

Machine learning in disease screening, diagnosis, and surveillance

Edited by

Yi-Ju Tseng and Yu-Hsiu Lin

Published in

Frontiers in Public Health



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83252-039-0
DOI 10.3389/978-2-83252-039-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Machine learning in disease screening, diagnosis, and surveillance

Topic editors

Yi-Ju Tseng — National Yang Ming Chiao Tung University, Taiwan

Yu-Hsiu Lin — National Chung Cheng University, Taiwan

Citation

Tseng, Y.-J., Lin, Y.-H., eds. (2023). *Machine learning in disease screening, diagnosis, and surveillance*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83252-039-0

Table of contents

- 05 **Deep-Learning Approach to Predict Survival Outcomes Using Wearable Actigraphy Device Among End-Stage Cancer Patients**
Tien Yun Yang, Pin-Yu Kuo, Yaoru Huang, Hsiao-Wei Lin, Shwetambara Malwade, Long-Sheng Lu, Lung-Wen Tsai, Shabbir Syed-Abdul, Chia-Wei Sun and Jeng-Fong Chiou
- 14 **Application of Machine Learning for the Prediction of Etiological Types of Classic Fever of Unknown Origin**
Yongjie Yan, Chongyuan Chen, Yunyu Liu, Zuyue Zhang, Lin Xu and Kexue Pu
- 25 **Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records**
Jingfeng Chen, Chonghui Guo, Menglin Lu and Suying Ding
- 43 **A Web-Based Prediction Model for Cancer-Specific Survival of Middle-Aged Patients With Non-metastatic Renal Cell Carcinoma: A Population-Based Study**
Jie Tang, Jinkui Wang, Xiudan Pan, Xiaozhu Liu and Binyi Zhao
- 54 **Development and Validation of a Nomogram to Predict Cancer-Specific Survival in Elderly Patients With Papillary Renal Cell Carcinoma**
Chenghao Zhanghuang, Jinkui Wang, Zhigang Yao, Li Li, Yucheng Xie, Haoyu Tang, Kun Zhang, Chengchuang Wu, Zhen Yang and Bing Yan
- 65 **Microcalcification Discrimination in Mammography Using Deep Convolutional Neural Network: Towards Rapid and Early Breast Cancer Diagnosis**
Yew Sum Leong, Khairunnisa Hasikin, Khin Wee Lai, Norita Mohd Zain and Muhammad Mokhzaini Azizan
- 78 **m6A Regulator-Mediated Methylation Modification Patterns and Characteristics in COVID-19 Patients**
Xin Qing, Qian Chen and Ke Wang
- 92 **An Explainable AI Approach for the Rapid Diagnosis of COVID-19 Using Ensemble Learning Algorithms**
Houwu Gong, Miye Wang, Hanxue Zhang, Md Fazla Elahe and Min Jin
- 104 **A Machine Learning Algorithm for Predicting the Risk of Developing to M1b Stage of Patients With Germ Cell Testicular Cancer**
Li Ding, Kun Wang, Chi Zhang, Yang Zhang, Kanlirong Wang, Wang Li and Junqi Wang
- 115 **Machine learning-assisted prediction of pneumonia based on non-invasive measures**
Clement Yaw Effah, Ruoqi Miao, Emmanuel Kwateng Drokow, Clement Agboyibor, Ruiping Qiao, Yongjun Wu, Lijun Miao and Yanbin Wang

- 130 **Gray wolf optimization-extreme learning machine approach for diabetic retinopathy detection**
Musatafa Abbas Abbood Albadr, Masri Ayob, Sabrina Tiun, Fahad Taha AL-Dhief and Mohammad Kamrul Hasan
- 146 **Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation**
Jiajin He, Jinhua Li, Siqing Jiang, Wei Cheng, Jun Jiang, Yun Xu, Jiezhe Yang, Xin Zhou, Chengliang Chai and Chao Wu
- 156 **Development and validation of a prognostic nomogram for adult patients with renal sarcoma: A retrospective study based on the SEER database**
Yongkun Zhu, Weipu Mao, Guangyuan Zhang, Si Sun, Shuchun Tao, Tiancheng Jiang, Qingbo Wang, Yuan Meng, Jianping Wu and Ming Chen
- 169 **Machine learning for identifying benign and malignant of thyroid tumors: A retrospective study of 2,423 patients**
Yuan-yuan Guo, Zhi-jie Li, Chao Du, Jun Gong, Pu Liao, Jia-xing Zhang and Cong Shao
- 180 **Predictive models based on machine learning for bone metastasis in patients with diagnosed colorectal cancer**
Tianhao Li, Honghong Huang, Shuocun Zhang, Yongdan Zhang, Haoren Jing, Tianwei Sun, Xipeng Zhang, Liangfu Lu and Mingqing Zhang
- 191 **Development and validation of chest CT-based imaging biomarkers for early stage COVID-19 screening**
Xiao-Ping Liu, Xu Yang, Miao Xiong, Xuanyu Mao, Xiaoqing Jin, Zhiqiang Li, Shuang Zhou and Hang Chang
- 203 **Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa**
Michael T. Mapundu, Chodziwadziwa W. Kabudula, Eustasius Musenge, Victor Olago and Turgay Celik
- 223 **Application of machine learning and natural language processing for predicting stroke-associated pneumonia**
Hui-Chu Tsai, Cheng-Yang Hsieh and Sheng-Feng Sung
- 234 **Three-dimensional evaluation using CBCT of the mandibular asymmetry and the compensation mechanism in a growing patient: A case report**
Monica Macri and Felice Festa
- 245 **Development of a new category system for the profile morphology of temporomandibular disorders patients based on cephalograms using cluster analysis**
Rui Zhu, Yun-Hao Zheng, Zi-Han Zhang, Pei-Di Fan, Jun Wang and Xin Xiong



Deep-Learning Approach to Predict Survival Outcomes Using Wearable Actigraphy Device Among End-Stage Cancer Patients

Tien Yun Yang^{1†}, Pin-Yu Kuo^{2†}, Yaoru Huang^{3,4,5†}, Hsiao-Wei Lin³, Shwetambara Malwade⁶, Long-Sheng Lu^{4,5,7,8}, Lung-Wen Tsai⁹, Shabbir Syed-Abdul^{6,10,11*}, Chia-Wei Sun^{2*} and Jeng-Fong Chiou^{3,4,8,12*}

OPEN ACCESS

Edited by:

Chen Lin,
National Central University, Taiwan

Reviewed by:

Chien-Chang Chen,
National Central University, Taiwan
Yu-Hsiu Lin,
National Chung Cheng
University, Taiwan

*Correspondence:

Shabbir Syed-Abdul
drshabbir@tmu.edu.tw
Chia-Wei Sun
chiaweisun@nctu.edu.tw
Jeng-Fong Chiou
solomanc@tmu.edu.tw

[†]These authors share first authorship

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 07 July 2021

Accepted: 18 November 2021

Published: 09 December 2021

Citation:

Yang TY, Kuo P-Y, Huang Y, Lin H-W,
Malwade S, Lu L-S, Tsai L-W,
Syed-Abdul S, Sun C-W and
Chiou J-F (2021) Deep-Learning
Approach to Predict Survival
Outcomes Using Wearable Actigraphy
Device Among End-Stage Cancer
Patients.
Front. Public Health 9:730150.
doi: 10.3389/fpubh.2021.730150

¹ School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan, ² Biomedical Optical Imaging Lab, Department of Photonics, College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, ³ Department of Hospice and Palliative Care, Taipei Medical University Hospital, Taipei, Taiwan, ⁴ Department of Radiation Oncology, Taipei Medical University Hospital, Taipei, Taiwan, ⁵ Graduate Institute of Biomedical Materials and Tissue Engineering, College of Biomedical Engineering, Taipei Medical University, Taipei, Taiwan, ⁶ International Center for Health Information Technology, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, ⁷ Clinical Research Center, Taipei Medical University Hospital, Taipei, Taiwan, ⁸ TMU Research Center of Cancer Translational Medicine, Taipei Medical University, Taipei, Taiwan, ⁹ Department of Medical Research, Taipei Medical University Hospital, Taipei, Taiwan, ¹⁰ Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, ¹¹ School of Gerontology and Health Management, College of Nursing, Taipei Medical University, Taipei, Taiwan, ¹² Department of Radiology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

Survival prediction is highly valued in end-of-life care clinical practice, and patient performance status evaluation stands as a predominant component in survival prognostication. While current performance status evaluation tools are limited to their subjective nature, the advent of wearable technology enables continual recordings of patients' activity and has the potential to measure performance status objectively. We hypothesize that wristband actigraphy monitoring devices can predict in-hospital death of end-stage cancer patients during the time of their hospital admissions. The objective of this study was to train and validate a long short-term memory (LSTM) deep-learning prediction model based on activity data of wearable actigraphy devices. The study recruited 60 end-stage cancer patients in a hospice care unit, with 28 deaths and 32 discharged in stable condition at the end of their hospital stay. The standard Karnofsky Performance Status score had an overall prognostic accuracy of 0.83. The LSTM prediction model based on patients' continual actigraphy monitoring had an overall prognostic accuracy of 0.83. Furthermore, the model performance improved with longer input data length up to 48 h. In conclusion, our research suggests the potential feasibility of wristband actigraphy to predict end-of-life admission outcomes in palliative care for end-stage cancer patients.

Clinical Trial Registration: The study protocol was registered on ClinicalTrials.gov (ID: NCT04883879).

Keywords: palliative care, performance status, survival prediction, prognostic accuracy, wearable technology, deep learning, long short-term memory networks, actigraphy

INTRODUCTION

Accurate survival prediction is highly valued in the clinical practice of end-of-life care. It enables better communication and preparation for impending death, helps avoid futile medical treatment, and facilitates optimal palliative care quality for patients, families, and physicians altogether (1–3). Several validated prognostic tools are available, including Palliative Prognostic Score (PaP) (4–6), Palliative Prognostic Index (PPI) (7–9), Prognosis in Palliative care study (PiPS) score (10, 11), and Glasgow Prognostic Score (12, 13). These scoring systems employ a combination of subjective clinical parameters and/or objective biomarkers to generate survival predictions. Among the parameters used by these prognostic tools, the evaluation of patient performance status (PS) stands as a predominant component. Commonly used PS assessment tools include Karnofsky Performance Status (KPS) (14), Eastern Cooperative Oncology Group (ECOG) Performance Status (15), and Palliative Performance Scale (PPS) (16). However, applications of these evaluation tools are subjective in nature and require trained healthcare professionals for assessments. These characteristics inevitably lead to issues including intraobserver or interobserver variability (17, 18), overestimating or underestimating (19), discontinuous evaluations of activity status, as well as inconvenient implementation in contexts without healthcare professionals.

With the advent of wearable activity monitoring technology, we are now granted convenient and objective methods for the evaluation of patient functional status. Wearable monitors also enable constant documentation of a patient's activity status which could be retrospectively examined and validated. Because of these benefits, monitoring technologies have been applied in different research areas and yielded valuable information on the relationship between activity status and diseases in clinical fields of gynecology (20), surgery (21), pulmonary (22), nephrology (23), and psychology (24). In addition, a study by Gresham et al. also applied objective PS evaluation in a group of advanced cancer patients, which identified correlations between objective activity data of patients and clinical outcomes of adverse events, hospitalization, and overall survival (25). However, no previous study had employed objective PS data for survival prognostication.

In this study, wearable actigraphy devices were applied in a group of end-stage cancer patients for objective measurement of their activity status. We hypothesized that the objective activity data recorded by the wearable devices contained information to help predict in-hospital death of end-stage cancer patients on their hospital admissions. A deep-learning-based prediction model was developed to analyze activity data and suggest survival outcomes of patients. Furthermore, the prognostic accuracy of the proposed activity monitoring and survival prediction model was compared to a current PS evaluation tool, KPS, and a complex prognostic tool, PPI. Finally, we explored and described the applicability, potential, and limitations of the objective activity data recorded by wearable devices as a simple prognostic parameter in clinical settings.

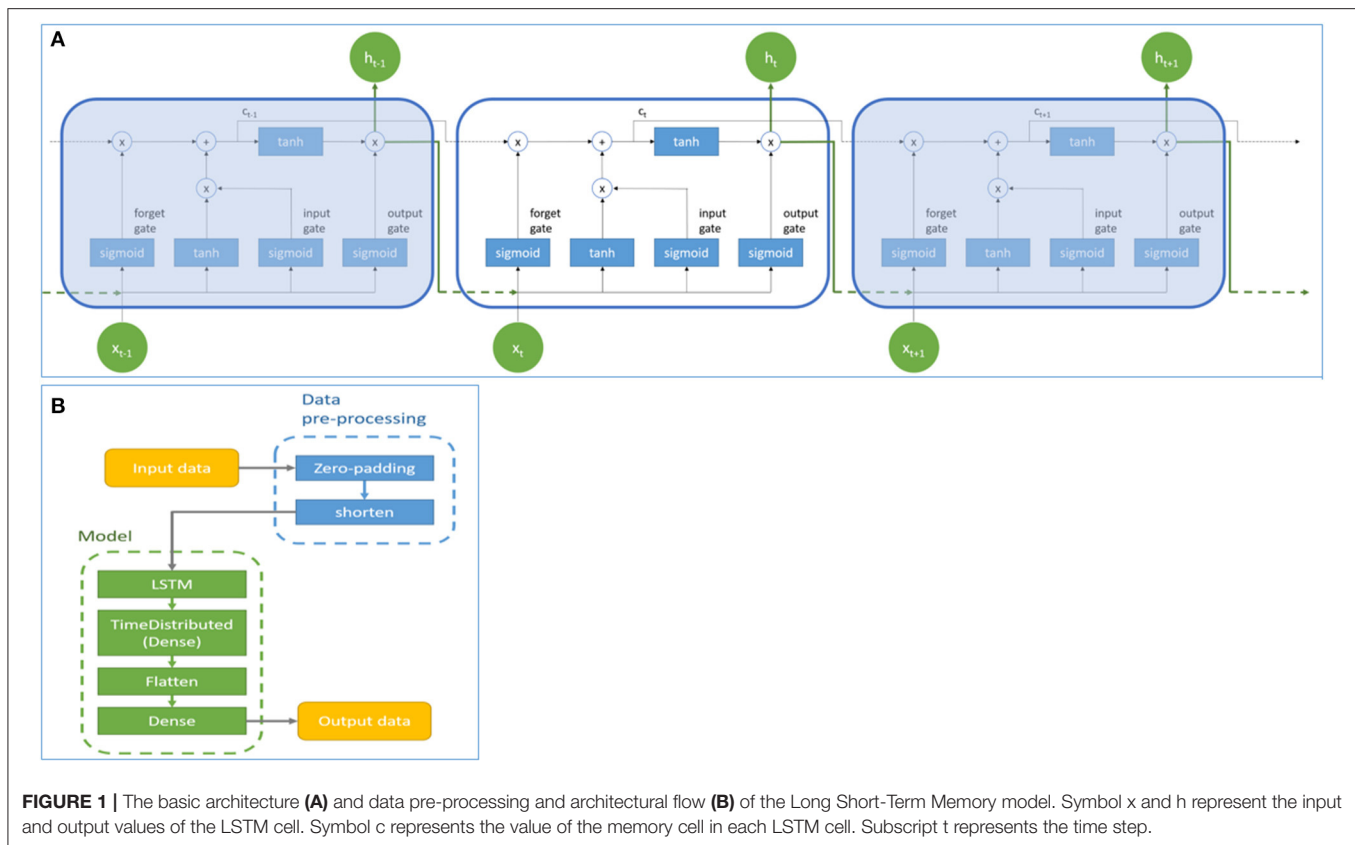
MATERIALS AND METHODS

Study Setting, Participants, and Procedures

The study was conducted in the hospice care unit of Taipei Medical University Hospital (TMUH) from December 2019 to December 2020. Patients with terminal illnesses were admitted to the unit for palliative care and management of pain and other symptoms. Participants aged > 20 years who had at least one diagnosis of end-stage solid tumor diseases and consented to receive hospice care were recruited. Patients with diagnoses of leukemia or carcinoma of unknown primary, patients with evident signs of approaching death upon admission, patients with no vital signs upon admission, or patients who continued to receive aggressive treatment were excluded from this study. After admission to the hospice care unit, patients and their caregivers were first visited and assessed by registered hospice specialist doctors and nurses. If the patient met the criteria mentioned above, they would be invited to participate in the study. Participants would only be recruited once the informed consent was signed by themselves or their legally authorized representative. The study was approved by the ethical committee of the Taipei Medical University-Joint Institutional Review Board (TMU-JIRB No. N201910041).

Clinical data including age, gender, diagnosis, and comorbidities were collected after successful recruitment. Patients were asked to wear a wristband actigraphy device on their hands without intravenous lines. The wearable actigraphy devices (model no. XB40ACT, K&Y lab, Taipei, Taiwan) used in this study is a tiny gadget that weighs 7 g with dimensions of 44*19*8 mm and has been previously validated (26) and applied in a sleep quality study among cancer patients (27). The monitor collects three-dimensional data of gravitational acceleration, angular change, and spin change of the patient's hand motion every second and transforms them into three statistical parameters: physical activity, angle, and spin. Participants were instructed to wear the devices throughout their hospital stay except showering time because they were not water-resistant. The information was also forwarded to their caregivers.

Subsequently, subjective PS assessments using the KPS and prognostic evaluations using the PPI were done by two trained specialists. The KPS system is an established tool designed for PS evaluation. The score collaboratively takes ambulation, activity, evidence of disease, self-care, the requirement of assistance, and progression of disease into consideration with a scale that ranges from normal activity (100) to death (0) (14). In addition to PS assessments, we applied a complex prognostic tool based on evaluations of PS and other clinical symptoms, namely PPI, starting from July 16, 2020. We were only able to conduct the PPI assessments due to the participation of an additional specialist, who undertook extra work derived from evaluations of patients' clinical symptoms. PPI considers PS and clinical symptoms of oral intake, edema, dyspnea at rest, and delirium, to generate an overall prognostication. According to the original study, the results range from 0 to 15, and a PPI > 6.0 estimates a survival time of fewer than 3 weeks (7). The same group of specialists conducted all KPS and PPI assessments to ensure



interpersonal consistency. After the initial consultation, patient activity data recorded by the actigraphy devices would be synced and uploaded every 2–3 days until the patient was discharged from the hospital. Survival outcomes were documented as either death or discharged in stable condition at the end of each patient's hospital stay.

Data Pre-processing and LSTM-Based Deep Learning Model

The data collected by the actigraphy device is a time series with three features: physical activity, angle, and spin. The issue of variations in each patient's data length was managed by zero paddings until the maximum length of the time series was reached. To avoid vanishing gradients in the deep learning model, we opted for an average value of 20 timesteps and shortened the time series to <500 timesteps.

In this study, we trained a long short-term memory (LSTM)-based deep learning model to predict the clinical status of patients at discharge, which was either death or discharged in stable condition. Recurrent neural networks (RNN) is a deep learning method well-suited to deal with time series structure (28, 29). However, the vanishing gradient problem of RNN made the tool suboptimal for long time-series data (30) for which the LSTM, a particular type of RNN, was used to resolve the issue. Compared to RNN, the LSTM architecture is more resistant to vanishing gradients and allows robust processing of long time-series data (31, 32). The performance of LSTM has been validated

in disciplines of economic, financial, stock market forecasting, and even stress forecasting using survey data and physiology parameters. In these studies, LSTM demonstrated lower error rates (33), lower variance (34), and higher accuracy (34, 35) than other analytical methods. A study by Umematsu et al. also showed that LSTM could generate satisfactory results based on objective data measured by wearable devices and phones (35).

Figure 1A showed the basic architecture of the LSTM model. Symbol x and h represent the input value and the output value of the LSTM cell, respectively. The value in the memory cell in each LSTM cell is c . The subscripts of x , h , and c represented different time points. Each LSTM cell contains an input gate, forget gate, and output gate. The input gate determines whether the neuron writes input values into the memory cell. The forget gate determines whether the memory cell formats memory values. The output gate determines whether the neuron reads the values in the memory cell. The hyperbolic tangent function (\tanh) and sigmoid function (σ) are activating functions in LSTM. In this study, the prediction model was based on the LSTM cell to process the three-dimensional time-series data. Data pre-processing and model architecture flows were presented in Figure 1B. The model consisted of an LSTM layer, a dense layer wrapped with TimeDistributed, a flatten layer, and a dense layer. Parameters were adjusted according to different model structures and are presented in the results section. It should be noted that the model was designed to generate survival predictions based solely on activity data of patients, therefore, demographic and

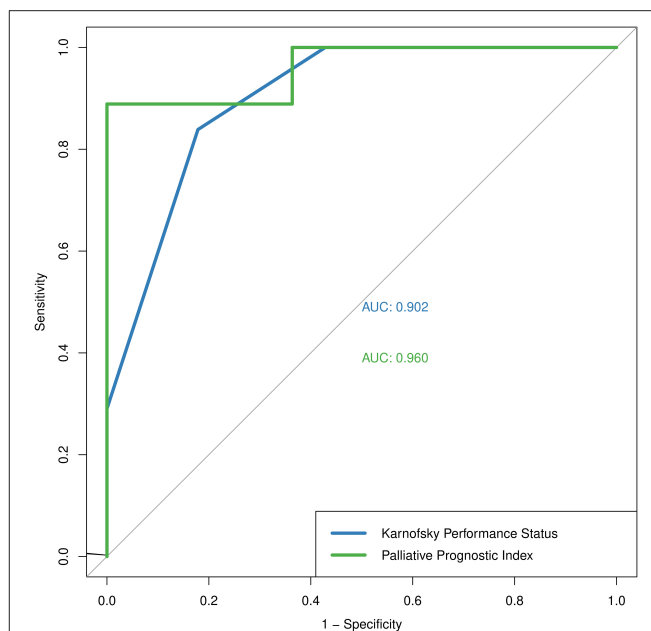
TABLE 1 | Patient demographics and characteristics at baseline visit.

Characteristics (N = 60)	Value	
Age, years		
• Mean	72.9	
• SD	12.2	
• Range	45–94	
Sex, N (%)		
• Male	37 (61.67%)	
• Female	23 (38.33%)	
Primary tumor site, N (%)		
• Gastrointestinal system	26 (43.33%)	
• Lung	12 (20.00%)	
• Genitourinary system	10 (16.67%)	
• Gynecological system	5 (8.33%)	
• Breast	3 (5%)	
• Head and neck	2 (3.33%)	
• Central nervous system	2 (3.33%)	
Patients with comorbidities, N (%)	46 (76.67%)	
Length of hospital stay, days		
• Median (IQR)	10 (5–15)	
Patient status at discharge		
• Death	28 (46.67%)	
• In stable condition	32 (53.33%)	
KPS (N = 59)		
	Death	Discharged in stable condition
• KPS < 50%	23 (38.98%)	5 (8.47%)
• KPS ≥ 50%	5 (8.47%)	26 (44.07%)
PPI (N = 20)		
	Death	Discharged in stable condition
• PPI > 6.0	8 (40.00%)	0 (0.00%)
• PPI ≤ 6.0	1 (5.00%)	11(55.00%)

clinical data of patients (such as comorbidities) were not utilized by the model.

Statistical Analysis

Patient characteristics were summarized using descriptive statistics. The clinical outcomes of participants were determined at the end of their hospital stay as binary results: death (1) or discharged in stable condition (0). We adopted a validated cutoff value of 50% for KPS (36) and a cutoff value of 6.0 for PPI as suggested by the original study (7). A receiver operating characteristic (ROC) curve analysis was also conducted to identify optimal cutoff values based on our dataset. The predictive accuracy of KPS and PPI were presented as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), overall accuracy, and the area under the receiver operating characteristic (ROC) curve (AUC). Additionally, an exploratory analysis was conducted to investigate the predictive correlation between KPS and the LSTM model. Correlation between the two variables was calculated using the Pearson correlation coefficient. Statistical analyses were computed using Python version 3.6 and R software version 4.0.2.

**FIGURE 2** | The Receiver Operating Characteristic curve of Karnofsky Performance Status (blue) and Palliative Prognostic Index (green).

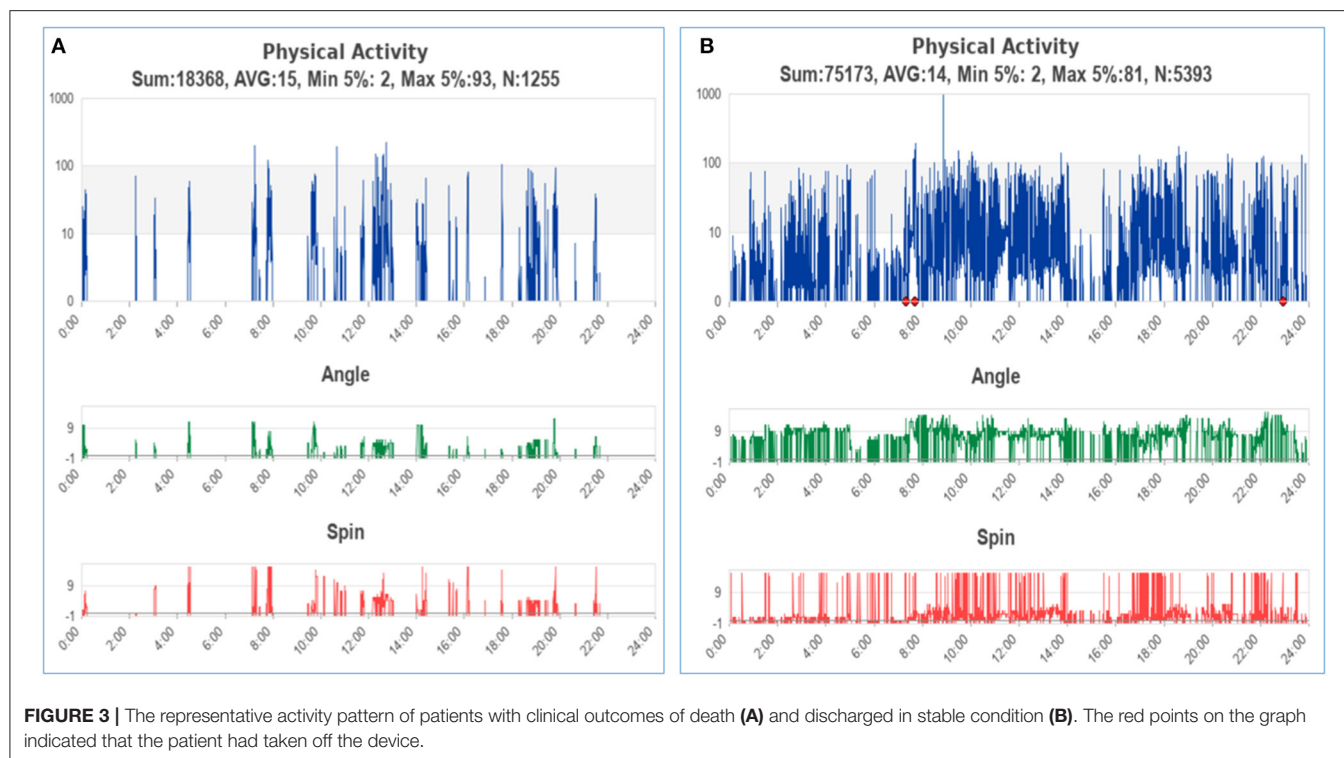
RESULTS

Demographics of Study Population

From December 11, 2019, to December 10, 2020, 60 patients admitted to the hospice care unit of TMUH were eligible for study recruitment and consented to participate. Patient characteristics, information on KPS and PPI, and their clinical outcomes at discharge were presented in **Table 1**. The mean age was 72.9 years old (SD 12.2), and 62% were male. Gastrointestinal tumors were the most common malignancies, followed by lung, genitourinary, gynecological, breast, head and neck, and CNS cancers. Seventy seven percent of participants had one or more comorbidities, consisting of hypertension, diabetes mellitus, hyperlipidemia, coronary artery diseases, cerebral infarctions, and others. The median length of hospital stay of patients was 10 (IQR 5–15) days. Twenty eight (47%) patients died at the end of their hospice care stay, whereas 32 (53%) patients were discharged from the hospice care unit in stable condition. It should be noted that one case was discharged against medical advice and deemed as discharged in stable condition. KPS assessments were available or 59 participants, with 28 of them having a KPS score < 50% at admission. PPI assessments were available for 20 participants, with 8 of them having a PPI score > 6.0 on admission.

Prognostic Accuracy of KPS and PPI

The absolute numbers of the true positive, false positive, false negative, and true negative of KPS and PPI assessments are presented in **Table 1**. True positive was defined as participants with KPS < 50% or PPI > 6.0 at baseline visit and death at the end of their hospital stay. The predictive performance of KPS score based on binary outcomes had an overall predictive accuracy of



83.1% (95% CI 71.0–91.6%), sensitivity of 82.1% (95% CI 63.1–93.9%), specificity of 83.9% (95% CI 66.3–94.5%), PPV of 82.1% (95% CI 63.1–93.9%), NPV of 83.9% (95% CI 66.3–94.5%), and AUC of 0.902. The predictive performance of PPI score based on binary outcomes had an overall predictive accuracy of 95.0% (95% CI 75.1–99.9%), sensitivity of 88.9% (95% CI 51.8–99.7%), specificity of 100% (95% CI 71.5–100%), PPV of 100% (95% CI 63.1–100%), NPV of 91.7 (95% CI 61.5–99.8%), and AUC of 0.960. The discrimination thresholds identified by the ROC curve analysis correlated with the cutoff values we initially adopted for both KPS and PPI (Figure 2).

Activity Dataset Description and Splitting

The representative activity pattern recorded by the wearable wristband was shown in Figure 3. Figure 3A belonged to a participant who died at the end of the hospital stay, while Figure 3B belonged to a participant who was discharged in stable condition. Although activity data of patients were recorded throughout their hospital stay, the LSTM-based prediction model only employed data of the initial 48 h for prognostic applicability in clinical settings. After excluding recordings with tracking interruption or data volume of fewer than 48 h, the final dataset included activity data of 44 participants, with 21 deaths and 23 discharged in stable condition at the end of hospital stay, respectively. The maximum length of data after zero-padding is 9,640.

All data was fed into the model after data pre-processing; thus, sampling rates and strides were not defined. We first conducted a preliminary analysis to investigate the feasibility and performance of the model. In the preliminary analysis, the data

TABLE 2 | Details of the dataset for the preliminary and final LSTM models.

	Training dataset	Validation dataset	Testing dataset	Total
Preliminary model				
Discharged in stable condition	15 (34.09%)	-	8 (18.18%)	23
Death	15 (34.09%)	-	6 (13.64%)	21
Total	30	-	14	44
Final model				
Discharged in stable condition	16 (36.36%)	4 (9.09%)	3 (6.82%)	23
Death	14 (31.82%)	4 (9.09%)	3 (6.82%)	21
Total	30	8	6	44

were divided into a training dataset and a testing dataset at a ratio of 7:3. The number of LSTM units for the preliminary model is 64, with a batch size of 8. We further divided data into training, validation, and testing datasets at a ratio of 7:2:1 in the final LSTM model to detect the possibility of overfitting. The number of LSTM units for the final model was 256, with a batch size of 16. The epochs of the preliminary and final model were 50 and 100, respectively. Both models adopted adam as the optimizer and the mean absolute error was used as the loss function. Dataset of the preliminary and final models are presented in Table 2.

Training of LSTM Survival Prediction Model

Based on the activity data recorded in the initial 48 h after admission, the preliminary model yielded an accuracy of 0.8667

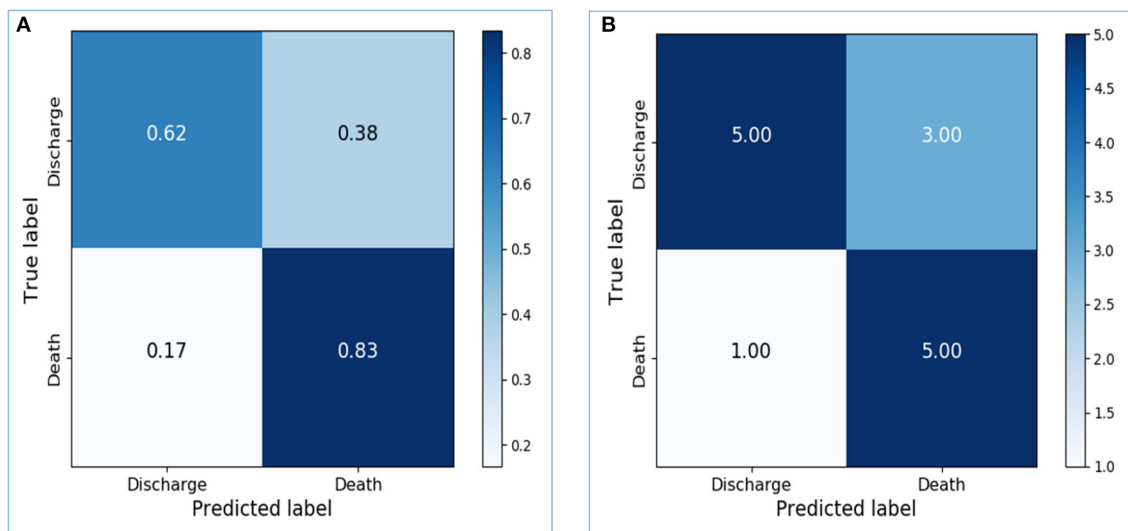


FIGURE 4 | Confusion matrices of the preliminary prediction model. **(A):** Confusion matrix of the testing dataset, with normalization. **(B):** Confusion matrix of the testing dataset, without normalization.

in the training dataset and 0.7143 in the testing dataset. The confusion matrix visualized the differences between model prediction and the ground truth. The variables used for the original and normalized confusion matrices were the same. In the normalized form of confusion matrices, the sum of each row is 1.0 and represents the correct prediction in terms of probability. **Figures 4A,B** illustrate the confusion matrices with normalization and without normalization, respectively. The sensitivity, specificity, PPV, NPV, and AUC of the model on the testing dataset were 0.8333, 0.625, 0.625, 0.8333, and 0.7292, respectively. These satisfactory results indicated the feasibility of LSTM in classifying time series data collected by wearable actigraphy devices without any physiological information.

The dataset was further sliced into training, validation, and testing data in the final model with appropriate parameters. The training accuracy increased to 0.9667, and the validation accuracy and testing accuracy were 0.75 and 0.8333, respectively. After increasing the LSTM units from 64 to 256, the performance of the model on the testing dataset was greatly improved. Confusion matrices of the final model were shown in **Figures 5A,B**. The sensitivity, specificity, PPV, NPV, and AUC of the model on the testing dataset were 1.0, 0.6667, 0.75, 1.0, and 0.8333, respectively.

The Impact of Data Length on LSTM Model Performance

Since activity data of the initial 48 h yielded favorable results, we further explored the performance of the model based on a shorter time series. The input data of the preliminary model and the final model were reduced from 48 to 24 h with the same parameters. The maximum length of data after zero-padding is 6,460. After reducing the time interval, the prognostic accuracy of both preliminary and final models decreased. The comparison of model performance based on 48 and 24 h is demonstrated in

Table 3. The finding indicated decreasing classification accuracy of the models with reducing time length of the input data.

DISCUSSION

The study proposed and examined the use of a wearable actigraphy device for survival prediction among end-stage cancer patients. Compared to the subjective PS evaluation by KPS, our results indicated that objective activity data recorded by the wearable devices also provided favorable prognostic accuracy when employing the LSTM model. The wearable actigraphy device employed in this study is a lightweight and low-cost device, and based on the results, provides convenient activity data for survival prediction in end-stage cancer patients. The findings of this study suggest implementing the wearable technology and the survival prediction model in end-of-life care to facilitate decision-making for clinicians and better preparation for patients and their families.

