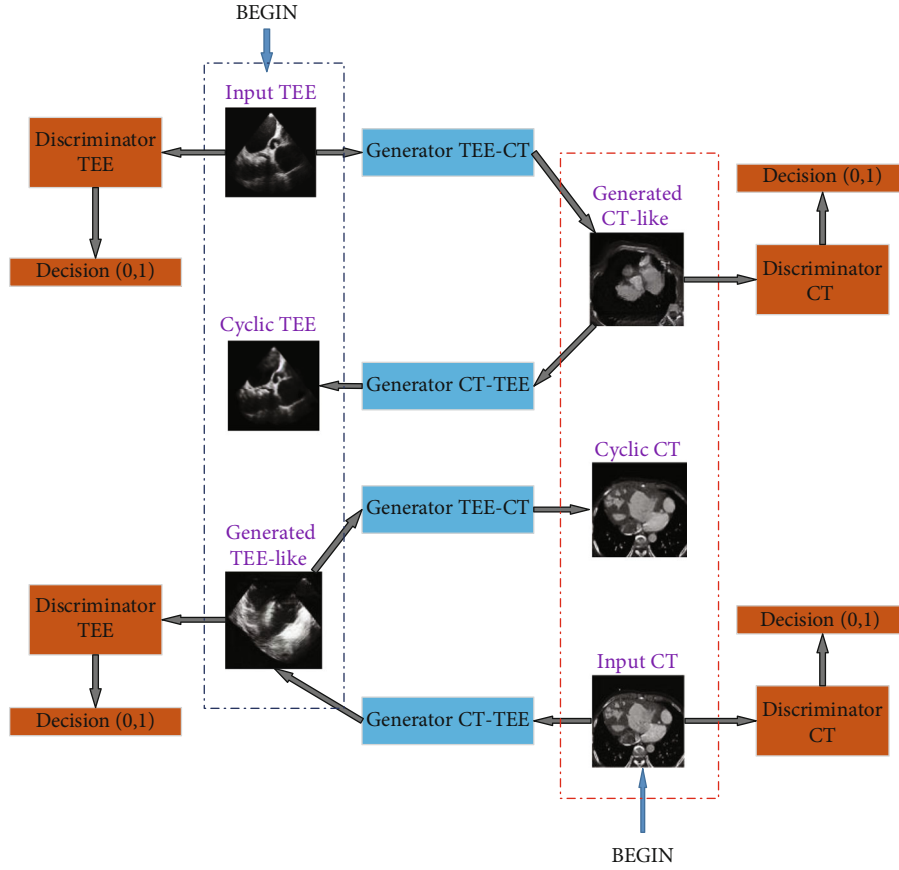


FIGURE 1: The architecture of CT-TEE image generation.

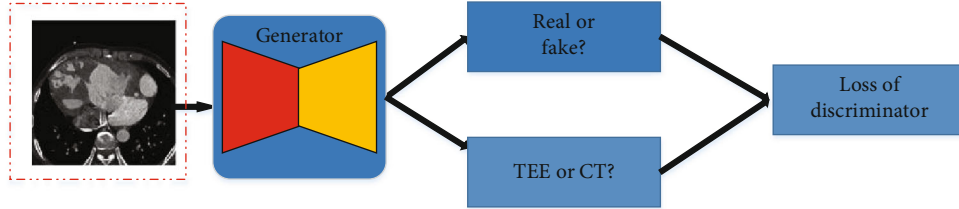


FIGURE 2: An illustration of discriminator.

$$\begin{aligned}
 L_{GAN}(G, D_1, D_2, X, Y) = & E_{y \sim P(Y)} [(D_1(y) - 1)^2] \\
 & + E_{y \sim P(Y)} [D_1(G(x))^2] \\
 & + \alpha \cdot (E_{y \sim P(Y)} [(D_2(y) - 1)^2]) \\
 & + E_{y \sim P(Y)} [D_2(G(x))^2],
 \end{aligned} \quad (1)$$

where G and D_1, D_2 denote the generator and discriminator, respectively. X and Y denote the CT and TEE domains, respectively. α is a weight which balances two parts of discriminative loss.

2.3. CT-TEE Registration Based on Generated Images. It is a great challenge to perform CT-TEE registration due to the appearance and morphological gaps between CT and TEE images. TEE and CT images are shown in Figures 3(a) and

3(d), respectively. It can be observed that the cardiac in TEE and CT images appears to have a different appearance and morphology. In the above paragraphs, we have introduced the CycleGAN to perform CT-TEE image translation, reducing the appearance and morphological gaps.

In this section, we will propose a learning-based registration method to reduce the appearance gap between TEE images and CT images by introducing a cycle adversarial network (CycleGAN). The whole framework can be illustrated by Figure 4.

In this framework, we first generate a TEE-like image from the corresponding CT image, and a CT-like image from the corresponding TEE image with the trained CycleGAN. By this network, the generated TEE-like images with the same morphological feature with the corresponding CT images of the same subject possess the similar appearance

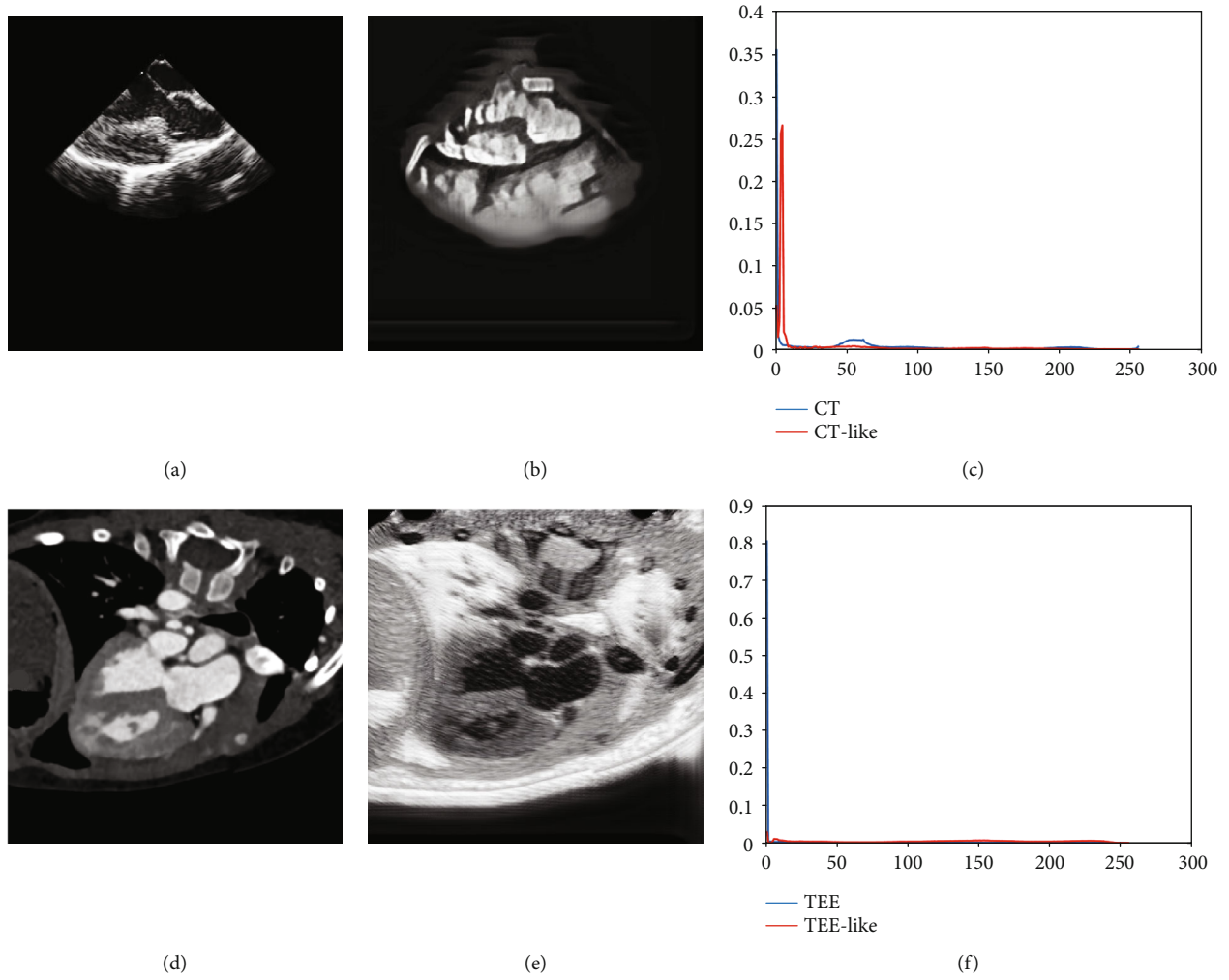


FIGURE 3: Comparisons between original images and generated images. (a) Original TEE image. (b) CT-like image. (c) Histograms of TEE image and CT-like image. (d) Original CT image. (e) TEE-like image. (f) Histograms of CT image and TEE-like image.

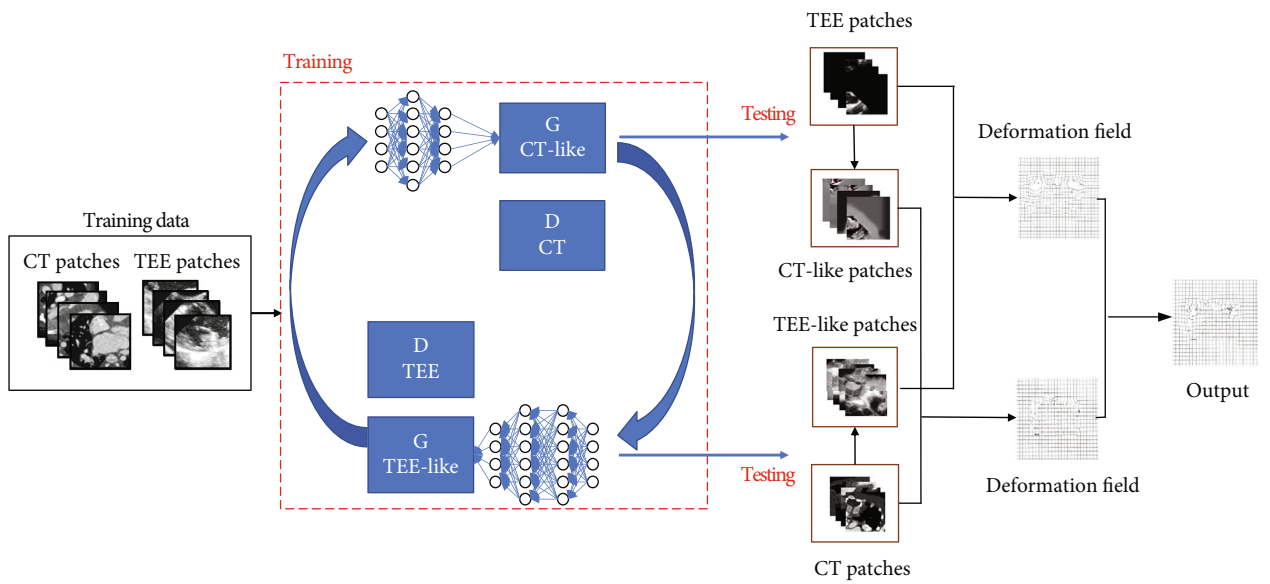


FIGURE 4: The whole framework of the proposed method.

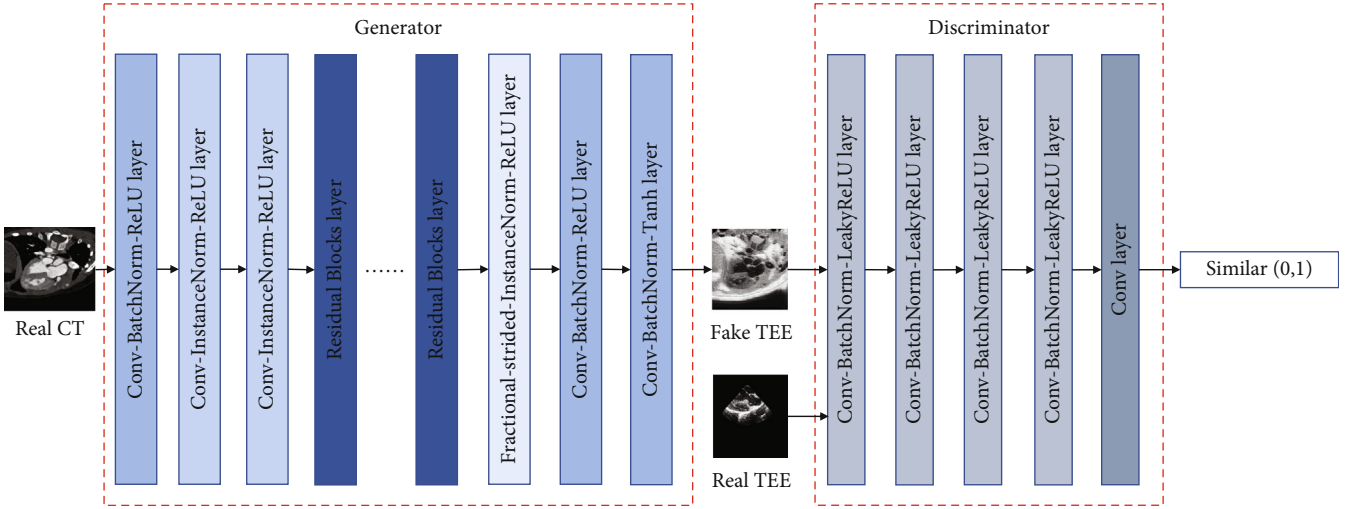


FIGURE 5: The network structure of CycleGAN.

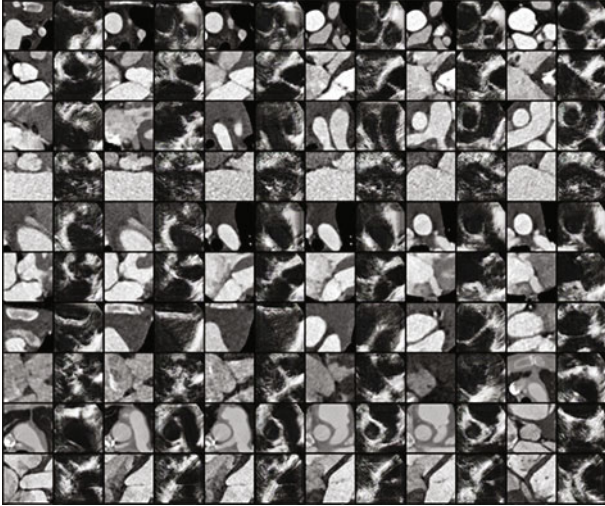


FIGURE 6: The TEE-like images generated from CT images.

features to TEE images, while the generated CT-like images with the same morphological features with the corresponding TEE images, possess the similar appearance features to CT images.

Then, we can perform the registration between CT images and TEE images through aligning TEE-like images with TEE images and CT images with CT-like images, respectively.

2.4. Training Network. The structure of the adversarial network we employ in this paper is illustrated in Figure 5. In this network, we employ a convolution network with 7 convolution-BatchNorm-ReLU layers as the generator. On the other hand, we introduce another convolution network with 5 Convolution-BatchNorm-ReLU layers as the discriminator.

To train CycleGANs between CT images and TEE images, we randomly draw patches with size 256×256 from the collected dataset, where 9750 and 15020 ones are collected from CT images and TEE images, respectively. The model is trained on a PC with Intel i7 9700K CPU,

64GB memory and NVIDIA TITAN Xp GPU. The platform is built on TensorFlow with Ubuntu 16.04. In the training process, the network is optimized by the Adam algorithm, where the batchsize and learning rate are set as 2 and 2×10^{-4} , respectively. The training loss becomes stable after 19055 iterations.

3. Experiment Results

In the following experiments, the proposed model is verified by the leave-one-out validation; i.e., a patient is taken out first as a testing subject and the rest subjects are employed to train the CycleGAN.

3.1. Image Generation Results. The comparisons between the original image and generated image are shown in Figure 3. The results show that the CT-like images generated from TEE images possess not only the same morphologic structure as original TEE images but also the similar appearance feature with CT images. The TEE-like image and TEE image tell the same story. Furthermore, the histogram in Figures 3(c) and 3(f) shows that the intensity distribution of the CT-like image is quite similar to the CT image, while the TEE-like image follows the similar distribution to the TEE image.

More TEE-like images are shown in Figure 6, where CT images are shown in odd columns and the corresponding TEE-like images are shown in even columns. It can be observed from numerous results that the framework in Figure 1 can reduce the gap between CT and TEE images which provides a good precondition for the following registration.

3.2. Image Registration Results. In the registration process, we employ the FLIRT [12] method as grid registration followed by Powell and Demons [4] as non-grid registration methods. Through the deformation field, we map the labels from CT images with the labels in TEE images of the same organ.

We show the registration results of two subjects in Figure 7. We applied FLIRT and Demons registrations to perform registration between CT and TEE images and

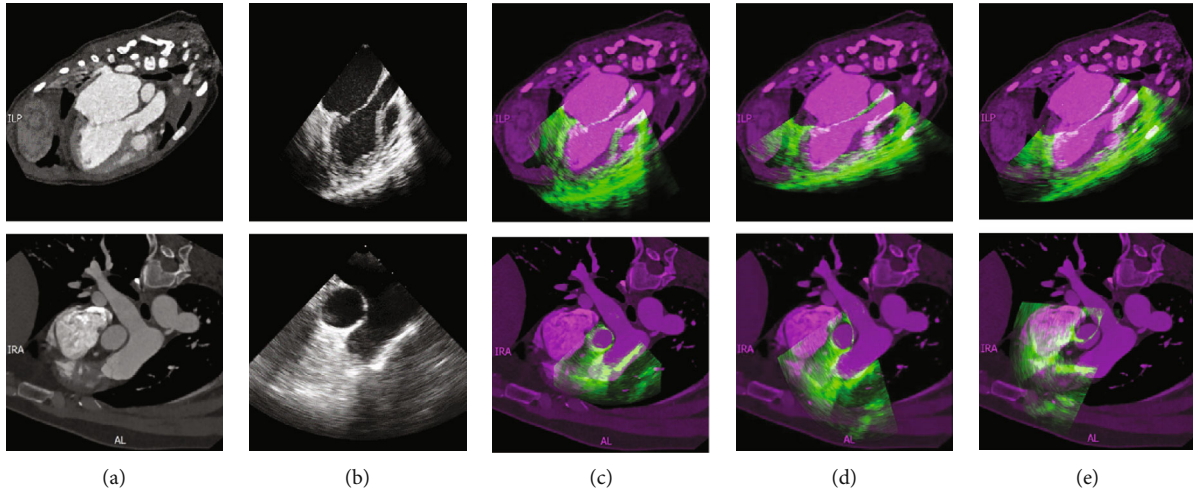


FIGURE 7: The registration result. (a) Original CT image. (b) Original TEE image. Results of (c) the proposed method. (d) FLIRT. (e) Demons registration.

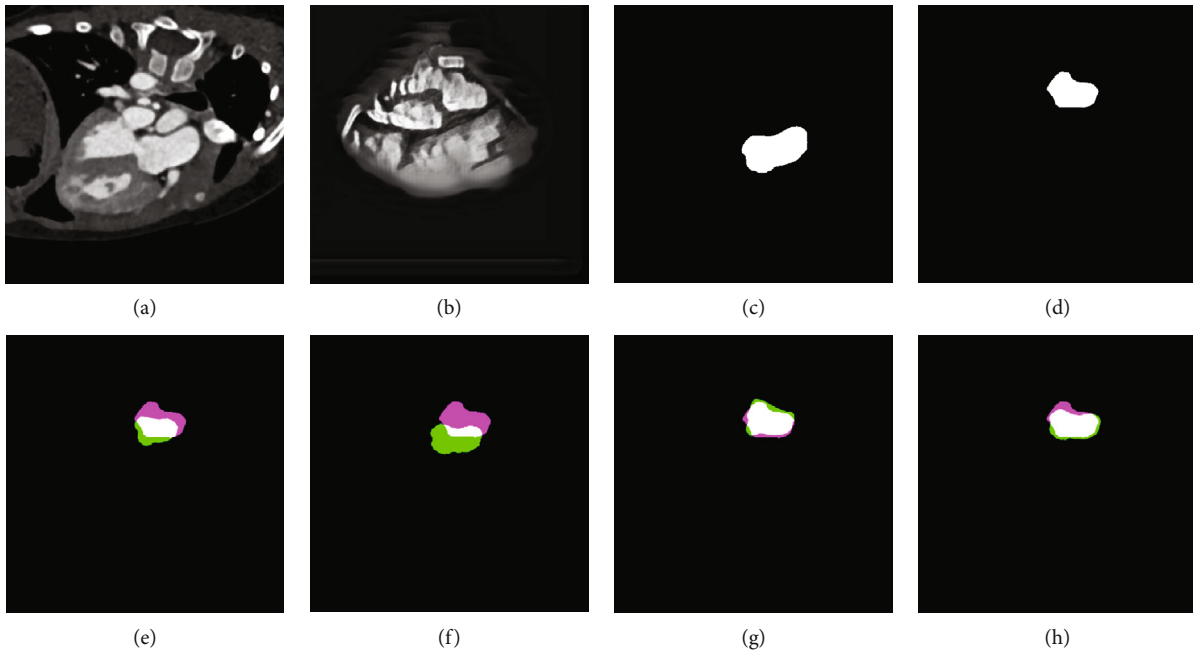


FIGURE 8: The registration of an adult subject. (a) Original CT image. (b) CT-like image. (c) Ground truth on CT domain. (d) Ground truth on TEE domain. (e) Demons registration on original image. (f) Powell registration on original image. (g) Demons registration on generated image. (h) Powell registration on generated image.

overlapped the CT image and the aligned TEE image. The results are shown in the panel (d) and (e). We can observe that two images are not well-aligned in both two subjects due to their appearance and morphological gaps. Instead, through CycleGAN, we perform the registration between the CT-like image and the CT image. It is shown in the panel (c) that we can get better alignment performance.

We also show another two examples including an adult subject and a teenager subject of registration results in Figures 8 and 9, respectively, where in the panels (e)-(h), the pink indicates the ground truth of TEE domain, the green indicates the mapped labels from CT domain to TEE domain, and the white indicates the overlap of them. It can be shown

from Figures 8(e) and 8(f) that the overlapping regions in Figures 8(g) and 8(h) are larger than (e) and (f), respectively. It means that the better alignments are obtained on generated images than original images. Similar visual results can also be obtained on the teenager subject. Our proposed method can tackle with both adult subjects and teenager subjects.

Furthermore, the registration performances are evaluated by the Dice ratio (DR), Hausdorff distance with percentile of 95% (HD95) [13] and the average symmetric surface distance (ASD) between the mapped labels through the deformation field and the ground truth in TEE domain. Hausdorff distance with percentile of 95% (95% HD) is based on the calculation of the 95th percentile of the

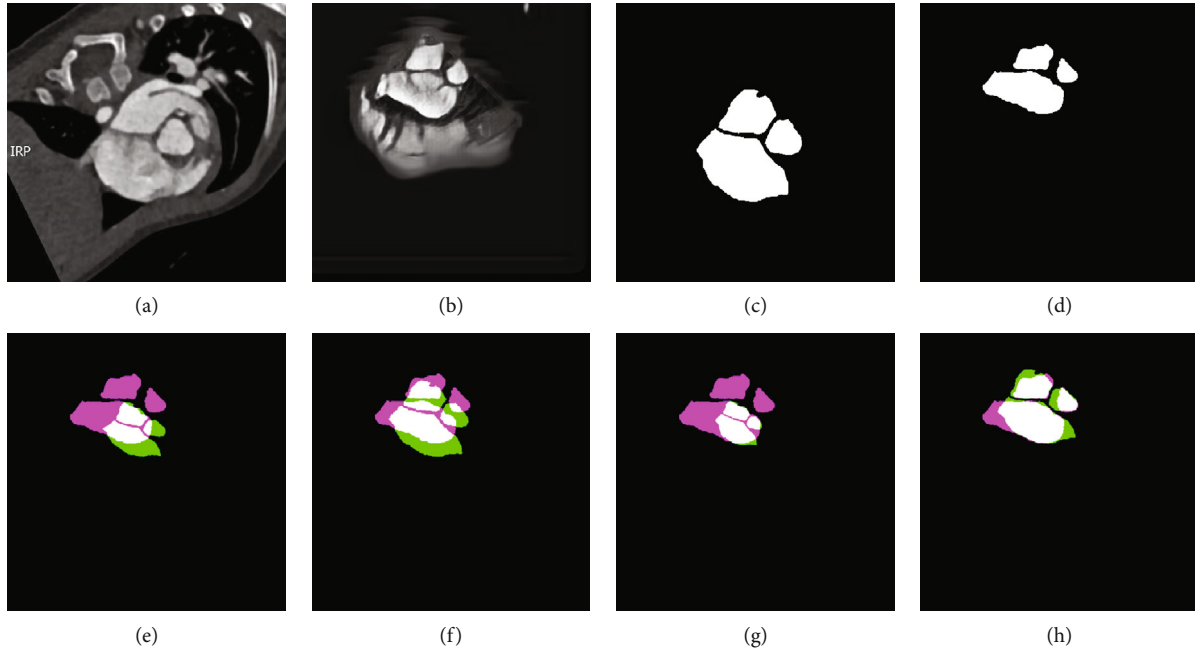


FIGURE 9: The registration of a teenager subject. (a) Original CT image. (b) CT-like image. (c) Ground truth on CT domain. (d) Ground truth on TEE domain. (e) Demons registration on original image. (f) Powell registration on original image. (g) Demons registration on generated image. (h) Powell registration on generated image.

TABLE 1: Quantitative evaluations of registrations.

	DR	HD95	ASD
Demons on OI	0.31 ± 0.22	76.50 ± 52.17	32.91 ± 37.98
Demons on GI	0.75 ± 0.09	33.83 ± 16.17	9.83 ± 3.75
Powell on OI	0.26 ± 0.29	84.81 ± 48.41	40.69 ± 28.38
Powell on GI	0.78 ± 0.05	32.16 ± 16.89	8.61 ± 4.62

Hausdorff distance between boundary points in two sets, Hausdorff distance means the maximum value of the shortest distance from a point set to another point set. Better results can get higher DR and lower HD95 and ASD. We show the mapped labels from the CT image and the ground truth on the TEE image, where we compare the Demons and Powell methods on both original images (OI) and generated images (GI). We list the means and standard deviations of evaluations on both teenager and adult registration results in Table 1. In terms of Dice Ratio, the values on the GI are much larger than the ones on the original images, while in terms of HD95 and ASD, the registrations on GI get lower ones than those on OI. It is demonstrated that classic nongrid registration methods are almost noneffective on original CT and TEE images alignment. Instead, the performances of registrations on generated images improve significantly than original images.

It is demonstrated from all of the above visual results and quantitative evaluations that our proposed method is effective for CT-TEE registration. It is benefited from the CycleGAN that reduces the appearance gaps between CT and TEE images. However, Demons is a nonrigid registration

method that is easy to fall into local optimum, and the registration result is usually not good when there is a certain difference in shape and size between two images, panels (g) and (h) in Figure 9 show the limitation of Demons.

4. Conclusion

It is quite significant to reduce the appearance gap between CT and TEE images which can benefit physicians and clinicians to get the anatomical information of ROIs in TEE images during the cardiac surgical operation. In this paper, we develop a CycleGAN-based registration method to align CT images with TEE images. The CycleGAN reduces the appearance gap between CT images and TEE images. Our proposed method is verified on 12 pairs of CT-TEE images. Both visual results and quantitative evaluations show that the performance of the registration with generated images is better than original images. It indicates that our proposed method can get reasonable registration results between CT and TEE images with the challenges of large appearance.

Data Availability

The data used in the article are from the Department of Cardiac Surgery, Structural Heart Disease, Guangdong General Hospital, Guangzhou, China.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by the Natural Science Foundation of Shaanxi Province under Grant No. 2019ZDLGY03-02-02, Research Plan of Improving Public Scientific Quality in Shaanxi Province, No. E219360001, and the Fundamental Research Funds for the Central Universities, No. JC2001.

References

- [1] X. Cao, Y. Gao, J. Yang, G. Wu, and D. Shen, "Learning-based multimodal image registration for prostate cancer radiation therapy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 1–9, Springer, 2016.
- [2] X. Cao, J. Yang, J. Zhang et al., "Deformable image registration based on similarity-steered cnn regression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 300–308, Springer, 2017.
- [3] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, "Adversarial similarity network for evaluating image alignment in deep learning based registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 739–746, Springer, 2018.
- [4] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.
- [5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, <http://arxiv.org/abs/1511.06434>.
- [6] D. Nie, R. Trullo, J. Lian et al., "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.
- [7] C. Tanner, F. Ozdemir, R. Profanter, V. Vishnevsky, E. Konukoglu, and O. Goksel, "Generative adversarial networks for mr-ct deformable image registration," 2018, <http://arxiv.org/abs/1807.07349>.
- [8] P. Yan, S. Xu, A. R. Rastinehad, and B. J. Wood, "Adversarial image registration with application for mr and trus image fusion," in *International Workshop on Machine Learning in Medical Imaging*, pp. 197–204, Springer, 2018.
- [9] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum, "Deep mr to ct synthesis using unpaired data," in *International workshop on simulation and synthesis in medical imaging*, pp. 14–23, Springer, 2017.
- [10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, Venice, Italy, 2017.
- [11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.
- [12] B. Fischer and J. Modersitzki, "Flirt: A flexible image registration toolbox," in *International Workshop on Biomedical Image Registration*, pp. 261–270, Springer, 2003.
- [13] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

Research Article

Systematic Identification of lncRNA-Associated ceRNA Networks in Immune Thrombocytopenia

Zhenwei Fan,¹ Xuan Wang ,² Peng Li ,³ Chunli Mei,¹ Min Zhang ,¹ Chunshan Zhao,¹ and Yan Song¹

¹Nursing College of Beihua University, Jilin 132013, China

²Department of Hematology, Affiliated Hospital of Beihua University, Jilin 132013, China

³Department of Oncology, Jilin Central Hospital, Jilin 132000, China

Correspondence should be addressed to Xuan Wang; xinqiancanhqh@163.com and Peng Li; gongyou4472720@163.com

Received 27 March 2020; Accepted 11 May 2020; Published 30 June 2020

Guest Editor: Tao Huang

Copyright © 2020 Zhenwei Fan 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.

Primary immune thrombocytopenia (ITP) is an autoimmune disease. However, the molecular mechanisms underlying ITP remained to be further investigated. In the present study, we analyzed a series of public datasets (including GSE43177 and GSE43178) and identified 468 upregulated mRNAs, 272 downregulated mRNAs, 134 upregulated lncRNAs, 23 downregulated lncRNAs, 29 upregulated miRNAs, and 39 downregulated miRNAs in ITP patients. Then, we constructed protein-protein interaction networks, miRNA-mRNA and lncRNA coexpression networks in ITP. Bioinformatics analysis showed these genes regulated multiple biological processes in ITP, such as mRNA nonsense-mediated decay, translation, cell-cell adhesion, proteasome-mediated ubiquitin, and mRNA splicing. We thought the present study could broaden our insights into the mechanism underlying the progression of ITP and provide a potential biomarker for the prognosis of ITP.

1. Introduction

Primary immune thrombocytopenia (ITP) is an autoimmune disease characterized by a decrease in platelets due to platelet destruction and insufficient platelet production [1, 2]. Previous studies had showed the increasing antiplatelet antibodies produced by B cells, and the aberrant functions of T lymphocytes were involved in regulating the progression of ITP [3]. However, the mechanisms regulating ITP progression remained to be further investigated.

In the past decades, increasing evidence showed more than 90% human genome could not be translated to proteins. Noncoding RNAs, such as miRNAs and lncRNAs, played important roles in the progression of human diseases [4]. miRNAs were a type of ncRNAs with 19-25 bps in length and regulated gene expression and protein translation by targeting 3'-UTR of mRNAs. Previous studies showed miRNAs were dysregulated and associated with the regulation of

ITP. For example, miR-99a expression was overexpressed in CD4+ cells [5], while expression of miR-182-5p and miR-183-5p was overexpressed in ITP. MIR130A was downregulated and suppressed TGFB1 and IL18 in ITP [6]. Meanwhile, MIR409-3p was also reported to be reduced in ITP samples [7]. Long noncoding RNAs (lncRNAs) are a class of ncRNAs longer than 200 nucleotides with no protein-coding potential. The roles of lncRNAs in autoimmune diseases were also implicated. Wang et al. found that lncRNA TMEVPG1 expression was lower than that in healthy control samples [8]. Liu et al. identified a total of 1177 and 632 lncRNAs were significantly upregulated or downregulated in ITP patients compared to normal samples [9].

In the present study, we screened differently expressed mRNAs, miRNAs, and lncRNAs in ITP compared to normal samples using two public datasets, GSE43177 and GSE43178. Then, bioinformatics analysis was employed to predict the potential functions of differently expressed mRNAs, miR-

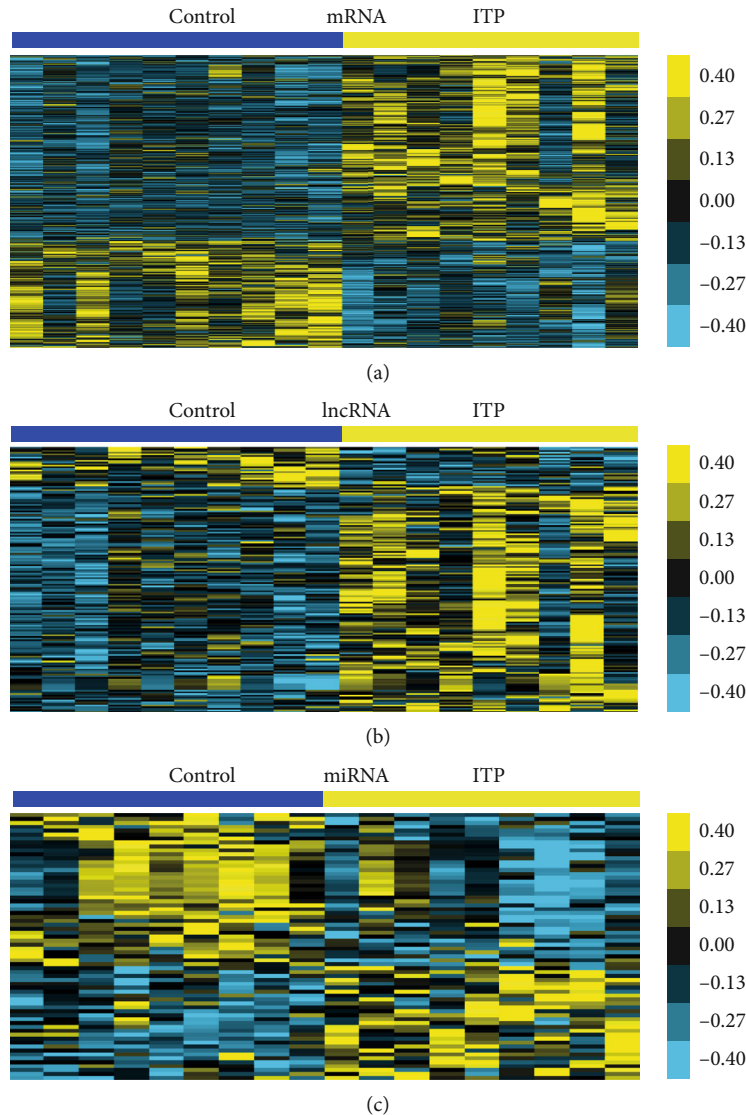


FIGURE 1: Heat map of differentially expressed mRNAs, lncRNAs, and miRNAs in immune thrombocytopenia. Heat map depicts different expression of (a) mRNAs, (b) lncRNAs, and (c) miRNAs in immune thrombocytopenia. Shades of yellow and deongaree represent log₂ gene expression values.

NA, and lncRNAs in ITP. This study could provide useful information for exploring therapeutic candidate targets and new molecular biomarkers for ITP.

2. Material and Methods

2.1. Microarray Data and Data Preprocessing. Gene expression datasets were obtained from the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo>) with accession numbers GSE43177 [10] and GSE43178 [10]. The 10 normal and 9 ITP samples were included in the GSE43177 dataset. Meanwhile, the 9 normal and 9 ITP samples were included in GSE43178 dataset.

2.2. lncRNA Classification Pipeline. In order to evaluate the expression of lncRNAs in microarray data, a pipeline was employed to identify the probe sets uniquely mapped to lncRNAs from the Affymetrix array. A total of 2448 anno-

tated lncRNA transcripts with corresponding Affymetrix probe IDs were obtained. The cutoff values used for selecting differentially expressed lncRNAs were fold change ≥ 2 and $P < 0.05$.

2.3. Prediction of the Targets of miRNAs. To obtain valuable insights into the potential mechanisms of miRNAs, a bioinformatics analysis was performed to identify the target genes of miRNAs using starBase. starBase is a database that combines data from six prediction programs: TargetScan, PicTar (<http://www.pictar.org/>), miRanda (<http://www.microrna.org/microrna/home.do>), PITA (<http://www.genie.weizmann.ac.il/index.html>), RNA22 (<http://www.cm.jefferson.edu/rna22/>), and CLIP-Seq (<http://www.starbase.sysu.edu.cn/>).

2.4. Functional Group Analysis. GO analysis and KEGG analysis were employed to determine the biological functions of

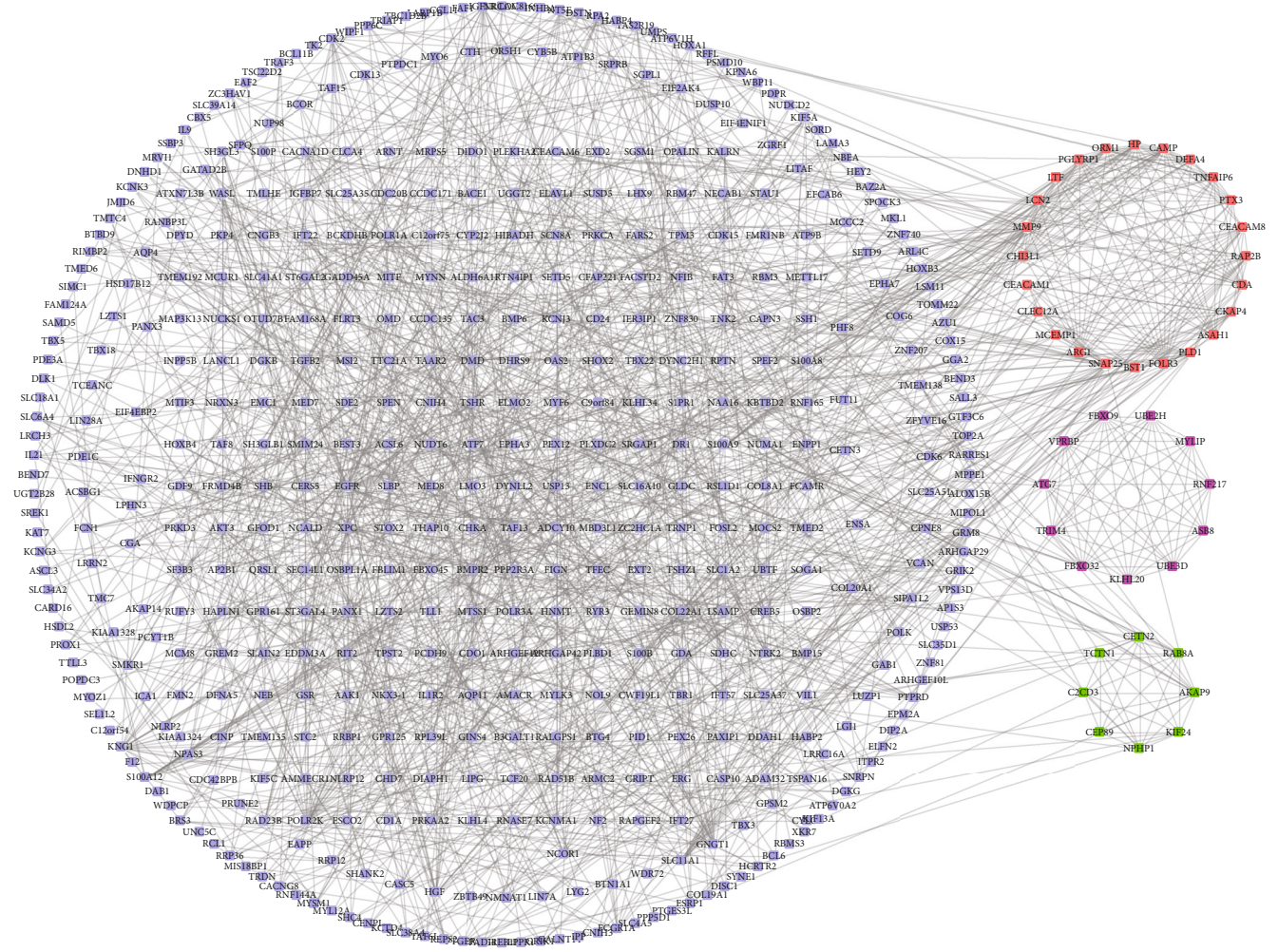


FIGURE 2: PPI network of differently expressed mRNAs in ITP. The PPI network consists of 404 mRNAs. The red subnetwork included 24 nodes and 132 edges. The green subnetwork included 11 nodes and 55 edges. And the purple subnetwork included 8 nodes and 28 edges.

the identified differentially expressed mRNAs, based on the freely available online MAS 3.0 system from CapitalBio Corporation (<http://bioinfo.capitalbio.com/mas3/>; Beijing, China). The P value (hypergeometric P value) denotes the significance of the pathway associated with the conditions. $P < 0.05$ was considered to indicate a statistically significant difference.

2.5. Protein-Protein Interaction Network Mapping. We followed the methods of Chen et al. [11]. The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [12] online software (<https://string-db.org>) was utilized to assess the potential interactions. The interactions of the proteins encoded by the differently expressed genes were searched using STRING online software, and the combined score of >0.4 was used as the cutoff criterion. Cytoscape software (<http://www.cytoscape.org>) was used for the visualization of the PPI network.

2.6. Construction of the Coexpression Network between Differentially Expressed mRNAs and lncRNAs. The Pearson correlation coefficient of DEG-lncRNA pairs was calculated

according to their expression values. The coexpressed DEG-lncRNA pairs with an absolute value of the Pearson correlation coefficient of ≥ 0.8 were selected, and the coexpression network was visualized by using Cytoscape software.

3. Result

3.1. Identification of Differently Expressed mRNAs, lncRNAs, and miRNAs in Immune Thrombocytopenia. First, we analyzed a public dataset GSE43177 to identify differently expressed mRNAs in ITP samples compared to healthy control samples. Subsequently, differential expression analysis was conducted by using GEO2R ($|\log_2FC| > 1$ and adj. P value < 0.05). A total 740 genes were identified as DEGs in ITP, including 468 upregulated genes and 272 downregulated genes. These upregulated and downregulated significant DEGs were present using hierarchical clustering (Figure 1(a)).

By reannotating the gene probes in GSE43177, we found that 1561 lncRNA probes were included in this dataset. Among them, 157 lncRNAs were found to be dysregulated in ITP. 134 lncRNAs were overexpressed and 23 lncRNAs

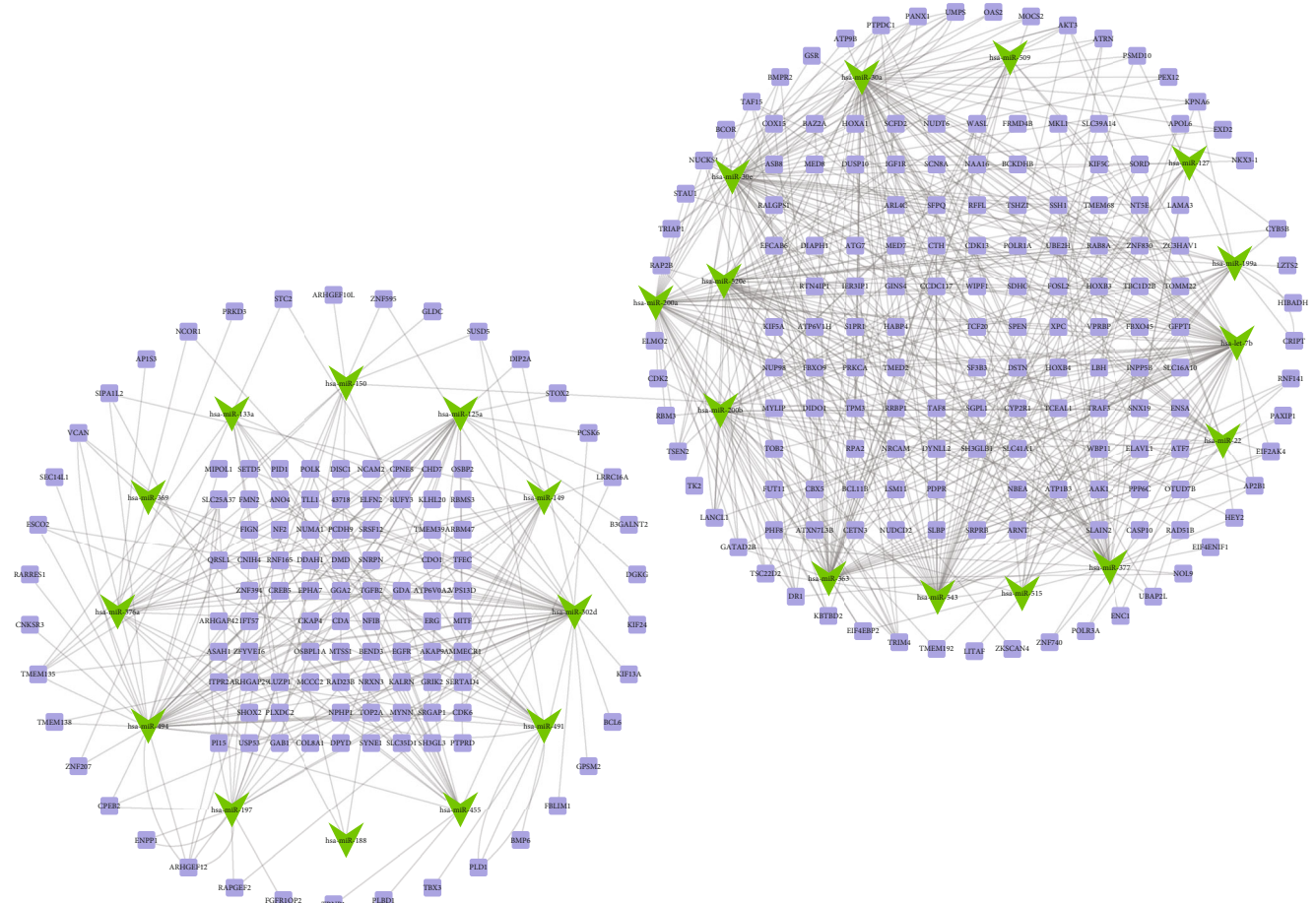


FIGURE 3: PPI network of differently expressed mRNA-target miRNA in ITP. The PPI network consists of 26 miRNAs and correlated 279 target mRNAs. The lilac dot represents mRNA; the green dot represents miRNA.

were downregulated in ITP samples compared to healthy control samples (Figure 1(b)).

Then, we analyzed a public dataset GSE43178 to identify differently expressed miRNAs in ITP. 68 miRNAs were observed to be differentially expressed, including 29 upregulated miRNAs and 39 downregulated miRNAs. The heat map of DEGs in the ITP and control stromal cells is shown in Figure 1(c).

3.2. Construction of the PPI Network Mediated by DEGs in ITP. Subsequently, the PPI network analyses were conducted to reveal the relationships among DEGs. As shown in Figure 2, a total of 404 nodes and 1391 interactions were identified in this PPI network. Interestingly, three sub-PPI networks (red network, green network, and purple network) were identified. The red network included 24 nodes and 132 edges. The green network included 11 nodes and 55 edges. And the purple network included 8 nodes and 28 edges. Seven DEGs played a more important regulatory role in this network by connecting with more than 10 different DEGs, including MMP9, LCN2, DYNLL2, CKAP4, FOLR3, FBXO32, and PLD1.

3.3. Construction of miRNA-DEG Networks in ITP. Furthermore, we used TargetScan and starBase [13] to predict the

downstream targets of differently expressed miRNAs in ITP. Then, a miRNA-DEG network was constructed using Cytoscape software (Figure 3). A total of 26 miRNAs and 279 mRNAs were included in this network. Interestingly, we found that hsa-miR-30a, hsa-let-7b, hsa-miR-30e, hsa-miR-200a, hsa-miR-520e, hsa-miR-494, hsa-miR-543, hsa-miR-302d, hsa-miR-377, hsa-miR-363, and hsa-miR-200b played crucial roles in ITP.

3.4. Construction of lncRNA-mRNA Coexpression Networks in ITP. In order to reveal the potential functions of lncRNAs in ITP, we first conducted lncRNA coexpression analysis based on their expression levels in ITP samples. Then, the lncRNA-mRNA pairs with the value of the absolute Pearson correlation coefficient ≥ 0.75 were selected for network construction. The lncRNA coexpression networks in ITP were constructed using Cytoscape 3.0 [14] (<http://www.cytoscape.org/>).

As presented in Figure 4, 136 lncRNAs, 430 mRNAs, and 1415 edges were contained in this coexpression network. Based on the coexpression network analysis, 8 lncRNAs (LOC101927237, LINC00515, LOC101927066, LOC440028, RP11-161D15.1, LOC101929312, AX747630, and LOC100506406) were identified as key regulators in

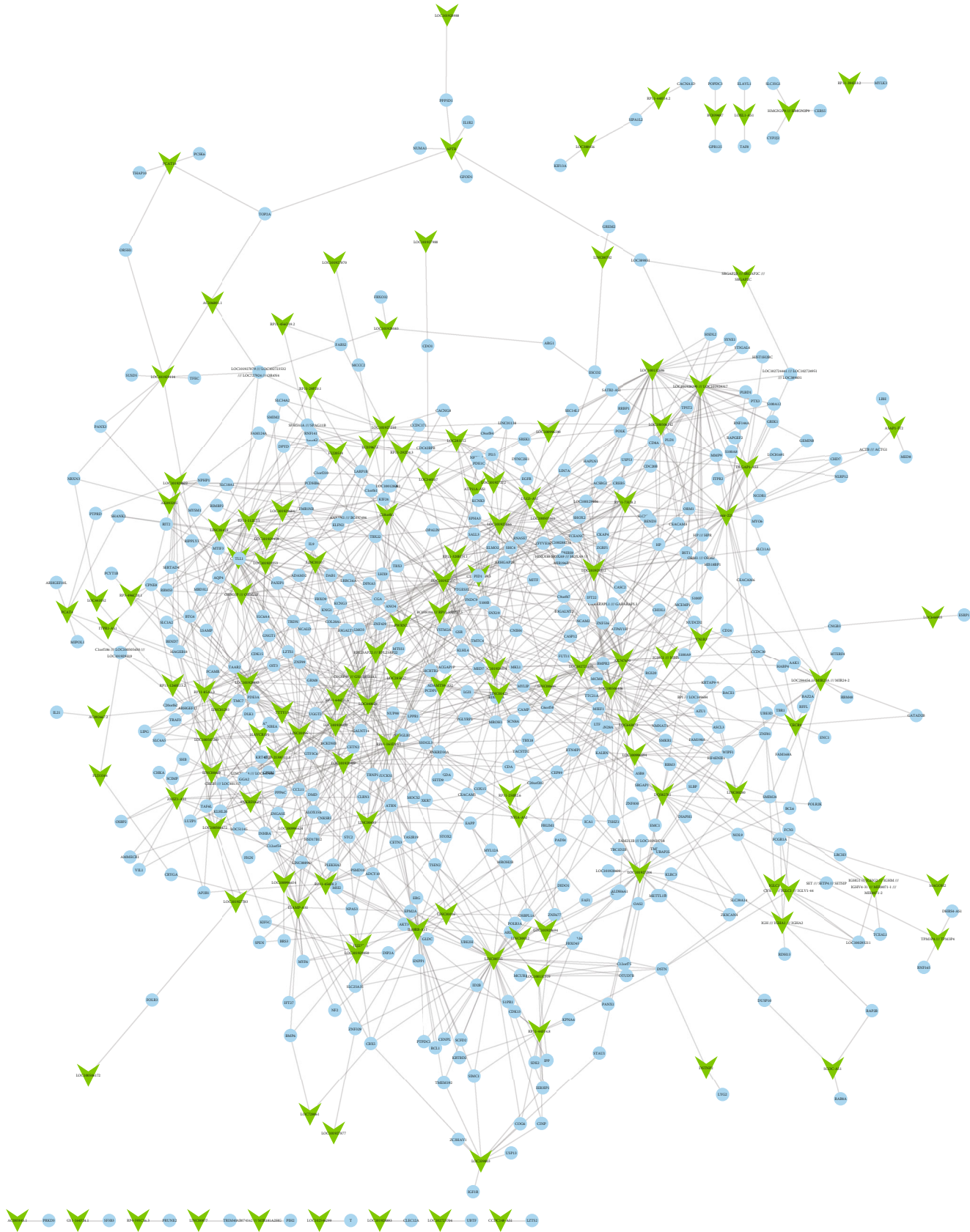


FIGURE 4: Coexpression networks of lncRNAs in ITP. The coexpression network consists of 136 lncRNAs and correlated 430 mRNAs. The blue dot represents mRNA; the green dot represents lncRNA.



FIGURE 5: GO analysis and KEGG analysis of mRNAs, miRNAs, and lncRNAs in ITP. (a) Biological process analysis of the related mRNAs. (b) KEGG pathway analysis of the related mRNAs. (c) Biological process analysis of the related miRNAs. (d) KEGG pathway analysis of the related miRNAs. (e) Biological process analysis of the related lncRNAs. (f) KEGG pathway analysis of the related lncRNAs.

ITP and regulated more than 55 dysregulated mRNAs in ITP (Figure 3).

3.5. Bioinformatics Analysis of mRNAs, miRNAs, and lncRNAs in ITP. In Figure 5, bioinformatics analysis showed DEGs in ITP were associated with the mRNA nonsense-mediated decay, translation, cell-cell adhesion, proteasome-mediated ubiquitin, and mRNA splicing, via spliceosome,

protein polyubiquitination, viral process, autophagy, rRNA processing, and macroautophagy. ITP-related miRNAs were involved in regulating the cytoskeleton-dependent intracellular transport, negative regulation of epithelial cell proliferation, protein localization, proteasome, nuclear DNA replication, nucleotide excision repair, branched-chain amino acid catabolic process, regulation of mitophagy, cellular response to cAMP, and negative regulation of transcription.

ITP-related lncRNAs were involved in regulating the positive regulation of inflammatory response, cellular response to cGMP, ephrin receptor signaling pathway, chronic inflammatory response, forelimb morphogenesis, stem cell population maintenance, cell junction assembly, positive regulation of cell growth, chemical synaptic transmission, and inflammatory response.

Bioinformatics analysis showed DEGs in ITP were associated with the oxytocin signaling pathway, glutamatergic synapse, choline metabolism in cancer, dopaminergic synapse, FoxO signaling pathway, hypertrophic cardiomyopathy (HCM), ovarian steroidogenesis, thyroid hormone synthesis, serotonergic synapse, and metabolic pathways. ITP-related miRNAs were associated with endocytosis, pyrimidine metabolism, prostate cancer, drug metabolism—other enzymes, FoxO signaling pathway, glioma, choline metabolism in cancer, thyroid hormone synthesis, hepatitis B, and metabolic pathways. ITP-related lncRNAs were associated with glutamatergic synapse, endocytosis, serotonergic synapse, dopaminergic synapse, arrhythmogenic right ventricular cardiomyopathy, platelet activation, estrogen signaling pathway, thyroid hormone synthesis, FoxO signaling pathway, and focal adhesion.

4. Discussion

ITP is an autoimmune disorder. The increasing antiplatelet antibodies produced by B cells, and the aberrant functions of T lymphocytes were involved in regulating the ITP. Previous studies revealed that the dysregulation of multiple genes, such as miRNAs and lncRNAs, contributed to the progression of ITP. For example, MIR130A, MIR409-3p, and lncRNA TMEVPG1 were downregulated in ITP. Moreover, Qian et al. identified a total of 1809 lncRNAs were significantly dysregulated in ITP patients compared to normal samples. Better understanding of the regulation of ITP is very crucial for the discovery of therapeutic targets for the treatment of this disease.

The present study screened differently expressed mRNAs, lncRNAs, and miRNAs in ITP. A total 740 genes were identified as DEGs in ITP, including 468 upregulated genes and 272 downregulated genes. Subsequently, a PPI network, including 404 nodes and 1391 interaction, was constructed to identify hub regulators in ITP. Seven DEGs played a more important regulatory role in this network by connecting with more than 10 different DEGs, including MMP9, LCN2, DYNLL2, CKAP4, FOLR3, FBXO32, and PLD1. This is the first time their regulatory roles in ITP were revealed. Notably, PLD1 had been demonstrated to play an important role in autoimmune diseases. PLD1 mediated lymphocyte adhesion and migration in autoimmune encephalomyelitis [15]. PLD1 regulated the expression of proinflammatory genes in rheumatoid arthritis synovial fibroblasts [16]. Bioinformatics analysis showed DEGs in ITP were associated with the mRNA nonsense-mediated decay, translation, cell-cell adhesion, proteasome-mediated ubiquitin, and mRNA splicing, via spliceosome, protein polyubiquitination, viral process, autophagy, rRNA processing, and macroautophagy.

Increasing evidence indicated that miRNAs and lncRNAs are essential in regulating gene expression, cell proliferation, apoptosis, and migration. However, the detail functions and special expression pattern of miRNAs and lncRNAs in ITP remained largely unclear. Meanwhile, we identified 134 upregulated lncRNAs, 23 downregulated lncRNAs, 29 upregulated miRNAs, and 39 downregulated miRNAs in ITP patients. Furthermore, we constructed the miRNA-DEG network and lncRNA coexpression network to explore their functions in ITP. Interestingly, 8 lncRNAs (LOC101927237, LINC00515, LOC101927066, LOC440028, RP11-161D15.1, LOC101929312, AX747630, and LOC100506406) were identified as key regulators in ITP. Among them, LINC00515 was reported to promote multiple myeloma autophagy and chemoresistance though the miR-140-5p/ATG14 axis [17]. However, the functions of most lncRNAs were unknown in human diseases. Bioinformatics analysis showed ITP-related lncRNAs were involved in regulating the positive regulation of inflammatory response, cellular response to cGMP, ephrin receptor signaling pathway, chronic inflammatory response, forelimb morphogenesis, stem cell population maintenance, cell junction assembly, positive regulation of cell growth, chemical synaptic transmission, and inflammatory response.

Several limitations should be noted in this study. First, this study was mainly based on bioinformatics analysis. Therefore, the functional validation should be conducted in the near future. Second, the sample size in this study was small. We should collect more clinical samples to detect the expression of the key mRNAs, miRNAs, and lncRNAs in the progression of ITP.

In conclusion, our integrative analysis identified key mRNAs, miRNAs, and lncRNAs in the progression of ITP. Bioinformatics analysis showed these genes regulated multiple biological processes in ITP, such as mRNA nonsense-mediated decay, translation, cell-cell adhesion, proteasome-mediated ubiquitin, and mRNA splicing. We thought the present study could broaden our insights into the mechanism underlying the progression of ITP and provide a potential biomarker for the prognosis of ITP.

Abbreviations

ITP:	Immune thrombocytopenia
lncRNAs:	Long noncoding RNAs
STRING:	Search Tool for the Retrieval of Interacting Genes/Proteins
GEO:	Gene Expression Omnibus.

Data Availability

Gene expression datasets were obtained from the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo>) with accession numbers GSE43177 and GSE43178.

Conflicts of Interest

The authors declare no financial conflicts of interest.

Authors' Contributions

The contributions of the authors involved in this study are as follows: guarantor of integrity of the entire study: Zhenwei Fan; study concepts: Xuan Wang; study design: Peng Li; definition of intellectual content: Peng Li; literature research: Min Zhang; clinical studies: Peng Li and Xuan Wang; experimental studies: Chunshan Zhao; data acquisition: Yan Song; data analysis: Chunshan Zhao; statistical analysis: Chunshan Zhao; manuscript preparation: Chunli Mei; manuscript editing: all authors; and manuscript review: all authors.

Acknowledgments

This work was funded by the Jilin Provincial Health Department in China (No. 2017ZC029) and Science and Technology Bureau Project of Jilin City (No. 201830557).

References

- [1] S. J. Barsam, B. Psaila, M. Forestier et al., "Platelet production and platelet destruction: assessing mechanisms of treatment effect in immune thrombocytopenia," *Blood*, vol. 117, no. 21, pp. 5723–5732, 2011.
- [2] K. Yazdanbakhsh, H. Zhong, and W. Bao, "Immune dysregulation in immune thrombocytopenia," *Seminars in Hematology*, vol. 50, pp. S63–S67, 2013.
- [3] D. A. Chistiakov, "Immunogenetics of Hashimoto's thyroiditis," *Journal of Autoimmune Diseases*, vol. 2, no. 1, pp. 1–21, 2005.
- [4] M. Esteller, "Non-coding RNAs in human disease," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 861–874, 2011.
- [5] S. C. Warth, K. P. Hoefig, A. Hiekel et al., "Induced miR-99a expression represses Mtor cooperatively with miR-150 to promote regulatory T-cell differentiation," *The EMBO Journal*, vol. 34, no. 9, pp. 1195–1213, 2015.
- [6] H. Zhao, H. Li, W. du et al., "Reduced *MIR130A* is involved in primary immune thrombocytopenia via targeting *TGFB1* and *IL18*," *British Journal of Haematology*, vol. 166, no. 5, pp. 767–773, 2014.
- [7] M. Chang, P. A. Nakagawa, S. A. Williams et al., "Immune thrombocytopenic purpura (ITP) plasma and purified ITP monoclonal autoantibodies inhibit megakaryocytopoiesis in vitro," *Blood*, vol. 102, no. 3, pp. 887–895, 2003.
- [8] J. Wang, H. Peng, J. Tian et al., "Upregulation of long noncoding RNA TMEVPG1 enhances T helper type 1 cell response in patients with Sjögren syndrome," *Immunologic Research*, vol. 64, no. 2, pp. 489–496, 2016.
- [9] W. J. Liu, J. Bai, Q. L. Guo, Z. Huang, H. Yang, and Y. Q. Bai, "Role of platelet function and platelet membrane glycoproteins in children with primary immune thrombocytopenia," *Molecular Medicine Reports*, vol. 14, no. 3, pp. 2052–2060, 2016.
- [10] M. Jernås, I. Nookaew, H. Wadenvik, and B. Olsson, "MicroRNA regulate immunological pathways in T-cells in immune thrombocytopenia (ITP)," *Blood*, vol. 121, no. 11, pp. 2095–2098, 2013.
- [11] L. Chen, Y. Zhang, Z. Rao, J. Zhang, and Y. Sun, "Integrated analysis of key mRNAs and lncRNAs in osteoarthritis," *Experimental and Therapeutic Medicine*, vol. 16, no. 3, pp. 1841–1849, 2018.
- [12] D. Szklarczyk, A. L. Gable, D. Lyon et al., "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, 2019.
- [13] J. H. Yang, J. H. Li, P. Shao, H. Zhou, Y. Q. Chen, and L. H. Qu, "starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data," *Nucleic Acids Research*, vol. 39, Supplement_1, pp. D202–D209, 2011.
- [14] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [15] M. A. Frohman, "The phospholipase D superfamily as therapeutic targets," *Trends in Pharmacological Sciences*, vol. 36, no. 3, pp. 137–144, 2015.
- [16] U. Müller-Ladner, J. Kriegsmann, B. N. Franklin et al., "Synovial fibroblasts of patients with rheumatoid arthritis attach to and invade normal human cartilage when engrafted into SCID mice," *American Journal of Pathology*, vol. 149, no. 5, pp. 1607–1615, 1996.
- [17] Y. Meng, R. Gao, J. Ma et al., "MicroRNA-140-5p regulates osteosarcoma chemoresistance by targeting HMG5 and autophagy," *Scientific Reports*, vol. 7, no. 1, p. 416, 2017.

Retraction

Retracted: miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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. Xu, Y. Ding, J. Yao et al., "miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 5807836, 10 pages, 2020.

Research Article

miR-215 Inhibits Colorectal Cancer Cell Migration and Invasion via Targeting Stearoyl-CoA Desaturase

Xinhua Xu,¹ Yan Ding,² Jun Yao,³ Zhiping Wei,³ Haipeng Jin,³ Chen Chen,³ Jun Feng^{ID},³ and Rongbiao Ying^{ID}³

¹Department of Pathology, Taizhou Cancer Hospital, Zhejiang Province, China

²Department of Radiotherapy Oncology, Taizhou Central Hospital, Zhejiang Province, China

³Department of Surgical Oncology, Taizhou Cancer Hospital, Zhejiang Province, China

Correspondence should be addressed to Jun Feng; charlifeng1980@163.com and Rongbiao Ying; roc619512@163.com

Xinhua Xu and Yan Ding contributed equally to this work.

Received 10 April 2020; Revised 22 May 2020; Accepted 26 May 2020; Published 30 June 2020

Guest Editor: Tao Huang

Copyright © 2020 Xinhua Xu 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.

Background. This study was aimed at exploring the effects of miR-215 and its target gene stearoyl-CoA desaturase (SCD) on colorectal cancer (CRC) cell migration and invasion. **Methods.** Here, we analyzed the relationship between miR-215 and SCD, as well as the regulation of miR-215 on CRC cells. We constructed wild-type and mutant plasmids of SCD to identify whether SCD was a target gene of miR-215 by using a luciferase reporter assay. The expression of miR-215 and SCD was detected by quantitative real-time polymerase chain reaction (qRT-PCR) and western blot, respectively. MTT, wound healing, and Transwell assays were applied to determine the effect of miR-215 on CRC cell proliferation, migration, and invasion. **Results.** It was found that miR-215 expression was significantly decreased in CRC tissue while SCD was highly expressed compared with those in adjacent normal tissue. The luciferase reporter assay indicated that SCD was a direct target gene of miR-215. Functional analysis revealed that miR-215 overexpression significantly inhibited CRC cell proliferation, migration, and invasion *in vitro*. In addition, the result of rescue experiments showed that overexpression of SCD could promote the proliferation, migration, and invasion of CRC cells, and the carcinogenic effect of SCD could be inhibited by miR-215. **Conclusions.** Taken together, our findings suggested that miR-215 could inhibit CRC cell migration and invasion via targeting SCD. The result could eventually contribute to the treatment for CRC.

1. Background

Colorectal cancer (CRC) is one of the common malignant cancers of the digestive tract, including the rectum and colon, which mainly occurs in the rectum and the junction between the rectum and sigmoid colon. CRC is also the third most frequently diagnosed cancer and the fourth leading cause of cancer-related death globally [1, 2]. Although cancer treatment strategies are increasingly developed and the treatment effect of CRC patients in early stages has been significantly improved during the past several decades, most patients have been already diagnosed in advanced stages. At present, surgical resection is the most effective method of treating CRC, but 25%-40% of patients still experience recurrence or metastasis. Therefore, it has become a hot topic to explore the invasion and migration

of CRC cells at the molecular level, which helps to find new effective treatment options and improve patients' survival rate.

Stearoyl-CoA desaturase (SCD) is a key enzyme for the formation of monounsaturated fatty acids from saturated fatty acids, and its main components include palmitoleic acid (C16:1) and oleic acid (C18:1) [3]. In recent years, SCD has been confirmed to play an important regulatory role in the occurrence and development of a variety of cancers. For example, decreased SCD expression can inhibit breast cancer progression through the β -catenin signaling pathway [4]. SCD can significantly promote the growth of lung cancer by activating EGFR/PI3K/AKT signaling in tumor tissue [5]. However, the regulatory mechanism of SCD in CRC remains unclear.

MicroRNAs (miRNAs) are a class of small noncoding RNAs that can regulate gene expression by facilitating

mRNA degradation or inducing translational repression [6]. The miRNA miR-215 has been proven to play an important role in tumorigenesis and tumor progression in many types of human cancers, such as epithelial ovarian cancer (EOC) [7], endometrial cancer [8], breast cancer [9], and non-small-cell lung cancer [10]. Recent studies show that overexpression of miR-215 markedly downregulates LEFTY2 protein expression level in Hec-1A cells and endometrial cancer tissue [11]. It is also reported that overexpression of LEFTY2 protein promotes epithelial-mesenchymal transition (EMT) and sensitizes Hec-1A cells to cisplatin treatment [11]. In addition, overexpression of miR-215 suppresses EOC growth and invasion by targeting NOB1 [12]. However, the relationship between miR-215 and SCD has not been reported yet.

In this study, we tested miR-215 expression in CRC cells and investigated the biological effect of miR-215 on migration and invasion of CRC cells. Here, we found that miR-215 exerted a suppressive effect on tumor migration and invasion by targeting SCD.

2. Methods

2.1. Microarray Analysis. We searched the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>) with “Colorectal cancer” as the key word, and two miRNA microarrays GSE110224 and GSE35834 were selected. GSE110224 included 17 normal samples and 17 tumor samples with histologically confirmed CRC. Total mRNAs were extracted for further processing with the Human Genome U133 Plus 2.0 Array (Affymetrix Inc., Santa Clara, CA, USA). GSE35834 contained 78 samples comprising 23 normal adjacent mucosa tissue samples and 55 CRC tumor samples, and GPL8786 Multispecies miRNA-1 Array (Affymetrix Inc., Santa Clara, CA, USA) was used as the sequencing platform. The datasets were analyzed with the R package “Limma.” $|\log_{2}FC| > 1$ and p value < 0.05 were set as the threshold for screening the differentially expressed genes (DEGs).

2.2. Analysis of the miRNAs That Regulate SCD. The miRNAs that regulate SCD were retrieved in the starBase V2.0 (<http://starbase.sysu.edu.cn/>), TargetScan (<http://www.targetscan.org/>), and miRTarBase (<http://mirtarbase.mbc.nctu.edu.tw/>) databases. A Venn diagram (<http://bioinformatics.psb.ugent.be/webtools/Venn/>) was used to find the intersections of the predicted results in the three databases.

2.3. Human Tissue Specimens. Paraffin-embedded pathological specimens from 30 CRC tumor and paired adjacent normal tissue samples were included in this study. Samples were obtained from Taizhou Cancer Hospital, Zhejiang Province, from July 2016 to June 2017. All the patients were diagnosed by pathological examination and had never received chemotherapy or radiotherapy before surgery. All the samples were collected with patients’ informed consent after approval from the Institute Research Medical Ethics Committee of Taizhou Cancer Hospital.

2.4. Cell Lines and Transfection. The CRC cell line HT29 was obtained from the Bena Culture Collection (Beijing, China) and was grown in Dulbecco’s Modified Eagle’s Medium (DMEM, Gibco) with 100 U/mL penicillin, 0.1 mg/mL streptomycin, and 10% fetal bovine serum (FBS). All cells were maintained in a humidified incubator with 5% CO₂ at 37°C until they were grown to a logarithmic phase. Cells (2×10^5 cells/well) were seeded in a six-well plate and then subjected to transfection by employing Lipofectamine 2000 (Invitrogen, Karlsruhe, Germany).

NC (transfected with negative sequence), miR-215 mimic, and miR-215 inhibitor were purchased from GeneCopoeia (Guangzhou, China). SCD overexpression (oe-SCD) and corresponding negative control (oe-NC) were constructed by lentiviral vectors.

2.5. Dual-Luciferase Reporter Gene Assay. Target sequences of wild-type (WT) and mutant (WUT) SCD 3’UTR were constructed artificially and ligated into the pmirGLO (Promega, Madison, USA) reporter plasmids with enzymes BamHI and XhoIII to obtain WT and MUT reporter plasmids. Afterwards, the two reporter plasmids were cotransfected with the miR-215 mimic or NC into the cancer cell line using Lipofectamine 2000. Relative luciferase activities were determined by the Dual-Luciferase® Reporter Assay System (Promega) following the instructions 48 h after transfection.

2.6. qRT-PCR. Total RNA was extracted from CRC cells, tumor tissue, and paired adjacent normal tissue using Trizol Reagent (Ambion, USA) according to the manufacturer’s instructions. The concentration and purity of RNA were determined with an ultraviolet spectrophotometer. RNA was reversely transcribed into cDNA by using RT-PCR Kit (ABI Company, 243 Forest City, CA, USA), and quantitative real-time- (qRT-) PCR was performed according to the manufacturer’s instructions of SYBR Premix Ex Taq II (TaKaRa). The relative expression level of RNA was calculated by the $2^{-\Delta\Delta CT}$ method with U6 and GAPDH as the internal reference for miR-215 and SCD, respectively. miR-215 stem-loop primers were as follows: 5’-CTCAACTGGTGTCTGGAGTCGGCAATTCAGTTGAGCGTCTGT-3’. The sequences of the PCR primers were as follows: miR-215 forward 5’-CTCAACTGGTGTCTGGAGTCGG-3’ and reverse 5’-ACAGGA AAATGACCTATGAATTGAC-3’, U6 forward 5’-GTAC AAAATACGTGACGTAGAAAG-3 and reverse 5’-GGTGTTCGTCCTTTCCAC-3’, SCD forward 5’-TCTAGCTCC TATACCACCACCA-3’ and reverse 5’-TCGTCTCCAAC TTATCTCCTCC-3’, and GAPDH forward 5’-GGAGCG AGATCCCTCCAAAAT-3’ and reverse 5’-GGCTGTGT CATACTTCTCATGG-3’.

2.7. Western Blot Analysis. Cells in the logarithmic growth phase were collected and lysed with the RIPA buffer containing protease and phosphatase inhibitors. Extracted proteins were loaded onto 8% SDS-PAGE at a voltage of 150 V after quantification. Thereafter, the proteins were transferred onto the PVDF membrane and incubated with the primary mouse

antibodies against SCD (1:500, Abcam) and GADPH (1:1000, Abcam), respectively, overnight at 4°C. After three times being washed with TBST, the membrane was probed with the secondary antibody rabbit anti-mouse IgG (HRP) (1:2000, Abcam) for 1 h at room temperature. Finally, the protein bands were detected using the ABI 7500 Real-Time PCR System (Applied Biosystems; Thermo Fisher Scientific, NY, USA).

2.8. MTT. The MTT assay was conducted to evaluate cell proliferation capacity. Cells were trypsinized after 24 h of culture in serum-free medium. The cells were counted with a hemocytometer, and the cell density was adjusted to 1×10^5 cells/mL. Then, the cells were seeded in 96-well plates at a density of 2×10^3 cells/well with the volume of 200 μ L. 5 mg/mL of MTT solution was added to each well at 24 h, 48 h, and 72 h, respectively. After incubation for 4 h, the reaction was stopped. The supernatant was discarded after centrifugation, and 100 μ L of DMSO was added to each well to promote crystal dissolution. The absorbance was measured at 490 nm. The assay was performed in triplicate.

2.9. Wound Healing Assay. CRC cells in the logarithmic phase were inoculated into a 6-well plate (2×10^5 cells/well) with marks on the back of the plate. After 24 h of culture, the cells covered the entire plate, and scratches were created perpendicular to the marks using a 10 μ L pipette. The detached cells were washed away with PBS, followed by the addition of serum-free DMEM. Cells were incubated in an incubator with 5% CO₂ at 37°C, and images at 0 h and 48 h were captured under an inverted microscope. Three parallel wells were prepared for each group.

2.10. Transwell Invasion Assay. Transwell inserts (Millipore, Bedford, Mass., USA) covered with Matrigel were used for a cell invasion assay. CRC cells in the logarithmic phase were harvested for digestion with trypsin, then washed once with PBS and resuspended in serum-free DMEM. The cell density was adjusted to 1×10^5 cells/mL. A total of 200 μ L of cell suspension was seeded in the upper chambers, and 500 μ L of DMEM containing 15% FBS was added to the lower chambers. The cells were then cultured in an incubator with 5% CO₂ at 37°C for 48 h. Cells invading the lower chambers were fixed with 95% ethanol for 10 min and stained with 0.1% crystal violet for 10 min, and PBS was used to remove the unstained cells. After observation under an inverted microscope, cells from 5 randomly selected fields were counted and the mean value was determined. This experiment was repeated three times. The number of cells that penetrated the Matrigel showed the invasion ability of cells in each group.

2.11. Statistical Analysis. Statistical analysis was conducted using SPSS 21.0 software and GraphPad Prism 6.0. Experiment data were recorded as the mean \pm standard deviation. Comparisons between two groups were assessed by Student's *t*-test, while comparisons among multiple groups were assessed by one-way ANOVA. Counting data were analyzed by the chi-square test. $P < 0.05$ was considered statistically significant.