PS evaluation can inform patients' clinical condition and treatment decisions in end-of-life care. However, subjective evaluation tools like KPS are seldomly used as a single predictor for patient survival; one of the reasons is the potential risk of measurement bias due to their subjective nature (37). As a result, studies examining the applicability of objective activity evaluation, such as measurements by wearable technology, are being conducted to investigate the usability of activity data for survival prediction. While several studies have identified associations between activity data of cancer patients and their clinical outcomes, such as unplanned healthcare encounters (38), adverse events, hospitalizations, and survival (25), no previous studies have utilized the activity data to build a prediction model that suggests survival outcomes. To our knowledge, this is the first study that applied objective activity data of patients in a

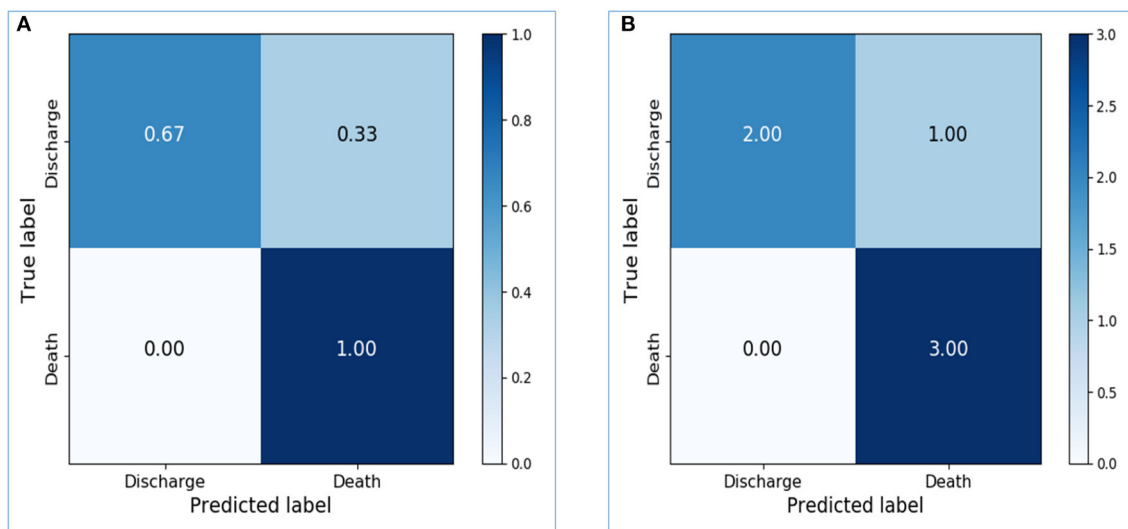


FIGURE 5 | Confusion matrices of the final prediction model. **(A)** Confusion matrix of the testing dataset, with normalization. **(B)** Confusion matrix of the testing dataset, without normalization.

TABLE 3 | Model performance with different input data lengths.

Model	Training ACC ^a	Validation ACC	Testing ACC	Sensitivity	Specificity	PPV ^b	NPV ^c	AUC ^d
Preliminary model 48 h	0.8667	N/A	0.7143	0.8333	0.625	0.625	0.8333	0.7292
Preliminary model 24 h	0.8333	N/A	0.6429	0.6667	0.625	0.5714	0.7143	0.6458
Final model 48 h	0.9667	0.75	0.8333	1.0	0.6667	0.75	1.0	0.8333
Final model 24 h	0.9333	0.625	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667

^aACC: accuracy.

^bPPV: positive predictive value.

^cNPV: negative predictive value.

^dAUC: area under the receiver operating characteristic curve.

deep-learning model to provide survival outcome predictions in the end-stage cancer population.

In this study, while KPS had comparable performance, PPI yielded a nearly impeccable result regarding prognostic accuracy. However, it should be noted that the accuracy of these prognostic tools, either KPS or PPI, relies heavily on the judgment of an experienced clinical practitioner. In comparison, the activity monitoring and survival prediction model proposed by this study, requires no clinical expertise but a wearable wristband. The advantage introduces two clinical implications: first, automatically-generated survival predictions can lessen healthcare practitioners' workload in clinical settings, and second, enable end-of-life care at places outside hospitals, such as hospice at home. The result also suggested that integrating activity evaluation and clinical parameters in a survival prediction model might facilitate better prognostic accuracy, and subsequent analysis should be conducted to investigate the feasibility of such a combination.

The activity data of only the initial 24 and 48 h since patients' hospital admission was employed to provide timely survival prediction and enable practicable use in the clinical settings. Activity recordings fewer than 24 h were not analyzed due to the

consideration of circadian rhythm (39). Circadian rhythms are part of the body's internal clock and are approximately 24 h a cycle. However, studies have shown that circadian rhythms can be disrupted by multiple factors, including the states of cancerous diseases (40); thus, we employed activity analysis of both 24 and 48 h to include at least a cycle of the circadian rhythm. Our findings showed that the predictions based on activity data of 48 h yielded better prognostic accuracy than 24 h in both preliminary and final models. While the better performance of the model may be attributed to the increasing length of data (41), the inclusion of at least a cycle of circadian rhythms might also serve as a constructive factor. Future studies examining the impact of circadian rhythm on activity data of end-stage cancer patients are thus warranted.

Though the study offers promising results of the deep-learning-based survival prediction model, the study still encompasses a few limitations. First, the issue of data discontinuity was noticeable. Probable causes include battery charging requirements and the non-waterproof characteristics of the device, as these monitors were removed during the showering time. Although the issue of data discontinuity and different data lengths were handled by data pre-processing,

future studies with better activity tracking devices and data quality are warranted. Second, the study was designed to provide patients' outcomes at the end of their hospital stay, either death or discharged in stable condition. Even though the survival time varied among participants regardless of their final survival outcomes, the proposed model only informed binary survival outcomes rather than the estimated survival time. Finally, we failed to adopt PPI assessments at the beginning of the study and thus, only applied the tool to the last 20 participants.

In conclusion, the study presented a wearable activity monitoring and survival prediction model for end-stage cancer patients in hospice care settings. Our survival prediction model provided satisfactory prognostic accuracy of patients' binary survival outcomes, death or discharged in stable condition, by using activity data of the initial 24 or 48 h on their hospital admission. The prognostic accuracy of the model was time-dependent, with models using activity data of 48 h yielding better results than those of 24 h. The automatically-generated survival prediction by the LSTM deep-learning model demonstrated feasibility in clinical settings and may benefit end-of-life care in settings without healthcare professionals.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Taipei Medical University-Joint Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

REFERENCES

- Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, McIntyre L, Tulsky JA. Factors considered important at the end of life by patients, family, physicians, and other care providers. *JAMA*. (2000) 284:2476–82. doi: 10.1001/jama.284.19.2476
- Steinhauser KE, Christakis NA, Clipp EC, McNeilly M, Grambow S, Parker J, et al. Preparing for the end of life: preferences of patients, families, physicians, and other care providers. *J Pain Symptom Manage*. (2001) 22:727–37. doi: 10.1016/S0885-3924(01)00334-7
- Kirk P, Kirk I, Kristjanson LJ. What do patients receiving palliative care for cancer and their families want to be told? A Canadian and Australian qualitative study. *BMJ*. (2004) 328:1343. doi: 10.1136/bmj.38103.423576.55
- Pirovano M, Maltoni M, Nanni O, Marinari M, Indelli M, Zaninetta G, et al. A new palliative prognostic score: a first step for the staging of terminally ill cancer patients. Italian multicenter and study group on palliative care. *J Pain Symptom Manage*. (1999) 17:231–9. doi: 10.1016/S0885-3924(98)00145-6
- Glare P, Virik K. Independent prospective validation of the PaP score in terminally ill patients referred to a hospital-based palliative medicine consultation service. *J Pain Symptom Manage*. (2001) 22:891–8. doi: 10.1016/S0885-3924(01)00341-4
- Tarumi Y, Watanabe SM, Lau F, Yang J, Quan H, Sawchuk L, et al. Evaluation of the palliative prognostic score (PaP) and routinely collected clinical

AUTHOR CONTRIBUTIONS

TY, YH, SM, L-SL, and SS-A conceived of the presented idea. YH, H-WL, L-SL, and J-FC were in charge of clinical evaluations and data collection. TY and SM performed the statistical analysis of clinical parameters. P-YK and L-WT performed data-processing and deep-learning prediction model building. L-SL and C-WS validated the analytical methods. SS-A, C-WS, and J-FC co-supervised the project. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was supported in part by Ministry of Science and Technology, Taiwan [Grant Numbers 108-2221-E-038-013, 110-2923-E-038-001-MY3, 110-5420-003-300, 110-2320-B-038-056, 109-2221-E-009-018-MY3, 109-2314-B-038-122, 109-2314-B-038-141, 109-2635-B-038-001, and 109-2314-B-038-072], Taipei Medical University, Taiwan [Grant Numbers 108-3805-009-110 and 109-3800-020-400], Ministry of Education, Taiwan [Grant Number 108-6604-002-400], and Wanfang hospital, Taiwan [Grant Number 106TMU-WFH-01-4].

ACKNOWLEDGMENTS

The authors thank the Ministry of Science and Technology and Ministry of Education of Taiwan, Taipei Medical University, and Wanfang hospital for financially supporting the project. The authors also thank TMU Research Center of Cancer Translational Medicine from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan. Finally, the authors thank the hospice care unit of Taipei Medical University Hospital for supportive cooperation with the research team.

data in prognostication of survival for patients referred to a palliative care consultation service in an acute care hospital. *J Pain Symptom Manage*. (2011) 42:419–31. doi: 10.1016/j.jpainsymman.2010.12.013

- Morita T, Tsunoda J, Inoue S, Chihara S. The palliative prognostic index: a scoring system for survival prediction of terminally ill cancer patients. *Support Care Cancer*. (1999) 7:128–33. doi: 10.1007/s005200050242
- Stone CA, Tiernan E, Dooley BA. Prospective validation of the palliative prognostic index in patients with cancer. *J Pain Symptom Manage*. (2008) 35:617–22. doi: 10.1016/j.jpainsymman.2007.07.006
- Kao CY, Hung YS, Wang HM, Chen JS, Chin TL, Lu CY, et al. Combination of initial palliative prognostic index and score change provides a better prognostic value for terminally ill cancer patients: a six-year observational cohort study. *J Pain Symptom Manage*. (2014) 48:804–14. doi: 10.1016/j.jpainsymman.2013.12.246
- Gwilliam B, Keeley V, Todd C, Gittins M, Roberts C, Kelly L, et al. Development of prognosis in palliative care study (PiPS) predictor models to improve prognostication in advanced cancer: prospective cohort study. *BMJ Support Palliat Care*. (2012) 2:63–71. doi: 10.1136/bmjspcare.2012.d4920rep
- Kim ES, Lee JK, Kim MH, Noh HM, Jin YH. Validation of the prognosis in palliative care study predictor models in terminal cancer patients. *Korean J Fam Med*. (2014) 35:283–94. doi: 10.4082/kjfm.2014.35.6.283
- Forrest LM, McMillan DC, McArdle CS, Angerson WJ, Dunlop DJ. Evaluation of cumulative prognostic scores based on the systemic inflammatory response

- in patients with inoperable non-small-cell lung cancer. *Br J Cancer*. (2003) 89:1028–30. doi: 10.1038/sj.bjc.6601242
13. Laird BJ, Kaasa S, McMillan DC, Fallon MT, Hjermstad MJ, Fayes P, et al. Prognostic factors in patients with advanced cancer: a comparison of clinicopathological factors and the development of an inflammation-based prognostic system. *Clin Cancer Res*. (2013) 19:5456–64. doi: 10.1158/1078-0432.CCR-13-1066
 14. Karnofsky DA, Burchenal JH. The clinical evaluation of chemotherapeutic agents in cancer. In: MC M, editor. *Evaluation of Chemotherapeutic Agents*. New York, NY: Columbia University Press (1949). p. 196.
 15. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the eastern cooperative oncology group. *Am J Clin Oncol*. (1982) 5:649–55. doi: 10.1097/00000421-198212000-00014
 16. Anderson F, Downing GM, Hill J, Casorso L, Lerch N. Palliative performance scale (PPS): a new tool. *J Palliat Care*. (1996) 12:5–11. doi: 10.1177/082585979601200102
 17. Blagden SP, Charman SC, Sharples LD, Magee LR, Gilligan D. Performance status score: do patients and their oncologists agree? *Br J Cancer*. (2003) 89:1022–7. doi: 10.1038/sj.bjc.6601231
 18. Kelly CM, Shahrokni A. Moving beyond karnofsky and ECOG performance status assessments with new technologies. *J Oncol*. (2016) 2016:6186543. doi: 10.1155/2016/6186543
 19. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ*. (2000) 320:469–72. doi: 10.1136/bmj.320.7233.469
 20. Michael JC, El Nokali NE, Black JJ, Rofey DL. Mood and ambulatory monitoring of physical activity patterns in youth with polycystic ovary syndrome. *J Pediatr Adolesc Gynecol*. (2015) 28:369–72. doi: 10.1016/j.jpog.2014.10.010
 21. Langenberg S, Schulze M, Bartsch M, Gruner-Labitzke K, Pek C, Köhler H, et al. Physical activity is unrelated to cognitive performance in pre-bariatric surgery patients. *J Psychosom Res*. (2015) 79:165–70. doi: 10.1016/j.jpsychores.2015.03.008
 22. Kawagoshi A, Kiyokawa N, Sugawara K, Takahashi H, Sakata S, Miura S, et al. Quantitative assessment of walking time and postural change in patients with COPD using a new triaxial accelerometer system. *Int J Chron Obstruct Pulmon Dis*. (2013) 8:397–404. doi: 10.2147/COPD.S49491
 23. Carvalho EV, Reboredo MM, Gomes EP, Teixeira DR, Roberti NC, Mendes JO, et al. Physical activity in daily life assessed by an accelerometer in kidney transplant recipients and hemodialysis patients. *Transplant Proc*. (2014) 46:1713–7. doi: 10.1016/j.transproceed.2014.05.019
 24. Wielopolski J, Reich K, Clepce M, Fischer M, Sperling W, Kornhuber J, et al. Physical activity and energy expenditure during depressive episodes of major depression. *J Affect Disord*. (2015) 174:310–6. doi: 10.1016/j.jad.2014.11.060
 25. Gresham G, Hendifar AE, Spiegel B, Neeman E, Tuli R, Rimel BJ, et al. Wearable activity monitors to assess performance status and predict clinical outcomes in advanced cancer patients. *NPJ Digit Med*. (2018) 1:27. doi: 10.1038/s41746-018-0032-6
 26. Kuo TBJ, Li JY, Chen CY, Lin YC, Tsai MW, Lin SP, et al. Influence of accelerometer placement and/or heart rate on energy expenditure prediction during uphill exercise. *J Mot Behav*. (2018) 50:127–33. doi: 10.1080/00222895.2017.1306481
 27. Barsasella D, Syed-Abdul S, Malwade S, Kuo TBJ, Chien M-J, Núñez-Benjumea FJ, et al. Sleep quality among breast and prostate cancer patients: a Comparison between subjective and objective measurements. *Healthcare*. (2021) 9:785. doi: 10.3390/healthcare9070785
 28. Kolen JF, Kremer SC. *A Field Guide to Dynamical Recurrent Networks*. New York, NY: John Wiley & Sons (2001).
 29. Lingras P, Sharma S, Zhong M. Prediction of recreational travel using genetically designed regression and time-delay neural network models. *Transportation Res Record*. (2002) 1805:16–24. doi: 10.3141/1805-03
 30. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*. (1994) 5:157–66. doi: 10.1109/72.279181
 31. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735
 32. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput*. (2000) 12:2451–71. doi: 10.1162/089976600300015015
 33. Siarni-Namini S, Tavakoli N, Namin AS, editors. A comparison of ARIMA and LSTM in forecasting time series. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Orlando, FL (2018). doi: 10.1109/ICMLA.2018.00227
 34. Karmiani D, Kazi R, Nambisan A, Shah A, Kamble V, editors. Comparison of predictive algorithms: backpropagation, SVM, LSTM and kalman filter for stock market. In: *2019 Amity International Conference on Artificial Intelligence (AICAI)*. Dubai (2019). doi: 10.1109/AICAI.2019.8701258
 35. Umematsu T, Sano A, Taylor S, Picard RW, editors. Improving students' daily life stress forecasting using LSTM neural networks. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. Chicago, IL (2019). doi: 10.1109/BHI.2019.8834624
 36. Hwang SS, Scott CB, Chang VT, Cogswell J, Srinivas S, Kasimis B. Prediction of survival for advanced cancer patients by recursive partitioning analysis: role of karnofsky performance status, quality of life, and symptom distress. *Cancer Invest*. (2004) 22:678–87. doi: 10.1081/CNV-200032911
 37. Jang RW, Caraiscos VB, Swami N, Banerjee S, Mak E, Kaya E, et al. Simple prognostic model for patients with advanced cancer based on performance status. *J Oncol Pract*. (2014) 10:e335–41. doi: 10.1200/JOP.2014.001457
 38. Nilanon T, Nocera LP, Martin AS, Kolatkar A, May M, Hasnain Z, et al. Use of wearable activity tracker in patients with cancer undergoing chemotherapy: toward evaluating risk of unplanned health care encounters. *JCO Clin Cancer Inform*. (2020) 4:839–53. doi: 10.1200/CCI.20.00023
 39. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. (2003) 26:342–92. doi: 10.1093/sleep/26.3.342
 40. Sulli G, Lam MTY, Panda S. Interplay between circadian clock and cancer: new frontiers for cancer treatment. *Trends Cancer*. (2019) 5:475–94. doi: 10.1016/j.trecan.2019.07.002
 41. Ergen T, Kozat SS. Online training of LSTM networks in distributed systems for variable length data sequences. *IEEE Trans Neural Netw Learn Syst*. (2018) 29:5159–65. doi: 10.1109/TNNLS.2017.2770179

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Yang, Kuo, Huang, Lin, Malwade, Lu, Tsai, Syed-Abdul, Sun and Chiou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Application of Machine Learning for the Prediction of Etiological Types of Classic Fever of Unknown Origin

Yongjie Yan¹, Chongyuan Chen², Yunyu Liu³, Zuyue Zhang¹, Lin Xu¹ and Kexue Pu^{1*}

¹ School of Medical Informatics, Chongqing Medical University, Chongqing, China, ² Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, Chongqing, China, ³ Medical Records and Statistics Office, The Second Affiliated Hospital of Chongqing Medical University, Chongqing, China

OPEN ACCESS

Edited by:

Yi-Ju Tseng,
National Central University, Taiwan

Reviewed by:

Chakrapani M,
Kasturba Medical College,
Mangalore, India
Abdolrazagh Hashemi Shahraki,
University of Florida, United States
Farzaneh Dastan,
Shahid Beheshti University of Medical
Sciences, Iran

*Correspondence:

Kexue Pu
pukexue@cqmu.edu.cn

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 23 October 2021

Accepted: 08 December 2021

Published: 24 December 2021

Citation:

Yan Y, Chen C, Liu Y, Zhang Z, Xu L
and Pu K (2021) Application of
Machine Learning for the Prediction of
Etiological Types of Classic Fever of
Unknown Origin.
Front. Public Health 9:800549.
doi: 10.3389/fpubh.2021.800549

Background: The etiology of fever of unknown origin (FUO) is complex and remains a major challenge for clinicians. This study aims to investigate the distribution of the etiology of classic FUO and the differences in clinical indicators in patients with different etiologies of classic FUO and to establish a machine learning (ML) model based on clinical data.

Methods: The clinical data and final diagnosis results of 527 patients with classic FUO admitted to 7 medical institutions in Chongqing from January 2012 to August 2021 and who met the classic FUO diagnostic criteria were collected. Three hundred seventy-three patients with final diagnosis were divided into 4 groups according to 4 different etiological types of classical FUO, and statistical analysis was carried out to screen out the indicators with statistical differences under different etiological types. On the basis of these indicators, five kinds of ML models, i.e., random forest (RF), support vector machine (SVM), Light Gradient Boosting Machine (LightGBM), artificial neural network (ANN), and naive Bayes (NB) models, were used to evaluate all datasets using 5-fold cross-validation, and the performance of the models were evaluated using micro-F1 scores.

Results: The 373 patients were divided into the infectious disease group ($n = 277$), non-infectious inflammatory disease group ($n = 51$), neoplastic disease group ($n = 31$), and other diseases group ($n = 14$) according to 4 different etiological types. Another 154 patients were classified as undetermined group because the cause of fever was still unclear at discharge. There were significant differences in gender, age, and 18 other indicators among the four groups of patients with classic FUO with different etiological types ($P < 0.05$). The micro-F1 score for LightGBM was 75.8%, which was higher than that for the other four ML models, and the LightGBM prediction model had the best performance.

Conclusions: Infectious diseases are still the main etiological type of classic FUO. Based on 18 statistically significant clinical indicators such as gender and age, we constructed and evaluated five ML models. LightGBM model has a good effect on predicting the etiological type of classic FUO, which will play a good auxiliary decision-making function.

Keywords: fever of unknown origin, machine learning, etiology, retrospective analysis, LightGBM algorithm

INTRODUCTION

Fever of unknown origin (FUO) is a difficult and active medical topic in the diagnosis and treatment of difficult and complicated diseases in internal medicine, and it is a challenging problem for physicians (1, 2). Currently, there are four categories of FUOs: classic FUO, FUO in hospitalized patients, FUO in patients with agranulocytosis, and FUO in patients with human immunodeficiency virus (HIV) infection (3, 4). Among them, classic FUO is the most common, which is defined as a disease that lasts for >3 weeks, has a body temperature of >38.3°C at least three times, and cannot be diagnosed after systematic and comprehensive examinations in the outpatient or inpatient department of the hospital for >1 week (5, 6). There are >200 kinds of causes of classic FUO (7). For clinicians, because of its complex etiology, lack of characteristic clinical signs, and inadequate laboratory tests, the diagnosis is very difficult (8). The etiological categories of classic FUO are infectious disease, non-infectious inflammatory disease (NIID), neoplastic disease, and others, and the treatment methods vary greatly, including anti-infective drugs, hormones, and chemotherapy (9–11). With the development of immunohistopathology and modern imaging (12, 13), the diagnosis of classic FUO has become easier, but the final diagnosis is often difficult and up to 50% of cases cannot be confirmed (8, 11, 14, 15).

The diagnostic process of a classic FUO includes four steps: to determine whether it belongs to classic FUO, a first stage of primary screening, a second stage of specific examination, and treatment (including symptomatic and diagnostic treatment) (4). Among them, the first stage (etiological screening) includes improving medical history collection, physical examination, and non-invasive laboratory and auxiliary examinations in line with local medical standards. After the first stage of screening, some patients are diagnosed and some patients offer no diagnostic clues and enter the second stage, which requires further specific examinations. The second phase of the process is more complex, partly invasive, and more expensive. Therefore, the first stage of etiology screening is very important. If the etiology of a FUO can be classified into one category, no matter the disease that caused the FUO, the direction of diagnosis can be determined, which is of great significance to physicians (16, 17). Previous studies of classic FUO have focused on the etiology, prognosis, or diagnosis of classic FUO (18, 19). So far, few researchers have studied the etiological causes of classic FUO from the perspective of clinical prediction models and machine learning (ML) (16). In recent years, ML has been widely used in the medical field and has achieved good results in disease diagnosis, risk assessment, and other factors (20–22).

In this study, the clinical data and etiological types of classic FUO patients were retrospectively analyzed, and a predictive model of FUO etiology was established to help clinicians make reasonable decisions in the diagnosis of classic FUO, improve diagnostic accuracy, and reduce the misdiagnosis rate.

MATERIALS AND METHODS

Materials

The clinical data of 527 patients with classic FUO admitted to seven medical institutions in Chongqing from January 2012 to August 2021 were selected. The selected patients, whose ages ranged from 14 years old upwards, had each been hospitalized for more than a week with a fever higher than 38.3°C (101°F) that had occurred on several occasions and had persisted for at least 21 days (4, 8). Patients diagnosed with HIV infection before hospitalization, patients with immunodeficiency disorders, and pregnant women were screened out (4, 8). Of the 527 patients with classic FUO, 373 were finally diagnosed and 154 were not diagnosed at discharge. A total of 373 patients with classic FUO were divided into four groups according to their diagnosis and medical record information: infectious disease, NIID, neoplastic disease, and other diseases groups.

The index system of this study included general information (gender and age), past history (operation history and history of blood transfusion), accompanying symptoms (headache/consciousness disorders, nasal obstruction, sore throat, abdominal pain, arthralgia, muscle pain, and rash), physical (lymphadenopathy, hepatomegaly, and splenomegaly) and laboratory examinations [globulin, red blood cell (RBC), lactate dehydrogenase (LDH), C-reactive protein (CRP), procalcitonin (PCT), erythrocyte sedimentation rate (ESR), monocyte, basophils, eosinophils, lymphocyte, white blood cell (WBC), alkaline phosphatase (ALP), platelet (PLT), alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma-glutamyltransferase (GGT)], and the final diagnosis of etiological types.

The research protocol was approved by the Medical Research Ethics Committee of Chongqing Medical University.

Statistical Analysis

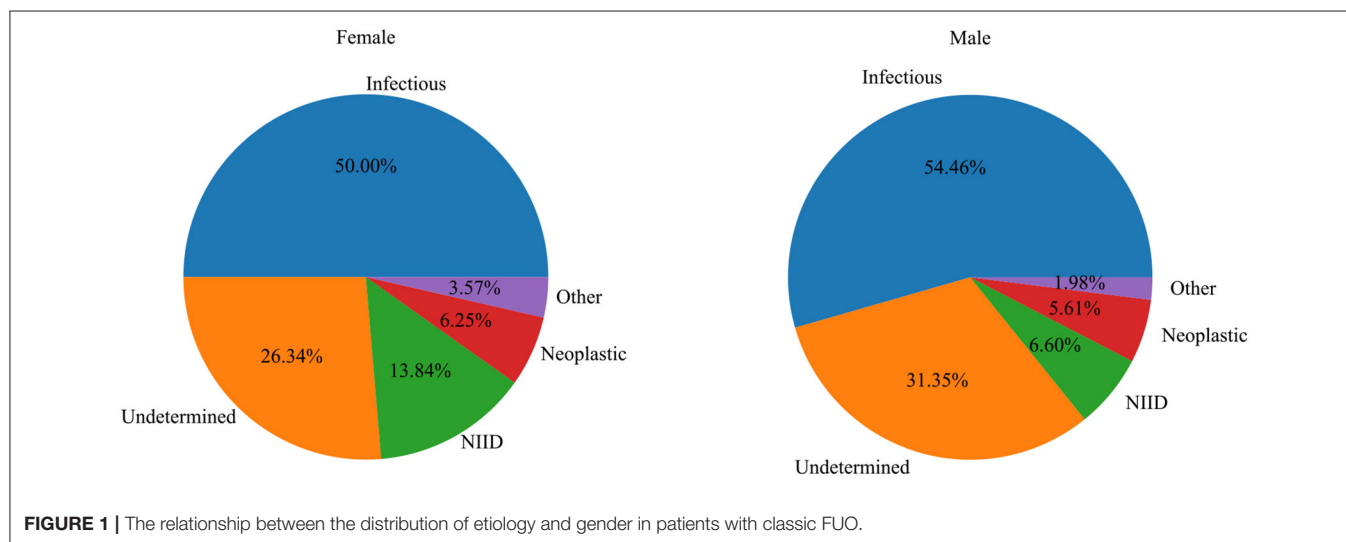
SPSS 25.0 statistical software was used for data processing. The continuity index was analyzed with a normality test, the median (M) and quartile (P_{25} , P_{75}) were used to express a non-normal distribution, the Kruskal–Wallis test was used to compare between groups. The normal distribution was expressed by $\bar{x} \pm s$. The analysis of variance was used to compare multiple groups, and the least significance difference (LSD) method was used for comparisons between two groups. The classification index was expressed by rate (%), and the comparison between groups was performed with a χ^2 -test. $P < 0.05$ was considered statistically significant. We used Python (version 3.7.3) for algorithm development.

Machine Learning

This study was based on the aforementioned differences that were statistically significant indicators to build the model. In order to determine the best model for classifying etiological types in this study, we compared the performance of the following representative ML classification algorithms: RF, SVM, LightGBM, ANN, and NB. For each algorithm, we used the 5-fold cross-validation method to split the data, each time using

TABLE 1 | Percentages of causes of classic FUO ranked by age.

Age	Infectious diseases (%)	Non-infectious inflammatory disease (%)	Neoplastic diseases (%)	Other diseases (%)	Undetermined (%)	Total
<20	13 (72.2)	3 (16.7)	0 (0)	0 (0)	2 (11.1)	18
20–39	48 (47.5)	9 (8.9)	4 (3.9)	7 (6.9)	33 (32.8)	101
40–59	99 (49.8)	25 (12.6)	12 (6.0)	5 (2.5)	58 (29.1)	199
≥60	117 (56.0)	14 (6.7)	15 (7.2)	2 (1.0)	61 (29.1)	209
Total	277 (52.5)	51 (9.7)	31 (5.9)	14 (2.7)	154 (29.2)	527



the training set to train the model and verify the performance of the model on the test set data. Because the predicted etiological types of this study had four categories and the categories were imbalanced, we evaluated the performance of the model using micro-F1. micro-F1 is suitable for multi-classification problems and unbalanced data, and higher values represent better model performance. The calculation method for micro-F1 is as follows (taking four categories as an example):

- a) Total $Recall_{mi} = \frac{TP_1 + TP_2 + TP_3 + TP_4}{TP_1 + TP_2 + TP_3 + TP_4 + FN_1 + FN_2 + FN_3 + FN_4}$;
- b) Total $Precision_{mi} = \frac{TP_1 + TP_2 + TP_3 + TP_4}{TP_1 + TP_2 + TP_3 + TP_4 + FP_1 + FP_2 + FP_3 + FP_4}$;
- c) Calculate $micro\ F1\ score = 2 \frac{Recall_{mi} \times Precision_{mi}}{Recall_{mi} + Precision_{mi}}$

wherein TP_i refers to a true positive of class i ; FP_i refers to a false positive of class i ; TN_i refers to a true negative of class i ; and FN_i refers to a false negative of class i .

RESULTS

Brief Introduction of the Cases Selected for the Study

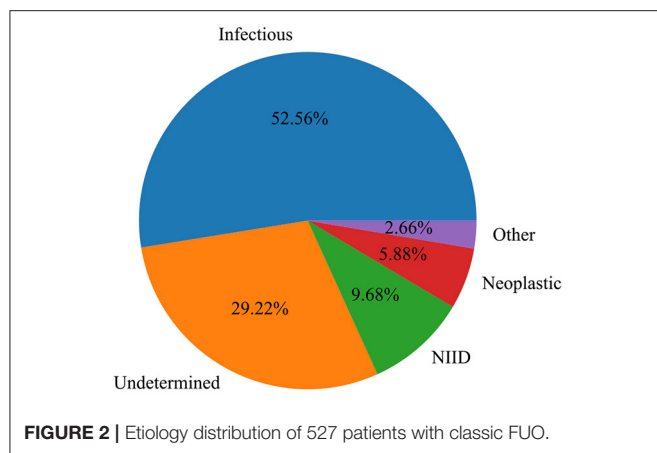
A total of 527 patients with classic FUO were collected from seven medical institutions in Chongqing, including 303 men (57.5%) and 224 women (42.5%). Of the patients, 3.4% ($n = 18$), 19.2% ($n = 101$), 37.8% ($n = 199$), and 39.6% ($n = 209$)

were <20, 20–39, 40–59, and ≥60 years, respectively. **Table 1**, **Figure 1** show the distribution of classic FUO etiologies by age and gender, respectively.

Infectious disease ($n = 277$; 52.5%) and NIID ($n = 51$; 9.7%) were the most common causes of classic FUO (**Figure 2**). Infectious diseases included bacterial ($n = 193$), tuberculosis ($n = 46$), and other bacterial infections ($n = 2$) and viral ($n = 21$), fungal ($n = 12$), parasitic ($n = 1$), and other pathogen infections ($n = 2$). The most common NIIDs were hemophagocytic syndrome ($n = 12$), anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis ($n = 9$), systemic lupus erythematosus ($n = 9$), and Adult-onset Still's disease ($n = 7$). Thirty-one cases (5.9%) were diagnosed as neoplastic diseases, of which eight cases were lymphoma. Other causes, such as subacute thyroiditis ($n = 9$) and drug fever ($n = 3$), were diagnosed in 14 patients (2.7%). A total of 29.2% ($n = 154$) of the patients remained undiagnosed at discharge (**Table 2**).

Test of the Difference in the Indexes of Patients With Classic FUO With Different Etiologies

There was a significant difference in the proportion of male and female patients with classic FUO among the four groups ($\chi^2 = 8.24$, $P < 0.05$). Male patients with FUO were common in the infectious and neoplastic disease groups, whereas female patients with FUO were common in the NIID and other diseases groups.



There was a significant difference in age among the groups ($H = 9.34$, $P < 0.05$). The age of patients with tumor disease was the oldest [57.00 (43.50, 67.50)], whereas the age of patients with other diseases was the youngest [42.00 (32.50, 50.75)]. Regarding their past history, there was significant difference between patients with or without an history of blood transfusion and patients diagnosed with different types of causes ($\chi^2 = 27.59$, $P < 0.001$). There were significant differences in concomitant symptoms and physical examinations among the four groups ($P < 0.05$), except for nasal obstruction ($\chi^2 = 2.66$, $P = 0.447$), abdominal pain ($\chi^2 = 5.79$, $P = 0.122$), and splenomegaly ($\chi^2 = 1.39$, $P = 0.708$). In terms of laboratory tests, RBC ($F = 6.97$, $P < 0.001$), LDH ($H = 12.37$, $P = 0.006$), PCT ($H = 15.69$, $P = 0.001$), monocyte ($H = 12.26$, $P = 0.007$), lymphocyte ($H = 8.51$, $P = 0.037$), ALP ($H = 9.83$, $P = 0.020$), AST ($H = 10.21$, $P = 0.017$), and GGT ($H = 8.70$, $P = 0.033$) were performed. The results are shown in **Tables 3, 4**.

Prediction Model Performance

There were significant differences in the 18 characteristics including age and gender among the four different etiological types of patients with classic FUO. On the basis of the aforementioned indicators, five ML models were constructed and the whole dataset was included in the analyses. **Table 5** shows the results of the five ML models. We mainly compared the sizes of the micro-F1 values. The micro-F1 value of each ML algorithm was the average of the five results in the 5-fold cross-validation. The micro-F1 of LightGBM was 75.8%, which was significantly higher than that of the other four ML algorithms (74.4, 73.4, 70.8, and 71.0%, respectively), and LightGBM has the best performance evaluation.

In order to better understand the contribution of each variable in our modeling results, we chose the LightGBM model with the best performance evaluation to present. Each variable was evaluated using Gini Importance, which is commonly used in ensembles of decision trees as a measure of a variable's impact in predicting a label that also takes into account the estimated error in randomly labeling an observation according to the known

TABLE 2 | Etiology distribution of 527 patients with classic FUO.

Etiology	N (%)	Etiology	N (%)
Infectious diseases	277 (52.5)	Other pathogenic infections	2 (0.4)
Bacterial infections	193 (36.6)	Mycoplasmal pneumonia	2 (0.4)
Respiratory system infection	116 (22.0)	Non-infectious inflammatory disease	51 (9.7)
Bloodstream infection	30 (5.7)	Hemophagocytic syndrome	12 (2.3)
Urinary tract infection	21 (4.0)	Systemic lupus erythematosus	9 (1.7)
Biliary tract infection	6 (1.1)	ANCA-associated vasculitis	9 (1.7)
Liver abscess	6 (1.1)	Adult onset still disease	7 (1.3)
Cellulitis	6 (1.1)	Sjogren syndrome	4 (0.8)
Pressure ulcers infection	2 (0.4)	Rheumatoid arthritis	2 (0.4)
Reproductive tract infection	2 (0.4)	Undifferentiated connective tissue disease	2 (0.4)
Infective endocarditis	2 (0.4)	Gouty arthritis	1 (0.2)
Umbilical infection	1 (0.2)	Dermatomyositis	1 (0.2)
Intra-abdominal infection	1 (0.2)	Takayasu arteritis	1 (0.2)
Tuberculosis	46 (8.7)	Crohn's disease	1 (0.2)
Pulmonary tuberculosis	35 (6.6)	Autoimmune hemolytic anemia	1 (0.2)
Extrapulmonary tuberculosis	11 (2.1)	macrophage activation syndrome	1 (0.2)
Other bacterial infections	2 (0.4)	Neoplastic diseases	31 (5.9)
Typhoid	1 (0.2)	Lymphoma	8 (1.5)
Brucellosis	1 (0.2)	Lung carcinoma	6 (1.1)
Viral infections	21 (4.0)	Hepatoma	5 (0.9)
HIV	10 (1.9)	Castleman's disease	3 (0.6)
Epstein-Barr virus	5 (0.9)	Acute myelogenous leukemia	2 (0.4)
Hepatitis B	4 (0.8)	Colon cancer	2 (0.4)
other viral infections	2 (0.4)	Myelodysplastic Syndrome	1 (0.2)
Fungal infections	12 (2.3)	Renal carcinoma	1 (0.2)
Candida albicans	2 (0.4)	Cholangiocarcinoma	1 (0.2)
Pneumocystis carinii pneumonia	2 (0.4)	Multiple myeloma	1 (0.2)
Cryptococcus neoformans	2 (0.4)	Thyroid carcinoma	1 (0.2)
Pulmonary aspergillosis	1 (0.2)	Other diseases	14 (2.7)
Candida tropicalis	1 (0.2)	Subacute thyroiditis	9 (1.7)
Other fungal infections	4 (0.8)	Drug fever	3 (0.6)
Parasitic infections	1 (0.2)	Hyperthyroidism	1 (0.2)
Malaria	1 (0.2)	Necrotizing lymphadenitis	1 (0.2)
		Undetermined	154 (29.2)

label distributions (23). **Figure 3** shows the ranking of feature importance for all variables in the model. The results showed that age, PCT, ALP, AST, and GGT were the top five important features in the model, which made a great contribution to the prediction results.

For the LightGBM model defined as the final prediction model, the relationship between each variable and the prediction outcome for the model is illustrated in **Figure 3**. To determine the most salient features that drove the model predictions,

TABLE 3 | Test of the difference of indexes (continuous indexes) in patients with classic FUO of different etiological types.

Variable	Infectious diseases (%)	Non-infectious inflammatory disease (%)	Neoplastic diseases (%)	Other diseases (%)	χ^2	P
No. of cases	277	51	31	14		
Gender						
Male	165 (59.6%)	20 (39.2%)	17 (54.8%)	6 (42.9%)	8.24	0.041
Female	112 (40.4%)	31 (60.8%)	14 (45.2%)	8 (57.1%)		
Operation history						
Yes	109 (39.4%)	20 (39.2%)	13 (41.9%)	4 (28.6%)	0.76	0.858
No	168 (60.6%)	31 (60.8%)	18 (58.1%)	10 (71.4%)		
History of blood transfusion						
Yes	35 (12.6%)	8 (15.7%)	15 (48.4%)	1 (7.1%)	27.59	<0.001
No	242 (87.4%)	43 (84.3%)	16 (51.6%)	13 (92.9%)		
Headache/consciousness disorders						
Yes	66 (23.8%)	6 (11.8%)	3 (9.7%)	8 (57.1%)	16.32	<0.001
No	211 (76.2%)	45 (88.2%)	28 (90.3%)	6 (42.9%)		
Nasal obstruction						
Yes	9 (3.2%)	3 (5.9%)	0 (0.0%)	0 (0.0%)	2.66	0.447
No	268 (96.8%)	48 (94.1%)	31 (100.0%)	14 (100.0%)		
Sore throat						
Yes	25 (9.0%)	12 (23.5%)	1 (3.2%)	6 (42.9%)	23.96	<0.001
No	252 (91.0%)	39 (76.5%)	30 (96.8%)	8 (57.1%)		
Abdominal pain						
Yes	17 (6.1%)	6 (11.8%)	5 (16.1%)	2 (14.3%)	5.79	0.122
No	260 (93.9%)	45 (88.2%)	26 (83.9%)	12 (85.7%)		
Arthralgia						
Yes	19 (6.9%)	12 (23.5%)	3 (9.7%)	2 (14.3%)	14.09	0.003
No	258 (93.1%)	39 (76.5%)	28 (90.3%)	12 (85.7%)		
Muscle pain						
Yes	30 (10.8%)	12 (23.5%)	1 (3.2%)	0 (0.0%)	11.25	0.010
No	247 (89.2%)	39 (76.5%)	30 (96.8%)	14 (100.0%)		
Rash						
Yes	7 (2.5%)	5 (9.8%)	2 (6.5%)	2 (14.3%)	9.63	0.022
No	270 (97.5%)	46 (90.2%)	29 (93.5%)	12 (85.7%)		
Lymphadenopathy						
Yes	10 (3.6%)	9 (17.6%)	4 (12.9%)	0 (0.0%)	18.10	<0.001
No	267 (96.4%)	42 (82.4%)	27 (87.1%)	14 (100.0%)		
Hepatomegaly						
Yes	1 (0.4%)	3 (5.9%)	3 (9.7%)	0 (0.0%)	18.41	<0.001
No	276 (99.6%)	48 (94.1%)	28 (90.3%)	14 (100.0%)		
Splenomegaly						
Yes	9 (3.2%)	2 (3.9%)	2 (6.5%)	0 (0.0%)	1.39	0.708
No	268 (96.8%)	49 (96.1%)	29 (93.5%)	14 (100.0%)		

we calculated the SHapley Additive exPlanation (SHAP) values of the best-performing models for different etiological types. **Figures 4A–D** shows the important characteristics of each etiological type. For infectious diseases, age, lymphocyte, and RBC increased and ALP and LDH decreased in favor of the classifier to predict infectious diseases. For NIID, higher LDH, monocyte, and AST; younger age; and a lower lymphocyte were helpful to the classifier to predict NIID. For neoplastic diseases,

higher ALP, monocyte, and lymphocyte; older age; and previous history of blood transfusion were conducive to the classifier to predict the cause of neoplastic diseases. For other diseases, accompanied by headache or disturbance of consciousness and sore throat symptoms, younger age and lower PCT and GGT were conducive to the classifier to predict the cause of tumor diseases. Other important features of each etiological type are shown in **Figure 4**.

TABLE 4 | Test of difference of indexes (classification indexes) in patients with classic FUO of different etiological types.

Variable	Infectious diseases (%)	Non-infectious inflammatory disease (%)	Neoplastic diseases (%)	Other diseases (%)	F/H	P
Age [year, M (P ₂₅ , P ₇₅)]	55.00 (42.00, 68.00)	51.00 (40.50, 60.50)	57.00 (43.50, 67.50)	42.00 (32.50, 50.75)	9.34	0.025
Laboratory examination						
Globulin (g/L, $\bar{x} \pm s$)	31.89 \pm 6.95	33.20 \pm 5.77	32.37 \pm 7.66	33.47 \pm 3.50	0.44	0.725
RBC ($\times 10^{12}/L$, $\bar{x} \pm s$)	3.82 \pm 0.72	3.38 \pm 0.68	3.38 \pm 0.85	3.98 \pm 0.67	6.97	<0.001
LDH [U/L, M (P ₂₅ , P ₇₅)]	219.50 (157.75, 441.48)	340.00 (198.50, 629.50)	408.00 (244.45, 867.00)	191.00 (166.50, 251.58)	12.37	0.006
CRP [mg/L, M (P ₂₅ , P ₇₅)]	61.90 (20.23, 115.11)	48.23 (10.21, 130.42)	112.78 (62.32, 152.64)	24.94 (8.38, 130.35)	6.59	0.086
PCT [ng/ml, M (P ₂₅ , P ₇₅)]	0.21 (0.09, 0.77)	0.26 (0.10, 0.59)	0.43 (0.19, 1.89)	0.07 (0.05, 0.12)	15.69	0.001
ESR [mm/H, M (P ₂₅ , P ₇₅)]	57.00 (29.50, 83.50)	69.00 (36.00, 94.50)	70.00 (38.75, 91.00)	49.00 (26.00, 78.75)	3.33	0.344
Monocyte [$\times 10^9/L$, M (P ₂₅ , P ₇₅)]	0.47 (0.31, 0.66)	0.44 (0.15, 0.70)	0.71 (0.55, 0.99)	0.55 (0.41, 0.77)	12.26	0.007
Basophils [$\times 10^9/L$, M (P ₂₅ , P ₇₅)]	0.01 (0.01, 0.02)	0.01 (0.00, 0.02)	0.01 (0.01, 0.02)	0.01 (0.00, 0.02)	0.53	0.913
Eosinophils [$\times 10^9/L$, M (P ₂₅ , P ₇₅)]	0.04 (0.01, 0.10)	0.02 (0.00, 0.10)	0.03 (0.01, 0.12)	0.02 (0.00, 0.06)	1.67	0.645
Lymphocyte [$\times 10^9/L$, M (P ₂₅ , P ₇₅)]	0.95 (0.61, 1.41)	0.79 (0.55, 1.15)	0.92 (0.61, 1.71)	1.43 (0.94, 1.61)	8.51	0.037
WBC [$\times 10^9/L$, M (P ₂₅ , P ₇₅)]	7.20 (5.45, 10.51)	8.04 (5.20, 11.39)	10.27 (6.15, 16.94)	6.48 (5.52, 11.29)	3.15	0.370
ALP [U/L, M (P ₂₅ , P ₇₅)]	83.30 (65.50, 119.25)	92.00 (71.00, 134.00)	110.60 (96.50, 208.00)	96.50 (78.75, 128.50)	9.83	0.020
PLT [$\times 10^9/L$, M (P ₂₅ , P ₇₅)]	224.00 (172.00, 311.50)	214.00 (108.00, 303.00)	200.00 (86.00, 301.50)	334.00 (159.50, 408.00)	3.75	0.289
ALT [U/L, M (P ₂₅ , P ₇₅)]	24.00 (13.00, 41.00)	29.00 (14.00, 53.60)	21.00 (14.00, 40.00)	26.00 (21.00, 36.00)	1.89	0.597
AST [U/L, M (P ₂₅ , P ₇₅)]	26.00 (17.00, 44.00)	33.50 (23.00, 80.00)	30.00 (19.00, 53.00)	17.00 (14.40, 28.00)	10.21	0.017
GGT [U/L, M (P ₂₅ , P ₇₅)]	46.00 (23.00, 108.00)	51.00 (27.00, 115.00)	77.00 (53.00, 196.00)	33.00 (28.00, 71.00)	8.70	0.033

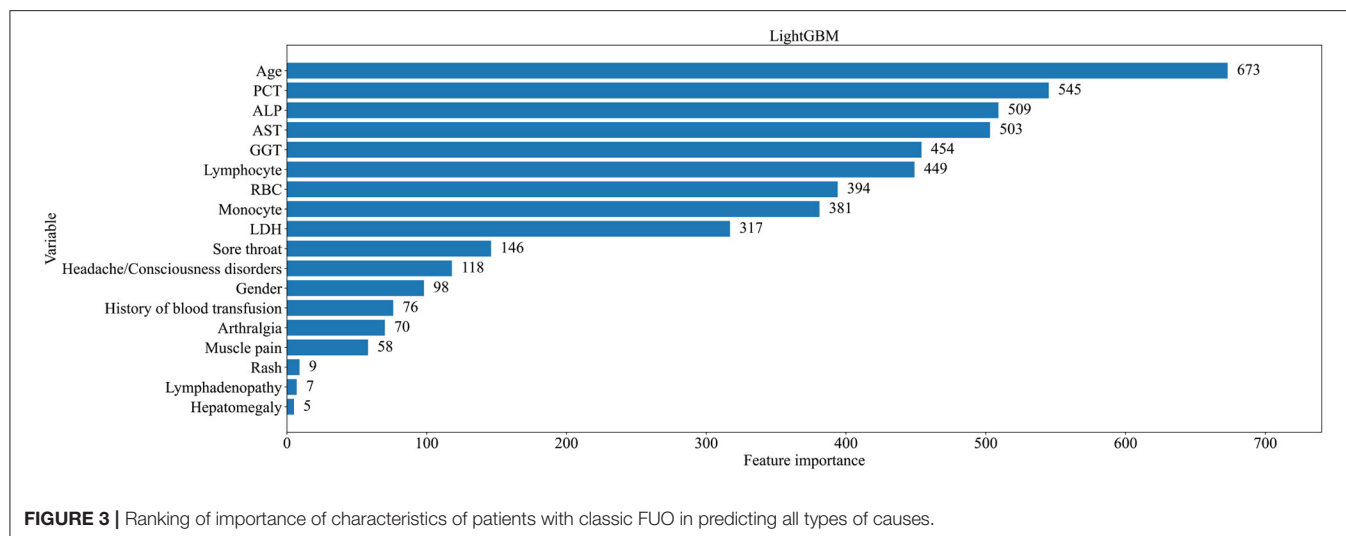
TABLE 5 | Comparison of five ML models.

Model	micro-F1 score, %	Recall _{mi} , %	Precision _{mi} , %
RF	74.4	74.4	74.4
SVM	73.4	73.4	73.4
LightGBM	75.8	75.8	75.8
ANN	70.8	70.8	70.8
NB	71.0	71.0	71.0

DISCUSSION

The etiological distribution of 527 patients with classic FUO was analyzed retrospectively, and the patients were divided into five groups according to the final diagnosis, including the group with unknown etiology of fever at discharge. Analysis showed that infectious diseases were the most common cause of classic FUO, followed by NIID. These results are consistent with most of the previous research results at home and abroad (8, 24–26), and the reasons for this phenomenon may be related to the non-standard use of antibiotics and drug resistance leading to disease persistence and changes in the rule of fever type. There were also different findings between this study and previous studies. First of all, the proportion of tuberculosis infection in infectious diseases was 16.6% (46/277), which was significantly lower than that of Li's study (30%) (27) but similar to that of Zhai's study (17.6%) (24). This change may be related to the strengthening of public awareness of tuberculosis and the improvement of medical conditions in recent years. Conversely, with the improvements

in diagnosis and treatments, most tuberculosis infections can be diagnosed clearly in the early stage, thus reducing the proportion of patients with tuberculosis with FUO. Second, this study showed that the proportion of HIV infection was 3.6%, which is higher than the previous research results (1%) (8), which may be related to the increase in floating populations, sexual attitudes and sexual behavior, sexual orientation changes, and other factors that increase the risk of HIV infection. This study found that the proportion of NIIDs was 9.7%, which was significantly lower than the results of Naito's research (30.6%) in 2013 (28). It may be due to the early use of relevant immunological indicators, which enabled the early diagnosis of autoimmune diseases with more typical symptoms, and no longer classified as classic FUO. In this study, neoplastic diseases accounted for 5.9% of classic FUO, which was significantly lower than 15%, as reported in the literature (1), which may be due to PET-CT and serum tumor markers that have been widely used in recent years (29, 30). Many malignant tumors can be diagnosed early, and the widespread use of early biopsy is also a reason for the reduction of neoplastic diseases with classic FUO. Among other diseases, subacute thyroiditis accounts for a considerable proportion (64.3%), which is in line with the findings of Popovska-Jovicić (31). Subacute thyroiditis rarely has persistent fever as the only clinical manifestation, generally have some related clinical manifestations (32), such as upper respiratory tract infection symptoms, weight loss, neck pain, fatigue, and anorexia. Routine thyroid color ultrasound, thyroid antibody tests, and thyroid function tests are rarely performed; therefore, thyroiditis is easily misdiagnosed as an upper respiratory tract infection. The manifestations of elderly patients with subacute thyroiditis

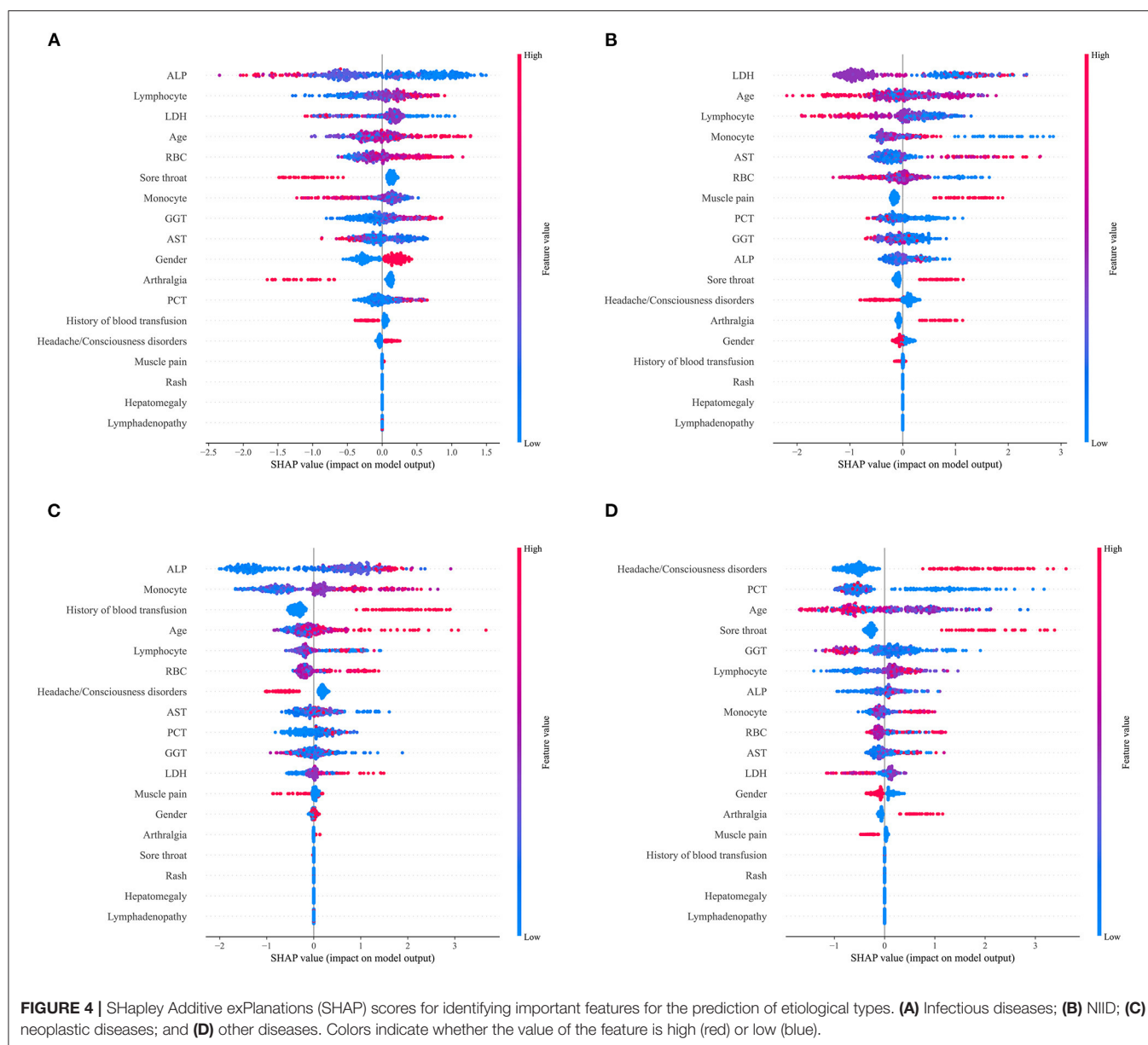


are often not obvious, and other underlying diseases may also have clinical signs of subacute thyroiditis, or patients cannot provide a good medical history. A total of 154 patients (29.2%) whose reason for fever was still not clear at the time of discharge from the hospital and who had early discharge due to economic or other personal reasons, had shorter hospitalization times, which led to inadequate examination and diagnostic treatment during hospitalization. A tentative diagnosis followed, and eventually, they were classified in the unclear group, which may be the reason for the high proportion of patients in this group.

In this study, 373 patients with classic FUO were divided into four groups and the groups were compared. The study found that male patients were more common and older in the tumor group than in the other three groups, whereas female patients were more common and younger in the other disease groups than in the other three groups. Regarding past history, accompanying symptoms and physical examination of patients with classic FUO had some special clinical signs that could provide clinical clues worthy of attention. Among them, arthralgia was very common. In this study, patients with classic FUO in all four groups with different etiological types had symptoms of arthralgia. The most common rheumatic diseases were heterogeneous diseases with joint, bone, and muscle pain as the main symptoms, which could have involved internal organs (33). Infectious arthritis in infectious diseases and some hematological tumors also have manifestations of arthralgia (34, 35). Rash is an important concomitant sign in patients with classic FUO, which may provide an important clue for the etiological diagnosis of classic FUO. In classic FUO, most diseases can be accompanied by clinical signs of skin rash (4), including (1) infectious diseases, such as Epstein-Barr virus infection, typhoid fever, and infective endocarditis; (2) NIID, such as systemic lupus erythematosus, dermatomyositis, and adult-onset Still's disease; (3) neoplastic diseases, such as lymphoma; and (4) other diseases, such as drug fever. In this

study, all patients in the infectious disease, NIID, and neoplastic disease groups had symptoms of lymphadenopathy. Lymph node enlargement is either localized or generalized (36). Localized lymphadenopathy involves a draining region, often caused by a non-specific inflammatory response of the tissue or organ in the draining region or by lymphatic metastasis of malignant tumors corresponding to the draining region. Direct invasion of infectious pathogens or immune response caused by infection, allergic or autoimmune diseases, and invasion of neoplastic diseases can lead to systemic lymphadenopathy. In laboratory examinations, we found that the levels of LDH, PCT, monocyte, ALP, and GGT in the neoplastic disease group were significantly higher than those in the other three groups, whereas the level of AST in the NIID group was higher than that in the other three groups, and the levels of RBC and lymphocyte in the other disease groups were higher than those in the other three groups. Among them, the higher PCT levels in the neoplastic disease group was an interesting finding. In general, increase concentration of blood PCT is associated with severe bacterial infection. However, the clinical interpretation of elevated PCT concentration in blood represents a great challenge in cancer patients since its values might be influenced by several factors such as the presence of metastasis or neuroendocrine function of malignant tissue (37). In these cases, PCT concentrations can be elevated regardless of infections, manifesting a poor specificity for bacterial infection. Matzaraki et al. (38) indicated that patients with solid tumors, metastasis, and no evidence of infection had markedly elevated PCT levels, especially those with generalized metastatic disease. Similarly, Liu et al. (39) show that in the absence of bacterial infection, PCT levels are elevated in patients with certain inflammatory conditions, such as Kawasaki disease, Adult-onset Still's disease and some cancers like medullary carcinoma of the thyroid and small-cell lung carcinoma.

On the basis of the aforementioned discussion, this study screened 18 indicators, such as gender and age, and constructed



a clinical prediction model of the etiological types of patients with classic FUO. The indicators included in the model are all from the indicators reported in the consensus on current management of fever of unknown origin, which adds reliability to the model we constructed. We compared five ML algorithms, all of which were tested using the 5-fold cross-validation method. These five ML algorithms are widely used in clinical prediction model construction. For example, SVM learning is widely used in cancer genomics (40). Compared to other ML algorithms, SVM is very powerful in identifying subtle patterns in complex data sets. However, there are also some shortcomings, such as slow training speed and difficult to understand the internal operation. Ivanović et al. (41) constructed an ANN model to predict

the lymph node status of clinical lymph node-negative breast cancer. ANN have the ability to adapt to variable interaction and non-linear correlation, but also have the constraints of opaque underlying model and difficult to explain (42). Yang et al. (43) constructed a response prediction model of breast cancer neoadjuvant chemotherapy based on NB algorithm. In their study, the NB algorithm showed higher predictive values than other algorithms. Each ML algorithm has its own advantages and disadvantages, but in our research data, the micro-F1 value of the LightGBM model was 75.8%, which was significantly higher than that of the other four ML algorithms. It is suggested that the LightGBM model has better predictive performance for the classification of etiological types of patients with classic

FUO. LightGBM is a distributed gradient lifting framework based on a decision tree algorithm, which has high efficiency and performance in dealing with binary classifications and multi-classification problems (44–46). LightGBM is an ensemble algorithm developed by Microsoft, which is superior to other machine learning methods for disease diagnosis in many cases (45). Fundamentally, this is achieved by combining multiple base classifiers into an ensemble model by learning the inherent statistics of the combined classifiers and, hence, outperforming the single classifiers. In addition, the RF model also achieved a high accuracy, micro-F1 was 74.4%, which was second only to LightGBM in the results of this study. RF is recognized as one type of ensemble learning method and are effective for the most classification and regression tasks (47), which further illustrates the advantages of ensemble learning methods. In this study, 18 indexes related to the etiological diagnosis of classic FUO were ranked in descending order of importance. Among them, the ranking of laboratory indicators can provide doctors with decision support for laboratory examination to a certain extent. We also calculated the SHAP value of the best performance model according to the cause category for explaining the model, and we could clearly see the influence of the characteristics of each cause type on the output of the model. In the following research, how to deal with the imbalanced data set and the small sample size problem is worth considering, because these problems affect the performance of the prediction model to some extent. Data imbalance is widespread in the real world, especially in medical big data, which affects the accuracy of medical diagnosis classification learning algorithm to a certain extent. In order to solve the problem of poor performance of medical diagnosis learning algorithms due to the serious shortage of minority samples, Han et al. (48) proposed a distribution-sensitive oversampling method for unbalanced large data, including the distribution-sensitive minority sample selection algorithm and the minority sample synthetic algorithm of weight adaptive adjustment, which improves the quality of newly generated minority samples. This may be a way to improve the accuracy of the model. In addition, few-shot learning is also a research direction that we should pay attention to. Few-shot learning is such a research topic that studies how to learn a new concept from few training data of this concept and has received significant attention from the machine learning community (49).

Our study has several limitations. First, this was a retrospective study, which had its own shortcomings, such as information bias. Second, the prediction model may have lacked generality because the 30 variables are still too few and many other variables were omitted because of the loss of too many values. Therefore, we hope to include more patients and variables in future studies. In addition, 154 cases with unknown etiology of fever were not included in the model, which do exist in the real world. Therefore, the accuracy in reality may be lower, and these situations should be taken into account in future studies.

CONCLUSIONS

In summary, this study retrospectively analyzed the clinical data of 527 patients with classic FUO from 7 medical institutions in Chongqing, discussed the differences of clinical indexes of 373 patients with classic FUO under 4 different etiological types, and introduced ML methods into the study of classic FUO to explore the application value of ML methods in the etiological diagnosis of classic FUO. The data of this study shows that infectious diseases are still the main etiological type of classic FUO. Based on 18 statistically significant clinical indicators such as gender and age, we constructed and compared 5 different ML algorithm models. The results show that compared with other algorithms, LightGBM is the best, and its micro-F1 value is 75.8%. We also use feature importance ranking and SHAP values to enhance the interpretability of the model. We believe that our model will provide clinicians with the most likely direction of etiological diagnosis in the diagnosis of classic FUO, assist clinicians to make reasonable decisions, improve the diagnostic accuracy of classic FUO, and reduce the misdiagnosis rate.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Research Ethics Committee of Chongqing Medical University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YY, CC, and KP participated in the research design and coordination and helped in drafting the manuscript. KP contributed in data acquisition. YY, CC, YL, ZZ, and LX analyzed the data. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by grants from the Natural Science Foundation of Chongqing (cstc2019jcyj-msxmX0027); the College of Medical Informatics, Chongqing Medical University, China, Student Research and Innovation Experiment Project (2019C011); the Philosophy and Social Sciences Innovation Team of Chongqing Medical University (ZX190101); and the Project of Innovative Research and Demonstration Base of Children's Medical Security in Children's Hospital Affiliated to Chongqing Medical University (NCRCHD- 2019-HP-04).

REFERENCES

- Kaya A, Ergul N, Kaya SY, Kilic F, Yilmaz MH, Besirli K, et al. The management and the diagnosis of fever of unknown origin. *Expert Rev Anti Infect Ther.* (2013) 11:805–15. doi: 10.1586/14787210.2013.814436
- Li JJ, Huang WX, Shi ZY, Sun Q, Xin XJ, Zhao JQ, et al. Comparison of classical diagnostic criteria and Chinese revised diagnostic criteria for fever of unknown origin in Chinese patients. *Ther Clin Risk Manag.* (2016) 12:1545–51. doi: 10.2147/TCRM.S97863
- Hayakawa K, Ramasamy B, Chandrasekar PH. Fever of unknown origin: an evidence-based review. *Am J Med Sci.* (2012) 344:307–16. doi: 10.1097/MAJ.0b013e31824ae504
- Zhang WH, Li TS. Consensus on current management of fever of unknown origin. *Shanghai Med J.* (2018) 41:385–400.
- Unger M, Karanikas G, Kerschbaumer A, Winkler S, Aletaha D. Fever of unknown origin (FUO) revised. *Wien Klin Wochenschr.* (2016) 128:796–801. doi: 10.1007/s00508-016-1083-9
- Wright WF, Mulders-Manders CM, Auwaerter PG, Bleeker-Rovers CP. Fever of Unknown Origin (FUO) - a call for new research standards and updated clinical management. *Am J Med.* (2021). doi: 10.1016/j.amjmed.2021.07.038
- Fusco FM, Pisapia R, Nardiello S, Cicala SD, Gaeta GB, Brancaccio G. Fever of unknown origin (FUO): which are the factors influencing the final diagnosis? A 2005–2015 systematic review. *BMC Infect Dis.* (2019) 19:653. doi: 10.1186/s12879-019-4285-8
- Zhou G, Zhou Y, Zhong C, Ye H, Liu Z, Liu Y, et al. Retrospective analysis of 1,641 cases of classic fever of unknown origin. *Ann Transl Med.* (2020) 8:690. doi: 10.21037/atm-20-3875
- Loizidou A, Aoun M, Klustersky J. Fever of unknown origin in cancer patients. *Crit Rev Oncol Hematol.* (2016) 101:125–30. doi: 10.1016/j.critrevonc.2016.02.015
- Mulders-Manders CM, Engwerda C, Simon A, van der Meer JWM, Bleeker-Rovers CP. Long-term prognosis, treatment, and outcome of patients with fever of unknown origin in whom no diagnosis was made despite extensive investigation: a questionnaire based study. *Medicine.* (2018) 97:e11241. doi: 10.1097/MD.00000000000011241
- Tan Y, Liu X, Shi X. Clinical features and outcomes of patients with fever of unknown origin: a retrospective study. *BMC Infect Dis.* (2019) 19:198. doi: 10.1186/s12879-019-3834-5
- Besson FL, Chaumet-Riffaud P, Playe M, Noel N, Lambotte O, Goujard C, et al. Contribution of (18)F-FDG PET in the diagnostic assessment of fever of unknown origin (FUO): a stratification-based meta-analysis. *Eur J Nucl Med Mol Imaging.* (2016) 43:1887–95. doi: 10.1007/s00259-016-3377-6
- Kouijzer IJE, Mulders-Manders CM, Bleeker-Rovers CP, Oyen WJG. Fever of unknown origin: the value of FDG-PET/CT. *Semin Nucl Med.* (2018) 48:100–7. doi: 10.1053/j.semnuclmed.2017.11.004
- Kouijzer IJ, Bleeker-Rovers CP, Oyen WJ. FDG-PET in fever of unknown origin. *Semin Nucl Med.* (2013) 43:333–9. doi: 10.1053/j.semnuclmed.2013.04.005
- Mulders-Manders CM, Simon A, Bleeker-Rovers CP. Rheumatologic diseases as the cause of fever of unknown origin. *Best Pract Res Clin Rheumatol.* (2016) 30:789–801. doi: 10.1016/j.berh.2016.10.005
- Jiang H, Li Y, Zeng X, Xu N, Zhao C, Zhang J, et al. Exploring fever of unknown origin intelligent diagnosis based on clinical data: model development and validation. *JMIR Med Inform.* (2020) 8:e24375. doi: 10.2196/24375
- Zhao MZ, Ruan QR, Xing MY, Wei S, Xu D, Wu ZH, et al. A diagnostic tool for identification of etiologies of fever of unknown origin in adult patients. *Curr Med Sci.* (2019) 39:589–96. doi: 10.1007/s11596-019-2078-3
- Naito T, Tanei M, Ikeda N, Ishii T, Suzuki T, Morita H, et al. Key diagnostic characteristics of fever of unknown origin in Japanese patients: a prospective multicentre study. *BMJ Open.* (2019) 9:e032059. doi: 10.1136/bmjopen-2019-032059
- Georga S, Exadaktylou P, Petrou I, Katsampoukas D, Mpalaris V, Moralidis EI, et al. Diagnostic value of (18)F-FDG-PET/CT in patients with FUO. *J Clin Med.* (2020) 9:2112. doi: 10.3390/jcm9072112
- Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. *Epilepsia.* (2019) 60:2037–47. doi: 10.1111/epi.16333
- Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. *JMIR Public Health Sur.* (2020) 6:e22400. doi: 10.2196/22400
- Jo YT, Joo SW, Shon SH, Kim H, Kim Y, Lee J. Diagnosing schizophrenia with network analysis and a machine learning method. *Int J Methods Psychiatr Res.* (2020) 29:e1818. doi: 10.1002/mpr.1818
- Kim A, Miano T, Chew R, Eggers M, Nonnemaker J. Classification of twitter users who tweet about e-cigarettes. *JMIR Public Health Sur.* (2017) 3:e63. doi: 10.2196/publichealth.8060
- Zhai YZ, Chen X, Liu X, Zhang ZQ, Xiao HJ, Liu G. Clinical analysis of 215 consecutive cases with fever of unknown origin: a cohort study. *Medicine.* (2018) 97:e10986. doi: 10.1097/MD.00000000000010986
- Lv KL, Xin XJ, Li CH. The etiology analysis of 548 cases patients of fever with unknown origin. *Electron J Emerg Infect Dis.* (2020) 5:258–61. doi: 10.19871/j.cnki.xfcrbzz.2020.04.009
- Tian D, Qi WJ, Wang C, Huang GW. Disease spectrum and etiology study on 347 patients with fever of unknown origin. *Clin J Med Officers.* (2020) 48:1433–6. doi: 10.16680/j.1671-3826.2020.12.14
- Li JB, Zhang JP, Chen BY. Etiological factors for 541 patients with fever of unknown origin: a retrospective analysis. *Chinese J Nosocomiol.* (2011) 21:1587–9.
- Naito T, Mizooka M, Mitsumoto F, Kanazawa K, Torikai K, Ohno S, et al. Diagnostic workup for fever of unknown origin: a multicenter collaborative retrospective study. *BMJ Open.* (2013) 3:e003971. doi: 10.1136/bmjopen-2013-003971
- Wang W, Xu X, Tian B, Wang Y, Du L, Sun T, et al. The diagnostic value of serum tumor markers CEA, CA19-9, CA125, CA15-3, and TPS in metastatic breast cancer. *Clin Chim Acta.* (2017) 470:51–55. doi: 10.1016/j.cca.2017.04.023
- Zauch JM, Chauvie S, Zaucha R, Biggii A, Gallamini A. The role of PET/CT in the modern treatment of Hodgkin lymphoma. *Cancer Treat Rev.* (2019) 77:44–56. doi: 10.1016/j.ctrv.2019.06.002
- Popovska-Jovicić B, Canović P, Gajović O, Raković I, Mijailović Z. Fever of unknown origin: most frequent causes in adults patients. *Vojnosanitetski Pregled.* (2016) 73:21–5. doi: 10.2298/VSP140820128P
- Bahowairath FA, Woodhouse N, Hussain S, Busaidi MA. Lesson of the month 1: subacute thyroiditis: a rare cause of fever of unknown origin. *Clin Med.* (2017) 17:86–7. doi: 10.7861/clinmedicine.17-1-86
- Kadavath S, Efthimiou P. Adult-onset Still's disease-pathogenesis, clinical manifestations, and new treatment options. *Ann Med.* (2015) 47:6–14. doi: 10.3109/07853890.2014.971052
- Ross JJ. Septic arthritis of native joints. *Infect Dis Clin North Am.* (2017) 31:203–18. doi: 10.1016/j.idc.2017.01.001
- Stephens DM, Byrd JC. How I manage ibrutinib intolerance and complications in patients with chronic lymphocytic leukemia. *Blood.* (2019) 133:1298–307. doi: 10.1182/blood-2018-11-846808
- Gaddey HL, Riegel AM. Unexplained lymphadenopathy: evaluation and differential diagnosis. *Am Fam Phys.* (2016) 94:896–903.
- Durnas B, Watek M, Wollny T, Niemirowicz K, Marzec M, Bucki R, et al. Utility of blood procalcitonin concentration in the management of cancer patients with infections. *Onco Targets Ther.* (2016) 9:469–75. doi: 10.2147/OTT.S95600
- Matzaraki V, Alexandraki KI, Venetsanou K, Piperi C, Myrianthefs P, Malamos N, et al. Evaluation of serum procalcitonin and interleukin-6 levels as markers of liver metastasis. *Clin Biochem.* (2007) 40:336–42. doi: 10.1016/j.clinbiochem.2006.10.027
- Liu W, Sigdel KR, Wang Y, Su Q, Huang Y, Zhang YL, et al. High level serum procalcitonin associated gouty arthritis susceptibility: from a southern chinese han population. *PLoS One.* (2015) 10:e0132855. doi: 10.1371/journal.pone.0132855
- Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genom Proteomics.* (2018) 15:41–51. doi: 10.21873/cgp.20063
- Ivanović D, Kupusinac A, Stokić E, Doroslovački R, Ivetić D. ANN prediction of metabolic syndrome: a complex puzzle that will be completed. *J Med Syst.* (2016) 40:264. doi: 10.1007/s10916-016-0601-7

42. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol.* (2019) 188:2222–39. doi: 10.1093/aje/kwz189
43. Yang L, Fu B, Li Y, Liu Y, Huang W, Feng S, et al. Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a Naive Bayes algorithm. *Comput Methods Programs Biomed.* (2020) 192:105458. doi: 10.1016/j.cmpb.2020.105458
44. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep.* (2020) 10:11981. doi: 10.1038/s41598-020-68771-z
45. Rufo DD, Debelee TG, Ibenthal A, Negera WG. Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics.* (2021) 11:1714. doi: 10.3390/diagnostics11091714
46. Zeng H, Yang C, Zhang H, Wu Z, Zhang J, Dai G, et al. A LightGBM-Based EEG analysis method for driver mental states classification. *Comput Intell Neurosci.* (2019) 2019:3761203. doi: 10.1155/2019/3761203
47. Wang Y, Xia ST, Tang Q, Wu J, Zhu X. A novel consistent random forest framework: bernoulli random forests. *IEEE Trans Neural Networks Learn Syst.* (2018) 29:3510–23. doi: 10.1109/TNNLS.2017.2729778
48. Han W, Huang Z, Li S, Jia Y. Distribution-sensitive unbalanced data oversampling method for medical diagnosis. *J Med Syst.* (2019) 43:39. doi: 10.1007/s10916-018-1154-8
49. Lai N, Kan M, Han C, Song X, Shan S. Learning to learn adaptive classifier-predictor for few-shot learning. *IEEE Trans Neural Networks Learn Syst.* (2021) 32:3458–70. doi: 10.1109/TNNLS.2020.3011526

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Yan, Chen, Liu, Zhang, Xu and Pu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records

Jingfeng Chen^{1,2*}, Chonghui Guo^{2*}, Menglin Lu² and Suying Ding¹

¹ Health Management Center, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, ² School of Economics and Management, Institute of Systems Engineering, Dalian University of Technology, Dalian, China

OPEN ACCESS

Edited by:

Yi-Ju Tseng,
National Central University, Taiwan

Reviewed by:

Martin Hofmann-Apitius,
Fraunhofer Institute for Algorithms and
Scientific Computing (FHG), Germany
Hsin-Yao Wang,

Linkou Chang Gung Memorial
Hospital, Taiwan

*Correspondence:

Chonghui Guo
dlutguo@dlut.edu.cn
Jingfeng Chen
fccjchen@zzu.edu.cn

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 12 October 2021

Accepted: 21 December 2021

Published: 20 January 2022

Citation:

Chen J, Guo C, Lu M and Ding S
(2022) Unifying Diagnosis Identification
and Prediction Method Embedding
the Disease Ontology Structure From
Electronic Medical Records.
Front. Public Health 9:793801.
doi: 10.3389/fpubh.2021.793801

Objective: The reasonable classification of a large number of distinct diagnosis codes can clarify patient diagnostic information and help clinicians to improve their ability to assign and target treatment for primary diseases. Our objective is to identify and predict a unifying diagnosis (UD) from electronic medical records (EMRs).