3. Results

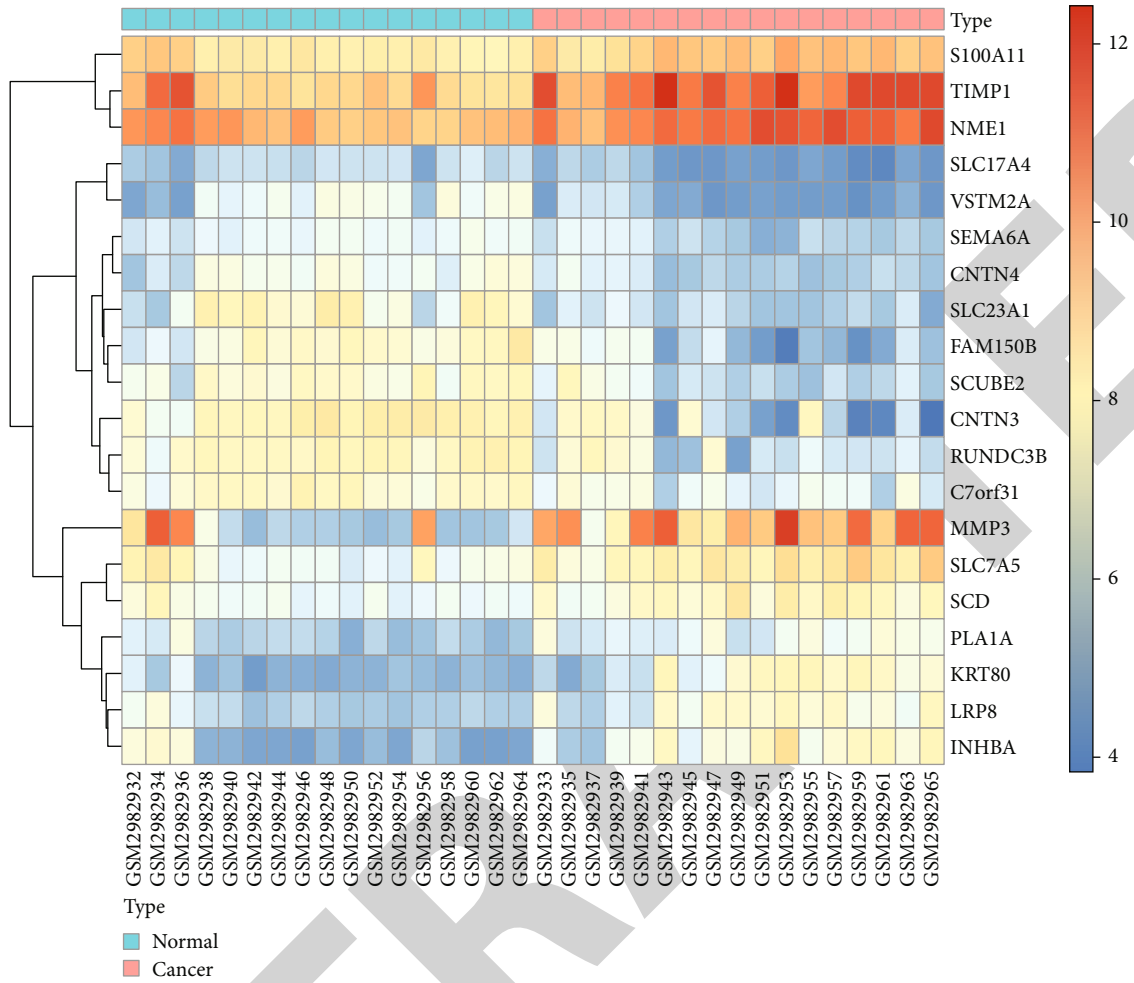
3.1. SCD Is Upregulated in CRC. The transcriptome expression data of CRC were analyzed by the bioinformatics method. The results showed that SCD was significantly upregulated in CRC samples compared with normal samples (Figures 1(a) and 1(b)). At the same time, the qRT-PCR result showed that the expression level of SCD mRNA in CRC tissue was significantly higher than that in adjacent tissue (Figure 1(c)).

3.2. SCD Is a Direct Target Gene of miR-215. In order to explore the underlying mechanism of SCD in CRC cells, we firstly analyzed the GSE35834 dataset and obtained 20 miRNAs with significant differential expression in CRC (Figure 2(a)). Then, the starBase V2.0, TargetScan, and miRTarBase databases were used to predict the potential upstream miRNAs for SCD (Figure 2(b)). According to bioinformatics databases, there was a binding site of miR-215 and 3'-UTR of SCD (Figure 2(c)). In addition, the expression level of miR-215 in CRC tissue was significantly decreased (Figure 2(d)).

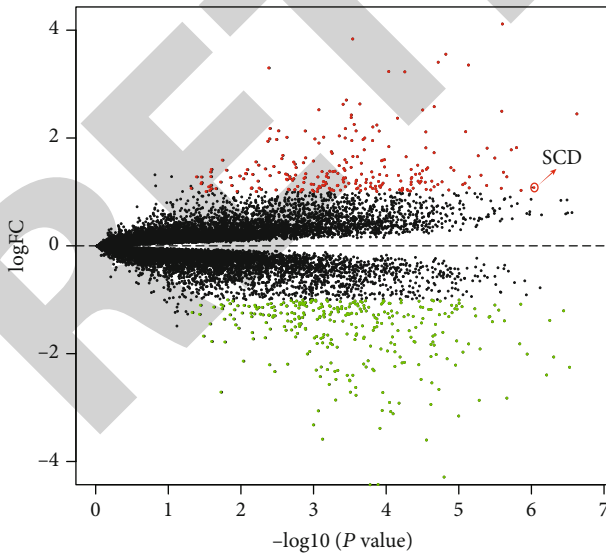
To confirm whether SCD was a direct target gene of miR-215 in CRC cells, we conducted the luciferase reporter assay in HT29 cells 48 h after cotransfection of WT or MUT SCD and miR-215 mimic or NC. The result exhibited that overexpression of miR-215 reduced the luciferase activity of WT SCD in HT29 cells but had no effect on MUT SCD (Figure 2(e)). Additionally, overexpression of miR-215 significantly inhibited the mRNA and protein expression of SCD (Figures 2(f) and 2(g)). Subsequently, we analyzed the correlation between miR-215 and SCD expression in clinical specimens. The result indicated that there was a negative correlation between them (Figure 2(h)). These results indicated that SCD was a direct target gene of miR-215.

3.3. miR-215 Inhibits CRC Cell Proliferation, Migration, and Invasion In Vitro. According to the aberrant expression of miR-215 in CRC cells, we speculated that it might regulate cancer cell proliferation, invasion, and migration. In order to test our hypothesis, we overexpressed and inhibited miR-215 in HT29 cells by transfecting the miR-215 mimic and inhibitor, respectively. qRT-PCR was used to confirm that the miR-215 expression was decreased significantly in the miR-215 inhibitor group and increased in the miR-215 mimic group (Figure 3(a)). After transfection, we used the MTT assay to evaluate the cell proliferation capacity at 24 h, 48 h, and 72 h. The result observed low proliferation activity of HT29 cells after overexpressing miR-215 (Figure 3(b)). Meanwhile, we investigated whether miR-215 affected CRC cell invasion and migration. Wound healing and Transwell invasion assays were performed in HT29 cells transfected with NC, miR-215 mimic, and miR-215 inhibitor. The results demonstrated that overexpression of miR-215 significantly inhibited cell invasion and migration (Figures 3(c) and 3(d)).

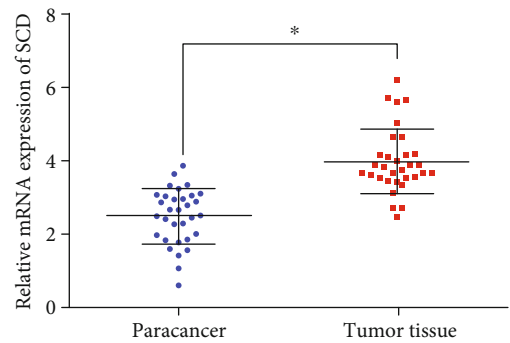
3.4. miR-215 Mediates the Migration and Invasion of CRC Cells via Targeting SCD. In order to verify that miR-215 can regulate the cellular function of CRC cells by inhibiting the



(a)

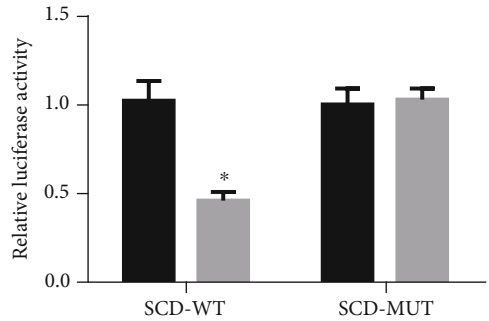
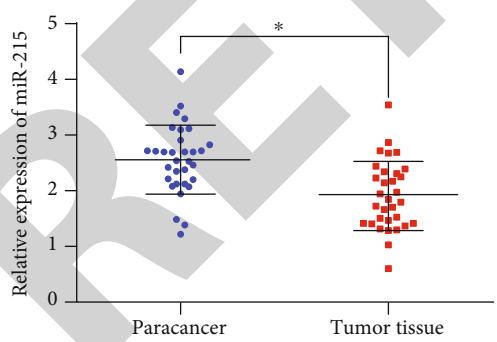
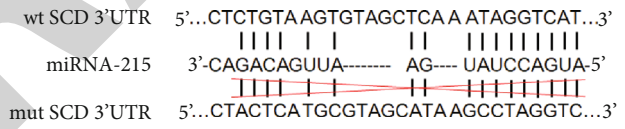
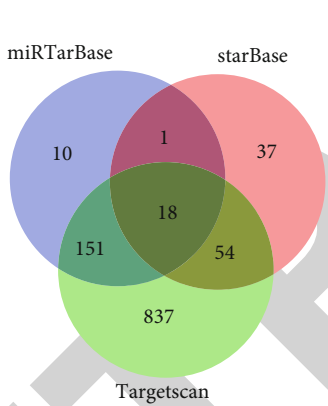
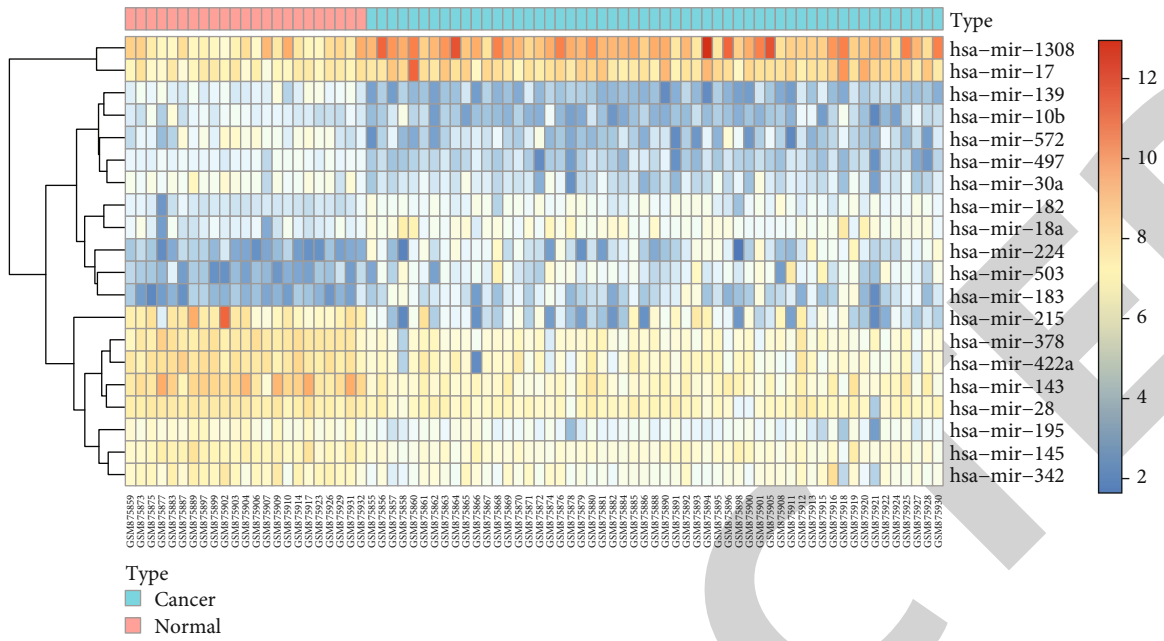


(b)



(c)

FIGURE 1: SCD is upregulated in CRC, and the number unit is expression \log_2 . (a) The top 20 DEGs in GSE110224. (b) The SCD gene is significantly upregulated in CRC samples. (c) qRT-PCR is used to detect the expression of the SCD gene in cancer tissue and paracancerous tissue ($n = 30$, $*P < 0.05$).



(d) (e)

FIGURE 2: Continued.

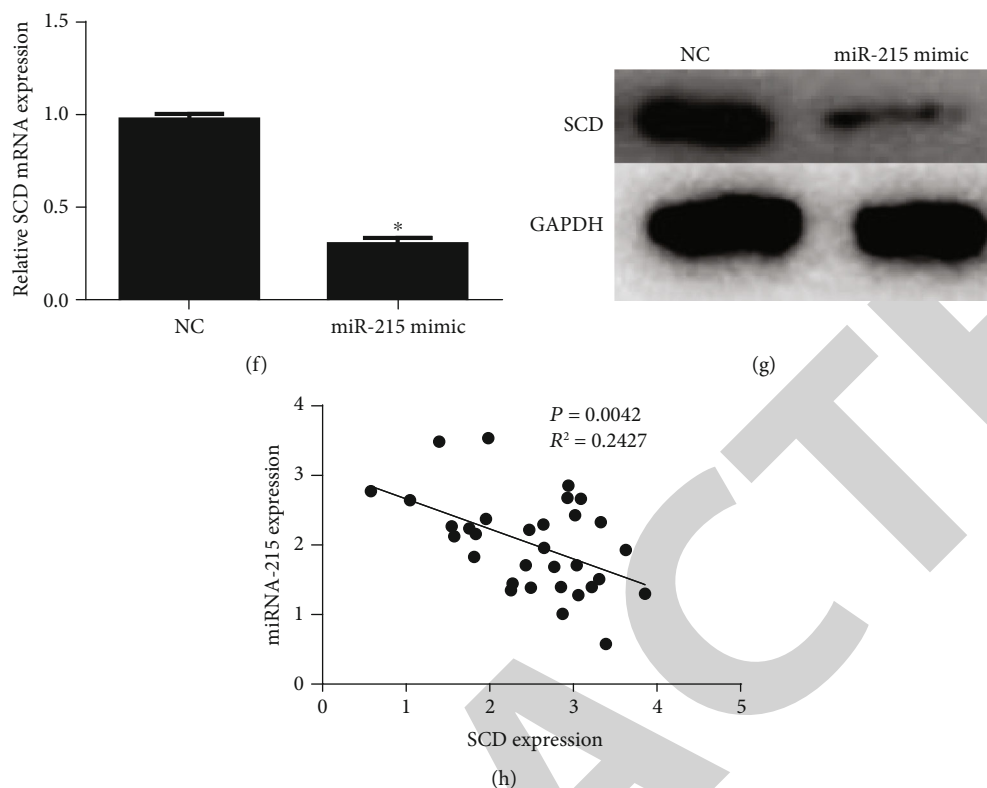


FIGURE 2: SCD is a direct target gene of miR-215. (a) The top 20 DE miRNAs in GSE35834. (b) The Venn diagram of miRNAs that regulate SCD. (c) The targeting sites of miR-215 on SCD 3'UTR and the corresponding mutant sequences. (d) qRT-PCR is used to detect the expression of miR-215 in cancer tissue and paracancerous tissue ($n = 30$, $*P < 0.05$). (e) Relative luciferase activity in HT29 cells after cotransfection with WT or MUT SCD and miR-215 mimic or NC ($*P < 0.05$, $**P < 0.01$, compared to NC). (f) SCD expression on mRNA level in HT29 cells transfected with the miR-215 mimic or NC ($*P < 0.05$, compared to NC). (g) SCD expression on the protein level in HT29 cells transfected with the miR-215 mimic or NC. (h) The correlation between miR-215 and SCD expression in clinical specimens ($*P < 0.05$, compared to NC).

expression of SCD, we conducted the rescue experiments. Firstly, we detected the mRNA and protein levels of SCD in three transfected HT29 cell lines (NC+oe-NC, NC+oe-SCD, and miR-215 mimic+oe-SCD), and the results showed that the elevated expression of SCD was downregulated by miR-215 (Figures 4(a) and 4(b)). Then, we measured the proliferation of cancer cells through the MTT assay and discovered that overexpression of SCD promoted the proliferation of cancer cells, and its promoting effect could be reversed by overexpression of miR-215 (Figure 4(c)). In addition, we used the wound healing assay and Transwell invasion assay to detect the migration and invasion of cells. The results indicated that the overexpression of SCD significantly promoted the migration and invasion of HT29 cells, while the simultaneous overexpression of miR-215 and SCD attenuated such promoting effect (Figures 4(d) and 4(e)). Therefore, we believed that miR-215 could inhibit the proliferation and migration of CRC by downregulating the expression of SCD.

4. Discussion

More than 1.2 million patients are diagnosed with CRC each year. The mortality of CRC is the fourth highest of all the cancer deaths, and the disease tends to be younger in epi-

demology. Therefore, studying the mechanism of occurrence, invasion, and migration of CRC can improve the therapeutic effect of CRC patients. Accumulating evidence shows that miRNAs are closely related to the proliferation, invasion, migration, and recurrence of various tumors, and they can be used as effective molecular markers as well as therapeutic targets for cancer diagnosis, prognosis, and treatment [13–15].

In this study, we found that SCD was highly expressed in CRC by bioinformatics. The literature on SCD has also showed that SCD is upregulated in multiple cancers, such as ovarian cancer [16], breast cancer [17], and liver cancer [18]. Next, in order to verify the result of bioinformatics, we detected the expression of SCD in normal tissue and CRC tissue and confirmed that SCD was upregulated in CRC tissue. It is universally known that miRNAs exert their biological functions by regulating the expression of target genes [19]. Therefore, we used bioinformatics to further explore the miRNAs that could target SCD, and it was found that miR-215 and SCD had binding sites. miR-215 is a widely studied miRNA that has been confirmed to targetedly inhibit various mRNAs, including ARFGEF1 [20], RUNX1 [21], KDM1B [22], and ZEB2 [10]. In order to verify that SCD was a downstream target of miR-215, we detected the

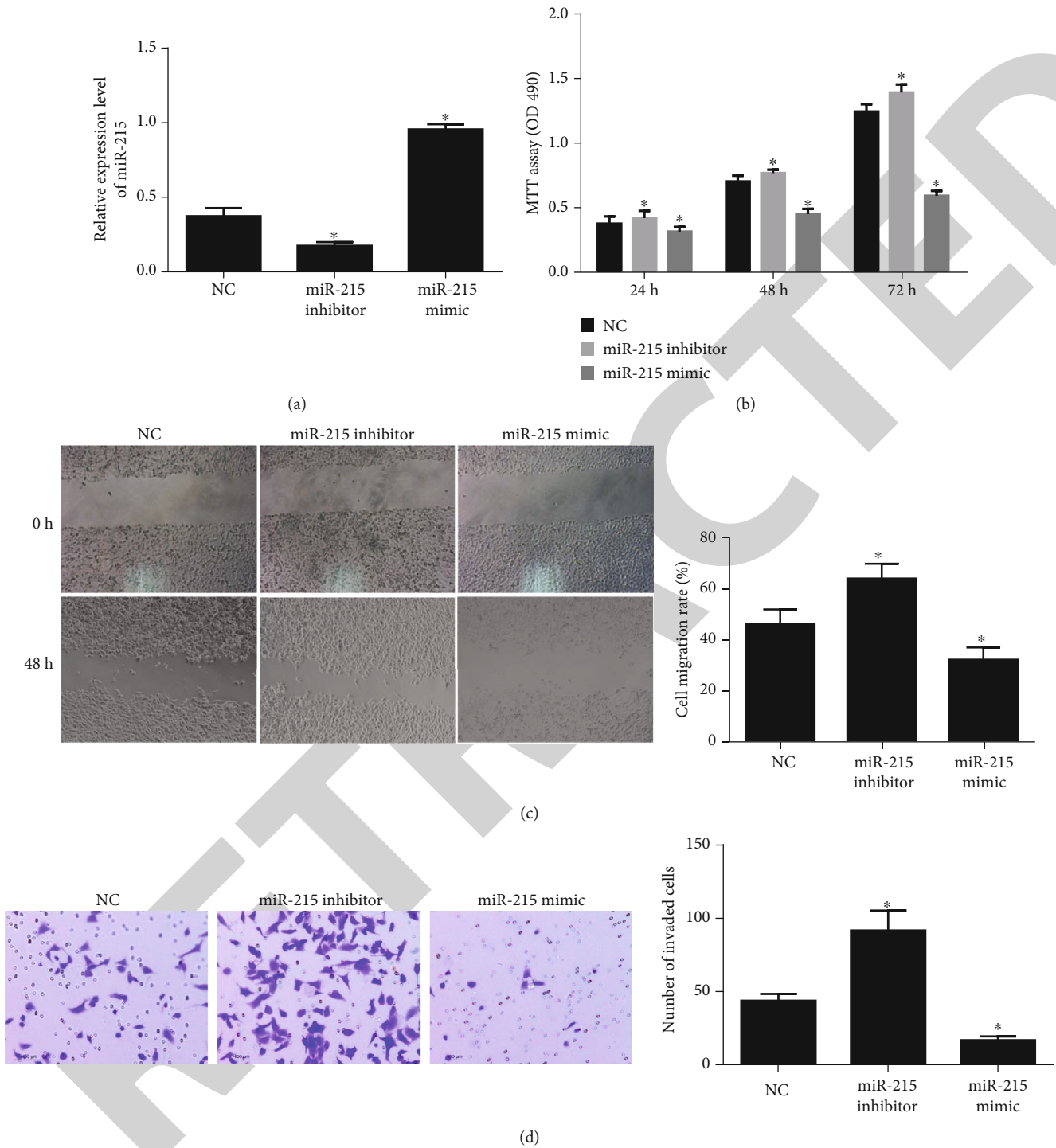


FIGURE 3: miR-215 inhibits CRC cell proliferation, invasion, and migration *in vitro*. (a) qRT-PCR is used to confirm the transfection efficiency of miR-215. (b) Cell proliferation is determined by the MTT assay. (c) Cell migration is determined by the wound healing assay. (d) The invasion of HT29 cells is examined using the Transwell assay, and the representative images are presented (* $P < 0.05$, compared to NC).

expression of miR-215 in normal tissue and CRC tissue by qRT-PCR and found that miR-215 was downregulated in CRC, and the expression of miR-215 was negatively correlated with SCD expression in CRC tissue. Meanwhile, the dual-luciferase assay verified the binding sites of miR-215 on SCD. In addition, we detected the expression of SCD after overexpressing miR-215 in CRC cells, and it was also discov-

ered that miR-215 could inhibit the expression of SCD in CRC cells. Thus, SCD was fully confirmed to be a downstream target of miR-215. Published literature indicates that the dysregulated miRNAs and mRNAs may have the function of regulating the occurrence and development of CRC [23–25]. Therefore, we further observed the effect of miR-215 on the proliferation, migration, and invasion of CRC

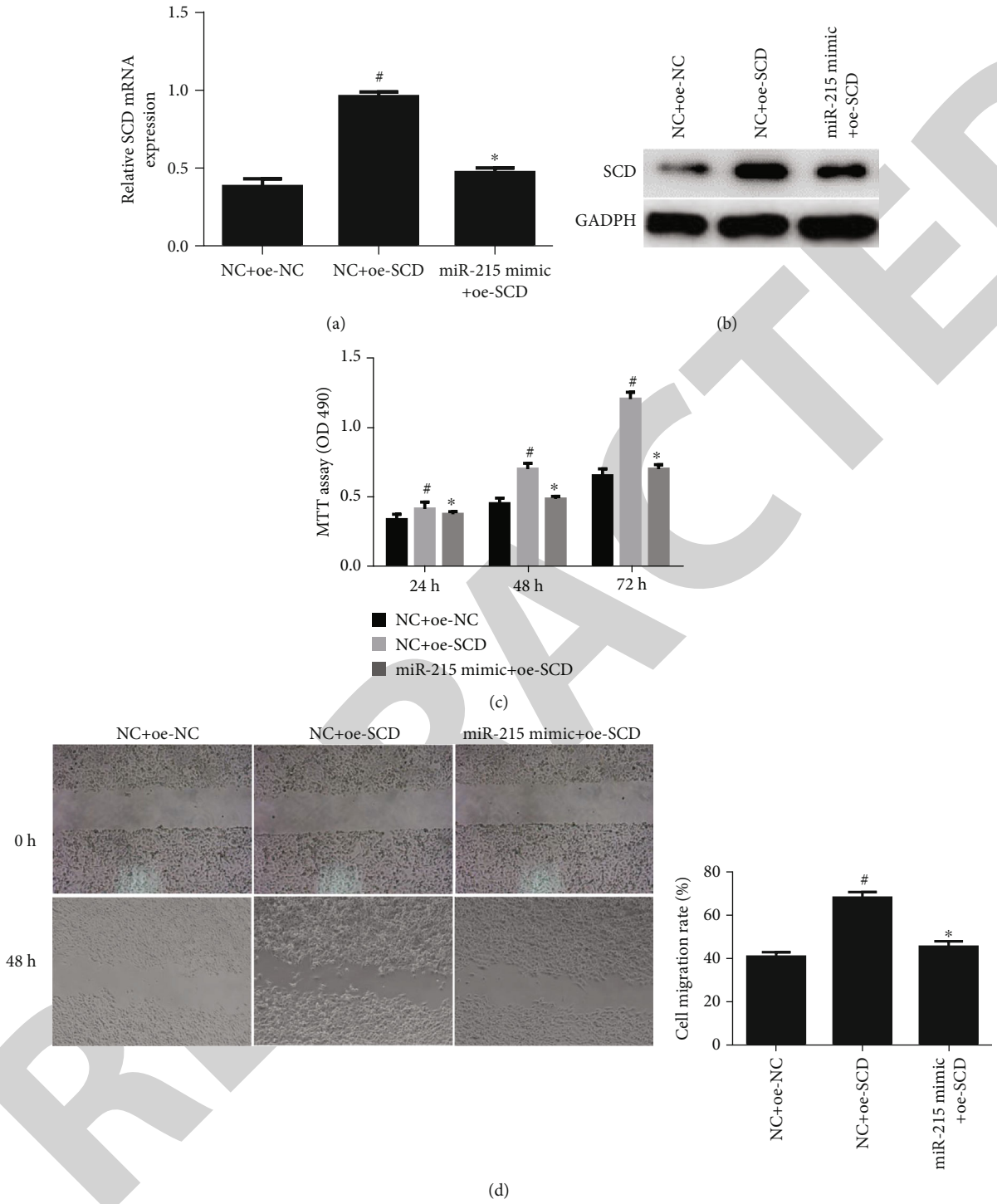
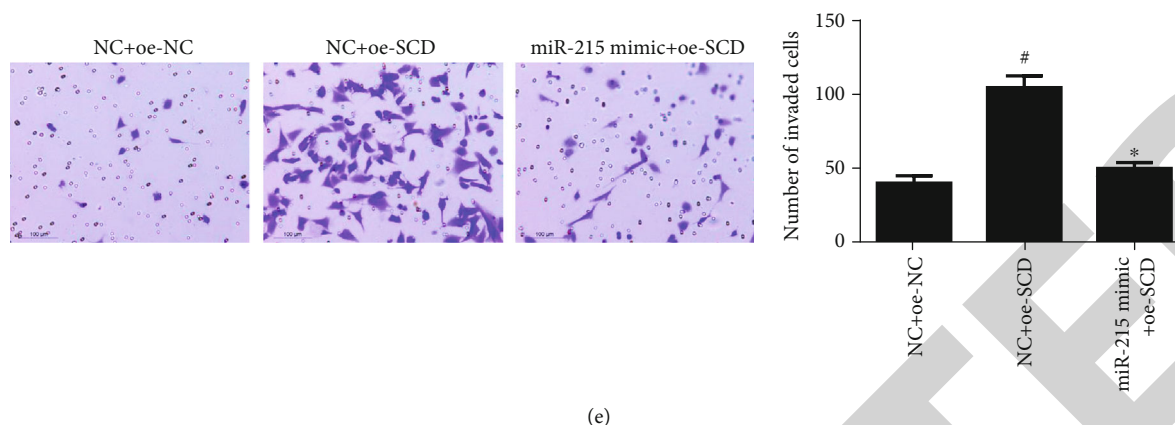


FIGURE 4: Continued.



(e)

FIGURE 4: miR-215 mediates the proliferation, migration, and invasion of CRC cells by targeting SCD. (a) SCD mRNA expression is measured by qRT-PCR. (b) SCD protein expression is detected by western blot. (c) Cell proliferation is determined by the MTT assay. (d) Cell migration is examined by the wound healing assay. (e) The invasion of HT29 cells is examined using the Transwell assay, and the representative images are presented ([#] $P < 0.05$, compared to NC+oe-NC; ^{*} $P < 0.05$, compared to NC+oe-SCD; NC is the control of the miR-215 mimic, and oe-NC is the control of oe-SCD).

cells through MTT, wound healing, and Transwell invasion assays and found that miR-215 had a significant inhibitory effect on CRC cells, while SCD could attenuate the inhibitory effect of miR-215 on CRC cells. The above results suggested that miR-215 could inhibit the proliferation, migration, and invasion of CRC cells by targeting SCD.

In summary, we demonstrated that the expression of miR-215 was downregulated in CRC tissue compared with adjacent normal colorectal tissue. miR-215 impaired CRC cell proliferation, migration, and invasion *in vitro* by inhibiting the expression of SCD. The results of this study contribute to the improvement of the understanding of the molecular mechanism underlying CRC development and progression and provide potential new therapeutic targets for the management of CRC.

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 have no competing interests.

Authors' Contributions

XHX, YD, JY, ZPW, JF, and RBY contributed to the study design. XHX, YD, HPJ, CC, JF, and RY are involved in the literature search. XHX, YD, JY, ZPW, HPJ, CC, JF, and RBY acquired the data. XHX, JF, and RBY wrote the article. All revised the article and gave the final approval of the version.

References

- [1] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, vol. 66, no. 4, pp. 683–691, 2017.
- [2] N. Keum and E. L. Giovannucci, "Epidemiology of colorectal cancer," in *Pathology and Epidemiology of Cancer*, M. Loda, L. Mucci, M. Mittelstadt, M. Hemelrijck, and M. Cotter, Eds., pp. 391–407, Springer, Cham, 2017.
- [3] D. A. Bluemke and J. A. C. Lima, "Using MRI to probe the heart in hypertrophic cardiomyopathy," *Radiology*, vol. 294, no. 2, pp. 287–288, 2020.
- [4] D. Mauvoisin, C. Charfi, A. M. Lounis, E. Rassart, and C. Mounier, "Decreasing stearyl-CoA desaturase-1 expression inhibits β -catenin signaling in breast cancer cells," *Cancer Science*, vol. 104, no. 1, pp. 36–42, 2013.
- [5] K. She, S. Fang, W. du et al., "SCD1 is required for EGFR-targeting cancer therapy of lung cancer via re-activation of EGFR/PI3K/AKT signals," *Cancer Cell International*, vol. 19, no. 1, p. 103, 2019.
- [6] E. A. Miska, "How microRNAs control cell division, differentiation and death," *Current Opinion in Genetics & Development*, vol. 15, no. 5, pp. 563–568, 2005.
- [7] G. Ge, W. Zhang, L. Niu, Y. Yan, Y. Ren, and Y. Zou, "miR-215 functions as a tumor suppressor in epithelial ovarian cancer through regulation of the X-chromosome-linked inhibitor of apoptosis," *Oncology Reports*, vol. 35, no. 3, pp. 1816–1822, 2016.
- [8] M. Karaayvaz, C. Zhang, S. Liang, K. R. Shroyer, and J. Ju, "Prognostic significance of miR-205 in endometrial cancer," *PLoS One*, vol. 7, no. 4, article e35158, 2012.
- [9] S. W. Zhou, B. B. Su, Y. Zhou et al., "Aberrant miR-215 expression is associated with clinical outcome in breast cancer patients," *Medical Oncology*, vol. 31, no. 11, p. 259, 2014.
- [10] Y. Hou, J. Zhen, X. Xu et al., "miR-215 functions as a tumor suppressor and directly targets ZEB2 in human non-small cell lung cancer," *Oncology Letters*, vol. 10, no. 4, pp. 1985–1992, 2015.
- [11] X. Gao, Y. Cai, and R. An, "miR215 promotes epithelial to mesenchymal transition and proliferation by regulating LEFTY2 in endometrial cancer," *International Journal of Molecular Medicine*, vol. 42, no. 3, pp. 1229–1236, 2018.
- [12] Y. Lin, Y. Jin, T. Xu, S. Zhou, and M. Cui, "MicroRNA-215 targets NOB1 and inhibits growth and invasion of epithelial

Research Article

Prediction of High-Risk Types of Human Papillomaviruses Using Reduced Amino Acid Modes

Xinnan Xu,¹ Rui Kong,¹ Xiaoqing Liu,² Pingan He ,³ and Qi Dai ¹

¹College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

²College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China

³College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

Correspondence should be addressed to Qi Dai; daialiu04@yahoo.com

Received 1 March 2020; Accepted 22 April 2020; Published 18 June 2020

Guest Editor: Lei Chen

Copyright © 2020 Xinnan Xu 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 human papillomavirus type plays an important role in the early diagnosis of cervical cancer. Most of the prediction methods use protein sequence and structure information, but the reduced amino acid modes have not been used until now. In this paper, we introduced the modes of reduced amino acids to predict high-risk HPV. We first reduced 20 amino acids into several nonoverlapping groups and calculated their structure and physicochemical modes for high-risk HPV prediction, which was tested and compared with the existing methods on 68 samples of known HPV types. The experiment result indicates that the proposed method achieved better performance with an accuracy of 96.49%, indicating that the reduced amino acid modes might be used to improve the prediction of high-risk HPV types.

1. Introduction

Cervical cancer is a cancer with a higher morbidity and mortality rate among women worldwide [1]. There are about 500,000 new cases of cervical cancer each year, with 280,000 deaths [2], which has become the second largest female cancer [3, 4]. Studies have indicated that human papillomavirus (HPV) infection is closely related to the occurrence and development of cervical cancer, and certain types of HPV cause abnormal tissue growth in the form of papilloma [5–7].

Human papillomavirus belongs to the papillomavirus family. It is an icosahedral, uncoated particle composed of double-stranded DNA of approximately 8,000 nucleotide base pairs [8, 9]. The circular DNA is about 55 nm in diameter [10–13]. To date, there are more than 150 types of human papillomavirus (HPV), and some new HPV types will be found when there are significant homologous differences between some new HPV types and defined HPV types [14–16]. Epidemiological studies have shown a strong correlation between genital HPV and cervical cancer. Genital HPV can be divided into three types according to its relative malignancy: low-risk type, intermediate-risk type, and

high-risk type. The clinical association studies usually use two types of HPV: high-risk and low-risk. Low-risk types are associated with low-grade lesions, while high-risk viral types are more closely related to high-grade cervical lesions and cancer [17]. High-risk types included HPV-16, HPV-18, HPV-26, HPV-31, HPV-33, HPV-35, HPV-39, HPV-45, HPV-51-53, HPV-56, HPV-58, HPV-59, HPV-66, HPV-68, HPV-70, HPV-73, HPV-82, and HPV-85 [18]. HPV-16 and HPV-18 accounted for 62.6% and 15.7% of cervical cancers [19], respectively. Therefore, the identification of high-risk HPV has become an important part of the diagnosis and treatment of cervical cancer.

Up to now, many epidemiological and experimental methods can identify HPV types [5, 20–22], mainly using polymerase chain reaction (PCR) technology, and be applied to rapid detection of clinical samples. With the rapid growth of human papillomavirus (HPV) data and sensitivity requirements, we need a reliable and effective calculation method to predict the high-risk types of HPV directly.

In recent years, several computational models have been proposed to predict high-risk HPV types. Eom et al. studied the sequence fragments and introduced genetic algorithms

to predict the HPV types [23]. Joung et al. used support vector machines to predict the HPV types based on the hidden Markov model [24, 25]. Park et al. proposed to use decision trees to predict human papillomavirus types [26]. Kim and Zhang calculated the distance of amino acid pairs and further predict the risk types of HPV based on E6 proteins [7, 9]. Kim et al. proposed a set of support vector machines (GSVM) for the classification of HPV types using the differential molecular sequence of protein secondary structure [13]. Esmaeili et al. used ROC to classify HPV types based on Chou's pseudo amino acid composition [27]. Alemi et al. compared the physicochemical properties between the high- and low-risk HPV types, and they used support vector machines to predict the high-risk HPV types [28].

These methods have performed well in the prediction of high-risk HPV types, but the challenge of extracting HPV information remains. The information widely used in the prediction of high-risk types of HPV is based on sequence information, but the information limited to the characteristics of 20 AAs and their reduction groups has not been explored so far. In this paper, we proposed a novel method to predict high-risk types of HPVs based on the reduced amino acid modes. We classified 20 amino acids into several groups and extract their structure and chemical properties. These extracted features were used to predict the high-risk type of HPVs based on a support vector machine. Through some experiments and comparative analysis, we want to evaluate the efficiency of the proposed method, as well as the efficiency of various reduced amino acid modes.

2. Materials and Methods

2.1. Datasets. There are eight open reading frames that encode early and late genes of the HPVs [11]. The early and late genes have polyA signal 1 and polyA signal 2. The produce of the late genes are L1 and L2 proteins which affect the viral capsid structure [12], while early genes are transformed into E1-E7 proteins. We constructed seven protein databases of the HPVs whose sequences are downloaded from the Los Alamos National Laboratory (LANL). Each protein has 72 HPV types. If a certain type of protein lacks the sequences of HPVs, we downloaded the missing sequence from the National Biotechnology Information Center. Since the E4 protein cannot be found in the National Biotechnology Information Center, its total number is 71. According to an HPV compendium, seventeen HPV types are classified as high-risk types (HPV-16, HPV-18, HPV-31, HPV-33, HPV-35, HPV-39, HPV-45, HPV-51, HPV-52, HPV-56, HPV-58, HPV-59, HPV-61, HPV-66, HPV-67, HPV-68, and HPV-72), and the remaining is low-risk type [13].

2.2. Reduced Amino Acids (RedAAs). 20 amino acids have subtle differences, but some of them have similar basic structures and functions. AAindex is a database of physical and biochemical indicators of amino acids established by Tomii and Kanehisa [29]. It mainly includes three parts: AAindex 1, AAindex 2, and AAindex 3. AAindex 1 is a database that describes the physicochemical and biological properties of amino acids. AAindex 2 is the matrix of amino acid muta-

tion, and AAindex 3 is the protein contact potential statistics. These data are from published articles. We mainly used AAindex 1 to calculate the correlation coefficient as the distance between the two indicators. AAindex 1 currently contains 544 indexes, and this article selected 522 indexes. These 522 characteristics are further divided into 7 categories: (A)—alpha and turn propensities, (B)—beta propensity, (C)—composition, (H)—hydrophobicity, (P)—physicochemical properties, and (O)—other properties [29].

Here, we introduced BLOSUM62 to classify amino acids to simplify sequence analysis [30]. We denote the i th group as X_i and denote its j th amino acid as $X_i(j)$. Using BLOSUM62, we calculated the similarity score $S(X_i(j), R_k)$ between $X_i(j)$ and the k th amino acid R_k as follows:

$$S(X_i(j), R_k) = \text{Blosum}(X_i(j), R_k), \quad (1)$$

where $\text{Blosum}(X_i(j), R_k)$ denotes the substitution value between $X_i(j)$ and R_k . Then, we summed up all scores of different groups as the score between Seq_s and Seq_0 :

$$S = \sum_{i=1}^N \left[\sum_{j=1}^{g_s(i)} \sum_{k=1}^{g_0(i)} m_i(k) S(X_i(j), R_k) \right] / g_s(i), \quad (2)$$

where $g_0(i)$ is the i th group size of Seq_0 , $g_s(i)$ is the i th group size of Seq_s , $m_i(k)$ is the total number of R_k occurrences in Seq_0 , and N is the group size. S measures the degree of retention of parent sequence information. Given a size N group, we analyzed all amino acid groups and calculated the similarity score between the parent sequence and the reduced sequence. The reduced alphabets were selected according to their scores. For example, 20 AAs are reduced into 9 RedAAs ($\{C\}$, $\{G\}$, $\{P\}$, $\{IMLV\}$, $\{AST\}$, $\{NH\}$, $\{YFW\}$, $\{DEQ\}$, and $\{RK\}$) in the BLOSUM62 matrix.

2.3. Reduced Amino Acid Modes (RedAA Modes). 20 amino acids were divided into the following nonoverlapping groups according to their physicochemical properties in AAindex, and four types of the reduced amino acid modes were calculated as protein structural and physicochemical features.

2.3.1. Content Modes. The first mode is associated with the content-specific features, including the distribution of the RedAA and RedAA pattern in protein sequences.

(1) *K-mer.* Protein sequences and peptides can be seen as a collection of symbols, and their characteristics can be analyzed by the frequency of their small fragments. k -mers are k consecutive characters in reduced proteins, and a sliding window of length m can be used to calculate their frequencies [31–33], moving from position 1 to $m - k + 1$ with one base at a time. It allows the overlaps of the k -mers and is calculated as

$$f_{w_{\text{RedAA}}} = \frac{\text{Count}_{w_{\text{RedAA}}}}{\sum_{x \in \mathfrak{R}} \text{Count}_x}, \quad (3)$$

where $\text{Count}_{w_{\text{RedAA}}}$ is the occurrence number of the k -mer w_{RedAA} and \mathfrak{R} is k -mer set of the RedAAs.

(2) *RCTD*. “Composition (C),” “Transition (T),” and “Distribution (D)” are three descriptors of RedAAs, which are defined as follows [34, 35]:

Composition: it can be regarded as a single monomer of the reduced sequence, and the sequence components are described by calculating the percentage of each RedAA.

Transition: it can be used as the conversion of RedAA I and A by calculating the frequency of I followed by A :

$$T_{IA} = \frac{\text{Count}_{IA} + \text{Count}_{AI}}{N - 1}, \quad (4)$$

where Count_{IA} and Count_{AI} are the “ IA ” and “ AI ” numbers, respectively, in the reduced sequence with length N .

Distribution: it describes the RedAA distribution in the reduced sequence, including the specified coding categories: 25%, 50%, 75%, and 100%.

(3) *PRseAAC*. Type I PRseAAC and type II PRseAAC are widely used pseudoreduced AA compositions (PRseAAC) [36–38].

Type I PRseAAC was proposed by Kuo-Chen Chou, which is defined as follows:

$$\begin{aligned} \text{PRseAAC1}_u &= \frac{f_u}{\sum_{i=1}^R f_i + w \sum_{j=1}^\lambda \theta_j}, \quad u \leq R, \\ \text{PRseAAC1}_u &= \frac{w \theta_u}{\sum_{i=1}^R f_i + w \sum_{j=1}^\lambda \theta_j}, \quad R \leq u \leq R + \lambda, \end{aligned} \quad (5)$$

where f_i is the RedAA frequency and w is the weighting factor. θ_i is calculated as

$$\begin{aligned} \theta_\lambda &= \frac{1}{N - \lambda} \left(\sum_{i=1}^{N-\lambda} \Theta(R_i, R_{i+\lambda}) \right), \\ \Theta(R_i, R_j) &= \frac{(\text{SH}_1(R_i) - \text{SH}_1(R_j))^2 + (\text{SH}_2(R_i) - \text{SH}_2(R_j))^2 + (\text{SH}_3(R_i) - \text{SH}_3(R_j))^2}{3}, \\ \text{SH}_i(\text{RedAA}_i) &= \frac{H_i(\text{RedAA}) - \left(\sum_{j=1}^R H_i(j)/R \right)}{\sqrt{\sum_{t=1}^R \left(H_i(t) - \left(\sum_{j=1}^R H_i(j)/R \right) \right)^2 / R}}, \end{aligned} \quad (6)$$

where $H_i(\text{RedAA})$ is the RedAAs’ property and R is the RedAA size.

Type II PRseAAC can be calculated as

$$\begin{aligned} \text{PRseAAC2}_u &= \frac{f_u}{\sum_{i=1}^R f_i + w \sum_{j=1}^\lambda \tau_j}, \quad u \leq R, \\ \text{PRseAAC2}_u &= \frac{w \tau_u}{\sum_{i=1}^R f_i + w \sum_{j=1}^\lambda \tau_j}, \quad R \leq u \leq R + \lambda, \\ \tau_{2\lambda-1} &= \frac{1}{N - \lambda} \sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^1, \\ \tau_{2\lambda} &= \frac{1}{N - \lambda} \sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^2, \\ H_{i,j}^1 &= \text{SH}_1(\text{RedAA}_i) \text{SH}_1(\text{RedAA}_j), \\ H_{i,j}^2 &= \text{SH}_2(\text{RedAA}_i) \text{SH}_2(\text{RedAA}_j), \end{aligned} \quad (7)$$

where f_i is the RedAA frequency, w is the weighting factor, $\text{SH}_i(\text{RedAA})$ is the RedAAs’ property, R is the RedAA size, and N is the sequence length.

2.3.2. Correlation Mode. The second RedAA mode is based on the characteristics of correlation, which describes the correlation among the RedAAs. In the proposed RedAA mode, three different autocorrelation features are implemented: normalized Moreau–Broto autocorrelation (NMB) [39], Moran autocorrelation (M) [40], and Geary autocorrelation (G) [41].

(1) *NMB*. The RedAA NMB is defined as

$$\text{NMB}(d) = \frac{\sum_{i=1}^{N-d} P_i^{\text{RedAA}} P_{i+d}^{\text{RedAA}}}{N - d}, \quad (8)$$

where P_i^{RedAA} denotes the RedAA property at position i of the sequence, d is the autocorrelation lag, and N is the sequence length.

(2) M . The RedAA M can be calculated as

$$M(d) = \frac{1/(N-d) \sum_{i=1}^{N-d} (P_i^{\text{RedAA}} - \bar{P}^{\text{RedAA}}) (P_{i+d}^{\text{RedAA}} - \bar{P}^{\text{RedAA}})}{1/N \sum_{i=1}^N (P_i^{\text{RedAA}} - \bar{P}^{\text{RedAA}})^2},$$

$$\bar{P}^{\text{RedAA}} = \frac{1}{N} \sum_{i=1}^N P_i^{\text{RedAA}}, \quad (9)$$

where P_i^{RedAA} denotes the RedAA property at position i of the sequence, d is the autocorrelation lag, and N is the sequence length.

(3) G . The RedAA G is defined as

$$G(d) = \frac{1/(2(N-d)) \sum_{i=1}^{N-d} (P_i^{\text{RedAA}} - P_{i+d}^{\text{RedAA}})^2}{1/N \sum_{i=1}^N (P_i^{\text{RedAA}} - \bar{P}^{\text{RedAA}})^2}, \quad (10)$$

$$\bar{P}^{\text{RedAA}} = \frac{1}{N} \sum_{i=1}^N P_i^{\text{RedAA}},$$

where P_i^{RedAA} denotes the RedAA property at position i of the sequence, d is the autocorrelation lag, and N is the sequence length.

2.3.3. *Order Mode*. The order mode reflects the physical and chemical interaction among the RedAA pairs. There are two kinds of order modes: sequence coupling score and quasi-sequence score [42].

(1) *Sequence Coupling Score*. The sequence coupling score is calculated:

$$\tau_d^{\text{RedAA}} = \sum_{i=1}^{N-d} d_{i,i+d}^{\text{RedAA}}, \quad (11)$$

where $d_{i,i+d}^{\text{RedAA}}$ is the Schneider-Wrede physicochemical distance or Grantham chemical distance between the RedAAs at positions i and $i+d$ and $1 \leq d \leq N$.

(2) *Quasi-Sequence Score*. The quasi-sequence score of the RedAA is defined:

$$\kappa_{\text{RedAA}} = \frac{f_{\text{RedAA}}}{\sum_{i=1}^R f_{\text{RedAA}_i} + w \sum_{d=1}^M \tau_d^{\text{RedAA}}}, \quad (12)$$

where f_{RedAA_i} is the RedAA frequency and w denotes the weighting factor.

The quasi-sequence score can be calculated as

$$\kappa_{\tau} = \frac{w \tau_d^{\text{RedAA}}}{\sum_{i=1}^R f_{\text{RedAA}_i} + w \sum_{d=1}^M \tau_d^{\text{RedAA}}}, \quad (13)$$

where τ is the sequence coupling score, f_{RedAA_i} is the RedAA frequency, and w denotes the weighting factor.

2.3.4. *Position Mode*. The position mode represents the distribution of RedAA positions of protein sequences based on the coefficient of variations [32, 43]. First, we converted the protein sequence into a digital sequence $N(\text{RedAA})$ and calculated the probabilities $P_{\text{RedAA}}(\xi)$ of the separation distance ξ between two adjacent RedAAs. The mean $E_{(\text{RedAA})}(\xi)$ and variance $D_{(\text{RedAA})}(\xi)$ are defined:

$$E_{(\text{RedAA})}(\xi) = \sum_{\xi} \xi \times P_{(\text{RedAA})}(\xi), \quad (14)$$

$$D_{(\text{RedAA})}(\xi) = E_{(\text{RedAA})}(\xi^2) - [E_{(\text{RedAA})}(\xi)]^2.$$

We then calculated the positional information $C_{(\text{RedAA})}(\xi)$:

$$C_{(\text{RedAA})}(\xi) = \frac{E_{(\text{RedAA})}(\xi)}{\sqrt{D_{(\text{RedAA})}(\xi)}}, \quad (15)$$

where $C_{(\text{RedAA})}(\xi)$ is the reciprocal of the coefficient of variation (CV) which compares the degree of change between two datasets, even if there are large differences between their means. In this paper, it was denoted as the RedAA position characteristics.

2.4. *Prediction Algorithm*. $Y = [y_1, y_2, \dots, y_n]^T$ is an HPV label set, $y_i = 1$ is from the high-risk type, and $y_i = 2$ is from the low-risk type. We used x_{ij} to represent the j th features of the RedAA modes of the i th HPV sample, where $j = 1, 2, \dots, m$. All of the features of the RedAA modes for all HPV samples are denoted as

$$X = \begin{matrix} & \text{index1} & \text{index2} & \cdots & \text{indexm} \\ \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \end{matrix}. \quad (16)$$

We used a support vector machine (SVM) to predict the HPV type, which is expressed as follows:

$$\begin{aligned} \min_{w,b,\xi} \quad & J(w, b, \xi) = \frac{1}{2} (w^T w) + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \begin{cases} y_i [w^T \varphi(x_i) + b] \geq 1 - \xi_i, & i = 1, 2, \dots, n, \\ \xi_i \geq 0, & i = 1, 2, \dots, n, \end{cases} \end{aligned} \quad (17)$$

where w is a linear combination of a set of nonlinear data conversion:

$$w = \sum_{i=1}^n \alpha_i y_i \varphi(x_i), \quad (18)$$

where b denotes the bias term, C denotes some regularization parameters, and ξ_i is the training error. The above problem can be expressed:

$$\begin{aligned} \max_{\alpha} \quad & J(\alpha) = \max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j) \\ \text{subject to} \quad & \begin{cases} \sum_{i=1}^n \alpha_i y_i = 0, & i = 1, 2, \dots, n, \\ 0 \leq \alpha_i \leq C, & i = 1, 2, \dots, n. \end{cases} \end{aligned} \quad (19)$$

Here, the Gaussian kernel function is used to calculate $\varphi(x_i)^T \varphi(x_j)$ instead of $\varphi(x_i)$ and $\varphi(x_j)$. The separation problem can be expressed:

$$\begin{aligned} f(x) &= \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b, \\ y(x) &= \text{sign}[f(x)]. \end{aligned} \quad (20)$$

The training model can predict the risk type of the test sample $x \in R^m$ according to the following formula:

$$y(x) = \begin{cases} 1, & \text{if } f(x) > 0, \\ 2, & \text{if } f(x) \leq 0. \end{cases} \quad (21)$$

$y(x) = 1$ indicates that the sample x belongs to the high-risk type; otherwise, it belongs to the low-risk type. In order to obtain a better model, we used a simple grid search strategy based on 10-fold cross-validation to find the optimal model for each dataset.

3. Results and Discussion

3.1. Evaluation Measures. There are three popular methods to evaluate the efficiency of prediction models: subsampling test, independent test, and jackknife test. Since the jackknife test can evaluate the efficiency of various predictor variables, we used it to evaluate the efficiency of the proposed method and calculated the class accuracies and overall accuracies:

$$\begin{aligned} \text{specificity(accuracy of high-risk type)} &= \frac{a}{a+c}, \\ \text{sensitivity(accuracy of low-risk type)} &= \frac{d}{b+d}, \\ \text{accuracy of totality} &= \frac{a+d}{a+b+c+d} \cdot 100\%, \end{aligned} \quad (22)$$

where a denotes true positives, c denotes false positives, d denotes true negatives, and b denotes false negatives.

3.2. HPV Classification. We used the jackknife test to evaluate the performance of the proposed RedAA modes. We divided the 20 amino acids into 5 to 19 groups and calculated their RedAA modes as protein features and then input them into the support vector machine to predict the HPV type. Table 1 shows the tagged HPV types and the predicted results.

It can be seen from Table 1 that the 65 HPV types predicted by our method are consistent with the actual types and have better performance. However, HPV-72 is predicted to be low-risk but is actually high-risk, and HPV-30 is predicted to be high-risk but is actually low-risk. For further comparison, we compared our results with Kim et al.'s results [13]. For Kim et al.'s prediction, HPV-56 was predicted to be potentially high-risk, and we predicted it to be high-risk; HPV-53 and HPV-73 were predicted to be potentially high-risk, but in our results, they were low-risk. Phylogenetic analysis showed that HPV-30 was closely related to the established oncogenic type HPV-56, suggesting that HPV-30 was more likely to be a high-risk type. The results show that the proposed method is more consistent with the actual risk type.

We further compared our method with the following method: SVM based on the mismatch [24], SVM classifier based on the linear kernel [13], SVM based on the gap spectral kernel (Gap) [7], BLAST model [13] and integrated SVM (Ensemble) [13], and two text prediction methods based on AdaCost [26] and naive Bayes [26]. The accuracy of our method reaches 96.49%, while the accuracy of the integrated SVM is 94.12%, the accuracy of the SVM based on the unmatched kernel is 92.70%, the accuracy of the SVM based on the linear kernel is 90.28%, and the accuracy of BLAST reaches 91.18%. As for the text prediction method, AdaCost [26] has an accuracy rate of 93.05%, while naive Bayes [26] has an accuracy rate of 81.94%. The comparison also shows that the RedAA model is more effective in classifying the risk types of human papillomaviruses.

3.3. The Performance of the Early and Late Proteins in HPV Type Prediction. Early HPV proteins contain E1, E2, E4, E5, E6, and E7, and late proteins include L1 and L2 [3, 5]. Information commonly used for high-risk and low-risk HPV prediction includes information on protein sequences, secondary structure, and pseudoamino acid composition, in which most of them use E6, E7, or L1 protein [23–28]. In this paper, we used seven protein datasets of early and late proteins in HPV type prediction and compared their performance. Figure 1 compares the accuracy of each category and the overall accuracy based on early and late proteins.

Figure 1 shows that the prediction accuracy of low-risk types is higher than that of high-risk types, except for E5 protein. L1 protein outperforms other HPV proteins in the prediction of low-risk types. L2 protein performs best in high-risk type predictions. The above research shows that E6, E7, L1, and L2 proteins are closely related to high-risk HPV and play an important role in the occurrence and development of diseases [14]. The function of L1 protein in low-

TABLE 1: Comparison of the real risk types (REAL) and the prediction results using the proposed approach.

Types	Real	Predicted	Types	Real	Predicted	Types	Real	Predicted	Types	Real	Predicted
HPV-39	High	High	HPV-7	Low	Low	HPV-34	Low	Low	HPV-50	Low	Low
HPV-72	High	Low	HPV-30	Low	High	HPV-44	Low	Low	HPV-5	Low	Low
HPV-33	High	High	HPV-73	Low	Low	HPV-43	Low	Low	HPV-20	Low	Low
HPV-51	High	High	HPV-6	Low	Low	HPV-32	Low	Low	HPV-23	Low	Low
HPV-16	High	High	HPV-27	Low	Low	HPV-24	Low	Low	HPV-19	Low	Low
HPV-56	High	High	HPV-13	Low	Low	HPV-8	Low	Low	HPV-47	Low	Low
HPV-18	High	High	HPV-55	Low	Low	HPV-48	Low	Low	HPV-22	Low	Low
HPV-59	High	High	HPV-2	Low	Low	HPV-12	Low	Low	HPV-25	Low	Low
HPV-52	High	High	HPV-10	Low	Low	HPV-49	Low	Low	HPV-9	Low	Low
HPV-35	High	High	HPV-42	Low	Low	HPV-15	Low	Low	HPV-36	Low	Low
HPV-68	High	High	HPV-28	Low	Low	HPV-21	Low	Low	HPV-41	Low	Low
HPV-58	High	High	HPV-40	Low	Low	HPV-4	Low	Low	HPV-63	Low	Low
HPV-31	High	High	HPV-3	Low	Low	HPV-65	Low	Low	HPV-1	Low	Low
HPV-66	High	High	HPV-11	Low	Low	HPV-37	Low	Low	HPV-80	Low	Low
HPV-45	High	High	HPV-29	Low	Low	HPV-38	Low	Low	HPV-77	Low	Low
HPV-61	High	High	HPV-74	Low	Low	HPV-60	Low	Low	HPV-76	Low	Low
HPV-67	High	High	HPV-53	Low	Low	HPV-17	Low	Low	HPV-75	Low	Low

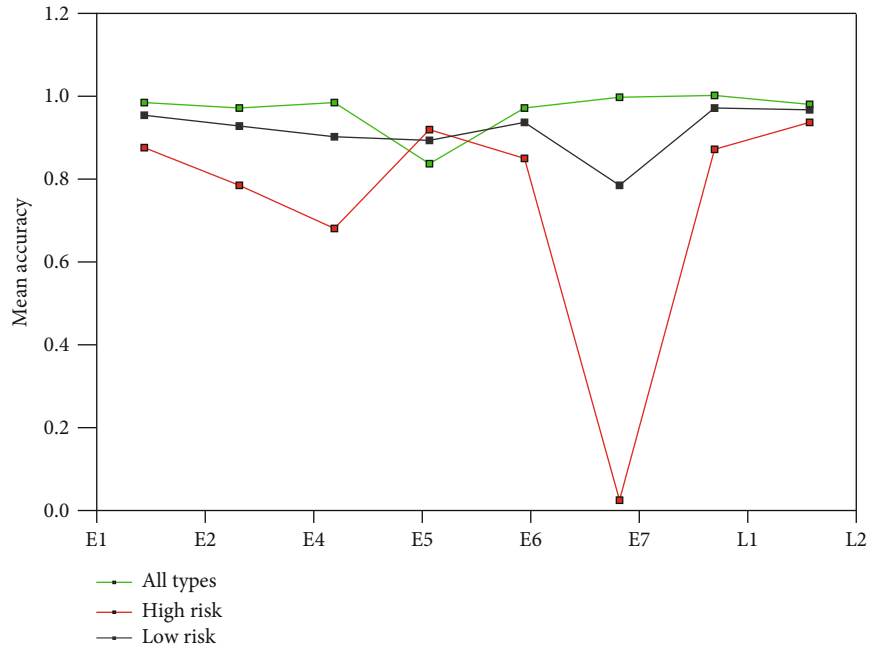


FIGURE 1: Comparison of prediction accuracy of each class based on all the early and late proteins.

risk and high-risk types is not exactly the same. L1 protein in the high-risk type exists in the form of integration, and L1 gene product self-assembly efficiency is low. L1 protein in the low-risk type exists in the form of free tissue, with high self-assembly efficiency. In high-risk typing, if L1 protein mutates, L1 protein cannot combine with L2 protein to form capsid protein and then cannot assemble HPV-infected virus particles. When HPV enters the host cell, the viral DNA replicates in large quantities and can integrate with the host cell DNA, resulting in host cell infection, infinite value addition,

and cell immortalization. The results show that L1 protein performs better in the prediction of high-risk HPV types, while L2 protein is more suitable for low-risk HPV types.

3.4. Influence of the Physicochemical Properties of Amino Acids. The proposed method reduced 20 AAs into several nonoverlapping groups, which relies heavily on the physical and biochemical indices of amino acids. The 522 characteristics of AAindex are divided into seven categories according to their physical and biochemical features [29]. The largest

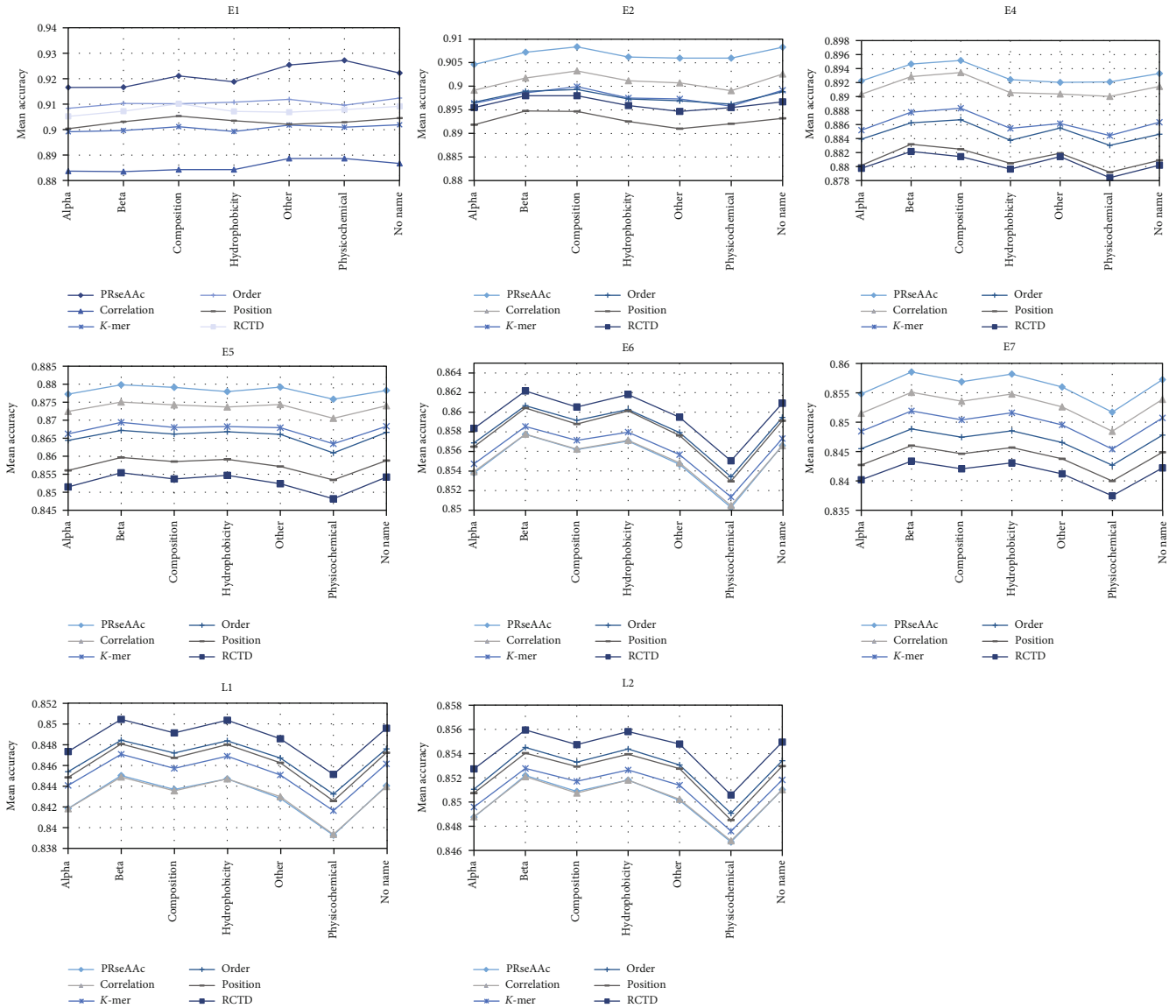


FIGURE 2: Comparison of the mean of the overall accuracies of HPV type prediction based on seven physicochemical property classes and six RedAA modes for all the early and late proteins.

group is hydrophobicity and the second largest group is alpha and turn propensities, and the sizes of the other four groups are relatively small. For each HPV protein, we used 522 physicochemical properties to calculate six kinds of reduced AA modes. For each class of the physicochemical properties of amino acids, we calculated their mean of the overall accuracies of HPV type prediction. The comparison of different physicochemical property classes and the RedAA modes is shown in Figure 2.

From Figure 2, it can be found that the proposed prediction has no obvious preference among 7 classes of physicochemical properties for E1 proteins. As for E2 proteins, composition is the best of the six reduced AA modes. For E4 proteins, the physicochemical properties of beta and composition are better. For the reduced AA mode position and RCTD, the physicochemical properties of beta are better in prediction, but composition is better for the other four

modes. The results of E5, E6, E7, L1, and L2 proteins are similar to those of E2 proteins, and the six reduced AA modes show better performance in beta physicochemical properties. These results indicate that E5, E6, E7, L1, and L2 proteins have a preference for beta physicochemical properties to reduce amino acids and calculate the six reduced AA modes in HPV type prediction.

3.5. Comparison of the Reduced Amino Acid Modes. In order to evaluate the performance of different modes, we used 522 physicochemical properties to calculate the RedAA modes of all the early and late proteins and calculated their average of the overall accuracies of HPV type prediction, which is shown in Figure 2. Figure 2 shows that six RedAA modes have the same preference trend among seven classifications of the physicochemical properties. As for E1, E2, E4, E5, and E7 proteins, PRseAAC is better than the other RedAA

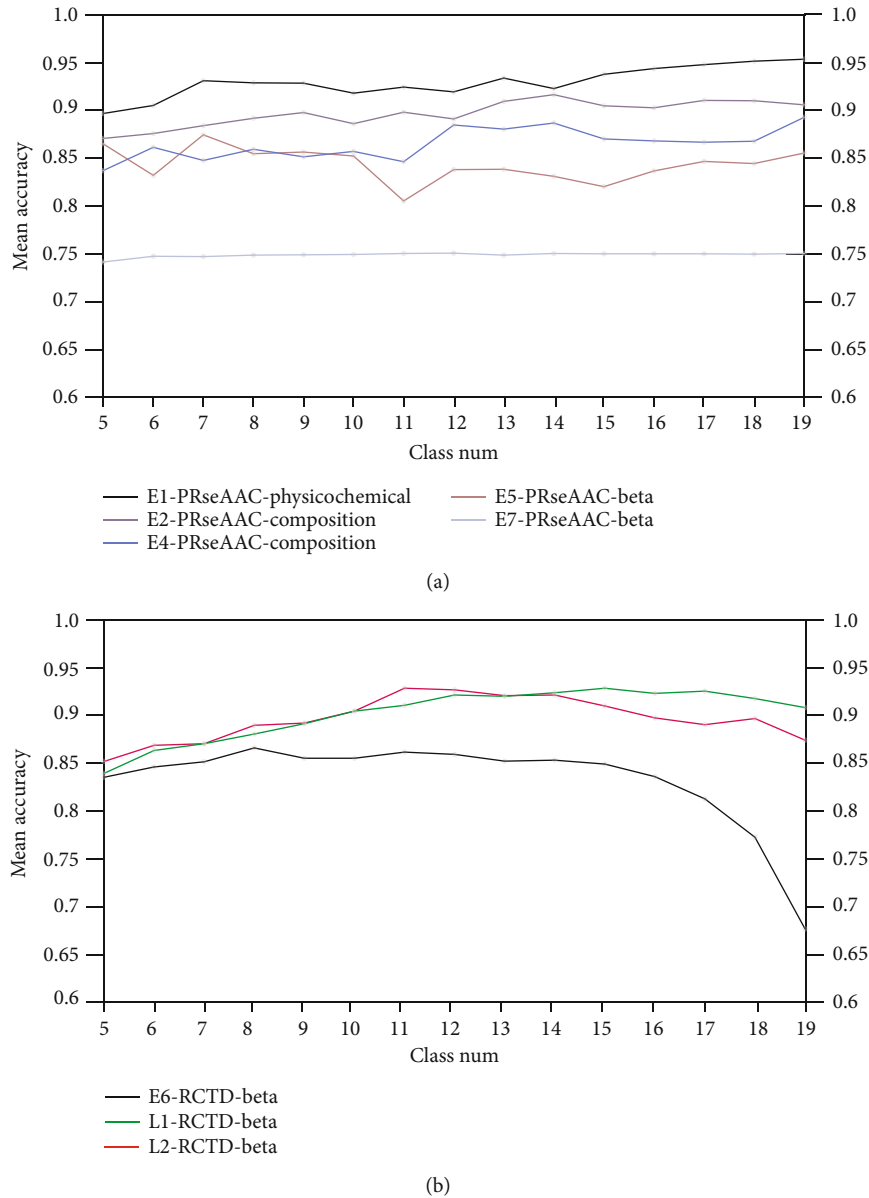


FIGURE 3: Performance comparison of the RedAA modes PRseAAC and RTCD with different reduced amino acids: (a) the average accuracies of the PRseAAC and RTCD with 5-19 reduced amino acids for E1, E2, E4, E5, and E7 and (b) the average accuracies of the PRseAAC and RTCD with 5-19 reduced amino acids for E6, L1, and L2.

modes, and the average accuracy of its prediction of HPV typing is also significantly higher than the average of other RedAA modes. As for E6, L1, and L2 proteins, RTCD outperforms the other five RedAA modes. In addition, PRseAAC and RTCD show better performance in beta physicochemical properties of the amino acids.

3.6. Influence of the Number of Reduced Amino Acids. The proposed method used the structural and physicochemical features of reduced amino acids, which reduces the dimension of input information and improves the efficiency of the prediction model. However, it should be noted that the RedAA modes are associated with the number of reduced amino acids. In order to discuss the influence of the RedAA size, we reduced 20 amino acids into 5-19 classes based on

522 physicochemical properties and calculated their RedAA modes PRseAAC and RTCD for of all the early and late proteins. The average accuracies of the RedAA modes PRseAAC and RTCD with 5-19 RedAAs are summarized in Figure 3.

Figure 3 shows the accuracy of HPV type prediction with the increase in reduced amino acids when combining the PRseAAC and physicochemical properties of amino acids for E1 proteins, and the best-performing PRseAAC achieves 95.378% accuracy with 19 reduced amino acids. For E2 proteins, the prediction model achieves the best performance with the PRseAAC and the physical and physicochemical properties of the composition class when amino acids are reduced to 14 classes. As for E5 and E7, PRseAAC achieves 87.18% and 75.07% accuracies when 20 amino acids are reduced to 7 and 12 classes, respectively. For E6, L1, and L2

proteins, the combination of the RCTD and beta physicochemical properties achieves best performances with 8, 15, and 11 reduced amino acids, respectively.

4. Conclusion

Genital papillomavirus is closely related to cervical cancer, especially high-risk HPV. Therefore, the identification of the HPV risk type is of great significance for the cervical cancer. We proposed a computational method for the prediction of the high-risk HPV based on the RedAA modes. With the help of the physicochemical properties of the amino acids, we reduced 20 amino acids into several nonoverlapping groups and calculated the structure and physicochemical characteristics of reduced AAs (RedAA) as the RedAA modes. We used reduced sequence information to predict high-risk types of HPV. Experiments with 68 known HPV types show that the proposed method has better performance than previous methods.

The first contribution is that L1 protein performs better in the prediction of high-risk HPV types, while L2 protein is more suitable for low-risk HPV types. The second contribution can be indicated from the influence of the physicochemical properties of amino acids; we noticed that E5, E6, E7, L1, and L2 proteins have a preference for beta physicochemical properties to reduce amino acids. The third contribution can be deduced from the comparison of the reduced amino acid modes; we found that the PRseAAC and RTCD outperform the other four RedAA modes and show better performance in beta physicochemical properties of the amino acids. The final contribution can be seen from the influence of the number of reduced amino acids; we noticed that the combination of the RCTD and beta physicochemical properties achieves the best performances with 8, 15, and 11 reduced amino acids for E6, L1, and L2 proteins, respectively.

Data Availability

All the data used to support the findings of this study are available from the Los Alamos National Laboratory (<https://pave.niaid.nih.gov/lanl-archives>).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61772028) and research Grants from Zhejiang Provincial Natural Science Foundation of China (LY20F020016).

References

- [1] E. K. Yim and J. S. Park, "Role of proteomics in translational research in cervical cancer," *Expert Review of Proteomics*, vol. 3, no. 1, pp. 21–36, 2014.
- [2] O. Peralta-Zaragoza, V. H. Bermúdez-Morales, C. Pérez-Plasencia, J. Salazar-León, C. Gómez-Cerón, and V. Madrid-Marina, "Targeted treatments for cervical cancer: a review," *OncoTargets and Therapy*, vol. 5, pp. 315–328, 2012.
- [3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: a Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.
- [4] D. Forman, C. de Martel, C. J. Lacey et al., "Global burden of human papillomavirus and related diseases," *Vaccine*, vol. 30, no. 5, pp. F12–F23, 2012.
- [5] F. X. Bosch, M. M. Manos, N. Munoz et al., "Prevalence of human papillomavirus in cervical cancer: a worldwide perspective," *Journal of the National Cancer Institute*, vol. 87, no. 11, pp. 796–802, 1995.
- [6] M. H. Schiffman, H. M. Bauer, R. N. Hoover et al., "Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia," *Journal of the National Cancer Institute*, vol. 85, no. 12, pp. 958–964, 1993.
- [7] S. Kim and J.-H. Eom, "Prediction of the human papillomavirus risk types using gap-spectrum kernels," *LNCS*, vol. 3973, pp. 710–715, 2006.
- [8] C. L. Pang and F. Thierry, "Human papillomavirus proteins as prospective therapeutic targets," *Microbial Pathogenesis*, vol. 58, pp. 55–65, 2013.
- [9] S. Kim and B.-T. Zhang, "Human papillomavirus risk type classification from protein sequences using support vector machines," *LNCS*, vol. 3907, pp. 57–66, 2006.
- [10] J. Haedicke and T. Iftner, "Human papillomaviruses and cancer," *Radiotherapy and Oncology*, vol. 108, no. 3, pp. 397–402, 2013.
- [11] J. Peng, L. Gao, J. Guo et al., "Type-specific detection of 30 oncogenic human papillomaviruses by genotyping both E6 and L1 genes," *Journal of Clinical Microbiology*, vol. 51, no. 2, pp. 402–408, 2013.
- [12] M. S. Longworth and L. A. Laimins, "Pathogenesis of human papillomaviruses in differentiating epithelia," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 362–372, 2004.
- [13] S. Kim, J. Kim, and B. T. Zhang, "Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures," *Computers in Biology and Medicine*, vol. 39, no. 2, pp. 187–193, 2009.
- [14] E. M. de Villiers, C. Fauquet, T. R. Broker, H. U. Bernard, and H. zur Hausen, "Classification of papillomaviruses," *Virology*, vol. 324, no. 1, pp. 17–27, 2004.
- [15] K. Münger, A. Baldwin, K. M. Edwards et al., "Mechanisms of human papillomavirus-induced oncogenesis," *Journal of Virology*, vol. 78, no. 21, pp. 11451–11460, 2004.
- [16] M. L. Eide and H. Debaque, "HPV detection methods and genotyping techniques in screening for cervical cancer," *Annales de Pathologie*, vol. 32, no. 6, pp. e15–e23, 2012.
- [17] M. F. Janicek and H. E. Averette, "Cervical cancer: prevention Diagnosis, and Therapeutics," *CA: A Cancer Journal for Clinicians*, vol. 51, no. 2, pp. 92–114, 2001.
- [18] M. D. Kaspersen, P. B. Larsen, H. J. Ingerslev et al., "Identification of multiple HPV types on spermatozoa from human sperm donors," *PLoS One*, vol. 6, no. 3, article e18095, 2011.
- [19] P. Guan, R. Howell-Jones, N. Li et al., "Human papillomavirus types in 115,789 HPV-positive women: a meta-analysis from cervical infection to cancer," *International Journal of Cancer*, vol. 131, no. 10, pp. 2349–2359, 2012.

- [20] H. Furumoto and M. Irahara, "Human papilloma virus (HPV) and cervical cancer," *Journal of Medical Investigation*, vol. 49, no. 3-4, pp. 124-133, 2002.
- [21] R. D. Burk, G. Y. F. Ho, L. Beardsley, M. Lempa, M. Peters, and R. Bierman, "Sexual behavior and partner characteristics are the predominant risk factors for genital human papillomavirus infection in young women," *The Journal of Infectious Diseases*, vol. 174, no. 4, pp. 679-689, 1996.
- [22] N. Muñoz, F. X. Bosch, S. de Sanjosé et al., "Epidemiologic classification of human papillomavirus types associated with cervical cancer," *New England Journal of Medicine*, vol. 348, no. 6, pp. 518-527, 2003.
- [23] J.-H. Eom, S.-B. Park, and B.-T. Zhang, "Genetic mining of DNA sequence structures for effective classification of the risk types of human papillomavirus(HPV)," in *Neural Information Processing*, N. R. Pal, N. Kasabov, R. K. Mudi, S. Pal, and S. K. Parui, Eds., pp. 1334-1343, Springer, Berlin, Heidelberg, 2004.
- [24] J.-G. Joung, O. Sok June, and B.-T. Zhang, "Prediction of the risk types of human papillomaviruses by support vector machines," in *PRICAI 2004: Trends in Artificial Intelligence*, pp. 723-731, Springer, Berlin, Heidelberg, 2004.
- [25] J.-G. Joung, O. Sok June, and B.-T. Zhang, "Protein sequence-based risk classification for human papillomaviruses," *Computers in Biology and Medicine*, vol. 36, no. 6, pp. 656-667, 2006.
- [26] S. B. Park, S. H. Wang, and B. T. Zhang, "Mining the risk types of human papillomavirus (HPV) by AdaCost," in *Lecture Notes in Computer Science*, pp. 403-412, Springer, Berlin, Heidelberg, 2003.
- [27] M. Esmaili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203-209, 2010.
- [28] M. Alemi, H. Mohabatkar, and M. Behbahani, "In silico comparison of low- and high-risk human papillomavirus proteins," *Applied Biochemistry and Biotechnology*, vol. 172, no. 1, pp. 188-195, 2014.
- [29] K. Tomii and M. Kanehisa, "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins," *Protein Engineering*, vol. 9, no. 1, pp. 27-36, 1996.
- [30] 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.
- [31] M. Bhasin and G. P. S. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *The Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262-23266, 2004.
- [32] Q. Dai, Y. Li, X. Q. Liu, Y. H. Yao, Y. J. Cao, and P. He, "Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position," *BMC Bioinformatics*, vol. 14, no. 1, p. 152, 2013.
- [33] Q. Dai, L. Wu, and L. H. Li, "Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features," *Journal of Computational Chemistry*, vol. 32, no. 16, pp. 3393-3398, 2011.
- [34] J. Cui, L. Han, H. Lin et al., "Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties," *Molecular Immunology*, vol. 44, no. 5, pp. 866-877, 2007.
- [35] L. Y. Han, C. J. Zheng, B. Xie et al., "Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness," *Drug Discovery Today*, vol. 12, no. 7-8, pp. 304-313, 2007.
- [36] Y. L. Chen and Q. Z. Li, "Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 248, no. 2, pp. 377-381, 2007.
- [37] H. B. Shen and K. C. Chou, "Using ensemble classifier to identify membrane protein types," *Amino Acids*, vol. 32, no. 4, pp. 483-488, 2007.
- [38] X. Q. Yu, X. Q. Zheng, T. G. Liu, Y. Dou, and J. Wang, "Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation," *Amino Acids*, vol. 42, no. 5, pp. 1619-1625, 2012.
- [39] Z. P. Feng and C. T. Zhang, "Prediction of membrane protein types based on the hydrophobic index of amino acids," *Journal of Protein Chemistry*, vol. 19, no. 4, pp. 269-275, 2000.
- [40] D. S. Horne, "Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities," *Biopolymers*, vol. 27, no. 3, pp. 451-477, 1988.
- [41] R. R. Sokal and B. A. Thomson, "Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population," *American Journal of Physical Anthropology*, vol. 129, no. 1, pp. 121-131, 2006.
- [42] K. C. Chou, "Prediction of protein subcellular locations by incorporating quasi-sequence-order effect," *Biochemical and Biophysical Research Communications*, vol. 278, no. 2, pp. 477-483, 2000.
- [43] S. L. Zhang, Y. Y. Liang, and X. G. Yuan, "Improving the prediction accuracy of protein structural class: approached with alternating word frequency and normalized Lempel-Ziv complexity," *Journal of Theoretical Biology*, vol. 341, pp. 71-77, 2014.

Research Article

Construction and Comprehensive Analysis of Dysregulated Long Noncoding RNA-Associated Competing Endogenous RNA Network in Moyamoya Disease

Xuefeng Gu ^{1,2}, Dongyang Jiang,³ Yue Yang,^{4,5} Peng Zhang ⁶, Guoqing Wan,^{1,2} Wangxian Gu,² Junfeng Shi,² Liying Jiang,² Bing Chen,⁷ Yanjun Zheng,² Dingsheng Liu ⁸, Sufen Guo ^{4,5} and Changlian Lu ²

¹Research Department, Shanghai University of Medicine & Health Science Affiliated Zhoupu Hospital, Shanghai, China

²Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine & Health Sciences, Shanghai, China

³Department of Cardiology, Pan-Vascular Medicine Institute, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China

⁴Key Laboratory of Cancer Prevention and Treatment of Heilongjiang Province, Mudanjiang Medical University, Mudanjiang, China

⁵Department of Pathology, Hongqi Hospital Affiliated to Mudanjiang Medical University, Mudanjiang, China

⁶School of Clinical Medicine, Shanghai University of Medicine & Health Sciences, Shanghai, China

⁷Department of Neurosurgery, Affiliated Hospital of Guangdong Medical University, Zhanjiang, Guangdong, China

⁸Department of Oncology and Hematology, Shanghai University of Medicine & Health Sciences Affiliated Zhoupu Hospital, China

Correspondence should be addressed to Dingsheng Liu; 13770396508@163.com, Sufen Guo; goldenpot@163.com, and Changlian Lu; lvcl@sumhs.edu.cn

Received 22 March 2020; Accepted 9 May 2020; Published 13 June 2020

Guest Editor: Lei Chen

Copyright © 2020 Xuefeng Gu 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.

Background. Moyamoya disease (MMD) is a rare cerebrovascular disease characterized by chronic progressive stenosis or occlusion of the bilateral internal carotid artery (ICA), the anterior cerebral artery (ACA), and the middle cerebral artery (MCA). MMD is secondary to the formation of an abnormal vascular network at the base of the skull. However, the etiology and pathogenesis of MMD remain poorly understood. **Methods.** A competing endogenous RNA (ceRNA) network was constructed by analyzing sample-matched messenger RNA (mRNA), long non-coding RNA (lncRNA), and microRNA (miRNA) expression profiles from MMD patients and control samples. Then, a protein-protein interaction (PPI) network was constructed to identify crucial genes associated with MMD. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analyses were employed with the DAVID database to investigate the underlying functions of differentially expressed mRNAs (DEmRNAs) involved in the ceRNA network. CMap was used to identify potential small drug molecules. **Results.** A total of 94 miRNAs, 3649 lncRNAs, and 2294 mRNAs were differentially expressed between MMD patients and control samples. A synergistic ceRNA lncRNA-miRNA-mRNA regulatory network was constructed. Core regulatory miRNAs (miR-107 and miR-423-5p) and key mRNAs (STAT5B, FOSL2, CEBPB, and CXCL16) involved in the ceRNA network were identified. GO and KEGG analyses indicated that the DEmRNAs were involved in the regulation of the immune system and inflammation in MMD. Finally, two potential small molecule drugs, CAY-10415 and indirubin, were identified by CMap as candidate drugs for treating MMD. **Conclusions.** The present study used bioinformatics analysis of candidate RNAs to identify a series of clearly altered miRNAs, lncRNAs, and mRNAs involved in MMD. Furthermore, a ceRNA lncRNA-miRNA-mRNA regulatory network was constructed, which provides insights into the novel molecular pathogenesis of MMD, thus giving promising clues for clinical therapy.

1. Introduction

Moyamoya disease (MMD) is a rare cerebrovascular disease characterized by chronic progressive occlusion or stenosis of the bilateral internal carotid artery (ICA), the anterior cerebral artery (ACA), and the middle cerebral artery (MCA) [1, 2]. MMD is secondary to the formation of an abnormal vascular network at the base of the skull. Because the abnormal vascular network of the skull base looks like “smoke” on cerebral angiography images, it is called “moyamoya disease” [3]. The MMD incidence rate in Eastern Asian countries is higher [4], and it mainly occurs in children and young adults, peaking at the ages of 5 to 9 and 35 to 45 years [5]. MMD can seriously affect the mental and physical health of patients. However, the etiology and pathogenesis of MMD remain poorly understood; it may be related to genetics, inflammation, immune response, and environmental factors [6–11].

Many studies have reported that the ring finger protein 213 (RNF213) gene is an important susceptibility gene for MMD in East Asia, especially the p.R4810K variant [12–17]. However, MMD also occurs in patients without mutations in RNF213. To date, new candidate risk-MMD genes, such as the vascular smooth muscle cell-specific isoform of α -actin (ACTA2) [18, 19], endothelial nitric oxide synthase (eNOSase) [20], soluble guanylyl cyclase alpha subunit (GUCY1A3) [21], matrix metalloproteinases (MMPs) [22–26], tissue inhibitor of metalloproteinases (TIMPs) [23, 24], transforming growth factor β 1 (TGF- β 1) [27], Sortilin 1 (SORT1) [28], Connexin 43 (Cx43) [29], and caveolin-1 (Cav-1) [30, 31], have been continuously reported to be associated with MMD.

Moreover, with the development of microarray and sequencing technology, investigators have begun to explore factors other than direct disease-causing genes, including noncoding RNAs (ncRNAs). Gao et al. revealed the expression profile of lncRNAs and mRNAs in MMD patients in 2016 [9], and Dai et al. analyzed miRNAs in the serum of MMD patients and healthy controls in 2012 [32]. miRNAs can posttranscriptionally regulate gene expression by binding to MREs (miRNA-response elements) of their target transcript. mRNAs, lncRNAs, and other RNA transcripts could act as endogenous miRNA sponges to inhibit miRNA function. These interactions illustrate the famous ceRNA hypothesis presented by Salmena in 2011 [33], which gave us a new “language” in different types of RNA transcripts. After that, the ceRNA hypothesis was applied to many fields [34]. The Linc2GO database was constructed by Liu et al. in 2013 [35]. StarBase v2.0 was published by Li et al. to predict miRNA-ceRNA interactions [36]. Moreover, continued analysis of ceRNA networks would deepen our knowledge about how different subtypes of noncoding RNAs work with each other.

In this study, a comprehensive analysis of the miRNA, mRNA, and lncRNA expression profiles in MMD was done, and then, we constructed MMD-specific ceRNA networks using a large cohort from an online database. As far as we know, this is the first study to establish a ceRNA lncRNA-miRNA-mRNA network in MMD, which provides novel

insight into the molecular pathogenesis of MMD, thus giving promising clues for clinical therapy. In addition, core regulatory miRNAs (miR-107 and miR-423-5p) and key mRNAs (STAT5B, FOSL2, CEBPB, and CXCL16) were enriched in immune system/inflammation biological processes, indicating their potential role in MMD.

2. Materials and Methods

2.1. Data Collection. miRNA microarray data were downloaded from Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) in NCBI (The National Center for Biotechnology Information). GEO is an unrestricted open access repository that provides high-throughput microarray and next-generation sequence datasets that have been submitted by researchers around the world. GSE45737 is a miRNA expression profile of the serum from 10 MMD patients and 10 normal healthy controls [32]. The lncRNA and mRNA expression profiles in blood samples from 15 MMD patients and 10 healthy controls were kindly provided by a collaborating academician, Zhao [9].

2.2. Identification of Differentially Expressed RNAs in MMD Patients Compared to Healthy Controls. R software with packages ggplot2, edgeR, and pheatmap (<http://bioconductor.org/bioclite>. R) was adopted to identify differentially expressed RNAs (DERs). In brief, datasets were standardized after conversion of formats, variance normalization, and the addition of missing values as well as statistical testing of differentially expressed probes. The expression levels of all targets, including mRNA, miRNA, and lncRNA, within the datasets were subjected to analysis with R. The threshold was set as a P value < 0.05 and $|\log_2 FC| > 1$. According to these criteria, DERs were identified for further analysis.

2.3. Gene Ontology and Pathway Enrichment Analyses. The Database of Annotation, Visualization and Integrated Discovery (DAVID, <http://david.ncifcrf.gov>) is a public database with comprehensive online tools for functional annotation. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases that contain information about genomes, biological pathways, diseases, and chemical substances [37]. Gene Ontology (GO) is an international standardized gene functional classification system that offers a dynamically updated controlled vocabulary and a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component, and biological process [38].

In the present study, GO and KEGG pathway enrichment analyses were performed using DAVID. $P < 0.05$ was considered statistically significant.

2.4. Construction of the ceRNA (lncRNA-miRNA-mRNA) Regulatory Network. The prediction of miRNA-mRNA interactions was performed on the open-source platform Encyclopedia of RNA Interactomes (ENCORI, <http://starbase.sysu.edu.cn>) [36]. The unique algorithm of ENCORI enables all obtained interactions to be confirmed by at least one other major RNA-RNA prediction website, such as miRanda, PicTar, or TargetScan. In addition to sequence matching, the

prediction was approved by multidimensional sequencing data. All these features make ENCORI a reliable source for predicting RNA-RNA interaction, especially the miRNA-mRNA interaction. Two other databases, miRcode (<http://www.mircode.org>) and DIANA (<http://carolina.imis.athena-innovation.gr>), were applied in the study for predicting miRNA-lncRNA interactions. Afterwards, all interactions were input into Cytoscape (version 3.7.2, <http://cytoscape.org>) to visualize ceRNA regulatory networks. The flow chart can be seen in Figure 1.

2.5. Protein-Protein Interaction (PPI) Network. All DEGs were imported into STRING 10.5, which is a search tool used to identify gene interactions (<https://string-db.org/>). The PPIs were used to construct a network, which was visualized by using Cytoscape software 3.6 (<http://www.cytoscape.org>). The color of edges in the network indicate protein-protein associations: light blue and purple indicate known interactions from curated database and experimentally determined, respectively; dark green/red/dark blue indicate predicted interactions by gene neighborhood/gene fusions/gene co-occurrence, respectively; and light green/black/blue indicate text mining/coexpression/protein homology.

2.6. Gene Expression Signature Analysis with a Connectivity Map. The DEGs were used to perform gene expression signature analysis with connectivity maps (CMap, clue.io). The upregulated and downregulated genes were used as tags, changed into probe IDs referred to Affymetrix U133 GeneChip and uploaded into the CMap database to calculate their values from other drug-target datasets. According to the similarity of gene expression profiles, pairs of gene expression signatures and targeted drugs were used to obtain a value. If the value was a positive number, the target drug would have an effect that was similar to that of the MMD-induced gene expression signature. If the value was a negative number, the targeted drug would have an effect that was opposite that of the MMD-induced gene expression signature; namely, the targeted drug might have an effect that could be useful in treatment.

2.7. Statistical Analysis. We used SPSS 11.0 (SPSS, Chicago, IL) to analyze the dataset from the microarray experiments. All data are represented as the mean \pm SD. Statistical significance was determined at $P < 0.05$.

3. Result and Discussion

3.1. Differentially Expressed mRNAs, miRNAs, and lncRNAs between MMD Patients and Healthy Controls. After differential expression analysis, a total of 2294 DE mRNAs were screened between MMD patients and healthy controls, 865 of which were downregulated and 1429 of which were upregulated in MMD patients. (Table S1, Figure 2(a)). Several genes reported in previous studies in MMD, such as HIF1 α ($\log_2FC = 1.214$), SORT1 ($\log_2FC = 1.628$), and MMP9 ($\log_2FC = 2.40$), are marked in Figure 2. HIF1 α was found to be overexpressed in the intima of the MCA of MMD patients. HIF1 α is a master transcriptional regulator of the adaptive response to hypoxia. Under hypoxic conditions,

HIF1 α translocates to the nucleus, where the HIF1 complex (HIF α /HIF β) binds to the hypoxia-response element and activates the expression of many genes that can increase oxygen delivery and respond to oxygen deprivation in MMD [7]. MMP9 belongs to a family of zinc-binding proteolytic enzymes that are capable of degrading all the components of the extracellular matrix in a variety of physiologic and pathophysiological conditions. Fujimura et al. inferred that the higher expression of MMP9 in MMD patients may play an integrated role in physiologic and pathologic angiogenesis and to the instability of the cerebral vascular structure [39]. SORT1 is another gene reported to be associated with MMD. Increased expression of SORT1 inhibited endothelial cell tube formation and regulated major angiogenic factors and MMP9 expression, implying that SORT1 participated in the pathogenesis of MMD [28].

In addition, 94 DE miRNAs and 3649 DE lncRNAs from GEO datasets were identified. Representative DERs are shown in Figure S1 (a-d).

3.2. Construction of a Competing Endogenous RNA Regulatory Network. The ENCORI database was employed to screen potential interactions between DERs. A synergistic, competitive module of the ceRNA network was constructed separately according to upregulated or downregulated DE mRNAs, which contained 84 nodes in the upregulated group and 66 nodes in the downregulated group. In addition, there were 68 mRNA-miRNA interactions and 16 lncRNA-miRNA interactions in the upregulated group (Figure 3(a)). In the downregulated group, there were 61 interactions between mRNAs and miRNAs and 35 interactions between lncRNAs and miRNAs (Figure 3(b)). The ceRNA network was generated using Cytoscape, as previously discussed.

Based on the network organization, we found that miR-107 competed with 16 mRNAs and 4 lncRNAs (LINC02434, AL589642.1, AC003092.1, and AL035425.3) in the module (Figure 3(b)). A previous study showed that miR-107 is upregulated in response to low-oxygen conditions [40]. Subsequently, miR-107 was found to be abnormally expressed in several cancers, such as PDAC. When miR-107 expression was downregulated in PDAC, cell migration and invasion were inhibited, implying the important role of miR-107 in tumor cell activity [41]. Furthermore, they found that the expression of caveolin-1 was upregulated by a miR-107 inhibitor. Caveolin-1 was reported to be associated with negative remodeling in MMD through the inhibition of angiogenesis in endothelial cells and the induction of apoptosis in VSMCs [30, 31]. Another study by Meng et al. found that miR-107 can inhibit endothelial progenitor cell (EPC) differentiation via HIF1 β [42]. HIF1 β is another subunit of HIF1 that generally heterodimerizes with HIF1 α . Together, they play key roles during hypoxic conditions, which are similar to the conditions in MMD: low oxygen because of vascular occlusion. EPCs can differentiate into mature endothelial cells and play important roles in the recovery of endothelial function and tissue repair. The role of EPCs reflects the mixed state of vascular obstruction and abnormal angiogenesis in the pathogenesis of MMD [43]. The ceRNA network near miR-107 reveals that FoxC1 is one of the potential

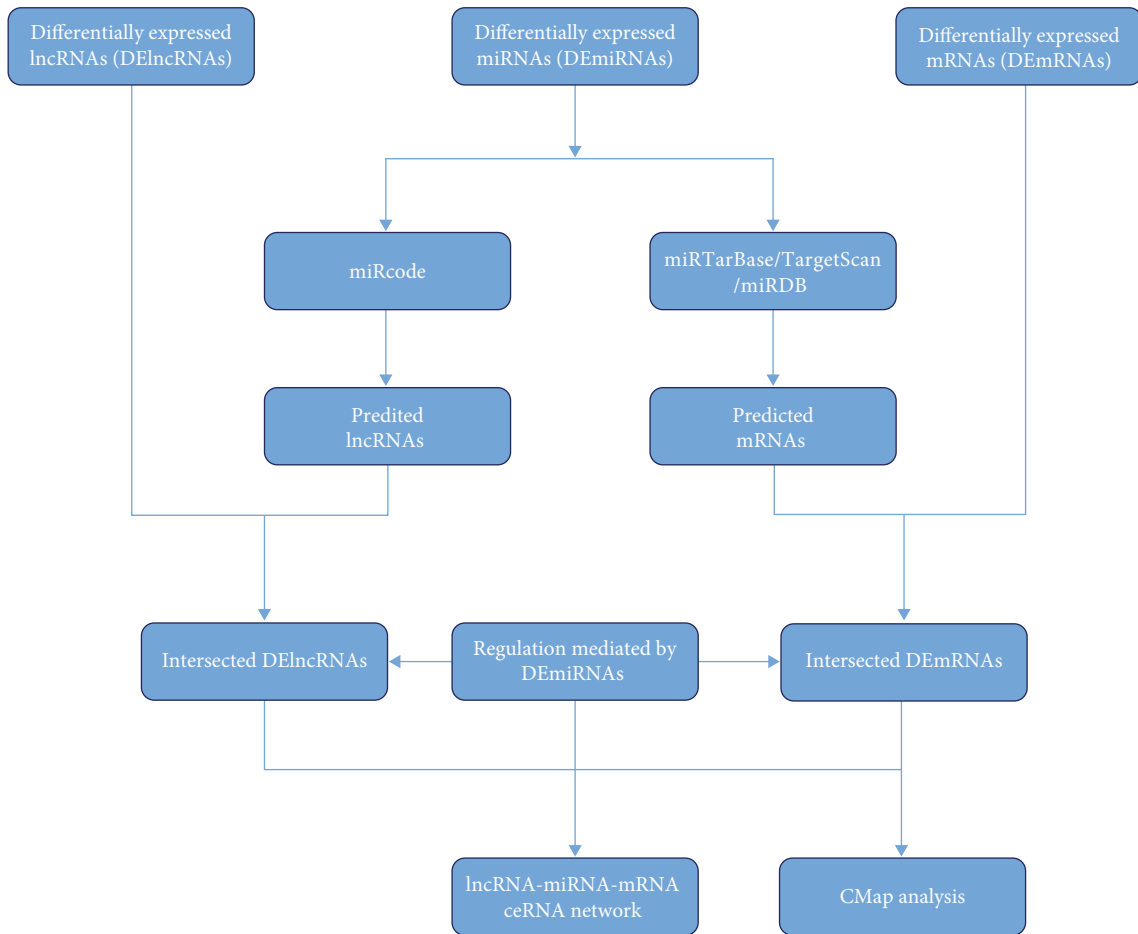


FIGURE 1: Flowchart of the lncRNA-miRNA-mRNA ceRNA network analysis.

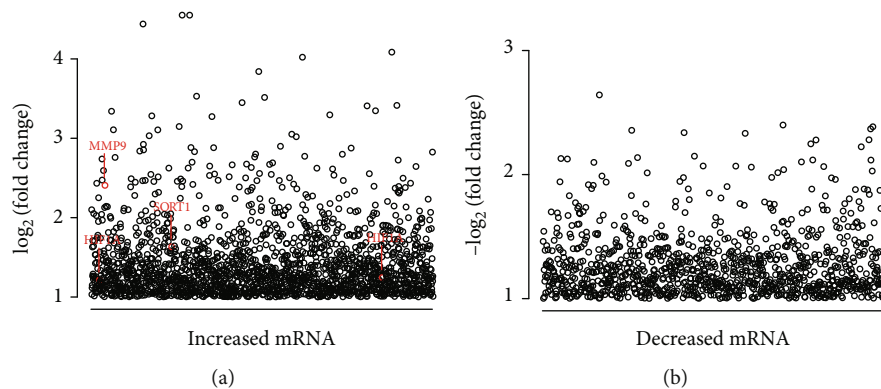
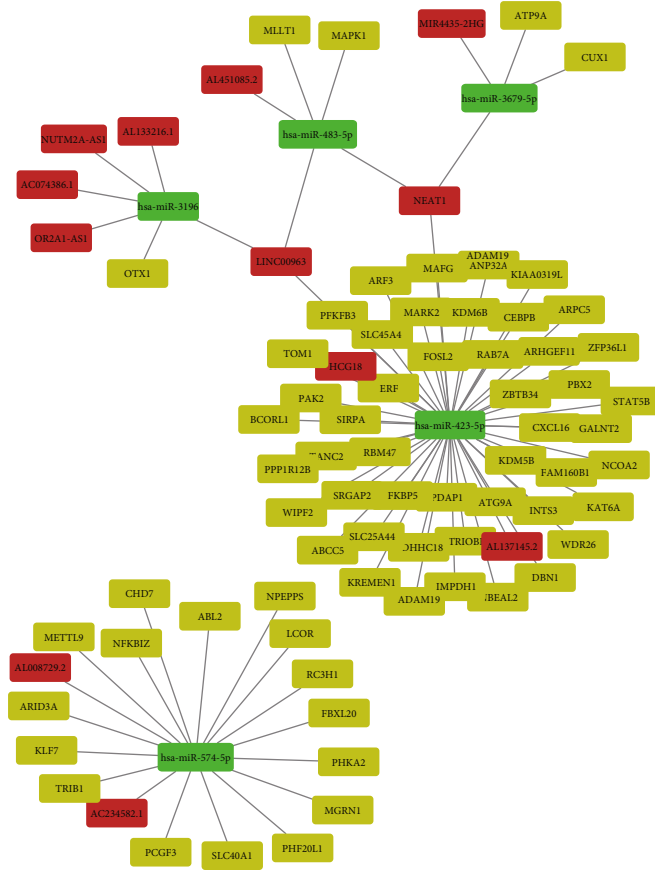


FIGURE 2: Differential mRNA expression between MMD patients and controls. $\log_2FC > 1$ ($P < 0.05$) (MMP9, SORT1, and HIF1 α are marked in red). X-axis shows that the probes of mRNA are arranged in sequence.

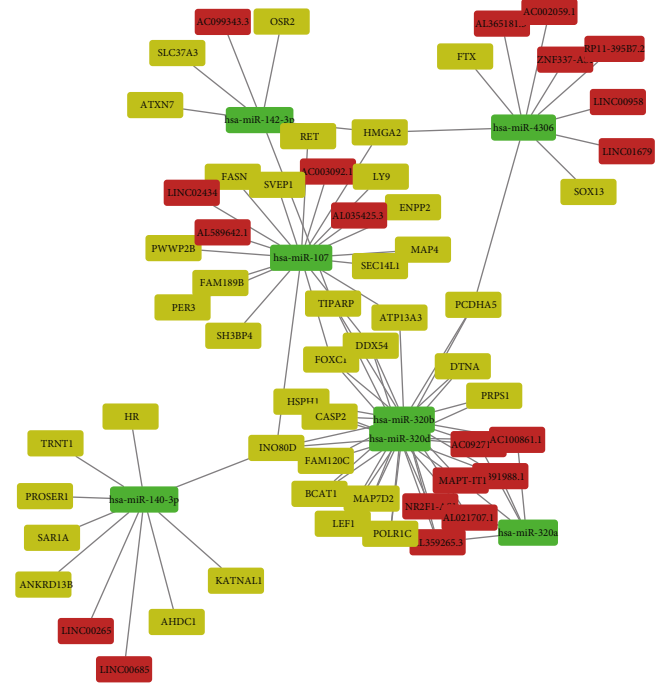
downstream target genes, and it is necessary in the process of vascular development, involving arterial specification and lymphatic sprouting. Abnormal expression of FoxC1 leads to unusual angiogenesis in many tissues [44, 45].

In addition, miR-423-5p competed with 36 mRNAs (CXCL16, FOSL2, etc.) and 4 lncRNAs (NEAT1, HCG18, AL137145.2, and LINC00963) in the module (Figure 3(a)). miR-423-5p was reported to play important roles in the inhibition of the cell proliferation and invasion of cancer cells

such as colon cancer and ovarian carcinoma [46, 47]. Therefore, the downregulation of miR-423-5p in MMD patients may increase the proliferation of vascular smooth muscle cells, which is one likely reason for vessel occlusion. In addition, numerous studies focusing on NEAT1's role in cancer biology found that this lncRNA plays a crucial role in carcinogenesis [48]. NEAT1 mainly works as a ceRNA by sponging antitumor miRNAs [49]. NEAT1 is also involved in immune system responses, viral diseases, and



(a)



(b)

FIGURE 3: The lncRNA-miRNA-mRNA ceRNA network in MMD. (a) ceRNA network based on upregulated mRNAs involved ceRNA. (b). ceRNA network based on downregulated mRNAs involved ceRNA. Notes: red rectangles represent DElncRNAs, green rectangles represent DEMiRNAs, and yellow rectangles represent DEMRNAs.

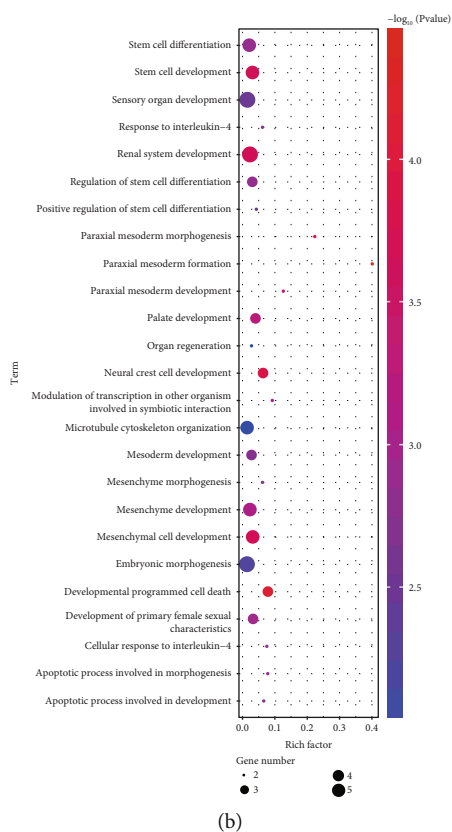
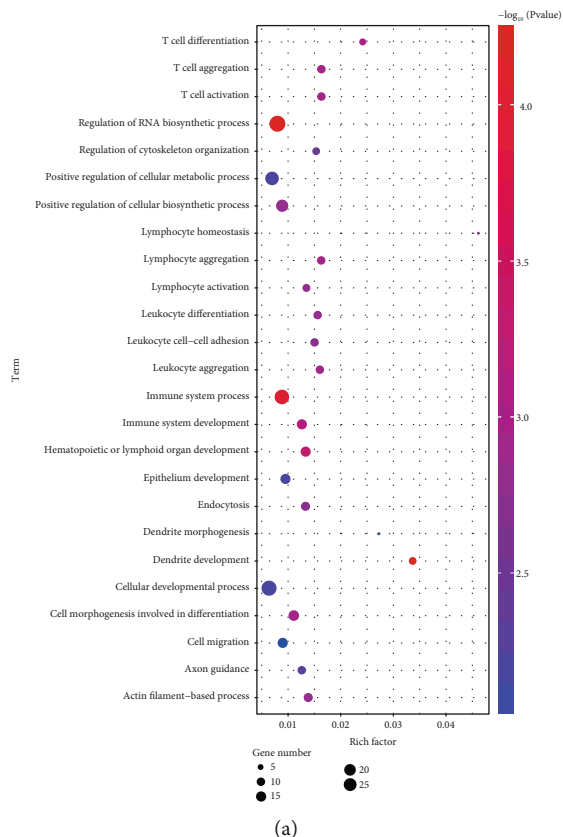


FIGURE 4: Enriched gene ontology terms for biological process based on the DEMRNAs involved in the ceRNA network.

TABLE 1: KEGG pathway enrichment of all DEmRNAs involved in the ceRNA network.

(a) KEGG pathway enrichment of decreased mRNA involved in the ceRNA network

KEGG ID	KEGG term	P value	Symbols
05216	Thyroid cancer	0.0012614	RET, LEF1
00230	Purine metabolism	0.0353649	PRPS1, POLR1C
04141	Protein processing in the endoplasmic reticulum	0.0365787	HSPH1, SAR1A

(b) KEGG pathway enrichment of increased mRNA involved in the ceRNA network

KEGG ID	KEGG term	P value	Symbols
04012	ErbB signaling pathway	0.000117	ABL2, PAK2, MAPK1, STAT5B
04270	Vascular smooth muscle contraction	0.004943	PPP1R12B, MAPK1, ARHGEF11
04380	Osteoclast differentiation	0.006506	FOSL2, MAPK1, SIRPA
04360	Axon guidance	0.006648	PAK2, MAPK1, SRGAP2
05221	Acute myeloid leukemia	0.012833	MAPK1, STAT5B
05131	Shigellosis	0.014609	MAPK1, ARPC5
04062	Chemokine signaling pathway	0.018775	MAPK1, STAT5B, CXCL16
05211	Renal cell carcinoma	0.018968	PAK2, MAPK1
05220	Chronic myeloid leukemia	0.020529	MAPK1, STAT5B
04810	Regulation of actin cytoskeleton	0.025711	PAK2, MAPK1, ARPC5
04666	Fc gamma R-mediated phagocytosis	0.032876	MAPK1, ARPC5
04660	T cell receptor signaling pathway	0.042378	PAK2, MAPK1

neurodegeneration disorders [50]. To study FOSL2, also named Fra 2, Maurer et al. created Fra 2 knockout mice and found that the mice developed pulmonary arterial occlusion due to vascular SMC proliferation and inflammation and pulmonary fibrosis [51, 52]. All of the above results imply that the ceRNA lncRNA-miRNA-mRNA regulatory network we constructed provides many new clues regarding MMD pathogenesis.

3.3. Functional Annotation of the mRNAs Involved in the ceRNA Network. After the ceRNA network was established with the help of the DAVID database, functional annotation and pathway analysis of this small group of DEmRNAs were performed to identify potential candidate pathways or biological processes related to MMD.

As shown in Figure 4, some of pathways require our attention, and processes related to the immune response and inflammatory reaction, including immune system process, T cell aggregation, T cell activation, lymphocyte aggregation, and lymphocyte activation, were significantly enriched. Additionally, another enrichment also occurred in biological processes associated with cell development and differentiation, including paraxial mesoderm development and mesenchymal cell differentiation; these results suggest important roles for these biological activities in MMD. The Kyoto Encyclopedia of Genes and Genomes showed that DEmRNAs were enriched in chemokine signaling, ErbB signaling, axon guidance, and vascular smooth muscle contraction (Table 1).

Recently, many studies have shown that immunological/inflammatory factors are involved in the occurrence and

development of MMD. According to IHC staining, there were T cells and macrophages infiltrating in the stenosed and thickened vascular intima of MMD patients [53]. The abnormal deposition of IgG in the elastic layer of the ICA and MCA suggests that the infiltration of immune cells and the damage to the immune functions are related to MMD [54]. Moreover, the overexpression of inflammatory factors in MMD patients, such as MCP-1, IL-1 β , and SDF-1 α , suggests that inflammation may also affect the progression of MMD [55]. Consistently, in this study, several mRNAs that encode critical inflammatory molecules, such as chemokines and cytokines, were dysregulated and were determined to be DEmRNAs in MMD patients. Nevertheless, although varied mRNAs were clearly enriched in terms of GO analysis, there were few found in the ceRNA network. However, several important genes involved in the regulation of inflammation in MMD were modulated by ceRNAs. CXCL16 is considered to be an important pathogenic mediator of atherosclerosis (clinical severity is graded according to the severity of carotid stenosis) [56]. CXCL16 is a vascular-derived factor that induces angiogenesis [57]. CXCL16 also exists in a soluble form and interacts with its specific chemokine receptor, CXCR6, to recruit the migration of activated T cells into the inflammatory tissue [58]. As shown in Figure 3(b), four potential lncRNAs, including LINC00963, NEAT1, HCG18, and AL137145.2, could act as ceRNAs to regulate CXCL16 through miR-107. The work on this interesting ceRNA network remains to be done in the future.

3.4. Protein-Protein Interaction (PPI) Network. As shown in Figure 5, a PPI network for DEmRNA-involved ceRNA

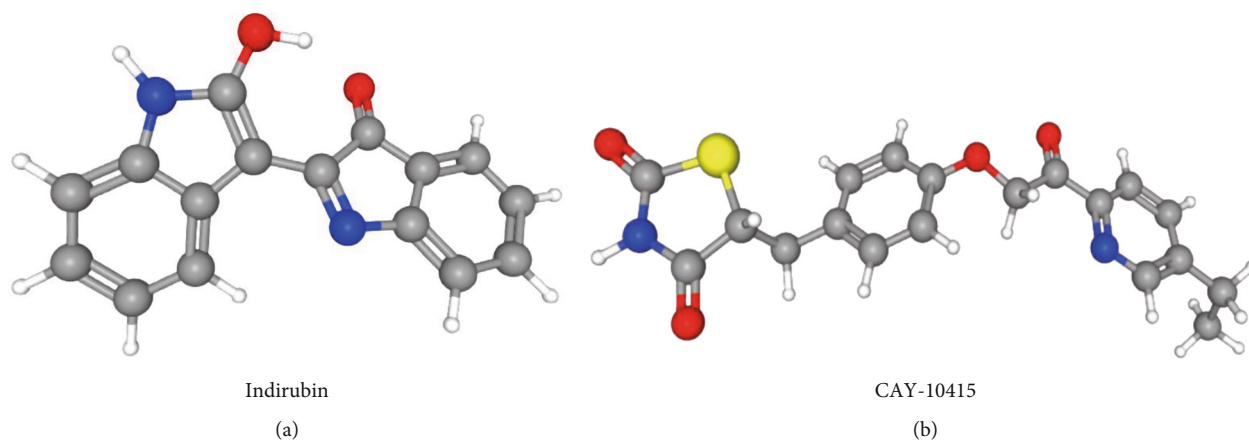


FIGURE 6: Potential molecular drugs. (a) Indirubin. (b) CAY-10415.

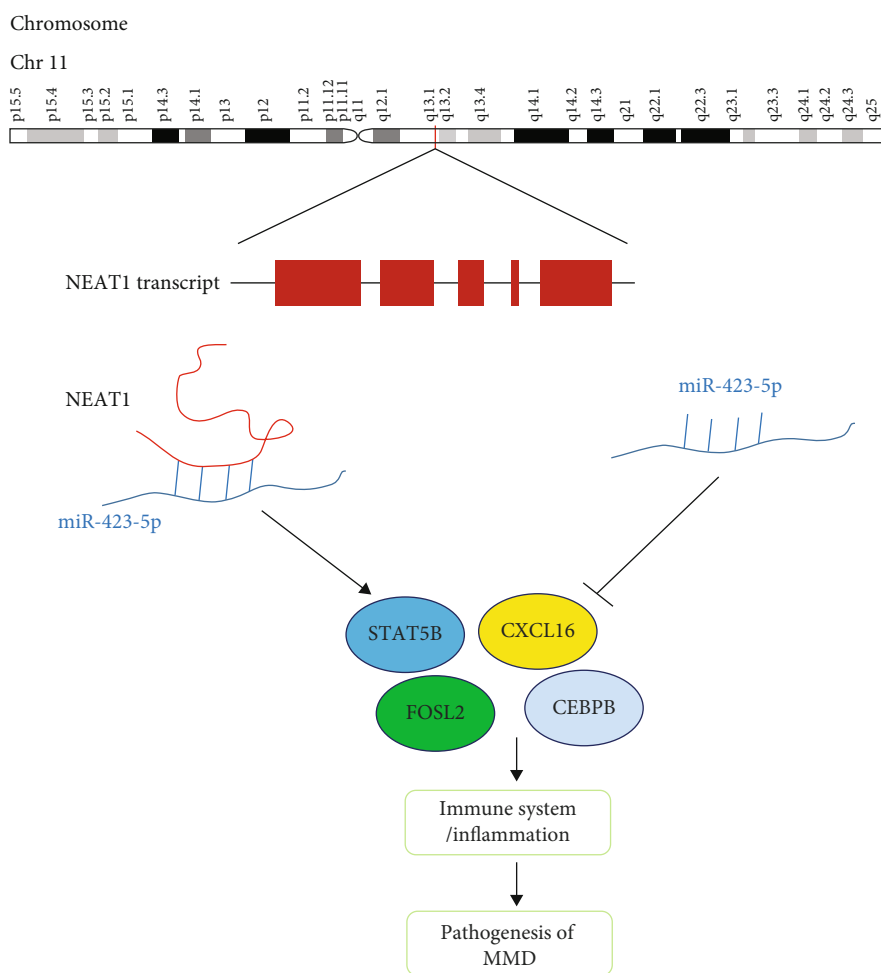


FIGURE 7: The potential mechanism of a DElncRNA sponging miR-423-5p.

networks was constructed by Cytoscape software. It is important to highlight that some striking genes, such as MAK1, STAT5B, CEBPB, FOSL2, PAK2, and ABL2, play vital key roles in MMD. These interesting genes were also shown in Table 1, such as ABL2, PAK2, MAPK1, and STAT5B were enriched in the ErbB signaling pathway. After the identifica-

tion of the overlap between the above genes, chemokine signaling, T cell receptor signaling, and ErbB signaling shed some light on the pathogenesis of MMD.

3.5. *Potential Small Molecule Drugs.* All the DE mRNAs involved in the ceRNA regulatory network in MMD were

analyzed by CMap to identify small molecule drugs. Strong negative correlations were found between MMD and enzasaurin, cyproheptadine, flupirtine, indirubin, and mitoglitazone (CAY-10415); strong positive correlations were found between MMD and flavokavain-b, CGS-20625, vinburnine, apicidin, and cytochalasin-d (Table S5). The drugs that had a strong negative correlation with the pathogenesis of MMD might have therapeutic effects on MMD (Table 2). CAY-10415 and indirubin gained our attention. The structures of the two potential molecular drugs were investigated using the PubChem database (Figure 6). CAY-10415 is a member of a new class of compounds that modulate mitochondrial pyruvate carrier (MPC), a key controller of cellular metabolism that influences mTOR activation [59]. It is commonly known that CAY-10415 can be used as an insulin sensitizer, and it can play this role without activating PPAR α . Therefore, CAY-10415 can avoid negative side effects observed in currently used insulin sensitizers, such as pioglitazone and rosiglitazone. CAY-10415 has been used in Alzheimer's disease patients [60]. It is generally accepted that insulin sensitizers can not only improve diabetes but also improve blood lipid disorders, reduce the level of free fatty acids in plasma, reduce the effect of fat toxicity, and indirectly protect the function of β cells [61]. By inhibiting the proliferation and migration of vascular smooth muscle cells and reducing the intima-media thickness of arteries, it can play a protective role in the intima. Likewise, indirubin, a red isomer of indigo, is the active ingredient of the traditional Chinese drug *Danggui Longhui Wan*, which was used for the treatment of chronic myelocytic leukemia (CML) [62]. Enzyme-based *in vitro* studies have observed that indirubin and its derivatives, such as indirubin-3'-monoxime, indirubin-5-sulfonate, and indirubin-3'-monoxime-5-sulphonic acid, are potential inhibitors of CDKs [63]. Furthermore, different indirubin derivatives showed antiangiogenesis activity by blocking VSMC proliferation and endothelial cell function through the inhibition of the STAT signaling pathway and reduction of neointima formation *in vivo* [64]. All of the above findings suggest that CAY-10415 and indirubin may be used in MMD patients to avoid vascular aberration and occlusion.

4. Conclusions

In summary, using bioinformatics analysis of candidate RNAs, the present study identified a series of clearly altered lncRNAs, miRNAs, and mRNAs involved in MMD. Furthermore, a ceRNA lncRNA-miRNA-mRNA regulatory network was constructed, which provides a novel insight into the molecular pathogenesis of MMD, thus giving promising clues for clinical therapy. In addition, core regulatory miRNAs (miR-107 and miR-423-5p) and key mRNAs (STAT5B, FOSL2, CEBPB, and CXCL16) were enriched in immune system/inflammation biological processes, indicating their potential role in MMD (Figure 7). In the future, more attention should be paid to the validation of competing endogenous RNA interactions with experimental techniques.

Finally, two potential small molecule drugs, CAY-10415 and indirubin, were identified by CMap to be candidate drugs for treating MMD.

Data Availability

(1) The miRNA microarray data used to support the findings of this study have been deposited in the Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) in NCBI (GSE45737). (2) The lncRNA and mRNA microarray data included in this study are available upon request by contact with the corresponding author. The data were kindly provided by a collaborating Prof. Zhao.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Xuefeng Gu, Dongyang Jiang and Yue Yang contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81772829, 81272376, and 81830052), the Key projects for collaborative innovation of Shanghai University of Medicine & Health Sciences, Construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400), the Shanghai Municipal Education Commission (Class II Plateau Disciplinary Construction Program for Medical Technology of SUMHS, 2018-2020), the Natural Science Foundation of Heilongjiang Province (LH2019H119), the Natural Science Foundation of Guangdong Province (2016A030313680), Academic Leader Training Program of Pudong New District Health Bureau of Shanghai (PWRd2015-09), and the Funding Scheme for Training Young Teachers in Shanghai Colleges (ZZJKYX19009).

Supplementary Materials

Table S1: DE mRNA list in excel file. Table S2: DELncRNA list in excel file. Table S3: increased miRNA list in excel file. Table S4: decreased miRNA list in excel file. Table S5: drugs in excel file. Figure S1: DELncRNA and DemiRNA in tiff format. (*Supplementary Materials*)

References

- [1] T. Kondo, "Moyamoya disease," *Canadian Medical Association Journal*, vol. 190, no. 46, article E1364, 2018.
- [2] S. Shang, D. Zhou, J. Ya et al., "Progress in moyamoya disease," *Neurosurgical Review*, vol. 43, no. 2, pp. 371–382, 2020.
- [3] M. Fujimura, O. Y. Bang, and J. S. Kim, "Frontiers of Neurology and Neuroscience," in *Moyamoya disease*, vol. 40, pp. 204–220, 2016.
- [4] J. S. Kim, "Moyamoya disease: epidemiology, clinical features, and diagnosis," *Journal of Stroke*, vol. 18, no. 1, pp. 2–11, 2016.

- [5] X. Y. Bao, Q. N. Wang, Y. Zhang et al., "Epidemiology of moyamoya disease in China: single-center, population-based study," *World Neurosurgery*, vol. 122, pp. e917–e923, 2019.
- [6] S. Newman, J. H. Boulter, J. G. Malcolm, I. Pradilla, and G. Pradilla, "Outcomes in patients with moyamoya syndrome and sickle cell disease: a systematic review," *World Neurosurgery*, vol. 135, pp. 165–170, 2020.
- [7] Q. Ma, L. Li, B. Yu et al., "Circular RNA profiling of neutrophil transcriptome provides insights into asymptomatic moyamoya disease," *Brain Research*, vol. 1719, pp. 104–112, 2019.
- [8] M. Zhao, F. Gao, D. Zhang et al., "Altered expression of circular RNAs in moyamoya disease," *Journal of the Neurological Sciences*, vol. 381, pp. 25–31, 2017.
- [9] F. Gao, L. Yu, D. Zhang, Y. Zhang, R. Wang, and J. Zhao, "Long noncoding RNAs and their regulatory network: potential therapeutic targets for adult moyamoya disease," *World Neurosurgery*, vol. 93, pp. 111–119, 2016.
- [10] J. Yu, J. Zhang, J. Li, J. Zhang, and J. Chen, "Cerebral hyperperfusion syndrome after revascularization surgery in patients with moyamoya disease: systematic review and meta-analysis," *World Neurosurgery*, vol. 135, pp. 357–366.e4, 2020.
- [11] O. Y. Bang, M. Fujimura, and S. K. Kim, "The pathophysiology of moyamoya disease: an update," *Journal of Stroke*, vol. 18, no. 1, pp. 12–20, 2016.
- [12] F. Kamada, Y. Aoki, A. Narisawa et al., "A genome-wide association study identifies RNF213 as the first moyamoya disease gene," *Journal of Human Genetics*, vol. 56, no. 1, pp. 34–40, 2011.
- [13] W. Liu, D. Morito, S. Takashima et al., "Identification of RNF213 as a susceptibility gene for moyamoya disease and its possible role in vascular development," *PLoS One*, vol. 6, no. 7, article e22542, 2011.
- [14] S. Miyatake, N. Miyake, H. Touho et al., "Homozygous c.14576G>A variant of RNF213 predicts early-onset and severe form of moyamoya disease," *Neurology*, vol. 78, no. 11, pp. 803–810, 2012.
- [15] E. H. Kim, M. S. Yum, Y. S. Ra et al., "Importance of RNF213 polymorphism on clinical features and long-term outcome in moyamoya disease," *Journal of Neurosurgery*, vol. 124, no. 5, pp. 1221–1227, 2016.
- [16] W. Liu, T. Hitomi, H. Kobayashi, K. O. U. J. I. H. HARADA, and A. Koizumi, "Distribution of moyamoya disease susceptibility polymorphism p.R4810K in RNF213 in east and south-east Asian populations," *Neurologia medico-chirurgica*, vol. 52, no. 5, pp. 299–303, 2012.
- [17] W. Liu, H. Hashikata, K. Inoue et al., "A rare Asian founder polymorphism of raptor may explain the high prevalence of moyamoya disease among east Asians and its low prevalence among Caucasians," *Environmental Health and Preventive Medicine*, vol. 15, no. 2, pp. 94–104, 2010.
- [18] D. C. Guo, C. L. Papke, V. Tran-Fadulu et al., "Mutations in smooth muscle alpha-actin (ACTA2) cause coronary artery disease, stroke, and Moyamoya disease, along with thoracic aortic disease," *The American Journal of Human Genetics*, vol. 84, no. 5, pp. 617–627, 2009.
- [19] C. Roder, V. Peters, H. Kasuya et al., "Analysis of ACTA2 in European moyamoya disease patients," *European Journal of Paediatric Neurology*, vol. 15, no. 2, pp. 117–122, 2011.
- [20] Y. S. Park, K. T. Min, T. G. Kim et al., "Age-specific eNOS polymorphisms in moyamoya disease," *Child's Nervous System*, vol. 27, no. 11, pp. 1919–1926, 2011.
- [21] S. Wallace, D. C. Guo, E. Regalado et al., "Disrupted nitric oxide signaling due to GUCY1A3 mutations increases risk for moyamoya disease, achalasia and hypertension," *Clinical Genetics*, vol. 90, no. 4, pp. 351–360, 2016.
- [22] H. Li, Z. S. Zhang, W. Liu et al., "Association of a functional polymorphism in the MMP-3 gene with moyamoya disease in the Chinese Han population," *Cerebrovascular Diseases*, vol. 30, no. 6, pp. 618–625, 2010.
- [23] X. Wang, Z. Zhang, W. Liu et al., "Impacts and interactions of PDGFRB, MMP-3, TIMP-2, and RNF213 polymorphisms on the risk of moyamoya disease in Han Chinese human subjects," *Gene*, vol. 526, no. 2, pp. 437–442, 2013.
- [24] Y. S. Park, Y. J. Jeon, H. S. Kim et al., "The GC + CC genotype at position -418 in TIMP-2 promoter and the -1575GA/-1306CC genotype in MMP-2 is genetic predisposing factors for prevalence of moyamoya disease," *BMC Neurology*, vol. 14, no. 1, 2014.
- [25] H. S. Kang, J. H. Kim, J. H. Phi et al., "Plasma matrix metalloproteinases, cytokines and angiogenic factors in moyamoya disease," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 81, no. 6, pp. 673–678, 2010.
- [26] S. Sonobe, M. Fujimura, K. Niizuma et al., "Increased vascular MMP-9 in mice lacking RNF213: moyamoya disease susceptibility gene," *NeuroReport*, vol. 25, no. 18, pp. 1442–1446, 2014.
- [27] C. Roder, V. Peters, H. Kasuya et al., "Polymorphisms in TGFB1 and PDGFRB are associated with moyamoya disease in European patients," *Acta Neurochirurgica*, vol. 152, no. 12, pp. 2153–2160, 2010.
- [28] H. Y. Sung, J. Y. Lee, A. K. Park et al., "Aberrant promoter hypomethylation of Sortilin 1: a moyamoya disease biomarker," *Journal of Stroke*, vol. 20, no. 3, pp. 350–361, 2018.
- [29] J. Liao, T. Hong, J. Xu, E. Zeng, B. Tang, and W. Lai, "Expression of Connexin43 in cerebral arteries of patients with moyamoya disease," *Journal of Stroke and Cerebrovascular Diseases*, vol. 27, no. 4, pp. 1107–1114, 2018.
- [30] O. Y. Bang, J. W. Chung, S. J. Kim et al., "Caveolin-1, Ring finger protein 213, and endothelial function in Moyamoya disease," *International Journal of Stroke*, vol. 11, no. 9, pp. 999–1008, 2016.
- [31] J. W. Chung, D. H. Kim, M. J. Oh et al., "Cav-1 (Caveolin-1) and arterial remodeling in adult moyamoya disease," *Stroke*, vol. 49, no. 11, pp. 2597–2604, 2018.
- [32] D. Dai, Q. Lu, Q. Huang et al., "Serum miRNA signature in moyamoya disease," *PLoS One*, vol. 9, no. 8, article e102382, 2014.
- [33] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?," *Cell*, vol. 146, no. 3, pp. 353–358, 2011.
- [34] D. S. Sardina, S. Alaimo, A. Ferro, A. Pulvirenti, and R. Giugno, "A novel computational method for inferring competing endogenous interactions," *Briefings in Bioinformatics*, vol. 18, no. 6, pp. 1071–1081, 2017.
- [35] K. Liu, Z. Yan, Y. Li, and Z. Sun, "Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis," *Bioinformatics*, vol. 29, no. 17, pp. 2221–2222, 2013.
- [36] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "star-Base v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data," *Nucleic Acids Research*, vol. 42, no. D1, pp. D92–D97, 2013.

- [37] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.
- [38] R. Huntley, E. Dimmer, D. Barrell, D. Binns, and R. Apweiler, "The gene ontology annotation (GOA) database," *Nature Precedings*, 2009, <https://doi.org/10.1038/npre.2009.3154.1>.
- [39] M. Fujimura, M. Watanabe, A. Narisawa, H. Shimizu, and T. Tominaga, "Increased expression of serum matrix metalloproteinase-9 in patients with moyamoya disease," *Surgical Neurology*, vol. 72, no. 5, pp. 476–480, 2009.
- [40] R. Kulshreshtha, M. Ferracin, S. E. Wojcik et al., "A microRNA signature of hypoxia," *Molecular and Cellular Biology*, vol. 27, no. 5, pp. 1859–1867, 2007.
- [41] J. Xiong, D. Wang, A. Wei et al., "Deregulated expression of miR-107 inhibits metastasis of PDAC through inhibition PI3K/Akt signaling via caveolin-1 and PTEN," *Experimental Cell Research*, vol. 361, no. 2, pp. 316–323, 2017.
- [42] S. Meng, J. Cao, L. Wang et al., "MicroRNA 107 partly inhibits endothelial progenitor cells differentiation via HIF-1 β ," *PLoS One*, vol. 7, no. 7, article e40323, 2012.
- [43] K. H. Jung, K. Chu, S. T. Lee et al., "Circulating endothelial progenitor cells as a pathogenetic marker of moyamoya disease," *Journal of Cerebral Blood Flow and Metabolism*, vol. 28, no. 11, pp. 1795–1803, 2008.
- [44] S. Seo, H. P. Singh, P. M. Lactal et al., "Forkhead box transcription factor FoxC1 preserves corneal transparency by regulating vascular growth," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 6, pp. 2015–2020, 2012.
- [45] S. Seo, H. Fujita, A. Nakano, M. Kang, A. Duarte, and T. Kume, "The forkhead transcription factors, Foxc1 and Foxc2, are required for arterial specification and lymphatic sprouting during vascular development," *Developmental Biology*, vol. 294, no. 2, pp. 458–470, 2006.
- [46] W. Jia, T. Yu, Q. An, X. Cao, and H. Pan, "MicroRNA-423-5p inhibits colon cancer growth by promoting caspase-dependent apoptosis," *Experimental and Therapeutic Medicine*, vol. 16, no. 2, pp. 1225–1231, 2018.
- [47] X. Tang, X. Zeng, Y. Huang et al., "miR-423-5p serves as a diagnostic indicator and inhibits the proliferation and invasion of ovarian cancer," *Experimental and Therapeutic Medicine*, vol. 15, no. 6, pp. 4723–4730, 2018.
- [48] C. Klec, F. Prinz, and M. Pichler, "Involvement of the long noncoding RNA NEAT1 in carcinogenesis," *Molecular Oncology*, vol. 13, no. 1, pp. 46–60, 2019.
- [49] K. Zhou, C. Zhang, H. Yao et al., "Knockdown of long non-coding RNA NEAT1 inhibits glioma cell migration and invasion via modulation of SOX2 targeted by miR-132," *Molecular Cancer*, vol. 17, no. 1, p. 105, 2018.
- [50] F. Prinz, A. Kapeller, M. Pichler, and C. Klec, "The implications of the long non-coding RNA NEAT1 in non-cancerous diseases," *International Journal of Molecular Sciences*, vol. 20, no. 3, 2019.
- [51] Y. Asano, "Recent advances in animal models of systemic sclerosis," *The Journal of Dermatology*, vol. 43, no. 1, pp. 19–28, 2016.
- [52] B. Maurer, N. Busch, A. Jüngel et al., "Transcription factor fos-related antigen-2 induces progressive peripheral vasculopathy in mice closely resembling human systemic sclerosis," *Circulation*, vol. 120, no. 23, pp. 2367–2376, 2009.
- [53] J. Masuda, J. Ogata, and C. Yutani, "Smooth muscle cell proliferation and localization of macrophages and T cells in the occlusive intracranial major arteries in moyamoya disease," *Stroke*, vol. 24, no. 12, pp. 1960–1967, 1993.
- [54] R. Lin, Z. Xie, J. Zhang et al., "Clinical and immunopathological features of moyamoya disease," *PLoS One*, vol. 7, no. 4, article e36386, 2012.
- [55] G. Ni, W. Liu, X. Huang et al., "Increased levels of circulating SDF-1 α and CD34⁺ CXCR4⁺ cells in patients with moyamoya disease," *European Journal of Neurology*, vol. 18, no. 11, pp. 1304–1309, 2011.
- [56] J. W. Shi, H. L. Yang, D. X. Fan et al., "The role of CXC chemokine ligand 16 in physiological and pathological pregnancies," *American Journal of Reproductive Immunology*, vol. 83, no. 4, article e13223, 2020.
- [57] X. Yu, R. Zhao, S. Lin et al., "CXCL16 induces angiogenesis in autocrine signaling pathway involving hypoxia-inducible factor 1 α in human umbilical vein endothelial cells," *Oncology Reports*, vol. 35, no. 3, pp. 1557–1565, 2016.
- [58] A. Ma, X. Pan, Y. Xing, M. Wu, Y. Wang, and C. Ma, "Elevation of serum CXCL16 level correlates well with atherosclerotic ischemic stroke," *Archives of Medical Science*, vol. 10, no. 1, pp. 47–52, 2014.
- [59] A. Ghosh, T. Tyson, S. George et al., "Mitochondrial pyruvate carrier regulates autophagy, inflammation, and neurodegeneration in experimental models of Parkinson's disease," *Science Translational Medicine*, vol. 8, no. 368, article 368ra174, 2016.
- [60] R. Shah, D. Matthews, R. Andrews et al., "An evaluation of MSDC-0160, a prototype mTOT modulating insulin sensitizer, in patients with mild Alzheimer's disease," *Current Alzheimer Research*, vol. 11, no. 6, pp. 564–573, 2014.
- [61] N. Rohatgi, H. Aly, C. A. Marshall et al., "Novel insulin sensitizer modulates nutrient sensing pathways and maintains β -Cell phenotype in human islets," *PLoS One*, vol. 8, no. 5, article e62012, 2013.
- [62] J. L. Lai, Y. H. Liu, C. Liu et al., "Indirubin inhibits LPS-induced inflammation via TLR4 abrogation mediated by the NF- κ B and MAPK signaling pathways," *Inflammation*, vol. 40, no. 1, pp. 1–12, 2017.
- [63] A. V. Schwaiberger, E. H. Heiss, M. Cabaravdic et al., "Indirubin-3'-monoxime blocks vascular smooth muscle cell proliferation by inhibition of signal transducer and activator of transcription 3 signaling and reduces neointima formation in vivo," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 30, no. 12, pp. 2475–2481, 2010.
- [64] T. Blažević, A. V. Schwaiberger, C. E. Schreiner et al., "12/15-lipoxygenase contributes to platelet-derived growth factor-induced activation of signal transducer and activator of transcription 3," *The Journal of Biological Chemistry*, vol. 288, no. 49, pp. 35592–35603, 2013.

Research Article

Influences of Daily Life Habits on Risk Factors of Stroke Based on Decision Tree and Correlation Matrix

Zeguo Shao,^{1,2} Yuhong Xiang,¹ Yingchao Zhu,³ Aiqin Fan,⁴ and Peng Zhang^{5,6} 

¹School of Medical Instrumentation, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China

²Center for Intelligent Medical Electronics (CIME), Fudan University, Shanghai 201318, China

³Nursing Department, Shanghai Pudong New District Zhoupu Hospital, Shanghai 201318, China

⁴Pudong New Area Lingqiao Community Health Service Center, Shanghai 200137, China

⁵School of Clinical Medicine, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China

⁶Shanghai General Practice Medical Education and Research Center, Shanghai 201318, China

Correspondence should be addressed to Peng Zhang; zhangp@sumhs.edu.cn

Received 13 March 2020; Accepted 29 April 2020; Published 1 June 2020

Guest Editor: Lei Chen

Copyright © 2020 Zeguo Shao 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.

Purpose. To explore the influences of smoking, alcohol consumption, drinking tea, diet, sleep, and exercise on the risk of stroke and relationships among the factors, present corresponding knowledge-based rules, and provide a scientific basis for assessment and intervention of risk factors of stroke. **Methods.** The decision tree C4.5 algorithm was optimized and utilized to establish a model for stroke risk assessment; then, the main risk factors of stroke (including hypertension, dyslipidemia, diabetes, atrial fibrillation, body mass index (BMI), history of stroke, family history of stroke, and transient ischemic attack (TIA)) and daily habits (e.g., smoking, alcohol consumption, drinking tea, diet, sleep, and exercise) were analyzed; corresponding knowledge-based rules were finally presented. Establish a correlation matrix of stroke risk factors and analyze the relationship between stroke risk factors. **Results.** The accuracy of the established model for stroke risk assessment was 87.53%, and the kappa coefficient was 0.8344, which was superior to that of the random forest and Logistic algorithm. Additionally, 37 knowledge-based rules that can be used for prevention of risk factors of stroke were derived and verified. According to in-depth analysis of risk factors of stroke, the values of smoking, exercise, sleep, drinking tea, alcohol consumption, and diet were 6.00, 7.00, 8.67, 9.33, 10.00, 10.60, and 10.75, respectively, indicating that their influence on risk factors of stroke was reduced in turn; on the one hand, smoking and exercise were strongly associated with other risk factors of stroke; on the other hand, sleep, drinking tea, alcohol consumption, and diet were not firmly associated with other risk factors of stroke, and they were relatively tightly associated with smoking and exercise. **Conclusions.** Establishment of a model for stroke risk assessment, analysis of factors influencing risk factors of stroke, analysis of relationships among those factors, and derivation of knowledge-based rules are helpful for prevention and treatment of stroke.

1. Introduction

Stroke is an acute cerebrovascular disease, associating with the characteristics of high morbidity, high disability, and high mortality. It is a refractory disease that imposes a major threat to human health and life [1]. At present, there are no effective treatments for stroke. Prevention is still the most feasible strategy to reduce the harm of stroke and reduce its social burden, especially with respect to high global incidence and potential risk factors of stroke [2]. The risk factors of

stroke are divided into intervention factors (e.g., smoking, alcohol consumption, and body mass index (BMI)) and non-intervention factors (e.g., age, gender, ethnicity, and genetic attributes) according to whether the risk can be changed through intervention [3]. Hence, studying the intervention factors is of great significance for the prevention of stroke. In addition, we previously found that the interventional risk factors for stroke appeared more in people's daily lives and behavioral habits [4, 5]. Unhealthy lifestyles can trigger or increase the risk of stroke, and moderate lifestyle changes

TABLE 1: Subjects' clinical data.

Type of data	Risk factor of stroke	Field	Data distribution
Clinical diagnosis	Hypertension	Hyte	y: 1242, n: 3782, uncertain: 575
	Dyslipidemia	Dysl	y: 511, n: 4508, uncertain: 580
	Diabetes	Diab	y: 403, n: 4618, uncertain: 578
	Atrial fibrillation	AF	y: 75, n: 4940, uncertain: 584
Medical history and family history	Family history of stroke	FSH	y: 449, n: 4460, uncertain: 690
	History of stroke	SH	y: 165, n: 4730, uncertain: 704
	TIA	TIA	y: 95, n: 4350, uncertain: 1154
Demographic information	Gender	Gen	M: 2491, F: 3108
	Age	Age	Refer to Figure 1
Physical examination	BMI	BMIc	B1: 205, B2: 2926, B3: 1760, B4: 520, B5: 150, uncertain: 38
Daily habits	Smoking	Smok	y: 1192, n: 4379, null: 28
	Alcohol consumption	Alco	y: 1065, n: 4500, null: 34
	Drinking tea	Tea	y: 1563, n: 3997, null: 39
	Diet	DT	C1: 2812, C2: 263, C3: 2181, null: 370
	Sleep	Sleep	TS: 366, TB: 4958, BL: 205, null: 70
	Exercise sport	Sport	C1: 1518, C2: 1624, C3: 2275, null: 182

“y” means “yes,” “n” indicates “no,” and definitions of the types of BMI, diet, sleep, and exercise are presented in Figure 1. In Figure 1, we sometimes use fields to represent their corresponding stroke risk factors.

may reduce the risk of stroke as well [6]. Therefore, numerous scholars suggested that further studies should be carried out to provide effective interventions to guide and improve people’s lifestyle, so as to reduce the risk and incidence of stroke [7–9]. However, in 2019, Altobelli et al. analyzed the relevant literature and found that research in this area was conducted in only a limited number of developed countries, and there were very few reports on the impact of lifestyle and dietary habits on risk factors of stroke [10]. In China, Huang et al. conducted relevant research and demonstrated that a healthy lifestyle (high fruit intake, quitting smoking, doing housework, and good sleep quality) may reduce the chance of recurrence of first-onset ischemic stroke [11]. Although the risk factors of stroke in daily life habits are not the main risk factors of stroke, they are closely associated with the main risk factors [12].

The present study was aimed at the Chinese population, and large-scale and multidimensional stroke data were collected through modern information technology. The optimized decision tree algorithm was used to analyze risk factors of stroke in daily life habits, derive knowledge-based rules, and establish a model for stroke risk assessment to analyze relationships among risk factors of stroke.

2. Materials and Methods

2.1. Data Collection and Pretreatment. We established a whole-course stroke management network system via collection of large-scale data from Shanghai suburban population, involving nearly 10,000 people, in which 5599 valid data were finally acquired. The data included subjects’ demographic characteristics, physical examination, family medical history, treatment history, personal diet and lifestyle habits,

sleep and breathing, psychological status, quality of life, and stroke knowledge. In order to facilitate classification of stroke, we also designed a rapid stroke screening form and performed statistical analysis. We preliminarily extracted and integrated data and determined 16 risk factors of stroke for further analysis. As shown in Table 1, among 5599 data collected, there were 2491 males and 3108 females, subjects’ minimum and maximum age were 18 and 89 years old, respectively. The age- and gender-based data are shown in Figure 1.

As illustrated in Figure 1, [18,30] indicates that age is 18 years old or older and less than 30 years old; F and M denote female and male, respectively; and PN is the number of individuals.

The present research analyzed the risk factors of smoking, alcohol consumption, drinking tea, diet, sleep, sport, and BMI. The above-mentioned factors were defined as follows:

- (i) Smoking: those who have smoked for 6 months or more in their lifetime were marked as “y”; otherwise, they are denoted as “n”
- (ii) Alcohol consumption: those who have drunk no less than twice/week and no less than 80 ml each time were marked as “y”; otherwise, they were denoted as “n”
- (iii) Drinking tea: those who have drunk tea at least 3 days/week were marked as “y”; otherwise, they were denoted as “n”
- (iv) Diet: the daily food ingredients are mainly sugars, fats, or proteins, which were marked with “C1,” “C2,” and “C3,” respectively

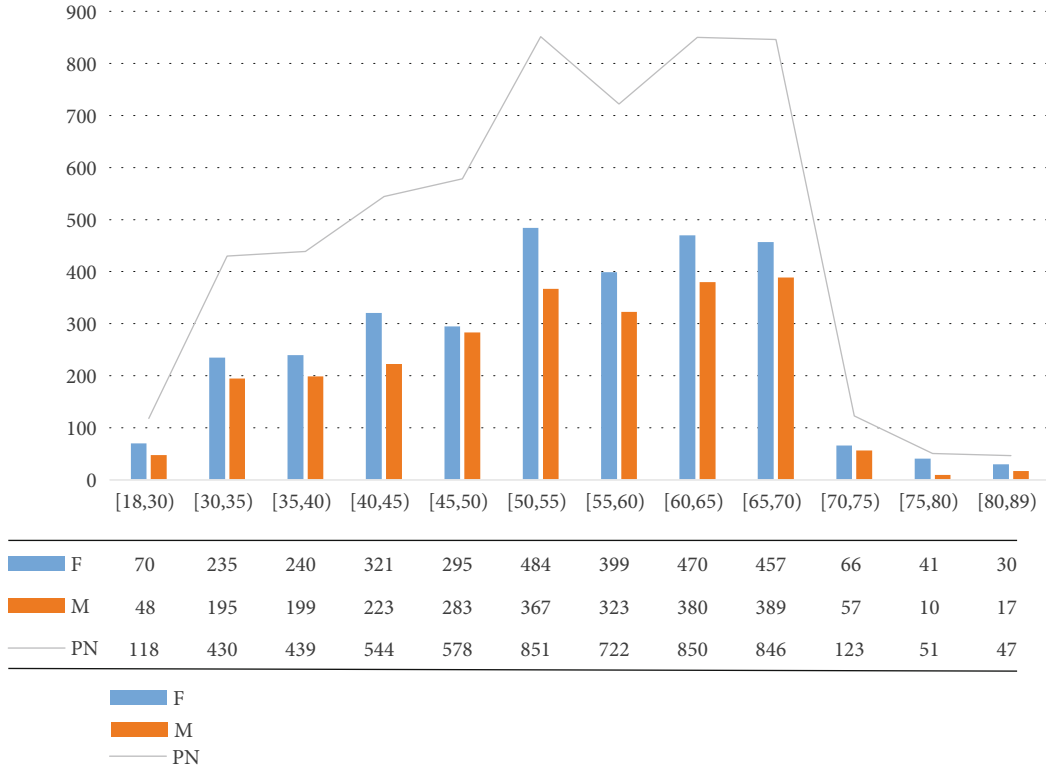


FIGURE 1: Distribution of age- and gender-based data.

- (v) Sport: those who have exercised sport more than 3 times/week and more than 30 min each time, demonstrating regular level of sport, marked as “C1”; those who have exercised sport 2-3 times/week, and 10-30 min each time, reflecting medium level of sport, marked as “C2”; those who have exercised less than or equal to 1 time/week and less than 10 min each time, indicating lower level of sport, marked “C3”
- (vi) BMI: since the WHO standards are not highly appropriate for Chinese people, the Chinese Reference Standards were formulated with reference to the WHO standards and are divided into five types: B1, B2, B3, B4, and B5 (Table 2)
- (vii) Sleep: duration of sleep in different ages can be divided into three types: very short-term, medium-term, and very long-term, which could be labelled as TS, TB, and TL, respectively, as shown in Figure 2

According to the rapid screening of risk factors of stroke (including hypertension, dyslipidemia, diabetes, atrial fibrillation, smoking history, BMI, sport, stroke history, family history of stroke, and transient ischemic attack (TIA)), refer to the Guidelines for Screening, Prevention and Control of Ischemic Stroke presented by the Ministry of Health of China (hereinafter referred to the guidelines), this study classified stroke risk into H, M, L, N, T, and Y levels, as summarized in Table 3.

2.2. *Decision Trees.* The decision tree is a popular, logic-based, easily interpretable, straightforward, and widely applicable method [13]. The classic decision tree algorithms include ID3, C4.5, and CART. In contrast to ID3, which can only handle discrete variables, C4.5 and CART can handle continuous variables, and they are not sensitive to incomplete data. In addition, the CART generates binary trees and the C4.5 algorithm generates multiple branches. Decision trees can generate interpretable knowledge rules, which can express relationship between factors. This is in line with our goal to explore relationships among the risk factors of stroke. Therefore, the C4.5 algorithm was selected in the current research. Details of the C4.5 algorithm were described in the following.

2.2.1. *C4.5 Algorithm.* In 1992, Ross Quinlan developed the C4.5 decision tree algorithm [14]. C4.5 constructs a decision tree as a learning model from the data samples. The divide-and-conquer approach is adopted for construction of decision tree models using a measure called information gain to select the attribute from the dataset for the tree.

(1) *Information Gain.* Suppose that there are C categories of data in the sample dataset D . The information entropy formula is as follows:

$$\text{Info}(D) = - \sum_{i=1}^c p_i \times \log_2(p_i), \quad (1)$$

TABLE 2: Sleep classification.

Age	Duration of sleep (hours)	Mark
<3 (months)	<14	TS
	14~17	TB
	>17	TL
1~2 (years old)	<11	TS
	11~14	TB
	>14	TL
6~13 (years old)	<9	TS
	9~11	TB
	>11	TL
14~17 (years old)	<8	TS
	8~10	TB
	<10	TL
18~64 (years old)	<6	TS
	6~10	TB
	<10	TL
>64 (years old)	<7	TS
	7~8	TB
	<8	TL

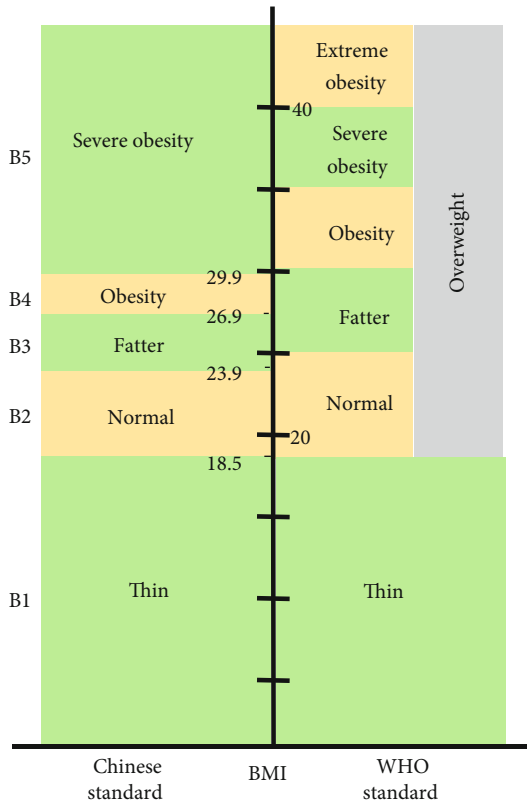


FIGURE 2: BMI classification.

where D represents the training dataset, C denotes the data class number, and p_i represents the ratio of the sample number in class i to all samples. When the attribute A is chosen as the node of the decision tree, the information entropy after the action of feature A is as follows:

$$\text{Info}_A(D) = - \sum_{j=1}^k \frac{|D_j|}{|D|} \times \text{Info}(D_j), \quad (2)$$

where k represents the data samples D divided into k parts.

(2) *Gain Ratio*. The information gain represents the value of the information entropy that the dataset D decreases after the action of the feature A . The formula is as follows:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D). \quad (3)$$

The information gain ratio is given by

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{Info}_A(D)}. \quad (4)$$

2.2.2. Improvement and Implementation of C4.5 Algorithm.

We used a decision tree algorithm to analyze the above-mentioned 16 risk factors of stroke (see Table 1). The decision tree is generated using the J48 (C4.5 algorithm implementation) in the Weka classifier algorithm. The confidence factor for the pruning is set to 0.25, and the minimum number of instances per leaf (minNumObj) is set to 1. The 10-fold cross-validation is additionally used to select and evaluate the model.

In order to solve imbalanced data problem and improve the robustness of the system, we, in the current study, presented SMOTE algorithm to improve the model. The SMOTE algorithm is an intelligent oversampling technique for unbalanced datasets proposed by Chawla et al. in 2002. It can effectively improve the overfitting phenomenon caused by traditional oversampling techniques and solve the problem of biased classification results. As illustrated in Figure 3, after classified dataset is preprocessed for equilibrium judgment, the number of records in each class is first counted to find out the maximum value (max) and minimum value (min) of the number of records and then quotient max and min, if $\text{max}/\text{min} < 3$. After the dataset is judged to be balanced, it is directly entered into the C4.5 classifier for classification. Otherwise, it is judged that the dataset is unbalanced and is entered into the SMOTE processor: first, the entire dataset is sampled, the sampling method is nonrepeatable sampling, the number is equal to the number of datasets, each record is randomly sorted, and then, SMOTE is used to generate new minority data. The effects of operations, such as filtering and sorting preprocessing on the SMOTE algorithm, are eliminated to ensure that the data obtained by SMOTE is obtained by randomly combining the major data and the minor data to avoid overfitting caused by the data generated by SMOTE only from the minor data. Then, the data are entered into the classification module.

TABLE 3: Definition of different levels of risk factors of stroke.

Type	Definition
Y	Have a history of stroke.
T	Has a previous transient ischemic attack.
H	The major risk factors defined in the guidelines are 2 items or more, or the major risk factors include 1 item, and the secondary risk factors involve 2 items or more.
M	The major risk factors defined in the guidelines include 1 item, and the secondary risk factors involve less than 2 items.
L	The main risk factors defined in the guidelines include 0 item, and the secondary risk factors involve 2 items or more.
N	The main risk factors defined in the guidelines include 0 item, and the secondary risk factors involve less than 2 items.

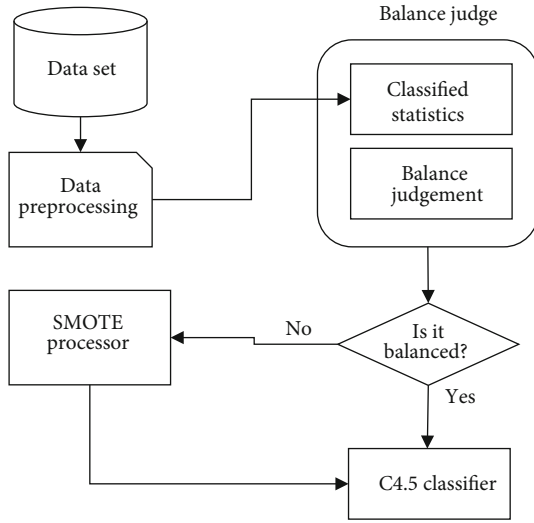


FIGURE 3: SMOTE+C4.5 classification model.

3. Results

The number of leaves of the tree was 98, while the size of the tree was 171 (Figures 4–8). The performance indexes of the tree are as follows: classification accuracy: 87.5281%; kappa statistic: 0.8344; mean absolute error: 0.0567; and root-mean-square error: 0.175.

To assess the performance of the proposed system for stroke risk classification, precision, recall, accuracy, and kappa were calculated, and 10-fold cross-validation was used. Equations (5)–(8) were presented to calculate precision, recall, accuracy, and kappa, respectively.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (6)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (7)$$

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e}. \quad (8)$$

Precision represents the correct positive prediction ratio to the whole positive samples. Recall is the correct positive prediction ratio to the whole positive predictions. Accuracy

is correct prediction ratio to the whole predictions. True positives (TPs) are positive cases that are correctly predicted as positive. False negatives (FNs) are positive cases that are incorrectly predicted as negative. True negatives (TNs) are negative cases that are correctly predicted as negative. False positives (FPs) are negative cases that are incorrectly predicted as positive. Meanwhile, kappa offers a more robust estimated performance of the proposed system compared with a simple agreement and gives an overall evaluation of all the cases. p_o is the relative observed agreement among the proposed system and the physician analysis, and p_e is the hypothetical probability of chance agreement.

Table 4 presents the confusion matrix of the classification result using optimized C4.5 algorithm. In order to evaluate the performance of the optimized C4.5 algorithm, the random forest and Logistic algorithm were implemented for making comparison. Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [15]. Logistic regression is a generalized linear regression analysis model, commonly used in data mining, automatic disease diagnosis, economic prediction, and other fields. The Logistic regression is good at analyzing linear relationships, and analyzing nonlinear relationships is worse than decision trees. In addition, it is sensitive to extreme values and easily affected by extreme values, and the decision tree performs better in this respect [16].

In the current study, the number of trees in the random forest was set to 100, and for each tree, the minimum number of instances for each leaf was set to 1. The Ridge value in the Logistic was set to $1.0E - 8$, and the maximum number of iterations to perform was set to -1. They all use tenfold cross-validation like decision trees. Tables 5 and 6 summarize the confusion matrix of classification results using random forest and Logistic algorithm, respectively.

Regardless of accuracy or kappa value, the optimized C4.5 is the highest among the three algorithms. The recall of the risk type “T” could achieve only 0.208 using the random forest algorithm, which was noticeably lower than 0.962 using the C4.5 algorithm. Figures 9–11 demonstrate that misclassification rate of risk type “T” is the lowest in optimized C4.5 algorithm among the three algorithms.