Methods: We screened 4,418 sepsis patients from a public MIMIC-III database and extracted their diagnostic information for UD identification, their demographic information, laboratory examination information, chief complaint, and history of present illness information for UD prediction. We proposed a data-driven UD identification and prediction method (UDIPM) embedding the disease ontology structure. First, we designed a set similarity measure method embedding the disease ontology structure to generate a patient similarity matrix. Second, we applied affinity propagation clustering to divide patients into different clusters, and extracted a typical diagnosis code co-occurrence pattern from each cluster. Furthermore, we identified a UD by fusing visual analysis and a conditional co-occurrence matrix. Finally, we trained five classifiers in combination with feature fusion and feature selection method to unify the diagnosis prediction.

Results: The experimental results on a public electronic medical record dataset showed that the UDIPM could extracted a typical diagnosis code co-occurrence pattern effectively, identified and predicted a UD based on patients' diagnostic and admission information, and outperformed other fusion methods overall.

Conclusions: The accurate identification and prediction of the UD from a large number of distinct diagnosis codes and multi-source heterogeneous patient admission information in EMRs can provide a data-driven approach to assist better coding integration of diagnosis.

Keywords: unifying diagnosis, disease ontology structure, set similarity measure, clustering, electronic medical records

INTRODUCTION

In medical practice, clinicians are encouraged to seek a unifying diagnosis (UD) that could explain all the patient's signs and symptoms in preference to providing several explanations for the distress being presented (1). A UD is a critical pathway to identify the correct illness and craft a treatment plan; thus, clinical experience and knowledge play an important role in the science of diagnostic reasoning. Generally, from a brief medical history from a patient, clinicians can use the intuitive system in their brain and rapidly reason the disease types, whereas for complex and multi-type abnormal results, clinicians must use the more deliberate and time-consuming method of analytic reasoning to deduce the UD, raising the risk of diagnostic errors (2).

To increase the accuracy of a UD, enhancing individual clinicians' diagnostic reasoning skills and improving health care systems are regarded as two important approaches to support clinicians through the diagnostic process. The former requires professional knowledge training and lifelong learning, whereas the latter mainly involves the development of information technology (3). For an individual clinician, an intelligent clinical decision support system is prone to acceptable and can help clinicians to improve their unifying diagnostic decisions (4). Recently, along with the widespread adoption of electronic medical records (EMRs), an extremely large volume of electronic clinical data has been generated and accumulated (5, 6). Meanwhile, artificial intelligence and big data analytic technology have been successfully applied to clinical diagnostic procedures and treatment regimen recommendation, which has resulted in new opportunities for intelligent clinical decision support systems that use data-driven knowledge discovery methods (7–10).

From the data mining perspective, a UD aims to classify a large number of distinct diagnosis codes reasonably according to the disease taxonomy and attempt to adopt a disease to summarize or explain various clinical manifestations of the disease. Therefore, the nature of a UD is diagnosis code assignment along with disease correlation exploitation. Diagnosis code assignment refers to the clinical decision process in which supervised methods are adopted to predict and annotate disease codes based on patients' medical history, signs and symptoms, and laboratory examination (11). According to the number of diagnosis codes that patients suffer from, diagnosis code assignment can be divided into single-label (12), multi-class (13), multi-label (14), and multi-task learning methods (15). However, although many novel supervised learning models have been proposed and can achieve high performance in terms of assigning diagnosis codes for new patients using frontier supervised methods, such as ensemble learning (16), reinforcement learning (17), and deep

learning (18), they cannot further explore disease co-occurrence relations for UD identification and prediction.

The coexistence of multiple diseases is pervasive in the clinical environment, particularly for patients in the intensive care unit (ICU) (19). According to the statistical results of the MIMIC-III database, which is a freely accessible critical care database, the average number of diagnosis codes for patients in the ICU is 11. Additionally, diagnosis codes are highly fine-grained, closely related, and extremely diverse (20). For example, the patient with admission identifier (ID) 100223 is assigned to 28 ICD-9 codes, and many diagnosis codes are similar, such as 276.2 (Acidosis, order: 15), 276.0 (Hyperosmolality and/or hyponatremia, order: 18), and 276.6 (Hyperpotassemia, order: 26). Thus, it is trivial and difficult for clinicians to make a consistent, accurate, concise, and unambiguous diagnostic decision reasonably.

Furthermore, although the inter-relation of diagnosis codes was considered in previous studies, the researchers commonly used the first three digits of ICD-9 codes to assign diagnosis codes for patients (21–23); hence, the complexity may increase and prediction performance may reduce when considering all digits of the ICD-9 codes. Additionally, in those studies, reasonable complicated and confused diagnosis codes could not be classified into a UD using a data-driven method. A UD is the basic principle of clinical diagnostic thinking. Its basic idea is that when a patient has many symptoms, if these symptoms can be explained by one disease, it will never explain different symptoms using multiple diseases (1). A UD reflects the integrity of the patient and the professionalism of clinicians; however, in previous studies, the main focus was on the UD of a category of diseases from the clinical perspective, such as mood/mental disorders (24), intracranial mesenchymal tumor (25), and arrhythmogenic right ventricular cardiomyopathy (26). In this study, we fully consider the fine-grained diagnosis codes (i.e., all digits) of patients, identify the UD from a group of patient diagnostic information using an unsupervised clustering method and predict the UD for new unseen patients using multi-class learning methods.

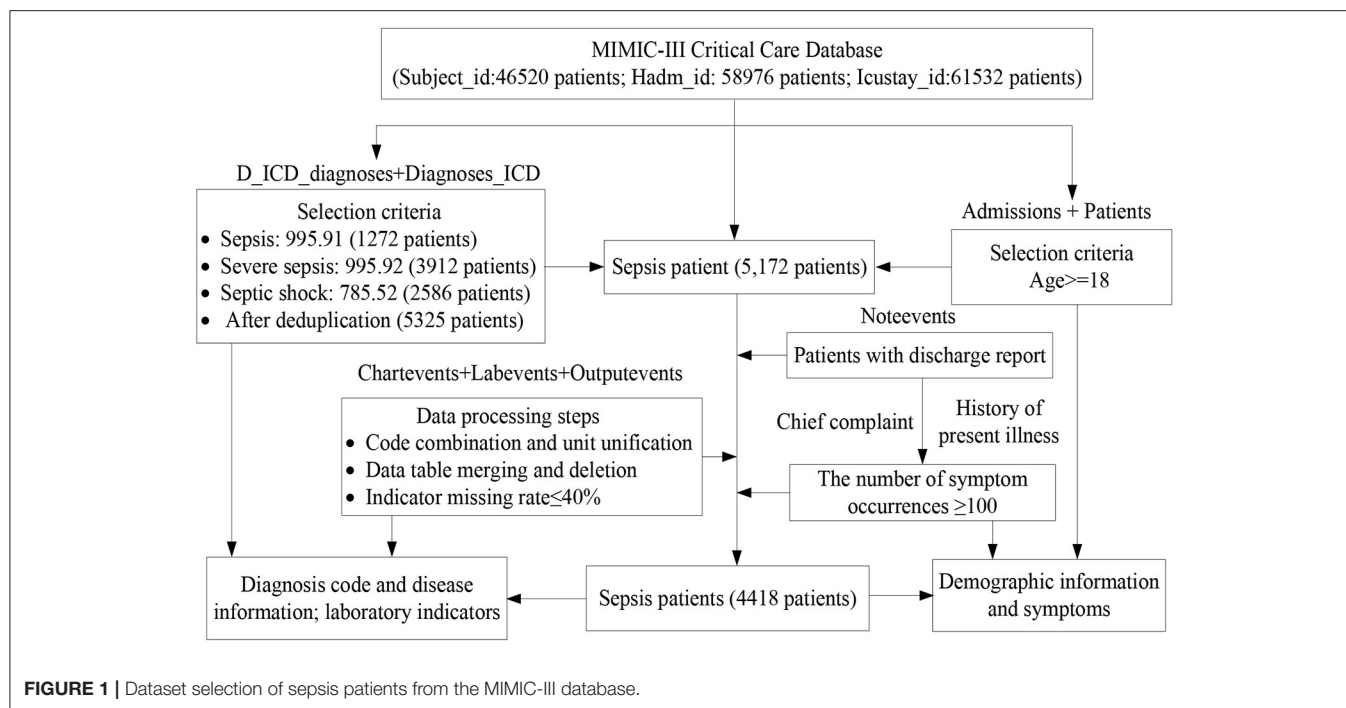
MATERIALS AND METHODS

Data Collection

We selected a dataset of sepsis patients from the MIMIC-III database, where sepsis is divided into general sepsis, severe sepsis, and septic shock (27, 28). **Figure 1** shows the detailed processes of data collection and preprocessing of sepsis patients, including the identification of sepsis patients, data extraction, data cleaning, and feature selection. Finally, we screened 4,418 sepsis patients and extracted their diagnostic information to unify the diagnosis identification, their demographic information, laboratory examination information, chief complaint, and history of present illness information, and obtain a UD prediction.

First, the diagnostic information of 4,418 sepsis patients mainly contained the patient hospital admission ID (Hadm-id), ICD-9 diagnosis code, order of diagnosis code, and a brief definition of the diagnosis codes, where the sum, maximum, minimum, and average numbers of diagnosis codes were 80501, 39, 3, and 18.3, respectively. Additionally, for the visualization,

Abbreviations: EMR, Electronic medical record; UDIPM, Unifying diagnosis identification and prediction method; CDSS, Clinical decision support system; ICU, Intensive care unit; IC, Information content; LCA, Least common ancestor; AP, Affinity propagation; SS, Sum of similarities; TDC, Typical diagnosis code; LCoP, LCA co-occurrence pattern; AOrd, Average order; TDCCoP, Typical diagnosis code co-occurrence pattern; CCoM, Conditional co-occurrence matrix; UD, Unifying diagnosis; Hadm-id, Hospital admission identifier; FM, Fusion method.

**TABLE 1 |** Feature information of the health condition of sepsis patients.

Information	Feature	Description (Range, Type)
Demographic information	Admission type	Emergency, elective, urgent (Nominal)
	Gender	Female, male (Nominal)
	Age	[18, 89] (Numeric)
Laboratory examination information	Potassium Level, PO2, serum bicarbonate level, temperature, sodium level, urine out foley, urea nitrogen, WBC, bilirubin level, GCSmotor, GCSeyes, HR, GCSverbal, NBP, RR, SPO2, hemoglobin, platelet count, creatinine	Minimum, maximum, median, mean, and variance value (Numeric)
Symptom information	Fever, abdominal pain, shortness of breath, nausea and vomiting, weakness, diarrhea, dizziness, palpitation, cough, fatigue, discomfort, dysuria, shock, weight change, loss of appetite, and night sweating	0, 1 (Nominal)
Related indicators	AIDS, hematologic malignancy, metastatic cancer	0, 1 (Nominal)
	SOFA, SAPS, and SAPS-II	Integer (Numeric)

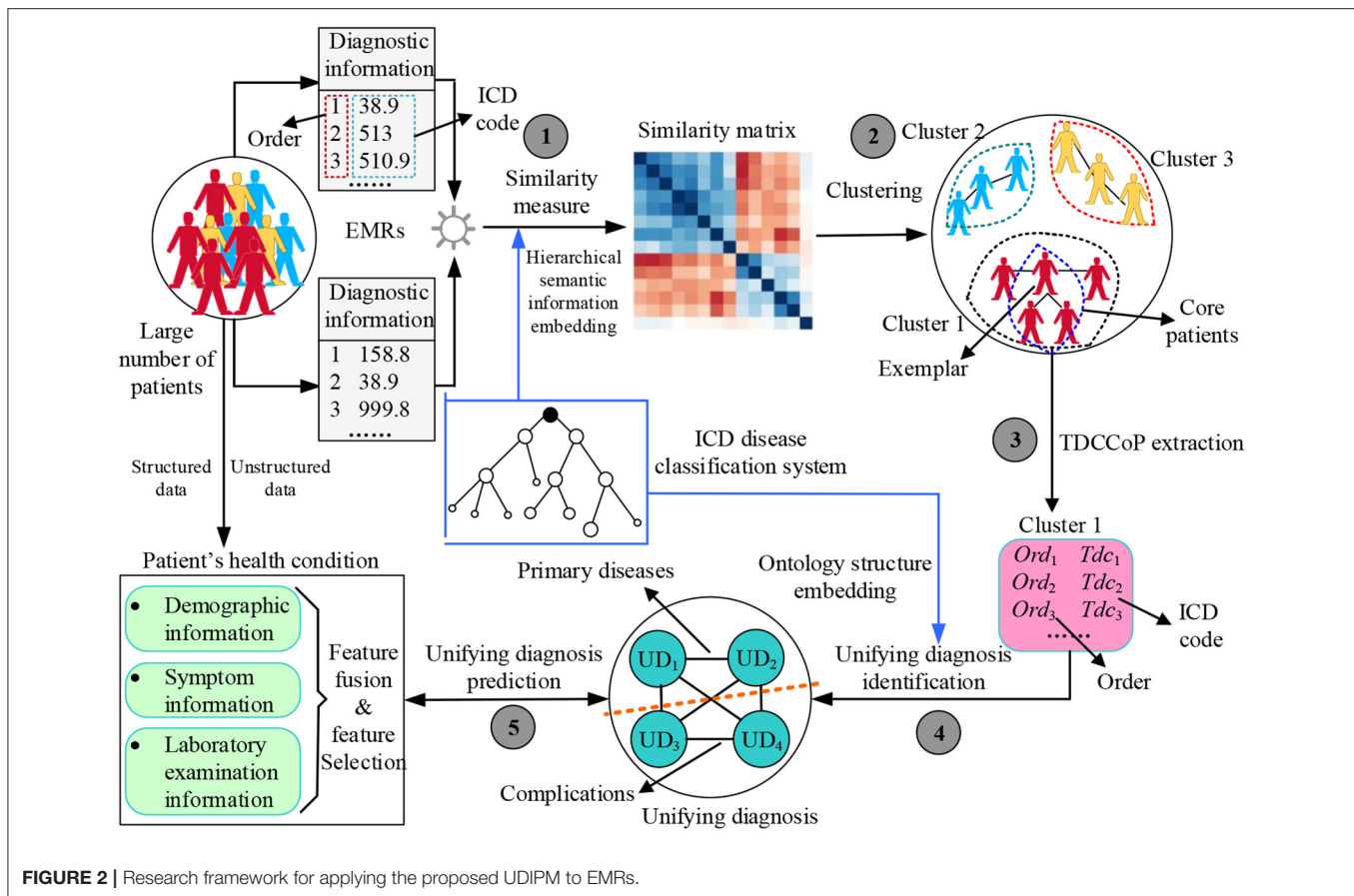
we removed duplicate diagnosis codes and converted the remaining 3,070 diagnosis codes into digital numbers from 1 to 3,070. The **Supplementary Table 1** shows the diagnostic information of two patients.

Then, for the health condition of patients admitted to hospital, we used the minimum, maximum, median, mean, and variance value as the 5-tuple features of each laboratory indicator, and designed a symptom identification method based on text analysis of patient discharge reports, including rule setting, text segmentation, text extraction, abbreviation dictionary construction, negative word recognition, case unification, word segmentation, stop word removal, and external symptom dictionary embedding (**Supplementary Figure 1**). Additionally, we added related indicators to measure patients' severity, such as AIDS, hematologic malignancy, metastatic cancer SOFA, SAPS, and SAPS-II. Finally, we obtained 120 features of the health

condition of sepsis patients in the experimental dataset, as shown in **Table 1**.

Method

Figure 2 shows the proposed UD identification and prediction method (UDIPM), which uses four types of information from EMRs. We adopt diagnostic information to identify the UD, and use demographic information, symptom information, and laboratory examination information to predict the UD. First, we apply a set of similarity measure methods to a large number of patients by embedding the semantic relation of the ICD classification system (Task 1 in **Figure 2**). Second, we apply a clustering algorithm to the similarity matrix to divide patients into different groups, and further obtain the exemplar and core patients of each cluster (Task 2 in **Figure 2**). Third, we extract the typical diagnosis code co-occurrence patterns (TDCCoP)



from each cluster by defining a threshold and a sorting function (Task 3 in **Figure 2**). Fourth, we combine the visual analysis and conditional co-occurrence matrix (CCoM) to identify the UD by selecting the optimal segmentation (Task 4 in **Figure 2**). Finally, after obtaining the health condition of the patient admitted to hospital, we obtain a UD prediction using multi-class classification methods (Task 5 in **Figure 2**).

Patient Similarity Measure Method

Many methods exist for measuring patient similarity (29, 30). In this study, considering the semantic relations of diagnosis codes in the ICD ontology structure, we adopt a set similarity measure method. First, we define patient diagnostic information as a series of ordered diagnosis codes. Then we reconstruct the ontology structure based on a disease classification system to easily measure patient similarity. Finally, we describe the process of the set similarity method, including the information content (IC) measure of diagnosis codes, diagnosis code similarity measure, and diagnosis code set similarity measure.

Patient's Diagnostic Information Representation

Diagnostic information refers to a record of disease diagnosis made by clinicians based on the health condition of a patient admitted to hospital. It is stored in the patient's EMR data in the form of a diagnosis code (e.g., ICD-9 and ICD-10). Because of the prevalence of disease complications, a patient's EMR is

typically annotated using multiple disease codes, and these codes have a certain priority (i.e., order). The higher the priority of the diagnosis code is, the more central and important the disease is for this patient, then the weaker conversely. Thus, patient diagnostic information can be represented as

$$D = \{(dc_1, \text{Ord}(dc_1)), (dc_2, \text{Ord}(dc_2)), \dots, (dc_i, \text{Ord}(dc_i)), \dots\}, \quad (1)$$

where dc_i and $\text{Ord}(dc_i)$ represent the i -th diagnosis code and its order, respectively.

Ontology Structure Construction

We automatically construct a five-level ICD-9 ontology structure, shown in **Figure 3**, in which level-0 is the virtual root node, level-1 has 19 chapters, level-2 has 129 sections, level-3 has ~1,300 categories (**Supplementary Figure 2**), and the last two levels are expanded to 10 types of sub-nodes under each node. For example, level-4 contains 550.0, 550.1, 550.2 (virtual code), 550.3 (virtual code), ... and 550.9, and level-5 includes 550.10, 550.11, 550.12, 550.13, 550.14 (virtual code), ... 550.19 (virtual code). More importantly, the actual diagnosis codes of patients belong to the ICD-9 ontology structure, whereas the virtual codes are only used to construct a complete ICD ontology structure and do not play a role in the actual similarity measure.

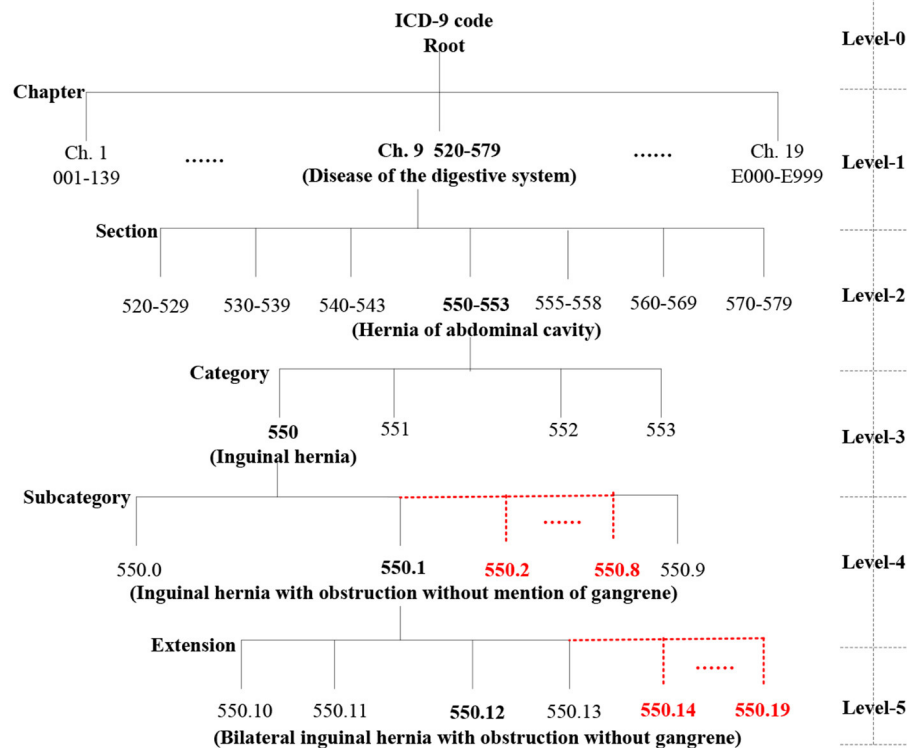


FIGURE 3 | Local ontology structure of ICD-9 codes.

Set Similarity Measure

Information Content Measure of Diagnosis Codes

In the ICD-9 ontology structure, each code represents a concept, and there is semantic similarity between classification concepts. Additionally, concepts on the same branch are more similar than those on different branches. Thus, we use the level depth measure method of the hierarchical tree (29), that is, we assign a value to each level of the ICD-9 ontology structure; the deeper the concept level, the larger the value. For an ICD-9 code dc_i , the IC is defined as

$$IC(dc_i) = \text{level}(dc_i \rightarrow \text{Root}), \quad (2)$$

where *Root* is the virtual root node and the function *level*(.) denotes the level depth from the ICD-9 code dc_i to the root node. Intuitively, the IC of the root node (level-0) is 0, the ICs of a chapter (level-1), section (level-2), category (level-3), subcategory (level-4), and extension (level-5) are 1, 2, 3, 4, and 5, respectively.

Code-Level Similarity Measure

For the IC of codes, there are several approaches to measure code-level similarity. We use the least common ancestor (LCA) of two codes to measure the similarity of diagnosis codes, defined as

$$s(dc_i, dc_j) = \frac{2IC(LCA(dc_i, dc_j))}{IC(dc_i) + IC(dc_j)}, \quad (3)$$

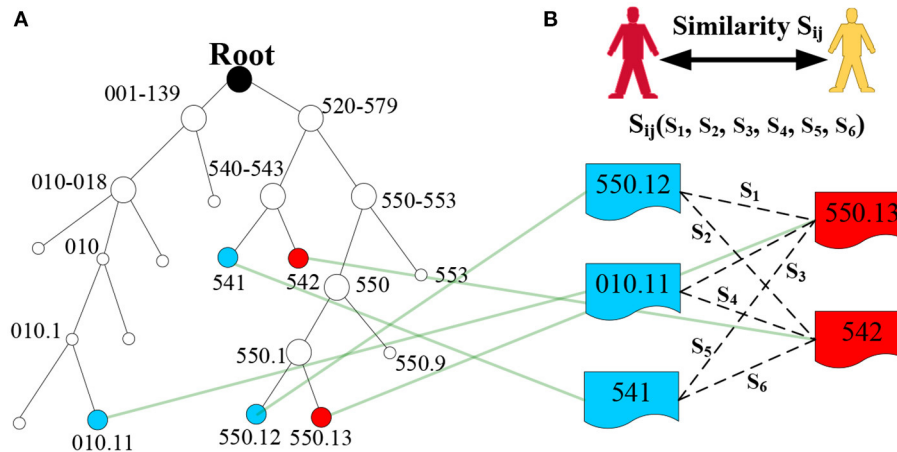
where dc_i and dc_j are two diagnosis codes, and $LCA(dc_i, dc_j)$ is the LCA of dc_i and dc_j . If $dc_i = dc_j$, then $LCA(dc_i, dc_j) = dc_i = dc_j$, and $IC[LCA(dc_i, dc_j)] = IC(dc_i) = IC(dc_j)$. If $dc_i \neq dc_j$ and $LCA(dc_i, dc_j) = \text{Root}$, then $IC[LCA(dc_i, dc_j)] = 0$.

To make this concept easier to understand, we provide a simple example in **Figure 4A**. Thus, $LCA(550.12, 550.13) = 550.1$, $LCA(541, 550.13) = 520-579$, $s = s_1(550.12, 550.13) = 2IC(550.1)/[IC(550.12) + IC(550.13)] = 2 * 4/(5 + 5) = 0.8$.

Code Set-Level Similarity Measure

In the EMR dataset, patient diagnostic information is typically a set of diagnosis codes. Thus, patient similarity can be transformed into the similarity of the diagnosis code set. Generally, for binary code-level similarity, we can use classical methods, such as Dice, Jaccard, cosine, and overlap, to calculate set-level similarity. However, these methods cannot fully embed semantic similarity. Thus, we use the most similar concept pair's average value to measure the set-level similarity (29), and the formula is defined as

$$S(D'_i, D'_j) = \frac{(\sum_{dc_{ig} \in D'_i} \min_{dc_{jh} \in D'_j} (1 - s(dc_{ig}, dc_{jh})) + \sum_{dc_{jh} \in D'_j} \min_{dc_{ig} \in D'_i} (1 - s(dc_{jh}, dc_{ig})))}{1 - \frac{|D'_i| + |D'_j|}{|D|}}, \quad (4)$$



where \mathbf{D}'_i and \mathbf{D}'_j are the diagnostic information of patient i and patient j , respectively, which does not consider the order of diagnosis codes; that is, $\mathbf{D}'_i = \{dc_{i1}, dc_{i2}, \dots, dc_{ig}, \dots\}$ and $\mathbf{D}'_j = \{dc_{j1}, dc_{j2}, \dots, dc_{jh}, \dots\}$. $|\mathbf{D}'_i|$ and $|\mathbf{D}'_j|$ are the number of diagnosis codes for patient i and patient j , and dc_{ig} and dc_{jh} are the g -th diagnosis code of patient i and the h -th diagnosis code of patient j , respectively. Finally, we obtain the similarity \mathbf{S}_{ij} of the two patients (**Figure 4B**), and similarity matrix \mathbf{S} for all patients in the EMRs using a set similarity measure method. The pseudocode of the patient similarity measure method is presented in **Algorithm 1**.

Algorithm 1 | Patient similarity measure method.

Input: $D'_i = \{dc_{i1}, dc_{i2}, \dots, dc_{iq}, \dots\}$, $i = 1, 2, \dots, N$

Output: Similarity matrix $\mathbf{S}_{N \times N}$

1. Construct the ICD ontology structure
2. **For** $i = 1 : N$ **do**
For $j = i + 1 : N$ **do**
 Compute $IC(dc_{i1}, dc_{i2}, ..., dc_{in}, ...)$, $IC(dc_{j1}, dc_{j2}, ..., dc_{jn}, ...)$, and diagnosis code similarity $s(dc_{ig}, dc_{jg}) = 2IC(LCA(dc_{ig}, dc_{jg})) / (IC(dc_{ig}) + IC(dc_{jn}))$ based on the ICD ontology structure, compute set similarity $S(\mathbf{D}_i, \mathbf{D}_j)$ using Eq. 4
3. Obtain the similarity matrix $\mathbf{S}_{N \times N}$ for N patients

Patient Clustering Algorithm

A clustering algorithm aims to divide patients into multiple groups based on the similarity matrix \mathbf{S} , requiring that patients in the same group are as similar as possible, and patients in different groups are as dissimilarity as possible (31, 32). In this study, considering the advantages, such as not predefining the number of clusters, the real existence of exemplars, and much lower error, we adopt affinity propagation (AP) clustering (33, 34).

AP clustering determines the number of clusters by controlling the input exemplar preferences (p), where p is more robust than K because p monotonically controls the perception granularity. Generally, p depends on the similarity matrix $\mathbf{S}_{N \times N}$, number of input patients (N), and p coefficient (p_{coe}), which is represented as

$$p = median(\mathbf{S}) - p_{coe} * N. \quad (5)$$

After patients are clustered, we identify K clusters (C_1, C_2, \dots, C_K), and define the popularity (i.e., support) of each cluster as

$$Support(C_k) = \frac{\sum_{j \in \{1, 2, \dots, N\}} \lambda(C(D'_j), E(C_k))}{N}, k = 1, 2, \dots, K, (6)$$

where $C(D_j')$ represents the cluster to which patient j belongs and $E(C_k)$ denotes the exemplar of C_k . $\lambda(\cdot)$ is an indicator function; if patient j belongs to C_k , then $\lambda[C(D_j'), E(C_k)] = 1$; otherwise, $\lambda[C(D_j'), E(C_k)] = 0$.

Additionally, we obtain the sum of similarities (SS), which is an important indicator used to evaluate clustering performance. The SS depends on the similarity matrix $\mathbf{S}_{N \times N}$, number of input patients (N), number of clusters (K), and corresponding exemplars, which is represented as

$$SS(K) = \sum_{i=1}^K \sum_{D'_i \in C_i} S(D'_i, E(C_i)). \quad (7)$$

Generally, the larger the SS value, the better the clustering performance. The pseudocode of the patient clustering algorithm is presented in **Algorithm 2**.

TDCCoP Extraction Method

In our previous studies, we proved that defining the core zone of a cluster is an effective approach to extract stable clustering results (35). Additionally, considering the complex semantic relations

Algorithm 2 | Patient clustering algorithm.**Input:** $\mathbf{S}_{N \times N}$, ρ_{coe} , step size ϵ **Output:** Optimal clustering number K^* , $E(\mathbf{C}_k)$, support (\mathbf{C}_k) , $SS(K^*)$

1. Initialize $\mu = 1$, $\rho_{coe}(\mu) = \rho_{coe} = 0$, ϵ
2. Run the AP clustering algorithm with $\mathbf{S}_{N \times N}$ and ρ ($\rho = \text{median}(\mathbf{S}) - \rho_{coe}(\mu) * N$)
3. Return the clustering number $K(\mu)$
4. **While** $K(\mu) < N$ and $K(\mu) > 1$ **do**
 $\mu = \mu + 1$, $\rho_{coe}(\mu) = \rho_{coe}(\mu - 1) + \epsilon$
 $\rho = \text{median}(\mathbf{S}) - \rho_{coe}(\mu) * N$
 Run the AP clustering algorithm with $\mathbf{S}_{N \times N}$ and ρ
 Return the clustering number $K(\mu)$ and $\rho_{coe}(\mu)$
5. Compute the distance $dp_{coe}(K) = \max[\rho_{coe}(\mu_i)] - \min[\rho_{coe}(\mu_j)]$ for the same K
6. Return the maximum $dp_{coe}(K)$ and the optimal clustering number K^*
7. Set $\rho_{coe} = 0.5 * \{\max[\rho_{coe}(\mu_i)] + \min[\rho_{coe}(\mu_j)]\}$ for K^*
8. Run the AP clustering algorithm with $\mathbf{S}_{N \times N}$ and ρ ($\rho = \text{median}(\mathbf{S}) - \rho_{coe} * N$)
9. Return $E(\mathbf{C}_k)$, support (\mathbf{C}_k) using Eq. 6, and $SS(K^*)$ using Eq. 7

among different diagnosis codes, the feature of a cluster cannot be fully described when the diagnostic information (cluster center or exemplar) of only one patient is used. Thus, we also define the core zone of each cluster to select a group of patients (i.e., core patients) using the k -nearest neighbor method, and further extract typical diagnosis codes (TDCs). For cluster \mathbf{C}_k , the core zone is defined as

$$\text{Core}_k = \{D'_j | S(D'_j, E(\mathbf{C}_k)) \geq \tau\}, \quad (8)$$

where $E(\mathbf{C}_k)$ is the exemplar of cluster \mathbf{C}_k and τ is a similarity threshold defined in advance, which aims to determine the number of core patients.

Then, for cluster \mathbf{C}_k , the occurrence probability of the diagnosis code dc_h can be represented as

$$\text{Prob}_k(dc_h) = \frac{\sum_{D'_j \in \text{Core}_k} \lambda(dc_h, D'_j)}{|\text{Core}_k|}, h = 1, \dots, H, \quad (9)$$

where $|\text{Core}_k|$ denotes the number of core patients in cluster \mathbf{C}_k . $\lambda(\cdot)$ is an indicator function; if the diagnostic information D'_j of patient j contains diagnosis code dc_h , then $\lambda(dc_h, D'_j) = 1$; otherwise, $\lambda(dc_h, D'_j) = 0$. H is the number of all diagnosis codes after duplicates are deleted.

After we calculate the probability of all diagnosis codes in the cluster \mathbf{C}_k , we define the TDC as

$$\text{Tdc}_h = \{dc_h | \text{Prob}_k(dc_h) > \delta_1\}, \quad (10)$$

where δ_1 is a threshold defined in advance to differentiate high-frequency and low-frequency diagnosis codes.

Based on all TDCs of the cluster \mathbf{C}_k , we further analyze the priority of TDCs by embedding the order of the patient

diagnostic information, that is, for patient j , $D_j = \{[dc_{j1}, \text{Ord}(dc_{j1})], [dc_{j2}, \text{Ord}(dc_{j2})], [dc_{jh}, \text{Ord}(dc_{jh})], \dots\}$ and $D'_j = \{dc_{j1}, dc_{j2}, dc_{jh}, \dots\}$. Thus, the average order (AOrd) of TDC Tdc_h is defined as

$$\text{AOrd}(\text{Tdc}_h) = \frac{\sum_{D'_j \in \text{Core}_k, \text{Tdc}_h \in D'_j} \text{Ord}_{D_j}(\text{Tdc}_h) \lambda(\text{Tdc}_h, D'_j)}{\sum_{D'_j \in \text{Core}_k, \text{Tdc}_h \in D'_j} \lambda(\text{Tdc}_h, D'_j)}, \quad h = 1, \dots, H', \quad (11)$$

where H' is the number of TDCs in cluster \mathbf{C}_k and $\text{Ord}_{D_j}(\text{Tdc}_h)$ denotes the order of TDC Tdc_h in the diagnostic information D_j of patient j . Generally, the smaller the AOrd of typical diagnostic codes, the more likely they are to be primary diseases.