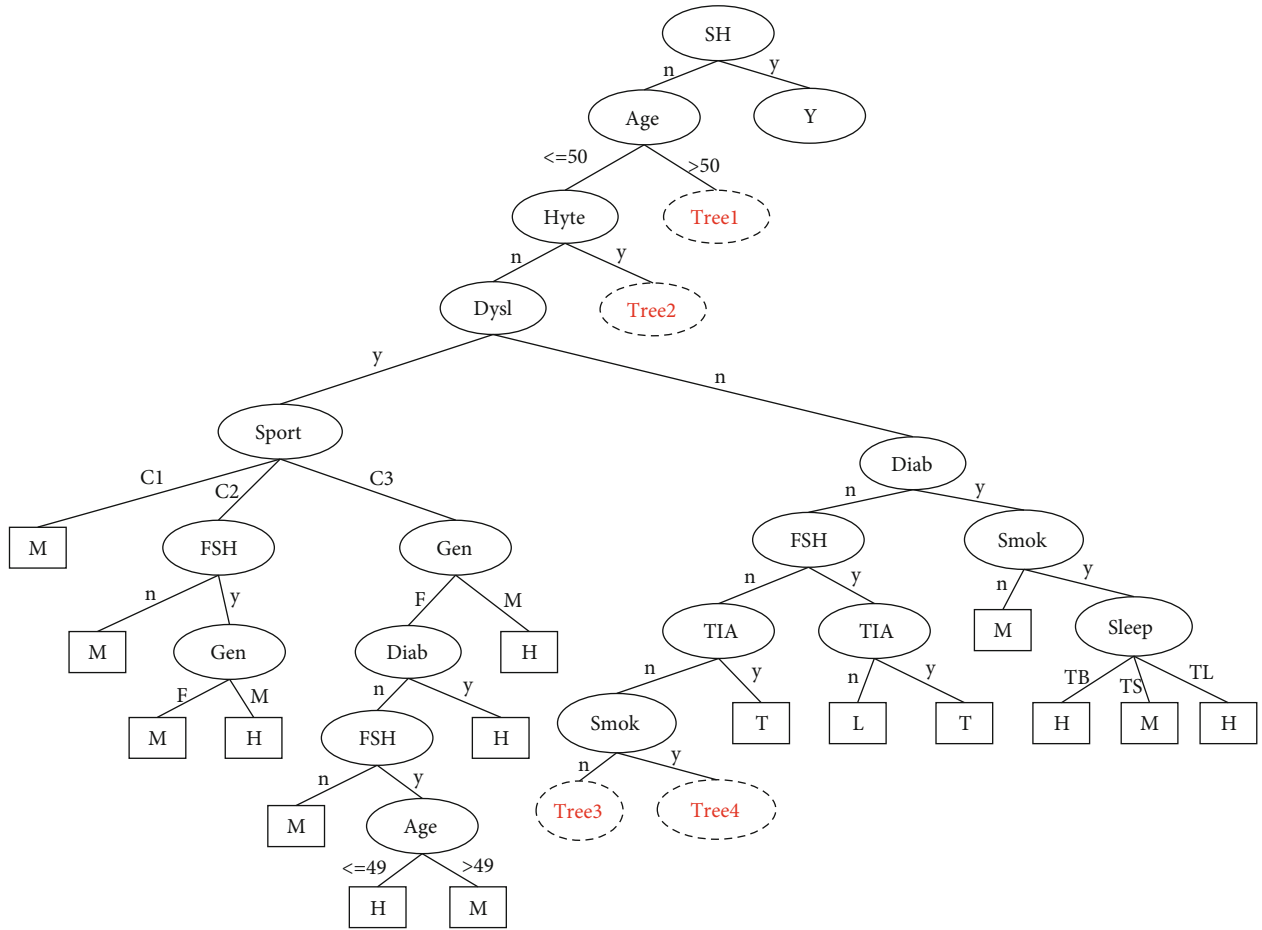


FIGURE 4: A decision tree to classify risk factors of stroke.

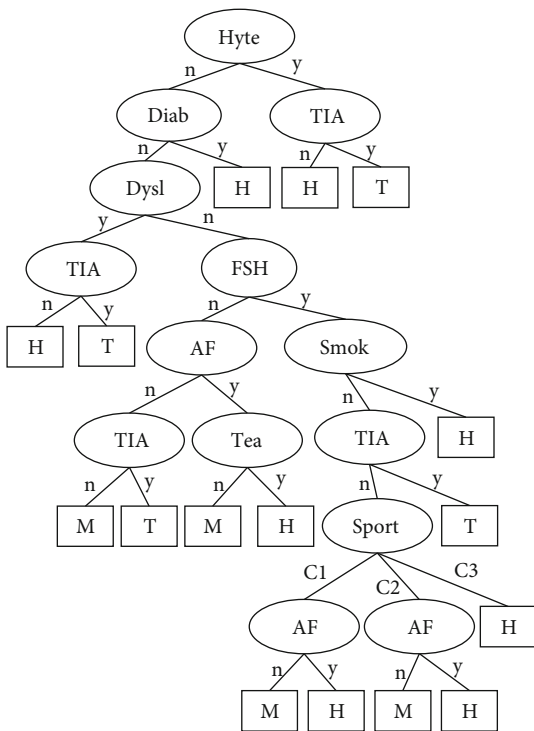


FIGURE 5: Decision tree #1 to classify risk factors of stroke.

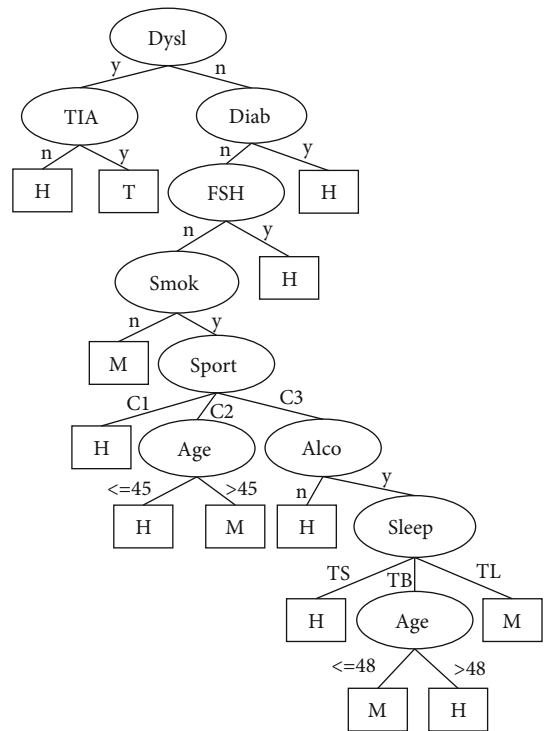


FIGURE 6: Decision tree #2 to classify risk factors of stroke.

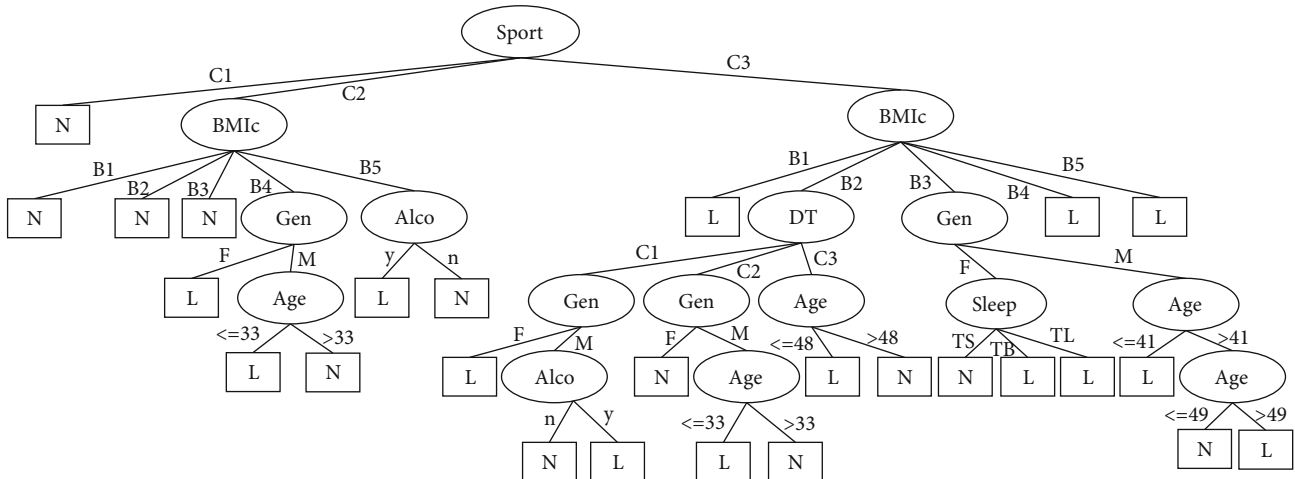


FIGURE 7: Decision tree #3 to classify risk factors of stroke.

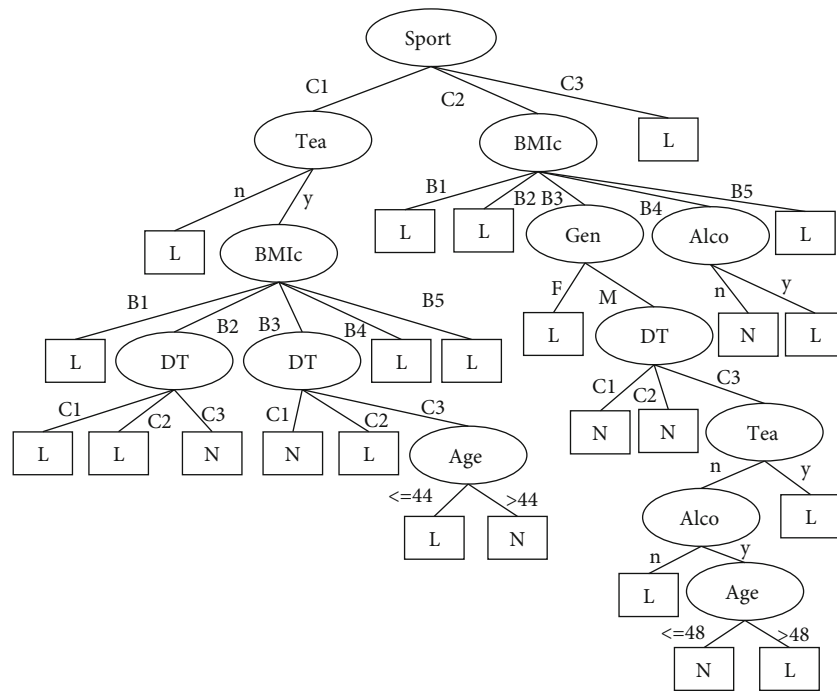


FIGURE 8: Decision tree #4 to classify risk factors of stroke.

Corresponding knowledge-based rules can be deduced from the decision tree. There were 98 knowledge-based rules deduced from the present case. There are 37 rules related to the 6 daily living habits (smoking, alcohol consumption, drinking tea, diet, sleep, and sport), which are illustrated in the Supplementary Information (available here).

4. Discussion

According to the previous decision tree, the average depth and frequency of each risk factor in the decision tree were calculated, as shown in Table 7. Values of risk factors for stroke (stroke history, hypertension, dyslipidemia, diabetes,

family history of stroke, TIA, smoking, atrial fibrillation, exercise, sleep, gender, BMI, drinking tea, age, and alcohol consumption) were increased, indicating that their influence on risk factors of stroke was relatively reduced. Simultaneously, the impact of daily living habits on risk factors of stroke was relatively insignificant, demonstrating that the influence of lifestyle habits and diet on risk factors of stroke is indirect.

We further analyzed the above-mentioned 98 knowledge-based rules for risk factors of stroke, in which risk factors were extracted from the knowledge-based rules. Within each set, the sum of the reciprocals of factors was used to represent the weight of each factor. All factor sets

TABLE 4: Confusion matrix achieved by the optimized C4.5 algorithm.

	Risk level analyzed by optimized C4.5 algorithm						Recall
	H	M	Y	T	N	L	
Risk level analyzed by physicians							
H	1288	127	0	0	0	0	0.910
M	44	1502	0	0	0	0	0.972
Y	0	0	165	0	0	0	1.000
T	2	0	0	51	0	0	0.962
N	0	0	0	0	679	255	0.727
L	0	0	0	0	182	596	0.766
Precision	0.966	0.922	1.000	1.000	0.789	0.700	
Accuracy				87.53%			
Kappa				0.8344			

TABLE 5: Confusion matrix achieved by the random forest algorithm.

	Risk level analyzed by random forest algorithm						Recall
	H	M	Y	T	N	L	
Risk level analyzed by physicians							
H	1300	115	0	0	0	0	0.919
M	72	1473	0	0	1	0	0.953
Y	6	0	158	0	0	1	0.958
T	24	6	0	11	3	9	0.208
N	0	0	0	0	699	235	0.748
L	0	0	0	0	239	539	0.693
Precision	0.927	0.924	1.000	1.000	0.742	0.688	
Accuracy				85.46%			
Kappa				0.8063			

TABLE 6: Confusion matrix achieved by the Logistic algorithm.

	Risk level analyzed by Logistic						Recall
	H	M	Y	T	N	L	
Risk level analyzed by physicians							
H	1289	124	0	1	0	1	0.911
M	97	1446	1	1	1	0	0.935
Y	0	0	164	0	0	1	0.994
T	5	0	1	46	1	0	0.868
N	0	0	0	0	690	244	0.739
L	0	1	0	0	214	563	0.724
Precision	0.927	0.920	0.988	0.958	0.762	0.696	
Accuracy				85.83%			
Kappa				0.8119			

and their weights will be described in the Supplementary Information. Within each set, every two factors formed a factor pair; the same factor pairs were weighted and summed together to form a factor-based relationship matrix, as shown in Table 8.

As illustrated in Table 8, it was unveiled that the risk factors of stroke, such as stroke history (SH), hypertension

(Hyte), dyslipidemia (Dysl), diabetes (Diab), and age (Age), have the highest correlation. Of the 6 daily habit factors we examined (smoking, alcohol consumption, tea, diet, sleep, and exercise), only the correlation of smoking (Smok) and sport (Sport) was higher than the average (1.95). This indicates that alcohol consumption, drinking tea, diet, and sleep are not strongly correlated with other factors. In addition,

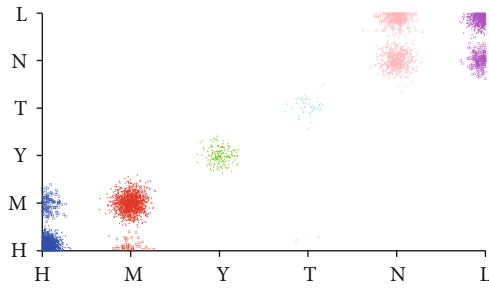


FIGURE 9: Illustration of errors of the optimized C4.5 algorithm.

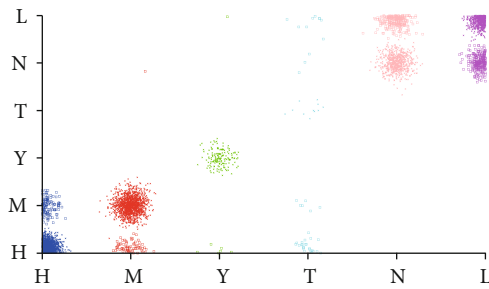


FIGURE 10: Illustration of errors of the random forest algorithm.

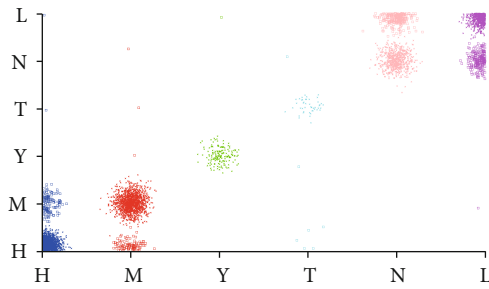


FIGURE 11: Illustration of errors of the Logistic algorithm.

regarding this weak correlation, the correlation values of alcohol consumption, drinking tea, diet, sleep, smoking, and sport were close to those of strong correlation categories (SH, HYTE, Dysl, Diab, and Age), as shown in Table 9.

4.1. Smoking and Sport. Of the 37 knowledge-based rules mentioned above, 30 rules included a “smoking” factor, suggesting that smoking significantly increases the risk factors of stroke. Yamagishi et al. demonstrated that smoking increases the risk of stroke in patients with hypertension [17], which is in line with our findings. In addition, the radar chart of the risk ratio of smoking to nonsmoking is also illustrated by Figure 12(a).

Of the 37 knowledge-based rules mentioned above, 35 contained “sport.” As displayed in Figure 12(b), there is no significant difference in the impact of high-intensity and medium-intensity exercise on risk factors of stroke. Exercise is the most common factor affecting the risk of stroke, and moderate exercise helps prevent stroke, which is consistent with the results of McDonnell et al.’s study [18].

Additionally, 28 knowledge-based rules contained both “smoking” and “sport” factors, indicating that smoking and sport are closely associated together, and further, doing exercise by smokers is beneficial to reduce the risk of stroke.

4.2. Alcohol Consumption and Drinking Tea. It was noted that individuals who drink alcohol have a significantly higher risk of stroke than nonalcohol consumers (Figure 12(c)). This is in line with Hu et al.’s outcome that heavy drinking can increase the risk of stroke, while moderate drinking has insignificant influence on the risk of stroke [19]. However, it is not an independent factor and is typically associated with hypertension, diabetes, and hypercholesterolemia.

Knowledge-based rules showed that drinking tea has no direct effect on the risk of stroke (Figure 12(d)), and similar to alcohol consumption, it can be related to BMI. Sosa et al. demonstrated that tea is highly beneficial to reduce the risk of stroke in obese people [20]. Zhang et al. conducted experiments on mice and concluded that drinking tea has a neuroprotective effect on hemorrhagic stroke [21]. In addition, we found that “tea=y” and “alco=y” do not simultaneously appear in the same rule in the present study, and the correlation value of 0.14 (Table 8) between them is also very insignificant, indicating that drinking tea and alcohol consumption have simultaneously no effect on the risk of stroke.

4.3. Diet. As shown in Figure 12(e), the effects of the three types of diet (mainly sugar, fat, and protein) on risk of stroke are not significantly different. According to the rules, these types are more concentrated in the “H” and “M” types, demonstrating that dietary structure has a certain influence on individuals with high risk of stroke. In addition, from the perspective of correlation value (Table 8), it has a relatively higher correlation with other factors compared with alcohol consumption, drinking tea, and sleep.

4.4. Sleep. As displayed in Figure 12(f), the risk of stroke is lower when duration of sleep is appropriate. Very long or short duration of sleep is not conducive to avoid the risk of stroke, which is consistent with Huang et al.’s findings, expressing that a good sleep quality helps reduce the risk of stroke [11, 22]. From the perspective of rules, sleep is associated with smoking, alcohol consumption, and sport, and from the perspective of correlation, sleep, smoking, and exercise are relatively correlated together. People who exercise less and are obese have an increased risk of stroke, if the duration of their sleep is extremely long. People who exercise less, as well as being smokers, and alcohol drinkers have a higher risk of stroke, if the duration of their sleep would be lower than normal level.

As shown in Figure 12(a), “YESp” stands for “smoking” and “Nop” stands for “nonsmoking.” As illustrated in Figure 12(b), “C1p,” “C2p,” and “C3p” represent three kinds of exercise: “C1,” “C2,” and “C3.” In Figure 12(c), “YESp” stands for “drinking,” and “Nop” denotes “no drinking.” As displayed in Figure 12(d), “YESp” stands for “drinking tea,” and “Nop” represents “no tea drinking.” As depicted in Figure 12(e), “C1p,” “C2p,” and “C3p” represent “C1,”

TABLE 7: Values of risk factors for stroke.

Risk factors	Depth/frequency														Average depth	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13		14
SH	1															0.00
Hyte			2													2.00
Dysl				2	1											3.33
Diab				4	2		1									3.71
FSH						4		1								5.40
TIA				1	1	1	2	2								5.43
Smok						1	2	1								6.00
AF							7			2						6.67
Sport					1			1	3							7.00
Sleep							1			1		1				8.67
Gen						1	1				3	2				9.00
BMI										3	1					9.25
Tea								1		1			1			9.33
Age		1							2		1	3	3		1	10.00
Alco									1		2		1	1		10.60
DT											1	3				10.75

TABLE 8: A factor-based relationship matrix.

	SH	Hyte	Dysl	Diab	FSH	TIA	Smok	AF	Sport	Sleep	Gen	BMIc	Tea	Age	Alco
Hyte	6.84														
Dysl	6.34	6.34													
Diab	5.71	5.71	5.46												
FSH	5.71	4.95	4.95	4.64											
TIA	3.91	3.91	3.66	3.46	3.29										
Smok	4.16	4.16	4.16	4.16	3.85	3.03									
AF	0.45	0.45	0.45	0.45	0.45	0.33	0.20								
Sport	4.49	4.49	4.49	3.82	3.98	2.90	3.44	0.20							
Sleep	0.42	0.42	0.42	0.42	0.27	0.08	0.42	0.00	0.27						
Gen	1.74	1.74	1.74	1.43	1.43	1.05	1.05	0.00	1.74	0.08					
BMIc	2.17	2.17	2.17	2.17	2.17	2.17	2.17	0.00	2.17	0.08	1.05				
Tea	0.60	0.60	0.60	0.60	0.60	0.48	0.48	0.13	0.48	0.00	0.22	0.38			
Age	6.84	6.84	6.34	5.71	4.95	3.91	4.16	0.45	4.49	0.42	1.74	2.17	0.60		
Alco	0.70	0.70	0.70	0.70	0.70	0.40	0.70	0.00	0.70	0.19	0.22	0.40	0.14	0.70	
DT	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.00	0.96	0.00	0.62	0.86	0.38	0.96	0.22

“C2,” and “C3,” respectively. As illuminated in Figure 12(f), “TSp,” “TBp,” and “TLp” denote “TS,” “TB,” and “TL,” respectively.

5. Conclusions

In the present study, we optimized the decision tree C4.5 algorithm to assess and analyze risk factors of stroke (stroke history, hypertension, dyslipidemia, diabetes, family history of stroke, TIA, smoking, atrial fibrillation, sport, sleep, gender, BMI, drinking tea, age, alcohol consumption, and diet) via 5599 valid data collected. The classification result

showed to have an accuracy of 87.5281% and a kappa coefficient of 0.8344. It also was noted that classification performance was higher than that of the random forest and Logistic algorithm. Then, we focused on 6 factors influencing daily life, such as smoking, alcohol consumption, drinking tea, sleep, and sport, and presented a series of knowledge-based rules that are conducive to guide patients to adjust individuals’ living habits. With further analysis of decision tree and knowledge-based rules, the independent influence of each factor and the relationship between the factors were analyzed. Different from other studies, we analyzed the relationship between smoking and exercise, among

TABLE 9: Factors with higher correlation values than the mean values within the group.

Smok	Sport		Sleep		Tea		Alco		DT		
Factors	Correlation	Factors	Correlation	Factors	Correlation	Factors	Correlation	Factors	Correlation	Factors	Correlation
SH	4.16	SH	4.49	SH	0.42	SH	0.60	SH	0.70	SH	0.96
Hyte	4.16	Hyte	4.49	Hyte	0.42	Hyte	0.60	Hyte	0.70	Hyte	0.96
Dysl	4.16	Dysl	4.49	Dysl	0.42	Dysl	0.60	Dysl	0.70	Dysl	0.96
Diab	4.16	Age	4.49	Age	0.42	Age	0.60	Age	0.70	Age	0.96
Age	4.16	FSH	3.98	Diab	0.42	Diab	0.60	Diab	0.70	Diab	0.96
FSH	3.85	Diab	3.82	Smok	0.42	FSH	0.60	FSH	0.70	FSH	0.96
Sport	3.44	Smok	3.44	FSH	0.27	Smok	0.48	Smok	0.70	Smok	0.96
TIA	3.03	TIA	2.90	Sport	0.27	Sport	0.48	Sport	0.70	Sport	0.96
						TIA	0.48			TIA	0.96
										BMI	0.86

The effects of the 6 daily habits (smoking, alcohol consumption, drinking tea, diet, sleep, and exercise) on stroke risk are discussed in the next sections.

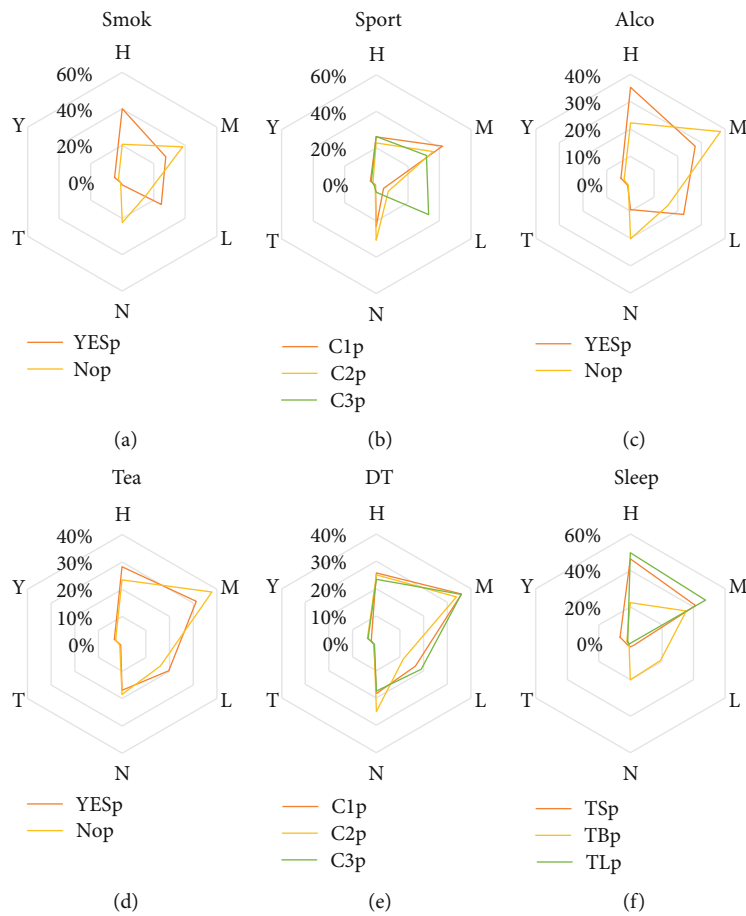


FIGURE 12: Radar charts illustrating the effects of daily life habits on risk factors of stroke.

alcohol consumption, drinking tea, and BMI, among diet, sport, and BMI, and among sleep, sport, smoking, and alcohol consumption and found that although these daily living habits cannot directly determine the risk of stroke (with low independent influence) they could be used to intervene the risk factors of stroke. On the one hand, smoking and exercise were strongly associated with other

risk factors of stroke; on the other hand, sleep, drinking tea, alcohol consumption, and diet were not firmly associated with other risk factors of stroke, and they were relatively tightly associated with smoking and exercise. However, further research needs to be conducted to indicate whether smoking and exercise play a significant role in the risk of stroke in daily habits.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by National Key R&D Program of China (Grant No. 2017YFE0112000), and the Collaborative Innovation Key Project of Shanghai University of Medicine & Health Sciences and Technology (no. E1-0200-18-201001).

Supplementary Materials

Supplementary Material 1: 37 knowledge-based rules related to the 6 daily living habits (smoking, alcohol consumption, drinking tea, diet, sleep, and sport). Material 2: table of factor sets and their weights. (*Supplementary Materials*)

References

- [1] V. L. Feigin, W. Wang, H. Fu et al., "Primary stroke prevention in China - a new approach," *Neurological Research*, vol. 37, no. 5, pp. 378–380, 2015.
- [2] B. Ovbiagele and M. N. Nguyen-Huynh, "Stroke epidemiology: advancing our understanding of disease mechanism and therapy," *Neurotherapeutics*, vol. 8, no. 3, pp. 319–329, 2011.
- [3] J. Guo, T. J. Guan, Y. Shen et al., "Lifestyle factors and gender-specific risk of stroke in adults with diabetes mellitus: a case-control study," *Journal of Stroke and Cerebrovascular Diseases*, vol. 27, no. 7, pp. 1852–1860, 2018.
- [4] P. A. Wolf, R. B. D'Agostino, A. J. Belanger, and W. B. Kannel, "Probability of stroke: a risk profile from the Framingham study," *Stroke*, vol. 22, no. 3, pp. 312–318, 1991.
- [5] J. Y. Chong and R. L. Sacco, "Risk factors for stroke, assessing risk, and the mass and high-risk approaches for stroke prevention," *CONTINUUM: Lifelong Learning in Neurology*, vol. 11, no. 4, pp. 18–34, 2005.
- [6] P. M. Rist, J. E. Buring, C. S. Kase, and T. Kurth, "Healthy lifestyle and functional outcomes from stroke in women," *American Journal of Medicine*, vol. 129, no. 7, pp. 715–724.e2, 2016.
- [7] R. R. Bailey, A. Phad, R. McGrath, and D. Haire-Joshu, "Prevalence of five lifestyle risk factors among U.S. adults with and without stroke," *Disability and Health Journal*, vol. 12, no. 2, pp. 323–327, 2019.
- [8] V. A. Hill, B. G. Vickrey, E. M. Cheng et al., "A pilot trial of a lifestyle intervention for stroke survivors: design of Healthy Eating and Lifestyle after Stroke (HEALS)," *Journal of Stroke and Cerebrovascular Diseases*, vol. 26, no. 12, pp. 2806–2813, 2017.
- [9] S. Lueders, B. Schrader, J. Baesecke et al., "ELITE study-nutrition, lifestyle and individual information for the prevention of stroke, dementia and heart attack-study design and cardiovascular status," *Deutsche Medizinische Wochenschrift*, vol. 144, no. 6, pp. e42–e50, 2019.
- [10] E. Altobelli, P. M. Angeletti, L. Rapacchietta, and R. Petrocelli, "Overview of meta-analyses: the impact of dietary lifestyle on stroke risk," *International Journal of Environmental Research and Public Health*, vol. 16, no. 19, p. 3582, 2019.
- [11] Z. X. Huang, X. L. Lin, H. K. Lu, X. Y. Liang, L. J. Fan, and X. T. Liu, "Lifestyles correlate with stroke recurrence in Chinese inpatients with first-ever acute ischemic stroke," *Journal of Neurology*, vol. 266, no. 5, pp. 1194–1202, 2019.
- [12] T. B. Cumming, E. Holliday, D. Dunstan, and C. English, "Television viewing time and stroke risk: Australian diabetes obesity and lifestyle study (1999-2012)," *Journal of Stroke and Cerebrovascular Diseases*, vol. 28, no. 4, pp. 963–970, 2019.
- [13] M. Ture, F. Tokatli, and I. Kurt, "Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2017–2026, 2009.
- [14] J. R. Quinlan, *C4.5 Programming for Machine Learning*, Morgan Kaufmann, 1993.
- [15] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.
- [16] T. Clifford, J. Bruce, T. Obafemi-Ajayi, and J. Matta, "Comparative analysis of feature selection methods to identify biomarkers in a stroke-related dataset," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 51–58, Siena, Italy, July 2019.
- [17] K. Yamagishi, H. Iso, A. Kitamura et al., "Smoking raises the risk of total and ischemic strokes in hypertensive men," *Hypertension Research*, vol. 26, no. 3, pp. 209–217, 2003.
- [18] M. N. McDonnell, S. L. Hillier, S. P. Hooker, A. Le, S. E. Judd, and V. J. Howard, "Physical activity frequency and risk of incident stroke in a national US study of blacks and whites," *Stroke*, vol. 44, no. 9, pp. 2519–2524, 2013.
- [19] D. Hu, J. Huang, Y. Wang, D. Zhang, and Y. Qu, "Dairy foods and risk of stroke: a meta-analysis of prospective cohort studies," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 24, no. 5, pp. 460–469, 2014.
- [20] P. M. Sosa, M. A. de Souza, and P. B. Mello-Carpes, "Green tea and red tea from *Camellia sinensis* partially prevented the motor deficits and striatal oxidative damage induced by hemorrhagic stroke in rats," *Neural Plasticity*, vol. 2018, 8 pages, 2018.
- [21] J. C. Zhang, H. Xu, Y. Yuan et al., "Delayed treatment with green tea polyphenol EGCG promotes neurogenesis after ischemic stroke in adult mice," *Molecular Neurobiology*, vol. 54, no. 5, pp. 3652–3664, 2017.
- [22] D. L. de Oliveira Diniz, P. R. Barreto, P. F. C. de Bruin, and V. M. S. de Bruin, "Wake-up stroke: clinical characteristics, sedentary lifestyle, and daytime sleepiness," *Revista Da Associação Médica Brasileira*, vol. 62, no. 7, pp. 628–634, 2016.

Research Article

CUL1-Mediated Organelle Fission Pathway Inhibits the Development of Chronic Obstructive Pulmonary Disease

Ran Li,¹ Feng Xu,² Xiao Wu,² Shaoping Ji,³ and Ruixue Xia² 

¹Department of Critical Care Medicine, Henan University Huaihe Hospital, No. 8 Baobei Street, Gulou District, Kaifeng 475000, China

²Department of Respiratory and Critical Care Medicine, Henan University Huaihe Hospital, No. 8 Baobei Street, Gulou District, Kaifeng 475000, China

³Cell Signal Transduction Laboratory and Institute of Biomedical Informatics, School of Basic Medical Sciences, Henan University, Kaifeng 475000, China

Correspondence should be addressed to Ruixue Xia; xiasi86592732783@163.com

Received 17 April 2020; Accepted 4 May 2020; Published 26 May 2020

Guest Editor: Tao Huang

Copyright © 2020 Ran Li 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.

Chronic obstructive pulmonary disease (COPD) is a global high-incidence chronic airway inflammation disease. Its deterioration will lead to more serious lung lesions and even lung cancer. Therefore, it is urgent to determine the pathogenesis of COPD and find potential therapeutic targets. The purpose of this study is to reveal the molecular mechanism of COPD disease development through in-depth analysis of transcription factors and ncRNA-driven pathogenic modules of COPD. We obtained the expression profile of COPD-related microRNAs from the NCBI-GEO database and analyzed the differences among groups to identify the microRNAs significantly associated with COPD. Then, their target genes are predicted and mapped to a protein-protein interaction (PPI) network. Finally, key transcription factors and the ncRNA of the regulatory module were identified based on the hypergeometric test. The results showed that CUL1 was the most interactive gene in the highly interactive module, so it was recognized as a dysfunctional molecule of COPD. Enrichment analysis also showed that it was much involved in the biological process of organelle fission, the highest number of regulatory modules. In addition, ncRNAs, mainly composed of miR-590-3p, miR-495-3p, miR-186-5p, and transcription factors such as MYC, BRCA1, and CDX2, significantly regulate COPD dysfunction blocks. In summary, we revealed that the COPD-related target gene CUL1 plays a key role in the potential dysfunction of the disease. It promotes the proliferation of fibroblast cells in COPD patients by mediating functional signals of organelle fission and thus participates in the progress of the disease. Our research helps biologists to further understand the etiology and development trend of COPD.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a respiratory system disease which is usually caused by chronic inflammation caused by respiratory pathogens such as viruses or bacteria [1]. It is usually accompanied by persistent respiratory symptoms and airflow limitations. As the main cause of disability and death worldwide, the deterioration of pathological inflammation will evolve into highly susceptible chronic bacterial pulmonary infection and acute recurrence of COPD (AECOPD) [2–4]. Moreover, it can induce chronic hypercapnia respiratory failure (CHRF)

and lung cancer [5, 6]. These all cause a significant health burden to patients, seriously impairing their quality of life, exercise ability, and lung function [7, 8]. In addition to susceptible adult groups, neonates with pulmonary and bronchial dysplasia (BPD) are thought to have a high latent risk of COPD [9]. COPD is the product of complex interaction between heredity, the environment, and other factors. In the field of genetics, identifying genetic variants that lead to disease progression is conducive to identifying risk factors, understanding potential disease mechanisms, and developing new therapies. Genome-wide association studies (GWAS) have successfully identified many loci related to

lung function, COPD, and asthma. Among them, lung function-related single nucleotide polymorphisms (SNPs) overlap considerably with SNPs that may be involved in COPD mechanisms [10]. Besides, air pollution seriously impairs lung and airway functions and induces many diseases (including lung cancer, COPD, cardiovascular disease, and malignant tumors) [11, 12].

At present, there are various reports about the molecular mechanisms of COPD. For example, PINK1-PARK2-mediated mitochondrial autophagy plays a major role in the pathogenesis of aging-related lung diseases such as chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF) [13]. Cytokines such as interleukin-1 (IL-1), IL-6, and IL-8 are key participants in the initiation and transmission of inflammation in chronic inflammatory airway diseases such as COPD [14]. Multipotent mesenchymal stem/stromal cells (MSCs) have strong self-renewal characteristics and the ability to differentiate into tissue-specific cells. It has obvious therapeutic potential in early clinical trials in acute respiratory distress syndrome (ARDS) and chronic obstructive pulmonary disease (COPD). It is useful to note that secretory proteins derived from MSC may become potential therapeutic agents for invasive lung diseases [15]. Besides, research on molecular functional pathways is also an important direction to reveal the pathogenesis and treatment mechanism of COPD. According to research reports, fibroblast growth factor 10 (Fgf10) located in the lung mesenchyme is essential for promoting epithelial cell regeneration after injury. The signaling pathway plays a regulatory role in different human lung diseases such as bronchopulmonary dysplasia (BPD), idiopathic pulmonary fibrosis (IPF), and COPD [16]. Abnormal regulation of Wnt/beta-catenin signal transduction is also closely related to COPD and other disease types. It can be used as a potential target for disease treatment, and the progress of Wnt activator is particularly important [17]. This series of experimental results has greatly deepened our understanding of the pathogenesis of COPD and encouraged us to conduct more in-depth research.

In order to have a deeper understanding of the underlying mechanism of COPD disease progression and related signaling pathways, we combined the frontier reports of smoking on potential disease risks and systematically analyzed the microarray expression profiles of healthy smokers and COPD smokers through involving healthy controls [18, 19]. It was found that CUL1 was highly expressed and significantly mediated the dysfunction module of COPD, especially in the enrichment analysis of the function and pathway. It was noted that CUL1 promoted the proliferation of fibroblast cells in COPD patients by activating organelle fission. Therefore, this study identified it as the core biomarker of COPD. In conclusion, the comprehensive and systematic analysis in this study revealed that CUL1 participates in the organelle fission pathway to inhibit the proliferation of fibroblast cells in COPD patients. This discovery will contribute to the understanding of the pathogenic mechanism of diseases in the medical community and also indicates the direction for scientific research to effectively curb the global spread of diseases.

2. Results

2.1. Predicting Target Genes Based on Targeting Relations of MicroRNAs. Differential expression analysis can screen genes related to the occurrence and development of COPD. Therefore, based on the microarray expression profiles of microRNAs, we screened for differentially expressed microRNAs between nonsmokers and healthy smokers and between nonsmokers and COPD smokers. A total of 123 differentially expressed microRNAs were obtained. These microRNAs may play an important role in the pathogenesis of COPD. Then, 9952 target genes were predicted according to the targeting relationship of microRNAs.

In order to observe the interaction between COPD target genes, we mapped it to human protein-protein interaction (PPIs) networks and obtained a PPI of target genes. This PPI network consists of 5878 gene nodes and 91496 edges. According to the principles of systems biology and molecular biology, it can be concluded that this PPI generalizes the molecular pathogenic mechanism of COPD to a certain extent.

2.2. High Interaction Module Characterizes Potential Dysfunction of COPD. In order to further explore the key pathways involved in COPD, we conducted a modular analysis of PPIs related to target genes. Based on the cohesion and neighbor selection algorithm, we identified 19 functional modules (Figure 1) with 1656 related genes. Relatively speaking, these interactive modules have more significant interaction relationships, which can better characterize the basic molecular mechanism of COPD. At the gene level, module genes represent a series of highly related genes. Genes in the same module may play similar biological functions or coregulate certain biological processes. From the point of view of systems biology, searching for modular genes with potential functions is actually a bridge between the functions of individual genes and the characteristics of global networks. In addition, each module may represent a pathway that mediates the onset of COPD. Therefore, the identification of gene function modules is the core of targeted COPD research and the key step to understanding its molecular mechanism.

In order to further explore the function of module genes in the pathogenesis of COPD, we analyzed the enrichment of the function and pathway of module genes (Figures 2(a) and 2(b)). Consequences of 23548 biological processes, 2879 cell components, 2849 molecular functions, and 944 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were obtained. It was found that the genes in the module were significantly enriched in various biological processes involving COPD, such as organelle fission, mitosis, and cell adhesion molecule binding. At the same time, module genes are also significantly involved in the PI3K-Akt signaling pathway, autophagy-animal and RNA transport, and other COPD-related signaling pathways. In addition, based on statistical analysis, we found that up to 18 modules were significantly enriched in the biological processes of organelle fission and regulation of binding, while mitotic mitosis, negative regulation of binding, and binding of cell adhesion molecules were

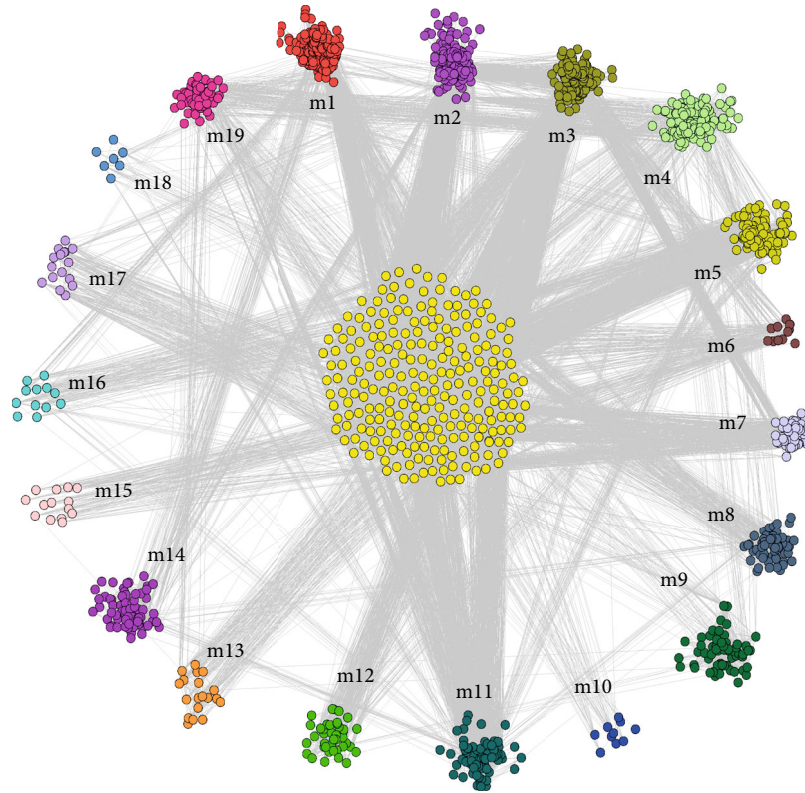


FIGURE 1: Highly interactive module characterizes 19 COPD highly interactive modules obtained from modular analysis of potential dysfunction of COPD. Different color circle dot groups represent 19 different module genes, and the center yellow dot group represents module overlap genes.

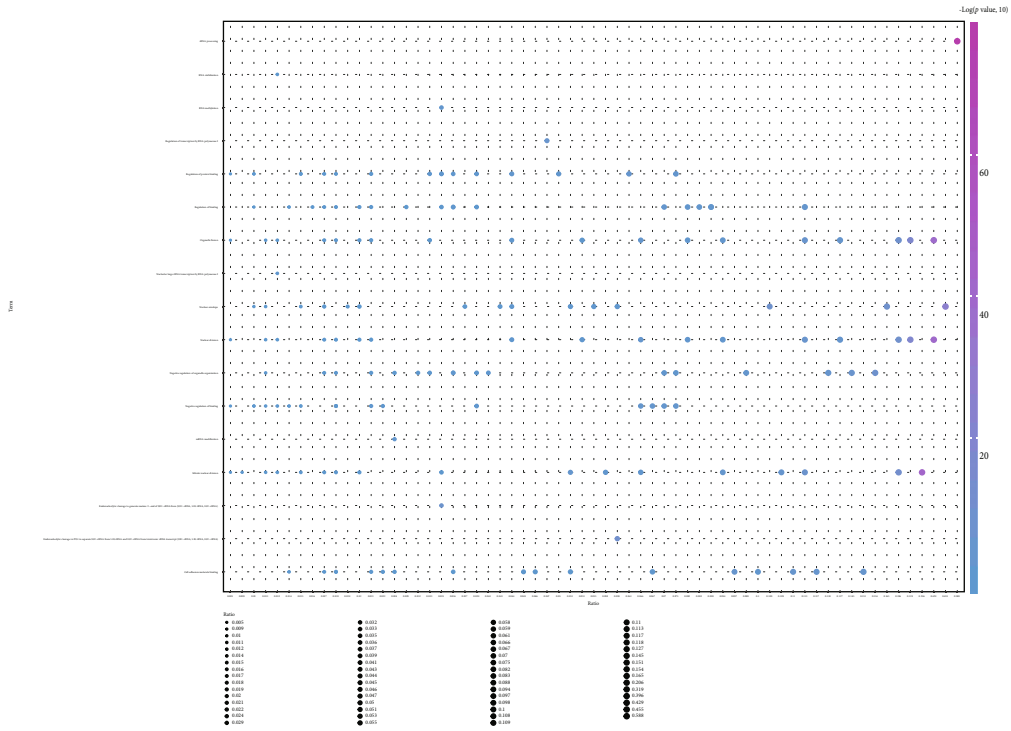
significantly enriched in 17 modules. In retrospect, we integrated 19 dysfunctional module genes and constructed a functional global network (Figure 2(c)). This functional network may imply the overall dysfunction mechanism of COPD.

2.3. Modular Introductory Gene May Be the Core Gene of COPD Disease. The modular approach has deepened our understanding of the basic molecular mechanism of COPD, but 1656 genes still fail to accurately represent the dysfunction mechanism of COPD. Therefore, in order to identify the genes that play a critical role in the dysfunction module, we first constructed a protein interaction subnet for the genes in the module. Then, based on the module subnet, we analyze the connectivity of nodes (Figure 3). According to regulations, genes with greater connectivity mean more active supervisory roles in a module, so in a module, genes with the greatest connectivity will be considered as intrinsic genes in dysfunctional modules. Depending on the order of connectivity, we find that the core gene *CUL1* of module 1 is the most prominent. It effectively targets other genes and drives dysfunctional modules and then mediates the occurrence of diseases. It plays an important role in the probable pathogenesis of COPD. Therefore, *CUL1* was identified as the core endogenous gene of COPD.

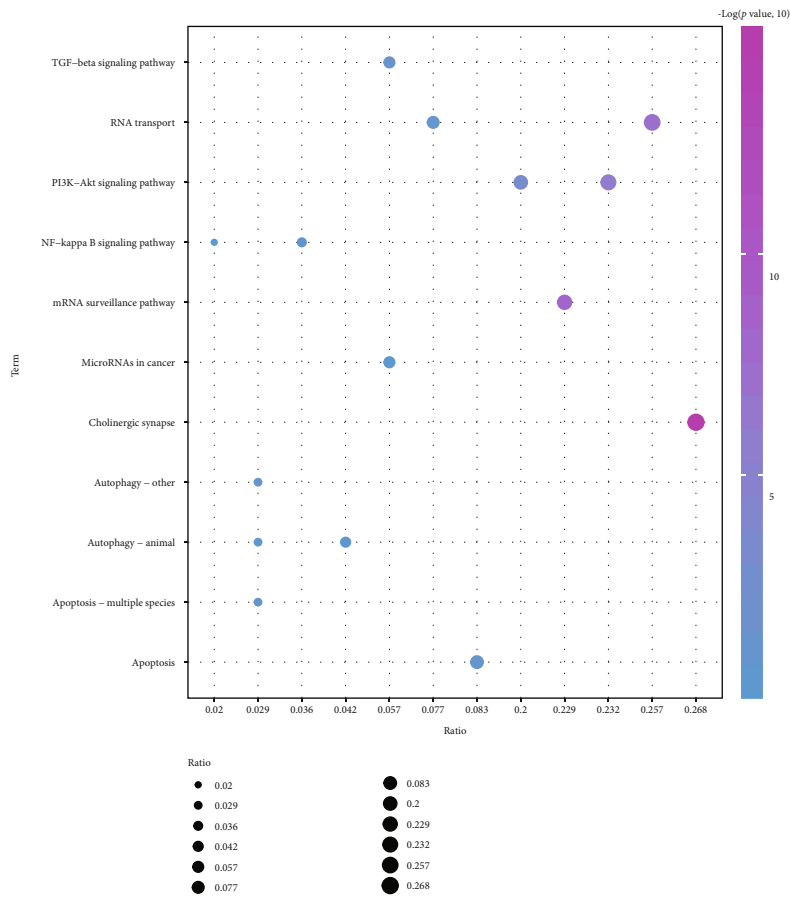
2.4. Modular ncRNA Pivot Mediates COPD Dysfunction. In systemic genetics, gene transcription and posttranscriptional

regulation have been taken into account as key regulators of disease occurrence and development, and ncRNA is recognized as a gene regulator. Although the regulation of a single or several ncRNAs on the pathogenesis of COPD has been confirmed by biologists, few studies have focused on their comprehensive regulation of dysfunctional modules. Scientific prediction of ncRNA pivot regulators in dysfunctional modules is advantageous for us to explore the transcriptional regulation mechanism of COPD. To this end, pivot analysis based on the targeting relationship between ncRNA and module genes was performed to explore ncRNA regulators causing module dysfunction. The predicted results (Figure 4) show that a total of 2511 ncRNAs involve 1360 ncRNA-module target pairs, which substantially regulate these COPD-related functional modules and affect the occurrence and development of diseases. In addition, the number of pivot regulatory modules was statistically analyzed. It was found that microRNA miR-590-3p drastically regulated 15 functional modules and played a central role in the potential dysfunction mechanism of COPD. miR-495-3p was identified to be significantly associated with 11 dysfunctional modules and played a major role in the pathogenesis of COPD. Other ncRNAs also show significant regulatory effects on modules, which may be a possibility pathogenic factor of COPD and play a potential role.

2.5. TF Pivot Driver Module Participates in COPD Dysfunction Mechanism. In addition to ncRNA, transcription factors are



(a)



(b)

FIGURE 2: Continued.

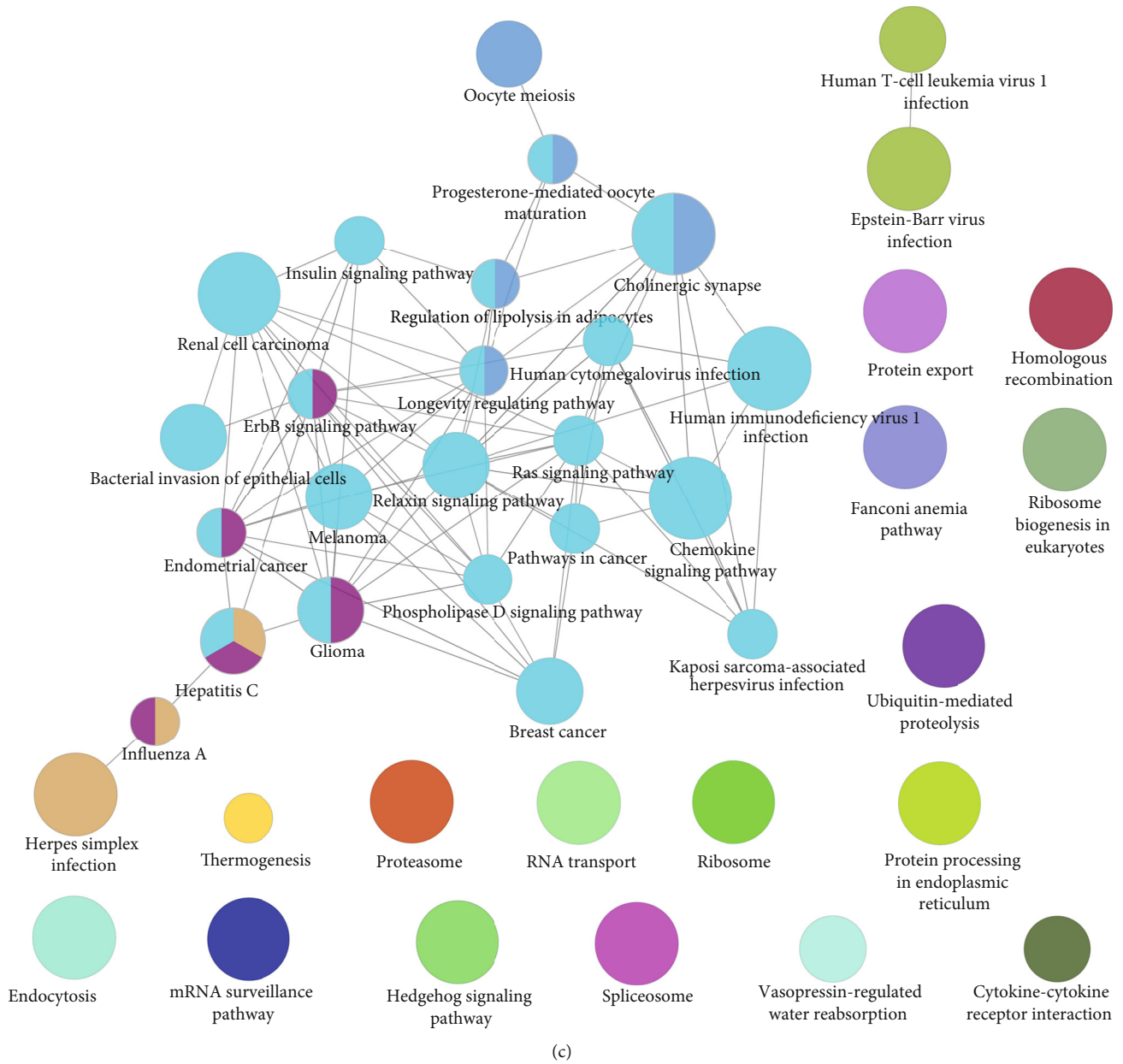


FIGURE 2: Functional and pathway enrichment analysis of modular genes (excerpts). (a). GO functional enrichment analysis of module genes. From blue to purple, the enrichment increased dramatically. The larger the circle, the larger the proportion of module genes in GO functional entry genes. (b). KEGG pathway enrichment analysis of modular genes. From blue to purple, the enrichment increased markedly. The larger the circle, the larger the proportion of module genes to KEGG pathway entry genes. (c). Network map of the functional pathway.

equally essential for the transcriptional regulation of genes. Numerous studies have shown that disordered expression of transcription factors may lead to the occurrence of various diseases. The occurrence of COPD is also closely related to the dysfunction of transcription factors, which are fully reflected in the regulation of dysfunctional modules. Based on the pivot analysis of transcription factors (Figure 5), we identified 55 transcription factors that may be associated with COPD dysfunction, involving 67 TF-module regulatory pairs. It is to be noted that statistical analysis of these TF-module regulatory pairs reveals that MYC significantly regulates four modules, while BRCA1 and CDX2 regulate two modules. These tran-

scription factors play an essential role in the occurrence and development of COPD. Additional transcription factors also show significant regulatory effects on modules, contributing to the pathogenesis of COPD, which may be a potential dysfunctional molecule of COPD.

3. Discussion

Chronic obstructive pulmonary disease (COPD) is one of the most common diseases in the world, and smoking is thought to be the main contributing factor to its pathogenesis and development [18]. Despite the fact that researchers have

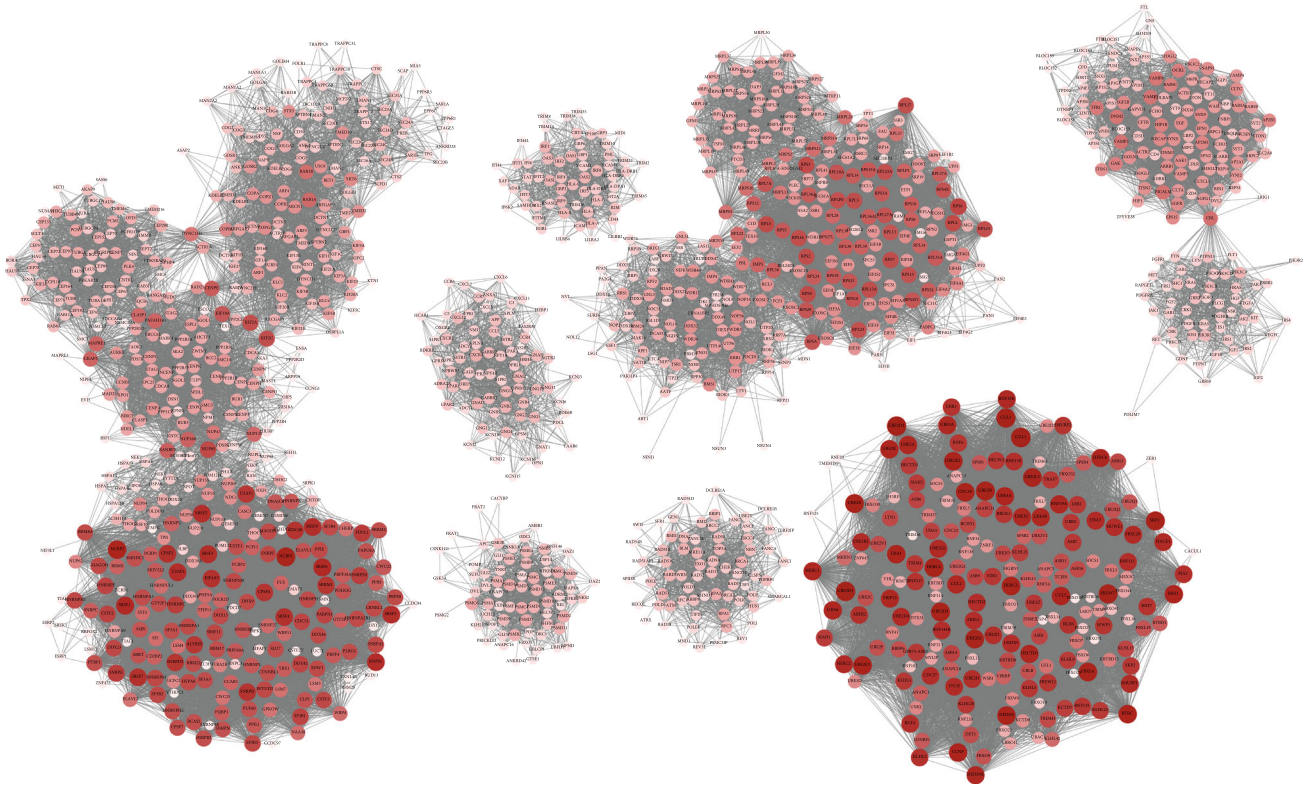


FIGURE 3: Highly interactive module-driven genes. Node colors from brown to dark red represent the connectivity of module genes from tiny to large, and each node group represents each module.

explored the etiology of COPD from various aspects, the potential relationship between COPD and tobacco substances and other factors remains unclear. In this study, we synthesized the multivariate analysis to determine key molecules and their disease-mediated functions. At the molecular level, we first construct the internal subnet of the interaction module and then analyze the connectivity of each module subnet and get the most connected intrinsic gene *CUL1*. Maximum connectivity of introductory genes means that these are the genes that interact the most with them in this module, that is to say, they result in pulling the whole body together. The slight abnormal expression of endogenous genes may bring about great changes in the module level. Cullin1 (*CUL1*) is a scaffold protein of ubiquitin E3 ligase Skp1-Cullin1-F-box protein complex (SCF). It is grouped and located in the nucleus (to a lesser extent in the cytoplasm), and its ubiquitousness involves cell cycle processes, signal transduction, and transcription. At the same time, *CUL1* is also a cancer-related gene which has attracted wide attention in the academic circles in recent years. It mainly affects the proliferation, invasion, and metastasis of cancer cells. It mediates the occurrence, development, and adverse prognosis of various diseases through related pathways, and it has been determined that it is a new diagnostic and prognostic marker for lung cancer and other cancers [20–23]. Taking into account these confirmed results, *CUL1* was identified as a dysfunctional molecule and potential biomarker of COPD. Nevertheless, the underlying relationship between its functions, pathways, and the physiological processes of

COPD has not been clearly elucidated, and we are required to conduct enrichment analysis of functional pathways to verify it.

At the segmental level, we note that modular genes are involved in the most significant biological process of organelle fission, up to 18 dysfunctional modules. Functions such as mitochondrial fission and fusion of immune-related organelles can directly affect health, leading to diseases such as aging, tumorigenesis, lung injury, and COPD. In addition, COPD is recognized as oxidative stress injury caused by long-term exposure to stimulants such as smoke inhalation, and reactive oxygen species (ROS) induce structural and functional mutations in airway epithelial mitochondria [24, 25]. Therefore, the dysfunctional molecule *CUL1* in the COPD highly interactive module mediates oxidative damage induced by the fission of functional organelles of the most prominent module, thus accelerating the proliferation of disease cells. Other significant functional pathways are also involved in the process of disease generation and pathogenesis to varying degrees, which need further experimental verification and analysis by future researchers.

Then, we predicted that 2512 ncRNAs participated in the occurrence and development of COPD through regulatory modules and verified their abnormal expression in COPD to varying degrees based on a different analysis. According to the statistical analysis, we determined that miR-590-3p had a significant effect on 15 dysfunctional modules, FENRR had significant regulation on 9 modules, miR-218-5p had regulation on 7 modules, and other ncRNAs

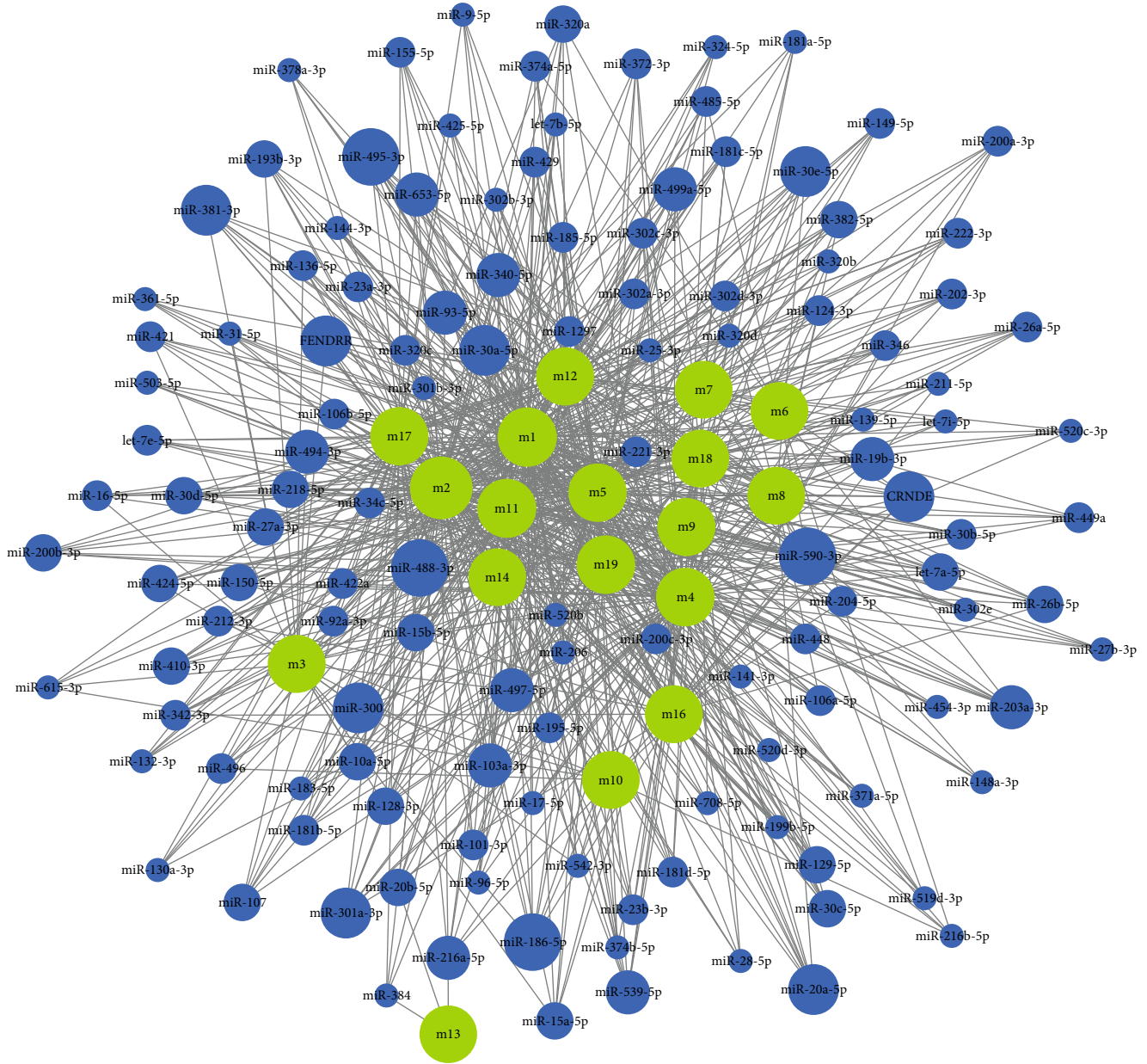


FIGURE 4: Regulation of the ncRNA pivot regulator on the dysfunction module. The green circle represents the module, the blue circle represents the ncRNA of the control module, and the circle size represents the number of control modules. The larger the circle, the more the number of regulations.

had regulation on different numbers of modules. Among them, downregulation of miR-590-3p is common especially in breast cancer, which is negatively correlated with SIRT1 expression and coinhibits cell survival and induces cancer cell apoptosis. It may be a possibility target for further development and more effective treatment of breast cancer, but its potential role in COPD has not been reported [26]. However, the forecast results of this study clearly indicate that microRNA-590-3p, as the ncRNA regulating the most COPD dysfunction module, may play a potential role in the pathogenesis of COPD, which can be used as a candidate factor for further molecular experimental validation research. Long-chain noncoding RNA (lncRNA) FENDRR can be

upregulated by CNVR_3425.1, which may be a potential target for COPD treatment [27]. In a comprehensive analysis of the microRNA-RNA-lncRNA network in nonsmoking and smoking COPD patients, the microRNA-218-5p and its interaction targets may be associated with the deterioration process of nonsmoking COPD [28].

Finally, we identified 55 transcription factors that substantially regulate COPD dysfunction in varying degrees. According to regulatory analysis, MYC significantly regulates four COPD dysfunction modules, and BRCA1 and E2F1 have regulatory effects on two modules. c-MYC, a transcription activator located in small pulmonary vessels, is highly expressed in COPD lung tissue. The abnormal apoptosis

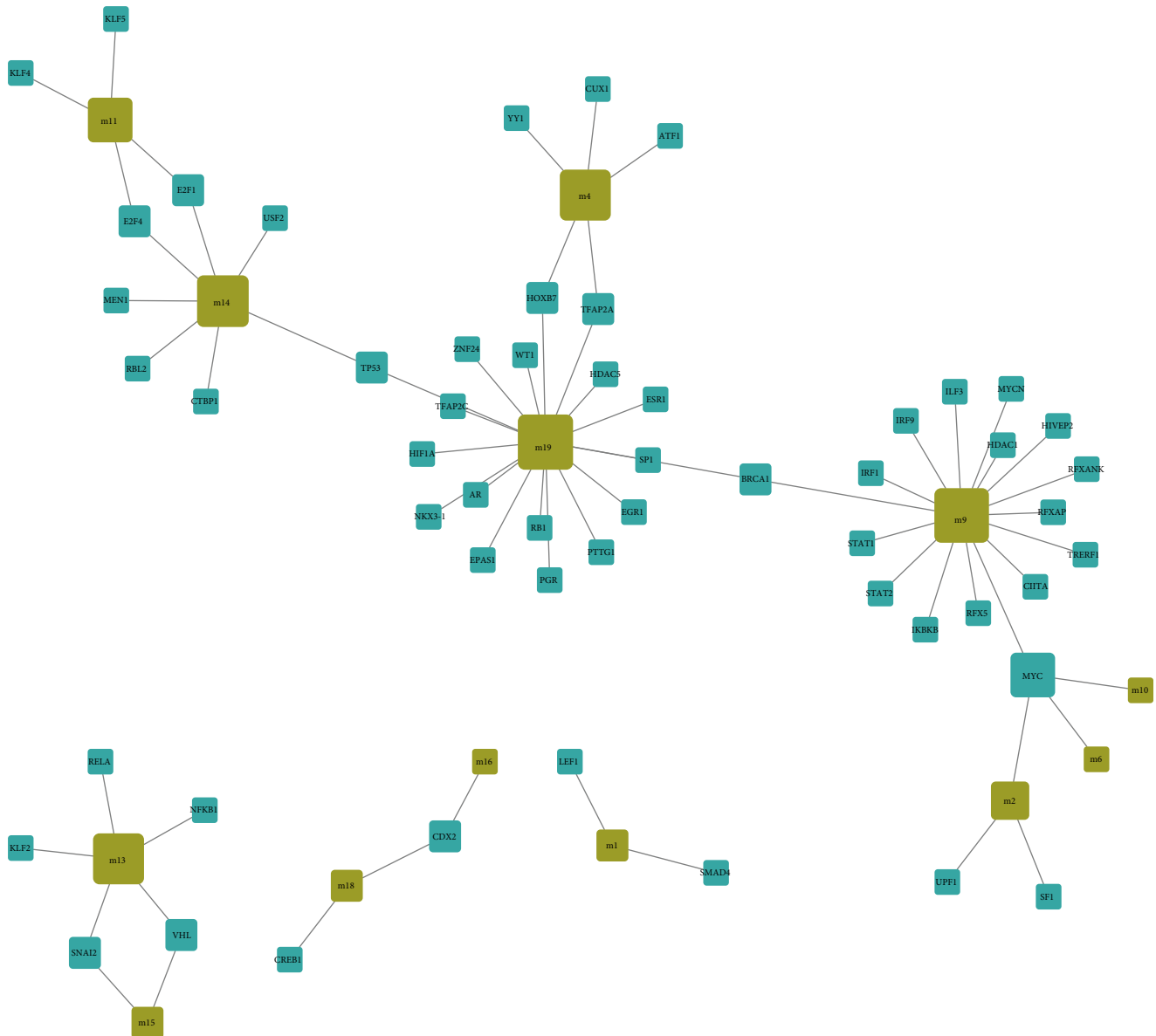


FIGURE 5: Regulation of the TF pivot regulator on the dysfunction module. The brown square represents the module. The blue square represents the transcription factor of the regulatory module.

and proliferative activity induced by c-MYC may contribute to the structural remodeling of COPD pulmonary vessels [28]. Moreover, c-MYC has been recognized as a carcinogen of lung cancer in molecular biology [29]. In addition, BRCA1 showed increased protein abundance expression in proteomics analysis, which is expected to be a candidate molecule for further exploration of COPD [30]. E2F1 is a transcription factor targeted by microRNA-197. Its molecular level increases in the arteries (PA) of COPD patients and mediates the role of microRNA-197 in vascular wall remodeling regulating the phenotype of smooth muscle cells (SMC) [31]. Other transcription factors that significantly regulate COPD dysfunction modules may also participate in the elementary process of COPD, which needs to be verified by experiments.

In conclusion, based on the modular analysis method, CUL1 was identified as the core endogenous gene of COPD,

which inhibits the deterioration of COPD by mediating the organelle fission pathway. COPD patients should reduce the incidence of cigarette smoking during treatment and quit smoking as soon as possible. The risk of relapse should also be paid attention to in subsequent rehabilitation. In addition, ncRNA and transcription factors mediating dysfunction modules were explored by combining transcriptional and posttranscriptional regulation. These findings will help to reveal the intricate molecular pathogenic mechanism of COPD and provide new candidate factors and a solid theoretical basis for subsequent research.

4. Materials and Methods

4.1. Data Resources. Firstly, we collected a set of microRNA expression profiles of COPD from the NCBI Gene

Expression Omnibus database (GEO Dataset) [32], the number of which is GSE56923. The data set included 8 non-smokers, 8 healthy smokers, and 8 COPD smokers. Secondly, we downloaded all the human protein-protein interaction data in the STRING V10 database [33] to construct the differentially targeted gene PPIs related to microRNAs. The STRING database is a universal search tool for retrieving interactive genes/proteins. It can help us find and annotate functional interactions in the life system. Then, we screened ncRNA-RNA (protein) interaction pairs with a score ≥ 0.5 from the RAID v2.0 database [34] for predicting target genes. At the same time, all human transcription factor target data are downloaded and used in the TRRUST V2 database [35] to predict hypothetical factors that regulate modular genes.

4.2. Difference Analysis. The differential expression analysis of microRNA expression profile data in this study was implemented by the R language limma package [36–38]. Firstly, the background correct function is used for background correction and standardization. Secondly, the method of normal between array function quantile normalization was used to filter out the control probe and the low expression probe. Then, based on the lmFit and eBayes functions, the default parameters are used to determine the differential expression of microRNAs in the dataset that are potentially involved in the pathogenesis of COPD.

4.3. Recognition Module Based on Protein Interaction Network. Modularization is essential for this study. Firstly, we use the screened ncRNA-RNA (protein) interaction pairs as background sets to search for differentially expressed target genes targeting for microRNAs. Cytoscape [39] visualization methods are used to observe the mapping of target genes into a human protein-protein interaction network more intuitively. Subsequently, interaction pairs containing only these genes were extracted, and a target gene PPI for COPD was constructed. Then, we use the plug-in ClusterONE [40] with default parameters to identify modules based on the cohesion algorithm and neighbor selection strategy. Finally, on the basis of modularization, we also conducted connectivity analysis among genes to screen out the most interactive endogenous genes in the module.

4.4. Functional and Pathway Enrichment Analysis. Exploring the functions and signaling pathways of gene involvement is often advantageous to study the molecular mechanism of diseases. Enrichment of genes in dysfunctional modules is an effective means to explore the underlying pathogenesis of COPD. Therefore, based on the R language clusterProfiler package [41], we performed enrichment analysis of the gene of the module with the Gene Ontology (GO) function (p value cutoff = 0.05, q value cutoff = 0.05) and KEGG pathway (p value cutoff = 0.05, q value cutoff = 0.05). In addition, we also used ClueGO plug-in [42] with default parameters in Cytoscape to analyze the functions of all modules' comprehensive network and build a functional network of COPD.

4.5. Pivot Analysis and Prediction of ncRNA and TF in Regulatory Module. We stipulate that the pivot regulator means that the number of targeting regulators between each

regulator and each module exceeds 2. Meanwhile, the significance of the interaction between the pivot regulator and the module is calculated by the hypergeometric test (p value < 0.01). In this study, we used the target data of ncRNA and TF as the background and combined them with the Python program to forecast the pivot analysis. We obtained pivot regulators of a meaningful regulatory dysfunction module.

Data Availability

All the data in this manuscript can be accessed in GSE56923.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

Ruixue Xia conceived and designed the study. Ran Li was responsible for the collection and analysis of the data. All authors interpreted the data, drafted the manuscript, and approved the final version of the manuscript.

Acknowledgments

This manuscript is supported from the Fund of Henan Provincial Department of Education (No. 20B320001).

Supplementary Materials

Table S1: differential gene expression in disease samples. Table S2: functions and pathways of modular gene participation. Table S3: the ncRNA pivot that regulates modules. Table S4: the TF pivot that regulates modules. (*Supplementary Materials*)

References






- [1] S. Makris and S. Johnston, "Recent advances in understanding rhinovirus immunity," *F1000Research*, vol. 7, 2018.
- [2] Y. C. Su, F. Jalalvand, J. Thegerström, and K. Riesbeck, "The interplay between immune response and bacterial infection in COPD: focus upon non-typeable *Haemophilus influenzae*," *Frontiers in Immunology*, vol. 9, 2018.
- [3] N. Pires, P. Pinto, N. Marçal et al., "Pharmacological treatment of COPD – New evidence," *Pulmonology*, vol. 25, no. 2, pp. 90–96, 2019.
- [4] H. Wang, D. Anthony, S. Selemidis, R. Vlahos, and S. Bozinovski, "Resolving viral-induced secondary bacterial infection in COPD: a concise review," *Frontiers in Immunology*, vol. 9, p. 2345, 2018.
- [5] S. van der Leest and M. L. Duiverman, "High-intensity non-invasive ventilation in stable hypercapnic COPD: evidence of efficacy and practical advice," *Respirology*, vol. 24, no. 4, pp. 318–328, 2019.
- [6] M. S. Eapen, P. M. Hansbro, A. K. Larsson-Callerfelt et al., "Chronic obstructive pulmonary disease and lung cancer: underlying pathophysiology and new therapeutic modalities," *Drugs*, vol. 78, no. 16, pp. 1717–1740, 2018.

- [7] R. E. K. Russell, "What does the TOVITO programme tell us about how we can manage COPD?," *Turkish Thoracic Journal*, vol. 19, no. 4, pp. 216–219, 2018.
- [8] K. Marsaa, S. Gundestrup, J. U. Jensen et al., "Danish respiratory society position paper: palliative care in patients with chronic progressive non-malignant lung diseases," *European Clinical Respiratory Journal*, vol. 5, no. 1, 2018.
- [9] J. L. Y. Cheong and L. W. Doyle, "An update on pulmonary and neurodevelopmental outcomes of bronchopulmonary dysplasia," *Seminars in Perinatology*, vol. 42, no. 7, pp. 478–484, 2018.
- [10] R. Hall, I. P. Hall, and I. Sayers, "Genetic risk factors for the development of pulmonary disease identified by genome-wide association," *Respirology*, vol. 24, no. 3, pp. 204–214, 2019.
- [11] D. E. Schraufnagel, J. R. Balmes, C. T. Cowl et al., "Air Pollution and Noncommunicable Diseases: A Review by the Forum of International Respiratory Societies' Environmental Committee, Part 2: Air Pollution and Organ Systems," *Chest*, vol. 155, no. 2, pp. 417–426, 2019.
- [12] K. C. Rajendra, S. D. Shukla, S. S. Gautam, P. M. Hansbro, and R. F. O'Toole, "The role of environmental exposure to non-cigarette smoke in lung disease," *Clinical and Translational Medicine*, vol. 7, no. 1, p. 39, 2018.
- [13] K. Tsubouchi, J. Araya, and K. Kuwano, "PINK1-PARK2-mediated mitophagy in COPD and IPF pathogenesis," *Inflammation and Regeneration*, vol. 38, no. 1, 2018.
- [14] J. Garth, J. Barnes, and S. Krick, "Targeting cytokines as evolving treatment strategies in chronic inflammatory airway diseases," *International Journal of Molecular Sciences*, vol. 19, no. 11, p. 3402, 2018.
- [15] A. Mohammadipoor, B. Antebi, A. I. Batchinsky, and L. C. Cancio, "Therapeutic potential of products derived from mesenchymal stem/stromal cells in pulmonary disease," *Respiratory Research*, vol. 19, no. 1, p. 218, 2018.
- [16] T. Yuan, T. Volckaert, D. Chanda, V. J. Thannickal, and S. P. De Langhe, "Fgf10 signaling in lung development, homeostasis, disease, and repair after injury," *Frontiers in Genetics*, vol. 9, p. 418, 2018.
- [17] P. Huang, R. Yan, X. Zhang, L. Wang, X. Ke, and Y. Qu, "Activating Wnt/ β -catenin signaling pathway for disease therapy: Challenges and opportunities," *Pharmacology & Therapeutics*, vol. 196, pp. 79–90, 2019.
- [18] Y. Jiang, X. Wang, and D. Hu, "Mitochondrial alterations during oxidative stress in chronic obstructive pulmonary disease," *International Journal of Chronic Obstructive Pulmonary Disease*, vol. Volume 12, pp. 1153–1162, 2017.
- [19] K. H. Lee, J. Jeong, Y. J. Koo, A. H. Jang, C. H. Lee, and C. G. Yoo, "Exogenous neutrophil elastase enters bronchial epithelial cells and suppresses cigarette smoke extract-induced heme oxygenase-1 by cleaving sirtuin 1," *The Journal of Biological Chemistry*, vol. 292, no. 28, pp. 11970–11979, 2017.
- [20] W. Liu, Y. Wang, C. Zhang, B. Huang, J. Bai, and L. Tian, "Cullin1 is up-regulated and associated with poor patients' survival in hepatocellular carcinoma," *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 4, pp. 4001–4007, 2015.
- [21] J. Deng, W. Chen, Y. du et al., "Synergistic efficacy of cullin1 and MMP-2 expressions in diagnosis and prognosis of colorectal cancer," *Cancer Biomarkers*, vol. 19, no. 1, pp. 57–64, 2017.
- [22] V. Benesova, V. Kinterova, J. Kanka, and T. Toralova, "Characterization of SCF-complex during bovine preimplantation development," *PLoS One*, vol. 11, no. 1, 2016.
- [23] Q. Cheng and G. Yin, "Cullin-1 regulates MG63 cell proliferation and metastasis and is a novel prognostic marker of osteosarcoma," *The International Journal of Biological Markers*, vol. 32, no. 2, pp. e202–e209, 2017.
- [24] L. Zhang, W. Wang, B. Zhu, and X. Wang, "Epithelial mitochondrial dysfunction in lung disease," *Adv. Exp. Med. Biol.*, vol. 1038, pp. 201–217, 2017.
- [25] C. Michaeloudes, P. K. Bhavsar, S. Mumby, K. F. Chung, and I. M. Adcock, "Dealing with stress: defective metabolic adaptation in chronic obstructive pulmonary disease pathogenesis," *Annals of the American Thoracic Society*, vol. 14, Supplement_5, pp. S374–S382, 2017.
- [26] Z. Abdolvahabi, M. Nourbakhsh, S. Hosseinkhani et al., "MicroRNA-590-3P suppresses cell survival and triggers breast cancer cell apoptosis via targeting sirtuin-1 and deacetylation of p53," *Journal of Cellular Biochemistry*, vol. 120, no. 6, pp. 9356–9368, 2019.
- [27] Y. Qian, Z. D. Mao, Y. J. Shi, Z. G. Liu, Q. Cao, and Q. Zhang, "Comprehensive analysis of miRNA-mRNA-lncRNA networks in non-smoking and smoking patients with chronic obstructive pulmonary disease," *Cellular Physiology and Biochemistry*, vol. 50, no. 3, pp. 1140–1153, 2018.
- [28] Q. Tao, Z. Zhang, and Y. Xu, "Apoptosis versus proliferation activities and relative mechanism in chronic obstructive pulmonary disease," *Zhonghua Yi Xue Za Zhi*, vol. 78, no. 8, pp. 574–577, 1998.
- [29] N. Martinet and Y. Martinet, "Application of molecular biology techniques to pneumology," *Revue Des Maladies Respiratoires*, vol. 7, no. 6, pp. 541–550, 1990.
- [30] M. Cabanski, B. Fields, S. Boue et al., "Transcriptional profiling and targeted proteomics reveals common molecular changes associated with cigarette smoke-induced lung emphysema development in five susceptible mouse strains," *Inflammation Research*, vol. 64, no. 7, pp. 471–486, 2015.
- [31] M. M. Musri, N. Coll-Bonfill, B. A. Maron et al., "MicroRNA dysregulation in pulmonary arteries from chronic obstructive pulmonary disease. Relationships with vascular remodeling," *American Journal of Respiratory Cell and Molecular Biology*, vol. 59, no. 4, pp. 490–499, 2018.
- [32] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, pp. D991–D995, 2013.
- [33] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. D1, pp. D447–D452, 2015.
- [34] Y. Yi, Y. Zhao, C. Li et al., "RAID v2.0: an updated resource of RNA-associated interactions across organisms," *Nucleic Acids Research*, vol. 45, no. D1, pp. D115–D118, 2017.
- [35] H. Han, J. W. Cho, S. Lee et al., "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Research*, vol. 46, no. D1, pp. D380–D386, 2018.
- [36] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1091–D1097, 2013.
- [37] M. E. Ritchie, B. Phipson, D. Wu et al., "limma powers differential expression analyses for RNA-sequencing and

- microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.
- [38] G. K. Smyth, “Linear models and empirical Bayes methods for assessing differential expression in microarray experiments,” *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–25, 2004.
- [39] P. Shannon, A. Markiel, O. Ozier et al., “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [40] T. Nepusz, H. Yu, and A. Paccanaro, “Detecting overlapping protein complexes in protein-protein interaction networks,” *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.
- [41] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, “clusterProfiler: an R package for comparing biological themes among gene clusters,” *OMICS*, vol. 16, no. 5, pp. 284–287, 2012.
- [42] B. Mlecnik, J. Galon, and G. Bindea, “Comprehensive functional analysis of large lists of genes and proteins,” *Journal of Proteomics*, vol. 171, pp. 2–10, 2018.

Research Article

Interpretable Learning Approaches in Resting-State Functional Connectivity Analysis: The Case of Autism Spectrum Disorder

Jinlong Hu ^{1,2}, Lijie Cao ^{1,2}, Tenghui Li ^{1,2}, Bin Liao ³, Shoubin Dong ^{1,2} and Ping Li⁴

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

²Communication and Computer Network Laboratory of Guangdong, South China University of Technology, Guangzhou, China

³College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China

⁴Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong, China

Correspondence should be addressed to Jinlong Hu; jlhu@scut.edu.cn

Received 26 March 2020; Accepted 5 May 2020; Published 18 May 2020

Guest Editor: Lin Lu

Copyright © 2020 Jinlong Hu 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.

Deep neural networks have recently been applied to the study of brain disorders such as autism spectrum disorder (ASD) with great success. However, the internal logics of these networks are difficult to interpret, especially with regard to how specific network architecture decisions are made. In this paper, we study an interpretable neural network model as a method to identify ASD participants from functional magnetic resonance imaging (fMRI) data and interpret results of the model in a precise and consistent manner. First, we propose an interpretable fully connected neural network (FCNN) to classify two groups, ASD versus healthy controls (HC), based on input data from resting-state functional connectivity (rsFC) between regions of interests (ROIs). The proposed FCNN model is a piecewise linear neural network (PLNN) which uses piecewise linear function LeakyReLU as its activation function. We experimentally compared the FCNN model against widely used classification models including support vector machine (SVM), random forest, and two new classes of deep neural network models in a large dataset containing 871 subjects from ABIDE I database. The results show the proposed FCNN model achieves the highest classification accuracy. Second, we further propose an interpreting method which could explain the trained model precisely with a precise linear formula for each input sample and decision features which contributed most to the classification of ASD versus HC participants in the model. We also discuss the implications of our proposed approach for fMRI data classification and interpretation.

1. Introduction

Autism spectrum disorder (ASD) is a subtype of extensive developmental disorder which is characterized by reciprocal social communication impairment as well as repetitive, restricted, and stereotyped behaviors [1]. The cause of ASD is uncertain, and the diagnosis is often difficult since the expressions of ASD symptoms are diverse and may vary over the course of development [2]. Functional magnetic resonance imaging (fMRI) is one of the most widespread approaches which is noninvasive and useful for understanding brain function [3]. fMRI has been applied recently for distinguishing ASD patients from healthy controls, and vari-

ous machine learning methods have been used to analyze fMRI data of brain disorder [4–7]. However, so far, it has been challenging to analyze fMRI data for brain disorder due to the data characteristics such as high dimensionality, structural complexity, nonlinear separability, and the sequential changes of traceable signals in each voxel [8].

Given the excellent learning capability and classification performance in many domains, deep learning methods have been recently applied to fMRI data from ASD patients [9–14]. Sólón et al. [9] investigated the patterns of functional connectivity that help to identify ASD participants from functional brain imaging data. They used stacked denoising autoencoders for the unsupervised pretraining

stage to extract a low-dimensional version from the ABIDE database and then applied the encoder weights to a multilayer perceptron for classification. The ABIDE (Autism Brain Imaging Data Exchange) database [15] contains a rich set of fMRI data that has aggregated functional and structural brain imaging data collected from multisite around the world (see Section 2.1 below for details). Guo et al. [10] stacked multiple sparse autoencoders for data dimension reduction and developed a feature selection method to select features with high discriminative power. Then, they used a softmax regression on top of the stacked sparse autoencoders for data classification. Eslami et al. [11] used an autoencoder and a single-layer perceptron to extract lower dimensional features, and the trained perceptron is used for the final round of classification. Brown et al. [12] proposed an element-wise layer based on BrainNetCNN [16] and used anatomically informed, data dependent, prior to regularize the weights of the layer.

Researchers are also trying to explain these models, by analyzing the discriminative features or potential neuroimaging biomarkers that contribute to the classification of ASD from healthy controls. Li et al. [17] trained a deep neural network to classify 3D fMRI volumes, developed a frequency-normalized sampling method to replace a ROI of the original image with the sampling data, and put it in the trained model to get a new prediction. Based on the different predicting performance, they used a statistical method to interpret the importance of the ROI. In the study of discovering imaging biomarkers for ASD [18], they went beyond looking at only individual features by using Shapley value explanation on interactive features' prediction power analysis. Guo et al. [10] proposed a deep neural network with a feature selection method from multiple trained sparse autoencoders, then developed Fisher's score-based biomarker identification method for their deep neural network using the rs-fMRI dataset in ABIDE I. These approaches all led to useful insights into the mechanism of deep learning models. However, such deep and nonlinear models are usually constructed as black boxes with complex network structure and hidden internal logic and are difficult to interpret with regard to how architecture decisions are consistently made by researchers [19].

In this study, we introduce an interpretable learning approach for resting-state functional connectivity analysis. We firstly propose an interpretable neural network model to distinguish between ASD participants and healthy controls (HC) based on resting-state functional connectivity (rsFC) of each subject. The proposed model is an interpretable fully connected neural network (FCNN), which uses piecewise linear function LeakyReLU as its activation function. It is a fully connected neural network including two hidden layers, input layer and output layer. Further, the proposed model is a piecewise linear neural network (PLNN) [20], which is mathematically equivalent to a set of local linear classifiers and could be interpreted precisely and consistently [19]. Secondly, taking advantage of the interpretation of PLNN, we propose an interpretable method which could explain the trained classification model with a precise linear formula for each input sample

and the decision features which contribute most to classify ASD versus HC in the model.

We experimentally compared the proposed FCNN model against widely used benchmark models including SVM, random forest (RF), and two new neural network models in classifying data from the multisite ABIDE I database [15]. The proposed FCNN model, based on input data from rsFC between regions of interests (ROIs) accord to the AAL atlas [21], achieved the highest accuracy 69.81% in the large dataset containing 871 subjects (403 ASD patients and 468 healthy controls). We also explained the most important features in the model.

2. Dataset and Preprocessing

2.1. Dataset. We chose the dataset from the Autism Brain Imaging Data Exchange (ABIDE) initiative [15] to confirm the approach proposed in this study. The ABIDE initiative has aggregated functional and structural brain imaging data collected from multiple sites around the world. The dataset used in this study contained 871 subjects acquired from 17 acquisition sites with different imaging protocols that met the imaging quality and phenotypic information criteria [22]. This dataset includes 403 individuals suffering from ASD and 468 healthy controls (HC).

2.2. Preprocessing. We downloaded the preprocessed resting-state fMRI data from the Preprocessed Connectomes Project (PCP) (<http://preprocessed-connectomes-project.org/abide/download.html>). The data [23] was preprocessed by the Configurable Pipeline for the Analysis of Connectomes (CPAC) pipeline that included the following procedure: slice timing correction, motion realignment, intensity normalization, regression of nuisance signals, band-pass filtering (0.01-0.1 Hz), and registration of fMRI images to standard anatomical space (MNI152). The detailed description of pipeline can be found at <http://preprocessed-connectomes-project.org/abide/Pipelines.html>. The data was parcellated into 116 regions of interests (ROIs) using the AAL atlas [21].

3. Proposed Approach

The flow chart of the proposed interpretable learning approach is shown in Figure 1. First, we propose the FCNN model for classifying ASD and healthy participants, including extracting the rsFC features, training the FCNN model, and validating the model. Second, we interpret the trained model with an easily explained linear formula for each subject, identifying the decision rsFC features for the ASD group from the data.

3.1. Feature Extraction. The resting-state fMRI data was preprocessed as described in Section 2. The brain was parcellated into 116 regions of interests (ROIs) according to the AAL atlas [21]. Then, the mean time series of each ROI was extracted for each subject, and the rsFCs between ROIs were measured by computing Pearson's correlation coefficient of the extracted time series. A 116×116 connectivity matrix was constructed for each subject, respectively.

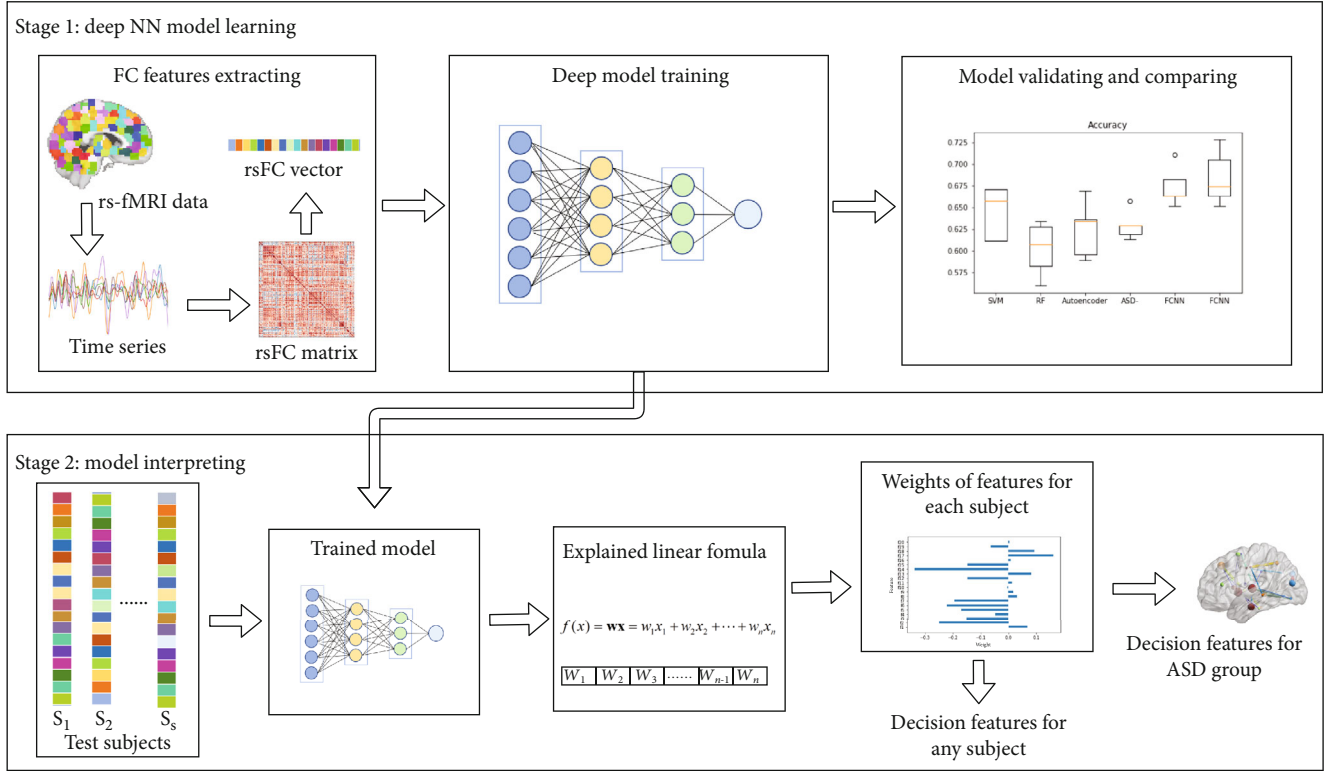


FIGURE 1: Flow chart of the proposed approach: learning and interpreting model on resting-state fMRI data.

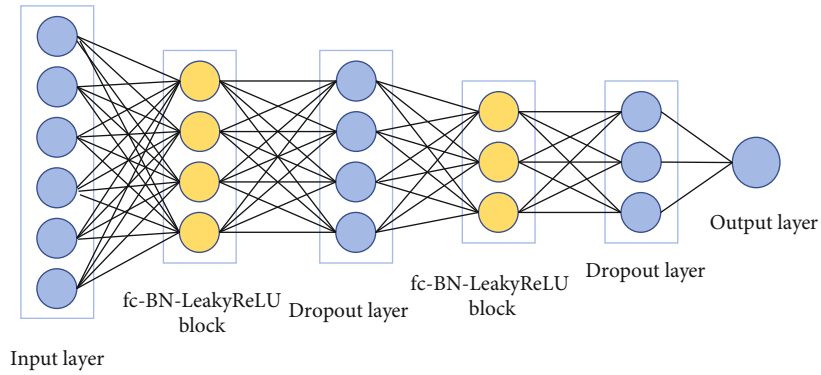


FIGURE 2: The architecture of the proposed FCNN model.

Fisher transformation was applied to the connectivity matrices to improve normality. The upper triangle values were then extracted and flattened into vectors, with the dimension of the feature vector which is $(116 \times (116 - 1)) / 2 = 6670$.

3.2. FCNN Model. The architecture of the proposed FCNN model is shown in Figure 2. The FCNN is a fully connected neural network and a piecewise linear neural network (PLNN), where the PLNN is a deep neural network in which the nonlinear activation function is a piecewise linear function with a constant number of pieces [20].

The FCNN model contains two fc-BN-LeakyReLU blocks, where the fc-BN-LeakyReLU block consists of a fully

connected (fc) layer followed by a Batch Normalization (BN) layer and LeakyReLU activation function.

LeakyReLU is a variant of rectified linear unit (ReLU) [24] which allows a small, positive gradient when the unit is not active [25]. For each hidden neuron u , LeakyReLU is defined as

$$f(u) = \begin{cases} u, & u \geq 0, \\ \alpha u, & u < 0, \end{cases} \quad (1)$$

where α represents slope coefficient. LeakyReLU is clearly a piecewise linear function.

In this study, for simplicity and clarity, we regarded a fc-BN-LeakyReLU block as a hidden layer. For a model with L

layers, a fc layer can be formulated as

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)}, \quad (2)$$

where $l \in \{1, \dots, L-1\}$; suppose there are n neurons in layer l and m neurons in layer $l+1$, $W^{(l)}$ is $m \times n$ weight matrix, $b^{(l)}$ is $m \times 1$ bias vector, and $a^{(l)}$ will be in Equation (3).

Then, the fc-BN-LeakyReLU block can be written as

$$a^{(l)} = f\left(\text{BN}\left(z^{(l)}\right)\right), \quad (3)$$

where $l \in \{2, \dots, L-1\}$ are hidden layers, $f(\bullet)$ is the LeakyReLU function, explicitly, and $a^{(1)}$ is the input instance x .

The sigmoid function is applied on the output layer to predict the probability of any given participant being an ASD patient. The number of units (nodes) is 6670, 64, 32, and 1, respectively, for input layer, two fully connected layers, and output layer. The dropout layer is added to avoid data overfitting, and the loss function uses binary cross entropy.

3.3. Interpreting Method. We interpret the trained neural network model with two stages: (i) computing the decision boundary of a fixed instance and the weight of features in linear formula for the instance and (ii) extracting and analyzing decision features of the trained model in the ASD group level.

In the first stage, we computed the decision boundary of a fixed instance x .

For each hidden neuron u , BN can be formulated as

$$y = \frac{\gamma}{\sqrt{\text{Var}[u] + \epsilon}} \bullet u + \left(\beta - \frac{\gamma E[u]}{\sqrt{\text{Var}[u] + \epsilon}} \right), \quad (4)$$

where γ and β are learned parameters [26]. In the test phase of the model, $\text{Var}[u]$ and $E[u]$ are fixed, so Equation (4) can be regarded as a linear function.

As shown in Equation (1), for hidden neurons with LeakyReLU activation function, there are two kinds of activation status that each corresponds to a corresponding linear function where the mapping relationship between $f(u)$ and u is linear. And it is proved that for a fixed PLNN model and a fixed instance x , the output of model $F(x)$ on an input x can be seen as a linear classifier [19], which can be formulated as

$$F(x) = \text{sigmoid}\left(\widehat{W}x + \widehat{b}\right), \quad (5)$$

where $\widehat{W} = \prod_{h=0}^{L-3} \widetilde{W}^{(L-1-h)} W^{(1)}$ is the coefficient vector of x and \widehat{b} is the constant intercept. For a fixed input instance x , $F(x)$ is a linear classifier whose decision boundary is explicitly defined by $\widehat{W}x + \widehat{b}$. Therefore, \widehat{W} are weights assigned to the features of x .

As for FCNN, we computed \widetilde{W} as follows: since BN can be regarded as a linear function in the test phase of model as discussed above, the Equation (3) can be rewritten

as

$$a^{(l)} = f\left(\widetilde{\gamma}^{(l)} \circ z^{(l)} + \widetilde{\beta}^{(l)}\right), \quad (6)$$

where $\widetilde{\gamma}^{(l)}$ is the constant slope, $\widetilde{\beta}^{(l)}$ is the constant intercept, for all $l \in \{2, \dots, L-1\}$. Since $f(\bullet)$ is the piecewise linear activation function, Equation (6) can be rewritten as

$$a^{(l)} = r^{(l)} \circ \left(\widetilde{\gamma}^{(l)} \circ z^{(l)} + \widetilde{\beta}^{(l)}\right) + t^{(l)}, \quad (7)$$

where $r^{(l)}$ is the constant slope and $t^{(l)}$ is the constant intercept. By plugging Equation (7) into Equation (2), we rewrite $z^{(l+1)}$ as

$$\begin{aligned} z^{(l+1)} &= W^{(l)} \left(r^{(l)} \circ \left(\widetilde{\gamma}^{(l)} \circ z^{(l)} + \widetilde{\beta}^{(l)}\right) + t^{(l)} \right) + b^{(l)} \\ &= \widetilde{W}^{(l)} z^{(l)} + \widetilde{b}^{(l)}, \end{aligned} \quad (8)$$

where $\widetilde{W}^{(l)} = W^{(l)} \circ r^{(l)} \circ \widetilde{\gamma}^{(l)}$ is an extended version of the Hadamard product.

In the second stage, based on the weights \widehat{W} for features of each test instance x , we could get the top K features with the highest weight. Then, we count the number of occurrences n^f of feature f in the top-k-feature-set from all the instances. By setting a threshold on n^f , we can get decision feature set F which contributes most to classify ASD versus HC in the model.

The whole flow of the interpreting method is formulated as in Algorithm 1. We firstly obtain the top-k-feature-set F_K^x for each instance x , and then, we obtain the decision feature set F by selecting the feature f whose occurrence number as a percentage of total instances is greater to the parameter ϵ . Meanwhile, we could also get the weights of all features for any specified test instance, which could help to explain the decision made by the trained model for the instance.

4. Classification Experiments

With the above approach and the model architecture, we conducted experiments on the ABIDE I dataset with 871 subjects and applied the interpretation algorithm to explain the results.

To evaluate the performance of the proposed method, we use sensitivity, specificity, accuracy, F1, and AUC as our metrics. These metrics are defined as follows:

$$\begin{aligned} \text{sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}, \\ \text{F1} &= \frac{2 \bullet \text{TP}}{2 \bullet \text{TP} + \text{FP} + \text{FN}}, \end{aligned} \quad (9)$$

Input: a well-trained FCNN; the set of test instances D ; parameter K , the number of top important features of instance x ; parameter ϵ , number of occurrences of feature f as a percentage of total instances.

Output: decision feature set F

1. Initialization: $F = \emptyset$, $F_K^D = \emptyset$
2. For each $x \in D$ do
3. Compute the weight \widehat{W}
4. Get the K top features F_K^X with the highest weight
5. $F_K^D \leftarrow F_K^D \cup F_K^X$
6. End for
7. For each feature f in F_K^D
8. Count the number of occurrences n^f of feature f
9. If $n^f > |D| * \epsilon$
10. $F \leftarrow F \cup f$
11. End for
12. Return F

ALGORITHM 1. A run-down flow for trained model interpreting.

TABLE 1: Classification performance using 5-fold cross-validation (mean \pm std).

	Accuracy	Sensitivity	Specificity	F1	AUC
SVM-linear	0.6441 \pm 0.0281	0.5856 \pm 0.0238	0.6946 \pm 0.0556	0.6039 \pm 0.0219	0.7053 \pm 0.0372
SVM-rbf	0.6624 \pm 0.0283	0.5631 \pm 0.0623	0.7478 \pm 0.0629	0.6055 \pm 0.0403	0.7059 \pm 0.0283
RF	0.6326 \pm 0.0416	0.4590 \pm 0.0428	0.7821 \pm 0.0442	0.5364 \pm 0.0506	0.6790 \pm 0.0339
Autoencoder+MLP [9]	0.6717 \pm 0.0217	0.6225 \pm 0.1601	0.7140 \pm 0.1124	0.6259 \pm 0.0784	0.6682 \pm 0.0293
ASD-DiagNet [11]	0.6900 \pm 0.0172	0.6277 \pm 0.0642	0.7436 \pm 0.0299	0.6504 \pm 0.0338	0.6857 \pm 0.0201
FCNN (without BN)	0.6889 \pm 0.0109	0.6204 \pm 0.0844	0.7479 \pm 0.0624	0.6456 \pm 0.0378	0.7099 \pm 0.0227
FCNN	0.6981 \pm 0.0169	0.6305 \pm 0.0474	0.7563 \pm 0.0182	0.6582 \pm 0.0287	0.7262 \pm 0.0308

where TP is defined as the number of ASD subjects that are correctly classified, FP is the number of normal subjects that are misclassified as ASD subjects, TN is defined as the number of normal subjects that are correctly classified, and FN is defined as the number of ASD subjects that are misclassified as normal subjects. Specifically, sensitivity measures the proportion of ASD subjects that are correctly identified as such; specificity measures the proportion of normal subjects that are correctly identified as such. AUC is defined as the area under the Receiver Operating Characteristic (ROC) curve.

4.1. Comparison Models. Given the above FCNN model, we use the following models as benchmarks for comparison.

SVM: support-vector machine (SVM) model with linear kernel and rbf kernel. The SVM method has been widely used to classify fMRI data for brain disorders. The parameters are chosen by grid search.

RF: random forest (RF) is an ensemble learning method for classification. The parameters are chosen by grid search.

Autoencoder+MLP: the model was proposed by Solon et al. [9]. Two stacked denoising autoencoders are pretrained; then, the encoder weights are applied to a multilayer perceptron (MLP), and the MLP is fine tuned to predict the proba-

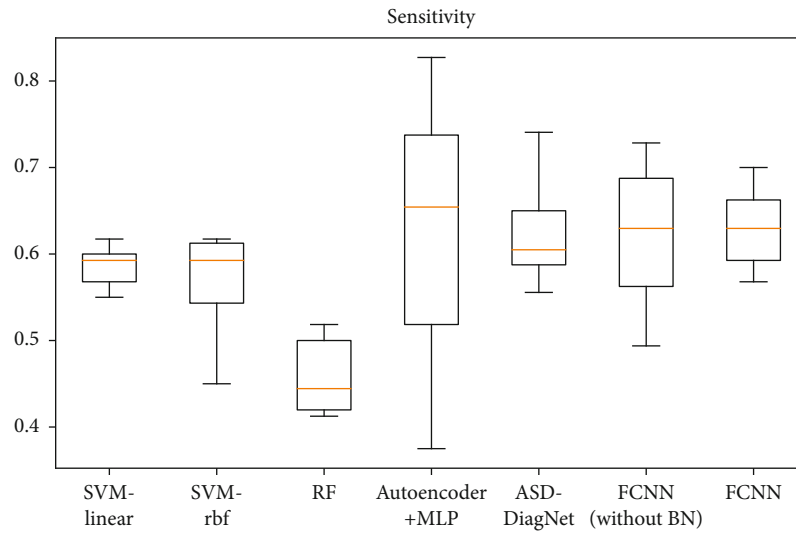
bility of any given participants being ASD. We applied the encoder weights to the MLP with the configuration: 6670-1000-600-2.

ASD-DiagNet: this method is proposed by Eslami et al. [11]. An autoencoder is used to extract a lower dimensional feature representation. Then, the feature representation is fed into a single-layer perceptron (SLP) with sigmoid function for classification. The autoencoder and SLP classifier are trained simultaneously. The input layer and output layer have 6670 units fully connected to a bottleneck of 1667 units from the hidden layer. Data augmentation using EROS similarity measure is applied with 5 nearest neighbors of each sample.

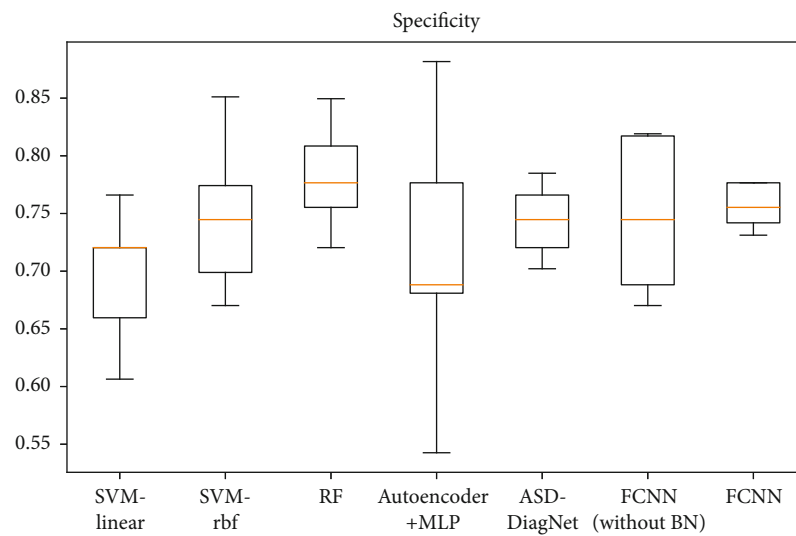
FCNN: the proposed FCNN model as described above in Figure 2. The model contains two fully connected layers: the first layer has 64 units and the second layer has 32 units. The dropout ratio is set to 0.8. We used the Adam optimizer with a learning rate of 0.0005.

For autoencoder+MLP [9] and ASD-DiagNet [11], we used their online code to evaluate the models.

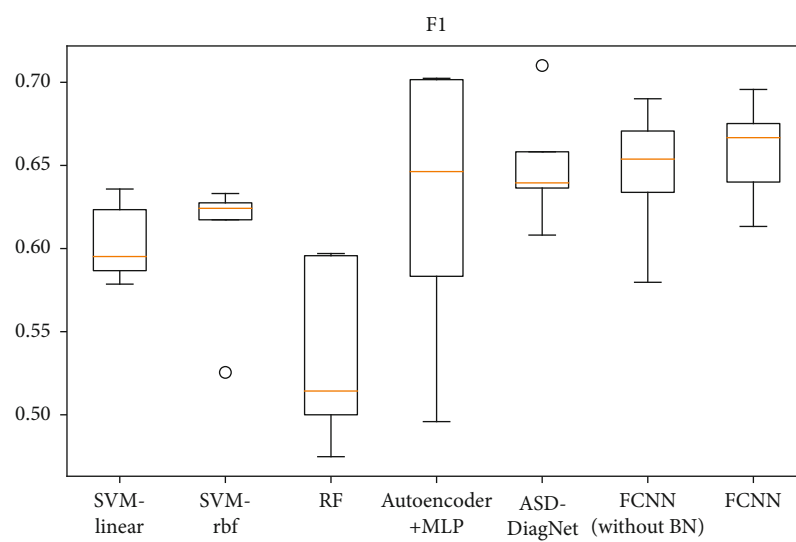
All functional connectivity features are flattened into one dimensional vector (see Figure 1), and the vectors are inputs in all model for training and classification. All the models were trained with 6670 functional connectivity features for each subject. We employed a 5-fold cross-validation setting



(a)



(b)



(c)

FIGURE 3: Continued.

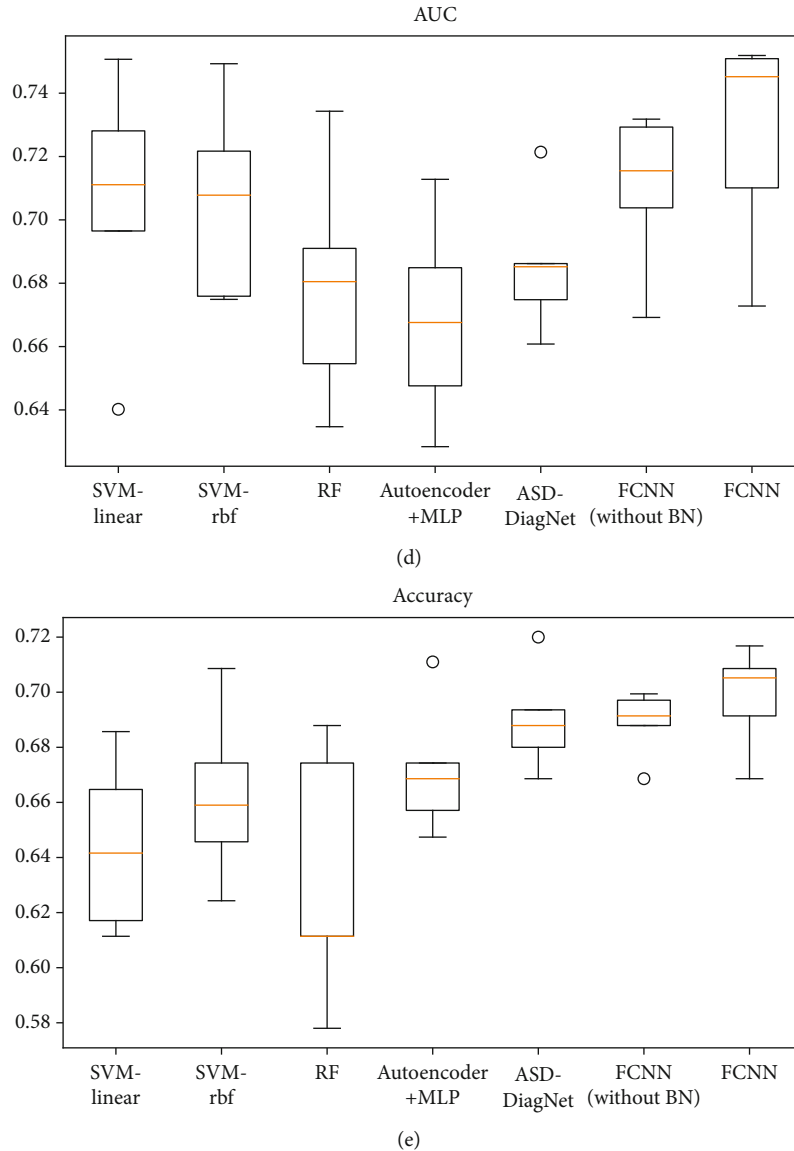


FIGURE 3: (a) Sensitivity, (b) specificity, (c) F1, (d) AUC, and (e) accuracy for classification task.

to evaluate the performance of all the models. The experiments were carried out on all 871 subjects including both ASD patients and healthy controls.

4.2. Classification Results. The classification results are shown in Table 1 and Figure 3. Box plots for sensitivity, specificity, F1, AUC, accuracy for classification task using 5-fold cross-validation are shown in Figure 3, where the middle line in each box represents the median value, and the circle represents the outlier.

The proposed FCNN model achieved the best performance on most evaluation metrics with accuracy of 69.81%, sensitivity of 63.05%, specificity of 75.63%, F1 of 65.82%, and AUC of 0.7262. The results showed that the deep learning models (FCNN, autoencoder+MLP, and ASD-DiagNet) have the better classification performance

in general than the traditional methods (SVM and RF) on the resting-state fMRI dataset. As for the method autoencoder+MLP [9], we would like to mention that they reported 70% accuracy in their paper; the performance we reported is not as good as theirs, maybe because the brain atlas we used is different.

We also compared the FCNN model with or without the BN (Batch Normalization) layer in Table 1. The results showed that the BN layer improves the performance and stability of the model.

5. Interpretation Experiments and Analysis

5.1. Model Interpreting for an Instance. According to Section 3.3, for a trained FCNN model and any instance x with n features, $x = \{x_1, x_2, \dots, x_{n-1}, x_n\}$, the fixed model can be

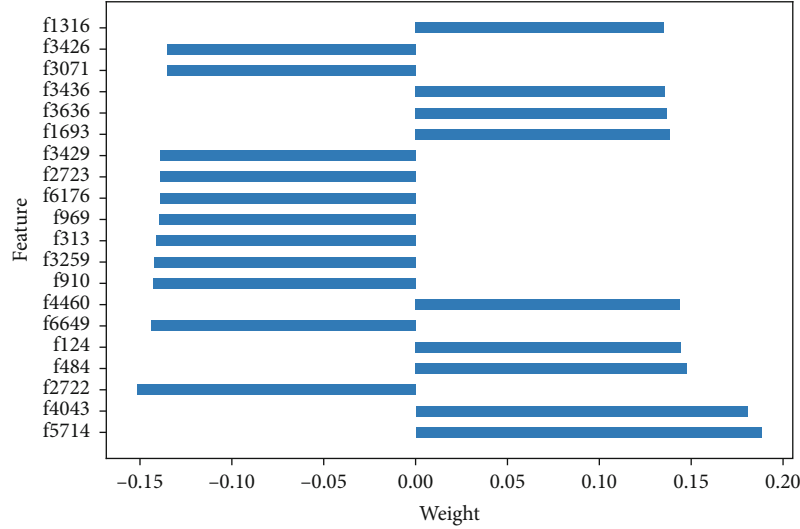


FIGURE 4: Weight visualization of some features of an instance.

formulated as a linear classifier with a fixed instance:

$$g(x) = \widehat{W}x = w_1x_1 + w_2x_2 + \dots + w_nx_n. \quad (10)$$

Since the number of layers L is 4 for the FCNN model we used in this paper, so the weight vector \widehat{W} can be computed as

$$\widehat{W} = \left(W^{(3)} \circ r^{(3)} \circ \tilde{y}^{(3)} \right) \left(W^{(2)} \circ r^{(2)} \circ \tilde{y}^{(2)} \right) W^{(1)}. \quad (11)$$

The trained model could be interpreted with linear formula for any instance. Given an instance, we can get the weight of each feature from the trained model according to Equations (10) and (11). Some feature weights of an instance are visualized in Figure 4. The vertical axis represents the feature index, and the horizontal axis represents the weight value. It can help to understand the prediction result according to the feature index which can correspond to the brain region involved in the feature.

5.2. Model Interpreting for the ASD Group. Based on the trained FCNN model, we used Algorithm 1 as described in Section 3.3 to extract the decision features of the model. We set the top-important feature parameter K from 5 to 300, with an interval of 5, and the parameter ε as 95%, and then, we get a set of decision features with different K .

5.2.1. Decision Feature Evaluation. To evaluate the quality of the decision features, we analyzed the FCNN model by setting the values of the decision features in instance x to zero and observed the changes of prediction of FCNN. We used metrics including sensitivity, accuracy, and the change of prediction probability (CPP) which is the absolute change of probability of classifying x as a positive instance, the number of label-changed instance (NLCI) which is the number of instances whose predicted label changes after being hacked.

For comparison, we also used the top N weighted features of linear-SVM to hack linear-SVM. The results are shown in Figure 5. It is shown that average CPP of FCNN is higher, and the NLCI of FCNN can be more than SVM with more decision features. And FCNN has considerable performance in sensitivity and accuracy.

For further comparison, we also applied the popular locally linear interpretation method (LIME) [27] to get the decision features in the trained FCNN model. Similar to Algorithm 1 in Section 3.3, we obtain the top K important features of each instance, and then, we obtain the decision feature set F by selecting the feature f whose occurrence number as a percentage of total instances is greater to the parameter ε . We set the same parameters (K from 5 to 300, with an interval of 5, and the parameter ε as 95%), and we did not obtain any decision feature. What is more, when we loosed the parameter ε to 20%, we also did not get any one feature. It means that the top 300 important features of the instance obtained by the LIME method are very different between instances in this model.

5.2.2. Decision Feature Analysis. When K is taken as 20, 15 decision features were obtained; we selected these 15 decision features as a case for further analysis. There are 23 brain regions (ROIs) of the AAL atlas that involved these 15 rsFC connections. These 15 rsFCs and 23 ROIs are shown in Table 2.

We computed the mean value of each rsFC of the ASD group and the HC group, respectively, as well as the mean difference of two groups. An independent two-sample t test was run on the means of the rsFC elements of two groups. The analysis is shown in Table 2. Among these 15 rsFCs, 2 rsFCs are statistically significant ($p < 0.05$) between the ASD and HC groups, and the rest of rsFCs are not statistically significant. It demonstrates that FCNN could find underlying features though the feature values are not statistically different between groups.

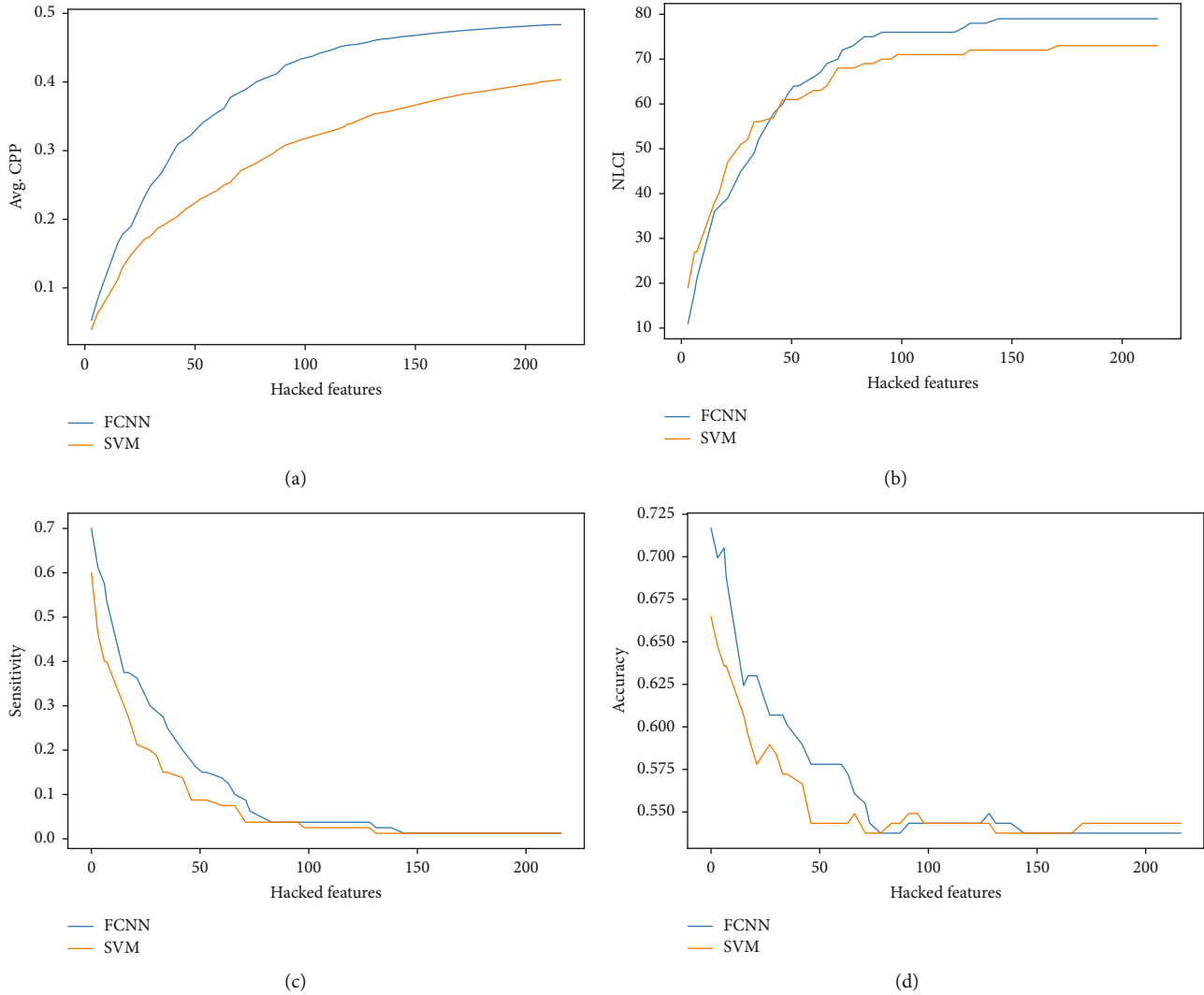


FIGURE 5: The performance of decision features on FCNN and SVM.

These 15 rsFC connections of the AAL atlas are visualized in Figure 6, where the label information is from the AAL atlas. The thicker connection indicates two regions are strongly correlated and vice versa. The figure was drawn with BrainNet Viewer [28] software.

5.2.3. Impact of Parameter ε . In order to evaluate the influence of parameter ε on the obtained decision features, we set the parameter K from 5 to 300, with an interval of 5, and the parameter ε from 70% to 95%, with an interval of 5%; then, N decision features were obtained accordingly. The result is shown in Figure 7. It is clear that the smaller the parameter ε , the more decision features will be obtained. While with a fixed K , the bigger the parameter ε , the fewer the decision features will be obtained.

6. Conclusion and Discussion

In this paper, we introduce an interpretable learning approach for resting-state functional connectivity analysis.

We firstly propose an interpretable FCNN to classify ASD from HC, based on rsFC features. We experimentally compared the FCNN model against widely used classification models including SVM, RF, and two new classes of deep neural network models in a large dataset containing 871 subjects from ABIDE I database. The results show the proposed FCNN model achieves the highest classification accuracy 69.81%.

We further propose an interpreting method which could explain the trained model with a precise linear formula for each input instance and identify decision features of the model which contributed most to the classification of ASD versus HC participants.

Though being focused on ASD analysis in this presentation, the proposed approach could be generalized to benefit many other brain science and medicine applications that involve deep neural networks. Particularly, this study offers a promising deep learning-based approach to explore potential biomarkers for assisting brain neurological disorder diagnosis and research.

TABLE 2: Analysis of 15 most significant rsFCs.

Connection ID	ROI number	Regions	ASD mean conn	Control mean conn	Mean difference	<i>p</i> value
1	72	Caudate_R	0.0919	0.0728	0.0192	0.4390
	107	Cerebelum_10_L				
2	44	Calcarine_R	0.7370	0.7256	0.0114	0.4920
	46	Cuneus_R				
3	2	Precentral_R	0.5996	0.5474	0.0522	0.0325
	12	Frontal_Inf_Oper_R				
4	50	Occipital_Sup_R	0.7175	0.7136	0.0038	0.8445
	52	Occipital_Mid_R				
5	5	Frontal_Sup_Orb_L	0.2666	0.2309	0.0357	0.1985
	36	Cingulum_Post_R				
6	16	Frontal_Inf_Orb_R	0.4435	0.4311	0.0124	0.6172
	90	Temporal_Inf_R				
7	13	Frontal_Inf_Tri_L	0.4219	0.4192	0.0027	0.9201
	16	Frontal_Inf_Orb_R				
8	6	Frontal_Sup_Orb_R	0.5359	0.4989	0.0371	0.2355
	26	Frontal_Med_Orb_R				
9	44	Calcarine_R	0.6847	0.6632	0.0215	0.3427
	50	Occipital_Sup_R				
10	64	SupraMarginal_R	0.3853	0.3586	0.0267	0.3485
	69	Paracentral_Lobule_L				
11	38	Hippocampus_R	0.2871	0.2618	0.0253	0.3651
	66	Angular_R				
12	36	Cingulum_Post_R	0.2296	0.2110	0.0185	0.5694
	43	Calcarine_L				
13	36	Cingulum_Post_R	0.2640	0.2474	0.0167	0.6089
	44	Calcarine_R				
14	38	Hippocampus_R	0.4710	0.4123	0.0586	0.0377
	86	Temporal_Mid_R				
15	68	Precuneus_R	0.3425	0.3111	0.0314	0.2958
	81	Temporal_Sup_L				

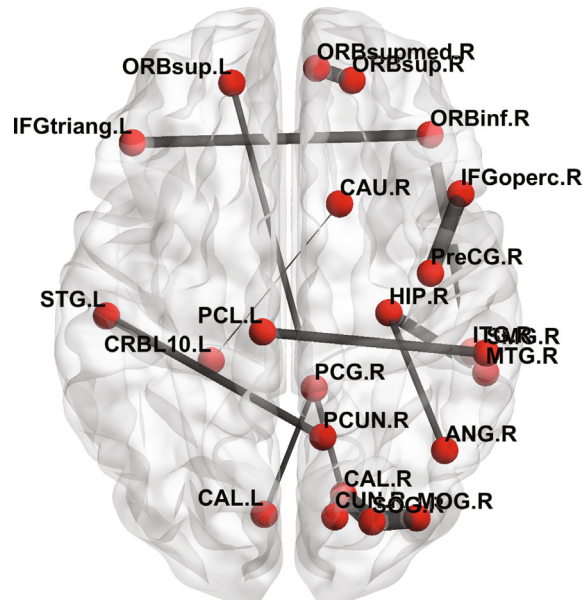


FIGURE 6: The visualization of 15 rsFCs from the ASD group.

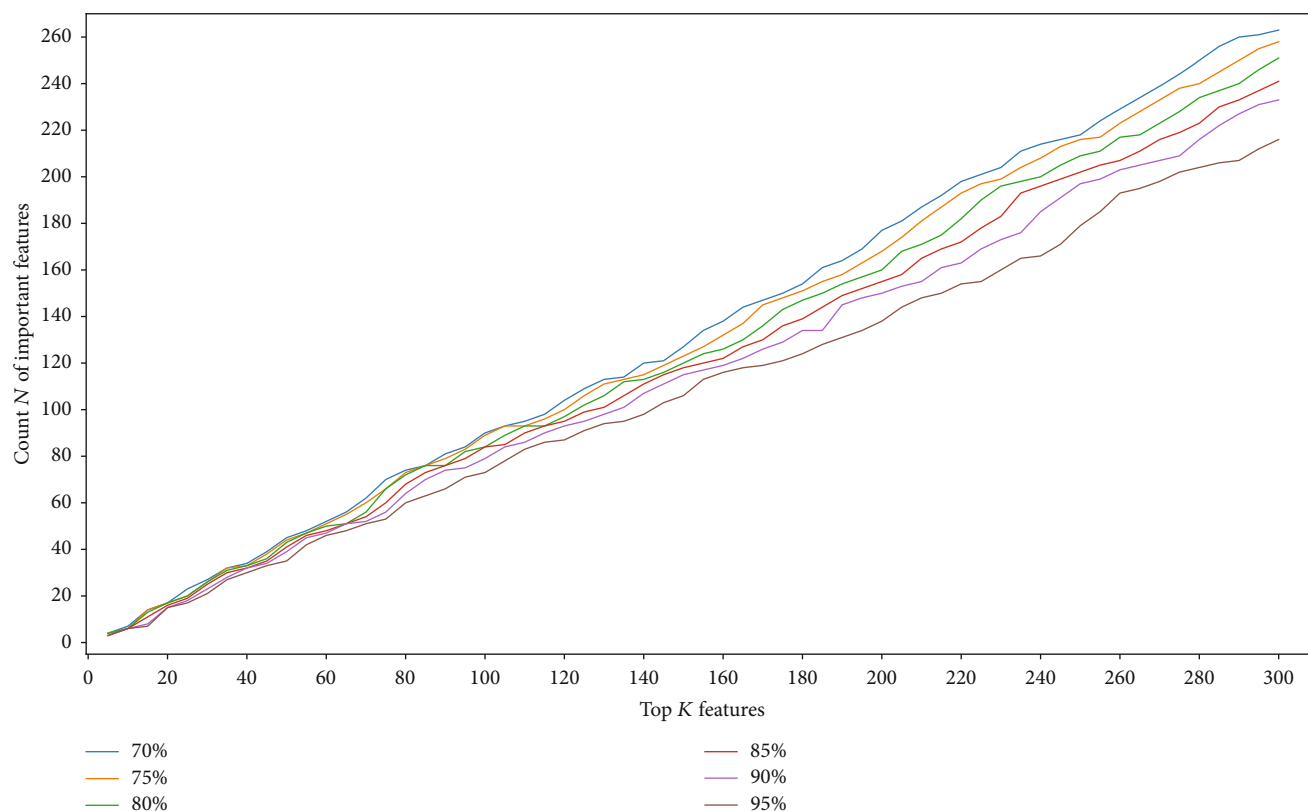


FIGURE 7: The number of decision features with different K and ϵ .

There are two limitations in the current work presented here. First, the dataset is limited to the 871 participants that contained ASD and HC. In order for this work to be more generalizable, it would be important to inspect and compare these initial findings with more fMRI data from more participants. Second, the proposed model is a compact fully connected neural network, given the number of layers and nodes in the model. Thus, it would be important to inspect the effectiveness of our interpreting approach for other types of neural network such as deeper and more complex architectures in the deep learning literature. Future work should focus on the accuracy and interpretation of our proposed approach for other large-scale fMRI data as well as other neuroimaging data based on brain disorders such as ASD.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request. The ABIDE I dataset analyzed during this study is available in the Preprocessed Connectomes Project website (<http://preprocessed-connectomes-project.org/abide/download.html>).

Conflicts of Interest

The authors have nothing to disclose.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Guangdong Province of China (grants #2018A030313309 and #2015A030308017), the Innovation Fund of Introduced High-End Scientific Research Institutions of Zhongshan (grant #2019AG031), the Fundamental Research Funds for the Central Universities, SCUT (grant #2019KZ20), and the Guangdong Pearl River Talents Plan Innovative and Entrepreneurial Team (grant #2016ZT06S220).

References

- [1] C. M. Freitag and T. A. Jarczok, "Autism spectrum disorders," in *Psychiatric Drugs in Children and Adolescents*, pp. 383–403, Springer, Vienna, 2014.
- [2] F. Volkmar, E. H. Cook, J. Pomeroy, G. Realmuto, and P. Tanguay, "Practice parameters for the assessment and treatment of children, adolescents, and adults with autism and other pervasive developmental disorders," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 38, no. 12, pp. 32S–54S, 1999.
- [3] S. Takerkart, G. Auzias, B. Thirion, and L. Ralaivola, "Graph-based inter-subject pattern analysis of fMRI data," *PLoS One*, vol. 9, no. 8, p. e104586, 2014.
- [4] R. Bhaumik, A. Pradhan, S. Das, and D. K. Bhaumik, "Predicting Autism Disorder Using Domain-Adaptive Cross-Site Evaluation," *Neuroinformatics*, vol. 16, no. 2, pp. 197–205, 2018.

- [5] H. Huang, D. Shen, and N. Carolina, "Enhancing the representation of functional connectivity networks by fusing multi-view information for autism spectrum disorder diagnosis," *Human Brain Mapping*, vol. 40, no. 3, pp. 833–854, 2018.
- [6] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage: Clinical*, vol. 7, pp. 359–366, 2015.
- [7] M. A. Just, V. L. Cherkassky, A. Buchweitz, A. Timothy, and T. M. Mitchell, "Identifying autism from neural representations of social interactions: neurocognitive markers of autism," *PLoS One*, vol. 9, no. 12, pp. 1–22, 2014.
- [8] B. Cao, X. Kong, and P. S. Yu, "A review of heterogeneous data mining for brain disorder identification," *Brain Informatics*, vol. 2, no. 4, pp. 253–264, 2015.
- [9] A. Sólón, A. Rosa, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.
- [10] X. Guo, K. C. Dominick, A. A. Minai, H. Li, C. A. Erickson, and L. J. Lu, "Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method," *Frontiers in neuroscience*, vol. 11, 2017.
- [11] T. Eslami, V. Mirjalili, A. Fong, A. Laird, and F. Saeed, "ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data," *Frontiers in Neuroinformatics*, vol. 13, pp. 1–8, 2019.
- [12] C. J. Brown, J. Kawahara, and G. Hamarneh, "Connectome priors in deep neural networks to predict autism," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 110–113, Washington, DC, USA, April 2018.
- [13] R. Anirudh and J. J. Thiagarajan, "Bootstrapping graph convolutional neural networks for autism spectrum disorder classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3197–3201, Brighton, United Kingdom, May 2019.
- [14] X.-a. Bi, Y. Liu, Q. Jiang, Q. Shu, Q. Sun, and J. Dai, "The diagnosis of autism spectrum disorder based on the random neural network cluster," *Frontiers in human neuroscience*, vol. 12, 2018.
- [15] A. Di Martino, C.-G. Yan, Q. Li et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.
- [16] J. Kawahara, C. J. Brown, S. P. Miller et al., "BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.
- [17] X. Li, N. C. Dvornek, J. Zhuang, P. Ventola, and J. S. Duncan, "Brain biomarker interpretation in ASD using deep learning and fMRI," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 206–214, Springer, Cham, 2018.
- [18] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, "Efficient interpretation of deep learning models using graph structure and cooperative game theory: application to asd biomarker discovery," in *International Conference on Information Processing in Medical Imaging*, pp. 718–730, Springer, Cham, 2019.
- [19] L. Chu, X. Hu, J. Hu, L. Wang, and J. Pei, "Exact and consistent interpretation for piecewise linear neural networks: a closed form solution," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1244–1253, 2018.
- [20] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," *Journal of Machine Learning Research*, vol. 20, no. 63, pp. 1–17, 2019.
- [21] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou et al., "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.
- [22] A. Abraham, M. P. Milham, A. di Martino et al., "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example," *NeuroImage*, vol. 147, pp. 736–745, 2017.
- [23] C. Cameron, B. Yassine, C. Carlton et al., "The Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives," *Frontiers in Neuroinformatics*, vol. 7, no. 4, pp. 1–19, 2013.
- [24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pp. 807–814, 2010.
- [25] A. L. Maas and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.
- [26] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, pp. 448–456, 2015.
- [27] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- [28] M. Xia, J. Wang, and Y. He, "BrainNet Viewer: a network visualization tool for human brain connectomics," *PLoS One*, vol. 8, no. 7, 2013.

Retraction

Retracted: Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle

Computational and Mathematical Methods in Medicine

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

Copyright © 2023 Computational and Mathematical Methods in Medicine. 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] W. Lu, Q. Pan, Y. Zhou, W. Chen, H. Zhang, and W. Qi, "Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 6896517, 4 pages, 2020.

Research Article

Development and Application of One Separation-Free Safety Tube on the Disposable Infusion Needle

Weifen Lu,¹ Qianli Pan,¹ Yinxin Zhou,¹ Wenyu Chen,¹ Hongyan Zhang,¹ and Weibo Qi^{1,2} 

¹Department of Respiration, First Hospital of Jiaxing (Affiliated Hospital of Jiaxing University), 314000 Jiaxing, China

²Department of Cardiothoracic Surgery, First Hospital of Jiaxing (Affiliated Hospital of Jiaxing University), 314000 Jiaxing, China

Correspondence should be addressed to Weibo Qi; qiweibo_abcd@163.com

Received 19 February 2020; Revised 15 April 2020; Accepted 27 April 2020; Published 16 May 2020

Guest Editor: Tao Huang

Copyright © 2020 Weifen Lu 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.

Objective. To develop a new type infusion set and apply it to the clinic, as well as explore its effectiveness in the prevention from needle stick injuries. **Methods.** A total of 200 inpatients who were in need of intravenous infusion with a disposable infusion needle were included and randomly divided into two groups: intervention group and control group. Disposable infusion needles with a separation-free safety tube were used in the intervention group, whereas conventional ones were used in the control group. Then, effects of the two types of infusion sets were observed and compared. **Results.** As for the operation time for infusion, it was (82.19 ± 1.80) seconds in the intervention group and (83.02 ± 1.83) seconds in the control group, with the difference statistically significant ($P < 0.05$). Besides, the exposure time of the needles after infusion in the intervention group was (3.36 ± 0.17) seconds while (18.85 ± 1.18) seconds in the control group; the difference between which was statistically significant ($P < 0.05$). In terms of the time for needle disposal, (18.60 ± 0.84) seconds was required in the intervention group, while for the control group, it took (18.85 ± 1.18) seconds, and the difference between two groups was of statistical significance as well ($P < 0.05$). Nevertheless, there was no statistically significant difference in the accidental slip rate of the needles as that turned out 0% in both groups ($P > 0.05$). It was worth noting that the block rate of the disposed needles in the intervention group was 100%. **Conclusion.** The separation-free safety tube on the disposable infusion needle could instantly block the sharp needle after infusion, which reduces the needle exposure time and lowers the risk of needle stick injuries. In the meantime, the safety tube is convenient to use, and its application can shorten the time for infusion and needle disposal, consequently improving the working efficiency of nurses. As the new type safety tube has above advantages and would not raise the risk of needle slippage, it is worthy of clinical promotion.

1. Introduction

According to the World Health Organization (WHO) [1], about 2,000,000 medical staffs suffer from infectious diseases caused by needle stick injuries (NSIs) each year, including hepatitis B virus (HBV), hepatitis C virus (HCV), and human immunodeficiency virus (HIV) [2]. Nowadays, NSIs have become the most serious occupational risk for medical staff. Our hospital is the First Hospital of Jiaxing affiliated to Jiaxing University which is the largest general hospital in Jiaxing city. In our hospital, we have over 2,200 staffs, among which 682 are doctors and 1,061 are nurses. Additionally, each year, there are more than 300 new resident doctors who have

received normalization training and over 300 medical students for clinical practice. Statistically, a total of 123 cases of NSIs happened in our hospital in 2018, including 57 cases that occurred in nurses accounting for 46.3%. The main medical sharp instrument responsible for NSIs turned out the intravenous infusion needles (41 cases) which accounted the highest as 33.3%. Among the 41 cases, 2 cases happened during punctuation, 2 cases occurred before reset after needle accidental slippage, 14 cases happened during the process of needle withdrawal, 9 cases happened on the way to the disposal room while the rest 14 cases occurred in the disposal room during needle processing. In order to reduce the incidence of NSIs caused by infusion needles, in this project,

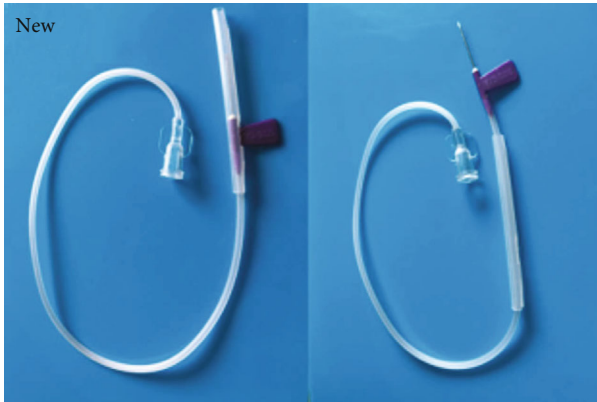


FIGURE 1: Pictures of the normal infusion set and the new type infusion set.

we developed a new type safety tube that could not only protect the needle before infusion but also instantly block it after infusion. Details are as follows.

2. Subject and Methods

2.1. Research Set. The conventional disposable infusion set originally has a protective cover on the needle, which prevents the package and the operator from being punctured and stabbed, and is usually discarded during normal operation. In this project, we transformed such protective cover into a new type one named safety tube. As shown in Figure 1, the inner diameter of the safety tube is slightly larger than the outside diameter of the needle body, and the length of the tube is slightly longer than the total length of the needle head plus the needle body. There is a slit along the length from the front to the middle of the tube. The open end (front) of the tube linked with the slit is in the shape of a “V,” the end of the slit at the middle of the tube is linked with a rectangular hole, and the joint is in the shape of an inverted “V.” The safety tube is in register with the needle, and the fin of the needle is out of the rectangular hole. During infusion, the fin is pushed to make the needle slide along the slit out of the tube, and after infusion, the needle slides back into the tube for safety. The safety tube has been commissioned to a qualified manufacturer and applied to the clinical trial. Meanwhile, it was approved by the hospital ethics committee.

2.2. Subject. 200 inpatients who were in need of infusion therapy from October to December 2018 were enrolled and randomly divided into the intervention group and the control group. There was no significant difference in age and disease diagnosis between the two groups ($P > 0.05$). In the study, we used disposable infusion needles with a new type separation-free safety tube for the patients in the intervention group and conventional ones with a self-contained protective cover in the control group. Patients in both groups volunteered to participate and had signed informed consent.

2.3. Operation Methods. Six ward nurses with proficiency in the conventional infusion operation were selected, including 2 nurses working for 1-3 years, 2 for 3-5 years, and 2 for over

5 years. Before the project, all of them were trained in the operation of this new type safety tube and qualified. In the control group, conventional disposable infusion needles were used. After breathing, the self-contained protective cover was discarded, and the needle was fixed after acupuncture, then separated from the infusion set into a sharps box at the end. Patients in the intervention group were treated with new disposable infusion sets. The specific operation steps were as follows: (1) Instead of being removed after breathing, the safety tube slid to the flexible tube when the fin of the needle was pushed along the slit to make the needle out of the tube. Then, the infusion operation was as the same as the control group. (2) After infusion, the needle was removed. The upper end of the flexible tube away from the needle was raised, and the needle side was lowered to make the safety tube slide down to the needle side. Then, the fin was pushed to slide along the slit into the rectangular hole. The needle was thus blocked, and the whole infusion set was disposed into a special collection bag.

2.4. Observation Indicators. (1) Infusion operation time: from the beginning of the breathing to the end of needle fixation after application. (2) Needle exposure time: from the time of needle withdrawal to the time of needle blocked. The latter refers to the time for needle blocked by the safety tube in the intervention group and for needle blocked in a sharps box in the control group. (3) Needle disposal time: from the time of needle withdrawal to the time of needle disposed, wherein the latter refers to the time for needle collection in a special collection bag in the intervention group and in a sharps box in the control group. (4) Accidental slip rate of the infusion needle: the percentage of the cases with accidental needle slippage in the total infusion cases. (5) Needle block rate in the intervention group after disposal: the percentage of the successfully blocked needles (premarked in certain color) in the total amount of needles.

2.5. Statistical Methods. The data were input by the statistician and analyzed using the SPSS 21.0 software. The measurement data were presented in the form of mean \pm standard deviation ($M \pm SD$). The enumeration data were expressed by frequency and percentage, and the ranked data were determined using the Mann-Whitney U of the nonparametric test. $P < 0.05$ was considered statistically significant.

3. Results

3.1. Infusion Operation Time. As shown in Table 1, the operation time in the intervention group was shorter than that in the control group, with a statistically significant difference ($P < 0.05$).

3.2. Needle Exposure Time. As shown in Table 2, the needle exposure time in the intervention group was significantly shorter than that in the control group, with a statistically significant difference ($P < 0.05$).

3.3. Needle Disposal Time. As shown in Table 3, the time for needle disposal after infusion in the intervention group was

TABLE 1: Infusion operation time (seconds).

Operation time	Median	Interquartile range	Z	P
Intervention group	82.19	1.69	-3.441	0.001
Control group	82.53	1.27		

TABLE 2: Needle exposure time (seconds).

Exposure time	Median	Interquartile range	Z	P
Intervention group	3.36	0.16	-12.219	<0.001
Control group	18.52	1.24		

TABLE 3: Needle disposal time (seconds).

Disposal time	Median	Interquartile range	Z	P
Intervention group	18.39	1.03	-2.151	0.031
Control group	18.52	1.24		

shorter than that in the control group, with a statistically significant difference ($P < 0.05$).

3.4. Others. The accidental slip rate of the infusion needles in the intervention group and the control group were both 0%, with no statistically significant difference ($P > 0.05$). In addition, 100 needles in the intervention group were all blocked after disposal with the block rate of 100%.

4. Discussion

According to the Centers for Disease Control and Prevention (CDC) statistics [3], 80%-90% healthy medical staffs with infectious diseases are caused by NSIs, of which 80% are nurses. In recent years, as the NSIs happen more often, it has been highly focused in medicine at home and abroad. Various protective measures thus have been studied, such as the development of nursing equipment with safe and protective sets and the application of needle-free products, which have made the incidence of NSIs reduced by 43% [4]. However, due to the high cost of such products, they have not been widely promoted in China at present. Disposable infusion needle is still the main infusion set used in most hospitals, and nurses still have to face the exposed needles with blood and body fluids of patients every day. In addition, nurses need to hold the needles by hand to the disposal room for needle separation after infusion completed. During this process, NSIs could easily happen in either the operators or other people. Among the 41 cases of NSIs in our hospital in 2018, 9 cases happened on the way to find a sharps box. The exposure time of the needle is positively related to the risk of NSIs. Therefore, shortening the needle exposure time is the key to reduce the risk of NSIs. The effective way is to place the sharps boxes in a place where nurses can conveniently reach, such as the bed end or the treatment cart configured with a sharps box in each ward, which could allow

nurses to dispose the used needles in time so as to reduce the exposure time. However, due to the factors like the national conditions, risk of sharps loss, and economic cost, this approach has not been implemented. In most hospitals in China, only the disposal room and the treatment cart in the nursing station are configured with sharps boxes. From the results of this study, the needle exposure time in the intervention group was significantly shorter than that in the control group ($P < 0.01$). The main reason might be that in the intervention group, the safety tube could slide to the flexible tube when acupuncture completed, and instantly slide back to the needle side after infusion completed, making the needle blocked in time with no necessity searching for a sharps box. The application of the safety tube greatly shortened the needle exposure time, thereby reducing the risk of NSIs. Besides, during the process of needle disposal, the needle should be separated from the infusion set using tools, with the needle placed in a sharps box and the infusion set in a special collection bag. Such treatment is apt to cause NSIs, and 14 cases among the total stick injury cases in our hospital in 2018 just happened during the disposal process. In this study, the blocked needles were disposed in a special collection bag together with the infusion set in the intervention group. Moreover, the needle block rate reached 100% before collection in 10-24 hours, suggesting that the separation-free type safety tube could effectively work, and the sharps box for separation could be no longer needed, which further lowered the incidence of NSIs.

As a large number of infusion operations have to be completed every day, nurses would spend much working time on it. From the data of this group, the infusion operation time of the intervention group was less than that of the control group, and the difference was statistically significant ($P < 0.01$). The main reason can be concluded as the difference in the handling methods for the safety tube in the intervention group and the protective cover in the control group. Moreover, the safety tube is convenient to operate and nurses can grasp it easily. For the protective cover in the control group, it should be removed from the needle and discarded into a waste pail. While for the safety tube in the intervention group, it only needs to slide to the flexible tube when acupuncture, which shortens the operation time. In addition, the needle disposal time after infusion in the intervention group was shorter than that in the control group, with a statistically significant difference ($P < 0.01$). The main reason is that in the control group, the needle should be separated from the infusion set into a sharps box after withdrawal for waste disposal. While in the intervention group, the needle is instantly blocked by the safety tube and placed into the nearest special collection bag. There is no need to process separation; thus, the needle disposal time is reduced. It can be seen that the application of the separation-free safety tube can improve the working efficiency of nurses. In the meantime, the subjects in our study gave no bad feedback and no accidental slippage occurred, indicating that the new type infusion needle is safe to be applied.

The self-contained protective cover of the conventional infusion needle will be discarded routinely before infusion. In this study, the protective cover that should have been

Research Article

Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy

Haiyan Liang,¹ Lei Chen ,^{1,2} Xian Zhao,¹ and Xiaolin Zhang¹

¹College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

²Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China

Correspondence should be addressed to Lei Chen; chen_lei1@163.com

Received 28 January 2020; Revised 14 April 2020; Accepted 23 April 2020; Published 9 May 2020

Academic Editor: Rafik Karaman

Copyright © 2020 Haiyan Liang 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.

Drugs are an important way to treat various diseases. However, they inevitably produce side effects, bringing great risks to human bodies and pharmaceutical companies. How to predict the side effects of drugs has become one of the essential problems in drug research. Designing efficient computational methods is an alternative way. Some studies paired the drug and side effect as a sample, thereby modeling the problem as a binary classification problem. However, the selection of negative samples is a key problem in this case. In this study, a novel negative sample selection strategy was designed for accessing high-quality negative samples. Such strategy applied the random walk with restart (RWR) algorithm on a chemical-chemical interaction network to select pairs of drugs and side effects, such that drugs were less likely to have corresponding side effects, as negative samples. Through several tests with a fixed feature extraction scheme and different machine-learning algorithms, models with selected negative samples produced high performance. The best model even yielded nearly perfect performance. These models had much higher performance than those without such strategy or with another selection strategy. Furthermore, it is not necessary to consider the balance of positive and negative samples under such a strategy.

1. Introduction

Drugs are always special products for the treatment of various diseases. However, a drug is also a double-edged sword; it can bring some unexpected negative effects, usually called side effects, when it produces therapeutic effects. Side effects are almost inevitable for all drugs. Determining the side effects of drugs as early as possible can decrease the risks both for patients and pharmaceutical companies. It is reported that side effects cause 100,000 deaths per year in the United States [1]. On the other hand, an unacceptable side effect is the major reason for the failure of drug development. Even some launched drugs (e.g., Rofecoxib) had to be withdrawn after their unacceptable side effects were discovered. Thus, it is urgent to design effective methods to determine the side effects of drugs. However, it takes a lot of time and is of high costs to ascertain the side effects of a given drug through clinical trials. With the development of computer science, lots of

advanced computational methods have been proposed, which give abundant resources to build effective computational models in this regard.

In recent years, many computational methods have been developed for predicting side effects of drugs. Among these methods, several of them built an individual classifier for each side effect [1–5]. They always took the drugs having a given side effect as positive samples and other drugs as negative samples. Clearly, to determine all side effects of a given drug, a large number of classifiers should be performed. Considering the fact that plenty of drugs have multiple side effects, some methods deemed the problem of predicting drug side effects as a multilabel classification problem [6–11]. It is a good idea to build a uniform frame to predict side effects of given drugs. However, these models are always complex and have high computational complexity. Different from the above methods, other methods built regression models for the prediction of drug side effects [12, 13].

Recently, some studies proposed a uniform binary classification model for predicting drug side effects [14–17]. They deemed the pairs of drugs and side effects as samples. A pair containing a drug and a side effect such that the drug has this side effect was termed as a positive sample and other pairs as negative samples. Because there were lots of negative samples, if all negative samples are selected, it is quite difficult to set up an effective prediction model. In some studies, they randomly selected some of them to build the model. It is clear that the utility of these constructed models relied on the selection of negative samples. Random selection of negative samples is not a rigorous way because some potential positive samples that have not been validated may be selected. Furthermore, selecting how many negative samples is also an important problem. It is necessary to design a refined strategy for picking up negative samples that are true negative samples with extreme high probabilities.

In this study, we did some work for selecting negative samples. A refined negative sample selection strategy was proposed to select high-quality negative samples. To this end, a drug network was constructed according to the chemical-chemical interaction (CCI) information retrieved from STITCH [18, 19]. Then, the random walk with restart (RWR) algorithm [20] was applied on the network to access high-quality negative samples. Based on obtained negative samples and positive samples retrieved from SIDER [21], classification models incorporating certain classification algorithms can be built, in which each sample was encoded into five features used in Zhao et al.’s study [14]. Three classification algorithms, random forest (RF) [22], support vector machine (SVM) [23], and artificial neural network (ANN), were adopted in this study. Several tests were performed to evaluate classification models with different classification algorithms and different quality negative samples. The best model gave the almost perfect classification. Furthermore, the proportion of positive and negative samples was not a problem when our negative sample selection strategy was used.

2. Materials and Methods

2.1. Materials. Drugs and their side effects used in this study were the same as those in our previous study [14]. In fact, this information was obtained from the well-known public database, SIDER [21]. The raw information contained a total of 888 drugs and 1385 side effects. With the same data cleaning procedures, we excluded the side effects with less than six drugs and drugs whose properties mentioned in Drug Properties and Associations were not available. Finally, 841 drugs and 824 side effects were accessed. In this study, the pairs of drugs and side effects were termed as samples. The above-mentioned drugs and side effects can comprise 57,058 pairs of drugs and side effects, which were deemed as positive samples. For convenience, these samples constituted the dataset PDS.

2.2. Negative Sample Selection Strategy. In Materials, the dataset PDS containing the positive pairs of drugs and side effects was constructed according to the information in SIDER. To construct the classification model, negative samples were necessary. In our previous study [14], negative

samples were produced by randomly pairing drugs and side effects. Here, a refined strategy was proposed, which can generate high-quality negative samples.

2.2.1. Drug Network. It has been reported in several studies that interacting chemicals are more likely to share similar properties [24–29]. It is feasible to adopt such information for investigating drug side effects because side effect is one of the important properties of drugs. In this study, we used the information of CCI to construct a drug network.