Finally, after obtaining TDCs and their AOrd, we define a sorting function to determine TDCCoP, which is represented as

$$\begin{aligned} \text{TDCCoP}_k &= \text{Sort}((\text{Tdc}_1, \text{AOrd}_k(\text{Tdc}_1)), \dots, (\text{Tdc}_{H'}, \text{AOrd}_k(\text{Tdc}_{H'}))) \\ &= \{(\text{Tdc}_1, \text{Ord}'(\text{Tdc}_1)), \dots, (\text{Tdc}_{H'}, \text{Ord}'(\text{Tdc}_{H'}))\}, \quad (12) \end{aligned}$$

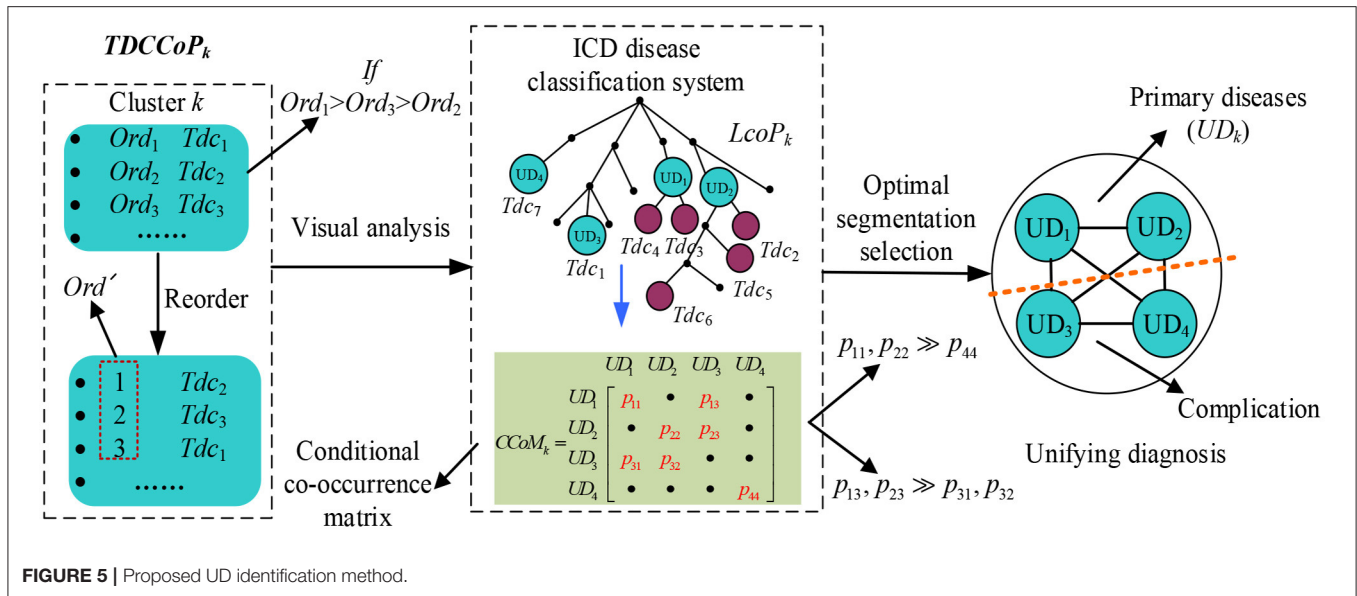
where $\text{Ord}'(\text{Tdc}_h)$ is the new order of Tdc_h . For example, if cluster \mathbf{C}_k has only three TDCs (e.g., Tdc_1 , Tdc_2 , and Tdc_3) and its AOrd are 5.3, 7.8, and 3.8, respectively, then after sorting, the TDCCoP_k is $\{(\text{Tdc}_3, 1), (\text{Tdc}_1, 2), (\text{Tdc}_2, 3)\}$. The pseudocode of the TDCCoP extraction method is presented in **Algorithm 3**.

Algorithm 3 | TDCCoP extraction method.**Input:** \mathbf{C}_k , $E(\mathbf{C}_k)$, Core_k , \mathbf{D}_j , \mathbf{D}'_j , H , $k = 1, 2, \dots, K$ **Output:** TDCCoP_k , $k = 1, 2, \dots, K$

1. Initialize $k = 1$, τ , $h = 1$, δ_1 , $i = 1$
2. **For** $k = 1: K$ **do**
 $\text{Core}_k = \{D'_j | S(D'_j, E(\mathbf{C}_k)) \geq \tau\}$
For $h = 1: H$ **do**
 $\text{Prob}_k(dc_h) = \sum_{D'_j \in \text{Core}_k} \lambda(dc_h, D'_j) / |\text{Core}_k|$
If $\text{Prob}_k(dc_h) > \delta_1$ **then**
 $\text{Tdc}_h \leftarrow dc_h$
 $i = i + 1$
 $H' = i$
For $h = 1: H'$ **do**
 $\text{AOrd}_k(\text{Tdc}_h) = \sum_{\text{Tdc}_h \in D'_j} \text{Ord}(\text{Tdc}_h) / |D'_j|$
 $\text{TDCCoP}_k = \text{Sort}((\text{Tdc}_1, \text{AOrd}_k(\text{Tdc}_1)), \dots, (\text{Tdc}_{H'}, \text{AOrd}_k(\text{Tdc}_{H'})))$
3. Return TDCCoP_k , $k = 1, 2, \dots, K$

UD Identification Method

To identify a UD, categorizing the TDCCoP of each cluster reasonably according to the disease taxonomy is a critical step. In this study, we propose a UD identification method, as shown in **Figure 5**. Specifically, for the TDCCoP_k of cluster k , we first visualize all TDCs in the reconstructed ICD ontology structure, and mark their orders. Then we use the LCA method to categorize these codes, and define their LCA and the corresponding orders. Furthermore, we calculate the CCoM



using patient diagnostic information to select the optimal segmentation between primary diseases and complications. Finally, we regard the identified primary diseases as the UD.

First, we define the LCA co-occurrence pattern (LCoP) of the $TDCCoP_k$ using visual analysis of the ICD ontology structure as

$$LCoP_k = \{d_i | d_i = LCA_{\{Tdc_1, Tdc_2, \dots\}}(Tdc_1, Tdc_2, \dots), d_i \neq Root\}. \quad (13)$$

Then we calculate the order of each d_i in $LCoP_k$ as

$$Ord(d_i) = \min_{d_j = LCA(Tdc_1, Tdc_2, \dots, Tdc_m)} (Ord'(Tdc_1), Ord'(Tdc_2), \dots, Ord'(Tdc_m)), \quad (14)$$

where m is the number of TDCs in $LCoP_k$ whose LCA is d_i .

Additionally, considering the causal relation between d_i and d_j in $LCoP_k$, we define the conditional co-occurrence probabilities $p_k(d_j/d_i)$ and $p_k(d_i/d_j)$ as

$$\begin{aligned} p_k(d_j/d_i) &= Freq_k(d_j, d_i) / Freq_k(d_i) \\ p_k(d_i/d_j) &= Freq_k(d_i, d_j) / Freq_k(d_j) \end{aligned} \quad (15)$$

where $Freq_k(d_i, d_j)$ and $Freq_k(d_j, d_i)$ denote the number of co-occurrences of d_i and d_j , respectively, and $Freq_k(d_i)$ denotes the number of occurrences of d_i in the cluster C_k .

Thus, for all diagnosis codes in $LCoP_k$, we generate a CCoM $CCoM_k$, where $CCoM_k(i, j) = p_k(d_j/d_i)$, $CCoM_k(j, i) = p_k(d_i/d_j)$, and the diagonal entry $CCoM_k(i, i) = p_k(d_i) = Freq_k(d_i) / |Core_k|$. If $CCoM_k(i, j) \gg CCoM_k(j, i)$ or $CCoM_k(i, i) \gg CCoM_k(j, j)$ exist, then d_j is more prone to occur after the occurrence of d_i ; thus, d_i is more likely to be a primary disease, whereas d_j will become a complication, and vice versa.

After analyzing the precedence relation of all diagnosis codes in $LCoP_k$ using $CCoM_k$, we obtain the optimal segmentation

between primary diseases and complications, and define the UD of cluster k as

$$UD_k = \{d_i | d_i \in LCoP_k, d_i \neq \text{Complication}\}, \quad (16)$$

where UD_k is a set of primary diseases. The pseudocode of the UD identification method is presented in **Algorithm 4**.

Algorithm 4 | UD identification method.

Input: $TDCCoP_k$, $Core_k$, \mathbf{D}_j , \mathbf{D}'_j , $k = 1, 2, \dots, K$, ICD ontology structure

Output: UD_k , $k = 1, 2, \dots, K$

1. Initialize $i = 1$, call the ICD ontology structure

2. **For** $k = 1: K$ **do**

While $Tdc \in TDCCoP_k$ **do**

$d_i = LCA(Tdc_1, Tdc_2, \dots)$

$Ord(d_i) = \min(Ord'(Tdc_1), Ord'(Tdc_2), \dots)$

$i = i + 1$

$i \leftarrow i$

For $i_1 = 1: i$ **do**

For $i_2 = i_1 + 1: i$ **do**

$p_k(d_{i_1}) = \sum_{d_{i_2} = LCA(Tdc_1, \dots, Tdc_g)} \lambda(Tdc_g, \mathbf{D}'_j) / |Core_k|$

$p_k(d_{i_2}) = \sum_{d_{i_1} = LCA(Tdc_1, \dots, Tdc_h)} \lambda(Tdc_h, \mathbf{D}'_j) / |Core_k|$

$p_k(d_{i_2}/d_{i_1}) = (\sum_{d_{i_1} = LCA(\dots, Tdc_g), d_{i_2} = LCA(\dots, Tdc_h)} \lambda(Tdc_g, Tdc_h, \mathbf{D}'_j) / |Core_k|) / p_k(d_{i_1})$

$p_k(d_{i_1}/d_{i_2}) = (\sum_{d_{i_1} = LCA(\dots, Tdc_g), d_{i_2} = LCA(\dots, Tdc_h)} \lambda(Tdc_g, Tdc_h, \mathbf{D}'_j) / |Core_k|) / p_k(d_{i_2})$

If $p_k(d_{i_1}) \gg p_k(d_{i_2}) \parallel p_k(d_{i_2}/d_{i_1}) \gg p_k(d_{i_1}/d_{i_2}) \parallel Ord(d_{i_1}) << Ord(d_{i_2})$ **then**

$UD_k \leftarrow d_{i_1}$

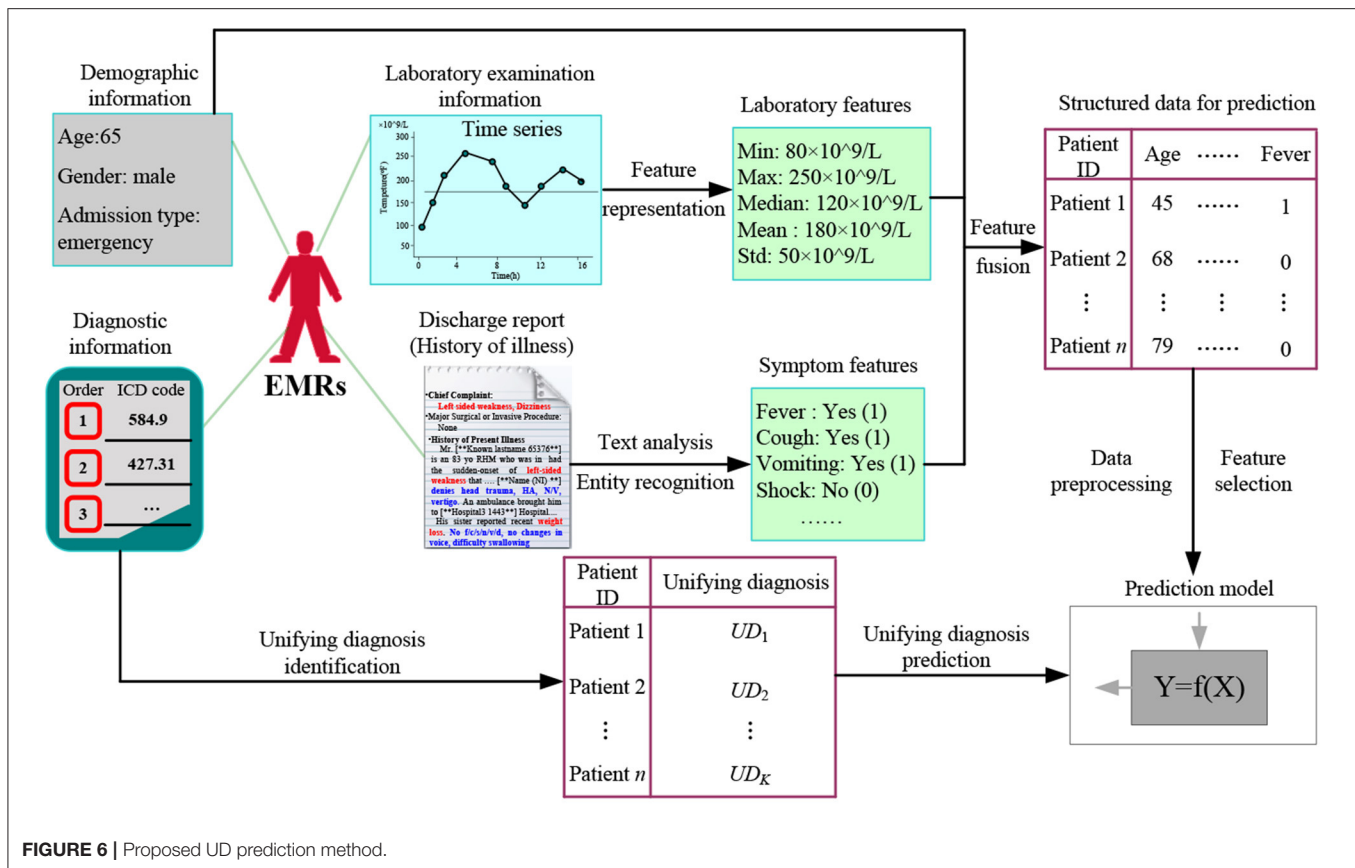
Else

$UD_k \leftarrow d_{i_2}$

3. **Return** UD_k , $k = 1, 2, \dots, K$

UD Prediction Method

After identifying the UD, we further study the prediction task based on the health condition of a patient admitted to hospital,



exploring the important features to assign the most possible UD to new patients. **Figure 6** shows the proposed UD prediction method. First, we extract three categories of features using time series feature representation and text analysis methods, and fuse them in structured data for further prediction. Then after data pre-processing and feature selection, we label all patients with a UD. Finally, we adopt classical prediction models to perform the UD prediction task.

Patient's Health Condition Representation

The health condition of a patient admitted to hospital includes demographic information, symptom information, and laboratory examination information, which play crucial roles for clinicians in diagnosing disease types, evaluating disease severity, and designing a treatment regimen.

Demographic Information

Demographic information mainly includes the date of birth, age, gender, admission type, marital status, occupation, and residence, defined as

$$De = \{De^{Age}, De^{Gender}, De^{Admission\ Type}, De^{Marital\ Status}, \dots\} \quad (17)$$

Symptom Information

Symptom information is recorded in the chief complaint and history of present illness in the form of text, where the chief complaint is the most painful part of the disease process,

including the main symptoms and onset time. The history of present illness describes the entire process for the patient after suffering from diseases, including occurrence, development, evolution, diagnosis, and treatment. Thus, the patient's symptom information can be represented as

$$Sy = \{Sy^{Fever}, Sy^{Weakness}, Sy^{Diarrhea}, \dots\} \quad (18)$$

Laboratory Examination Information

Laboratory examination refers to an indirect judgment of the health condition as a result of measuring specific components of blood and body fluids using instruments. Laboratory indicators typically have the characteristics of a time series, particularly for patients in the ICU. Thus, we use the minimum value, maximum value, median value, mean value, and variance of laboratory indicators to represent the time series, defined as

$$LE = \{(\min(LE^{WBC}), \max(LE^{WBC}), \text{med}(LE^{WBC}), \text{mean}(LE^{WBC}), \text{var}(LE^{WBC})), \dots\} \quad (19)$$

Finally, we obtain the health condition of a patient admitted to hospital using a feature fusion method, that is, $X = \{De; Sy; LE\}$.

Information Gain-Based Feature Selection

Before predicting the UD, to remove noisy data, reduce the complexity and dimensionality of the dataset, and achieve

accurate results, it is essential to apply feature selection methods to identify useful features. Therefore, feature selection is an important step that improves the clarity of the data and decreases the training time of prediction models (4). In this study, we use the information gain (IG) method to measure the importance of features and eliminate some irrelevant features. Then we compute the IG of feature x_i as

$$\begin{aligned} IG(x_i) &= H(Y) - H(Y/x_i) \\ &= -\sum_{k=1}^K P(y_k) \log P(y_k) + \sum_{k=1}^K P(y_k/x_i) \log P(y_k/x_i), \end{aligned} \quad (20)$$

where feature $x_i \in X$, $Y = \{UD_1, \dots, UD_k, \dots, UD_K\}$, $y_k \in Y$, $H(Y)$, and $H(Y/x_i)$ denote the information entropy and conditional information entropy given feature x_i for a UD classification, and $P(y_k)$ and $P(y_k/x)$ denote the probability of y_k and condition probability of y_k given feature x_i , respectively.

Thus, we obtain the important features as

$$X' = \{x_i | IG(x_i) > \delta_2\}, \quad (21)$$

where δ_2 is a threshold defined in advance to differentiate the important and unimportant features using the IG method.

Prediction Model Establishment

After obtaining the feature representation and UD result of each patient, we generate a standard dataset (Y and X') and establish a prediction model [$Y = f(X')$]. In this study, we apply five classifiers to achieve a UD prediction: logistic regression, decision tree, random forest, SVM, and extreme gradient boosting (XGBoost). In the prediction process, we adopt the Z-fold cross-validation (CV) method, which randomly partitions the initial dataset into Z mutually exclusive subsets, and perform training and testing Z times. We set Z to 5 or 10. Then we compute the average CV error to determine the prediction model as

$$CVError_Z = \frac{1}{Z} \sum_{z=1}^Z L_z = \frac{1}{Z} \sum_{z=1}^Z \frac{1}{m_z} \sum_{j=1}^{m_z} (\hat{y}_j - y_j)^2, \quad (22)$$

where L_z and m_z are the average CV error and number of the z-th testing dataset, and y_j and \hat{y}_j are the real and predicted UDs of the j-th patient, respectively.

Additionally, we identify distinctive features of different unifying diagnoses by analyzing the feature importance ranking results.

Parameter Setting

In our experiment, we set 5 parameters in advance. First, we set p_{coe} in Eq. 5 to select the number of clusters, and then τ in Eq. 8, which is a similarity threshold to determine the number of core patients (i.e., |Core|). We discuss both parameters based on the stability of the experimental results. We set δ_1 in Eq. 10 to 0.3 to obtain TDCs, and δ_2 in Eq. 21 to 0.005 to select the important features. We set the last parameter Z in Eq. 22 to 10 to perform the 10-fold CV method. In particular, before UD prediction, we

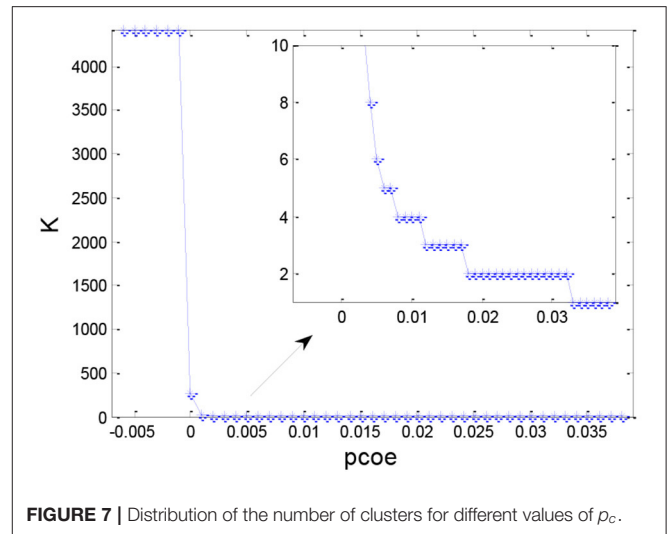


FIGURE 7 | Distribution of the number of clusters for different values of p_c .

used data pre-processing methods, that is, data normalization and smoothing for imbalanced classes.

RESULTS

Selection of the Cluster Number

After obtaining the set similarity measure based on the ontology structure for 4,418 sepsis patients, we obtained the similarity matrix S and used the AP clustering algorithm to divide all the patients into multiple groups. Figure 7 shows the distribution of the number of clusters under different values of p_c . Generally, the number of clusters decreased as the preference coefficient increased. The most stable number of clusters was two when p_c ranged from 0.018 to 0.032. Thus, we selected two clusters ($p_c = 0.025$) to identify TDCs and extract TDCoPs from each cluster.

Stability Analysis of TDCs

After applying the AP clustering algorithm, we first divided the 4,418 sepsis patients into two clusters, where cluster 1 and 2 contained 1,391 and 3,027 patients with a support of 31.48% and 68.52%, respectively. Then we analyzed the stability of the TDCs in Eq. 10 using a set of different numbers of core patients in Eq. 8 (|Core|=100, 200, 400, 500, 800, and all patients), as shown in Figure 8, Supplementary Figure 3.

From the distribution of TDCs in Figure 8, Supplementary Figure 3, the results showed that the stable range of core patients was from 400 to 800 (five codes in cluster 1 and 12 codes in cluster 2) because the number of TDCs and their distributions were approximately coincident. Specifically, compared with the stable TDCs, more TDCs were identified when the number of core patients was set to 100 and 200 (14 codes in cluster 2), such as the digital number 71 (276, disorders of fluid electrolyte and acid-base balance) and digital number 490 [V58.610, long-term (current) use of anticoagulants] (Supplementary Figures 3A,B). Digital number 99 (995.91, sepsis) was identified in cluster 1, and another three codes (486, 276.2, and 250) were not identified in cluster 2

(Supplementary Figure 3E) when we used all patients in the two clusters to extract TDCs. Thus, in the next experiment, we set the number of core patients to 800 to extract the TDCCoPs.

TDCCoP Extraction From Each Cluster

Using the clustering results, we finally determined two clusters, selected 800 core patients from each cluster, and set δ to 0.3 in Eq. 10 to identify TDCs and extract TDCCoPs. Figure 9 shows the co-occurrence relation and AOrd of all TDCs in two TDCCoPs, and Table 2 provides a detailed description of all TDCs in the two TDCCoPs.

To summarize, the experimental results indicated that there were 12 types of TDCs in the two TDCCoPs, where TDCCoP₁ and TDCCoP₂ had 5 and 12 codes, respectively. Specifically, the two TDCCoPs had similarities and differences. There were

three similarities: (1) Five types of TDCs were the same, that is, 518.81, 38.9, 785.52, 584.9, and 995.92. (2) The AOrd of all TDCs in the same TDCCoPs were similar, for example, the AOrd of four TDCs in TDCCoP₁ were all below 6, whereas those of the TDCs in TDCCoP₂ were over 7. (3) The TDCs 38.9 (septicemia), 785.52 (septic shock), and 995.92 (severe sepsis) had the highest occurrence probability in the two TDCCoPs. There were also three differences: (1) TDCCoP₂ identified more TDCs than TDCCoP₁. (2) The occurrence probabilities of TDCs in TDCCoP₁ were larger than those in TDCCoP₂. (3) The AOrd of the same TDC were different in the two TDCCoPs, for example, 518.81 (acute respiratory failure) in the two TDCCoPs was 4.145 and 7.665, respectively. Additionally, septicemia (38.9) was a high-frequency and primary disease in sepsis patients, which is a life-threatening complication that can occur when bacteria from another infection enters the blood and spreads throughout the body.

Furthermore, using Eq. 12 and Algorithm 3, we extracted the TDCCoPs of the two clusters described in Table 2, that is, TDCCoP₁ = {(38.9, 1), (785.52, 2), (518.81, 3), (584.9, 4), (995.92, 5)} and TDCCoP₂ = {(584.9, 1), (38.9, 2), (518.81, 3), (599.0, 4), (428.0, 5), (486.0, 6), (401.9, 7), (785.52, 8), (276.2, 9), (995.92, 10), (427.31, 11), (250.0, 12)}. Thus, from a reordering perspective, acute kidney failure, septicemia, and acute respiratory failure were probably the primary diseases in the two TDCCoPs.

UD Identification Based on TDCCoPs

After obtaining TDCCoPs, we visualized all the TDCs in the ICD-9 ontology structure. First, we categorized them using the LCA method to identify LCoPs using Eq. 13. Consider TDCCoP₂ as an example. The visualization result is shown in Figure 10. Clearly, we identified LCoP₂ with seven types of diseases, which are light green color, and computed the order of the new diseases using Eqs 13, 14: diseases of the genitourinary system (580–629, order: 1), septicemia (38.9, order: 2), diseases of the respiratory

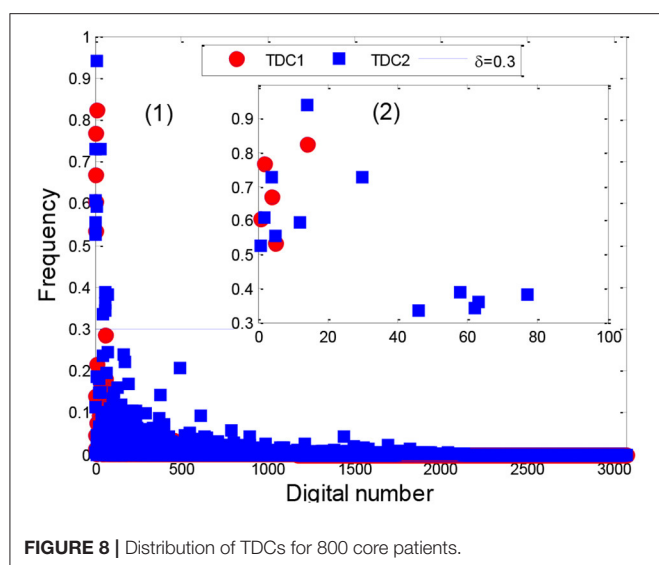


FIGURE 8 | Distribution of TDCs for 800 core patients.

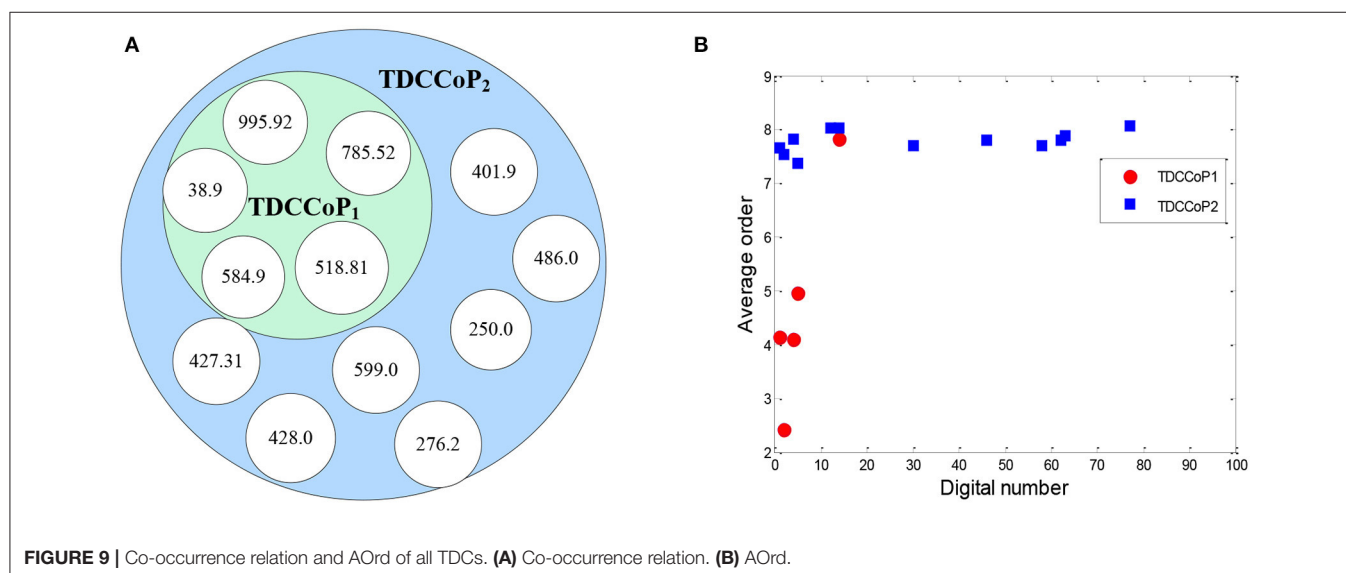
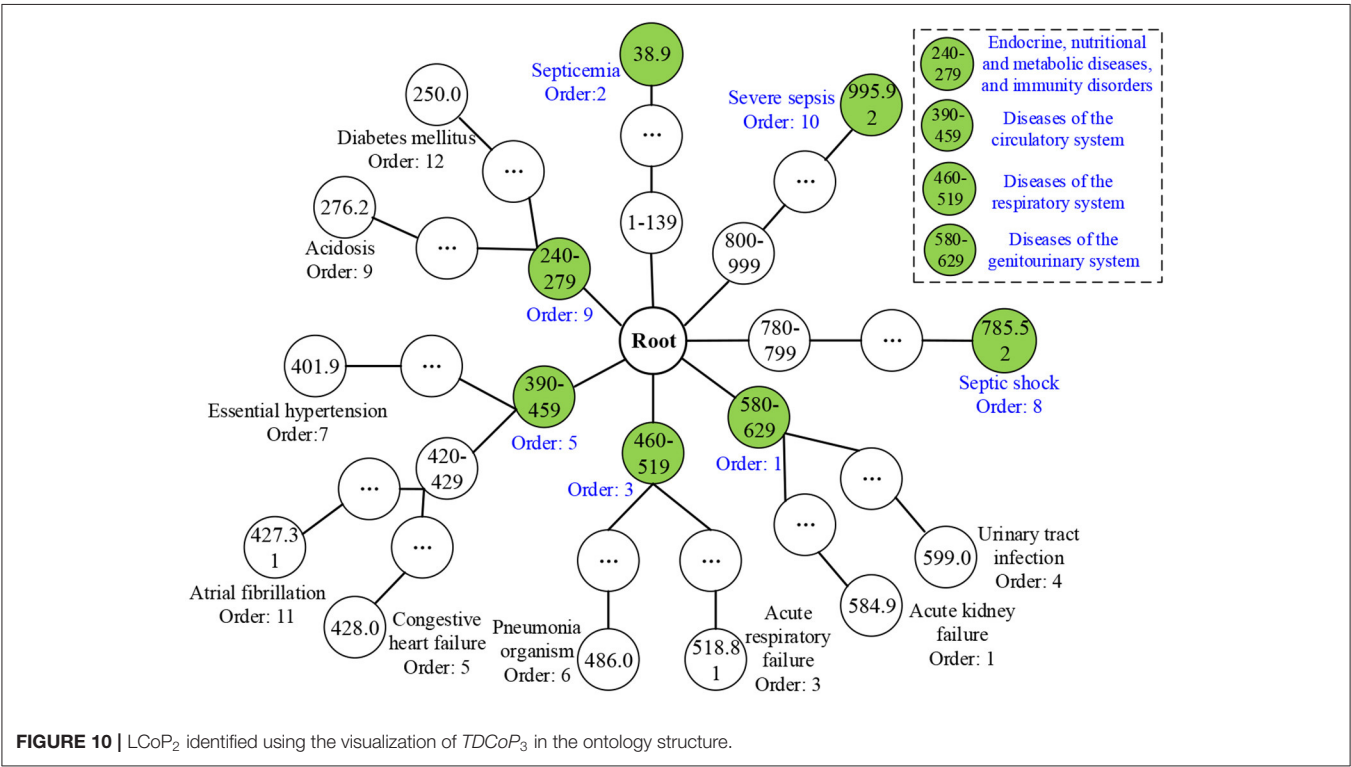


FIGURE 9 | Co-occurrence relation and AOrd of all TDCs. (A) Co-occurrence relation. (B) AOrd.

TABLE 2 | Detailed description of three TDCs.

TDCCOP	Digital number	TDC	Definition of diagnosis code	Occurrence frequency	Average order	Re-order
TDCCOP ₁ (1391)	1	518.81	Acute respiratory failure	0.604	4.145	3
	2	38.9	Septicemia	0.769	2.411	1
	4	785.52	Septic shock	0.669	4.090	2
	5	584.9	Acute kidney failure	0.534	4.956	4
	14	995.92	Severe sepsis	0.824	7.816	5
TDCCOP ₂ (3027)	1	518.81	Acute respiratory failure	0.526	7.665	3
	2	38.9	Septicemia	0.608	7.545	2
	4	785.52	Septic shock	0.729	7.813	8
	5	584.9	Acute kidney failure	0.554	7.377	1
	12	427.31	Atrial fibrillation	0.593	8.038	11
	14	995.92	Severe sepsis	0.941	8.031	10
	30	428.0	Congestive heart failure	0.729	7.703	5
	46	486.0	Pneumonia organism	0.334	7.805	6
	58	599.0	Urinary tract infection	0.389	7.701	4
	62	401.9	Essential hypertension	0.343	7.807	7
	63	276.2	Acidosis	0.360	7.875	9
	77	250.0	Diabetes mellitus without complication	0.383	8.062	12



system(460–519, order: 3), diseases of the circulatory system (390–459, order: 5), septic shock (785.52, order: 8), endocrine, nutritional, and metabolic diseases, and immunity disorders (240–279, order: 9), and severe sepsis (995.92, order: 10).

Then we calculated the CCoM₂ of the LCoP₂ based on the diagnostic information of 800 core patients in cluster 2, as described in Table 3. First, the conditional probabilities $p(\{390-459, 995.92\} / \{580-629, 38.9, 460-519\})$ colored red were significantly larger than the values $p(\{580-629, 38.9, 460-519\} / \{390-459, 995.92\})$ colored blue, which indicates that diseases of the genitourinary system (580–629, order: 1), septicemia (38.9, order: 2), and diseases of the respiratory system (460–519, order: 3) were more likely to be primary diseases, whereas diseases of the circulatory system (390–459, order: 5) and

severe sepsis (995.92, order: 10) were probably complications. Second, the orders of septic shock (785.52, order: 8) and endocrine, nutritional, and metabolic diseases, and immunity disorders (240–279, order: 9) were also larger than those of the first three diseases. Thus, diseases of the respiratory system (460–519, order: 3) and diseases of the circulatory system (390–459, order: 5) were likely to be the optimal segmentation between primary diseases and complications, and the first three diseases were considered to be the UD (UD₂) of cluster 2.

UD Prediction Based on Patient Admission Information

After we applied feature fusion and feature selection using the IG method, we further performed five classifications to predict a UD based on patient admission information and identify important features for the constructed prediction models. **Figure 11** shows the classification performance of the proposed UDIPM, including the area under the ROC curve (AUC), accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1), and **Figure 12** presents the 10 most important features identified using the random forest method (**Supplementary Figure 4**).

TABLE 3 | CCoM₂ of the LCoP₂.

$p_2(d_j/d_i)$	580-629	38.9	460-519	390-459	785.52	240-279	995.92
580-629 (1)	0.75	0.60	0.64	0.92	0.71	0.61	0.94
38.9 (2)	0.73	0.61	0.73	0.93	0.72	0.63	0.94
460-519 (3)	0.72	0.67	0.66	0.92	0.71	0.62	0.94
390-459 (5)	0.74	0.61	0.66	0.93	0.71	0.60	0.94
785.52 (8)	0.99	0.60	0.65	0.93	0.73	0.60	0.97
240-279 (9)	0.74	0.63	0.67	0.91	0.71	0.61	0.94
995.92 (10)	0.74	0.61	0.66	0.92	0.75	0.61	0.94

Values in brackets are the orders of the seven diseases, bold values on the master diagonal denote the occurrence probabilities of the seven diseases, and values in red and blue are conditional probabilities for distinguishing between primary diseases and complications.

The experimental results indicated that the proposed UDIPM achieved better prediction performance, where the AUC values were all above 0.8, except for the decision tree method. Similarly, the best Acc, Pre, Rec, and, F1 among all classifications was XGBoost, at ~80%, followed by random forest, SVM, and logistic regression, whereas the decision tree was last, at ~66%. Consider the random forest as an example. We obtained the feature importance results to better understand the prediction model. First, we found that demographic information (i.e., age) and laboratory examination information were more important than symptom information. Then some disease severity indicators were very important, such as SAPS and SAPS-II. Finally, the variance distribution (i.e., Var) of the laboratory examination indicators was more important than the mean, median, minimum, and maximum values. To summarize, the proposed UDIPM not only identified a UD from patient

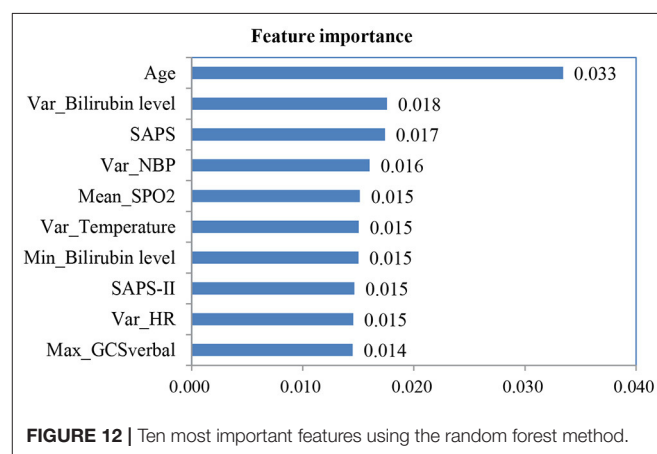


FIGURE 12 | Ten most important features using the random forest method.

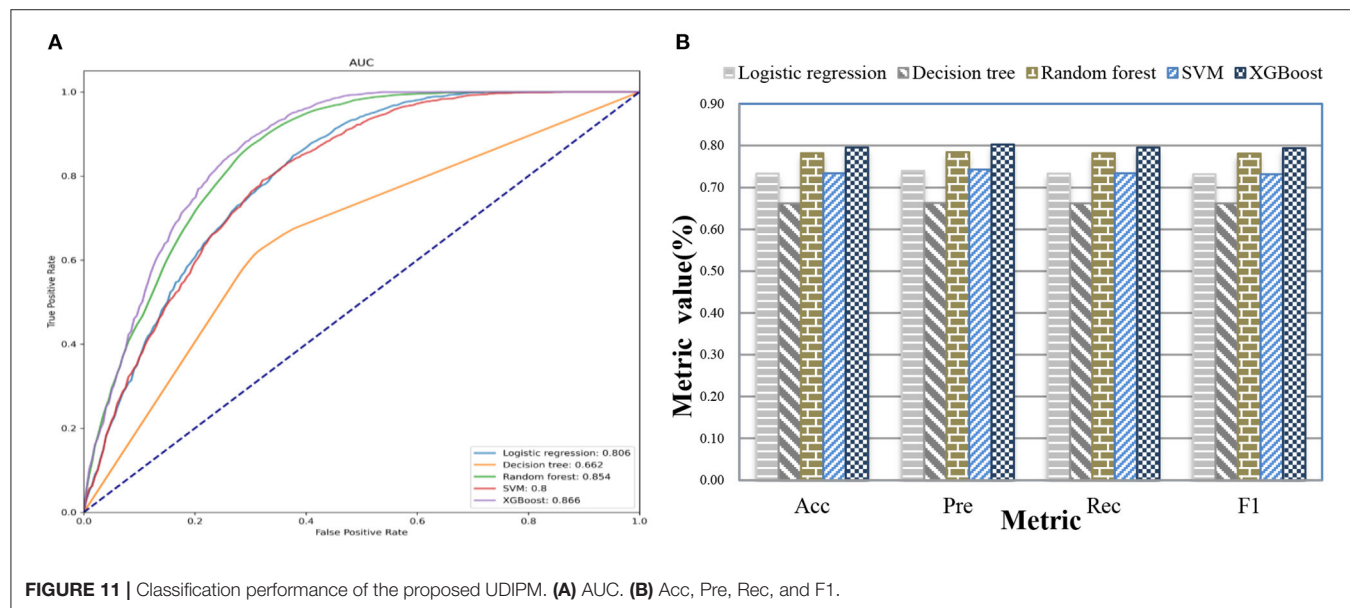


FIGURE 11 | Classification performance of the proposed UDIPM. (A) AUC. (B) Acc, Pre, Rec, and F1.

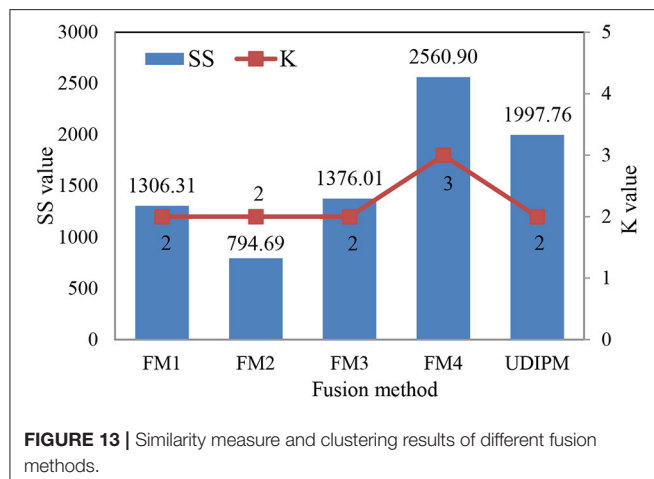
TABLE 4 | Evaluation methods and metrics used in our experiment.

Method name	Set similarity measure	Clustering	Classification
The proposed method (UDIPM)	Set similarity based on ontology	AP clustering	Logistic regression
Fusion method 1 (FM1)	Dice = $2 A \cap B / A + B $		Decision tree
Fusion method 2 (FM2)	Jaccard = $ A \cap B / A \cup B $		Random forest
Fusion method 3 (FM3)	Cosine = $ A \cap B / \sqrt{ A \cdot B }$		SVM
Fusion method 4 (FM4)	Overlap = $ A \cap B / \min(A , B)$		XGBoost
			AUC
			Acc = $(TP + TN)/N$
			Pre = $TP/(TP + FP)$
			Rec = $TP/(TP + FN)$
			F1 = $2Pre \cdot Rec / (Pre + Rec)$

Metric

SS (Eq. 7)

A and B are the diagnosis code sets of two patients, the Dice method is the same as the proposed UDIPM when we do not consider the disease ontology structure and replace the code similarity with $s = (dc_i, dc_j) = \begin{cases} 1, & \text{if } dc_i = dc_j \\ 0 & \text{otherwise} \end{cases}$, true positive (TP) and true negative (TN) measure the ability of classifier models to predict the UD, false positive (FP) and false negative (FN) identify the number of false predictions generated by the models, and we used FM to determine the prediction performance.

**FIGURE 13 |** Similarity measure and clustering results of different fusion methods.

diagnostic information but also predicted a UD based on the health condition of a patient admitted to hospital.

DISCUSSION

In this study, we conducted various experiments to demonstrate the efficiency of the proposed UDIPM when compared with other methods. Specifically, the proposed UDIPM fused three methods: a set similarity measure method, clustering, and classification algorithms. For the set similarity measure method, we selected Dice, Jaccard, cosine, and overlap as comparative methods, and used SS in Eq. 7 as a performance metric based on the AP clustering results. For the classification algorithms, we selected logistic regression, decision tree, random forest, SVM, and XGBoost. Additionally, we used AUC, Acc, Pre, Rec, and F1 as performance metrics to measure the effectiveness of the classification algorithms. The evaluation methods and metrics are described in detail in Table 4.

The detailed experimental results are shown in Figure 13, Table 5. Specifically, for the set similarity measure, we first

TABLE 5 | Classification results of different fusion methods.

Fusion method	Classification algorithm	Metric				
		Acc	Pre	Rec	F1	AUC
FM1 (Dice)	Logistic regression	0.725	0.739	0.725	0.721	0.782
	Decision tree	0.682	0.683	0.682	0.682	0.682
	Random forest	0.779	0.782	0.779	0.778	0.851
	(Jaccard)	0.722	0.763	0.722	0.711	0.778
	XGBoost	0.804	0.818	0.804	0.802	0.860
FM3 (Cosine)	Logistic regression	0.734	0.743	0.734	0.732	0.804
	Decision tree	0.682	0.683	0.682	0.682	0.682
	Random forest	0.786	0.790	0.786	0.785	0.859
	SVM	0.736	0.752	0.736	0.732	0.801
	XGBoost	0.813	0.821	0.813	0.812	0.884
FM4 (Overlap)	Logistic regression	0.465	0.437	0.421	0.411	0.628
	Decision tree	0.388	0.370	0.371	0.369	0.529
	Random forest	0.467	0.434	0.400	0.371	0.620
	SVM	0.471	0.384	0.404	0.350	0.626
	XGBoost	0.481	0.451	0.423	0.404	0.629
UDIPM	Logistic regression	0.733	0.740	0.733	0.732	0.806
	Decision tree	0.662	0.663	0.662	0.662	0.662
	Random forest	0.782	0.784	0.782	0.781	0.854
	SVM	0.734	0.743	0.734	0.732	0.800
	XGBoost	0.795	0.803	0.795	0.794	0.866

Bold values denote the first and second-highest performance using the UDIPM.

selected the optimal number of clusters using AP clustering algorithms, and then computed the SS value based on the clustering results (Algorithm 2). The experimental results indicated that the optimal numbers of clusters for four FMs were 2, 2, 2, and 3 (Supplementary Figure 5), and the proposed UDIPM achieved the second-highest SS value of 1997.86; it was only below FM4 (Figure 13). The reason is that the SS value increased as the cluster number increased. Interestingly, although the similarities of FM1 and FM2 were different, they had the same clustering results.

For the classification results obtained using the 10-fold CV method in **Table 5**, the proposed method achieved the second-highest performance using logistic regression, random forest, and SVM, and the third-highest performance using the decision tree and XGBoost. More importantly, all metrics of the proposed UDIPM were higher than those of FM4. Therefore, from the overall performance evaluation in combination with the set similarity measure, clustering, and classification, the UDIPM was an effective method for identifying and predicting a UD from EMRs.

Further, for all fusion methods, the results of performance comparison indicated that both XGBoost and random forest were superior to other classification algorithms in terms of the Acc, Pre, Rec, F1, and AUC. The main reason is that XGBoost and random forest are ensemble learning algorithms by combining multiple classifiers, which can often achieve more significant generalization performance than a single classifier. Specifically, XGBoost is an improved algorithm based on the gradient boosting decision tree, which can efficiently construct boosted trees and run in parallel. XGBoost works by combining a set of weaker machine learning algorithms to obtain an improved machine learning algorithm as a whole (36). XGBoost has been shown to perform exceptionally well in a variety of tasks in the areas of bioinformatics and medicine, such as the lysine glycation sites prediction for *Homo sapiens* (37), the chronic kidney disease diagnosis (38), and the risk prediction of incident diabetes (39). Also, random forest classifier is an ensemble algorithm, which combines multiple decorrelated decision tree prediction variables based on each subset of data samples (40). In general, random forest shows better performance in disease diagnosis than many single classifiers (41).

CONCLUSION

In this study, we proposed a UDIPM embedding the disease ontology structure to identify and predict a UD from EMRs to assist better coding integration of diagnosis in the ICU. We discussed many critical issues, including a formal representation of multi-type patient information, symptom feature extraction from an unstructured discharge report, ICD ontology structure reconstruction for semantic relation embedding, multi-level set similarity measure for generating a patient similarity matrix, number of cluster selections using AP clustering, stability of the extracted TDC and TDCCoP from each cluster, optimal split line determination for identifying a UD based on visual analysis and the CCoM of LCoP, feature fusion and selection using the IG-based method, and the performance evaluation of UD prediction using five classifiers. We verified the proposed UDIPM on 4,418 sepsis patients in the ICU extracted from the MIMIC-III database. The results showed that the highest stability cluster number and largest range of TDCs were 2 and 400–800, respectively, the UD of cluster 2 was diseases of the genitourinary system (580–629, order: 1), septicemia (38.9, order: 2), and diseases of the respiratory system (460–519, order: 3), and the best AUC and Acc, Pre,

Rec, and F of the UD prediction were 0.866, 0.795, 0.803, 0.795, and 0.794, respectively, which were better than those of other fusion methods from the overall view of SS and prediction performance.

STUDY LIMITATIONS

The proposed UDIPM can identify and predict a UD from EMRs; however, there remain several topics for future work. First, the order of diagnosis codes should be considered in the patient similarity measure by way of different weights because of the importance of primary diseases. Then some state-of-the-art feature selection and classification models should be implemented to improve the prediction accuracies of the UD. Additionally, we hope to make progress on many of the valuable suggestions made by clinicians regarding our implemented method and experimental results.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: <https://mimic.mit.edu/docs/iii/tables/>.

AUTHOR CONTRIBUTIONS

JC, CG, and SD conceived and designed the study and revised the manuscript. JC and ML carried out the experiments and drafted the manuscript. All the authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71771034, 72101236, and 71421001), the Scientific and Technological Innovation Foundation of Dalian (Grant No. 2018J11CY009), the Henan Province Youth Talent Promotion Project (Grant No. 2021HYTP052), the Henan Province Medical Science and Technology Research Plan (Grant No. LHGJ20200279), and the Henan Province Key Scientific Research Projects of Universities (Grant No. 21A320035).

ACKNOWLEDGMENTS

We would like to thank the MIT Laboratory for Computational Physiology and collaborating research groups for providing the freely available database (MIMIC-III). We thank Maxine Garcia, Ph.D., from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2021.793801/full#supplementary-material>

REFERENCES

- Herman J. The unifying diagnosis. *Scand J Prim Health.* (1994) 12:68–9. doi: 10.3109/02813439409003677
- Xie J, Jiang J, Wang Y, Guan Y, Guo X. Learning an expandable EMR-based medical knowledge network to enhance clinical diagnosis. *Artif Intell Med.* (2020) 107:101927. doi: 10.1016/j.artmed.2020.101927
- Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. *Lancet Digit Health.* (2021) 3:e383–e96. doi: 10.1016/S2589-7500(21)00005-4
- Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf Fusion.* (2020) 63:208–22. doi: 10.1016/j.inffus.2020.06.008
- Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs) A survey. *ACM Comput Surv.* (2018) 50:1–40. doi: 10.1145/3127881
- Lin AL, Chen WC, Hong JC. Electronic health record data mining for artificial intelligence healthcare. *Artif Intell Med.* (2021) 133–50. doi: 10.1016/B978-0-12-821259-2.00008-9
- Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature.* (2020) 585:193–202. doi: 10.1038/s41586-020-2669-y
- Myszczyńska MA, Ojames PN, Lacoste AM, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol.* (2020) 16:440–56. doi: 10.1038/s41582-020-0377-8
- Guo C, Chen J. Big data analytics in healthcare: data-driven methods for typical treatment pattern mining. *J Syst Sci Syst Eng.* (2019) 28:694–714. doi: 10.1007/s11518-019-5437-5
- Piri S. Missing care: a framework to address the issue of frequent missing values; The case of a clinical decision support system for Parkinson's disease. *Decis Support Syst.* (2020) 136:113339. doi: 10.1016/j.dss.2020.113339
- Wang S, Li X, Chang* X, Yao L, Sheng QZ, Long G. Learning multiple diagnosis codes for ICU patients with local disease correlation mining. *ACM T Knowl Discov D (TKDD).* (2017) 11:1–21. doi: 10.1145/3003729
- Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput Meth Programs Biomed.* (2019) 177:141–53. doi: 10.1016/j.cmpb.2019.05.024
- Gour N, Khanna P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomed Signal Proces.* (2021) 66:102329. doi: 10.1016/j.bspc.2020.102329
- Trigueros O, Blanco A, Lebera N, Casillas A, Perez A. Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. *Int J Med Inform.* (2021) 157:104615. doi: 10.1016/j.ijmedinf.2021.104615
- Zhang Y, Yang Q. A survey on multi-task learning. *IEEE Trans Knowl Data Eng.* (2021). doi: 10.1109/TKDE.2021.3070203. [Epub ahead of print].
- Yu K, Xie X. Predicting hospital readmission: a joint ensemble-learning model. *IEEE J Biomed Health.* (2019) 24:447–56. doi: 10.1109/JBHI.2019.2938995
- Li T, Wang Z, Lu W, Zhang Q, Li D. Electronic health records based reinforcement learning for treatment optimizing. *Inf Syst.* (2022) 104:101878. doi: 10.1016/j.is.2021.101878
- Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Med Inf.* (2021) 9:e23230. doi: 10.2196/23230
- Sareen J, Olafson K, Kredentser MS, Bienvenu OJ, Blouw M, Bolton JM, et al. The 5-year incidence of mental disorders in a population-based ICU survivor cohort. *Crit Care Med.* (2020) 48:e675–e83. doi: 10.1097/CCM.00000000000004413
- Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* (2016) 3:1–9. doi: 10.1038/sdata.2016.35
- Diao X, Huo Y, Zhao S, Yuan J, Cui M, Wang Y, et al. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *Int J Med Inform.* (2021) 153:104543. doi: 10.1016/j.ijmedinf.2021.104543
- Wu Y, Zeng M, Fei Z, Yu Y, Wu F-X, Li M. KAICD: a knowledge attention-based deep learning framework for automatic ICD coding. *Neurocomputing.* (2020) 469:376–83. doi: 10.1016/j.neucom.2020.05.115
- Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. In: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.* New Orleans, LA (2018). p. 409–16.
- Malhi GS, Bell E, Boyce P, Mulder R, Porter RJ. Unifying the diagnosis of mood disorders. *Aust N Z J Psychiatry.* (2020) 54:561–5. doi: 10.1177/0004867420926241
- Sloan EA, Chiang J, Villanueva-Meyer JE, Alexandrescu S, Eschbacher JM, Wang W, et al. Intracranial mesenchymal tumor with FET-CREB fusion-A unifying diagnosis for the spectrum of intracranial myxoid mesenchymal tumors and angiomatoid fibrous histiocytoma-like neoplasms. *Brain Pathol.* (2021) 31:e12918. doi: 10.1111/bpa.12918
- Liang JJ, Goodsell K, Grogan M, Ackerman MJ. LMNA-mediated arrhythmogenic right ventricular cardiomyopathy and charcot-marie-tooth type 2B1: a patient-discovered unifying diagnosis. *J Cardiovasc Electrophysiol.* (2016) 27:868–71. doi: 10.1111/jce.12984
- Zhu Y, Zhang J, Wang G, Yao R, Ren C, Chen G, et al. Machine learning prediction models for mechanically ventilated patients: analyses of the MIMIC-III database. *Front Med.* (2021) 8:662340. doi: 10.3389/fmed.2021.662340
- Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Med Inform Decis.* (2020) 20:1–10. doi: 10.1186/s12911-020-01271-2
- Jia Z, Lu X, Duan H, Li H. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Med Inform Decis.* (2019) 19:1–11. doi: 10.1186/s12911-019-0807-y
- Jia Z, Zeng X, Duan H, Lu X, Li H, A. patient-similarity-based model for diagnostic prediction. *Int J Med Inform.* (2020) 135:104073. doi: 10.1016/j.ijmedinf.2019.104073
- Park S, Xu H, Zhao H. Integrating multidimensional data for clustering analysis with applications to cancer patient data. *J Am Stat Assoc.* (2021) 116:14–26. doi: 10.1080/01621459.2020.1730853
- Lopez-Martinez-Carrasco A, Juarez JM, Campos M, Canovas-Segura B. A methodology based on Trace-based clustering for patient phenotyping. *Knowl Based Syst.* (2021) 232:107469. doi: 10.1016/j.knosys.2021.107469
- Chen J, Sun L, Guo C, Wei W, Xie Y, A. data-driven framework of typical treatment process extraction and evaluation. *J Biomed Inform.* (2018) 83:178–95. doi: 10.1016/j.jbi.2018.06.004
- Liu Y, Liu J, Jin Y, Li F, Zheng T. An affinity propagation clustering based particle swarm optimizer for dynamic optimization. *Knowl Based Syst.* (2020) 195:105711. doi: 10.1016/j.knosys.2020.105711
- Chen J, Sun L, Guo C, Xie Y. A fusion framework to extract typical treatment patterns from electronic medical records. *Artif Intell Med.* (2020) 103:101782. doi: 10.1016/j.artmed.2019.101782
- Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *PLoS ONE.* (2021) 16:e0246306. doi: 10.1371/journal.pone.0246306
- Yu B, Qiu W, Chen C, Ma A, Jiang J, Zhou H, et al. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics.* (2020) 36:1074–81. doi: 10.1093/bioinformatics/btz734
- Ogunleye A, Wang Q-G. XGBoost model for chronic kidney disease diagnosis. *IEEE ACM T COMPUT BI.* (2019) 17:2131–40. doi: 10.1109/TCBB.2019.2911071
- Wu Y, Hu H, Cai J, Chen R, Zuo X, Cheng H, et al. Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Front Public Health.* (2021) 9:626331. doi: 10.3389/fpubh.2021.626331
- Mueller SQ. Pre-and within-season attendance forecasting in Major League Baseball: a random forest approach. *Appl Econ.* (2020) 52:4512–28. doi: 10.1080/00036846.2020.1736502

41. Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y. An improved random forest-based rule extraction method for breast cancer diagnosis. *Appl Soft Comput.* (2020) 86:105941. doi: 10.1016/j.asoc.2019.105941

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of

the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Chen, Guo, Lu and Ding. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

NOTATION

dc_i	i -th diagnosis code
$Ord(dc_i)$	Order of dc_i
$LCA(dc_i, dc_j)$	Least common ancestor of dc_i and dc_j
$s(dc_i, dc_j)$	Similarity of diagnosis code dc_i and dc_j
$S(D_i', D_j')$	Similarity of diagnostic information of patients i and j
S	Patient similarity matrix based on diagnostic information
p_c	p coefficient to control the input exemplar preferences
K	Number of clusters
C_k	k -th cluster, $k = 1, 2, \dots, K$
$E(C_k)$	Exemplar of cluster C_k
$Core_k, Core_k $	Core zone and the number of patients in C_k
$Prob_k(dc_h)$	Occurrence probability of the diagnosis code dc_h in C_k
$AOrd_k(Tdc_h)$	Average order of the typical diagnosis code dc_h in C_k
$Ord'(Tdc_h)$	New order of the typical diagnosis code dc_h
$TDCCoP_k$	k -th typical diagnosis code co-occurrence pattern
$LCoP_k$	k -th least common ancestor co-occurrence pattern
CCoM_k	Conditional co-occurrence matrix for all diseases in $TDCoP_k$
UD_k	k -th unifying diagnosis
$IG(x_i)$	Information gain of feature x_i
$CVError_Z$	Average error using Z -fold cross-validation



A Web-Based Prediction Model for Cancer-Specific Survival of Middle-Aged Patients With Non-metastatic Renal Cell Carcinoma: A Population-Based Study

Jie Tang^{1†}, Jinkui Wang^{2†}, Xiudan Pan¹, Xiaozhu Liu³ and Binyi Zhao^{3*}

¹ Department of Biostatistics and Epidemiology, School of Public Health, Shenyang Medical College, Shenyang, China,

² Department of Urology, Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing Key Laboratory of Pediatrics, China International Science and Technology Cooperation Base of Child Development and Critical Disorders, National Clinical Research Center for Child Health and Disorders (Chongqing), Children's Hospital of Chongqing Medical University, Chongqing, China, ³ Department of Cardiology, The Second Affiliated Hospital of Chongqing Medical University, Chongqing, China

OPEN ACCESS

Edited by:

Yi-Ju Tseng,
National Central University, Taiwan

Reviewed by:

Sharnil Pandya,
Symbiosis International
University, India
Abdul Rehman Javed,
Air University, Pakistan

*Correspondence:

Binyi Zhao
zq720128@163.com

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 26 November 2021

Accepted: 17 January 2022

Published: 24 February 2022

Citation:

Tang J, Wang J, Pan X, Liu X and
Zhao B (2022) A Web-Based
Prediction Model for Cancer-Specific
Survival of Middle-Aged Patients With
Non-metastatic Renal Cell Carcinoma:
A Population-Based Study.
Front. Public Health 10:822808.
doi: 10.3389/fpubh.2022.822808

Background: Renal cell carcinoma (RCC) is one of the most common cancers in middle-aged patients. We aimed to establish a new nomogram for predicting cancer-specific survival (CSS) in middle-aged patients with non-metastatic renal cell carcinoma (nmRCC).

Methods: The clinicopathological information of all patients from 2010 to 2018 was downloaded from the SEER database. These patients were randomly assigned to the training set (70%) and validation set (30%). Univariate and multivariate COX regression analyses were used to identify independent risk factors for CSS in middle-aged patients with nmRCC in the training set. Based on these independent risk factors, a new nomogram was constructed to predict 1-, 3-, and 5-year CSS in middle-aged patients with nmRCC. Then, we used the consistency index (C-index), calibration curve, and area under receiver operating curve (AUC) to validate the accuracy and discrimination of the model. Decision curve analysis (DCA) was used to validate the clinical application value of the model.

Results: A total of 27,073 patients were included in the study. These patients were randomly divided into a training set ($N = 18,990$) and a validation set ($N = 8,083$). In the training set, univariate and multivariate Cox regression analysis indicated that age, sex, histological tumor grade, T stage, tumor size, and surgical method are independent risk factors for CSS of patients. A new nomogram was constructed to predict patients' 1-, 3-, and 5-year CSS. The C-index of the training set and validation set were 0.818 (95% CI: 0.802-0.834) and 0.802 (95% CI: 0.777-0.827), respectively. The 1-, 3-, and 5-year AUC for the training and validation set ranged from 77.7 to 80.0. The calibration curves of the training set and the validation set indicated that the predicted value is highly consistent with the actual observation

value, indicating that the model has good accuracy. DCA also suggested that the model has potential clinical application value.

Conclusion: We found that independent risk factors for CSS in middle-aged patients with nmRCC were age, sex, histological tumor grade, T stage, tumor size, and surgery. We have constructed a new nomogram to predict the CSS of middle-aged patients with nmRCC. This model has good accuracy and reliability and can assist doctors and patients in clinical decision making.

Keywords: nomogram, middle-aged patients, nmRCC, cancer-specific survival, SEER, online application

INTRODUCTION

In recent years, renal cell carcinoma (RCC) incidence has gradually increased, accounting for 2–3% of adult malignant tumors (1). The incidence of RCC in the United States is about 9.1 per 100,000, and the mortality rate is 3.5 per 100,000 (2). It has been reported in the literature that 15% of patients with RCC diagnosed for the first time have developed distant metastases, and another 10–20% of patients with localized RCC eventually develop metastatic RCC (3, 4). The incidence of RCC in men is higher than that in women, about 1.65:1 (5, 6). In 2016, there were 6,700 new diagnoses of RCC in the United States, and 14,240 patients died of renal cancer (7). The prognosis of nmRCC is good, but the 5-year survival rate of metastatic RCC is about 10%, and the median survival time is only 10 months (8). A comprehensive treatment method based on surgery is advocated for localized RCC (9). However, 20–30% of patients with localized RCC still relapse after surgery (10). Therefore, evaluation of the progression, metastasis and prognosis of RCC is critical in clinical management.

At present, studies have shown that clinicopathological factors such as age, sex, and tumor size are related to the prognosis of RCC (11, 12). Guo et al. found that the right RCC has a better prognosis than the left (13). Wang et al. constructed a nomogram to predict the survival of RCC patients with bone metastases and found that age, sex, marriage, tumor histology grade, T stage, N stage, surgery, and radiotherapy are independent risks factors for patients (14). Li et al. developed a nomogram to predict the risk of distant metastasis in patients with RCC (15). Yue et al. found that age is a critical factor in the prognosis of patients with metastatic RCC; elderly patients have a worse prognosis than younger patients (16).

At present, artificial intelligence has been widely used in the medical field. Awais et al. (17) use texture analysis to classify abnormal areas of the mouth and promote the development of oral cancer treatment. Mishra et al. (18) use intelligent drive for multistage assessment of mental disorders to help patients with mental illness. Although various kinds of nomograms have been widely used in clinical practice, the accuracy and specificity of these nomograms are very worrying. We aimed to establish a specific nomogram to predict survival in middle-aged patients with renal cell carcinoma. This study used big data based on the Cox regression model to construct a simple nomogram, which is as convenient as possible for users to operate under the premise of ensuring accuracy.

RCC has become significant cancer endangering the health of the population. Accurate prediction of the survival of cancer patients is the key to improving the survival time and quality of life of patients with RCC. At present, using big medical data to establish a prediction model has become an essential means to predict the survival of cancer patients. The nomogram is a user-friendly graphical digital model that can accurately predict the occurrence of a given event based on the numerical estimation of multiple single variables (19). Middle-aged patients with RCC have a good prognosis without distant metastasis. However, accurate prognostic assessment can answer patient consultations and help doctors and patients make clinical decisions. Therefore, we aim to establish a nomogram to predict the CSS of middle-aged patients with nmRCC.

PATIENTS AND METHODS

Data Source and Data Extraction

We downloaded the clinical-pathological data of the patients from the National Cancer Institute's Surveillance, Epidemiology, and Final Results (SEER) project, including patients who were diagnosed with nmRCC in the United States from 2010 to 2018 between 40 and 60 years old. The data of this study can be obtained from the SEER database (<http://seer.cancer.gov/>). The SEER database is a public database that contains 18 cancer registries and covers ~28% of the American population (20). Patient information can be obtained on the database, including demographic information, tumor characteristics, and survival status. Because the data we used is publicly available, and the patient's personal information is not identifiable, our study did not require ethical approval and informed consent. Our research method complies with the rules of the SEER database.

The patient's demographic information and clinical-pathological data include age, sex, race, year of diagnosis, marriage, tumor laterality, tumor histological type, histological grade, T stage, type of surgery, radiotherapy chemotherapy, and survival time. Inclusion criteria: (1) age 40–60 years; (2) pathological diagnosis of renal cell carcinoma (ICD-O-3 codes 8260, 8310, 8312, 8317); (3) diagnosis year 2010–2018. Exclusion criteria: (1) unknown race; (2) unknown tumor size; (3) unknown surgical method; (4) unknown T stage; (5) survival time <1 month; (6) unknown cause of death. The flow chart of patient screening is shown in **Figure 1**.

The year of diagnosis was divided into 2010-2014 and 2015-2018. The race included white, black, and other races (American Indian/AK Native, Asian/Pacific Islander). Tumor grades have grade I (highly differentiated), grade II (moderately differentiated), grade III (poorly differentiated), and grade IV (undifferentiated). The pathological types of RCC include renal clear cell carcinoma, renal papillary adenocarcinoma, renal chromophobe cell carcinoma, and unclassified renal cell carcinoma. According to the SEER operation code, the operation was divided into local tumor excision (code 10-27), partial nephrectomy (PN, code 30) and radical nephrectomy (RN, code 40-80).

Univariate and Multivariate Cox Regression Analysis

The patients were randomly divided into a training set (70%) and a validation set (30%). In the training set, univariate and multivariate Cox regression models were used to analyze independent risk factors for survival of nmRCC patients, and the hazard ratio (HR) and 95% confidence interval (CI) were recorded.

Nomogram Construction for 1-, 3-, and 5-Year CSS

The identified independent risk factors were used to construct a nomogram to predict 1-, 3-, and 5-year CSS in nmRCC patients. All independent risk factors were imported into the nomogram based on the Cox regression model. The risk weights of various variables and the degree of risk are accurately displayed in the nomogram.

Nomogram Validation

The calibration curve was used to test the accuracy of the prediction model, and we used 1,000 bootstrap samples for internal validation. The 1-, 3-, and 5-year areas under the receiver operating curve (AUC) of the training set and the validation

set were used to test the accuracy and discrimination of the prediction model. Similarly, we used the consistency index (C-index) to test the discriminative power of the model.

Clinical Utility

Decision curve analysis (DCA) is a new calculation method that estimates the net benefits under various risk thresholds to evaluate the clinical value of the model (21). DCA was used to assess the clinical application value of the nomogram and compare it with T staging. In addition, according to the nomogram score, patients were divided into a high-risk group and a low-risk group. Kaplan-Meier curve and log-rank test were used to compare the survival differences of patients in different groups.

Statistical Analysis

The count data was described by frequency (%), and the chi-square test and non-parametric you test were used to compare groups. Measurement data (age, tumor size) were expressed using mean and standard deviation, and a non-parametric test (*U*-test) was used for differences between groups. The Cox regression model was used to analyze the risk factors of patient survival, and the Kaplan-Meier curve and log-rank test were used to compare the survival differences of patients between groups. All statistical analysis uses SPSS 26.0 and R software 4.1.0. *P*-value < 0.05 was considered to be statistically different.

RESULTS

Clinical Features

According to the inclusion and exclusion criteria, a total of 27,073 patients were included in the study. These patients were randomly divided into a training set ($N = 18,990$) and a validation set ($N = 8,083$). **Table 1** shows the clinicopathological characteristics of all patients. The average age of the patients was 52.4 years, 20,993 patients were white (77.5%), 17,531 patients

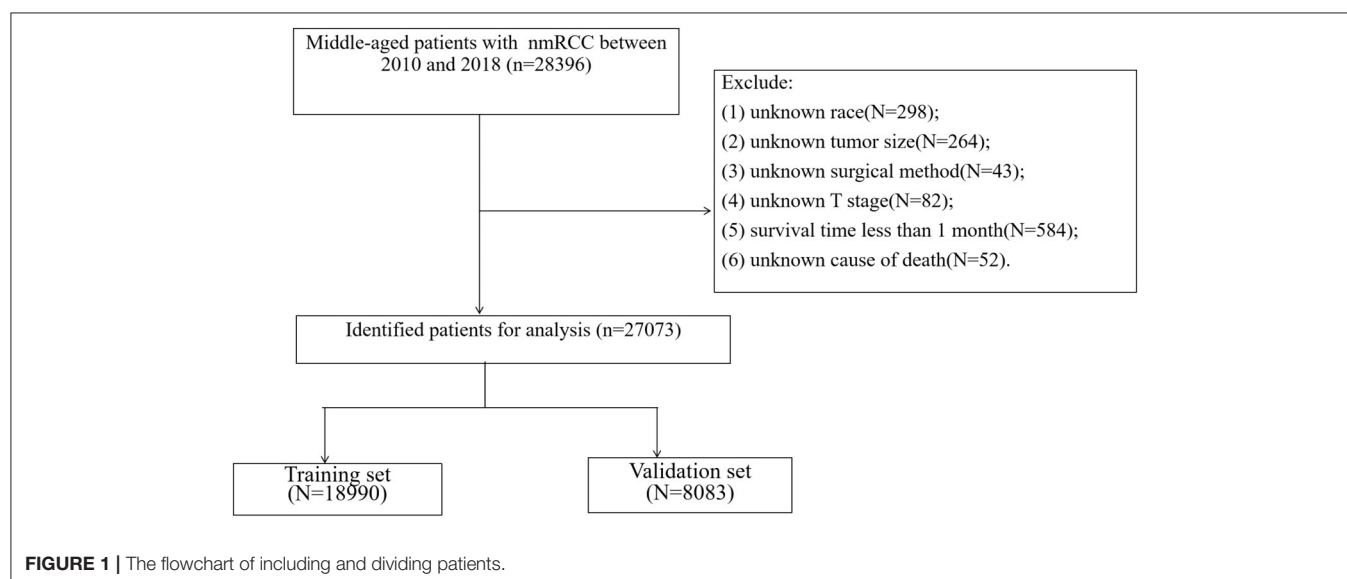


TABLE 1 | Clinicopathological characteristics of patients with nmRCC.

	All	Training set	Validation set	P
	N = 27,073	N = 18,990	N = 8,083	
Age	52.4 (5.61)	52.4 (5.61)	52.4 (5.61)	0.724
Race				0.651
White	20,993 (77.5%)	14,754 (77.7%)	6,239 (77.2%)	
Black	4,288 (15.8%)	2,985 (15.7%)	1,303 (16.1%)	
Other	1,792 (6.62%)	1,251 (6.59%)	541 (6.69%)	
Sex				0.644
Male	17,531 (64.8%)	12,314 (64.8%)	5,217 (64.5%)	
Female	9,542 (35.2%)	6,676 (35.2%)	2,866 (35.5%)	
Year of diagnosis				0.718
2010-2014	14,784 (54.6%)	10,356 (54.5%)	4,428 (54.8%)	
2015-2018	12,289 (45.4%)	8,634 (45.5%)	3,655 (45.2%)	
Marriage				0.152
No	11,058 (40.8%)	7,810 (41.1%)	3,248 (40.2%)	
Married	16,015 (59.2%)	11,180 (58.9%)	4,835 (59.8%)	
Grade				0.143
I	2,747 (10.1%)	1,894 (9.97%)	853 (10.6%)	
II	12,412 (45.8%)	8,797 (46.3%)	3,615 (44.7%)	
III	6,048 (22.3%)	4,196 (22.1%)	1,852 (22.9%)	
IV	928 (3.43%)	651 (3.43%)	277 (3.43%)	
Unknown	4,938 (18.2%)	3,452 (18.2%)	1,486 (18.4%)	
T				0.172
T1a	14,571 (53.8%)	10,178 (53.6%)	4,393 (54.3%)	
T1b	6,105 (22.6%)	4,314 (22.7%)	1,791 (22.2%)	
T2	5,015 (18.5%)	3,557 (18.7%)	1,458 (18.0%)	
T3	1,350 (4.99%)	917 (4.83%)	433 (5.36%)	
T4	32 (0.12%)	24 (0.13%)	8 (0.10%)	
Laterality				0.784
Left	13,090 (48.4%)	9,171 (48.3%)	3,919 (48.5%)	
Right	13,983 (51.6%)	9,819 (51.7%)	4,164 (51.5%)	
Histologic type				0.866
Clear cell	17,534 (64.8%)	12,270 (64.6%)	5,264 (65.1%)	
Papillary	4,011 (14.8%)	2,832 (14.9%)	1,179 (14.6%)	
Chromophobe	1,776 (6.56%)	1,249 (6.58%)	527 (6.52%)	
Not classified	3,752 (13.9%)	2,639 (13.9%)	1,113 (13.8%)	
Tumor size	46.0 (31.9)	46.1 (31.8)	45.7 (32.1)	0.323
Surgery				0.843
No	1,161 (4.29%)	811 (4.27%)	350 (4.33%)	
Local tumor excision	1,042 (3.85%)	742 (3.91%)	300 (3.71%)	
Partial nephrectomy	11,750 (43.4%)	8,222 (43.3%)	3,528 (43.6%)	
Radical nephrectomy	13,120 (48.5%)	9,215 (48.5%)	3,905 (48.3%)	
Chemotherapy				0.266
No/unknown	26,702 (98.6%)	18,740 (98.7%)	7,962 (98.5%)	
Yes	371 (1.37%)	250 (1.32%)	121 (1.50%)	
Radiation				0.531
No/unknown	27,015 (99.8%)	18,952 (99.8%)	8,063 (99.8%)	
Yes	58 (0.21%)	38 (0.20%)	20 (0.25%)	

were male (64.8%), and 16,015 patients were married (59.2%). There were 14,784 (54.6%) patients diagnosed in 2010-2014. Patients with tumor grades I, II, III, and IV was 2,747 (10.1%),

12,412 (45.8%), 6,048 (22.3%), and 928 (3.43%), respectively. 14,571 (53.8%) tumors with T1a stage, 17,534 (64.8%) with the histopathological type of renal clear cell carcinoma, and the

average tumor diameter were 46.0 mm. Most patients underwent surgery, 11,750 (43.4%) patients underwent PN, and 13,120 (48.5%) patients underwent RN. Most of the patients did not receive radiotherapy and chemotherapy, 26,702 (98.6%) patients did not receive chemotherapy, and 27,015 (99.8%) patients did not receive radiotherapy. There was no significant difference between the clinical-pathological information of the patients in the training set and the validation set.