The CCI information was retrieved from STITCH (<http://stitch.embl.de/>, version 4.0) [18, 19], an online public database collecting known and predicted interactions between chemicals and proteins. These interactions were obtained by the evidence derived from experiments, databases, and the literature. Thus, they can widely measure the associations between chemicals and proteins. Each CCI in STITCH is assigned five scores, titled by “similarity,” “experimental,” “database,” “textmining,” and “combined_score,” with a range between 1 and 999. In detail, the first four scores measure the associations of chemicals according to their structures, activities, reactions and cooccurrence in the literature, while the last one integrates all above scores. Clearly, the last score can widely and accurately evaluate the linkages between chemicals. Thus, we used such score to construct the drug network. For formulation, let us denote the “combined_score” of chemicals c_1 and c_2 as $Q(c_1, c_2)$.

The constructed drug network took 841 drugs as nodes, and two drugs were adjacent if and only if they can comprise a CCI with a “combined_score” larger than zero. Furthermore, to indicate the different strength of edges, each edge with d_1 and d_2 as endpoints was assigned a weight that was defined as $Q(d_1, d_2)$.

2.2.2. Random Walk with Restart Algorithm. The RWR algorithm is a powerful and widely used network ranking algorithm [20, 27, 30–33]. In this algorithm, the walker randomly moves from a seed node set to other nodes in the network. When the algorithm stops, each node in the network receives a probability, which can be deemed as an important indicator representing the essential associations to seed nodes. Given a seed node set SN, the RWR algorithm first constructs a probability vector, denoted as p_0 , in which the probability for each node in SN is defined as $1/|SN|$, while probabilities for other nodes are set to zero. The RWR algorithm repeatedly updates this probability vector until it becomes stable. Let p_t represent such probability vector after the t -th iteration has been executed. Then, the probability vector p_{t+1} is updated by the following equation:

$$p_{t+1} = (1 - \lambda)A^T p_t + \lambda p_0, \quad (1)$$

where λ was set to 0.8, as used in other studies [27, 32–34], in this study and A represents the columnwise normalized adjacency matrix of the network. When $\|p_{t+1} - p_t\|_{L_1} < \theta$, the update procedure stops, and p_{t+1} is picked up as the output of the RWR algorithm. In this study, θ was set to 10^{-6} .

The refined negative sample selection strategy is based on the above-mentioned drug network and RWR algorithm. For

each drug side effect, we picked up the drugs owning such side effect as seed nodes of the RWR algorithm. Then, the RWR algorithm was applied on the drug network. When the RWR algorithm stopped, each node in the network was assigned a probability. It is clear that a node (drug) with a high probability had a strong association with seed nodes, thereby inferring that such node had a high probability of owning the side effect. On the contrary, nodes (drugs) with low probabilities were less likely to own the side effect. Given a threshold ε of the probability, drugs receiving the probabilities less than ε can be extracted, and they were paired with the side effect as the candidate negative samples. After considering all side effects, a negative sample set, denoted by NDS, was built by collecting all candidate negative samples for each side effect. This set was combined with PDS to constitute the training dataset.

2.3. Drug Properties and Associations. To encode each pair of drugs and side effects, we employed five drug properties, which were also used in our previous study [14]. Based on each property, a score evaluating the associations between two drugs can be obtained. Here, a brief description is given. The detailed description can be found in our previous study [14].

2.3.1. Drug Association in Fingerprint. A drug can be represented by a SMILES (simplifying the molecular linear input specification) string [35], from which its fingerprints (ECFP_4) were extracted via RDKit [36]. Then, the Tanimoto coefficient is adopted to quantify the association between two drugs based on their fingerprints. For formulation, the thus-obtained association between drugs d_1 and d_2 is denoted by $W^f(d_1, d_2)$.

2.3.2. Drug Association in Structure. Apart from the SMILES strings to represent drugs, drugs can also be represented by a graph [37]. Then, the association between two drugs can be assessed according to the sizes of two graphs and their maximum common subgraph. The online tool "SIMCOMP" in KEGG adopts such scheme to evaluate the associations of drugs [38]. The score between d_1 and d_2 obtained by "SIMCOMP" is denoted by $W^s(d_1, d_2)$.

2.3.3. Drug Association in ATC Code. In the Anatomical Therapeutic Chemical (ATC) classification system, each drug is assigned one or more five-level ATC codes. According to the ATC codes of two drugs, their associations can be quantified. Detailed descriptions can be found in [14]. $W^c(d_1, d_2)$ is used to represent the associations of drugs in terms of their ATC codes.

2.3.4. Drug Literature Association. The drug association can further be assessed by text-mining methods. Here, we adopted such association reported in STITCH [18, 19]. For drugs d_1 and d_2 , their association is denoted by $W^l(d_1, d_2)$.

2.3.5. Drug Association in Target Protein. A drug has one or more target proteins. This information can be represented by a 0-1 vector. Then, the association of two drugs can be quan-

tified by the direction cosine of corresponding vectors. Let us denote such association between d_1 and d_2 by $W^t(d_1, d_2)$.

2.4. Feature Construction. Based on the five types of drug associations mentioned in Drug Properties and Associations, we used the "similarity" concept to extract features for each sample. For each type of drug association, one feature was extracted. Here, we gave a procedure for extracting one feature from the drug association in the fingerprint. Others can be obtained in a similar way.

For a sample containing a drug d and side effect s , let S be a drug set consisting of drugs owning side effect s . The feature derived from the drug association in the fingerprint for such sample was defined as

$$Q^f(d, s) = \max \left\{ W^f(d, d') \mid d' \in S - \{d\} \right\}, \quad (2)$$

where $W^f(d, d')$ indicated the strength of the association between d and d' according to their fingerprints (see Drug Properties and Associations for detail). Obviously, a high $Q^f(d, s)$ meant the drug d was highly related to drugs owning side effect s . Thus, it was more likely to have side effect s .

Finally, each sample can be represented by a 5-dimension vector.

2.5. Classification Algorithm. Selecting a proper classification algorithm is very important for constructing an efficient classification model. This study adopted three classic algorithms: RF [22], SVM [23], and ANN. To quickly implement these algorithms, three tools "RandomForest," "SMO," and "MultilayerPerceptron" in Weka [39] were employed. For convenience, these tools were executed with their default parameters. Although the performance of models can be improved if more proper parameters were tried for each of the above-mentioned algorithms, it is not the keynote of this study. In our study, we tried to prove that the quality of negative samples selected by the proposed negative sample selection strategy was high no matter which algorithm was chosen as the prediction engine.

2.6. Performance Measurement. This study modeled a binary classification model for the prediction of drug side effects. For a binary classification problem, several measurements can be calculated to evaluate the performance of the model. In this study, we used the following measurements: *Recall* (also known as Sensitivity (SN) and *true positive rate* (TPR)), *false positive rate* (FPR), *Specificity* (SP), *prediction accuracy* (ACC), *Matthews correlation coefficient* (MCC) [40], *Precision*, and *F1-measure* [41]. Their formulations are as follows:

$$\text{Recall} = \text{SN} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

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

$$\begin{aligned}
ACC &= \frac{TP + TN}{TP + FN + FP + TN}, \\
MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP)}}, \\
Precision &= \frac{TP}{TP + FP}, \\
F1\text{-measure} &= \frac{2 \times Precision \times Recall}{Precision + Recall},
\end{aligned} \tag{3}$$

where TP and TN represent true positive and true negative, respectively, while FP and FN indicate false positive and false negative, respectively.

Besides, to fully evaluate the performance of different classification models, we further employed a receiver operating characteristic (ROC) curve and a precision-recall (PR) curve. By setting several thresholds for predicting positive samples, a series of TPRs, FPRs, and Precisions can be obtained. The ROC curve takes the TPR as the y -axis and FPR as the x -axis. Likewise, the PR curve sets the Precision as the y -axis and Recall as the x -axis. The areas under these two curves, called AUROC and AUPR, respectively, can be further calculated to assess the performance of the model. Clearly, a high AUROC or AUPR indicates high performance.

3. Results and Discussion

In this study, a binary classification model incorporating a refined negative sample selection strategy was proposed to predict drug side effects. The whole procedures are illustrated in Figure 1. This section gave detailed testing results of different models and made further analysis.

3.1. Negative Samples with Different Thresholds of Probability. The negative sample selection strategy applied the RWR algorithm to the drug network and extracted negative samples according to the threshold ε of the probability. We tried nine values of ε to construct nine different NDSs. The numbers of negative samples under different values of ε are listed in Table 1. It can be observed that the numbers of negative samples followed an increasing trend with the increasing of ε .

According to the principle of the RWR algorithm, it can be inferred that negative samples obtained by small ε were of high quality. To confirm this, based on nine thresholds listed in Table 1, we divided 333,797 negative samples selected by setting the threshold $\varepsilon = \varepsilon_9$ into nine parts ($[0, \varepsilon_1]$, $(\varepsilon_1, \varepsilon_2]$, $(\varepsilon_2, \varepsilon_3]$, $(\varepsilon_3, \varepsilon_4]$, $(\varepsilon_4, \varepsilon_5]$, $(\varepsilon_5, \varepsilon_6]$, $(\varepsilon_6, \varepsilon_7]$, $(\varepsilon_7, \varepsilon_8]$, and $(\varepsilon_8, \varepsilon_9]$). Then, for each negative sample with the drug d and side effect s , the ‘‘combined_score’’ between d and drugs owing side effect s was extracted. For each part, we counted the proportions of such scores in ten intervals from 0 and 999, which are illustrated in Figure 2, where Figure 2(a) considers zero scores, whereas Figure 2(b) excludes these scores. It can be observed from Figure 2 that all scores were zeros for the first four parts, indicating that

the drug in each of these samples had no direct links to drugs owning the side effect in the same sample. It is suggested that this drug shared such side effect with a quite low probability. For the following five parts, they contained more and more high scores, implying that in some samples, drugs had direct links to those sharing the side effects and these links became stronger. Thus, it can be deduced that drugs had the side effects with higher probabilities than the samples in the first four parts. In Figure 2, we also counted the scores of positive samples. Score distributions of some of the first parts were quite different from those of the positive samples, and with increase of the probability, the distribution became more and more similar to that of the positive samples. This suggested that with the increase of the part index, samples became more and more similar to positive samples. With the above analysis, it can be partly concluded that the quality of samples decreased with the increase of the part index. Thus, with the increase of the threshold, quality of selected negative samples became worse and worse because more and more negative samples with low quality were poured in. Especially when $\varepsilon = 0$, 128,220 negative samples were of the highest quality.

3.2. Performance of the Models with the Highest Quality Negative Samples. As mentioned in Negative Samples with Different Thresholds of Probability, 128,220 negative samples were obtained when $\varepsilon = 0$. These samples were deemed to be of the highest quality. Based on them and three classification algorithms: RF, SVM and ANN, three models, named as RF, SVM, and ANN models, respectively, were built. Then, tenfold crossvalidation [42–45] was adopted to evaluate their performance. Six measurements, SN, SP, ACC, MCC, Precision, and F1-measure, mentioned in Performance Measurement, were calculated and are listed in Table 2. It can be observed that the RF model yielded the best performance. The MCC, ACC, and F1-measure obtained by the RF model were 0.943, 0.975, and 0.959, respectively. As for the SVM and ANN models, they produced perfect SPs and Precisions; however, their SNs were much lower, inducing much lower ACCs, MCCs, and F1-measures. In addition, we plotted the ROC curves and PR curves of these models, as shown in Figure 3. Clearly, the ROC curve of the RF model was always above those of the SVM and ANN models. It was also true for the PR curve. The AUROC and AUPR of the RF model was 0.986 and 0.983, respectively, indicating the high utility of the RF model. The AUROCs and AUPRs of the other two models were at least 10% lower than those of the RF model. Therefore, it is suggested to select RF as the classification algorithm for building the classification model.

3.3. Performance of the Models with Different Quality Negative Samples. Given different thresholds of the probability, we can obtain different negative samples and construct different models. As listed in Table 1, nine thresholds were tried in this study. For RF, we constructed nine RF models. These models were evaluated by tenfold crossvalidation. The results are listed in Table 3 from which we can see that the MCCs followed a general decreasing trend with the

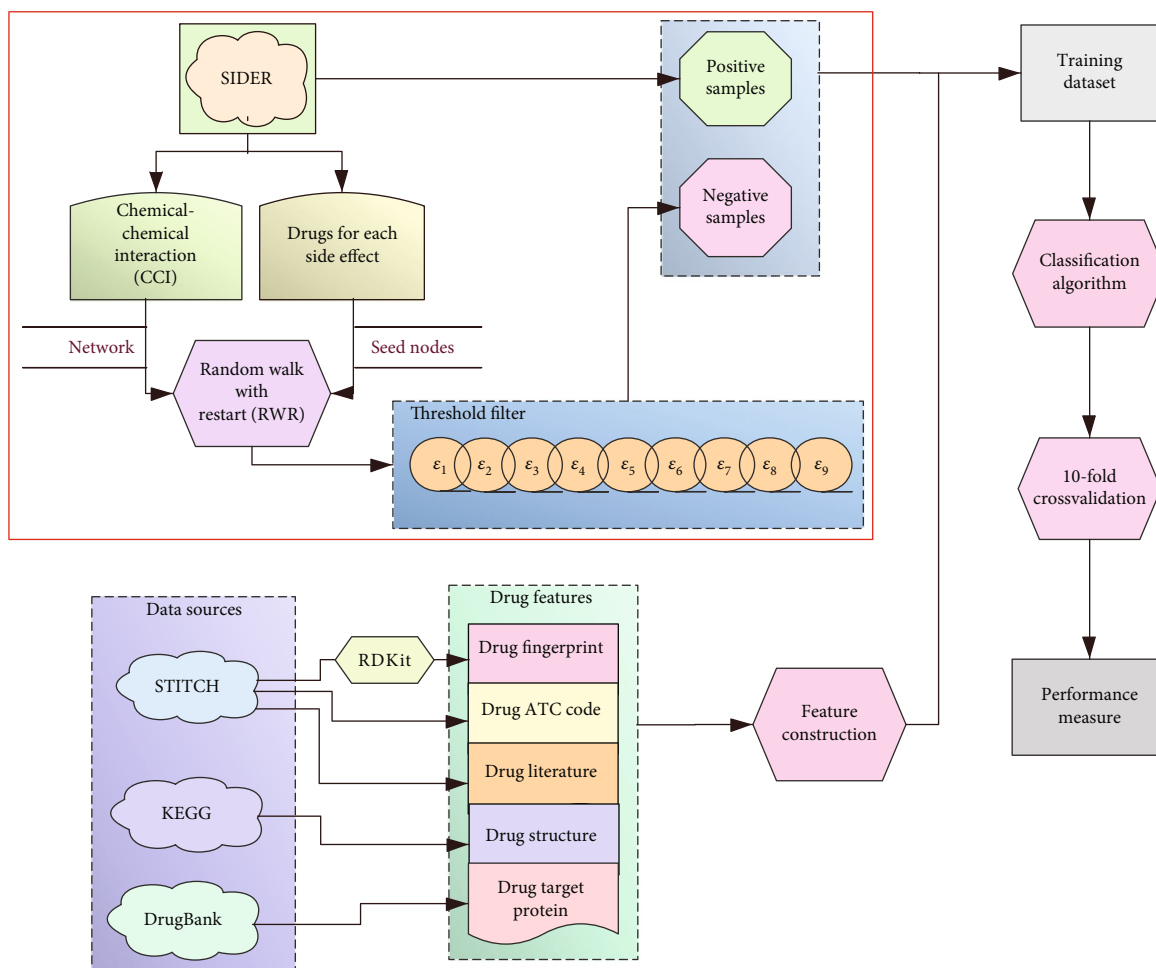


FIGURE 1: Entire procedures of the construction of classification models with a refined negative sample selection strategy.

TABLE 1: Numbers of negative samples under different thresholds of the probability.

Tag of threshold	Threshold	Number of negative samples	Times of positive samples
ϵ_1	0	128,220	2.25
ϵ_2	5×10^{-7}	131,301	2.30
ϵ_3	5×10^{-6}	152,596	2.67
ϵ_4	1×10^{-5}	173,822	3.05
ϵ_5	2×10^{-5}	216,256	3.79
ϵ_6	3×10^{-5}	259,566	4.55
ϵ_7	4×10^{-5}	294,260	5.16
ϵ_8	5×10^{-5}	317,971	5.57
ϵ_9	6×10^{-5}	333,797	5.85

increase of the threshold. ACC and F1-measure also followed such trend. It is reasonable because when the threshold increased, more and more negative samples were added and their quality became poorer. It can also be concluded from Table 3 that when the thresholds were small, the performance of the RF model followed a sharp decreasing trend

with the increase of the threshold, while this trend became alleviative when the thresholds were large. With the increase of the threshold, the added negative samples were poor enough which cannot influence the performance of the model a lot. Besides, we also plotted the ROC curves and PR curves yielded by these RF models, as shown in Figure 4. The AUROCs and AUPRs followed the similar trend of ACCs, MCCs, and F1-measures. Thus, it is better to use a small threshold for determining negative samples.

To prove that the above results were not special for RF, we also did the same tests for SVM and ANN. The predicted results are provided in Table S1 and S2. The ROC curves and PR curves are available in Figure S1 and S2. All results were almost identical to those of the RF models, indicating that with the increase of the threshold, the quality of negative samples became poorer. It is suggested to extract negative samples with a small threshold.

3.4. Analysis of the Models on Balanced and Imbalanced Datasets. For the problem investigated in this study, it is a dilemma to determine the number of negative samples. Based on our negative sample selection strategy, it is not a problem. The negative samples can be determined by a proper threshold, which suggests choosing a small threshold.

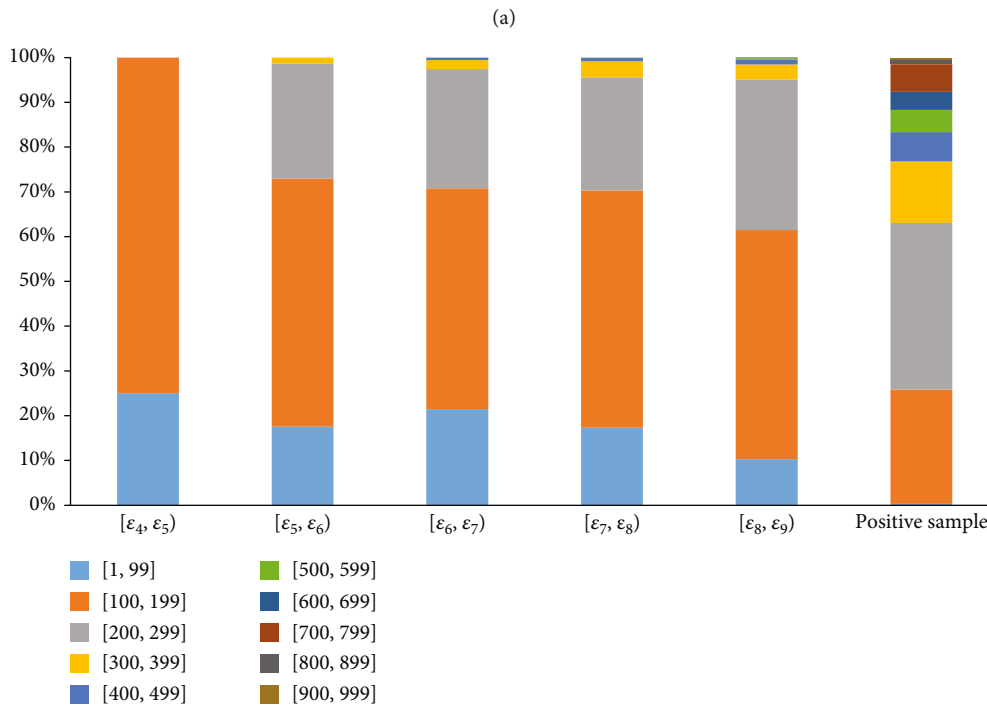
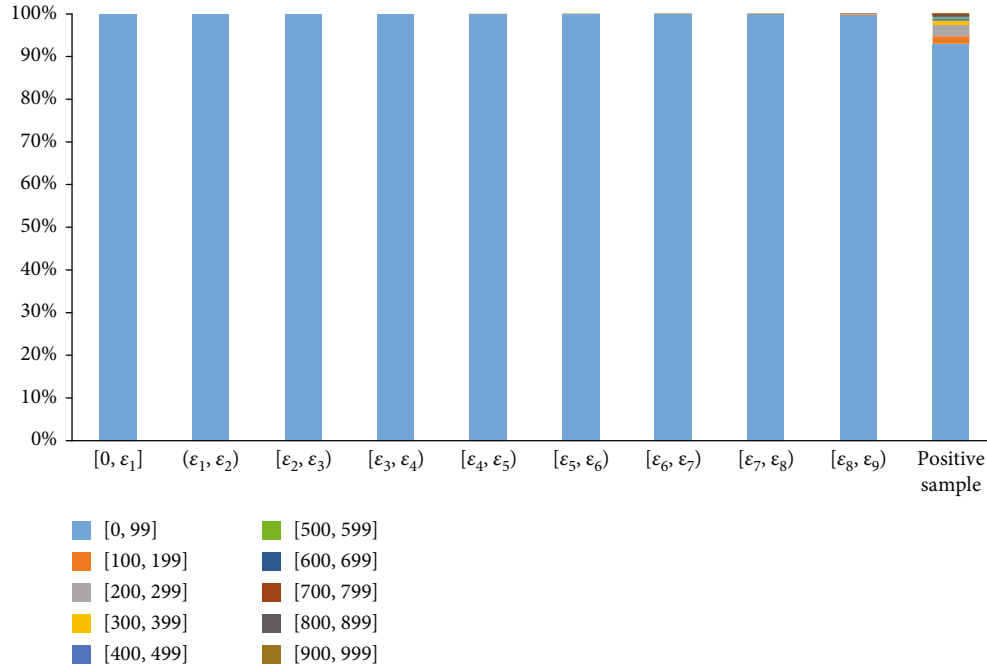


FIGURE 2: Distribution of “combined_score” of drugs and drugs sharing the side effects for negative samples in nine parts and positive samples. (a) Zero scores were included; (b) zero scores were not included.

TABLE 2: The performance of three models with the highest quality negative samples.

Model	SN	SP	ACC	MCC	Precision	F1-measure
RF model	0.923	0.999	0.975	0.943	0.997	0.959
SVM model	0.656	1.000	0.894	0.754	1.000	0.792
ANN model	0.670	1.000	0.898	0.764	1.000	0.802

It can be seen that the above-constructed models all used much more negative samples than positive samples (Table 1). For example, when $\varepsilon = 0$, the negative samples were more than twice the positive samples. With the increase of ε , the negative samples became more and more. Thus, the above-constructed models were all based on imbalanced datasets. This section proved that when the threshold was given, the proportion of positive and negative samples cannot be considered. To this end, we did the following tests.

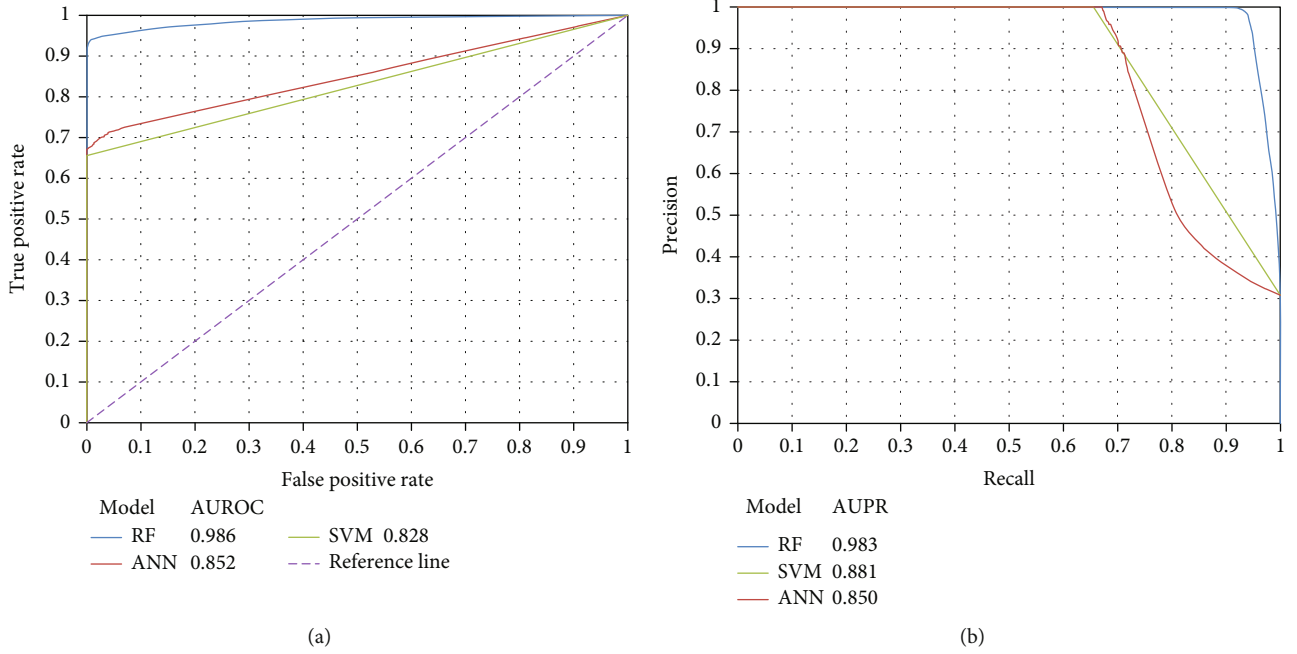


FIGURE 3: The ROC curves and PR curves of three models with the highest quality negative samples. (a) The ROC curves; (b) the PR curves.

TABLE 3: The performance of the RF models with different quality negative samples.

Threshold	SN	SP	ACC	MCC	Precision	F1-measure
ε_1	0.923	0.999	0.975	0.943	0.997	0.959
ε_2	0.910	0.978	0.958	0.899	0.948	0.929
ε_3	0.816	0.960	0.921	0.796	0.883	0.849
ε_4	0.751	0.964	0.912	0.754	0.873	0.808
ε_5	0.668	0.975	0.911	0.715	0.877	0.758
ε_6	0.622	0.982	0.917	0.698	0.884	0.730
ε_7	0.605	0.986	0.924	0.695	0.890	0.720
ε_8	0.594	0.987	0.927	0.691	0.891	0.713
ε_9	0.588	0.989	0.930	0.694	0.901	0.712

For a given threshold ε , we can obtain several negative samples. Among them, we randomly selected negative samples, which were as many as positive samples. These selected negative samples were combined with positive samples to construct a balanced dataset. Because the selection of negative samples may influence the results, we constructed four additional datasets in the same way. Thus, five balanced datasets, denoted by $BD_1^\varepsilon, \dots, BD_5^\varepsilon$, were constructed. Furthermore, we constructed five imbalanced datasets in a similar way. These datasets contained negative samples twice as many as positive samples. These imbalanced datasets were denoted by $IBD_1^\varepsilon, \dots, IBD_5^\varepsilon$. A RF model was built based on each of the above-mentioned datasets and evaluated by tenfold crossvalidation. The performance of these RF models on balanced datasets is shown in Figure 5. It can be observed that with the increase of the threshold, the performance of RF models decreased, which conformed to the results in Performance of the Models with Different Quality Negative Sam-

ples. Furthermore, the performance of the RF models on imbalanced datasets is illustrated in Figure 6, giving the same conclusion. In addition, SVM and ANN models were also constructed on the above-mentioned balanced and imbalanced datasets. Their performance, evaluated by tenfold crossvalidation, is shown in Figure S3-S6. The same conclusion can be arrived at; that is, the performance of the models decreased when the threshold increased.

Given a threshold ε , three types of datasets were constructed. The first one contained all negative samples; the second one, imbalanced datasets, containing negative samples twice as many as positive samples; and the last one, balanced datasets, containing negative samples as many as positive samples. As shown in Table 1, the first type of dataset had the highest imbalanced degree, followed by the second and third ones. Here, we investigated the performance of RF models on these three types of datasets under different thresholds of the probability. The MCCs are illustrated in Figure 7. It is interesting that given a threshold, the model on the first type of datasets always provided the best performance although it contained much more negative samples than the other two types of datasets. The reason may be that negative samples under a certain threshold were quite similar for the RF model; thus, employing more negative samples can improve the performance. For the two other types of datasets, when the threshold was smaller than or equal to ε_6 , RF models on imbalanced datasets were superior to those on balanced datasets, while it became contrary when the threshold was larger than ε_6 . It is indicated that there existed a critical value to control the performance of the RF model on balanced and imbalanced datasets. Furthermore, we investigated the performance of the SVM and ANN models on three types of datasets. Obtained MCCs are illustrated in Figure S7 and S8. For the SVM and ANN models, their performance was

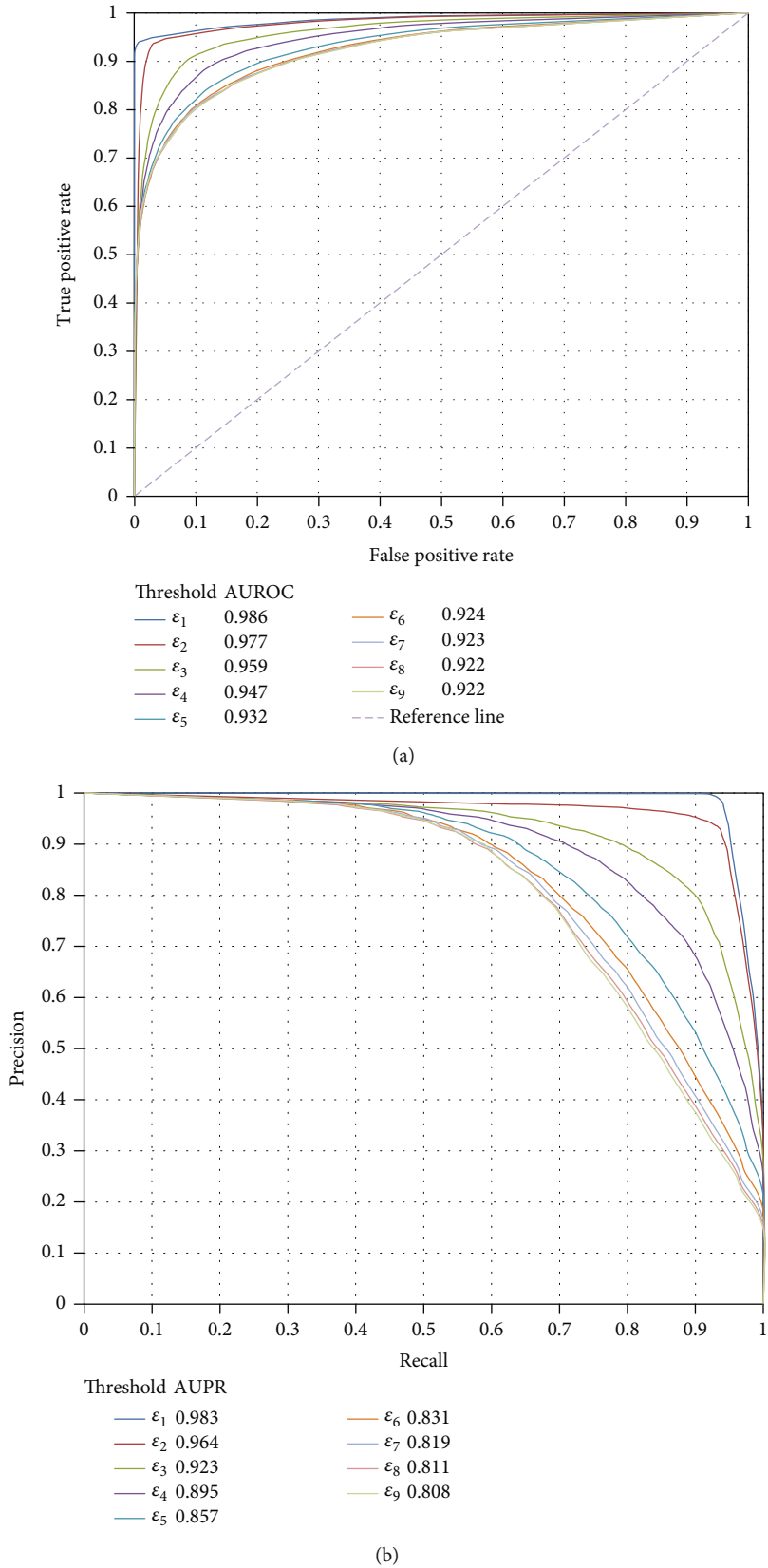
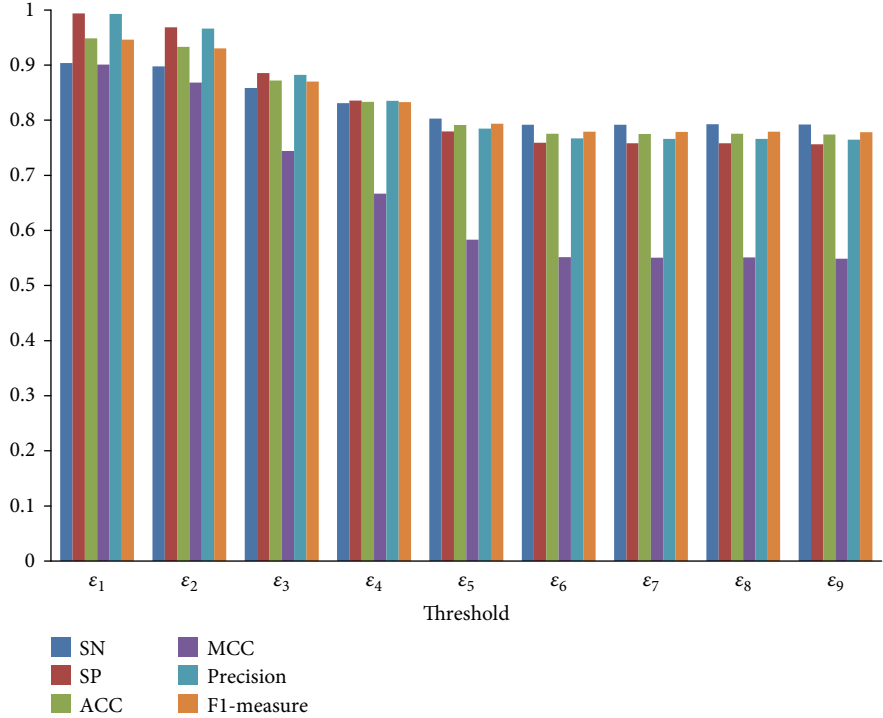
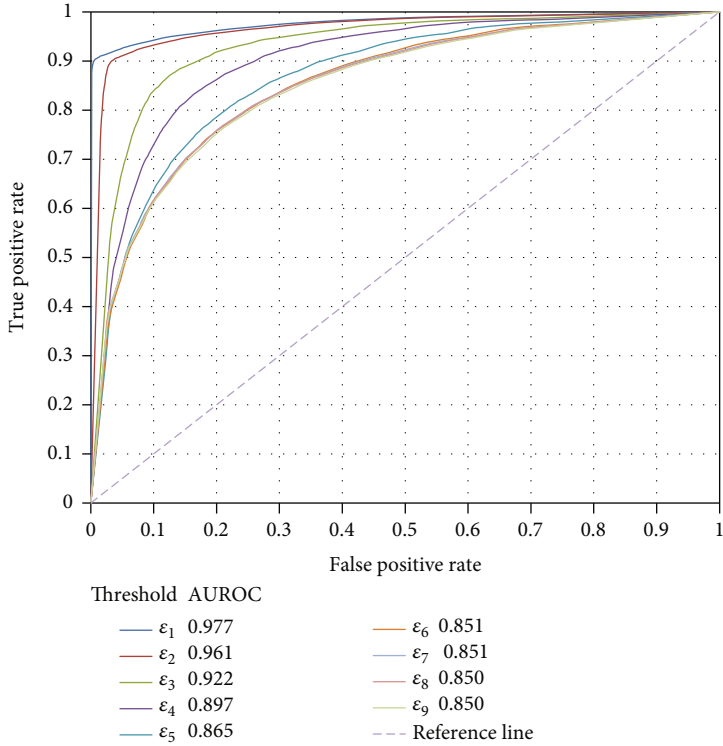


FIGURE 4: The ROC curves and PR curves of the RF models with different quality negative samples. (a) The ROC curves; (b) the PR curves.



(a)



(b)

FIGURE 5: Continued.

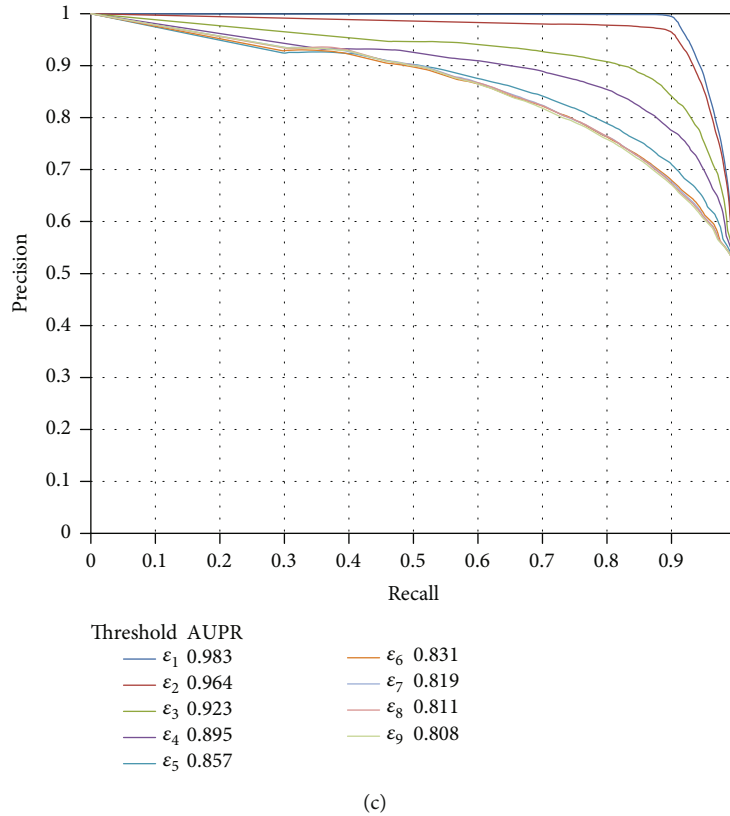


FIGURE 5: The performance of the RF models on balanced datasets, in which negative samples, as many as positive samples, are randomly selected under different thresholds. (a) Six measurements; (b) the ROC curves; (c) the PR curves.

not always best on the first type of datasets when the threshold was fixed. However, when the threshold was small (smaller than ε_3), the first type of dataset still yielded the best performance. For the second (imbalanced) and third (balanced) types of datasets, similar phenomena occurred. The only difference was the different critical values for SVM and ANN.

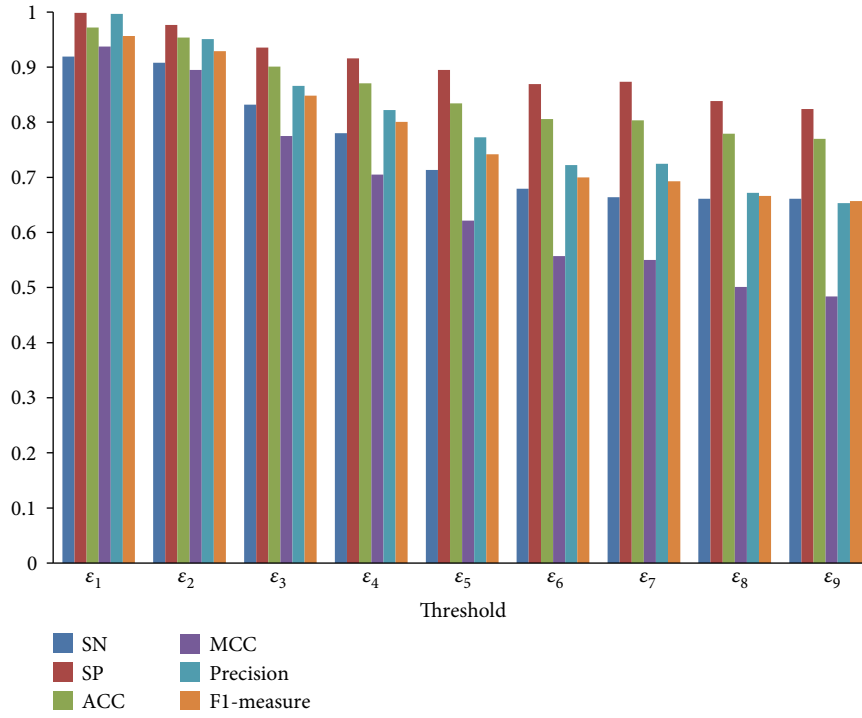
All in all, when we used threshold ε , which was suggested to be small in Performance of the Models with Different Quality Negative Samples, to determine the candidate negative samples, it is better to pick up all these candidates to construct the model, and it was not necessary to consider the proportion of positive and negative samples in this case.

3.5. Comparison of the Model without Negative Sample Selection. In this study, we proposed a refined negative sample selection strategy to extract high-quality negative samples. When using the highest quality negative samples, the RF model produced the best performance, listed in Table 2. If such strategy was not adopted, we randomly selected negative samples that were as many as positive samples to construct the RF model, which was identical to that in our previous study [14]. The predicted results yielded by the tenfold crossvalidation are listed in Table 4. It is easy to see that our model was much superior to the previous model. Each measurement was improved more than 10%. In detail, the ACC, MCC, and F1-measure improved about 20%, 40%, and 18%, respectively. Thus, the proposed negative sample

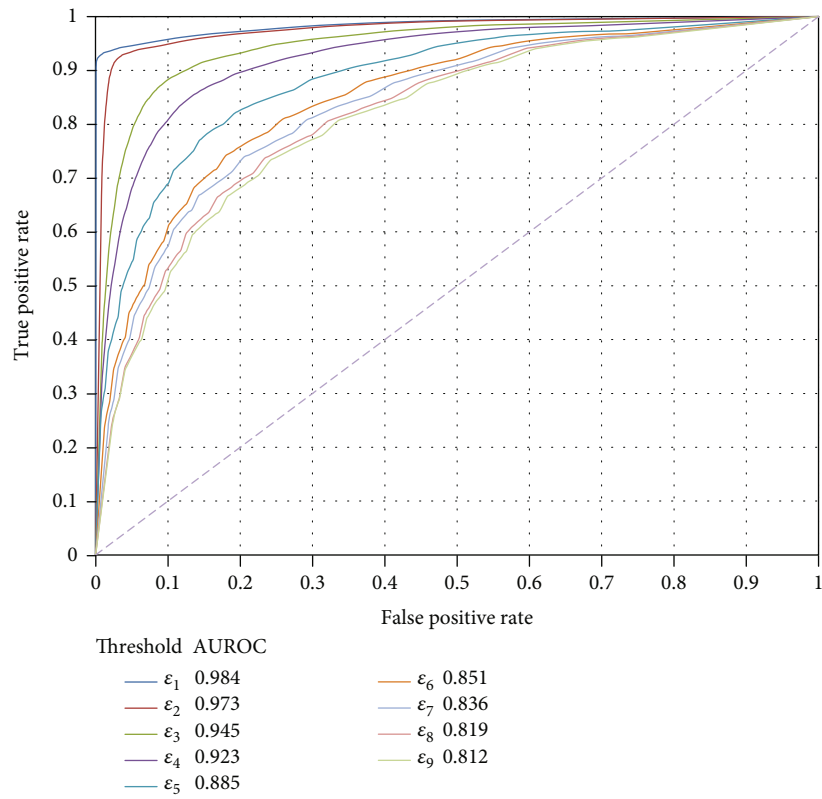
selection strategy can sharply improve the utility of the model. Furthermore, we also did the same comparisons for the SVM and ANN models. Predicted results are also listed in Table 4. The same conclusion can be obtained.

3.6. Comparison of the Model with Another Negative Sample Selection Strategy. In [46], another negative sample selection strategy, namely, finding reliable negative samples (FIRE), was proposed to improve the model for predicting protein-RNA interactions. This strategy was employed in this study to compare with the proposed strategy. We termed drugs as proteins and side effects as RNAs in FIRE. Furthermore, the “combined_score” between drugs was deemed as the protein-protein similarity score in FIRE. According to FIRE, each pair of drug and side effect that was not a positive sample was assigned a score. It was claimed in [46] that samples with low scores were of high quality. Thus, we picked up the pairs of drugs and side effects with zero scores as negative samples for making comparison, obtaining 355,634 negative samples. The negative samples with the highest quality (using threshold ε_1) filtered by our strategy were selected to make comparison. Several RF models were constructed on these two different negative sample sets and the same positive samples.

First, we compared the RF models with balanced positive and negative samples; that is, 57,058 negative samples were randomly selected from two negative sample sets, which were combined with the positive samples to construct RF models.



(a)



(b)

FIGURE 6: Continued.

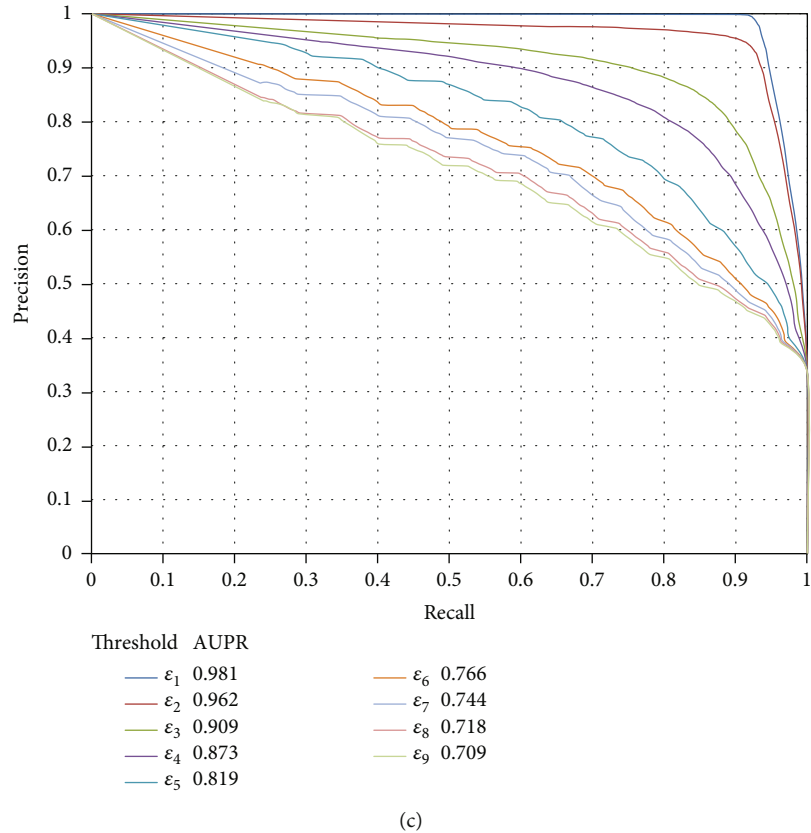


FIGURE 6: The performance of the RF models on imbalanced datasets, in which negative samples, twice as many as positive samples, are randomly selected under different thresholds. (a) Six measurements; (b) the ROC curves; (c) the PR curves.

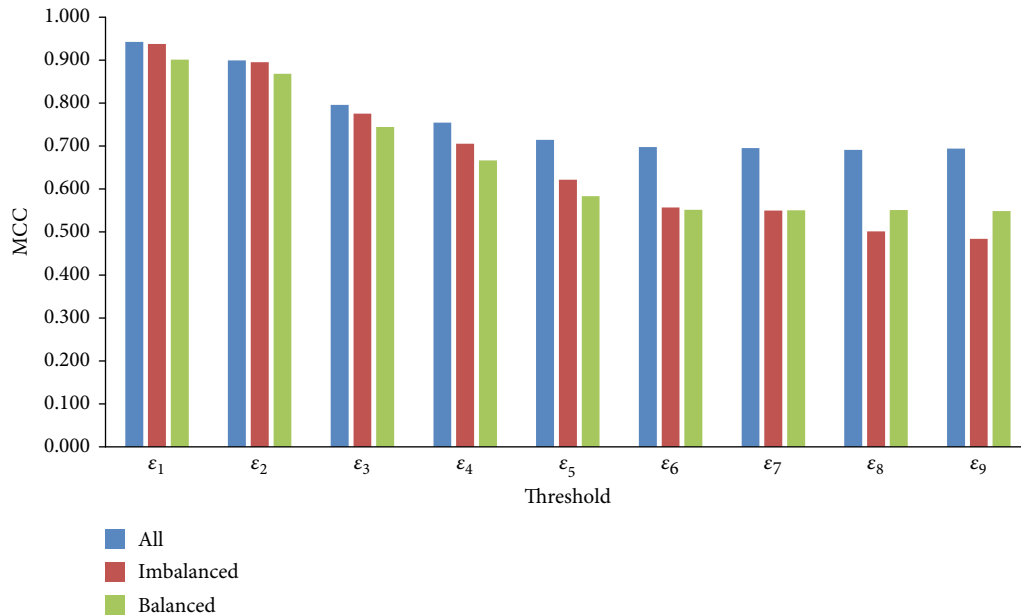


FIGURE 7: The MCCs yielded by the RF models on three types of datasets. “All” means that all negative samples under the given threshold are selected; “Imbalanced” indicates that negative samples, twice as many as positive samples, under the given threshold are randomly selected; and “Balanced” indicates that negative samples, as many as positive samples, under the given threshold are randomly selected.

Tenfold crossvalidation results are listed in Table 5. Clearly, the RF model obtained by the proposed strategy (called the proposed model in the following text for convenience) was

superior to that obtained by FIRE (called the FIRE model in the following text for convenience). The MCC was 10.7% higher. Furthermore, we also did the ROC and PR curve

TABLE 4: Comparison of the models with or without negative sample selection strategy.

Model	Negative sample selection strategy	SN	SP	ACC	MCC	Precision	F1-measure
RF model	$\sqrt{(\text{with } \epsilon_1)}$	0.923	0.999	0.975	0.943	0.997	0.959
	\times	0.791	0.759	0.775	0.550	0.766	0.778
SVM model	$\sqrt{(\text{with } \epsilon_1)}$	0.656	1.000	0.894	0.754	1.000	0.792
	\times	0.585	0.715	0.650	0.302	0.672	0.625
ANN model	$\sqrt{(\text{with } \epsilon_1)}$	0.670	1.000	0.898	0.764	1.000	0.802
	\times	0.631	0.695	0.663	0.332	0.682	0.650

TABLE 5: Comparison of RF models with two different negative sample selection strategies.

Negative sample selection strategy	Number of selected negative samples	SN	SP	ACC	MCC	Precision	F1-measure
Proposed strategy	57,058	0.904	0.994	0.949	0.901	0.993	0.946
	114,116	0.919	0.999	0.972	0.938	0.997	0.956
	128,220	0.923	0.999	0.975	0.943	0.997	0.959
FIRE	57,058	0.887	0.907	0.897	0.794	0.905	0.896
	114,116	0.844	0.954	0.917	0.811	0.901	0.872
	128,220	0.832	0.961	0.921	0.812	0.904	0.867
	355,634	0.745	0.992	0.957	0.812	0.934	0.829

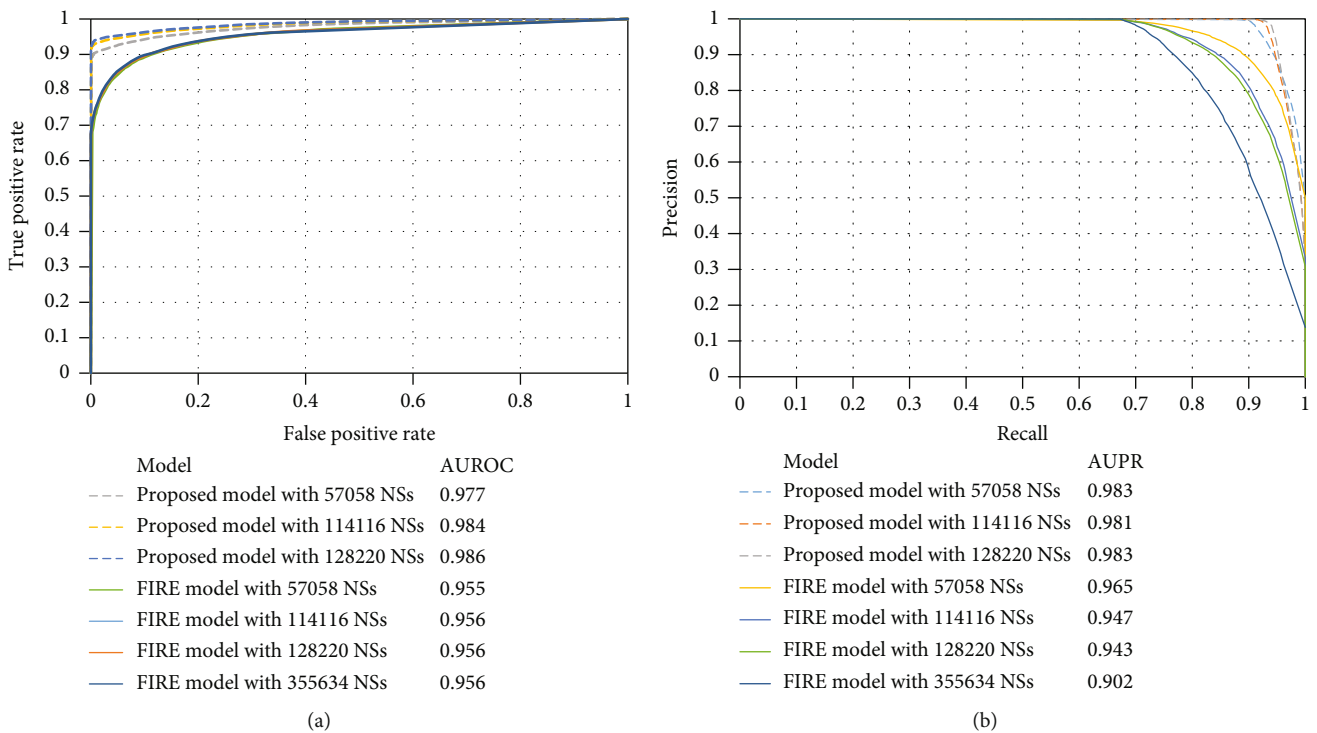


FIGURE 8: ROC and PR curves of models with different negative samples obtained by two different negative sample selection strategies. (a) The ROC curves; (b) the PR curves. The proposed model is constructed with negative samples obtained by the proposed strategy, whereas the FIRE model is built with negative samples obtained by FIRE. NS: negative sample.

analyses, which are shown in Figure 8. Clearly, the ROC and PR curves of the proposed model were always above the corresponding curves of the FIRE model. The AUROC and AUPR were 2.2% and 1.8% higher, respectively. Thus, the proposed model was better than the FIRE model. Second,

we compared the RF models with imbalanced positive and negative samples. In this case, negative samples were twice as many as positive samples. According to the results listed in Table 5 and Figure 8, the proposed model was also superior to the FIRE model. Third, the RF models with all samples

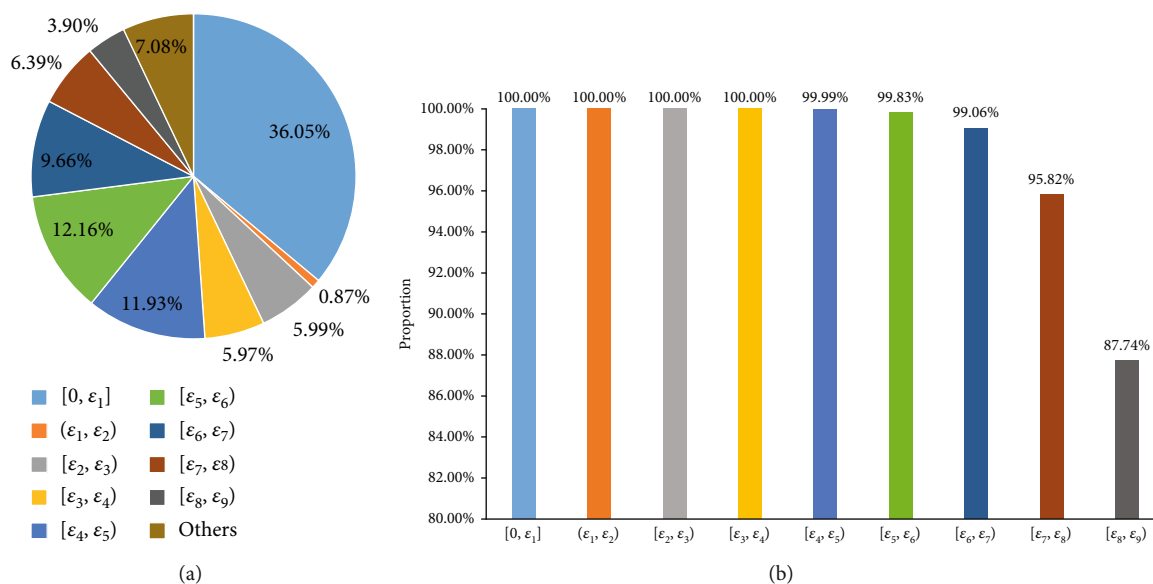


FIGURE 9: Breakdown of the negative samples selected by FIRE. (a) Distribution of such negative samples on nine parts of negative samples obtained by the proposed strategy. (b) Proportions of negative samples selected by FIRE in each part of the negative samples obtained by the proposed strategy.

in two negative sample sets were constructed to make comparison. The results are also listed in Table 5 and Figure 8. The MCC of the FIRE model was 0.812, which was 13.1% lower than that of the proposed model. As for AUROC and AUPR, they were 3% and 8.1% lower than those of the proposed model, respectively. It was also indicated that the proposed model was better than the FIRE model. Finally, considering the fact that negative samples selected by FIRE were much more than those filtered by the proposed strategy, we randomly selected 128,220 samples from the negative samples obtained by FIRE and used them to construct the RF model. The tenfold crossvalidation results are also listed in Table 5 and Figure 8. Clearly, such model was inferior to the proposed model with the same number of negative samples. According to the above arguments, the proposed models were superior to the FIRE models, proving that the proposed negative sample selection strategy can screen out negative samples with higher quality than FIRE.

From the above arguments, negative samples selected by the proposed strategy were of higher quality than those filtered by FIRE. Here, we provided an investigation to explain the reason. Figure 9(a) shows the distribution of 355,634 negative samples selected by FIRE on nine parts of negative samples mentioned in Negative Samples with Different Thresholds of Probability. Each of the nine parts contained several such negative samples. For example, the first part contained 36.05% such negative samples and the second part contained 0.87% such negative samples. This result suggested that our strategy can classify negative samples generated by FIRE into different parts, which contained negative samples with different quality. Furthermore, as shown in Figure 9(b), all negative samples generated by the proposed strategy with threshold ϵ_4 were selected by FIRE, and this proportion decreased with the increase of the threshold. Most negative samples in each part (more than 87%) were

also selected by FIRE. Thus, our strategy improved the evaluation scheme on negative samples and gave a more refined partition on negative samples. FIRE evaluated the quality of negative samples by only considering the direct links between drugs. If the distances between one drug and drugs sharing one side effect were all larger than one, such pair of drug and side effect was assigned a zero score and deemed as a negative sample with the highest quality by FIRE. FIRE did not consider the factor of distance. In fact, such pairs can be further classified. Pairs with long distances were clearly more likely to be actual negative samples. For the proposed strategy, it adopted the RWR algorithm to evaluate the quality of negative samples. Generally, pairs with long distances would be assigned low probabilities. Therefore, we can further classify negative samples selected by FIRE into many parts by setting different thresholds on the probability. However, FIRE cannot divide them because their scores were all zeros. All these induced the phenomenon shown in Figure 9, and it was the main reason why our strategy was better than FIRE.

4. Conclusions

This study proposed a novel negative sample selection strategy for the prediction of drug side effects. Under a small threshold, the negative samples are of high quality, indicating that it is not necessary to consider the balance of positive and negative samples. It is hopeful this strategy can give useful help for determining novel side effects of given drugs and new insights for dealing with similar biological and medicine problems.

Data Availability

The original data used to support the findings of this study are available at SIDER and in supplementary information files.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Shanghai [17ZR1412500], and Science and Technology Commission of Shanghai Municipality (STCSM) [18dz2271000].

Supplementary Materials

Table S1: the performance of the SVM models with different quality negative samples. Table S2: the performance of the ANN models with different quality negative samples. Figure S1: the ROC curves and PR curves of the SVM models with different quality negative samples. (A) The ROC curves; (B) the PR curves. Figure S2: the ROC curves and PR curves of the ANN models with different quality negative samples. (A) The ROC curves; (B) the PR curves. Figure S3: the performance of the SVM models on balanced datasets, in which negative samples, as many as positive samples, are randomly selected under different thresholds. (A) Six measurements; (B) the ROC curves; (C) the PR curves. Figure S4: the performance of the SVM models on imbalanced datasets, in which negative samples, twice as many as positive samples, are randomly selected under different thresholds. (A) Six measurements; (B) the ROC curves; (C) the PR curves. Figure S5: the performance of the ANN models on balanced datasets, in which negative samples, as many as positive samples, are randomly selected under different thresholds. (A) Six measurements; (B) the ROC curves; (C) the PR curves. Figure S6: the performance of the ANN models on imbalanced datasets, in which negative samples, twice as many as positive samples, are randomly selected under different thresholds. (A) Six measurements; (B) the ROC curves; (C) the PR curves. Figure S7: the MCCs yielded by the SVM models on three types of datasets. “All” means that all negative samples under the given threshold are selected; “Imbalanced” indicates that negative samples, twice as many as positive samples, under the given threshold are randomly selected; and “Balanced” indicates that negative samples, as many as positive samples, under the given threshold are randomly selected. Figure S8: the MCCs yielded by the ANN models on three types of datasets. “All” means that all negative samples under the given threshold are selected; “Imbalanced” indicates that negative samples, twice as many as positive samples, under the given threshold are randomly selected; and “Balanced” indicates that negative samples, as many as positive samples, under the given threshold are randomly selected. (*Supplementary Materials*)

References

- [1] E. Pauwels, V. Stoven, and Y. Yamanishi, “Predicting drug side-effect profiles: a chemical fragment-based approach,” *BMC Bioinformatics*, vol. 12, no. 1, p. 169, 2011.
- [2] M. Liu, Y. Wu, Y. Chen et al., “Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs,” *Journal of the American Medical Informatics Association*, vol. 19, no. e1, pp. e28–e35, 2012.
- [3] S. Jamal, S. Goyal, A. Shanker, and A. Grover, “Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models,” *Scientific Reports*, vol. 7, no. 1, p. 872, 2017.
- [4] S. Dey, H. Luo, A. Fokoue, J. Hu, and P. Zhang, “Predicting adverse drug reactions through interpretable deep learning framework,” *BMC Bioinformatics*, vol. 19, no. S21, p. 476, 2018.
- [5] Y. Zheng, H. Peng, S. Ghosh, C. Lan, and J. Li, “Inverse similarity and reliable negative samples for drug side-effect prediction,” *BMC Bioinformatics*, vol. 19, no. S13, 2019.
- [6] L. Chen, T. Huang, J. Zhang et al., “Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions,” *BioMed Research International*, vol. 2013, Article ID 485034, 8 pages, 2013.
- [7] E. Munoz, V. Novacek, and P. Y. Vandebussche, “Using drug similarities for discovery of possible adverse reactions,” *American Medical Informatics Association Annual Symposium Proceedings*, vol. 2016, pp. 924–933, 2016.
- [8] W. Zhang, Y. Chen, S. Tu, F. Liu, and Q. Qu, “Drug side effect prediction through linear neighborhoods and multiple data source integration,” in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 427–434, Shenzhen, China, 2016.
- [9] W. Zhang, F. Liu, L. Luo, and J. Zhang, “Predicting drug side effects by multi-label learning and ensemble learning,” *BMC Bioinformatics*, vol. 16, no. 1, 2015.
- [10] N. Atias and R. Sharan, “An algorithmic framework for predicting side effects of drugs,” *Journal of Computational Biology*, vol. 18, no. 3, pp. 207–218, 2011.
- [11] E. Muñoz, V. Nováček, and P.-Y. Vandebussche, “Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models,” *Briefings in Bioinformatics*, vol. 20, pp. 190–202, 2019.
- [12] Y. Yamanishi, E. Pauwels, and M. Kotera, “Drug side-effect prediction based on the integration of chemical and biological spaces,” *Journal of Chemical Information and Modeling*, vol. 52, no. 12, pp. 3284–3292, 2012.
- [13] Y. Niu and W. Zhang, “Quantitative prediction of drug side effects based on drug-related features,” *Interdisciplinary Sciences Computational Life Sciences*, vol. 9, no. 3, pp. 434–444, 2017.
- [14] X. Zhao, L. Chen, and J. Lu, “A similarity-based method for prediction of drug side effects with heterogeneous information,” *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.
- [15] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, “Predicting drug side effects with compact integration of heterogeneous networks,” *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.
- [16] Y. Ding, J. Tang, and F. Guo, “Identification of drug-side effect association via semi-supervised model and multiple kernel learning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, 2019.
- [17] Y. Ding, J. Tang, and F. Guo, “Identification of drug-side effect association via multiple information integration with centered kernel alignment,” *Neurocomputing*, vol. 325, pp. 211–224, 2019.
- [18] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, “STITCH: interaction networks of chemicals and proteins,” *Nucleic Acids Research*, vol. 36, pp. D684–D688, 2007.

- [19] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild et al., "STITCH 4: integration of protein-chemical interactions with user data," *Nucleic Acids Research*, vol. 42, no. D1, pp. D401-D407, 2013.
- [20] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949-958, 2008.
- [21] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, no. 1, p. 343, 2010.
- [22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
- [24] L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng, and K. C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS One*, vol. 7, no. 4, article e35254, 2012.
- [25] L. Chen, J. Lu, N. Zhang, T. Huang, and Y. D. Cai, "A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes," *Molecular BioSystems*, vol. 10, no. 4, pp. 868-877, 2014.
- [26] L. Chen, C. Chu, Y. H. Zhang et al., "Identification of drug-drug interactions using chemical interactions," *Current Bioinformatics*, vol. 12, no. 6, pp. 526-534, 2017.
- [27] L. Chen, T. Liu, and X. Zhao, "Inferring anatomical therapeutic chemical (ATC) class of drugs using shortest path and random walk with restart algorithms," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1864, no. 6, pp. 2228-2240, 2018.
- [28] L. L. Hu, C. Chen, T. Huang, Y. D. Cai, and K. C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PLoS One*, vol. 6, no. 12, article e29491, 2011.
- [29] Y. F. Gao, L. Chen, Y. D. Cai, K. Y. Feng, T. Huang, and Y. Jiang, "Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins," *PLoS One*, vol. 7, no. 9, article e45944, 2012.
- [30] L. Chen, Y.-H. Zhang, Z. Zhang, T. Huang, and Y.-D. Cai, "Inferring novel tumor suppressor genes with a protein-protein interaction network and network diffusion algorithms," *Molecular Therapy - Methods & Clinical Development*, vol. 10, pp. 57-67, 2018.
- [31] S. Lu, Y. Yan, Z. Li et al., "Determination of genes related to uveitis by utilization of the random walk with restart algorithm on a protein-protein interaction network," *International Journal of Molecular Sciences*, vol. 18, no. 5, p. 1045, 2017.
- [32] J. Zhang, Y. Suo, M. Liu, and X. Xu, "Identification of genes related to proliferative diabetic retinopathy through RWR algorithm based on protein-protein interaction network," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1864, pp. 2369-2375, 2018.
- [33] F. Yuan and W. Lu, "Prediction of potential drivers connecting different dysfunctional levels in lung adenocarcinoma via a protein-protein interaction network," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1864, pp. 2284-2293, 2018.
- [34] Y. Zhang, L. Dai, Y. Liu, Y. Zhang, and S. Wang, "Identifying novel fruit-related genes in *Arabidopsis thaliana* based on the random walk with restart algorithm," *PLoS One*, vol. 12, no. 5, article e0177017, 2017.
- [35] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31-36, 1988.
- [36] G. Landrum, *RDKit: open-source cheminformatics*, GitHub, 2006, <http://www.rdkit.org>.
- [37] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *Journal of the American Chemical Society*, vol. 125, no. 39, pp. 11853-11865, 2003.
- [38] M. Hattori, N. Tanaka, M. Kanehisa, and S. Goto, "SIMCOMP/SUBCOMP: chemical structure search servers for network analyses," *Nucleic Acids Research*, vol. 38, Supplement, pp. W652-W656, 2010.
- [39] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan, Kaufmann, San Francisco, CA, USA, 2nd edition, 2005.
- [40] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442-451, 1975.
- [41] D. Powers, "Evaluation: from precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37-63, 2011.
- [42] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International joint Conference on artificial intelligence*, pp. 1137-1145, Montreal, QC, Canada, 1995.
- [43] J.-P. Zhou, L. Chen, and Z.-H. Guo, "iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs," *Bioinformatics*, vol. 36, no. 5, pp. 1391-1396, 2020.
- [44] L. Chen, S. Wang, Y. H. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582-26590, 2017.
- [45] J. Che, L. Chen, Z. H. Guo, S. Wang, and Aorigele, "Drug target group prediction with multiple drug networks," *Combinatorial Chemistry & High Throughput Screening*, vol. 22, 2019.
- [46] Z. Cheng, K. Huang, Y. Wang, H. Liu, J. Guan, and S. Zhou, "Selecting high-quality negative samples for effectively predicting protein-RNA interactions," *BMC Systems Biology*, vol. 11, no. S2, p. 9, 2017.

Research Article

Fusion of FDG-PET Image and Clinical Features for Prediction of Lung Metastasis in Soft Tissue Sarcomas

Jin Deng , Weiming Zeng , Yuhu Shi , Wei Kong , and Shunjie Guo 

College of Information Engineering, Shanghai Maritime University, 1550 Haigang Ave., Shanghai 201306, China

Correspondence should be addressed to Weiming Zeng; zengwm86@163.com

Received 1 February 2020; Accepted 4 April 2020; Published 5 May 2020

Guest Editor: Lin Lu

Copyright © 2020 Jin Deng 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.

Extracting massive features from images to quantify tumors provides a new insight to solve the problem that tumor heterogeneity is difficult to assess quantitatively. However, quantification of tumors by single-mode methods often has defects such as difficulty in features extraction and high computational complexity. The multimodal approach has shown effective application prospects in solving these problems. In this paper, we propose a feature fusion method based on positron emission tomography (PET) images and clinical information, which is used to obtain features for lung metastasis prediction of soft tissue sarcomas (STSs). Random forest method was adopted to select effective features by eliminating irrelevant or redundant features, and then they were used for the prediction of the lung metastasis combined with back propagation (BP) neural network. The results show that the prediction ability of the proposed model using fusion features is better than that of the model using an image or clinical feature alone. Furthermore, a good performance can be obtained using 3 standard uptake value (SUV) features of PET image and 7 clinical features, and its average accuracy, sensitivity, and specificity on all the sets can reach 92%, 91%, and 92%, respectively. Therefore, the fusing features have the potential to predict lung metastasis for STSs.

1. Introduction

Sarcomas are a highly heterogeneous group of tumors classified according to the similar adult tissue types in tissue occurrence [1]. It is characterized by invasive or destructive growth that can recur and by distant metastasis [2]. As one of the sarcomas, soft tissue sarcomas (STSs) can occur anywhere in the body, and 59% of which originate in an extremity [3]. Unfortunately, 10%-20% of the patients with sarcomas or STSs have distant metastasis at the time of diagnosis. The metastasis rate is approximately 30%-40% in the course of follow-up, of which lung metastasis accounts for about 90% [4-6]. Moreover, there is a great deficiency in the cognition of the prognostic factors of lung metastatic tumor resection and the recurrence rate after resection is high [7]. Therefore, early screening and prediction of lung metastasis can help patients with STSs find corresponding self-treatment measures at an early stage and improve the survival rate of patients.

The most common method to evaluate the risk of lung metastasis is to study the heterogeneity of tumors from histopathological samples, while the biological relationships

between different clonal subgroups or clones and microenvironments in solid tumors such as STSs are still unclear, so that the information obtained from the samples is affected by the sampled region, which is not necessarily representative [8]. Therefore, it is hard to study the heterogeneity of tumors from the point of view at the molecular level because solid cancers are spatial and temporal heterogeneous. Lambin et al. have suggested that extracting a large number of features from medical images can solve this problem because the radiomic feature has the ability to capture intratumoural heterogeneity in a noninvasive [9]. Several studies have predicted the effect of lymph node metastasis and adjuvant radiotherapy or chemotherapy in preoperative colorectal cancer by using radiomic features [10, 11]. Corino et al. performed radiomic analysis of STSs to distinguish moderate and high lesions [12]. Valliã Res et al. extracted a large number of texture features from 2-deoxy-2-[18F]fluoro-D-glucose (FDG) positron emission tomography (PET) and magnetic resonance imaging (MRI) data for the construction of a STS lung metastasis prediction model [13]. However, the cost of acquiring multiple modal images at the same time is

high and may not be affordable for some patients. Also, the image acquisition of different modes is complex, and different image information sets tend to obtain several overlapping feature information. In addition, the acquisition of a large amount of information increases the complexity of the model construction and the time complexity of the operation. The features obtained from only a single image are limited, and more other accessible modal data may be overlooked such as clinical data. Actually, the fusion of image and other modal information can help us to obtain features from multiple aspects that are used to build a more accurate and stable model. There are a few studies that have used the multimodal method to achieve great results; for example, Aerts et al. quantified the tumor microenvironment by fusing imaging, gene, and clinical information to quantify tumor gene heterogeneity in the early stage [14]. Tingdan et al. developed and validated a clinical radiomics nomogram for preoperative prediction of lung metastasis in colorectal cancer patients [15].

PET image is a kind of reaction molecular metabolism imaging. When the disease is changing in the early stage of the molecular level and the morphology of the lesion area has not been abnormal, the lesion can be found by PET examination. Compared with other types of images, PET has the characteristics of high sensitivity, high specificity, and better security [16]. Therefore, this study extracted the features from two kinds of data including PET image and clinical data, and then the feature fusion was performed, which was applied to the subsequent prediction model construction. For the problem of model construction, it is expected that the predictive model can be as simple as possible and has a good predictive effect. The aim of the simple model is actually to expect it to be used as a representative feature as possible for the model construction. Hence, it is particularly important to choose higher importance features from a large number of features. There are two functions for feature selection including reducing the number of features and lowering the dimension of features, which are both used to make the model generalization more powerful by reducing the overfitting and by enhancing the understanding of features. Taking into account that the random forest algorithm has the ability to analyze the features of complex interaction classification and has good robustness for noise data or data with missing values, its variable importance measure can be used as a feature selection tool for high-dimensional data [17]. Therefore, the random forest method is applied to extract higher-contribution features from multimode data. Then, they are used as the input for a back propagation (BP) neural network with the superior ability of nonlinear mapping, self-learning, self-adaptive, generalization, and fault tolerance to construct the lung metastasis prediction model [18]. The results showed that the model constructed by combining the features of image and clinical data has a better performance in all data sets. Furthermore, it could be found that only the top PET features and clinical features achieved a higher accuracy rate of more than 90%. These features are strongly correlated with lung metastasis and may be used as a label for lung metastasis prediction of STSs.

TABLE 1: Patient and tumor characteristics.

Clinical parameters	
Age, years (mean \pm SD)	54.8 \pm 16.0
Gender, n (%)	
Male	24 (47.1)
Female	27 (52.9)
Histology, n (%)	
Malignant fibrous histiocytomas	17 (33.3)
Liposarcoma	11 (21.6)
Leiomyosarcoma	10 (19.6)
Synovial sarcoma	5 (9.8)
Extraskeletal bone sarcoma	4 (7.8)
Fibrosarcoma	1 (2.0)
Other	3 (5.9)
Grade, n (%)	
High	28 (54.9)
Intermediate	15 (29.4)
Low	5 (9.8)
Unknown	3 (5.9)
Metastases, n (%)	
Lung	19 (37.3)
Other	5 (9.8)
None	27 (52.9)
Time, days (mean \pm SD)	
Diagnosis to outcome	285.7 \pm 252.3
Diagnosis to last follow-up	849 \pm 447.4
Status, n (%)	
No evidence of disease	26 (51.0)
Alive with disease	9 (17.6)
Dead	15 (29.4)

Note: SD: standard deviation; n : number; diagnosis to outcome: days elapsed between the date of diagnosis of primary STS (biopsy) and the date of diagnosis of recurrence or metastases; diagnosis to last follow-up: days elapsed between the date of diagnosis of primary STS (biopsy) and the date of last-follow-up (or death, if applicable).

2. Methods

2.1. Data Sources and Preprocessing. The FDG-PET imaging data of 51 patients with STSs were included in this study, and the corresponding clinical data were downloaded from The Cancer Imaging Archive (TCIA). All patients underwent pretreatment FDG-PET scanned between November 2004 and November 2011, during which 19 patients developed lung metastases [13]. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median inplane resolution was 5.47×5.47 mm² (range: 3.91-5.47 mm). The details of samples could be found in Table 1.

We divided these samples into two groups, *LungMets* and *NoLungMets*. *NoLungMets* were patients that did not develop lung metastases and *LungMets* were patients that eventually developed lung metastases. The FDG-PET imaging data also contained contours of the 3D tumor region of each

sample drawn by an expert radiation oncologist. To better understand the difference of images between the two groups, the region of interest (ROI) is a tumor area which is extracted according to the lesion contour mask mapped onto the original image. Therefore, ROI volumes of 51 patients are obtained for further analysis.

2.2. Feature Extraction. Considering that the standard uptake values (SUV) as a semiquantitative index in PET can be effectively applied in the evaluation of benign and malignant tumors and evaluation of curative effect, the FDG-PET volume is first converted to SUV maps and then a square root transform is applied to help stabilize the PET noise in the image.

These features are extracted from three aspects, and all of which are derived from the ROI regions. In Table 2, there are 67 features including 5 SUV metrics features, 5 types of texture features, and 6 types of clinical features. The study uses five types of texture features, namely, global, gray-level co-occurrence matrix (GLCM) [19], gray-level run-length matrix (GLRLM) [20–23], gray-level size zone matrix (GLSZM) [20–23], and neighborhood gray-tone difference matrix (NGTDM) [24].