Univariate and Multivariate Cox Regression Analysis

All variables were included in univariate Cox regression analysis to screen out survival-related variables. We found that age (HR 1.05, 95%CI 1.03-1.06, $p < 0.001$), sex (HR 0.7, 95%CI 0.59-0.82, $p < 0.001$), tumor histological grade (HR 1.41, 95%CI 1.34-1.49, $p < 0.001$), T stage (HR 2.55, 95%CI 2.35-2.75, $p < 0.001$), tumor size (HR 1.01, 95%CI 1.01-1.01, $p < 0.001$), and surgery (HR 1.23, 95%CI 1.1-1.38, $p < 0.001$) were related to survival prognosis. These factors were included in the multivariate cox regression analysis and showed that all variables were independent prognostic risk factors (Table 2). In other words, these risk factors can be used as factors predicting CSS in patients with nmRCC.

Nomogram Construction for 1-Year, 3-Year, and 5-Year CSS

Based on the independent risk factors screened out by univariate and multivariate Cox regression analysis, we constructed a new nomogram to predict the 1-year, 3-year, and 5-year CSS of middle-aged patients with nmRCC (Figure 2). The nomogram showed that tumor size and T stage are the most significant factors affecting the patient's CSS, followed by surgery, histological tumor grade, and the final age and sex have little effect on the survival and prognosis of patients.

Validation of the Nomogram

The calibration curve showed that the 1-, 3-, and 5-year predicted values are highly consistent with the actual observed values in the training set, and the validation set are highly compatible with the existing experimental values, suggesting that our model has good accuracy (Figures 3A-F). The C-index in the training set and the validation set were 0.818 (95% CI: 0.802-0.834) and 0.802 (95% CI: 0.777-0.827), respectively, indicating that our prediction model has good discrimination. In the training set, the AUCs of the models that predict patients' 1-, 3-, and 5-year CSS are 0.796, 0.80, and 0.792, respectively (Figure 4A). In the validation set, the AUCs of the models that predict the patient's 1-, 3-, and 5-year CSS are 0.781, 0.795, and 0.777, respectively (Figure 4B). It also proved that the predictive model has good discrimination.

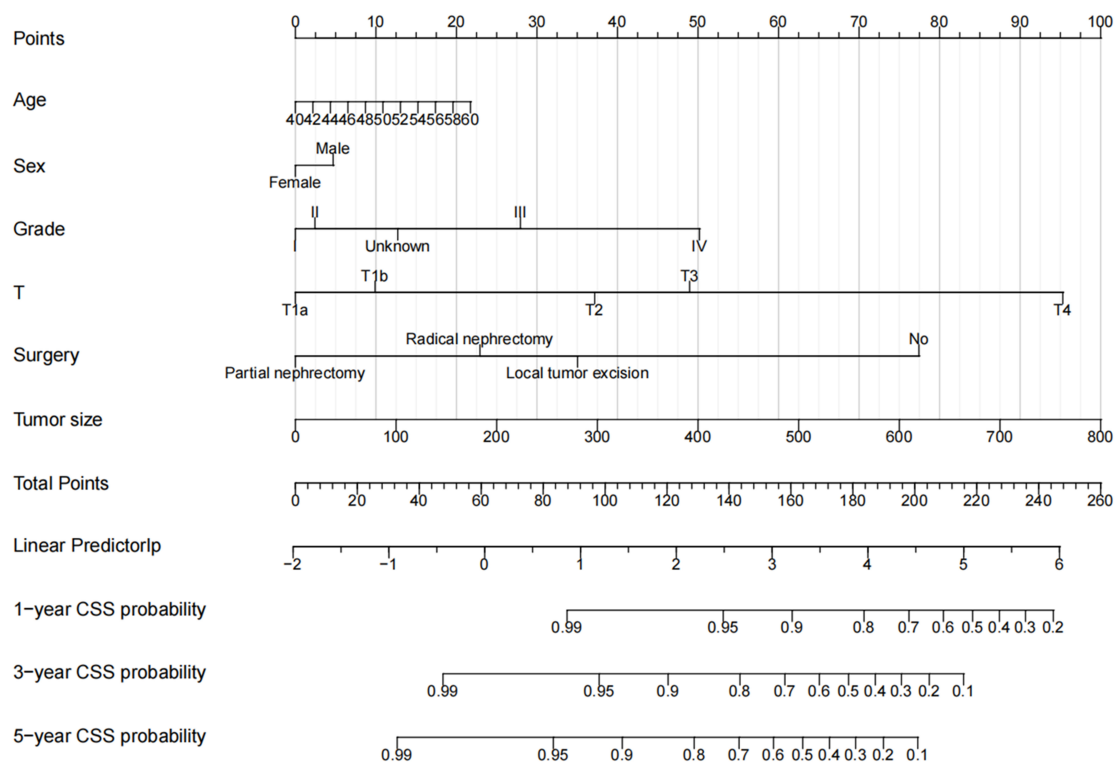


FIGURE 2 | Nomogram for 1-, 3-, and 5-year CSS of middle-aged patients with nmRCC.

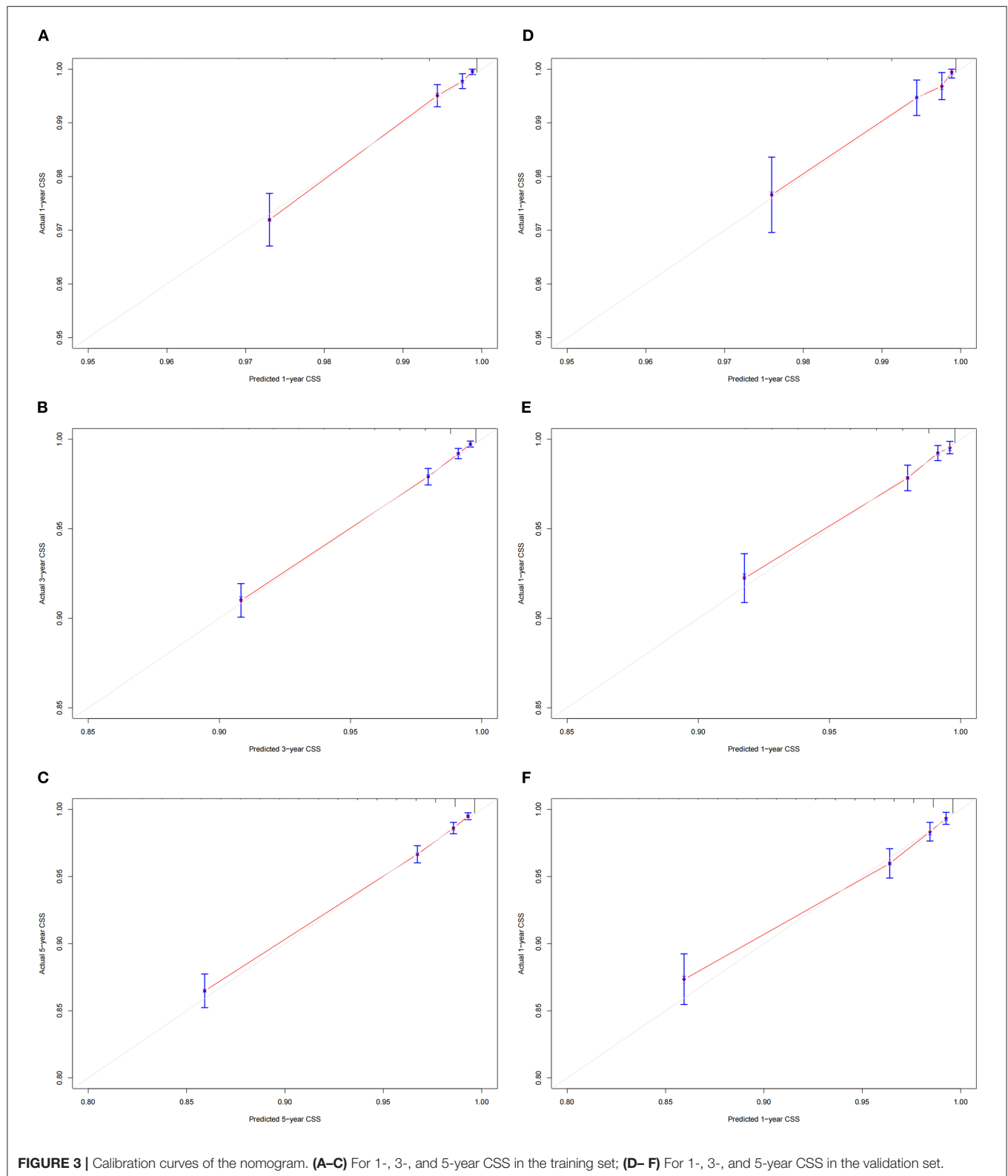


FIGURE 3 | Calibration curves of the nomogram. (A–C) For 1-, 3-, and 5-year CSS in the training set; (D–F) For 1-, 3-, and 5-year CSS in the validation set.

Clinical Application of the Nomogram

DCA suggested that the nomogram has a better clinical application value in the training and validation set, and it is

significantly better than T staging (Figures 5A,B). In addition, we had developed a risk stratification system. According to the score of each patient on the nomogram, all patients were divided

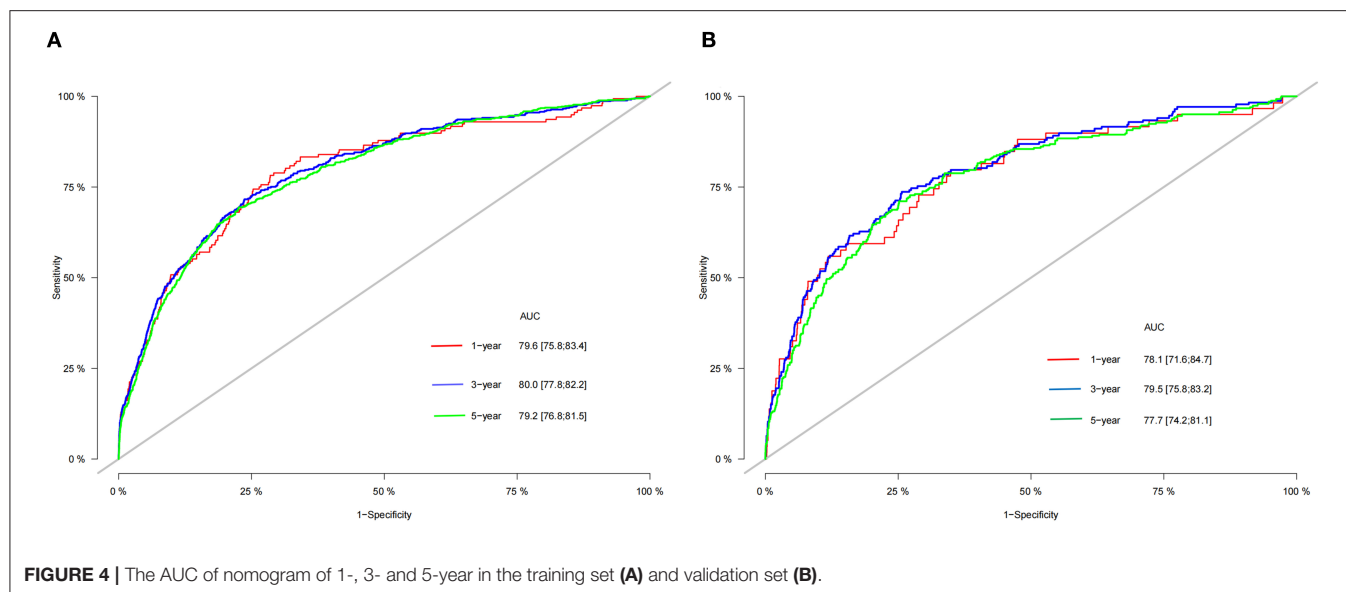


FIGURE 4 | The AUC of nomogram of 1-, 3- and 5-year in the training set (A) and validation set (B).

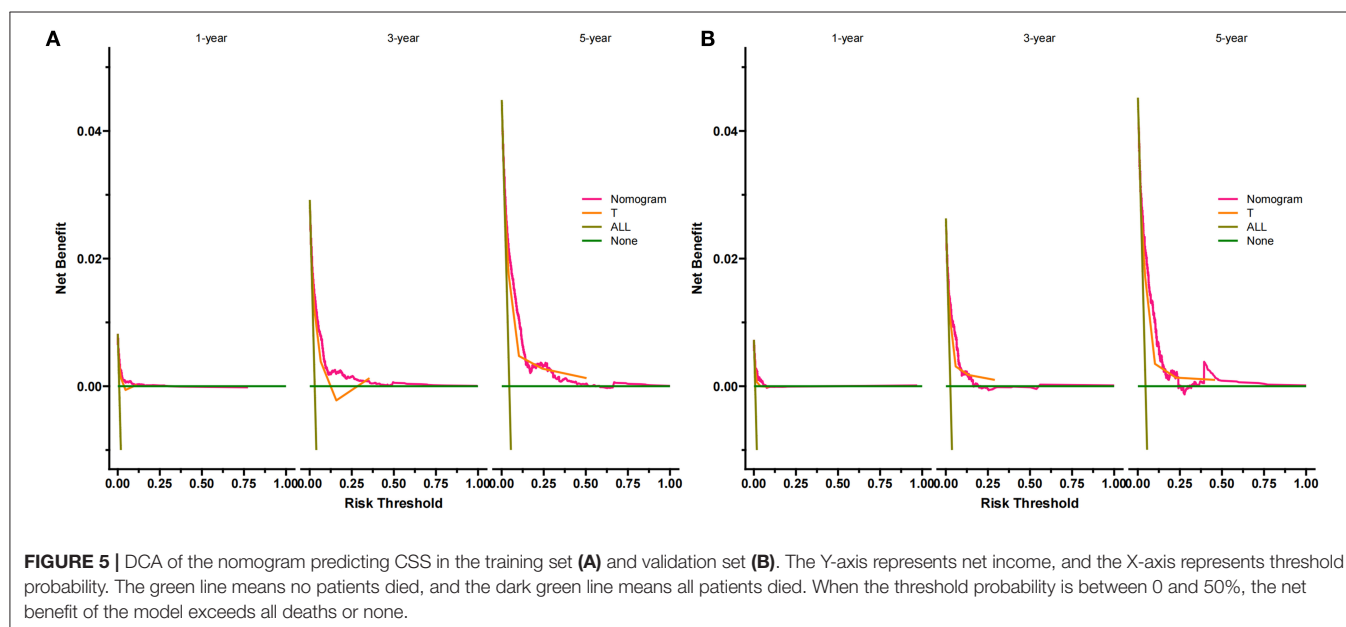


FIGURE 5 | DCA of the nomogram predicting CSS in the training set (A) and validation set (B). The Y-axis represents net income, and the X-axis represents threshold probability. The green line means no patients died, and the dark green line means all patients died. When the threshold probability is between 0 and 50%, the net benefit of the model exceeds all deaths or none.

into a low-risk group (total score ≤ 72.3) and a high-risk group (total score > 72.3).

According to the Kaplan-Meier curve, the high-risk group's 1-, 3-, and 5-year CSS rates were 97.2, 91.5, and 87.2%, respectively. The low-risk group's 1-, 3-, and 5-year CSS rates were 99.7, 98.7, and 97.8%. There was a significant difference in survival between the high-risk group and the low-risk patients in the training and validation set (Figure 6), indicating that our predictive model can accurately identify high-risk patients. In addition, we compared the survival differences of surgical methods in patients with different risk groups. We found that patients with surgery in the low-risk group had a higher survival rate than patients without surgery, including PN, RN, and local tumor excision (Figure 7A). However, although most patients chose RN in the high-risk

group, patients with PN and local tumor excision have a higher survival rate than RN (Figure 7B).

Online Application for CSS Prediction

Based on the nomogram we constructed, we developed a web application to predict the CSS of middle-aged patients with nmRCC. Visit <https://xiudanpan.shinyapps.io/DynNomapp/> to enter the website. Enter the patient's clinical characteristics, and we can obtain the CSS of the patient at each time.

DISCUSSION

RCC is a common tumor of the urinary system in the world, the incidence of women ranks ninth, and the incidence of

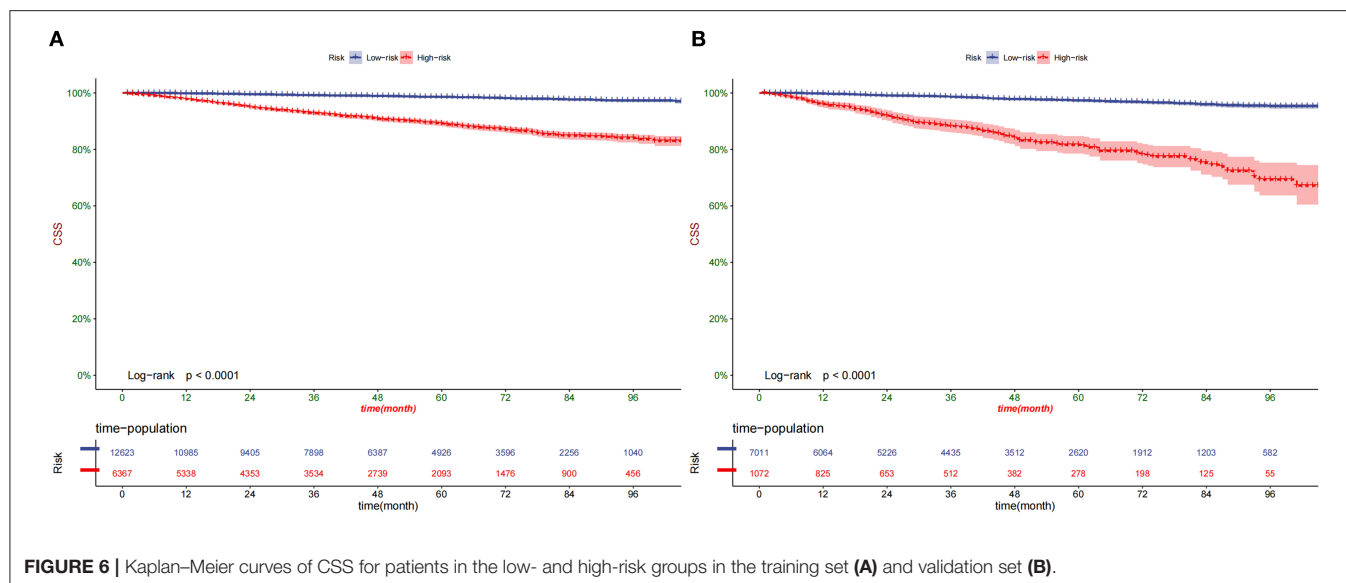


FIGURE 6 | Kaplan-Meier curves of CSS for patients in the low- and high-risk groups in the training set (A) and validation set (B).

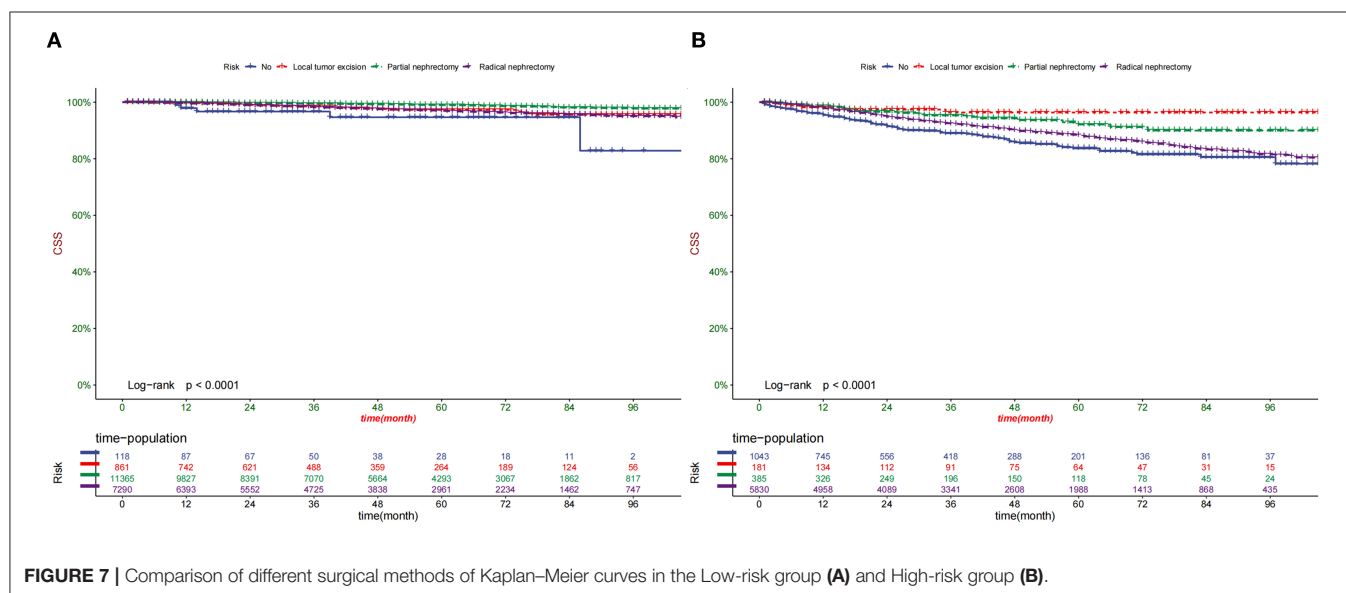


FIGURE 7 | Comparison of different surgical methods of Kaplan-Meier curves in the Low-risk group (A) and High-risk group (B).

men ranks seventh (22, 23). Although surgical treatment, immunotherapy, targeted therapy, and other RCC treatment methods are developing rapidly. However, due to the widespread local recurrence, distant metastasis, and drug tolerance of RCC, the prognosis of RCC patients is not very optimistic (24). To improve RCC patients' prognosis and quality of life, more and more renal cancer surgery risk scoring standards and renal cancer prognostic risk stratification have been established (25–27). The prognosis of early RCC is relatively good. The study reported that early asymptomatic RCC prediction is significantly better than that of symptomatic RCC (28). With improved health awareness and the popularization of health examinations, symptomatic kidney cancer is rare, and patients with advanced kidney cancer are more common. The proportion of early asymptomatic kidney cancer is gradually increasing, with reports ranging from 46.2 to

61% (29). It may be because of the recent increase in abdominal imaging, which is the main reason for the early diagnosis of asymptomatic kidney cancer (30). One study found that frequent use of CT for abdominal scans is associated with the risk of nephrectomy (31).

This study focused on middle-aged nmRCC patients and established a prognostic nomogram for predicting the CSS of middle-aged nmRCC patients for the first time. Because middle-aged patients' remaining lives with nmRCC are still very long, an accurate prognosis can help patients improve their survival rate and quality of life. Our constructed nomogram can accurately predict patients' 1-, 3-, 5-year CSS. According to the univariate and multivariate analysis of patients, age, sex, histological grade, T stage, surgery, and tumor size are independent risk factors.

TABLE 2 | Univariate and multivariate analyses of CSS in training set.

	Univariate			Multivariate		
	HR	95%CI	P	HR	95%CI	P
Age	1.05	1.03-1.06	<0.001	1.036	1.021-1.051	<0.001
Race						
White	Reference					
Black	1.075	0.911-1.267	0.391			
Other	0.998	0.774-1.288	0.989			
Sex						
Male	Reference			Reference		
Female	0.7	0.59-0.82	<0.001	0.859	0.729-1.013	0.07
Year of diagnosis						
2010-2014	Reference					
2015-2018	0.89	0.73-1.08	0.225			
Marriage						
No	Reference					
Married	0.88	0.76-1.02	0.088			
Grade						
I	Reference			Reference		
II	1.238	0.895-1.714	0.198	1.082	0.723-1.618	0.702
III	4.17	3.042-5.717	<0.001	2.468	1.658-3.673	<0.001
IV	14.972	10.711-20.93	<0.001	5.054	3.293-7.759	<0.001
Unknown	3.14	2.26-4.363	<0.001	1.508	0.99-2.297	0.055
T						
T1a	Reference			Reference		
T1b	1.982	1.628-2.413	<0.001	1.377	1.065-1.779	0.015
T2	6.509	5.565-7.613	<0.001	3.328	2.612-4.241	<0.001
T3	11.057	8.482-14.412	<0.001	4.861	3.46-6.83	<0.001
T4	74.388	36.566-151.329	<0.001	21.349	9.677-47.103	<0.001
Laterality						
Left	Reference					
Right	0.93	0.81-1.08	0.345			
Histologic type						
Clear cell	Reference					
Papillary	0.909	0.758-1.091	0.308			
Chromophobe	0.366	0.245-0.545	<0.001			
Not classified	1.414	1.209-1.655	<0.001			
Tumor size	1.01	1.01-1.01		1.004	1.003-1.005	<0.001
Surgery						
No	Reference			Reference		
Local tumor excision	0.174	0.111-0.273	<0.001	0.254	0.151-0.427	<0.001
Partial nephrectomy	0.08	0.062-0.103	<0.001	0.082	0.059-0.115	<0.001
Radical nephrectomy	0.421	0.345-0.514	<0.001	0.172	0.128-0.232	<0.001

Similar to other studies, our results also found that age is a critical factor in the prognosis of patients, even in middle-aged patients (32). Because the increase of age will bring about the weakening of the immune system, further causing the deterioration of the tumor and reducing the survival time of the patient (33). In our study, men have a higher incidence of kidney cancer and a higher mortality rate. Sex as a prognostic factor of patients may be related to hormone levels in the body, such as androgens and testosterone can cause specific cancers (34, 35).

Previous studies have found that tumor characteristics are also critical factors for patient survival, such as histological tumor grade, T stage, N stage, and distant metastasis (36). Our study found that tumor size and histological tumor grade are independent risk factors for patient prognosis. The histological grade is related to the stemness of the tumor. Previous studies have found that high-grade tumors are related to bladder cancer and prostate cancer (37, 38). Because high-grade tumors are often highly malignant and aggressive tumors. In addition, tumor size

is also associated with the patient's prognosis. The larger the tumor, the higher the risk of metastasis and invasion.

The TNM staging system is a standard staging system for all tumors. It is mainly determined by postoperative pathological results and clinical staging (39). According to the patient's tumor condition (T), lymph node (N), distant metastasis (M), the cancer is divided into different stages. Indeed, TNM staging is related to the patient's prognosis. The higher the stage, the worse the patient's prognosis. For nmRCC, there is no lymph node and distant metastasis, and only T staging can reflect the staging of the tumor. Our study found that the T stage is the most critical factor affecting the prognosis of patients. The higher the T stage, the worse the patient's prognosis. This also proved that T staging should be used as an essential component of the nomogram.

Tumor treatment mode is also an important prognostic factor for patients with RCC. Surgery, as the essential treatment method, is the most critical factor for the prognosis of renal cancer patients (40). The nomogram showed that patients with PN have the best prognosis, while those without surgery have the worst prognosis. Our risk stratification system suggested that most patients in the low-risk group choose PN and have a high survival rate. For high-risk patients, most patients choose RN. Although patients with RN and local tumor excision have a higher survival rate, this may be caused by selection bias. Because of more extensive and higher T-stage tumors, doctors and patients are more inclined to choose RN. And these patients will have worse outcomes.

This study used the identifiable variables in the SEER database to construct predictions of 1-, 3-, and 5-year CSS in middle-aged patients with nmRCC. The model has good accuracy and discrimination. The calibration curve of the nomogram indicated that the prediction accuracy of the prediction model is very high. The C-index and AUC of the nomogram are about 0.8, which stated that the discriminative accuracy of the prediction model is about 80% and proved that the model is reliable. This nomogram can predict the prognosis of middle-aged patients with nmRCC and provide a reliable basis for personalized treatment and monitoring.

This study used the identifiable variables in the SEER database to construct predictions of 1-, 3-, and 5-year CSS in middle-aged patients with nmRCC. The model has good accuracy and discrimination. The calibration curve of the nomogram indicated that the prediction accuracy of the prediction model is very high. The C-index and AUC of the nomogram are about 0.8,

which stated that the discriminative accuracy of the prediction model is about 80% and proved that the model is reliable. This nomogram can predict the prognosis of middle-aged patients with nmRCC and provide a reliable basis for personalized treatment and monitoring.

This study also has some limitations. First of all, we did not include some possible clinical factors, such as BMI, smoking, drinking, hypertension, genetic markers, etc. But we had included important clinical-pathological information, such as tumor stage, surgery and other vital factors, so our results will not be too biased. Secondly, our study was a retrospective cases study, and there may be some deviations that are difficult to adjust. Further prospective studies are necessary to validate our prediction model. Finally, we only used the data in the SEER database for internal validation, and the subsequent external proof is needed to validate the model's accuracy.

CONCLUSION

We found that independent risk factors for CSS in middle-aged patients with nmRCC were age, sex, histological tumor grade, T stage, tumor size, and surgery. We have constructed a new nomogram to predict the CSS of patients. This model has good accuracy and reliability and can assist doctors and patients in clinical decision making.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: <https://seer.Cancer.gov/>.

ETHICS STATEMENT

The data of this study is obtained from the SEER database. The patients' data is public and anonymous, so this study does not require ethical approval and informed consent.

AUTHOR CONTRIBUTIONS

JW, BZ, XL, and JT contributed to the conception and design. JW, BZ, and JT collected and analyzed the data. JW, BZ, XP, and JT drew the figures and tables. JT, XL, XP, and JW wrote the draft. JT, BZ, and XP contributed to manuscript writing and revision. All authors approved the final manuscript.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* (2019) 69:7-34. doi: 10.3322/caac.21551
2. Pantuck AJ, Zisman A, Belldgrun AS. The changing natural history of renal cell carcinoma. *J Urol.* (2001) 166:1611-23. doi: 10.1016/S0022-5347(05)65640-6
3. Janzen NK, Kim HL, Figlin RA, Belldgrun AS. Surveillance after radical or partial nephrectomy for localized renal cell carcinoma and management of recurrent disease. *Urol Clin North Am.* (2003) 30:843-52. doi: 10.1016/S0094-0143(03)00056-9
4. Schwaab T, Schwarzer A, Wolf B, Crocenzi TS, Seigne JD, Crosby NA, et al. Clinical and immunologic effects of intranodal autologous tumor lysate-dendritic cell vaccine with Aldesleukin (Interleukin 2) and IFN- α 2a therapy in metastatic renal cell carcinoma patients. *Clin Cancer Res.* (2009) 15:4986-92. doi: 10.1158/1078-0432.CCR-08-3240
5. Woldrich JM, Mallin K, Ritchey J, Carroll PR, Kane CJ. Sex differences in renal cell cancer presentation and survival: an analysis of the National Cancer Database, 1993-2004. *J Urol.* (2008) 179:1709-13. doi: 10.1016/j.juro.2008.01.024
6. Bergström A, Hsieh CC, Lindblad P, Lu CM, Cook NR, Wolk A. Obesity and renal cell cancer—a quantitative review. *Br J Cancer.* (2001) 85:984-90. doi: 10.1054/bjoc.2001.2040

7. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin.* (2016) 66:7-30. doi: 10.3322/caac.21332
8. De Meerleer G, Khoo V, Escudier B, Joniau S, Bossi A, Ost P, et al. Radiotherapy for renal-cell carcinoma. *Lancet Oncol.* (2014) 15:e170-7. doi: 10.1016/S1470-2045(13)70569-2
9. Frank I, Blute ML, Chevillet JC, Lohse CM, Weaver AL, Zincke H. An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: the SSIGN score. *J Urol.* (2002) 168:2395-400. doi: 10.1016/S0022-5347(05)64153-5
10. Cho HJ, Kim SJ, Ha US, Hong SH, Kim JC, Choi YJ, et al. Prognostic value of capsular invasion for localized clear-cell renal cell carcinoma. *Eur Urol.* (2009) 56:1006-12. doi: 10.1016/j.eururo.2008.11.031
11. Wu J, Zhang P, Zhang G, Wang H, Gu W, Dai B, et al. Renal cell carcinoma histological subtype distribution differs by age, gender, and tumor size in coastal Chinese patients. *Oncotarget.* (2017) 8:71797-804. doi: 10.18632/oncotarget.17894
12. Jung EJ, Lee HJ, Kwak C, Ku JH, Moon KC. Young age is independent prognostic factor for cancer-specific survival of low-stage clear cell renal cell carcinoma. *Urology.* (2009) 73:137-41. doi: 10.1016/j.urol.2008.08.460
13. Guo S, Yao K, He X, Wu S, Ye Y, Chen J, et al. Prognostic significance of laterality in renal cell carcinoma: a population-based study from the surveillance, epidemiology, and end results (SEER) database. *Cancer Med.* (2019) 8:5629-37. doi: 10.1002/cam4.2484
14. Wang K, Wu Z, Wang G, Shi H, Xie J, Yin L, et al. Survival nomogram for patients with bone metastatic renal cell carcinoma: a population-based study. *Int Braz J Urol.* (2021) 47:333-49. doi: 10.1590/s1677-5538.ibju.2020.0195
15. Li Y, Chen P, Chen Z. A population-based study to predict distant metastasis in patients with renal cell carcinoma. *Ann Palliat Med.* (2021) 10:4273-88. doi: 10.21037/apm-20-2481
16. Yue G, Deyu L, Lianyan T, Fengmin S, Mei G, Yajun H, et al. Clinical features and prognostic factors of patients with metastatic renal cell carcinoma stratified by age. *Aging.* (2021) 13:8290-305. doi: 10.18632/aging.202637
17. Awais M, Ghayvat H, Krishnan Pandarathodiyil A, Nabillah Ghani WM, Ramanathan A, Pandya S, et al. Healthcare professional in the loop (HPIL): classification of standard and oral cancer-causing anomalous regions of oral cavity using textual analysis technique in autofluorescence imaging. *Sensors.* (2020) 20:5780. doi: 10.3390/s202005780
18. Mishra S, Tripathy HK, Kumar Thakkar H, Garg D, Kotecha K, Pandya S. An explainable intelligence driven query prioritization using balanced decision tree approach for multi-level psychological disorders assessment. *Front Public Health.* (2021) 9:795007. doi: 10.3389/fpubh.2021.795007
19. Yan Y, Liu H, Mao K, Zhang M, Zhou Q, Yu W, et al. Novel nomograms to predict lymph node metastasis and liver metastasis in patients with early colon carcinoma. *J Transl Med.* (2019) 17:193. doi: 10.1186/s12967-019-1940-1
20. Cronin KA, Ries LA, Edwards BK. The surveillance, epidemiology, and end results (SEER) program of the National Cancer Institute. *Cancer.* (2014) 120(Suppl 23):3755-7. doi: 10.1002/cncr.29049
21. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak.* (2008) 8:53. doi: 10.1186/1472-6947-8-53
22. Rini BI, Campbell SC, Escudier B. Renal cell carcinoma. *Lancet.* (2009) 373:1119-32. doi: 10.1016/S0140-6736(09)60229-4
23. Scelo G, Larose TL. Epidemiology and risk factors for kidney cancer. *J Clin Oncol.* (2018) 36:JCO2018791905. doi: 10.1200/JCO.2018.79.1905
24. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021 [published correction appears in *CA Cancer J Clin.* (2021) 71:359]. *CA Cancer J Clin.* (2021) 71:7-33. doi: 10.3322/caac.21654
25. Beisland C, Guðbrandsdóttir G, Reisæter LA, Bostad L, Hjelle KM. A prospective risk-stratified follow-up programme for radically treated renal cell carcinoma patients: evaluation after eight years of clinical use. *World J Urol.* (2016) 34:1087-99. doi: 10.1007/s00345-016-1796-4
26. Waldert M, Klatte T. Nephrometry scoring systems for surgical decision-making in nephron-sparing surgery. *Curr Opin Urol.* (2014) 24:437-40. doi: 10.1097/MOU.0000000000000085
27. Park DS, Hwang JH, Kang MH, Oh JJ. Association between RENAL nephrometry score and perioperative outcomes following open partial nephrectomy under cold ischemia [published correction appears in *Can Urol Assoc J.* (2014) Mar-Apr;8:84] [published correction appears in *Can Urol Assoc J.* (2014) 8:84]. *Can Urol Assoc J.* (2014) 8:E137-41. doi: 10.5489/auaj.1372
28. Patar JJ, Leray E, Cindolo L, Ficarra V, Rodriguez A, De La Taille A, et al. Multi-institutional validation of a symptom based classification for renal cell carcinoma. *J Urol.* (2004) 172:858-62. doi: 10.1097/01.ju.0000135837.64840.55
29. Jayson M, Sanders H. Increased incidence of serendipitously discovered renal cell carcinoma. *Urology.* (1998) 51:203-5. doi: 10.1016/S0090-4295(97)00506-2
30. Maehara CK, Silverman SG, Lacson R, Khorasani R. Journal club: renal masses detected at abdominal CT: radiologists' adherence to guidelines regarding management recommendations and communication of critical results. *AJR Am J Roentgenol.* (2014) 203:828-34. doi: 10.2214/AJR.13.11497
31. Welch HG, Skinner JS, Schroek FR, Zhou W, Black WC. Regional variation of computed tomographic imaging in the United States and the risk of nephrectomy. *JAMA Intern Med.* (2018) 178:221-7. doi: 10.1001/jamainternmed.2017.7508
32. Mao W, Zhang Z, Huang X, Fan J, Geng J. Marital status and survival in patients with penile cancer. *J Cancer.* (2019) 10:2661-9. doi: 10.7150/jca.32037
33. Zeng C, Wen W, Morgans AK, Pao W, Shu XO, Zheng W. Disparities by race, age, and sex in the improvement of survival for major cancers: results from the National Cancer Institute Surveillance, epidemiology, and end results (SEER) program in the United States, 1990 to 2010. *JAMA Oncol.* (2015) 1:88-96. doi: 10.1001/jamaoncol.2014.161
34. Huang X, Shu C, Chen L, Yao B. Impact of sex, body mass index and initial pathologic diagnosis age on the incidence and prognosis of different types of cancer. *Oncol Rep.* (2018) 40:1359-69. doi: 10.3892/or.2018.6529
35. Key T, Appleby P, Barnes I, Reeves G, Endogenous Hormones and Breast Cancer Collaborative Group. Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. *J Natl Cancer Inst.* (2002) 94:606-16. doi: 10.1093/jnci/94.8.606
36. Zeng Y, Mayne N, Yang CJ, D'Amico TA, Ng CSH, Liu CC, et al. A nomogram for predicting cancer-specific survival of TNM 8th edition stage I non-small-cell lung cancer. *Ann Surg Oncol.* (2019) 26:2053-62. doi: 10.1245/s10434-019-07318-7
37. Abdel-Rahman O. Bladder cancer mortality after a diagnosis of nonmuscle-invasive bladder carcinoma. *Future Oncol.* (2019) 15:2267-75. doi: 10.2217/fon-2018-0861
38. Liauw SL, Ham SA, Das LC, Rudra S, Packiam VT, Koshy M, et al. Prostate cancer outcomes following solid-organ transplantation: a SEER-medicare analysis. *J Natl Cancer Inst.* (2020) 112:847-54. doi: 10.1093/jnci/djz221
39. Amin MB, Greene FL, Edge SB, Compton CC, Gershengwald JE, Brookland RK, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin.* (2017) 67:93-9. doi: 10.3322/caac.21388
40. Higuchi T, Yamamoto N, Hayashi K, Takeuchi A, Abe K, Taniguchi Y, et al. Long-term patient survival after the surgical treatment of bone and soft-tissue metastases from renal cell carcinoma. *Bone Joint J.* (2018) 100-B:1241-8. doi: 10.1302/0301-620X.100B9.BJJ-2017-1163.R3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Tang, Wang, Pan, Liu and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Development and Validation of a Nomogram to Predict Cancer-Specific Survival in Elderly Patients With Papillary Renal Cell Carcinoma