The corresponding feature vector is calculated and obtained for each feature. SUV-related features can be obtained by simple mathematical calculations. Each extraction method of texture feature has a corresponding calculation formula, which can be obtained by corresponding references [19–24]. Then, we calculated all the texture feature values corresponding to each sample by the formula of each texture feature. Furthermore, considering that the most clinical features are presented in text form, a coding method called one-hot is applied to extract text features [25]. For example, the *Sex* feature includes male and female, the male is represented as 1 0 and the female is denoted as 0 1. Similarly, feature *Status* is coded as Alive (1 0 0), Alive with disease (0 1 0) and Dead (0 0 1). These feature vectors make up a feature matrix with a size of 51×67 . Row denotes the sample and each column is a feature vector.

2.3. Feature Selection Based on Random Forest. A random forest is an integrated classifier with a set of decision tree classifiers that can be expressed as $\{h(X, \theta_k), k = 1, 2, \dots, K\}$, where $\{\theta_k\}$ is a random vector obeying independent and identical distribution. K represents the number of decision trees in random forest. The optimal classification result is determined by the voting of each decision tree classifier when given an argument X [17].

The variable importance assessment is a significant feature of random forest algorithms. In this study, we use a variable importance measure based on the classification accuracy of out-of-bag (oob) data. The evaluation criterion of this method is the average reduction of the classification accuracy after the slight disturbance of the independent variable of the oob data and classification accuracy before the disturbance.

Assuming that there are bootstrap samples $b = 1, 2, \dots, B$, where B represents the number of training samples, the

variable importance \overline{D}_j of feature X_i based on classification accuracy is calculated by the following steps: Firstly, the decision tree T_b is constructed based on training samples after setting the value of b to 1, and then the oob data is defined as L_b^{oob} . After that, the oob data is classified by using the T_b , and the number of correct classifications is calculated as R_b^{oob} . For feature $X_j (j = 1, 2, \dots, N)$, the value of the feature X_j in L_b^{oob} is disturbed, and the data set after disturbance is defined as L_{bj}^{oob} . The number of L_{bj}^{oob} is classified by T_b , and the count of correct classifications is calculated as R_{bj}^{oob} . The same steps are performed on other features. Finally, the importance \overline{D}_j of feature X_j is calculated by the formula

$$\overline{D}_j = \frac{1}{B} \sum_{b=1}^B (R_b^{oob} - R_{bj}^{oob}). \quad (1)$$

2.4. Back Propagation (BP) Neural Network Model. The principle of BP neural network is that the gradient descent method is used to adjust the weights and thresholds, so that the mean square error value of the actual output of the network and the expected output is minimal. The training simulation process is presented as follows.

Firstly, the BP neural network structure is determined and the input layer to the implicit layer weight value w_{ij} , the implicit value to the output layer weight v_{ij} , the implicit layer threshold θ_j , and the output layer threshold γ_t are assigned. Then the training samples (P^k, R^k) are randomly selected to provide to the network. After that, the input of each element of the implicit layer S_j is calculated by using the output sample P^k , the connection weight value w_{ij} , and the threshold value θ_j , and then the output B_j of the implicit layer unit is calculated by using the S_j through the transfer function as follows:

$$S_j = \sum_{i=1}^m w_{ij} p_i^k - \theta_j, \quad (2)$$

$$B_j = f(S_j).$$

Next, the output L_t of the output layer units is calculated using the output B_j , weight value v_{jt} , and threshold γ_t of the middle layer, and then the response C_t of the output layer unit is calculated by the transfer function so that the following formulas are obtained:

$$L_t = \sum_{j=1}^n v_{jt} B_j - \gamma_t, \quad (3)$$

$$C_t = f(L_t).$$

The generalization error d_t of the output layer can be calculated using the expected output R^k and the network actual output C_t , and the generalization error e_j of each unit in the

TABLE 2: SUV metrics features and Clinical features used in this study.

Type	Name	Description
SUV metrics	SUVmax	Maximum SUV of the tumour region
	SUVpeak	Average of the voxel with maximum SUV within the tumour region and its 26 connected neighbors
	SUVmean	Average SUV value of the tumour region
	aucCSH	Area under the curve of the cumulative SUV volume histogram describing the percentage of total tumour volume above a percentage threshold of maximum SUV
	PercentInactive	Percentage of the tumour region that is inactive. A threshold of $0.005 \times (\text{SUVmax})^2$ followed by closing and opening morphological operations were used to differentiate active and inactive regions on FDG-PET scans
Textures	Global	Variance
		Skewness
		Kurtosis
		Energy
		Contrast
	GLCM	Entropy
		Homogeneity
		Correlation
		SumAverage
		Variance
		Dissimilarity
		AutoCorrelation
		Textures
LRE (long run emphasis)		
GLN (gray-level nonuniformity)		
RLN (run-length nonuniformity)		
RP (run percentage)		
LGRE (low gray-level run emphasis)		
HGRE (high gray-level run emphasis)		
SRLGE (short run low gray-level emphasis)		
SRHGE (short run high gray-level emphasis)		
LRLGE (long run low gray-level emphasis)		
LRHGE (long run high gray-level emphasis)		
GLV (gray-level variance)		
RLV (run-length variance)		
GLSZM	SZE (small-zone emphasis)	
	LZE (large-zone emphasis)	
	GLN (gray-level nonuniformity)	
	ZSN (zone-size nonuniformity)	
	ZP (zone percentage)	
	LGZE (low gray-level zone emphasis)	
	HGZE (high gray-level zone emphasis)	
	SZLGE (small-zone low gray-level emphasis)	
	SZHGE (small-zone high gray-level emphasis)	
	LZLGE (large-zone low gray-level emphasis)	
LZHGE (large-zone high gray-level emphasis)		
GLV (gray-level variance)		
ZSV (zone-size variance)		

TABLE 2: Continued.

Type	Name	Description
		Coarseness
		Contrast
	NGTDM	Busyness
		Complexity
		Strength
	Age	Age
	Sex	Male
		Female
		Radiotherapy + surgery + chemotherapy
	Treatment	Radiotherapy + surgery
		Surgery + chemotherapy
		Alive
	Status	Alive with disease
		Dead
Clinical		High
	Grade	Intermediate
		Low
		Liposarcoma
		Leiomyosarcoma
		Synovial sarcoma
	MSKCC type	Malignant fibrous histiocyctomas
		Extraskelatal bone sarcoma
		Fibrosarcoma
		Other

middle layer can be also calculated based on three parameters including v_{jt} , d_t , and B_j .

$$\begin{aligned} d_t &= (r - C_t) \cdot C_t(1 - C_t) \\ e_j &= \left[\sum_{t=1}^p d_t \cdot v_{jt} \right] B_j(1 - B_j) \end{aligned} \quad (4)$$

Then, the connection weight v_{jt} and threshold γ_t are corrected by using the generalization error d_t of the output layer units and the output B_j of each unit in the middle layer.

$$\begin{aligned} v_{jt}(N+1) &= v_{jt}(N) + \alpha \cdot d_t \cdot B_j, \\ \gamma_t(N+1) &= \gamma_t(N) + \alpha \cdot d_t. \end{aligned} \quad (5)$$

Therefore, the fixed connection weight w_{ij} and threshold θ_j can be obtained as follows:

$$\begin{aligned} w_{ij}(N+1) &= w_{ij}(N) + \beta \cdot e_j \cdot p_i, \\ \theta_j(N+1) &= \theta_j(N) + \beta \cdot e_j, 0 < \beta < 1. \end{aligned} \quad (6)$$

Furthermore, the next train sample is randomly selected to provide to the network according to the previous method of training until the training samples are fully trained.

3. Results

3.1. Feature Selection Based on Random Forest. Feature selection can not only improve the performance of the model but also help us to understand the characteristics of the data and the underlying structure, which plays a significant role in the further improvement of model and algorithm. In this study, the number of trees in random forest is set to 250, and then there are 67 features including 5 SUV metrics features, 5 types of texture features, and 6 types of clinical features. In order to explore the contribution of these features to the prediction model, random forest was applied to sort these features. 36 features were selected whose importance values were more than 0.01. Moreover, a T test is used to verify the performance of random forest method in this study. 25 significant features were retained in these features by using the T test with a confidence level of 95% which are shown in Figure 1.

As shown in Figure 1, these selected features contain 6 clinical features and 19 image features including 5 SUV features and 14 types of texture features. Furthermore, 6 of the top 10 features are clinical features, including age, status, treatment, and MSKCC type, and the other features belong to image features namely SUV features, including SUVpeak, SUVmax, aucCSH, and PercentInactive. Therefore, there is no any texture feature.

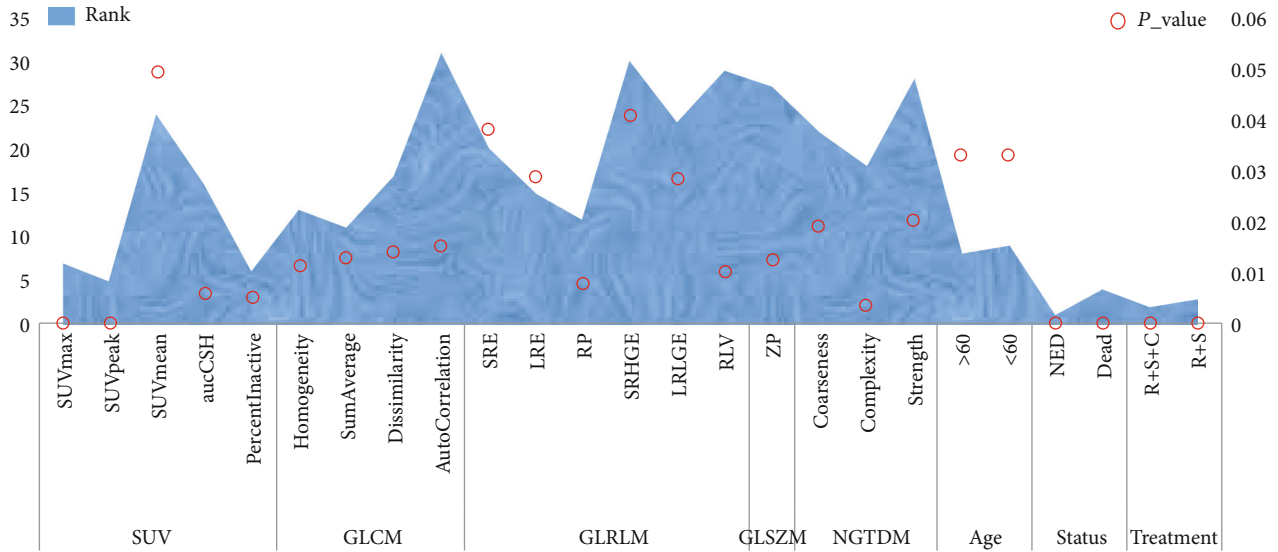


FIGURE 1: The 25 significant features based on random forest and T test. The transverse axis represents different characteristics, the main longitudinal axis represents the ranking of the feature, and the secondary longitudinal axis represents the significance level of the feature.

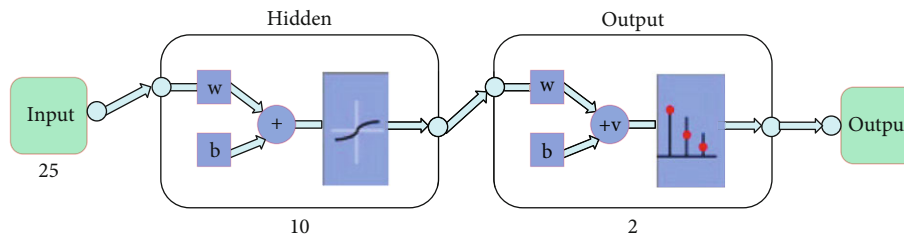


FIGURE 2: Neural network model. The numbers in the figure represent the number of neurons in each layer.

3.2. *BP Neural Network.* After obtaining those features, the BP neural network model was constructed, including 1 input layer, 1 hidden layer, and 1 output layer as shown in Figure 2.

For this neural network model, there are 25 most important features so that the number of neurons for input layer is 25. In order to make the model as simple as possible and the time complexity lower, the model in this paper is a simple three-layer model that is the hidden layer is single. The number of neurons for the output layer is 2 because the output layer contains two groups: *LungMets* and *NoLungMets*.

In this study, the *sigmoid* function was applied to be the activation function and the method of adaptive learning rate adjustment used to avoid local optimization and overfitting [26]. Then, 51 samples were randomly disrupted and divided into three types according to the proportion of 70%, 15%, and 15%, which included 35 training samples, 8 test samples, and 8 validation samples. When the number of iterations reached 1000 times or the gradient value is less than 0.001, the training model was considered to have been trained. Furthermore, in order to overcome the impact of small sample volume and sample specificity on the model, the samples were randomly divided 10 times and then repeated the above process in our study.

In order to measure the performance of the model, three indicators were used in this study, including accuracy, speci-

ficity, and sensitivity, as shown in Figure 3(a). In addition, the best validation performance of a randomly selected model is shown in Figure 3(b).

It is expected that the model has a high specificity and is sensitive on the basis of high accuracy. In other words, we hope that the model will have a better effect on both *LungMets* and *NoLungMets*. In terms of the total performance of the model, the average accuracy is 92%, and the specificity and sensitivity are 89% and 94%, respectively. Moreover, the model can achieve a good performance in training and validation set as expected. In fact, the results of the test set are more concerned because the model construction is based on the training set and the validation set; the test set is actually not involved in the construction of the model and completely independent of other sets. Therefore, the result of the test set is the standard of model performance evaluation; it can be seen from Figure 3(a) that the average accuracy, specificity, and sensitivity of the test set reached 90%, 87%, and 92%, respectively. In addition, the best validation performance is shown in Figure 3(b); the model has been trained after iterating 43 times, and the gradient mean square error is less than 0.001. At the same time, it can be seen on the validation set that the overall trend of the curve is also in the gradient drop. These evidences not only show that this model has good stability but also show that the

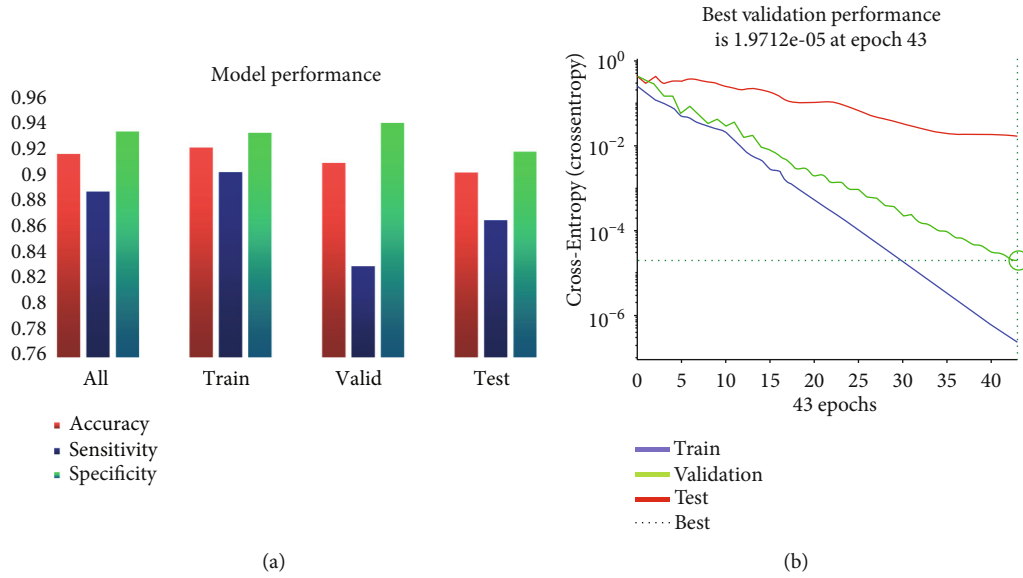


FIGURE 3: (a) The performance of model including accuracy, sensitivity, and specificity from different data sets. (b) Best validation performance of the neural network.

selected features can predict the lung metastasis in the soft tissue sarcomas.

4. Discussion

In order to further confirm whether the prediction effect after feature fusion is really better than that of the single type of features and to avoid the occurrence of chance, this study merely compares the final results of feature fusion with those of image features or clinical features as shown in Figure 4.

Compared with the original features, the effect of prediction is obviously improved based on the selected features by using random forest. For the test set, the prediction accuracy based on the original feature is 83%, while the model prediction accuracy after feature extraction reaches 90%, and the sensitivity of the model is increased by 16%. Therefore, the random forest method can effectively extract the features of higher contribution to the prediction model from the original features and can greatly improve the specificity and sensitivity of the prediction model. In addition, in order to verify the necessity of feature fusion based on multimodal data, the features of single modal data are used to establish a prediction model. The evaluation results of model according to the three measurement indices show that the prediction performance of the feature fusion model is better than that using a single class of modal data at the level of all data sets. For the test set, the multimodal features can obtain higher accuracy and specificity, and although image features also reveal a high average value of sensitivity, it cannot characterize its high sensitivity because the variance is actually too large. Moreover, the sensitivity is not even exceeding 70% in the training and test set because the image features alone are used to construct the model so that the model cannot be trained very well.

It is worth mentioning that the top 10 features of the 25 features include 3 SUV features and 4 types of clinical

features, but these do not contain texture features. Therefore, the prediction model with 10 features as the model input was constructed in this study, and it was compared with the previous model with 25 features as shown in Figure 5.

It can be seen from Figure 5(a) that the top ten features of the contribution ranking are mainly 5 types, of which the image features include three SUV features and the clinical features include the *Status*, *Treatment*, *Age*, and *MSKCC type*. In Figure 5(b), it can be found that the prediction accuracy of the model without texture features decreased by less than 1% compared to the model with 25 features, and the sensitivity increased significantly although the specificity does not seem to be as ideal. These results suggest that the effect of the 10 features is similar to that of the 26 features, which means that it may be not necessary to do a lot of complex texture feature calculations to obtain the same good prediction effect, while the basic clinical features and SUV features are easier to obtain to get than texture features.

For PET images, SUV is widely used in the identification and prognosis prediction of benign and malignant tumors. *SUVmax* is always used as the initial index of benign and malignant tumor due to the characteristics of simple operation, good repeatability and not affected by the sketch area of interest. Compared to *SUVmax*, *SUVpeak* overcomes the problem of insensitive to image noise but is sensitive to regions of interest, and *PercentInactive* denotes the percentage of the inactive tumour region. A threshold of $0.005 \times (SUVmax)^2$ followed by closing and opening morphological operations is used to differentiate active and inactive regions on FDG-PET scans. For patients with lung metastasis, a vast majority of tumors are at a low differentiation stage, and these characteristics of SUV are indicators to distinguish the low differentiation of tumors. Furthermore, we calculated the correlation between these top features and lung metastasis events, which is the label shown in Figure 6(a). In order to verify the relationship between the most contributing feature

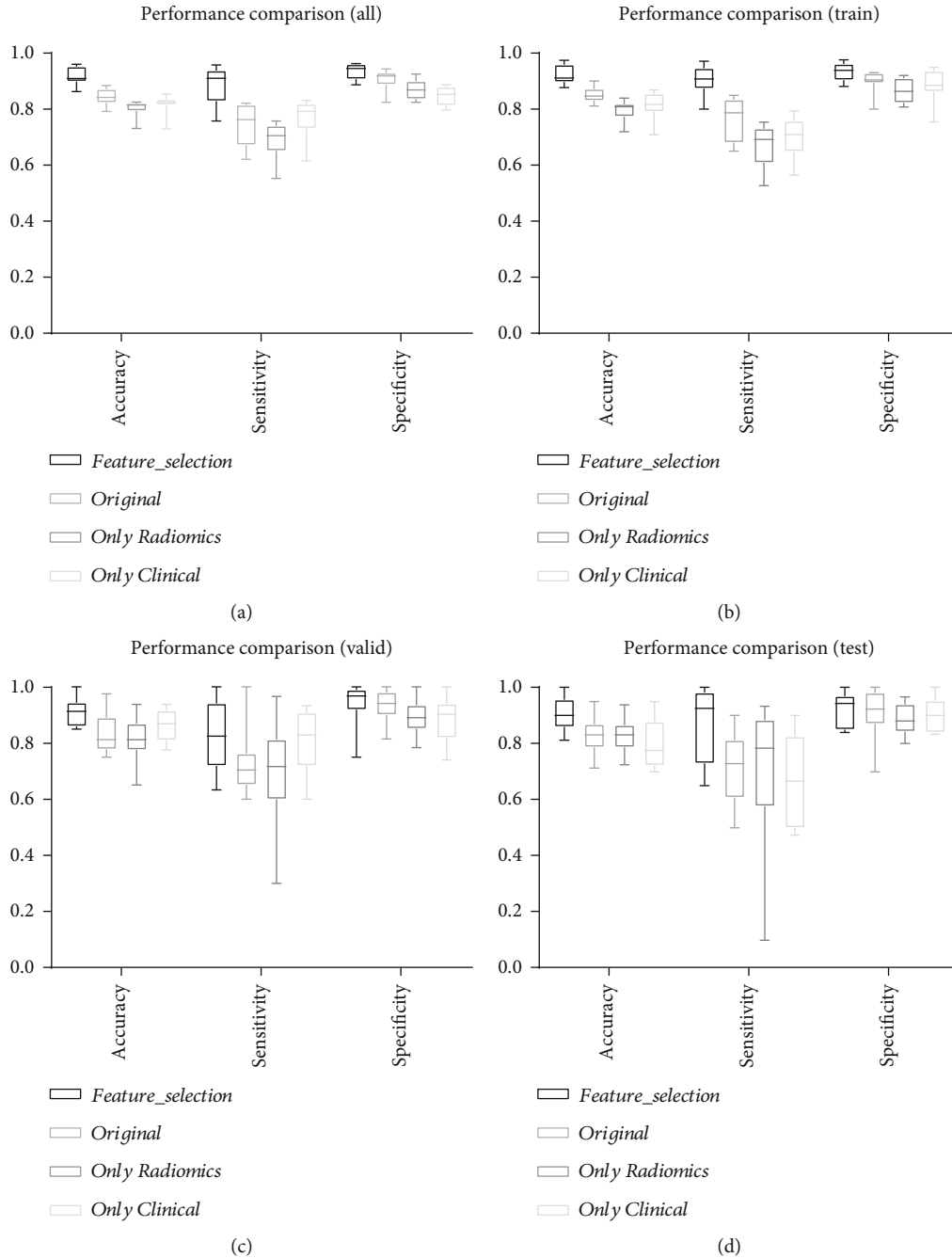


FIGURE 4: Performance comparison of feature selection. (a) represents the overall performance from the perspective of all data sets. (b–d) denote the performance from the three types of data sets including training, validation, and test set, respectively. *Feature_selection* represents 24 features selected by random forest and T test methods. *Original* denotes 67 features including 48 image features and 16 clinical features. *Only Radiomics* represents 48 image features, and *Only Clinical* denotes 16 clinical features.

Status and lung metastasis, the clinical data of 254 sarcoma samples with complete information was downloaded from the TCGA database, including 220 samples of *NoLungMets* and 34 *LungMets*, and then these sets of information were applied to calculate the survival curve of sarcoma patients, as shown in Figure 6(b).

It can be seen from Figure 6(a) that there is over medium even strong correlation between most features and label, especially for the *Status*, *Treatment*, and *SUV* features. More-

over, the correlation between features is weak, except for the features of the same types that are very strong, such as *SUVmax*, *SUVpeak*, *Age*, and *Status*. It is also completely understandable, such as the feature *Age* is either greater than 60 or less than 60. Therefore, these types of nonrelated features are highly representative and can be used as a feature of the prediction model effectively. Figure 6(b) shows that the data of *LungMets* and *NoLungMets* has a significant difference in patient survival time, which demonstrates that

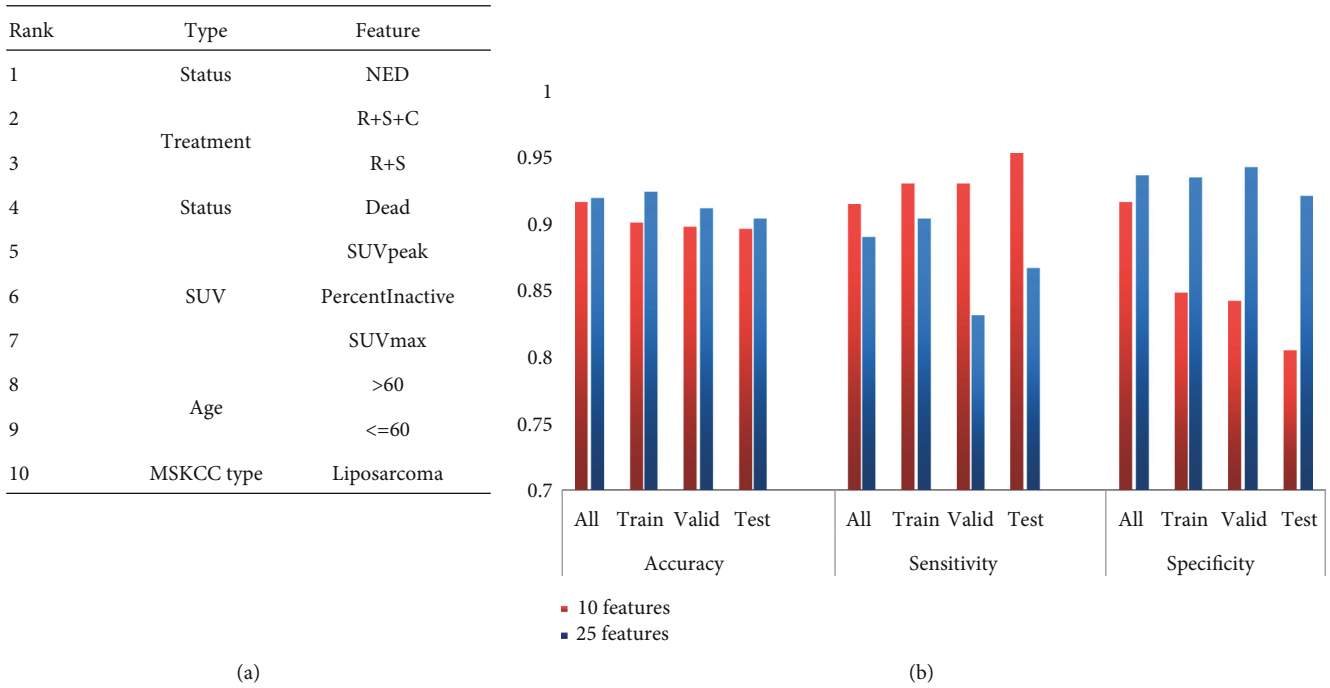


FIGURE 5: (a) The 10 top features of the contribution degree to prediction model. (b) Comparison of model performance based on different features.

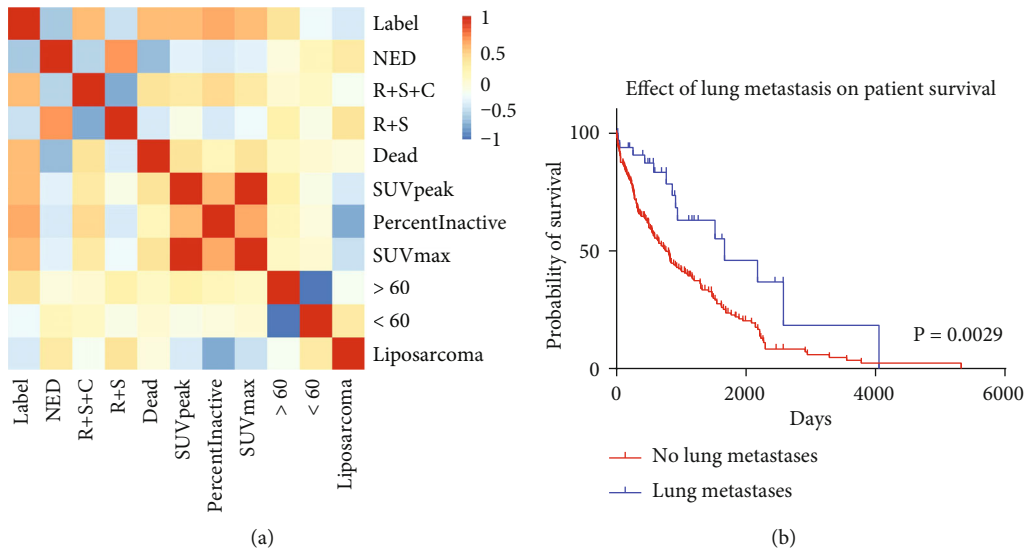


FIGURE 6: (a) Correlation between features and label. (b) Effect of lung metastasis on patient survival. The blue and red lines represent the survival probability of patients with *LungMets* and *NoLungMets*, respectively. The *x*-axis shows survival time of the patients, and the *y*-axis shows the survival probability of patients.

the correlation between survival state and lung metastasis is strong. In other words, our results suggest that the feature *Status* is very helpful for the prediction of lung metastasis, and it is easy to obtain the information about the state of survival in practical clinical.

With the development of comprehensive treatment of tumors and the prolongation of survival of cancer patients, the incidence of lung metastatic tumors is increasing. How-

ever, in the past, there were few studies on lung metastasis prediction of soft tissue tumors. Valliã`Res et al. used PET and MRI image data to construct different prediction models, and obtained a considerable prediction accuracy rate by selecting the optimal model [13]. Their study mainly used a large number of texture features with a large number of different parameters. There were more than 9000 texture features extracted from PET data according to different

parameters that resulted in excessive time complexity. In our study, the texture features under the optimal parameters were selected, which were merged with SUV features to construct the prediction model for the lung metastasis prediction as shown in Figure 5. It can be drawn from the figure that the performance of models constructed with only image features is significantly lower than that of images and clinical fusion features. Moreover, compared with the best model proposed in [13] based on PET data, the specificity of model was significantly improved, which overcomes the problem of using a large number of texture features. In fact, each texture feature corresponds to a texture feature algorithm, which was complex in the implementation of feature acquisition.

5. Conclusion

Based on the complementarity between different modal data features, extracting features from images and clinical data separately provides a new idea to construct predictive models. In this study, the texture and SUV features were extracted from the PET image, and the features of age, gender, and others were extracted from the clinical data. Then, all features were sorted by random forest and two-sample T test. The selected features were constructed using the BP neural network to predict the model. The results showed that the performance of multimodal feature fusion was better than that of the image data or clinical data alone. At the same time, the study further analyzed the top 10 significant features, and these features were applied to construct predictive model. It was found that the performance of the model could still achieve the previous effects without the presence of texture features, which were hard to obtain. Furthermore, the method proposed in this study could effectively select high-performance features to construct a prediction model of lung metastasis in STSs, and a high predictive performance was achieved in all data sets. In the future, we hope that this method can integrate more modal data to construct a more effective model to achieve better results, including molecular data such as genes and proteins. At the same time, this method can be extended to other prediction problems such as tumor staging and degree of tumor differentiation.

Data Availability

The data used to support the findings of this study are from previously reported studies and public database, which have been cited.

Conflicts of Interest

The authors declare no conflict of interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 31870979 and 61906117), Natural Science Foundation of Shanghai (No. 18ZR1417200), and Shanghai Sailing Program (No. 19YF1419000). We also would like to acknowledge the individuals and institutions

that have provided data for this collection: McGill University, Montreal, Canada—special thanks are due to Martin Vallières of the Medical Physics Unit.

References

- [1] L. A. Doyle, "Sarcoma classification: an update based on the 2013 World Health Organization classification of tumors of soft tissue and bone," *Cancer*, vol. 120, no. 12, pp. 1763–1774, 2014.
- [2] W. D. Tap, Z. Papai, B. A. van Tine et al., "Doxorubicin plus evofosfamide versus doxorubicin alone in locally advanced, unresectable or metastatic soft-tissue sarcoma (TH CR-406/SARC021): an international, multicentre, open-label, randomised phase 3 trial," *Lancet Oncology*, vol. 18, no. 8, pp. 1089–1103, 2017.
- [3] J. N. Cormier and R. E. Pollock, "Soft tissue sarcomas," *British Medical Journal*, vol. 54, no. 2, pp. 94–109, 2010.
- [4] A. J. Chou, E. S. Kleinerman, M. D. Krailo et al., "Addition of muramyl tripeptide to chemotherapy for patients with newly diagnosed metastatic osteosarcoma: a report from the Children's oncology group," *Cancer*, vol. 115, no. 22, pp. 5339–5348, 2010.
- [5] N. C. Daw, A. J. Chou, N. Jaffe et al., "Recurrent osteosarcoma with a single pulmonary metastasis: a multi-institutional review," *British Journal of Cancer*, vol. 112, no. 2, pp. 278–282, 2015.
- [6] K. Giuliano, T. Sachs, E. Montgomery et al., "Survival following lung metastasectomy in soft tissue sarcomas," *Thoracic and Cardiovascular Surgeon*, vol. 64, no. 2, pp. 150–158, 2016.
- [7] T. Treasure, M. Milošević, F. Fiorentino, and F. Macbeth, "Pulmonary metastasectomy: what is the practice and where is the evidence for effectiveness?," *Thorax*, vol. 69, no. 10, pp. 946–949, 2014.
- [8] D. L. Longo, "Tumor heterogeneity and personalized medicine," *The New England Journal of Medicine*, vol. 366, no. 10, pp. 956–957, 2012.
- [9] P. Lambin, E. Rios-Velazquez, R. Leijenaar et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.
- [10] Y. Q. Huang, C. H. Liang, L. He et al., "Development and validation of a Radiomics Nomogram for preoperative prediction of lymph node metastasis in colorectal Cancer," *Journal of Clinical Oncology*, vol. 34, no. 18, pp. 2157–2164, 2016.
- [11] K. Nie, L. Shi, Q. Chen et al., "Rectal Cancer: assessment of Neoadjuvant Chemoradiation outcome based on Radiomics of multiparametric MRI," *Clinical Cancer Research*, vol. 22, no. 21, pp. 5256–5264, 2016.
- [12] V. D. A. Corino, E. Montin, A. Messina et al., "Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 3, pp. 829–840, 2018.
- [13] M. Vallières, C. R. Freeman, S. R. Skamene, and I. el Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine & Biology*, vol. 60, no. 14, pp. 5471–5496, 2015.
- [14] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar et al., "Decoding tumour phenotype by noninvasive imaging using

- a quantitative radiomics approach,” *Nature Communications*, vol. 5, no. 1, article 4006, 2014.
- [15] T. Hu, S. P. Wang, L. Huang et al., “A clinical-radiomics nomogram for the preoperative prediction of lung metastasis in colorectal cancer patients with indeterminate pulmonary nodules,” *European Radiology*, vol. 29, no. 1, pp. 439–449, 2019.
- [16] S. N. Reske and J. Kotzerke, “FDG-PET for clinical use. Results of the 3rd German Interdisciplinary Consensus Conference, “Onko-PET III”, 21 July and 19 September 2000,” *European Journal of Nuclear Medicine*, vol. 28, no. 11, pp. 1707–1723, 2001.
- [17] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] L. U. Yan, “Optimization and application research of BP neural network,” *Journal of Beijing University of Chemical Technology*, vol. 28, no. 1, pp. 67–69, 2001.
- [19] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.
- [20] M. M. Galloway, “Texture analysis using gray level run lengths,” *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975.
- [21] A. Chu, C. M. Sehgal, and J. F. Greenleaf, “Use of gray value distribution of run lengths for texture analysis,” *Pattern Recognition Letters*, vol. 11, no. 6, pp. 415–419, 1990.
- [22] B. V. Dasarathy and E. B. Holder, “Image characterizations based on joint gray level—run length distributions,” *Pattern Recognition Letters*, vol. 12, no. 8, pp. 497–502, 1991.
- [23] G. Thibault, B. Fertil, C. Navarro et al., “Shape and texture indexes application to cell nuclei classification,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, no. 1, article 1357002, 2013.
- [24] M. Amadasun and R. King, “Textural features corresponding to textural properties,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 5, pp. 1264–1274, 1989.
- [25] P. Cerda, G. Varoquaux, and B. Kégl, “Similarity encoding for learning with dirty categorical variables,” *Machine Learning*, vol. 107, no. 8-10, pp. 1477–1494, 2018.
- [26] G. Song, J. Zhang, and Z. Sun, “the research of dynamic change learning rate strategy in BP neural network and application in network intrusion detection,” in *2008 3rd International Conference on Innovative Computing Information and Control*, Dalian, Liaoning, China, June 2008.

Research Article

In Silico Analysis Identifies Differently Expressed lncRNAs as Novel Biomarkers for the Prognosis of Thyroid Cancer

Yuansheng Rao, Haiying Liu, Xiaojuan Yan, and Jianhong Wang 

Department of Otorhinolaryngology, Beijing Anzhen Hospital, No. 2 Anzhen Road, Chaoyang District, Beijing 100029, China

Correspondence should be addressed to Jianhong Wang; ou70394137@163.com

Received 3 February 2020; Accepted 24 March 2020; Published 23 April 2020

Guest Editor: Tao Huang

Copyright © 2020 Yuansheng Rao 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.

Background. Thyroid cancer (TC) is one of the most common type of endocrine tumors. Long noncoding RNAs had been demonstrated to play key roles in TC. **Material and Methods.** The lncRNA expression data were downloaded from Co-lncRNA database. The raw data was normalized using the limma package in R software version 3.3.0. The differentially expressed mRNA and lncRNAs were identified by the linear models for the microarray analysis (Limma) method. The DEGs were obtained with thresholds of $|\log_{2}FC| > 1.5$ and $P < 0.001$. The hierarchical cluster analysis of differentially expressed mRNAs and lncRNAs was performed using CLUSTER 3.0, and the hierarchical clustering heat map was visualized by Tree View. **Results.** In the present study, we identified 6 upregulated and 85 downregulated lncRNAs in TC samples. Moreover, we for the first time identified 16 downregulated lncRNAs was correlated to longer disease-free survival time in patients with TC, including ATP1A1-AS1, CATIP-AS1, FAM13A-AS1, LINC00641, LINC00924, MIR22HG, NDUFA6-AS1, RP11-175K6.1, RP11-727A23.5, RP11-774O3.3, RP13-895J2.2, SDCBP2-AS1, SLC26A4-AS1, SNHG15, SRP14-AS1, and ZNF674-AS1. **Conclusions.** Bioinformatics analysis revealed these lncRNAs were involved in regulating the RNA metabolic process, cell migration, organelle assembly, tRNA modification, and hormone levels. This study will provide useful information to explore the potential candidate biomarkers for diagnosis, prognosis, and drug targets for TC.

1. Introduction

Thyroid cancer (TC) is one of the most common type of endocrine tumors [1]. A recent study showed the incidence of TC increased rapidly worldwide, especially in female. However, there was still lacking of effective biomarkers for the prognosis of TC. Over the past decades, several genes were identified to be related to the progression of TC and could serve as potential biomarkers for TC, such as RAS [2] and BRAF (V600E) [3] gene mutations. Moreover, with the development of the next-generation sequencing method, a series of public datasets were developed to explore the potential biomarkers and mechanisms underlying tumor progression in human cancers. For example, Wang et al. analyzed TCGA dataset and found lncRNA UNC5B-AS1 promoted TC growth and metastasis [4]. Identification of novel biomarkers is still an urgent need for the TC.

Long noncoding RNAs (lncRNAs) were reported to play important roles in tumorigenesis and cancer progression [5].

lncRNAs bound to chromatin, proteins, and RNAs to modulate cancer proliferation, apoptosis, autophagy, epithelial-mesenchymal transition (EMT), and metastasis [6]. In TC, ENST00000539653 promoted cancer progression via MAPK signaling. TUG1 regulated TC cell proliferation and EMT through targeting miR-145 [7]. A recent study showed antisense lncRNA COMET repression inhibited cell viability and invasiveness and induced sensitivity to vemurafenib in BRAF- and RET-driven TC [8]. Interestingly, emerging studies demonstrated lncRNAs could serve as potential prognostic or diagnostic biomarkers for human cancers. For instance, Zhang et al. reported that downregulation of DANCR is a biomarker for TC diagnosis [9]. Decreased EMX2OS expression was associated with unfavorable recurrence-free survival (RFS) in classical PTC [10].

In this study, we identified differently expressed lncRNAs using two public datasets, including Co-lncRNA database and GEPIA database [11]. Then, coexpression network analysis, gene ontology (GO) analysis, and Kyoto Encyclopedia of

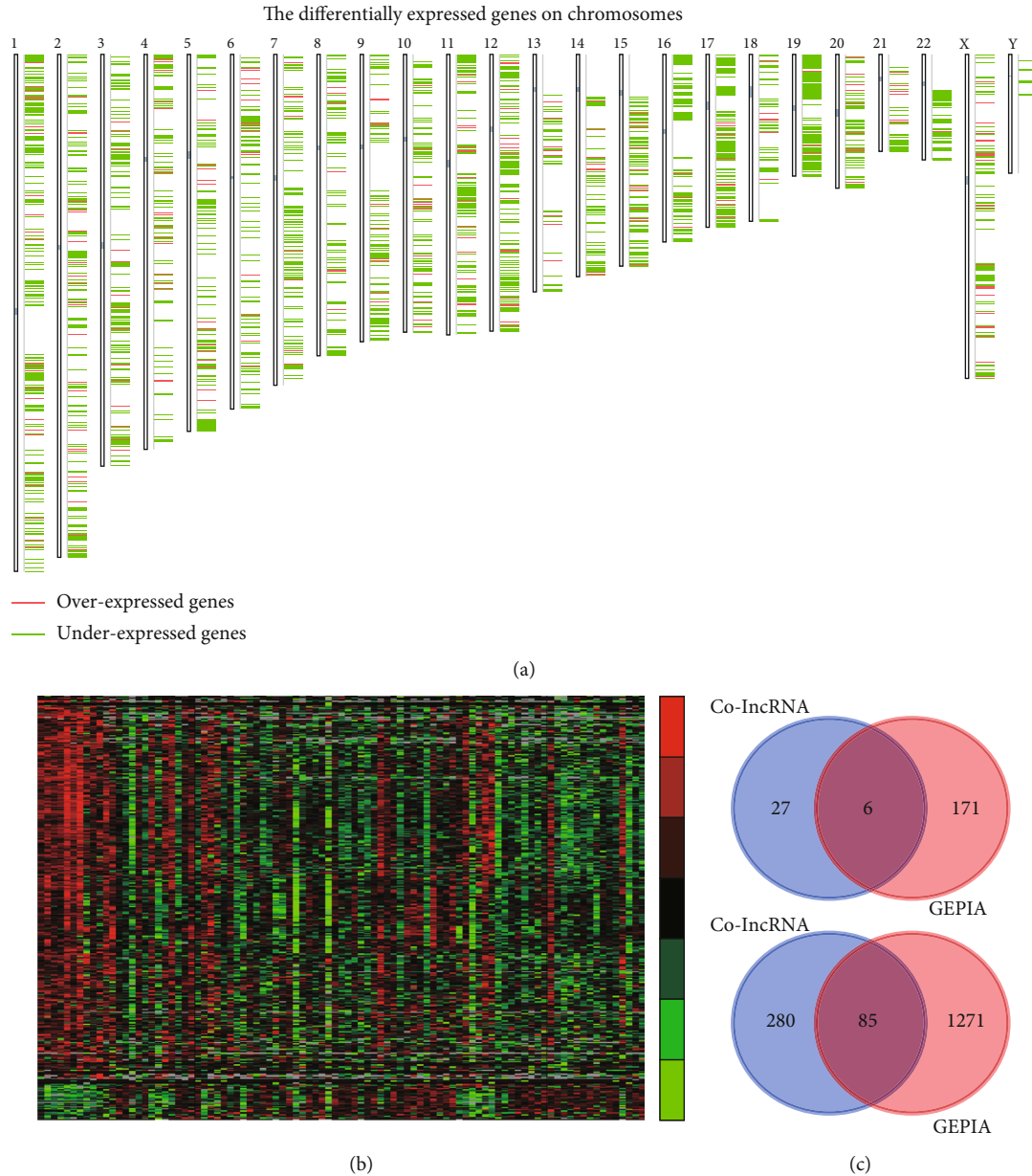


FIGURE 1: Identification of differentially expressed lncRNAs in TC. (a) Chromosomal distribution of differentially expressed genes in TC tissues using GEPIA database. (b) Hierarchical clustering analysis shows differential lncRNA expression between normal and TC samples by using Co-lncRNA database. (c, d) Venn diagrams display differentially expressed lncRNAs in both databases.

Genes and Genomes (KEGG) pathway analysis were used to evaluate the potential functions of these lncRNAs in TC. We thought this study could provide novel biomarkers for TC.

2. Materials and Methods

2.1. Public Dataset Analysis. The lncRNA expression data were downloaded from Co-lncRNA database. Co-lncRNA database included 12 normal samples and 83 TC samples. The raw data was normalized using the limma package in R software version 3.3.0 (<https://www.r-project.org/>). The differentially expressed mRNA and lncRNAs were identified by the linear models for microarray analysis (Limma)

method [12]. The DEGs were obtained with thresholds of $|\log_{2}FC| > 1.5$ and $P < 0.001$. The hierarchical cluster analysis of differentially expressed mRNAs and lncRNAs was performed using CLUSTER 3.0 [13], and the hierarchical clustering heat map was visualized by Tree View [14].

2.2. Coexpression Network Construction and Analysis. In this study, as Hu et al. [15] described, the Pearson correlation coefficient of DEG-lncRNA pairs was calculated according to the expression value of them. The coexpressed DEG-lncRNA pairs with the absolute value of Pearson correlation coefficient ≥ 0.75 were selected, and the coexpression network was established by using cytoscape software. Cytoscape

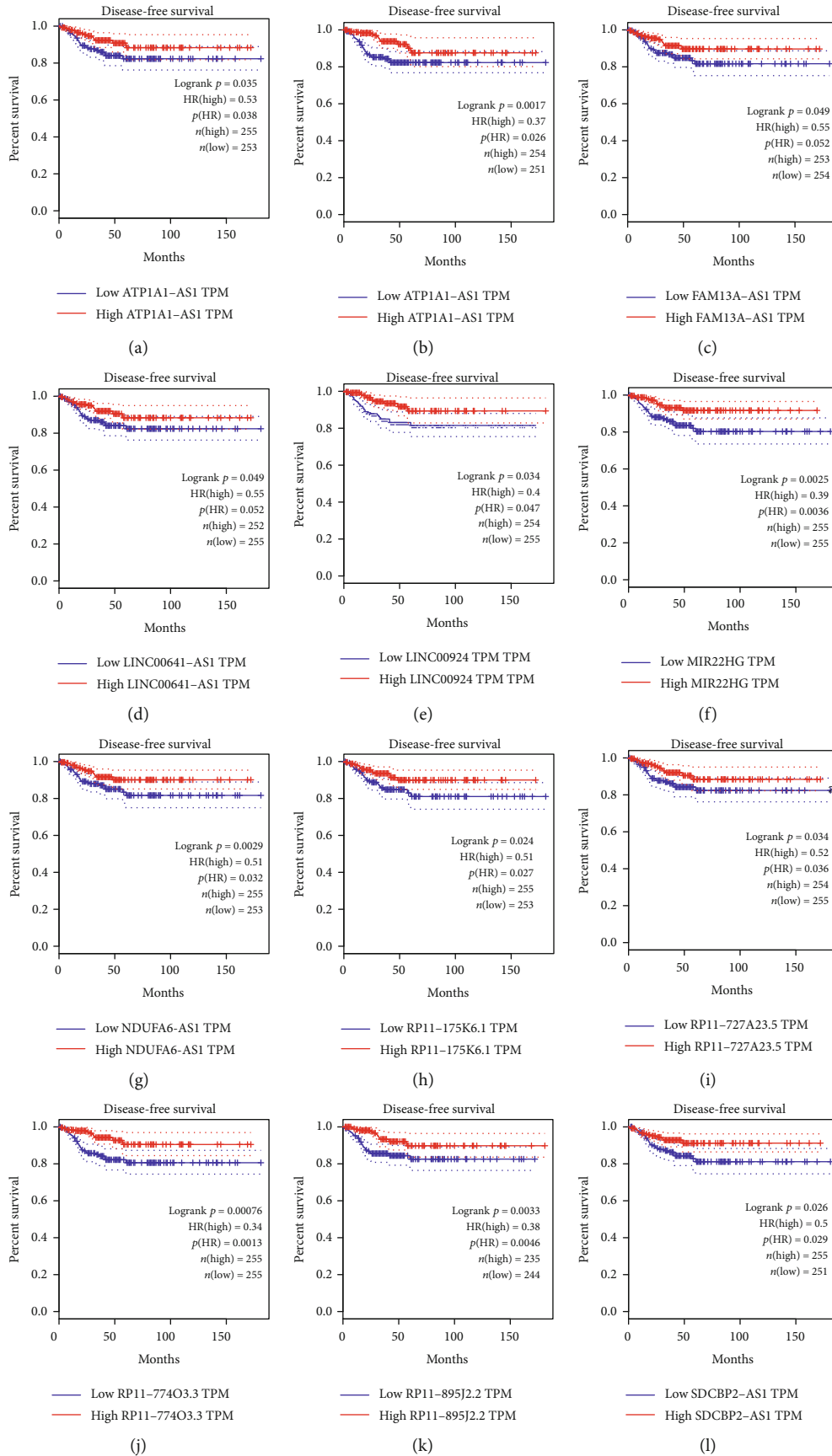


FIGURE 2: Continued.

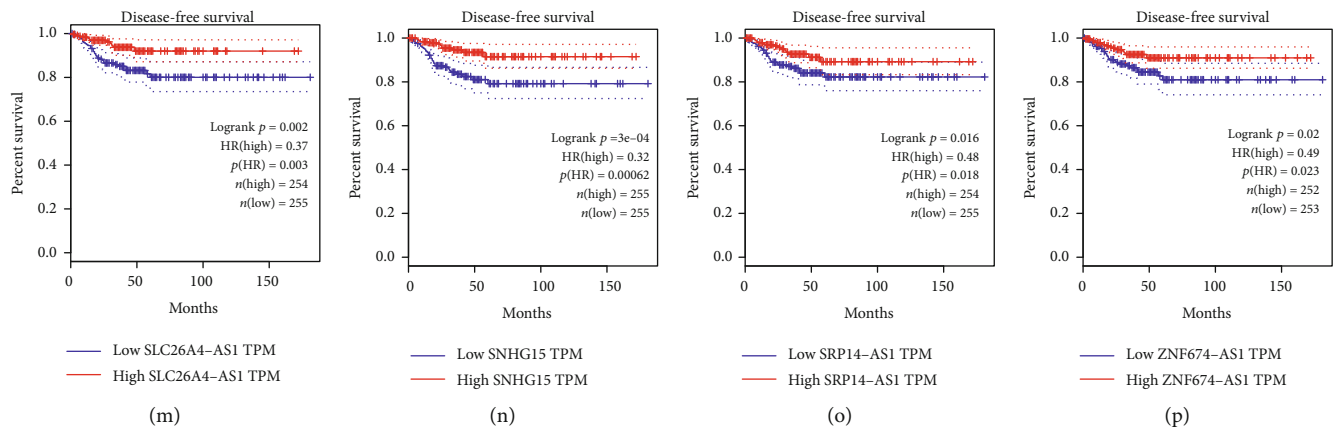


FIGURE 2: Identification of disease-free survival time related lncRNAs in TC. (a-p) Higher expression levels of ATP1A1-AS1 (a), CATIP-AS1 (b), FAM13A-AS1 (c), LINC00641 (d), LINC00924 (e), MIR22HG (f), NDUFA6-AS1 (g), RP11-175K6.1 (h), RP11-727A23.5 (i), RP11-774O3.3 (j), RP13-895J2.2 (k), SDCBP2-AS1 (l), SLC26A4-AS1 (m), SNHG15 (n), SRP14-AS1 (o), and ZNF674-AS1 (p) were significantly correlated to longer disease-free survival time in patients with TC.

MCODE plug-in (version 3.4.0, available online: <http://www.cytoscape.org/>) was applied for visualization of the coexpression networks.

Gene coexpression analysis could be applied to related genes of unknown function with GO or to analysis candidate disease genes or to predict transcriptional regulatory mechanism [16].

2.3. GO and KEGG Pathway Analyses. To identify functions of DEGs in smoking-related lung cancer, we performed GO function enrichment analysis in 3 functional ontologies: biological process (BP), cellular component (CC), and molecular function (MF). KEGG pathway enrichment analysis was also performed to identify pathways enriched in smoking-related lung cancer using the DAVID system (<https://david.ncifcrf.gov/>). The P value less than 0.05 was considered significant.

2.4. Survival Analysis. GEPIA database (<http://gepia.cancer-pku.cn/index.html>) was used to predict the correlation between candidate gene expression and overall survival (OS) time or disease-free survival (DFS) time. The median expression of target was selected as cutoff to divide all TC samples as high and low groups. The probability of survival was estimated using the Kaplan-Meier method. The log-rank test was used to compare differences in survival times.

2.5. Statistical Analysis. The numerical data were presented as mean \pm standard deviation (SD) of at least three determinations. Statistical comparisons between groups of normalized data were performed using T -test or Mann-Whitney U test according to the test condition. A $P < 0.05$ was considered statistical significance with a 95% confidence level.

3. Results

3.1. Identification of Differently Expressed lncRNAs in TC. GEPIA database was first analyzed. Our results identified 177 upregulated lncRNAs and 1359 downregulated lncRNAs in TC samples compared to normal tissues (Figure 1(a) and

Supplementary Table 1). By analyzing Co-lncRNA database, 399 lncRNAs were found to be dysregulated in TC. Among these lncRNAs, 33 lncRNAs were overexpressed and 366 lncRNAs were suppressed in TC tissues compared to normal tissues (Figure 1(b)).

By performing integrated analysis of Co-lncRNA and GEPIA databases, a total of 6 lncRNAs were found to be upregulated and 85 lncRNAs were found to be downregulated in TC samples (Figure 1(c)). CATIP-AS1 is the most significantly downregulated lncRNA, and RP11-280O1.2 is the most significantly upregulated lncRNA in TC.

3.2. Downregulated lncRNAs Were Correlated to Longer Disease-Free Survival Time in TC. Then, the GEPIA dataset was used to explore the correlation between lncRNA expression and overall survival time in TC. The median expression level of target gene was selected as the cutoff to divide all TC samples into high and low groups. Our analyses showed dysregulated lncRNAs were significantly correlated to the disease-free survival time in TC. Higher expression levels of ATP1A1-AS1, CATIP-AS1, FAM13A-AS1, LINC00641, LINC00924, MIR22HG, NDUFA6-AS1, RP11-175K6.1, RP11-727A23.5, RP11-774O3.3, RP13-895J2.2, SDCBP2-AS1, SLC26A4-AS1, SNHG15, SRP14-AS1, and ZNF674-AS1 were significantly correlated to longer disease-free survival time in patients with TC (Figures 2(a)-2(p)).

Of note, we found that ATP1A1-AS1, CATIP-AS1, FAM13A-AS1, LINC00641, LINC00924, MIR22HG, NDUFA6-AS1, RP11-175K6.1, RP11-727A23.5, RP11-774O3.3, RP13-895J2.2, SDCBP2-AS1, SLC26A4-AS1, SNHG15, SRP14-AS1, and ZNF674-AS1 were significantly downregulated in TC samples compared to normal tissues (Figures 3(a)-3(p)). These results suggested these lncRNAs may serve as tumor suppressors in TC.

3.3. Construction of Differently Expressed lncRNAs Regulating Coexpression Network in TC. Furthermore, we constructed differently expressed lncRNAs regulating coexpression network in TC. The Pearson correlation coefficients between

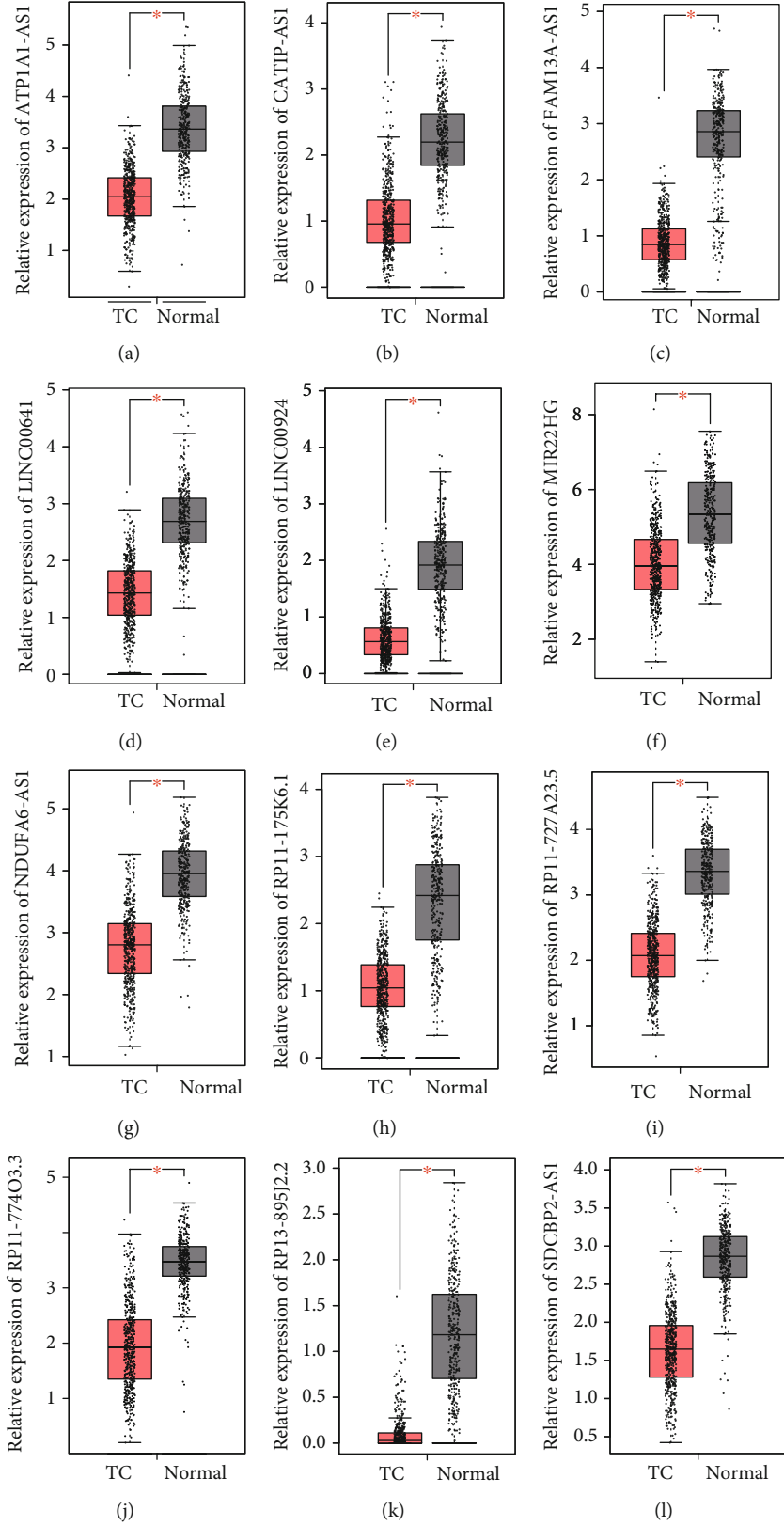


FIGURE 3: Continued.

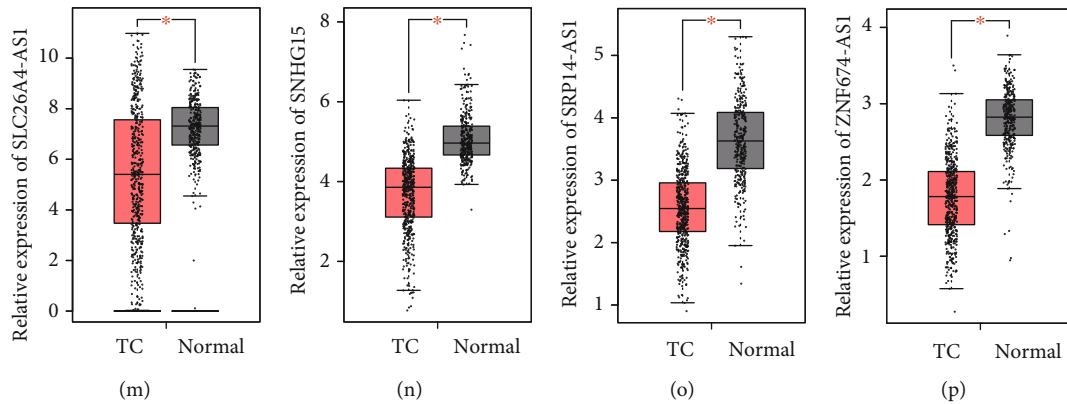


FIGURE 3: We identified downregulated lncRNAs in TC. (a-p) The expression levels of ATP1A1-AS1 (a), CATIP-AS1 (b), FAM13A-AS1 (c), LINC00641 (d), LINC00924 (e), MIR22HG (f), NDUFA6-AS1 (g), RP11-175K6.1 (h), RP11-727A23.5 (i), RP11-774O3.3 (j), RP13-895J2.2 (k), SDCBP2-AS1 (l), SLC26A4-AS1 (m), SNHG15 (n), SRP14-AS1 (o), and ZNF674-AS1 (p) were downregulated in TC samples compared to normal tissues.

lncRNA and mRNA was downloaded from GEPIA datasets. We selected the top 200 correlated genes as the potential targets of differently expressed lncRNAs. As shown in Figure 4, we found that this coexpression network contained 16 lncRNAs and 2698 mRNAs.

3.4. Bioinformatics Analysis of Differently Expressed lncRNAs in TC. Furthermore, we performed bioinformatics analysis for differentially expressed lncRNAs in TC (Figure 5). GO analysis showed that ATP1A1-AS1 [17] was involved in regulating the RNA metabolic process, the nucleic acid metabolic process, regulation of gene expression, transcription, DNA-templated, and cellular macromolecule metabolic process. FAM13A-AS1 [18] was involved in regulating RNA splicing and mRNA processing. LINC00641 was involved in regulating the RNA metabolic process; gene expression; mRNA processing; regulation of gene expression; RNA splicing; and mRNA splicing, via spliceosome, and transcription [19]. LINC00924 was associated with the regulation of cell migration, regulation of cellular component movement, regulation of locomotion, cell adhesion, circulatory system development, and locomotion. MIR22HG was involved in regulating organelle assembly, positive regulation of the RNA metabolic process, cilium organization, axoneme assembly, and regulation of gene expression [20]. RP11-175K6.1 was involved in regulating vasculature development, blood vessel development, circulatory system development, blood vessel morphogenesis, angiogenesis, and tube morphogenesis. RP11-727A23.5 was involved in regulating mRNA processing, RNA splicing, inner dynein arm assembly, cilium assembly, and gene expression. SDCBP2-AS1 was involved in regulating the tRNA process, tRNA methylation, methylation, macromolecule methylation, and tRNA modification [21]. SLC26A4-AS1 was involved in regulating regulation of hormone levels, the oxidation-reduction process, thyroid hormone generation, the hormone metabolic process, and the alpha-amino acid metabolic process [22]. SRP14-AS1 was involved in regulating cilium movement, determination of left/right symmetry, photoreceptor cell outer segment

organization, inner dynein arm assembly, and organelle assembly (Figure 5(a)-5(j)).

4. Discussion

Thyroid cancer is a rare but a highly lethal form of thyroid cancer, which needs more attention. And lncRNAs had been demonstrated to play key roles in the progression of most human cancers, including thyroid cancer. For instance, DGCR5 played as a tumor suppressor in TC though binding to miR-2861 [23]. SNHG16 promoted TC proliferation and invasion through modulation of miR-497 [24]. GAS8-AS1 inhibited TC growth through miR-135b-5p/CCND2 axis. Of note, lncRNAs were also found to be dysregulated in TC, suggesting the potential prognostic value of lncRNAs. For example, a bioinformatics analysis study showed that FAM95B1 and UCA1 were correlated with cervical lymph node metastasis, tumor staging, and TC prognosis. Lu et al. reported that the dysregulation of RUNDC3A-AS1, FOXD-AS1, RUNDC3A-AS1 and FOXD-AS1 was correlated to a shorter overall survival time in patients with TC [25]. However, only a small part of lncRNAs were reported in TC. The expression pattern and molecular functions of most lncRNAs in TC remained unknown.

In our study, silico analyses were performed to identify TC-related important lncRNA. Co-lncRNA and GEPIA databases were used to identify differently expressed lncRNAs in TC. There are a total of 6 upregulated and 85 downregulated lncRNAs in TC samples compared to normal tissues. Among these lncRNAs, only few lncRNAs were reported in previous studies. For example, NR2F1-AS1 was found to be upregulated in TC samples. In hepatocellular carcinoma, knockdown of NR2F1-AS1 significantly suppressed cancer invasion, migration, and in vivo tumor growth [26]. Moreover, we for the first time identified 16 downregulated lncRNAs was correlated to a longer disease-free survival time in patients with TC, including ATP1A1-AS1, CATIP-AS1, FAM13A-AS1, LINC00641, LINC00924, MIR22HG, NDUFA6-AS1, RP11-175K6.1,

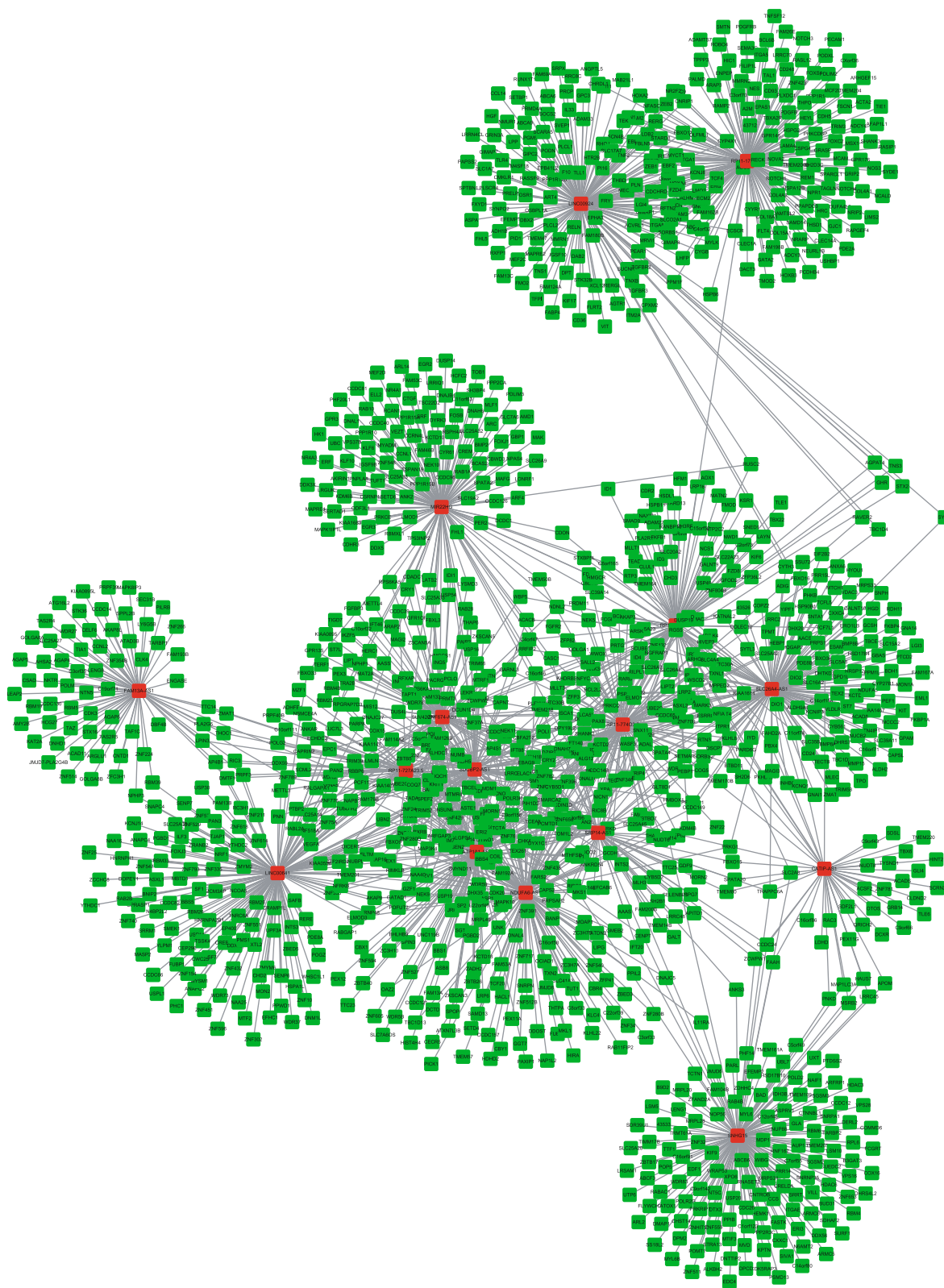


FIGURE 4: Coexpression network analysis of lncRNAs in TC. Coexpression network analysis of lncRNAs in TC. Red nodes: lncRNA; green nodes: mRNA.

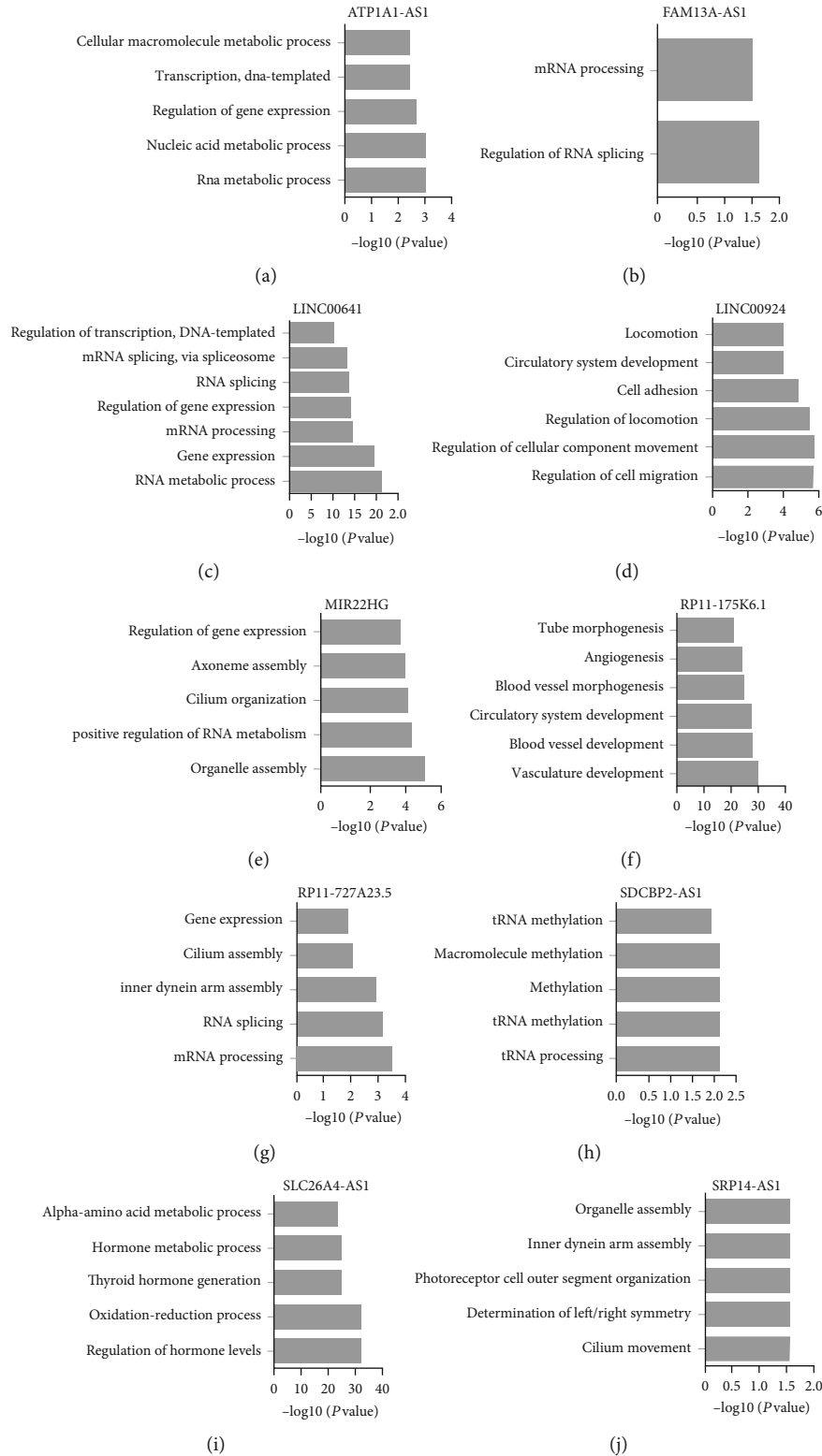


FIGURE 5: Bioinformatics analysis of differently expressed lncRNAs. (a-j) Bioinformatics analysis of ATP1A1-AS1 (a), FAM13A-AS1 (b), LINC00641 (c), LINC00924 (d), MIR22HG (e), RP11-175K6.1 (f), RP11-727A23.5 (g), SDCBP2-AS1 (h), SLC26A4-AS1 (i), and SRP14-AS1 (j) in TC.

RP11-727A23.5, RP11-774O3.3, RP13-895J2.2, SDCBP2-AS1, SLC26A4-AS1, SNHG15, SRP14-AS1, and ZNF674-AS1. The functions of these lncRNAs remained unclear.

ATP1A1-AS1 is a novel lncRNA. A previous study showed ATP1A1-AS1 is a negative regulator of Na/K-ATPase $\alpha 1$ and involved in regulating cell proliferation in human

kidney cells. LINC00641 was reported as a tumor suppressor in bladder cancer via sponging miR-197 [19]. SLC26A4-AS1 was found to be associated with overall survival in gastric cancer. SNHG15 was reported to be downregulated in thyroid cancer and acted as a tumor suppressor in TC [27].

LncRNA coexpression network was widely used to explore the potential roles of novel lncRNAs in TC. For example, Zhang et al. revealed that HCG11 was involved in regulating the MAPK signaling pathway and gene transcription through coexpression analysis [28]. In this study, we constructed a network including 19 downregulated lncRNAs and 2698 mRNAs. Bioinformatics analysis showed these lncRNAs played crucial roles in TC progression. For example, ATP1A1-AS1, RP11-727A23.5, and LINC00641 were involved in regulating the RNA metabolic process. FAM13A-AS1 was involved in regulating RNA splicing. LINC00924 was associated with the regulation of cell migration and cell adhesion. MIR22HG was involved in regulating organelle assembly. RP11-175K6.1 was involved in regulating vasculature development. SDCBP2-AS1 was involved in regulating tRNA modification. SLC26A4-AS1 was involved in regulating hormone levels.

In conclusion, we identified 6 upregulated and 85 downregulated lncRNAs in TC samples. Moreover, we for the first time identified 16 downregulated lncRNAs was correlated to a longer disease-free survival time in patients with TC, including ATP1A1-AS1, CATIP-AS1, FAM13A-AS1, LINC00641, LINC00924, MIR22HG, NDUFA6-AS1, RP11-175K6.1, RP11-727A23.5, RP11-774O3.3, RP13-895J2.2, SDCBP2-AS1, SLC26A4-AS1, SNHG15, SRP14-AS1, and ZNF674-AS1. Bioinformatics analysis revealed these lncRNAs were involved in regulating the RNA metabolic process, cell migration, organelle assembly, tRNA modification, and hormone levels. This study will provide useful information to explore the potential candidate biomarkers for diagnosis, prognosis, and drug targets for TC.

Abbreviations

TC:	Thyroid cancer
Lnc-RNAs:	Long noncoding RNAs
EMT:	Epithelial-mesenchymal transition
RFS:	Recurrence-free survival
GO:	Gene Ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
BP:	Biological process
CC:	Cellular component
MF:	Molecular function
OS:	Overall survival
DFS:	Disease-free survival.

Data Availability

All the data in this manuscript can be accessed in Co-LncRNA (<http://www.biobigdata.com/Co-LncRNA/>).

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

Jianhong Wang conceived and designed the study. Yuan-sheng Rao and Haiying Liu were responsible for the collection and analysis of the data. All authors interpreted the data, drafted the manuscript, and approved the final version of the manuscript.

Supplementary Materials

Supplementary table 1: the differently expressed lncRNAs in TC by using GEPIA database. (*Supplementary Materials*)

References

- [1] A. S. Al-Zahrani and K. Ravichandran, "Epidemiology of thyroid cancer: a review with special reference to Gulf Cooperation Council (GCC) states," *The Gulf Journal of Oncology*, vol. 2, pp. 17–28, 2007.
- [2] J. L. Bos, "Ras oncogenes in human cancer: a review," *Cancer Research*, vol. 49, no. 17, p. 4682, 1989.
- [3] D. Dankort, E. Filenova, M. Collado, M. Serrano, K. Jones, and M. McMahon, "A new mouse model to explore the initiation, progression, and therapy of BRAFV600E-induced lung tumors," *Genes & Development*, vol. 21, no. 4, pp. 379–384, 2007.
- [4] Y. Wang, A. Bhandari, J. Niu et al., "The lncRNA UNC5B-AS1 promotes proliferation, migration, and invasion in papillary thyroid cancer cell lines," *Human Cell*, vol. 32, no. 3, pp. 334–342, 2019.
- [5] M. Cong, J. Li, R. Jing, and Z. Li, "Long non-coding RNA tumor suppressor candidate 7 functions as a tumor suppressor and inhibits proliferation in osteosarcoma," *Tumour Biology*, vol. 37, no. 7, pp. 9441–9450, 2016.
- [6] L. Qiu, Q. Tang, G. Li, and K. Chen, "Long non-coding RNAs as biomarkers and therapeutic targets: Recent insights into hepatocellular carcinoma," *Life Sciences*, vol. 191, pp. 273–282, 2017.
- [7] H. Lei, Y. Gao, and X. Xu, "LncRNA TUG1 influences papillary thyroid cancer cell proliferation, migration and EMT formation through targeting miR-145," *Acta Biochimica et Biophysica Sinica*, vol. 49, no. 7, pp. 588–597, 2017.
- [8] A. Tahiri, K. Røe, A. H. Ree et al., "Differential inhibition of ex vivo tumor kinase activity by vemurafenib in BRAF(V600E) and BRAF wild-type metastatic malignant melanoma," *PLoS One*, vol. 8, no. 8, article e72692, 2013.
- [9] S. X. Yuan, J. Wang, F. Yang et al., "Long noncoding RNA DANCR increases stemness features of hepatocellular

- carcinoma by derepression of CTNNB1," *Hepatology*, vol. 63, no. 2, pp. 499–511, 2016.
- [10] C. A. Yang, S. Bauer, Y. C. Ho, F. Sotzny, J. G. Chang, and C. Scheibenbogen, "The expression signature of very long non-coding RNA in myalgic encephalomyelitis/chronic fatigue syndrome," *Journal of Translational Medicine*, vol. 16, no. 1, p. 231, 2018.
- [11] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic Acids Research*, vol. 45, no. W1, pp. W98–W102, 2017.
- [12] I. Diboun, L. Wernisch, C. A. Orengo, and M. Koltzenburg, "Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma," *BMC Genomics*, vol. 7, no. 1, p. 252, 2006.
- [13] Y. Zhai, Y. Li, J. Zhang et al., "Identification of the gene cluster for bistropolone-humulene meroterpenoid biosynthesis in *Phoma* sp.," *Fungal Genetics and Biology*, vol. 129, pp. 7–15, 2019.
- [14] R. D. Page, *TreeView*, Glasgow University, Glasgow, UK, 2001.
- [15] X. Hu, G. Shen, X. Lu, G. Ding, and L. Shen, "Identification of key proteins and lncRNAs in hypertrophic cardiomyopathy by integrated network analysis," *Archives of Medical Science*, vol. 15, no. 2, pp. 484–497, 2019.
- [16] X. Shi, T. Huang, J. Wang et al., "Next-generation sequencing identifies novel genes with rare variants in total anomalous pulmonary venous connection," *eBioMedicine*, vol. 38, pp. 217–227, 2018.
- [17] D. Zhang, P. Zhang, P. Yang et al., "Downregulation of ATP1A1 promotes cancer development in renal cell carcinoma," *Clinical Proteomics*, vol. 14, no. 1, p. 15, 2017.
- [18] Z. Jin, J. W. Chung, W. Mei et al., "Regulation of nuclear-cytoplasmic shuttling and function of Family with sequence similarity 13, member A (Fam13a), by B56-containing PP2As and Akt," *Molecular Biology of the Cell*, vol. 26, no. 6, pp. 1160–1173, 2015.
- [19] X. B. Wang, H. Wang, H. Q. Long, D. Y. Li, and X. Zheng, "LINC00641 regulates autophagy and intervertebral disc degeneration by acting as a competitive endogenous RNA of miR-153-3p under nutrition deprivation stress," *Journal of Cellular Physiology*, vol. 234, no. 5, pp. 7115–7127, 2019.
- [20] Z. Cui, X. An, J. Li, Q. Liu, and W. Liu, "LncRNA MIR22HG negatively regulates miR-141-3p to enhance DAPK1 expression and inhibits endometrial carcinoma cells proliferation," *Biomedicine & Pharmacotherapy*, vol. 104, pp. 223–228, 2018.
- [21] S. Talukdar, S. K. Das, A. K. Pradhan et al., "Novel function of MDA-9/Syntenin (SDCBP) as a regulator of survival and stemness in glioma stem cells," *Oncotarget*, vol. 7, no. 34, pp. 54102–54119, 2016.
- [22] C. Song, J. Zhang, Y. Liu et al., "Construction and analysis of cardiac hypertrophy-associated lncRNA-mRNA network based on competitive endogenous RNA reveal functional lncRNAs in cardiac hypertrophy," *Oncotarget*, vol. 7, no. 10, pp. 10827–10840, 2016.
- [23] R. Hu, W. Liu, H. Li et al., "A Runx2/miR-3960/miR-2861 regulatory feedback loop during mouse osteoblast differentiation," *Journal of Biological Chemistry*, vol. 286, no. 14, pp. 12328–12339, 2011.
- [24] W. Zhao, H. Fu, S. Zhang, S. Sun, and Y. Liu, "LncRNA SNHG16 drives proliferation, migration, and invasion of hemangioma endothelial cell through modulation of miR-520d-3p/STAT3 axis," *Cancer Medicine*, vol. 7, no. 7, pp. 3311–3320, 2018.
- [25] W. Lu, Y. Xu, J. Xu, Z. Wang, and G. Ye, "Identification of differential expressed lncRNAs in human thyroid cancer by a genome-wide analyses," *Cancer Medicine*, vol. 7, no. 8, pp. 3935–3944, 2018.
- [26] H. Huang, J. Chen, C. M. Ding, X. Jin, Z. M. Jia, and J. Peng, "LncRNA NR2F1-AS1 regulates hepatocellular carcinoma oxaliplatin resistance by targeting ABCC1 via miR-363," *Journal of Cellular and Molecular Medicine*, vol. 22, no. 6, pp. 3238–3245, 2018.
- [27] K. W. Chung, S. W. Kim, and S. W. Kim, "Gene expression profiling of papillary thyroid carcinomas in Korean patients by oligonucleotide microarrays," *Journal of the Korean Surgical Society*, vol. 82, no. 5, pp. 271–280, 2012.
- [28] Y. Xu, Y. Zheng, H. Liu, and T. Li, "Modulation of IGF2BP1 by long non-coding RNA HCG11 suppresses apoptosis of hepatocellular carcinoma cells via MAPK signaling transduction," *International Journal of Oncology*, vol. 51, no. 3, pp. 791–800, 2017.

Research Article

ACNNT3: Attention-CNN Framework for Prediction of Sequence-Based Bacterial Type III Secreted Effectors

Jie Li ¹, Zhong Li ^{1,2}, Jiesi Luo,³ and Yuhua Yao ⁴

¹*School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China*

²*School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China*

³*Key Laboratory for Aging and Regenerative Medicine, Department of Pharmacology, School of Pharmacy, Southwest Medical University, Luzhou 646000, China*

⁴*School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China*

Correspondence should be addressed to Zhong Li; lizhong@zstu.edu.cn and Yuhua Yao; yaoyuhua2288@163.com

Received 6 February 2020; Revised 9 March 2020; Accepted 17 March 2020; Published 2 April 2020

Guest Editor: Lei Chen

Copyright © 2020 Jie Li 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 type III secretion system (T3SS) is a special protein delivery system in Gram-negative bacteria which delivers T3SS-secreted effectors (T3SEs) to host cells causing pathological changes. Numerous experiments have verified that T3SEs play important roles in many biological activities and in host-pathogen interactions. Accurate identification of T3SEs is therefore essential to help understand the pathogenic mechanism of bacteria; however, many existing biological experimental methods are time-consuming and expensive. New deep-learning methods have recently been successfully applied to T3SE recognition, but improving the recognition accuracy of T3SEs is still a challenge. In this study, we developed a new deep-learning framework, ACNNT3, based on the attention mechanism. We converted 100 residues of the N-terminal of the protein sequence into a fusion feature vector of protein primary structure information (one-hot encoding) and position-specific scoring matrix (PSSM) which are used as the feature input of the network model. We then embedded the attention layer into CNN to learn the characteristic preferences of type III effector proteins, which can accurately classify any protein directly as either T3SEs or non-T3SEs. We found that the introduction of new protein features can improve the recognition accuracy of the model. Our method combines the advantages of CNN and the attention mechanism and is superior in many indicators when compared to other popular methods. Using the common independent dataset, our method is more accurate than the previous method, showing an improvement of 4.1-20.0%.

1. Introduction

Gram-negative bacteria can secrete proteins into host cells through a variety of secretion systems which affect the cell and its external environment. This process can be mediated by a variety of secretory systems, which can be divided into eight categories: type I to VIII secretory systems (T1SS-T8SS) [1]. Type I and III secretory systems are independent of signal sequences (sec), while types II and IV depend on signal sequences. The proteins secreted by the sec-dependent secretion system have a signal peptide sequence mainly composed of N-terminal hydrophobic amino acids which can guide the protein through the cell membrane.

When the protein reaches the periplasm, the signal peptide is cut off. Type II and IV secretion systems remove the N-terminal part of the secreted protein in the periplasm. The difference between systems is that proteins pass through the outer membrane in different ways. When protein secreted by the type II secretion system passes through the outer membrane, an additional set of inner membrane and outer membrane proteins is needed to assist, while the type IV secretion system includes a series of autotransporters which form a hole in the outer membrane to make the protein pass through, autolytically cut, and then release the protein. Neither the type I nor III secretion system processes the terminal amino acid of the secreted protein, nor does it appear that

the secreted protein stays in the periplasm. The secretion signal of the protein secreted by the type I secretion system is located at about 60 amino acids from the C-terminal end of the protein. This secretory signal appears to be subfamily specific, and the secreted proteins are not easily cut by proteolytic enzymes. The type V secretion system is related to the secretion of macromolecular proteins and may also be sec-dependent.

The type III secretion system is a transmembrane channel formed by the multicomponent protein complex that has been widely encoded in many Gram-negative bacteria including *Escherichia*, *Shigella*, *Yersinia*, *Salmonella*, and *Pseudomonas* [2, 3]. It can change the signal transduction [4] and innate immune response [5] of host cells by secreting proteins or injecting these virulent proteins directly into host cells. Type III secretion systems (T3SSs) have been widely studied because they are critical for virulence in various human pathogens. There are many *in vivo* and *in vitro* methods for predicting T3SEs, and while some of them obtain good predictions, the experiments are complicated and time-consuming.

Some machine-learning methods have successfully been used to predict T3SEs, such as the Naïve Bayes (NB) [6], artificial neural network (ANN) [7], support vector machine (SVM) [8, 9], and random forest (RF) [10]. The disadvantage of these machine-learning methods is that features must be defined in advance, the appropriate selection of features will affect the prediction accuracy, and the flexibility of model change or update is limited [11]. Many deep-learning methods have been recently proposed, such as LSTM [12], ResNet [13], DenseNet [14], and VGG16 [15], and these methods can also be used in bioinformatics and other related fields [16, 17]. The deep-learning method DeepT3 [11] trained deep CNN using only one-hot encoding as the model feature input and achieved good prediction results in terms of accuracy. Since only one feature is input and CNN cannot connect the sequence context well when extracting sequence features, this method can be improved upon for predicting T3SEs.

The attention mechanism has recently gained popularity in neural networks because it can weigh the input features to measure the importance of each feature to the object recognition. It has widely been applied for text and image classification [18, 19], machine translation [20], and bioinformatics [21]. In this study, we propose a method for predicting T3SEs using N-terminal sequences based on the Attention-CNN. Our model extracts features of one-hot and PSSM from 100 residues of the N-terminal sequence and fuses them as the feature input. The attention layer in the model can well connect the front and back of the sequence, and the CNN module can well extract the features of the sequence. We combine these two modules to make the entire framework learn the features of the sequence to their maximum extent. The results show that our method is effective in predicting T3SEs; not only can it accurately capture protein transport target information, but it also performs better than the existing methods.

2. Materials and Methods

2.1. Dataset. We collected a comprehensive dataset from multiple bacterial species known as T3SEs and non-T3SEs

from previous studies by Yang et al. [10], Wang et al. [22], and Tay et al. [23]. CD-HIT [24] with the sequence identity cutoff of 30% was used for sequence alignment to remove proteins with high similarity, and by skipping proteins with less than 100 amino acids, we obtained a balance dataset containing 283 T3 proteins and 311 non-T3 proteins.

We established our negative sample set by selecting type I to VIII secreted proteins of Gram-negative bacteria and removing type III secreted proteins and their homologues. In order to establish a 1:3 ratio of positive to negative [11], we randomly selected negative samples from the previous work of Dong et al. [8] and eliminated protein sequences with high similarity, resulting in a total of 835 negative samples.

There are two test sets used to evaluate our method. The independent dataset collected from Li et al. [11] contains 35 type III effectors and 86 non-type III effectors. The other test dataset is from the plant pathogen *P. syringae*. 85 type III effectors and 14 non-type III effectors that were not included in all models were collected from Baltrus et al. [25].

At present, most tools are based on the full sequence information of proteins or only 100 C-terminal residues [26]. In previous studies, N-terminal residues have been shown to also provide targeted information for protein transport [27–29], and the target information of T3SEs is usually located in the 50-100 N-terminal residues in different bacteria [30, 31]. Therefore, we have only used the N-terminal sequences in all the following calculations.

2.2. Feature Extraction. The feature input of our model is the combination of one-hot encoding and the PSSM of a protein. Each sequence is transformed into a one-hot matrix with 100 rows and 20 columns and a PSSM matrix with 100 rows and 20 columns, which are integrated into a combination matrix with 200 rows and 20 columns as the feature input. The 20 columns in the one-hot matrix correspond to 20 amino acids. One-hot encoding solves the problem of the classifier not effectively processing attribute data and expands the features to a certain extent. However, compared to PSSM, one-hot encoding is weaker with regard to protein feature extraction. Here, the introduction of PSSM enables the network model to better learn the characteristic preference of proteins, because PSSM features consider the position weight, number, and other parameters of each amino acid in the protein. PSSM also considers evolutionary information, so even the same residue may generate different characteristics, and it can effectively extract information from amino acid sequences. We used the PSI-BLAST [32] search database from UniprotKB/Swiss-Prot [33] to obtain the PSSM of the target protein. The matrix is an $L \times 20$ matrix, where L represents the total number of residues in the target protein's amino acid sequence. At the same time, we use 1, 2, 3, ..., 20 to represent the individual characters of the ordered 20 basic amino acids and get the corresponding number of columns. In summary, $U_{i \rightarrow j}^{\oplus}$ indicates the possibility that the i position of the amino acid sequence of the target protein is encoded as the basic amino acid j during the evolution process.

$$\text{PSSM} = \begin{bmatrix} U_{1 \rightarrow 1}^{\oplus} & U_{1 \rightarrow 2}^{\oplus} & \cdots & U_{1 \rightarrow 20}^{\oplus} \\ U_{2 \rightarrow 1}^{\oplus} & U_{2 \rightarrow 2}^{\oplus} & \cdots & U_{2 \rightarrow 20}^{\oplus} \\ \vdots & \vdots & \vdots & \vdots \\ U_{L \rightarrow 1}^{\oplus} & U_{L \rightarrow 2}^{\oplus} & \cdots & U_{L \rightarrow 20}^{\oplus} \end{bmatrix}. \quad (1)$$

2.3. Overview of Attention-CNN Model. The traditional CNN model includes convolution, pooling, and full connection layers, and it can be used to extract the sequence features of proteins. However, the sequence of a protein is more like a piece of text composed of amino acids, and since, when one amino acid may be related to the amino acids around it or those even farther away, it is not enough to extract these features using only the CNN mechanism. We also need to consider the information before and after the protein sequence and the correlation between discontinuous amino acids. Intuitively, an amino acid or a segment of amino acids may have a great impact on the protein sequence, so we can set a higher weight to this or this part of amino acids and have thus introduced the attention layer into the network.

Attention is a network structure model layer based on encode-decode, which has achieved satisfying prediction results compared to other traditional models in many fields including machine translation, picture description, and speech recognition. This implementation of the attention mechanism retains the intermediate output results of the input sequence of an LSTM encoder, then trains a model to selectively learn these inputs and associate the output sequence with them when the model outputs.

We added the attention and full connection layers in parallel after the convolution and pool layers, so that the model can not only take advantage of the mechanism of attention to learn the front and back features of the sequence, but also use the advantages of CNN feature extraction.

Our framework, ACNNT3, first uses multiple convolution and pooling layers to automatically learn protein sequence features, then takes the output feature vector as the input of the attention layer to calculate the score showing whether the neural network pays attention to the sequence features of the location. We define the output after convolution and pooling as a matrix $M^c (d \times q)$, where d is the number of convolution kernels, q is the whole position after sequence pooling, and the column j of the feature map matrix M^c can be viewed as a feature vector (denoted by V_j). W_j is the normalized importance score that is used to further average the columns of the feature map matrix M^c . The dense matrix output through the attention network is M^a , i.e.,

$$M^a = \sum_{j=1}^q w_j v_j, \quad (2)$$

$$w_j = \frac{\exp(e_j)}{\sum_{j=1}^q \exp(e_j)}, \quad (3)$$

where e_j is the importance score of the shared network and W_j is the relevant standardization score.

In order to integrate the features after convolution-pooling and the feature output by the attention layer, we first connect all the values in M^c and project them linearly to a value that represents the contribution of the whole sequence, represented by S^c , then we concatenate it with the dense representation M^a and input it into the logistic regression classifier to obtain the prediction score, namely,

$$\text{Pred}(s) = \text{sigm}(\text{concat}(M^a, S^c)), \quad (4)$$

where s represents a position in the integrated sequence.

$$S^c = \text{dense}(\text{pool}(\text{conv}(\text{encode}(s))))), \quad (5)$$

where $\text{encode}(\cdot)$, $\text{conv}(\cdot)$, $\text{pool}(\cdot)$, $\text{concat}(\cdot)$, $\text{dense}(\cdot)$, and $\text{sigmoid}(\cdot)$ represent the unification of one-hot and PSSM encoding, convolution, maximum pooling, concatenation, dense connection, and sigmoid operation, respectively. At the same time, for a specific sequence, we can also output a weight vector, i.e.,

$$\text{AttMap}(s) = (w_1, \dots, w_q). \quad (6)$$

This formula is used to express the attention of the model to each position of the input sequence.

2.4. Model Training. ACNNT3 is composed of a series of modules which use the fusion features of 100 amino acids at the N-terminal of the protein as input to predict T3SEs (Figure 1). The ACNNT3 model consists of convolution, pooling, attention, and fully connected layers. We use cross-validation to train our model and improve the generalization ability. The loss function uses a binary cross entropy loss function, and the optimizer uses the Adam algorithm. In Figure 2, we give the accuracy (ACC) comparison on the independent datasets under different epochs and batches. Since the dataset is not very large, the number of training epochs is set as 50 and the best batch value on the crossvalidation set is 10 as the optimal setting.

2.5. Performance Evaluation. We used 5-fold crossvalidation to estimate the classification performance of our model. Namely, we repeated the process five times and recorded the training parameters and average performance parameters for each time. The commonly used evaluation indexes for two-class classification are precision (PRE), sensitivity (SN), specificity (SP), F1 score, accuracy (ACC), and Matthew's correlation coefficient (MCC):

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (7)$$

$$\text{SN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (8)$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (9)$$

$$\text{F1 score} = 2 \times \frac{\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \quad (10)$$

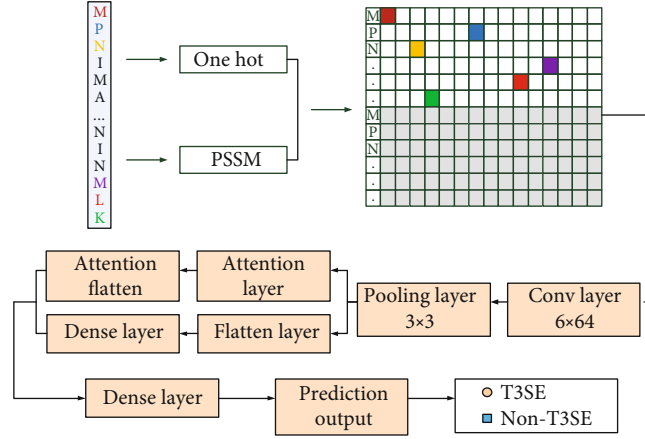


FIGURE 1: ACNNT3 architecture for T3SE prediction. Firstly, 64 1D convolution kernels with a length of 6 are convoluted to generate a 195×64 feature map, and then a 65×64 feature map is obtained through a 3×1 maximum pooling layer. The feature map is then input to the attention and full connection layers, and the two output results are combined to get 66 nodes. Finally, the 66 nodes are fully connected to the two output nodes, and the sigmoid function is used to activate to get the prediction probability of T3SE and non-T3SE.

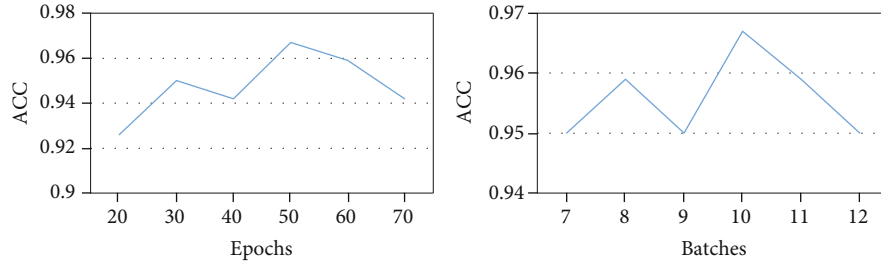


FIGURE 2: ACC comparison on the independent dataset under different epochs and batches.

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \quad (11)$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}}, \quad (12)$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative protein datasets, respectively.

The ROC curve is the relationship between the true positive and false positive rates, which is used to measure the comprehensive performance of different methods. The area under the ROC curve (AUC) is commonly used as a summary measure of diagnostic accuracy. In the ROC curve, the horizontal axis is the FPR (false positive rate, i.e., the ratio of wrongly predicted pairs over the total number of negative pairs), and the vertical axis is the TPR (true positive rate, i.e., the ratio of correctly predicted pairs over the total number of positive pairs). The maximum AUC is 1, which means a perfect prediction, and the AUC obtained by a random guess is 0.5.

3. Results

We have constructed a new prediction model to identify T3SEs by using a neural network that combines attention

with CNN. In order to study the influence of the negative sample set on performance, we divided the training set into two parts. The positive to negative ratio of training set 1 is 1:1, and the positive to negative ratio of training set 2 is 1:3. The ACNNT3 model was trained using training sets 1 and 2, respectively. To evaluate the classification performance of our ACNNT3 model, we use ROC and AUC as the evaluation criterion. The ROC charts of 5-fold crossvalidation curves under training sets 1 and 2 are shown in Figures 3(a) and 3(b). We can see that the ACNNT3 model achieved a good performance on the ROC chart. The mean AUC of the model is 0.95 on training set 1 and 0.98 on training set 2. These results show that our ACNNT3 model can accurately classify T3SEs and non-T3SEs on both training sets.

3.1. Comparison of Different Features on the Same Network. We take the one-hot single feature and the fused feature containing the one-hot matrix and PSSM as inputs, respectively, using ACNNT3 as the training model, and use the independent dataset to evaluate the two models. The results show that in all evaluation indexes, the model with the fusion feature is superior to the one with single feature training, thus verifying the proposed fusion feature's effectiveness (Figure 4). Compared to the one-hot single feature, the fusion feature is more comprehensive for the extraction of protein sequence information, and it can be seen from the

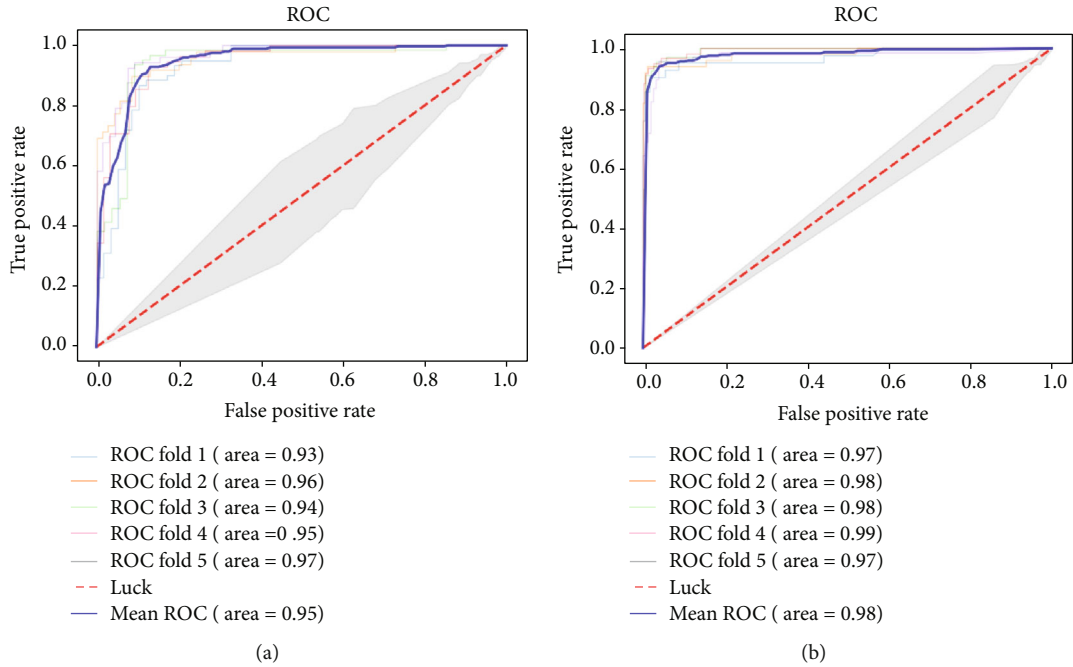


FIGURE 3: ROC curves on different training sets. (a) Use 5-fold crossvalidation experiment on training set 1. (b) Use 5-fold crossvalidation experiment on training set 2.

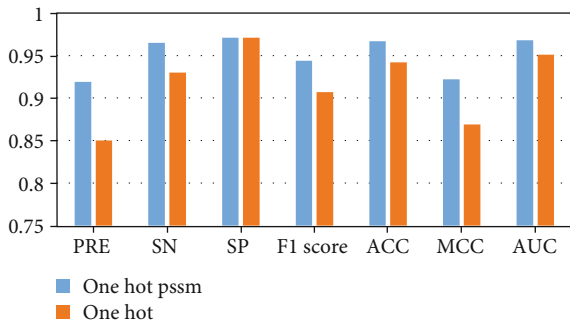


FIGURE 4: Comparison of experimental results of fusion feature and single feature under the same network model.

TABLE 1: Comparison with mainstream deep-learning methods.

Method	PRE	F1 score	ACC	MCC	AUC
ACNNT3	0.919	0.944	0.967	0.922	0.968
DenseNet	0.850	0.907	0.942	0.870	0.951
VGG16	0.846	0.892	0.934	0.847	0.937
ResNet	0.609	0.691	0.838	0.552	0.795
CNN	0.780	0.842	0.901	0.776	0.904
LSTM	0.875	0.933	0.959	0.909	0.961

The bold values indicate the best prediction results.

experimental results that two types of features have good compatibility with each other.

3.2. Comparison of Different Deep-Learning Methods. We compared the results from different popular network models using the independent dataset with the same feature input, as

TABLE 2: Comparison of ACNNT3 and DeepT3, Effective T3, BPBAac, and BEAN2 on an independent dataset.

Method	PRE	SN	SP	F1 score	ACC	MCC	AUC
ACNNT3-1	0.919	0.971	0.965	0.944	0.967	0.922	0.968
ACNNT3-2	0.711	0.914	0.849	0.800	0.868	0.716	0.882
DeepT3-1	0.825	0.943	0.919	0.880	0.926	0.830	0.974
DeepT3-2	0.643	0.771	0.825	0.701	0.810	0.569	0.896
Effective T3	0.542	0.839	0.741	0.658	0.767	0.521	0.803
BPBAac	0.944	0.548	0.988	0.694	0.871	0.656	0.902
BEAN2	0.674	0.935	0.835	0.784	0.862	0.706	0.865

The bold values indicate the best prediction results.

TABLE 3: Comparison of ACNNT3 and DeepT3, Effective T3, BPBAac, and BEAN2 on a *P. syringae* dataset.

Method	PRE	SN	SP	F1 score	ACC	MCC	AUC
ACNNT3-1	0.900	0.976	0.357	0.936	0.887	0.452	0.667
ACNNT3-2	0.872	0.988	0.143	0.926	0.866	0.265	0.565
DeepT3-1	0.905	0.962	0.429	0.932	0.884	0.472	0.838
DeepT3-2	0.913	0.924	0.500	0.918	0.860	0.437	0.763
Effective T3	0.906	0.906	0.428	0.906	0.838	0.334	0.810
BPBAac	0.875	0.494	0.571	0.631	0.505	0.046	0.562
BEAN2	0.883	0.988	0.083	0.938	0.884	0.271	0.607

The bold values indicate the best prediction results.

shown in Table 1. For a class of sequential processing problems, the addition of an attention layer makes the network model strengthen the connection before and after the amino

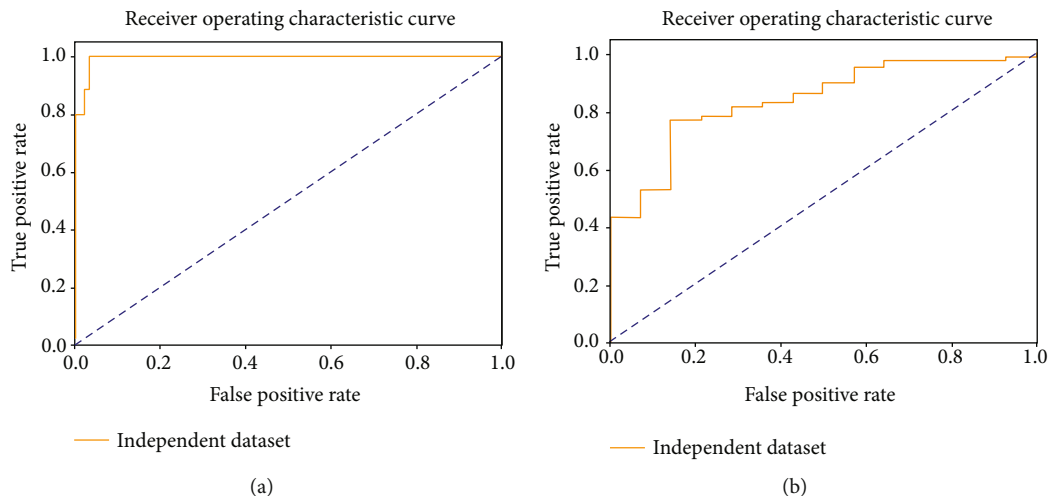


FIGURE 5: ROC curve of the best model selected from the 5-fold crossvalidation on two datasets. (a) ROC curve on a common independent dataset. (b) ROC curve on a *P. syringae* dataset.

acid and the attention of important information in the sequence. From the experimental results, it can be seen that our network model ACNNT3 is better than the existing deep-learning framework for predicting T3SEs in many indicators.

3.3. Comparison with Existing Methods. In order to evaluate the effectiveness of our method, we compared the ACNNT3 performance with four popular methods, DeepT3 [11], BPBAac [22], Effective T3 [6], and BEAN2 [34], on the same independent dataset. The parameter settings of these methods are the same as those used by Li et al. [11]. We found that our ACNNT3-1 model has a higher SN, F1 score, ACC, and MCC than the other four methods (Table 2). The results also show that our method achieved satisfactory performance in almost all indicators. For the important index of ACC, the accuracy of ACNNT3-1 is 0.967, which is 9.9%, 4.1%, 15.7%, 20.0%, 9.6%, and 10.5% higher than ACNNT3-2, DeepT3-1, DeepT3-2, Effective T3, BPBAac, and BEAN2, respectively. In another *P. syringae* dataset, our model still performed better than the existing methods on the index of ACC (Table 3). The accuracy of ACNNT3-1 is 0.887. We selected the best model in the fivefold crossvalidation and used the independent and *P. syringae* datasets to test it. We also obtained the ROC curves of the model on two test sets (Figure 5). Overall, our method has been shown to be superior to all the latest methods in T3SE prediction and is reliably stable.

4. Conclusion

We have proposed a new prediction model for Gram-negative bacteria type III secreted proteins based on a deep neural network. In order to better learn the feature preference of type III secreted proteins, we integrated the one-hot encoding and PSSM extracted from the protein primary sequence as the feature input and embedded the attention layer into CNN to improve the model's prediction ability. This method outperforms other existing methods on most

indicators, and using feature and network model comparisons, we have shown its advantages. In comparison with other popular methods, ACNNT3 is more accurate at predicting and recognizing T3SEs in the independent test set, which reflects its advantages and effectiveness. However, we found that ACNNT3's performance using the *P. syringae* dataset is not particularly obvious and was only slightly higher than the previous methods in terms of ACC and MCC. Our work in the future will focus on achieving better results in other experimental indicators and on applying this model for prediction using other large-scale datasets.

For easy implementation, all data used in this work and the source code for feature computing can be accessible at <https://github.com/Lijiesky/ACNNT3>.

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 by the National Natural Science Foundation of China under Grant Nos. 11671009 and 61762035, Zhejiang Provincial Natural Science Foundation of China under Grant Nos. LZ19A010002 and LY18F020027.

References

- [1] M. Desvaux, M. Hébraud, R. Talon, and I. R. Henderson, "Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue," *Trends in Microbiology*, vol. 17, no. 4, pp. 139–145, 2009.
- [2] G. R. Cornelis, "The type III secretion injectisome," *International Journal of Medical Microbiology*, vol. 4, no. 11, pp. 811–825, 2006.

- [3] S. Y. He, K. Nomura, and T. S. Whittam, "Type III protein secretion mechanism in mammalian and plant pathogens," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1694, no. 1-3, pp. 181–206, 2004.
- [4] G. N. Schroeder and H. Hilbi, "Molecular pathogenesis of *Shigella* spp.: controlling host cell signaling, invasion, and death by type III secretion," *Clinical Microbiology Reviews*, vol. 21, no. 1, pp. 134–156, 2008.
- [5] J. Engel and P. Balachandran, "Role of *Pseudomonas aeruginosa* type III effectors in disease," *Current Opinion in Microbiology*, vol. 12, no. 1, pp. 61–66, 2009.
- [6] R. Arnold, S. Brandmaier, F. Kleine et al., "Sequence-based prediction of type III secreted proteins," *PLoS Pathogens*, vol. 5, no. 4, article e1000376, 2009.
- [7] M. Löwer and G. Schneider, "Prediction of type III secretion signals in genomes of gram-negative bacteria," *PLoS One*, vol. 4, no. 6, article e5917, 2009.
- [8] X. Dong, Y.-J. Zhang, and Z. Zhang, "Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes," *PLoS One*, vol. 8, no. 2, article e56632, 2013.
- [9] S. Xie, Z. Li, and H. Hu, "Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization," *Gene*, vol. 642, no. 5, pp. 74–83, 2018.
- [10] X. Yang, Y. Guo, J. Luo, X. Pu, and M. Li, "Effective identification of gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles," *PLoS One*, vol. 8, no. 12, article e84439, 2013.
- [11] L. Xue, B. Tang, W. Chen, and J. Luo, "DeepT3: deep convolutional neural networks accurately identify gram-negative bacterial type III secreted effectors using the n-terminal sequence," *Bioinformatics*, vol. 35, no. 12, pp. 2051–2057, 2019.
- [12] H. Hu, Z. Li, A. Elofsson, and S. Xie, "A Bi-LSTM based ensemble algorithm for prediction of protein secondary structure," *Applied Sciences*, vol. 9, no. 17, article 3538, 2019.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [14] Z. Li, J. Zhu, X. Xu, and Y. Yao, "RDense: a protein-RNA binding prediction model based on bidirectional recurrent neural network and densely connected convolutional networks," *IEEE Access*, vol. 8, pp. 14588–14605, 2020.
- [15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," <http://arxiv.org/abs/1409.1556>.
- [16] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular Systems Biology*, vol. 12, no. 7, p. 878, 2016.
- [17] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.
- [18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," <https://www.aclweb.org/anthology/N16-1174.pdf>.
- [19] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2204–2212, 2014.
- [20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," <http://arxiv.org/abs/1409.0473>.
- [21] P. Verga, E. Strubell, and A. McCallum, "Simultaneously self-attending to all mentions for full-abstract biological relation extraction," <https://arxiv.org/pdf/1802.10569>.
- [22] Y. Wang, Q. Zhang, M. Sun, and D. Guo, "High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles," *Bioinformatics*, vol. 27, no. 6, pp. 777–784, 2011.
- [23] D. M. M. Tay, K. R. Govindarajan, M. A. Khan et al., "T3SEdb: data warehousing of virulence effectors secreted by the bacterial type III secretion system," *BMC Bioinformatics*, vol. 11, Suppl 7, p. S4, 2010.
- [24] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.
- [25] D. A. Baltrus, M. T. Nishimura, A. Romanchuk et al., "Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates," *PLoS Pathogens*, vol. 7, no. 7, article e1002132, 2011.
- [26] Y. An, J. Wang, C. Li et al., "Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI," *Briefings in Bioinformatics*, vol. 19, no. 1, pp. 148–161, 2018.
- [27] J. D. Bendtsen, L. J. Jensen, N. Blom, G. von Heijne, and S. Brunak, "Feature-based prediction of non-classical and leaderless protein secretion," *Protein Engineering Design and Selection*, vol. 17, no. 4, pp. 349–356, 2004.
- [28] C. Casper-Lindley, D. Dahlbeck, E. T. Clark, and B. J. Staskawicz, "Direct biochemical evidence for type III secretion-dependent translocation of the AvrBs2 effector protein into plant cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 8336–8341, 2002.
- [29] Y. Yang, J. Zhao, R. L. Morgan, W. Ma, and T. Jiang, "Computational prediction of type III secreted proteins from gram-negative bacteria," *BMC Bioinformatics*, vol. 11, article S47, 2010.
- [30] K. Schesser, E. Frithz-Lindsten, and H. Wolf-Watz, "Delineation and mutational analysis of the *Yersinia pseudotuberculosis* YopE domains which mediate translocation across bacterial and eukaryotic cellular membranes," *Journal of Bacteriology*, vol. 178, no. 24, pp. 7227–7233, 1997.
- [31] M. P. Sory, A. Boland, I. Lambermont, and G. R. Cornelis, "Identification of the YopE and YopH domains required for secretion and internalization into the cytosol of macrophages, using the *cyaA* gene fusion approach," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 26, pp. 11998–12002, 1996.
- [32] 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.
- [33] H. B. Shen and K. C. Chou, "Virus-mPLOC: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites," *Journal of Biomolecular Structure and Dynamics*, vol. 28, no. 2, pp. 175–186, 2010.
- [34] X. Dong, X. Lu, and Z. Zhang, "BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors," *Database*, vol. 2015, article bav064, 2015.

Research Article

HMMPred: Accurate Prediction of DNA-Binding Proteins Based on HMM Profiles and XGBoost Feature Selection

Xiuzhi Sang ¹, Wanyue Xiao ², Huiwen Zheng ³, Yang Yang ⁴, and Taigang Liu ¹

¹College of Information, Shanghai Ocean University, Shanghai 201306, China

²School of Information, Syracuse University, Syracuse, NY 13244, USA

³School of Engineering, University of Melbourne, Victoria 3010, Australia

⁴School of Information Management, Nanjing University, Nanjing 210023, China

Correspondence should be addressed to Yang Yang; njukyong@hotmail.com and Taigang Liu; tgliu@shou.edu.cn

Received 29 January 2020; Accepted 16 March 2020; Published 28 March 2020

Guest Editor: Lei Chen

Copyright © 2020 Xiuzhi Sang 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.

Prediction of DNA-binding proteins (DBPs) has become a popular research topic in protein science due to its crucial role in all aspects of biological activities. Even though considerable efforts have been devoted to developing powerful computational methods to solve this problem, it is still a challenging task in the field of bioinformatics. A hidden Markov model (HMM) profile has been proved to provide important clues for improving the prediction performance of DBPs. In this paper, we propose a method, called HMMPred, which extracts the features of amino acid composition and auto- and cross-covariance transformation from the HMM profiles, to help train a machine learning model for identification of DBPs. Then, a feature selection technique is performed based on the extreme gradient boosting (XGBoost) algorithm. Finally, the selected optimal features are fed into a support vector machine (SVM) classifier to predict DBPs. The experimental results tested on two benchmark datasets show that the proposed method is superior to most of the existing methods and could serve as an alternative tool to identify DBPs.

1. Introduction

DNA-binding proteins (DBPs), which can bind to and interact with DNA, play prominent roles in the structural composition of DNA and the regulation of genes. These proteins have a variety of biochemical functions in the cell and molecular biology, including the participation and regulation of various cellular processes, such as transcription, DNA replication, recombination, modification, and repair [1, 2]. Besides, DBPs are key components of steroids, antibiotics, and cancer drugs in the pharmaceutical industry [3]. Hence, the prediction of DBPs has become one of the research focuses in the field of protein science due to its significance in the related biological activities. In early studies, DBPs were normally identified by experimental techniques, such as filter binding assays, genetic analysis, X-ray crystallography, ChIP-chip analysis, and nuclear magnetic resonance (NMR) [4]. However, conventional experimental methods are often time-consuming and laborious. With the rapid increase of protein sequence data, there is a great need

to develop efficient computational methods to identify DBPs solely based on their primary sequences.

From the machine learning perspective, identification of DBPs is usually considered a binary classification problem. In recent years, many computational methods have been applied to solve this problem. These methods primarily focus on the following two aspects: (1) the construction of encoding schemes for protein sequences and (2) the application of classification algorithms. Many machine learning techniques have been adopted to perform the prediction of DBPs, including support vector machine (SVM) [5–7], random forest (RF) [8–10], naive Bayes classifier [4], ensemble classifiers [11–13], and deep learning [14–16]. Among these algorithms, SVM and RF have been widely used because of their excellent performance. The existing SVM-based predictive methods differ in encoding schemes for protein sequences. A great number of sequence features have been applied to represent protein sequences into fixed-length numeric vectors, such as amino acid composition (AAC) [17], dipeptide composition [18],

pseudo-AAC [19–22], position-specific score matrix (PSSM) profile [23–27], predicted secondary structure [28], and hidden Markov model (HMM) profile [29].

Numerous researches have proved that evolutionary information encoded in the PSSM profile is more informative than protein sequence alone [30]. The PSSM profiles have been widely used in bioinformatics, such as protein remote homology detection [31], protein fold recognition [32], and prediction of protein structural class [33]. Accordingly, PSSM-based feature descriptors have successfully enhanced the prediction accuracy of DBPs. For example, Kumar et al. [24] first adopted the PSSM profile to identify DBPs and constructed an SVM model called DNAbinder. Waris et al. [25] further developed a classifier by integrating the PSSM profile and other two protein representations, i.e., dipeptide composition and split AAC. Besides, the method of Wang et al. [26] applied the discrete cosine transform and the discrete wavelet transform to compress the PSSM profile and achieved excellent prediction performance. Wei et al. [9] proposed a powerful predictor called Local-DPP, which combined the local pseudo-PSSM features with the RF classifier. Recently, Zaman et al. [29] build a predictive model based on the HMM profile instead of the PSSM profile for the detection of DBPs and experimentally showed the effectiveness of the HMM-based features by using the jackknife test on the benchmark dataset. However, the method proposed by Zaman et al. performed relatively poorly on the independent dataset test [29]. It appears that evolutionary information in the form of HMM profile has not been adequately explored and there is still room for developing more effective feature extraction techniques to improve the prediction performance of DBPs.

To this end, we propose a novel method, called HMMPred, which utilizes features extracted solely from the HMM profile to further improve the prediction accuracy of DBPs. First, HMM profiles are transformed into fixed-length feature vectors with the joint use of three feature extraction methods including AAC, auto covariance transformation (ACT), and cross-covariance transformation (CCT). Next, the extreme gradient boosting (XGBoost) algorithm is adopted as a feature selection technique to pick the well-distinguished features. Finally, these selected optimal features are fed into an SVM classifier to make predictions. Validation results on two working datasets indicate that the proposed method performs better than most of the other existing predictors, especially the remarkably high accuracy on the independent dataset.

2. Materials and Methods

This section illustrates all details about our proposed method and the following flow chart (Figure 1) clearly presents the process framework of the method. This process involves both training and testing stages. For the training phase, the HMM profiles of query proteins are generated by running the HHblits program, which is an effective sequence alignment tool with less running time but higher sensitivity and accuracy than PSI-BLAST [34]. Next, features are extracted from the HMM profiles by fusing three techniques, i.e., AAC,

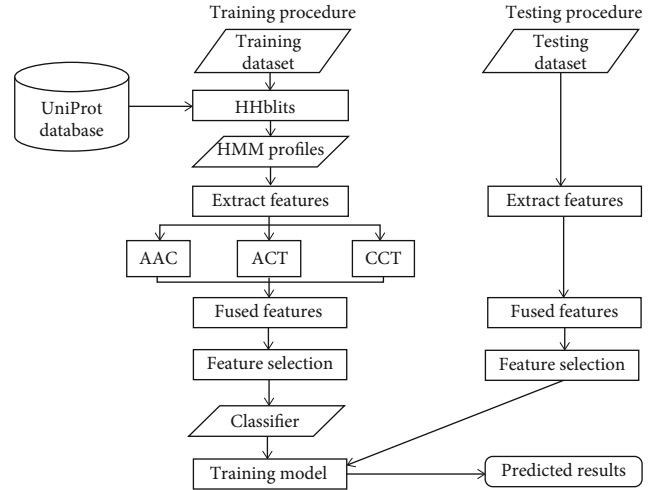


FIGURE 1: Framework of the proposed method for DBPs prediction.

ACT, and CCT. Then, the optimal features are selected and finally inputted into a classifier for the subsequent model training and DBPs prediction. For the testing phase, a series of procedures are similar to those in the previous part so that the prediction result can be obtained after feeding selected features into the training model, which is generated in the training stage.

2.1. Datasets. Two benchmark datasets, PDB1075 [22] and PDB186 [4], are used to measure the performance of the proposed method. The PDB1075 dataset which contains 525 DBPs and 550 non-DBPs is first applied for model training as well as testing by adopting cross-validation (CV) methods. On the other hand, the PDB186 dataset is adopted for an independent test to further evaluate the robustness and generalization ability of our predictor, which includes 93 DBPs and 93 non-DBPs. These protein sequences in the two datasets are selected from the Protein Data Bank [35] through a rigorous filtering procedure: (1) remove the sequences with a length of less than 50 amino acids or unknown residues such as “X”; and (2) cut off those sequences that have more than 25% sequence similarity with any other sequences.

2.2. Protein Sequence Representation

2.2.1. HMM Profiles. A previous study has shown that HMM profiles are more effective for DBPs prediction compared with PSSM profiles [29]. In this study, the HMM profile is generated by performing four iterations of HHblits against the newest UniProt database [36] with an E -value threshold of 0.001. Given a query protein of length L , the size of HMM profile is $L \times 30$. The values in HMM profile are converted to the range of (0, 1) by using the function $f(x) = 2^{-x/1000}$, where x is the original HMM value. Similar to the PSSM profile, we only use the first 20 columns of HMM profile.

2.2.2. Feature Extraction from HMM Profiles. Three feature extraction methods, i.e., AAC, ACT, and CCT, are adopted to transform HMM profiles into fixed-length feature vectors. It is well known that DNA-binding preference of a protein is closely related to its AAC features [17]. To

compute AAC features from the HMM profile, the following formula is used:

$$\bar{h}_j = \frac{1}{L} \sum_{i=1}^L h_{i,j} \quad (j = 1, 2, \dots, 20), \quad (1)$$

where L is the length of the protein sequence and $h_{i,j}$ represents the element at the i^{th} row and j^{th} column of the HMM profile. In this way, 20 AAC features are obtained in total.

Obviously, if only AAC features are used to represent the protein, all the sequence-order information would be lost. To solve this problem, we apply ACT and CCT to reflect the local sequence-order effect. These two techniques have been widely used to extract features from the PSSM profile [37–39]. Thus, in this work, ACT and CCT are also adopted to convert the HMM profile into two numerical vectors by using the following equations:

$$A_{j,g} = \frac{1}{L-g} \sum_{i=1}^{L-g} (h_{i,j} - \bar{h}_j) (h_{i+g,j} - \bar{h}_j), \quad (2)$$

$$C_{j,k,g} = \frac{1}{L-g} \sum_{i=1}^{L-g} (h_{i,j} - \bar{h}_j) (h_{i+g,k} - \bar{h}_k),$$

where $1 \leq j, k \leq 20$, $j \neq k$, and g is the lag. Hence, the number of ACT features is $20 \times G$, and the number of CCT features is $20 \times 19 \times G = 380 \times G$, where G is the maximum of g . As a result, each protein sequence can be represented as a $(20 + 400 \times G)$ -dimensional vector by fusing the AAC, ACT, and CCT features.

2.3. Feature Selection Algorithm. Feature selection plays a vital role in machine learning and pattern recognition, which can improve the performance of prediction models by removing irrelevant, noisy, and redundant information from the untreated features. In this study, we first obtain feature importance scores by applying RF and XGBoost algorithms individually. In the RF strategy, the importance of features is calculated by a total decrease in tree-node impurities from splitting off the predictor feature variable and is averaged over all sub-trees [40, 41]. The XGBoost method calculates an importance score for each feature based on its participation in making key decisions with boosted decision trees as suggested in [42]. Then, all of the features are ranked according to their importance scores. Finally, we select an optimal feature subset based on the ranked features. To the best of our knowledge, the XGBoost feature selection technique has not been explored for DBPs prediction.

2.4. Classification Algorithm. Two robust machine learning techniques, i.e., SVM and RF, are applied to perform the prediction of DBPs, which have been widely used for many classification tasks in the field of computational biology [43–46]. SVM is an outstanding classification method that is used to deal with a binary pattern recognition problem [47]. Its core idea is to find an optimal hyperplane as a decision surface, by maximizing the margin of separation between the two classes in the data. With the help of kernel tricks, SVM not only can

classify the linearly separable samples but also can handle classes with complex nonlinear decision boundaries. Popular kernels used with SVMs include linear, polynomial, sigmoid, and radial basis function (RBF). In this study, the RBF kernel is adopted due to its excellent performance in the previous tests and the values of parameters C and γ are optimized between 2^{-10} and 2^{10} based on the 10-fold CV using a grid search strategy.

RF, as an ensemble learning algorithm, is not only widely used in feature selection which is discussed before but also applied in classification [48]. It is composed of many decision trees, and each tree in the forest makes a judgment on the sample to determine whether it belongs to positive instances or negative ones. Then, all voting results from each tree are collected to finally classify the samples into the category with the maximum votes. The SVM and RF algorithms were implemented using the Python sklearn library [49]. All experiments in this study were carried out in version 3.7 of Python.

2.5. Performance Evaluation. The performance of HMMPred is evaluated by three commonly used tests: 10-fold CV and jackknife CV implemented on the PDB1075 dataset, and an independent test where the PDB1075 dataset is used to train the model and testing is on the PDB186 dataset. All results are reported using the following four performance metrics: sensitivity (SN), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC) [50, 51]. These metrics are formulated as follows:

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

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

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

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \quad (3)$$

where TN, FN, TP, FP, respectively, represent the number of true negative, false negative, true positive, and false positive samples predicted. In addition, we also compute the area under the receiver operating characteristic (ROC) curve (AUC), which is a preferred metric for evaluating the performance of a binary classifier.

3. Results and Discussion

3.1. The Impact of the Parameter g on Prediction Performance. The ACT and CCT features represent the average correlation of two amino acids separated by g positions along the query protein sequence. To investigate the impact of parameter g on the prediction performance, we compare the prediction results by increasing the value of g from 1 to 10 with an increment value of 1, using the RF classifier and the SVM classifier under two different evaluation methods

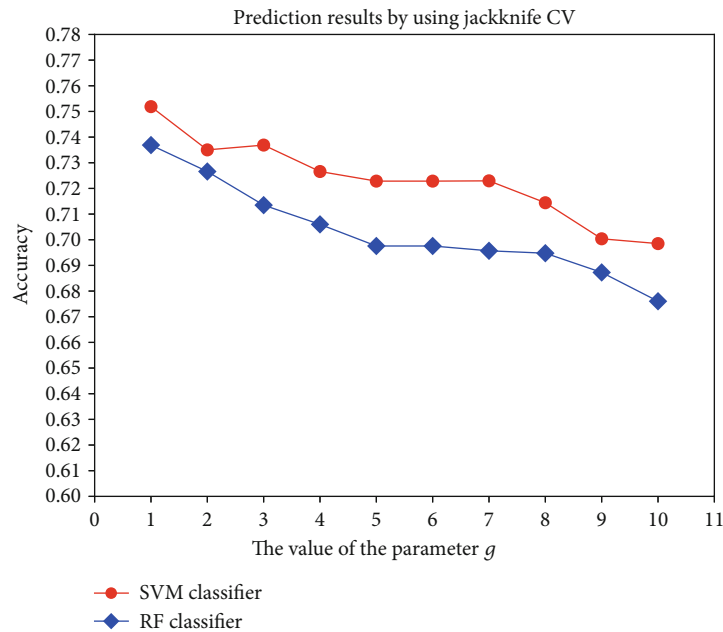
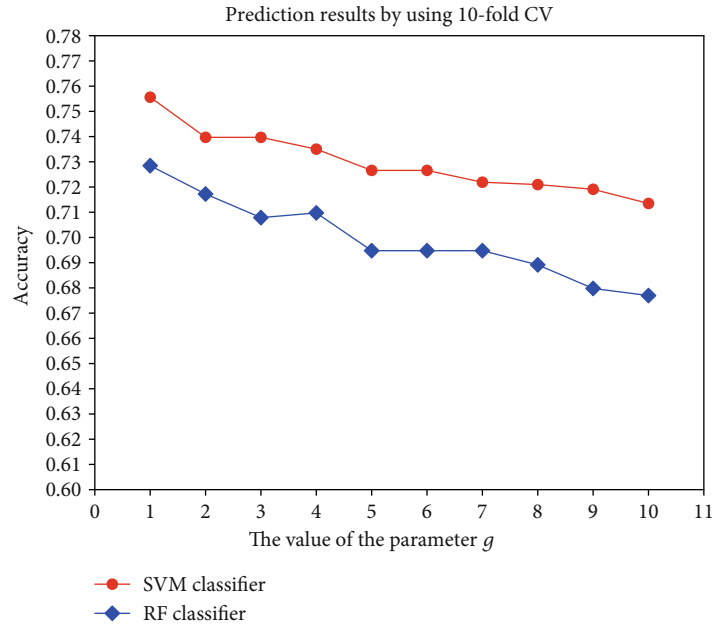


FIGURE 2: This shows how different g values affect the accuracies based on two CV methods. (a) The prediction results by using the 10-fold CV method. (b) The prediction results by using the jackknife CV method.

individually. Given that the accuracy rate is used as a crucial evaluation criterion in the model assessment (Figure 2), some insights into the selection of optimal g value and classifier are summarized below.

The following figures exhibit two striking traits. Firstly, the accuracy rate dwindles with the gradual increases of parameter g . Secondly, the accuracies of the SVM classifier are consistently better than those of the RF classifier. Referring to Figure 2(a), when the value of g is greater than 7, both SVM and RF classifiers show relatively poor performance. In

addition, the accuracies remain relatively stable with g ranging from 5 to 7. A similar conclusion could be drawn from Figure 2(b). On the other hand, the increment of G (i.e., the maximum of g) followed by the growth of feature dimension could cause issues of feature redundancy, additional computational cost, and extra time consumption. Hence, to make a trade-off between the accuracy rate and the number of feature dimension, keeping the maximum of g to 5 is recommended. Accordingly, the number of ACT features is 100 and the number of CCT features is 1900.

TABLE 1: Prediction results of SVM and RF classifiers based on the 10-fold CV.

Classifier	Feature extraction method	ACC	SN	SP	MCC	AUC
SVM	AAC	0.7893	0.8224	0.7582	0.5810	0.8586
	ACT	0.7004	0.6795	0.7200	0.3999	0.7492
	CCT	0.7678	0.7336	0.8000	0.5352	0.8309
	AAC+ACT+CCT	0.8034	0.8147	0.7927	0.6071	0.8717
RF	AAC	0.7772	0.8147	0.7418	0.5571	0.8600
	ACT	0.7369	0.7394	0.7345	0.4737	0.8022
	CCT	0.7566	0.7896	0.7255	0.5154	0.8232
	AAC+ACT+CCT	0.7781	0.8205	0.7382	0.5596	0.8437

TABLE 2: Prediction results of SVM and RF classifiers based on the jackknife CV.

Classifier	Feature extraction method	ACC	SN	SP	MCC	AUC
SVM	AAC	0.7912	0.8185	0.7655	0.5841	0.8663
	ACT	0.7004	0.6795	0.7200	0.3999	0.7641
	CCT	0.7650	0.7297	0.7982	0.5296	0.8373
	AAC+ACT+CCT	0.8015	0.8127	0.7909	0.6034	0.8806
RF	AAC	0.7930	0.8161	0.7618	0.5885	0.8705
	ACT	0.7369	0.7413	0.7327	0.4738	0.8125
	CCT	0.7547	0.7761	0.7345	0.5106	0.8299
	AAC+ACT+CCT	0.7706	0.8050	0.7382	0.5437	0.8539

3.2. *Comparative Analysis of Different Classifiers with Different CV Methods.* In this section, we further compare the performance between SVM and RF classifiers combined with four different feature extraction techniques including AAC, ACT, CCT, and AAC+ACT+CCT, respectively. Results on the PDB1075 dataset by using two CV tests are listed in Tables 1 and 2.

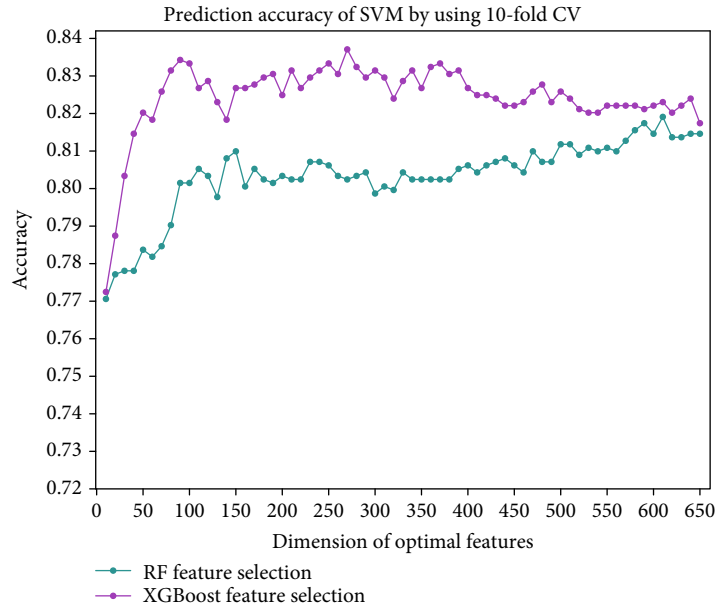
As shown in Table 1, the combination of SVM classifier with AAC+CCT+CCT features achieves the highest accuracy rate (0.8034) compared with others using the same classifier but with different features. Both MCC and AUC measures also give similar results. Meanwhile, the AAC feature and CCT feature obtain the highest SN and SP, respectively, suggesting that these two features are crucial to the identification of DBPs. For the RF classifier, AAC+ACT+CCT is also deemed to be the most appropriate method. Except for SP and AUC, the results of AAC+ACT+CCT consistently outperform the other three feature extraction methods. Apparently, the SVM classifier is more superior to the RF classifier in this experiment.

According to Table 2, similar conclusions can be reached by using the jackknife CV. For the SVM classifier, AAC+ACT+CCT is considered the optimal method with an accuracy rate of 0.8015. The RF classifier provides the accuracy rate of 0.7706 by using AAC+ACT+CCT features, which is higher than the cases with ACT and CCT features but is lower than the case with AAC features (0.7930). This suggests that multifeature fusion could generate irrelevant noise information and feature selection is necessary to enhance the prediction of DBPs in the next step.

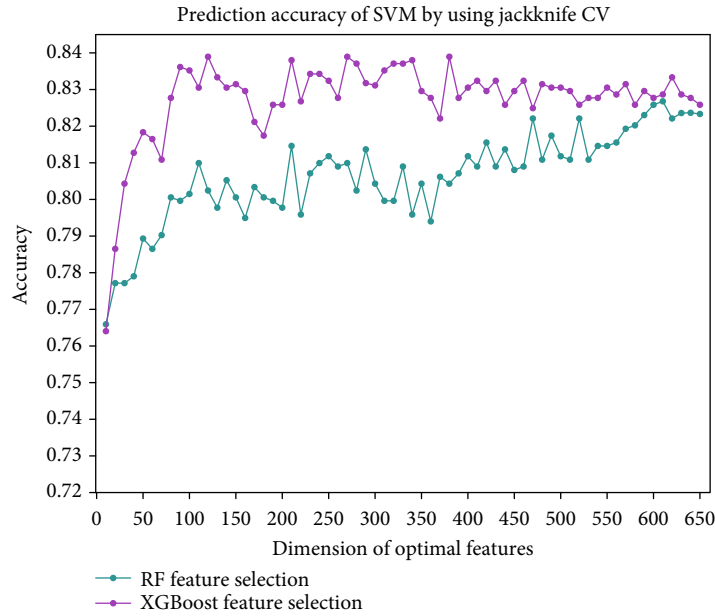
Therefore, after analysing the data obtained from the examinations above, the combination of the SVM classifier with the joint use of the AAC+ACT+CCT features is adopted in the subsequent analysis due to its finest achievement.

3.3. *Performance Analysis of Feature Selection.* By combining AAC+ACT+CCT features, we firstly obtain a 2020-dimensional vector for each protein. Then, these features are ranked according to their importance by applying RF and XGBoost techniques, respectively. To further determine the optimal feature subset, we calculate the accuracies for top K features by using the 10-fold CV and the jackknife CV, respectively, where $K = 10, 20, 30, \dots, 650$. The results on the PDB1075 dataset are illustrated in Figure 3. As can be observed from Figure 3(a), feature subsets ranked by the XGBoost method could obtain higher accuracies compared with the RF feature ranking technique. When $K = 270$, the highest accuracy of 0.8371 is achieved by using the 10-fold CV. Considering that Figure 3(b) also shows similar results, it is appropriate to pick the top 270 ranked features for the following analyses.

Table 3 further examines the effectiveness of the feature selection by comparing the prediction performance of the case without using feature selection, the case using RF feature ranking, and the case using XGBoost feature ranking. Two CV methods, i.e., 10-fold and jackknife, are tested on the PDB1075 dataset by running the SVM classifier, respectively. From Table 3, two main results emerge: (i) the feature selection technique can indeed help to effectively improve the performance of DBPs prediction; and (ii) the XGBoost



(a)



(b)

FIGURE 3: This illustrates how different feature subsets affect the accuracies by using two different feature selection methods. (a) The prediction accuracy of SVM based on the 10-fold CV test. (b) The prediction accuracy of SVM based on the jackknife CV test.

TABLE 3: Performance comparison before and after feature selection.

Feature selection	CV methods	ACC	SN	SP	MCC	AUC
Before	10-fold	0.8034	0.8147	0.7927	0.6071	0.8720
	Jackknife	0.8015	0.8127	0.7909	0.6034	0.8805
RF	10-fold	0.8221	0.8243	0.8200	0.6441	0.8819
	Jackknife	0.8267	0.8262	0.8272	0.6533	0.8946
XGBoost	10-fold	0.8371	0.8301	0.8436	0.6738	0.8896
	Jackknife	0.8390	0.8398	0.8382	0.6778	0.9018

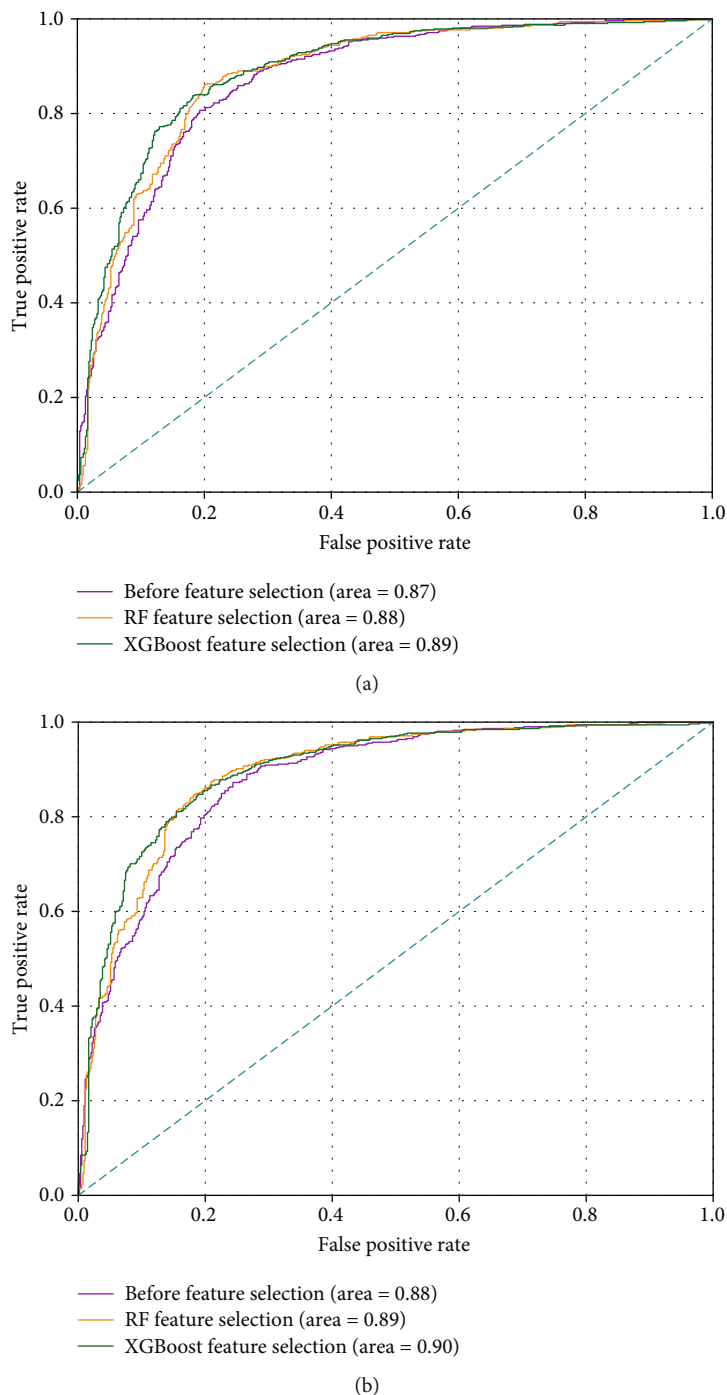


FIGURE 4: ROC curves of the SVM classifier before and after feature selection. (a) ROC curves based on the 10-fold CV. (b) ROC curves based on the jackknife CV.

algorithm may be able to provide better feature ranking than the RF method. We also plot the ROC curves for these experiments in Figure 4, which demonstrates the remarkably consistent findings.

3.4. Comparison with Existing Predictors. To objectively evaluate the effectiveness of the proposed method, we make comparisons with some existing predictors on the same datasets. These methods include DNAbinder [24], DNA-Prot [8],

iDNA-Prot [10], iDNA-Prot[dis [22], Kmer1+ACC [52], iDNAPro-PseAAC [27], PseDNA-Pro [19], Local-DPP [9], and HMMBinder [29]. The results of jackknife tests on the PDB1075 dataset are listed in Table 4. In addition, Table 5 illustrates five performance measures of various algorithms tested on the PDB186 independent dataset.

As shown in Table 4, the proposed method achieves the values of “ACC” (83.90%), “SP” (83.82%), “MCC” (0.68), and “AUC” (0.9018), which rank second on the benchmark

TABLE 4: Performance comparison on the PDB1075 dataset.

Methods	ACC	SN	SP	MCC	AUC
DNAbinder	0.7395	0.6857	0.7909	0.48	0.8140
DNA-Prot	0.7255	0.8267	0.5976	0.44	0.7890
iDNA-Prot	0.7540	0.8381	0.6473	0.50	0.7610
iDNA-Prot dis	0.7730	0.7940	0.7527	0.54	0.8260
Kmer1+ACC	0.7523	0.7676	0.7376	0.50	0.8280
iDNAPro-PseAAC	0.7656	0.7562	0.7745	0.53	0.8392
PseDNA-Pro	0.7655	0.7961	0.7363	0.53	—
Local-DPP	0.7920	0.8400	0.7450	0.59	—
HMMBinder	0.8633	0.8707	0.8555	0.72	0.9026
Our method	0.8390	0.8398	0.8382	0.68	0.9018

TABLE 5: Performance comparison on the independent dataset.

Methods	ACC	SN	SP	MCC	AUC
DNAbinder	0.6080	0.5700	0.6450	0.216	0.6070
DNA-Prot	0.6180	0.6990	0.5380	0.240	—
iDNA-Prot	0.6720	0.6770	0.6670	0.344	—
iDNA-Prot dis	0.7200	0.7950	0.6450	0.445	0.7860
Kmer1+ACC	0.7096	0.8279	0.5913	0.431	0.7520
iDNAPro-PseAAC	0.7150	0.8276	0.6022	0.442	0.7780
Local-DPP	0.7900	0.9250	0.6560	0.625	—
HMMBinder	0.6902	0.6153	0.7634	0.394	0.6324
Our method	0.8118	0.9462	0.6774	0.648	0.8715

dataset and are merely below those of HMMBinder. The Local-DPP algorithm, which explored local evolutionary information from the PSSM profile, gets the comparable SN of 84% to our method. This indicates that the PSSM profile indeed can provide important clues for predicting DBPs. It is worth mentioning that the Kmer1+ACC method applied the same strategy to extract AAC, ACT, and CCT features from the PSSM profile instead of the HMM profile. Judging from the results of performance comparison, the HMM profile could serve as a better source of information for the identification of DBPs. From the values reported in Table 5, the proposed method obtains the highest ACC, SN, MCC, and AUC among these methods by using the independent dataset test. It should be noted that the HMMBinder method could not provide desired optimal results on the testing set despite achieving the best SP value. This might lead us to believe that there is a risk of overfitting in the HMMBinder method.

In summary, the proposed method shows substantial improvements for identifying DBPs particularly on the independent test, which are attributed to the powerful feature fusion method from the HMM profile and the efficient feature selection by using the XGBoost technique.

4. Conclusion

In this paper, we propose a method called HMMPred, which makes an effective improvement on the existing HMM profile-based method to predict DBPs by integrating three

feature extraction techniques (i.e., AAC, ACT, and CCT) and adding the application of a prominent feature selection method called XGBoost. Then, the top 270-dimensional features are fed into the SVM classifier to train the model. Based on the comprehensive assessment, using the 10-fold CV, the jackknife CV, and the independent test, it is noteworthy that our method performs well compared to other existing methods and even achieves superior performance on the independent test. In our future work, we would like to develop a web server for the public use and continue to enhance the existing methods for achieving more precise identification of DBPs.

Data Availability

The datasets and source codes for this study are freely available to the academic community at: <https://github.com/taigangliu/HMMPred>.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Xiuzhi Sang, Wanyue Xiao, and Huiwen Zheng contributed equally to this work as co-first authors.

Acknowledgments

The authors would like to thank Dr. Xiaoguang Bao for his pertinent suggestions. This work was funded by the National Natural Science Foundation of China (grant numbers 11601324, 11701363).

References

- [1] R. E. Langlois and H. Lu, "Boosting the prediction and understanding of DNA-binding domains from sequence," *Nucleic Acids Research*, vol. 38, no. 10, pp. 3149–3158, 2010.
- [2] K. A. Jones, J. T. Kadonaga, P. J. Rosenfeld, T. J. Kelly, and R. Tjian, "A cellular DNA-binding protein that activates eukaryotic transcription and DNA replication," *Cell*, vol. 48, no. 1, pp. 79–89, 1987.
- [3] 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.
- [4] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian Naïve Bayes," *PLoS One*, vol. 9, no. 1, article e86703, 2014.
- [5] L. Nanni and S. Brahnam, "Set of approaches based on 3D structure and position specific-scoring matrix for predicting DNA-binding proteins," *Bioinformatics*, vol. 35, no. 11, pp. 1844–1851, 2019.
- [6] K. Qu, K. Han, S. Wu, G. Wang, and L. Wei, "Identification of DNA-binding proteins using mixed feature representation methods," *Molecules*, vol. 22, no. 10, p. 1602, 2017.
- [7] J. Zhang and B. Liu, "PSFM-DBT: identifying DNA-binding proteins by combing position specific frequency matrix and distance-bigram transformation," *International Journal of Molecular Sciences*, vol. 18, no. 9, 2017.
- [8] 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 & Dynamics*, vol. 26, no. 6, pp. 679–686, 2009.
- [9] 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.
- [10] 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 e24756, 2011.
- [11] 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, 2019.
- [12] X.-J. Liu, X.-J. Gong, H. Yu, and J. H. Xu, "A model stacking framework for identifying DNA binding proteins by orchestrating multi-view features and classifiers," *Genes*, vol. 9, no. 8, p. 394, 2018.
- [13] W. You, Z. Yang, G. Guo, X. F. Wan, and G. Ji, "Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble," *Knowledge-Based Systems*, vol. 163, pp. 598–610, 2019.
- [14] Y.-H. Qu, H. Yu, X.-J. Gong, J. H. Xu, and H. S. Lee, "On the prediction of DNA-binding proteins only from primary sequences: a deep learning approach," *PLoS One*, vol. 12, no. 12, article e0188129, 2017.
- [15] S. Chauhan and S. Ahmad, "Enabling full-length evolutionary profiles based deep convolutional neural network for predicting DNA-binding proteins from sequence," *Proteins*, vol. 88, no. 1, pp. 15–30, 2020.
- [16] S. Hu, R. Ma, and H. Wang, "An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences," *PLoS One*, vol. 14, no. 11, article e0225317, 2019.
- [17] G. B. Motion, A. J. M. Howden, E. Huitema, and S. Jones, "DNA-binding protein prediction using plant specific support vector machines: validation and application of a new genome annotation tool," *Nucleic Acids Research*, vol. 43, no. 22, article e158, 2015.
- [18] L. Nanni and A. Lumini, "Combing ontologies and dipeptide composition for predicting DNA-binding proteins," *Amino Acids*, vol. 34, no. 4, pp. 635–641, 2008.
- [19] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.
- [20] S. Adilina, D. M. Farid, and S. Shatabda, "Effective DNA binding protein prediction by using key features via Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 460, pp. 64–78, 2019.
- [21] 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.
- [22] 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, no. 9, article e106691, 2014.
- [23] X. Fu, W. Zhu, B. Liao, L. Cai, L. Peng, and J. Yang, "Improved DNA-binding protein identification by incorporating evolutionary information into the Chou's PseAAC," *IEEE Access*, vol. 6, pp. 66545–66556, 2018.
- [24] M. Kumar, M. M. Gromiha, and G. P. Raghava, "Identification of DNA-binding proteins using support vector machines and evolutionary profiles," *BMC Bioinformatics*, vol. 8, no. 1, p. 463, 2007.
- [25] M. Waris, K. Ahmad, M. Kabir, and M. Hayat, "Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix," *Neurocomputing*, vol. 199, pp. 154–162, 2016.
- [26] Y. Wang, Y. Ding, F. Guo, L. Wei, and J. Tang, "Improved detection of DNA-binding proteins via compression technology on PSSM information," *PLoS One*, vol. 12, no. 9, article e0185587, 2017.
- [27] B. Liu, S. Wang, and X. Wang, "DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation," *Scientific Reports*, vol. 5, no. 1, article 15479, 2015.
- [28] S. Y. Chowdhury, S. Shatabda, and A. Dehzangi, "iDNAProt-ES: identification of DNA-binding proteins using evolutionary and structural features," *Scientific Reports*, vol. 7, no. 1, article 14938, 2017.
- [29] R. Zaman, S. Y. Chowdhury, M. A. Rashid, A. Sharma, A. Dehzangi, and S. Shatabda, "HMMBinder: DNA-binding

- protein prediction using HMM profile based features,” *BioMed Research International*, vol. 2017, Article ID 4590609, 10 pages, 2017.
- [30] J. Wang, B. Yang, J. Revote et al., “POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles,” *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, 2017.
- [31] B. Liu, D. Zhang, R. Xu et al., “Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection,” *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.
- [32] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, “A trigram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition,” *IEEE Transactions on Nanobioscience*, vol. 13, no. 1, pp. 44–50, 2014.
- [33] T. Liu, Y. Qin, Y. Wang, and C. Wang, “Prediction of protein structural class based on gapped-dipeptides and a recursive feature selection approach,” *International Journal of Molecular Sciences*, vol. 17, no. 1, p. 15, 2016.
- [34] M. Remmert, A. Biegert, A. Hauser, and J. Söding, “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment,” *Nature Methods*, vol. 9, no. 2, pp. 173–175, 2012.
- [35] H. M. Berman, J. Westbrook, Z. Feng et al., “The protein data bank,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.
- [36] The UniProt Consortium, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, 2017.
- [37] X. Li, T. Liu, P. Tao, C. Wang, and L. Chen, “A highly accurate protein structural class prediction approach using auto cross covariance transformation and recursive feature elimination,” *Computational Biology and Chemistry*, vol. 59, pp. 95–100, 2015.
- [38] Q. Dong, S. Zhou, and J. Guan, “A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation,” *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, 2009.
- [39] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, “Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles,” *Amino Acids*, vol. 42, no. 6, pp. 2243–2249, 2012.
- [40] J. T. Wassan, H. Wang, F. Browne, and H. Zheng, “Phy-PMRFI: phylogeny-aware prediction of metagenomic functions using random forest feature importance,” *IEEE Transactions on Nanobioscience*, vol. 18, no. 3, pp. 273–282, 2019.
- [41] K. J. Archer and R. V. Kimes, “Empirical characterization of random forest variable importance measures,” *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.
- [42] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, USA, 2016.
- [43] J. P. Zhou, L. Chen, and Z. H. Guo, “iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs,” *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.
- [44] X. Zhang, L. Chen, Z.-H. Guo, and H. Liang, “Identification of human membrane protein types by incorporating network embedding methods,” *IEEE Access*, vol. 7, pp. 140794–140805, 2019.
- [45] J. Li, L. Lu, Y.-H. Zhang et al., “Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine,” *Cancer Gene Therapy*, vol. 27, no. 1-2, pp. 56–69, 2020.
- [46] Y. Yao, X. Li, B. Liao et al., “Predicting influenza antigenicity from Hemagglutinin sequence data based on a joint random forest method,” *Scientific Reports*, vol. 7, no. 1, p. 1545, 2017.
- [47] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [48] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [49] F. Pedregosa, G. Varoquaux, A. Gramfort et al., “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [50] B. Li, L. Cai, B. Liao, X. Fu, P. Bing, and J. Yang, “Prediction of protein subcellular localization based on fusion of multi-view features,” *Molecules*, vol. 24, no. 5, 2019.
- [51] L. Chen, S. Wang, Y.-H. Zhang et al., “Identify key sequence features to improve CRISPR sgRNA efficacy,” *IEEE Access*, vol. 5, pp. 26582–26590, 2017.
- [52] Q. Dong, S. Wang, K. Wang, X. Liu, and B. Liu, “Identification of DNA-binding proteins by auto-cross covariance transformation,” in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 470–475, Washington, DC, USA, 2015.