OPEN ACCESS

Edited by:

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

Reviewed by:

Hairong He,
The First Affiliated Hospital of Xi'an
Jiaotong University, China
Kaibo Guo,
Zhejiang Chinese Medical
University, China

*Correspondence:

Zhen Yang
yangzhen@ety.cn
Bing Yan
yanbing@ety.cn

†These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 12 February 2022

Accepted: 14 March 2022

Published: 04 April 2022

Citation:

Zhanghuang C, Wang J, Yao Z, Li L,
Xie Y, Tang H, Zhang K, Wu C, Yang Z
and Yan B (2022) Development and
Validation of a Nomogram to Predict
Cancer-Specific Survival in Elderly
Patients With Papillary Renal Cell
Carcinoma.
Front. Public Health 10:874427.
doi: 10.3389/fpubh.2022.874427

Chenghao Zhanghuang^{1,2,3†}, Jinkui Wang^{2†}, Zhigang Yao¹, Li Li³, Yucheng Xie⁴,
Haoyu Tang¹, Kun Zhang¹, Chengchuang Wu¹, Zhen Yang^{5*} and Bing Yan^{1,3*}

¹ Department of Urology, Kunming Children's Hospital (Children's Hospital Affiliated to Kunming Medical University), Kunming, China, ² Department of Urology, Chongqing Key Laboratory of Children Urogenital Development and Tissue Engineering, Chongqing Key Laboratory of Pediatrics, Ministry of Education Key Laboratory of Child Development and Disorders, National Clinical Research Center for Child Health and Disorders, China International Science and Technology Cooperation Base of Child Development and Critical Disorders, Children's Hospital of Chongqing Medical University, Chongqing, China, ³ Yunnan Key Laboratory of Children's Major Disease Research, Kunming Children's Hospital (Children's Hospital Affiliated to Kunming Medical University), Kunming, China, ⁴ Department of Pathology, Kunming Children's Hospital (Children's Hospital Affiliated to Kunming Medical University), Kunming, China, ⁵ Department of Oncology, Yunnan Children Solid Tumor Treatment Center, Kunming Children's Hospital (Children's Hospital Affiliated to Kunming Medical University), Kunming, China

Objective: Papillary renal cell carcinoma (pRCC) is the second most common type of renal cell carcinoma and an important disease affecting older patients. We aimed to establish a nomogram to predict cancer-specific survival (CSS) in elderly patients with pRCC.

Methods: Patient information was downloaded from the Surveillance, Epidemiology, and End Results (SEER) project, and we included all elderly patients with pRCC from 2004 to 2018. All patients were randomly divided into a training cohort and a validation cohort. Univariate and multivariate Cox proportional risk regression models were used to identify patient independent risk factors. We constructed a nomogram based on a multivariate Cox regression model to predict CSS for 1-, 3-, and 5- years in elderly patients with pRCC. A series of validation methods were used to validate the accuracy and reliability of the model, including consistency index (C-index), calibration curve, and area under the Subject operating curve (AUC).

Results: A total of 13,105 elderly patients with pRCC were enrolled. Univariate and multivariate Cox regression analysis suggested that age, tumor size, histological grade, TNM stage, surgery, radiotherapy and chemotherapy were independent risk factors for survival. We constructed a nomogram to predict patients' CSS. The training and validation cohort's C-index were 0.853 (95%CI: 0.859–0.847) and 0.855 (95%CI: 0.865–0.845), respectively, suggesting that the model had good discrimination ability. The AUC showed the same results. The calibration curve also indicates that the model has good accuracy.

Conclusions: In this study, we constructed a nomogram to predict the CSS of elderly pRCC patients, which has good accuracy and reliability and can help doctors and patients make clinical decisions.

Keywords: nomogram, papillary renal cell carcinoma, cancer-specific survival, elderly patients, SEER

BACKGROUND

Renal cell carcinoma (RCC) is the most common Renal malignant tumor in adults, accounting for 90% of renal tumors (1). RCC is divided into three main types based on histological features, with papillary renal cell carcinoma (pRCC) being the second most common type, accounting for ~10 to 15% of the total number of diseases. Clear cell renal cell carcinoma (ccRCC) accounts for 70–80% of these cases, and chromophobe renal cell carcinoma (cRCC) remains in the rest (2, 3). According to pathological features, pRCC is divided into two main subtypes: Type I papillary renal cell carcinoma is characterized by unique basophilic papillary cells. In contrast, Type II is characterized by many papillary cells, and the cytoplasm of type II pRCC is eosinophilic (4). It is worth noting that compared with other RCC, pRCC has special clinical manifestations, biological behaviors and pathological morphology, and its diagnosis and treatment are also different from other RCCs, which are still controversial (5, 6).

Around the world, 400,000 people are diagnosed with RCC every year (1), and the elderly over 60 years old account for more than 75% of the cases (7). In addition, with the aggravation of population aging and the extension of life expectancy, the incidence rate of renal cancer in the elderly is also increasing year by year (8). At present, the prognosis of pRCC is still poor, especially for advanced patients, and there is no effective treatment (9). Therefore, it is particularly important to judge the prognosis of elderly pRCC patients accurately.

Traditionally, TNM staging has been regarded as the main criteria for the prognosis of various malignant tumors. However, it is not enough to cover the biological characteristics of various malignant tumors nor to validate the survival outcome (10). Other clinical variables, such as age, sex, race, grade, surgical treatment, adjuvant therapy, and molecular characteristics, may also impact the outcome of cancer patients.

In recent years, the nomogram prediction model, including UISS (11), SSIGN (12), etc., is considered to be one of the most accurate methods for tumor prediction (13). However, there are no relevant reports of these clinical variables on elderly pRCC cases at the present stage (14). The objective of this retrospective study was to investigate the clinicopathological features associated with the prognosis of elderly pRCC patients collected from the Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute. We then used these features to construct a nomogram to predict cancer-specific survival of patients with pRCC.

PATIENTS AND METHODS

Data Source and Data Extraction

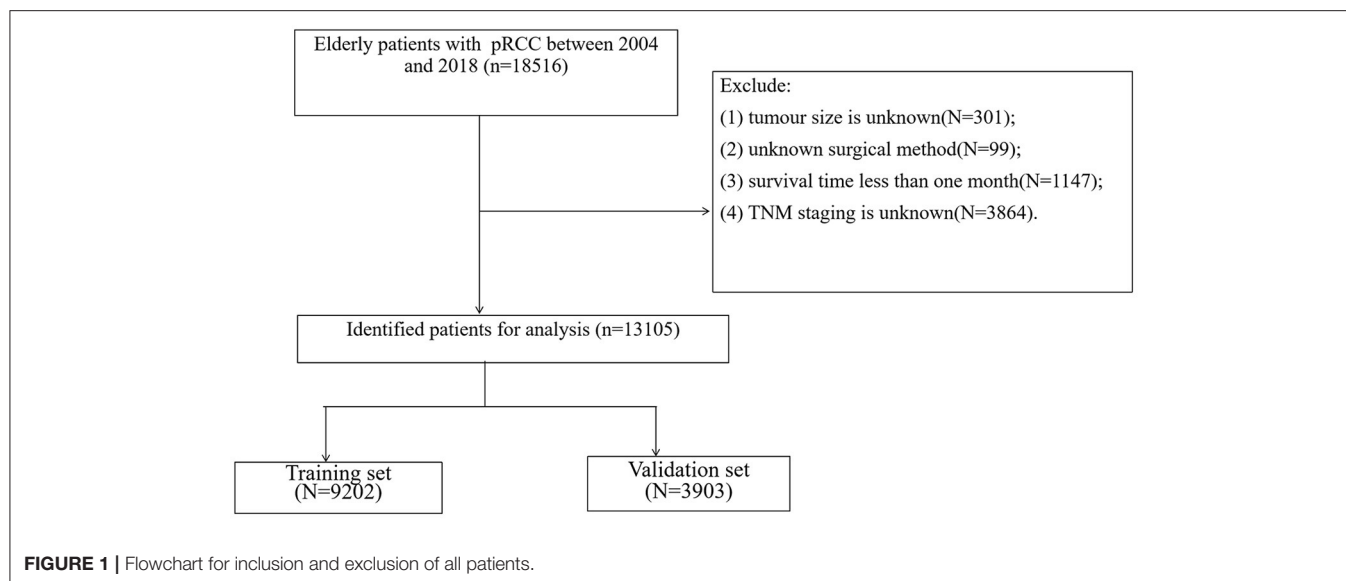
We downloaded clinicopathological information of all patients with pRCC from 2004 to 2018 to the SEER database. SEER data is the national cancer database of the United States, consisting of 18 cancer registries covering ~30% of the national population. Clinicopathological information and follow-up data for all cancer patients are publicly available from the SEER database. Patient personal information is not identifiable, and SEER database information is publicly available, so we do not need to obtain ethical approval and informed consent from patients. Our research methods strictly follow the rules of SEER data.

We collected the basic information of the patient, including age, gender, race, year of diagnosis, marital status; we collected the patient's clinical-pathological information, including the tumor size, laterality, histological grade, TNM staging, surgery, radiation therapy, chemotherapy, patients with follow-up information including living status, the cause of death and survival time. Inclusion criteria: (1) pathological diagnosis of papillary renal cell carcinoma (ICD-O-3 code, 8260); (2) Age ≥ 65 ; (3) Unilateral renal tumor. Exclusion criteria: (1) TNM staging is unknown; (2) Tumor size is unknown; (3) Unknown surgical method; (4) Survival time < 1 month. The screening flow chart of all patients is shown in **Figure 1**.

The patients' marital status was divided into married and unmarried (single, divorced, widowed); Patients' races were divided into white, black, and others (American Indian /AK Native, Asian/Pacific Islander). The years of diagnosis were divided into between 2004 and 2010 and between 2011 and 2018. The histological grades of the patients included grade I (well differentiated), grade II (moderately differentiated), grade III (poorly differentiated), and grade IV (undifferentiated). The surgical classification of patients included non-surgical (surgical code 0), local tumor resection (surgical code 10–27), partial nephrectomy (surgical code 30), and radical nephrectomy (surgical code 40–80).

Nomogram Development and Validation

All patients enrolled were randomly assigned to a training cohort (70%) or a validation cohort (30%). In the training cohort, we used a univariate Cox regression model to pre-screen the influencing factors of patients' prognoses. We then used a multivariate Cox proportional risk regression model to determine the independent risk factors for CSS in patients. Based on a multivariate Cox proportional risk regression model, we constructed a new nomogram to predict CSS at 1-, 3-, and 5 years in patients with pRCC. Then, we use a series of validation methods to test the accuracy and discrimination of the prediction model. We used consistency index (C-index) and



area under the receiver operating curve (AUC) to test the model's discrimination. Calibration curves of 1,000 bootstrap samples were used to validate the model's accuracy.

Clinical Utility

A decision analysis curve (DCA) is a new algorithm to calculate the net benefits of models under different thresholds. DCA was used to validate the clinical utility of the nomogram. In addition, we calculated the value of risk for each patient based on the nomogram and used truncation values to divide all patients into high-risk and low-risk groups. Kaplan-Meier (K-M) curves and log-rank tests were used to determine differences in survival among groups.

Statistical Analysis

Continuous variables (age, tumor size) were described by means and variance, and comparisons between groups were performed by chi-square or non-parametric *U*-tests. Count data were expressed by frequency (%), and a chi-square test was used to compare groups. Univariate and multivariate Cox proportional regression models analyzed the survival and prognostic factors. All statistical analyses were conducted by SPSS 26.0 and R software 4.1.0. A *P* value <0.05 was considered statistically significant.

RESULTS

Clinical Features

Based on inclusion and exclusion criteria, a total of 13,105 elderly patients with pRCC were included. All patients were divided into a training cohort (*N* = 9250) and a validation cohort (*N* = 3855). The mean age of the patients was 75.2 ± 7.57 years, and there were 10936 (83.4%) white patients, 7594 (57.9%) male patients, and 7089 (54.1%) married patients. There were 768 (5.86%) patients at grade I, 2560 (19.5%) at grade II, 1685 (12.9%) at grade III, and 497 (3.79%) at grade

IV. There were 5794 (65.8%) patients with stage T1a, 11983 (91.4%) patients with stage N0, and 10665 (81.4%) patients with stage M0. Local tumor excision, partial nephrectomy and radical nephrectomy were performed in 1269 (9.68%), 1519 (11.6%), and 4521 (34.5%) patients, respectively. 1,085 (8.28%) patients underwent chemotherapy, and 638 (4.87%) patients underwent radiotherapy. The clinicopathological information of all patients was shown in **Table 1**, and there was no significant difference between the training and validation cohorts.

Univariate and Multivariate Cox Regression Analysis

We analyzed patient prognostic factors using univariate and multivariable Cox regression models. The univariate Cox regression model showed that age, year of diagnosis, race, marriage, histological grade, tumor size, TNM stage, surgery, radiotherapy, and chemotherapy influenced patients' CSS. Multivariate Cox regression analysis showed that age, histological grade, TNM stage, tumor size, surgery, radiotherapy and chemotherapy were prognostic factors affecting patients' CSS. Cox regression analysis results are shown in **Table 2**.

Nomogram Construction for 1, 3, and 5-Year CSS

The essence of the nomogram is to visualize the multivariate Cox regression analysis. Therefore, we constructed a nomogram based on multivariate Cox regression analysis to predict CSS in elderly patients with pRCC (**Figure 2**). As shown in the figure, tumor size and TNM stage are the biggest factors affecting the prognosis of patients, followed by surgery, radiotherapy and chemotherapy. In addition, age and histological grade are also important factors. The larger the tumor, the higher the risk of death, and the higher the TNM stage, the higher the risk of death. Patients with partial nephrectomy had the lowest risk, and

TABLE 1 | Clinicopathological characteristics of elderly patients with pRCC.

	All	Training cohort	Validation cohort	
	<i>N</i> = 13105	<i>N</i> = 9,202	<i>N</i> = 3,903	<i>p</i>
Age				0.024
65–74	6,847 (52.2%)	4,762 (51.7%)	2,085 (53.4%)	
75–84	4,432 (33.8%)	3,110 (33.8%)	1,322 (33.9%)	
≥85	1,826 (13.9%)	1,330 (14.5%)	496 (12.7%)	
Race				0.404
White	10,936 (83.4%)	7,658 (83.2%)	3,278 (84.0%)	
Black	1,444 (11.0%)	1,036 (11.3%)	408 (10.5%)	
Other	725 (5.53%)	508 (5.52%)	217 (5.56%)	
Sex				0.337
Male	7,594 (57.9%)	5,307 (57.7%)	2,287 (58.6%)	
Female	5,511 (42.1%)	3,895 (42.3%)	1,616 (41.4%)	
Marital				0.002
Married	7,088 (54.1%)	4,885 (53.1%)	2,203 (56.4%)	
Unmarried or Domestic Partner/Single	1,874 (14.3%)	1,341 (14.6%)	533 (13.7%)	
Separated/Divorced/ Widowed	4,143 (31.6%)	2,976 (32.3%)	1,167 (29.9%)	
Year of diagnosis				0.683
2004–2010	6,125 (46.7%)	4,312 (46.9%)	1,813 (46.5%)	
2010–2018	6,980 (53.3%)	4,890 (53.1%)	2,090 (53.5%)	
Laterality				0.660
Left	6,440 (49.1%)	4,510 (49.0%)	1,930 (49.4%)	
Right	6,665 (50.9%)	4,692 (51.0%)	1,973 (50.6%)	
Grade				0.652
I	768 (5.86%)	531 (5.77%)	237 (6.07%)	
II	2,560 (19.5%)	1,785 (19.4%)	775 (19.9%)	
III	1,685 (12.9%)	1,167 (12.7%)	518 (13.3%)	
IV	497 (3.79%)	347 (3.77%)	150 (3.84%)	
Unknown	7,595 (58.0%)	5,372 (58.4%)	2,223 (57.0%)	
T				0.925
T1a	5,794 (44.2%)	4,070 (44.2%)	1,724 (44.2%)	
T1b	3,011 (23.0%)	2,121 (23.0%)	890 (22.8%)	
T2	1,606 (12.3%)	1,137 (12.4%)	469 (12.0%)	
T3	2,607 (19.9%)	1,813 (19.7%)	794 (20.3%)	
T4	87 (0.66%)	61 (0.66%)	26 (0.67%)	
N				0.295
N0	11,983 (91.4%)	8,430 (91.6%)	3,553 (91.0%)	
N1	1,122 (8.56%)	772 (8.39%)	350 (8.97%)	
M				0.724
M0	10,665 (81.4%)	7,481 (81.3%)	3,184 (81.6%)	
M1	2,440 (18.6%)	1,721 (18.7%)	719 (18.4%)	
Tumor size				0.963
<40 mm	6,109 (46.6%)	4,284 (46.6%)	1,825 (46.8%)	
41–80 mm	4,680 (35.7%)	3,293 (35.8%)	1,387 (35.5%)	
>80 mm	2,316 (17.7%)	1,625 (17.7%)	691 (17.7%)	
Surgery				0.125
No	5,796 (44.2%)	4,110 (44.7%)	1,686 (43.2%)	
Local tumor excision	1,269 (9.68%)	911 (9.90%)	358 (9.17%)	
Partial nephrectomy	1,519 (11.6%)	1,048 (11.4%)	471 (12.1%)	
Radical nephrectomy	4,521 (34.5%)	3,133 (34.0%)	1,388 (35.6%)	
Chemotherapy				1.000
No/Unknown	12,020 (91.7%)	8,440 (91.7%)	3,580 (91.7%)	
Yes	1,085 (8.28%)	762 (8.28%)	323 (8.28%)	
Radiation				0.757
No/Unknown	12,467 (95.1%)	8,758 (95.2%)	3,709 (95.0%)	
Yes	638 (4.87%)	444 (4.83%)	194 (4.97%)	

TABLE 2 | Proportional subdistribution hazard analyses of CSS in training cohort.

	CSS		
	HR	95%CI	P
Age			
65–74			
75–84	1.20	1.09–1.32	<0.001
≥85	1.50	1.32–1.7	<0.001
Race			
White			
Black	0.94	0.81–1.08	0.35
Other	0.89	0.75–1.06	0.18
Sex			
Male			
Female	0.87	0.79–0.95	0.001
Marital			
Married			
Unmarried or Domestic Partner/Single	1.05	0.93–1.19	0.4
Separated/Divorced/ Widowed	1.10	1–1.21	0.56
Year of diagnosis			
2004–2010			
2010–2018	0.88	0.81–0.95	0.002
Laterality			
Left			
Right	1.08	1–1.17	0.057
Grade			
I			
II	0.95	0.73–1.23	0.7
III	1.37	1.06–1.78	0.017
V	1.76	1.31–2.37	<0.001
Unknown	1.19	0.92–1.52	0.18
T			
T1a			
T1b	1.56	1.19–2.04	0.001
T2	2.00	1.54–2.6	<0.001
T3	2.38	1.87–3.04	<0.001
T4	2.03	1.24–3.32	0.005
N			
N0			
N1	1.49	1.32–1.68	<0.001
M			
M0			
M1	4.32	3.84–4.87	<0.001
Tumor size			
<40 mm			
41–80 mm	1.26	0.99–1.59	0.06
>80 mm	1.44	1.13–1.82	0.003
Surgery			
No			
Local tumor excision	0.47	0.37–0.58	<0.001
Partial nephrectomy	0.30	0.24–0.39	<0.001
Radical nephrectomy	0.49	0.42–0.56	<0.001
Chemotherapy			
No/Unknown			
Yes	0.99	0.88–1.12	0.92
Radiation			
No/Unknown			
Yes	1.20	1.04–1.38	0.013

patients without surgery had the highest risk. In addition, the older the patient, the higher the risk of death.

Validation of the Nomogram

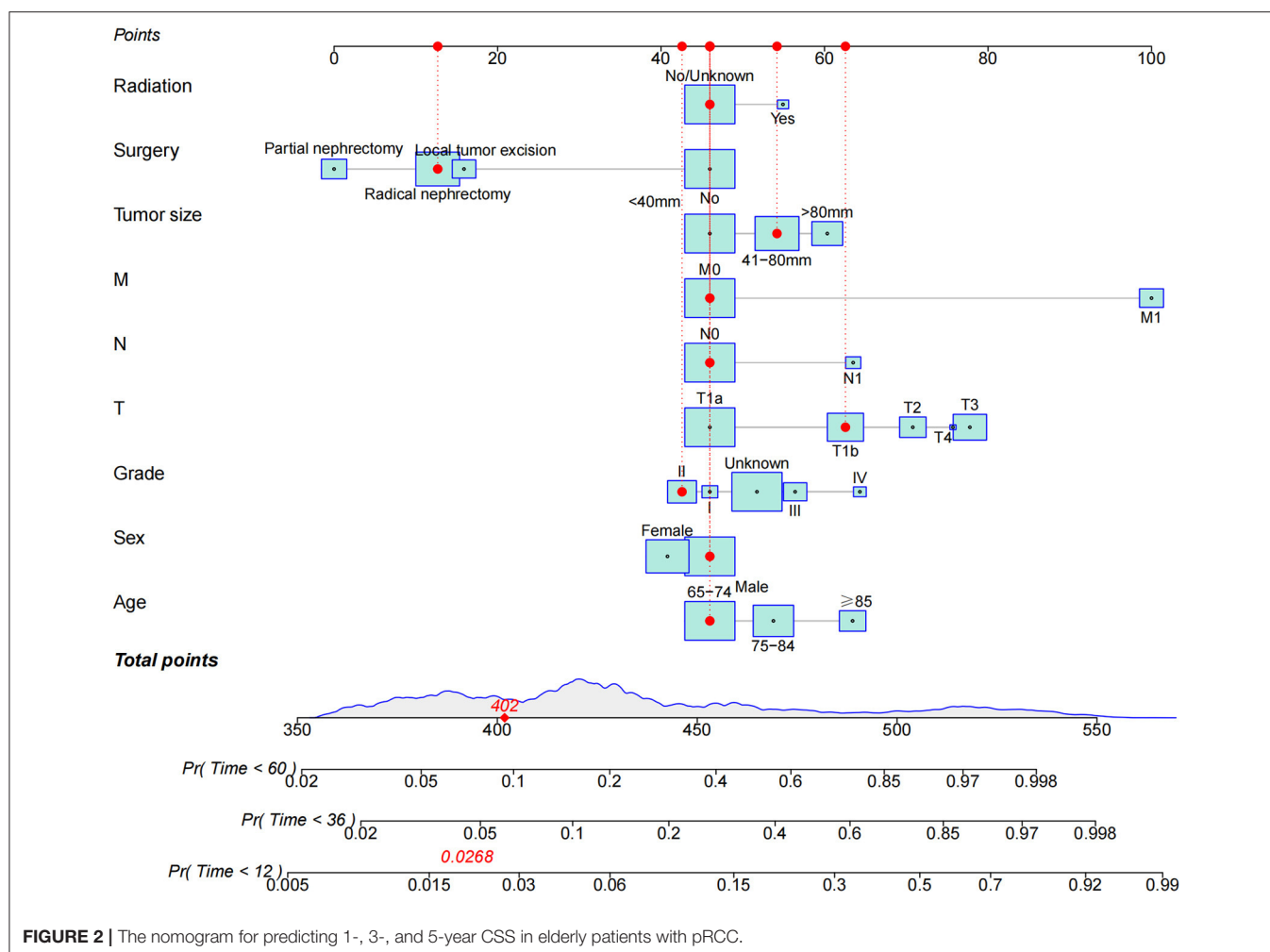
We first use the C-index to validate the discrimination of the prediction model. In the training cohort and validation cohort, the C-index was 0.853 (95%CI: 0.859–0.847) and 0.855 (95%CI: 0.865–0.845), respectively. The results showed that the nomogram had good discrimination. The calibration curve was also used to validate the accuracy of the model. The calibration curve showed that the predicted value of the nomogram was highly consistent with the actual observed value, indicating that the prediction model had good accuracy (Figure 3). In the training cohort, the nomogram's 1-, 3- and 5-year AUC values were 91.5, 91.5 and 90.2, respectively. In the validation cohort, the nomogram's 1-, 3- and 5-year AUC values were 92.1, 91.2 and 90.3, respectively. It shows that the nomogram has good discrimination (Figure 4).

Clinical Application of the Nomogram

DCA was used to test the clinical application value of the prediction model. DCA showed that the nomogram had potential clinical application value and was more practical than the traditional TNM staging (Figure 5). Based on the nomogram, we calculated the risk values of all patients and divided them into the high-risk group using ROC cut-off values (total score > 95.7) and the low-risk group (total score ≤95.7). The K-M curve showed that the survival rate of patients in the high-risk group was significantly lower than that in the low-risk group (Figure 6). In the high-risk group, 1-, 3-, and 5-year survival rates were 64.7, 47.9, and 42.2%, respectively. In the low-risk group, 1-, 3-, and 5-year survival rates were 98.4, 95.7, and 92.2%, respectively. In addition, we analyzed surgical procedures in the high-risk and low-risk groups. In the low-risk group, survival was highest in patients who received partial nephrectomy and lowest in radical nephrectomy. In the high-risk group, survival was highest who underwent radical nephrectomy and lowest for those who did not (Figure 7).

DISCUSSION

RCC accounts for about 2% of all cancer diagnoses and deaths worldwide, with higher rates in developed countries. RCC is the most common type of renal malignancy, accounting for more than 90%. pRCC accounts for 10–20% of all renal cell carcinomas. However, compared with other types of RCC, pRCC lacks specific clinical manifestations and associated symptoms, and more importantly, pRCC does not have typical radiographic findings. In addition, some elderly patients may present with perirenal abscesses due to weakened immunity. It brings great difficulties to the diagnosis and treatment of pRCC for clinicians (15). According to recent reports, the overall prognosis of pRCC is slightly better than that of clear cell renal carcinoma and chromophobe renal carcinoma (16). However, in clinical practice, in addition to TNM staging, there is currently a lack of a model that can accurately predict the prognosis of elderly patients with pRCC.

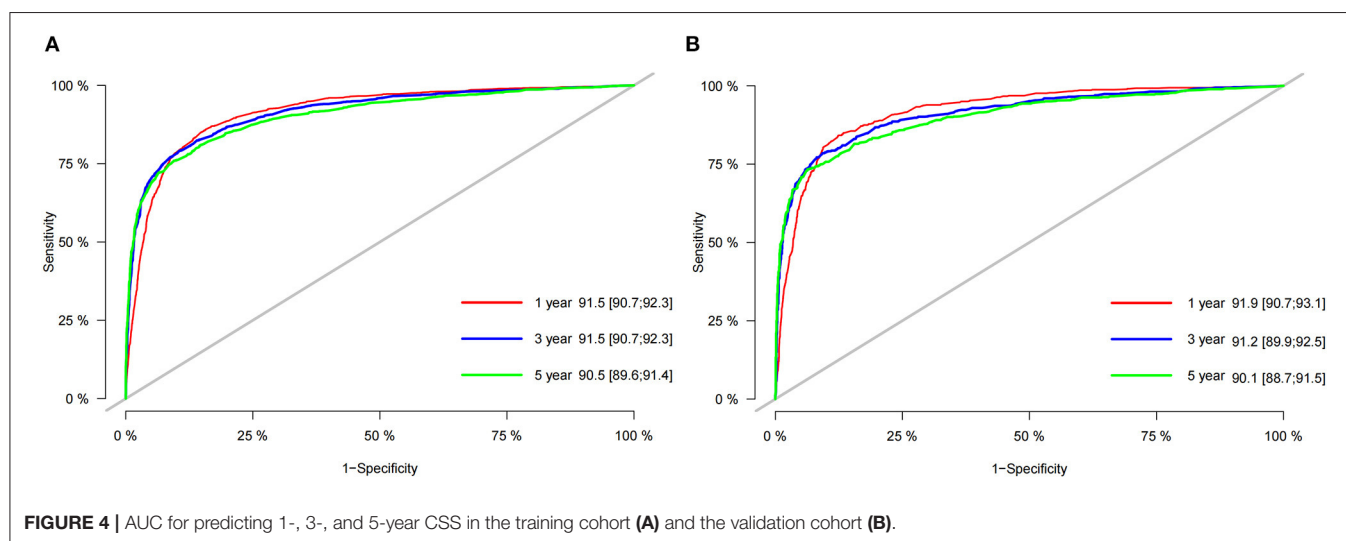
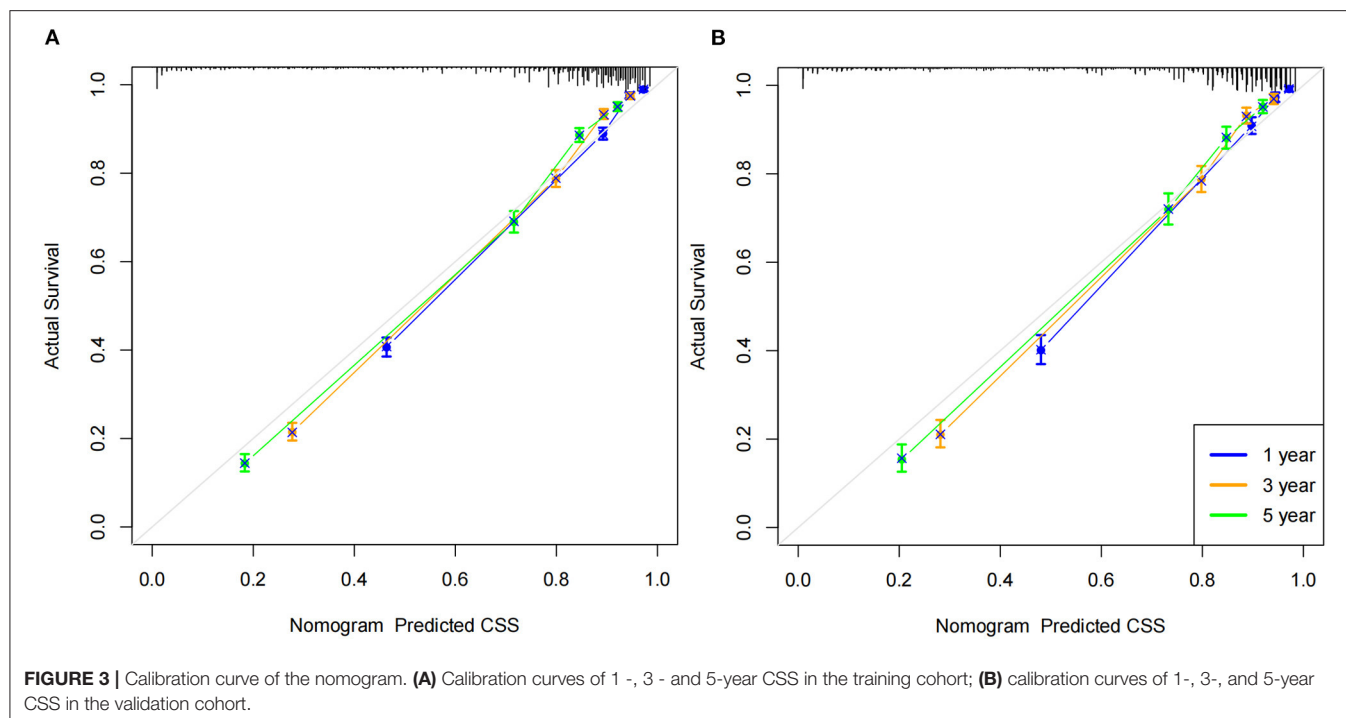


Nomogram is a data-based graphical computing tool that can estimate the risk of a disease based on staging systems such as the American Joint Commission on Cancer (AJCC) and other key risk factors related to prognosis (17). Compared with traditional TMN staging, nomogram has better accuracy in prognostic prediction and can provide better advice and help for clinicians in diagnosis and treatment (18). To our knowledge, there have been no reports on the prognosis of elderly patients with pRCC. In addition, due to the relatively low incidence of pRCC, it is difficult to collect a large sample size for single-center studies of this disease to draw reliable conclusions (19). Therefore, it is particularly important to establish a more reliable and accurate predictive model for pRCC in the elderly. This study collected data from the SEER Database, a large sample database established in 1973. At present, the database covers 18 countries and regions, effectively avoiding the lack of sample size and single type (20).

In this study, we established and validated a new nomogram to accurately predict CSS in elderly pRCC. Previous studies have found that pRCC has a higher incidence and worse survival rate in elderly patients (21). Our study also confirmed that age is a key factor in the development of pRCC in the elderly. As we

age, it is well known that the risk of genetic mutations leading to cancer increases. Studies have shown that age plays a key role in the survival rate of various cancers (22, 23). Huang et al. found by propensity matching comparison that pRCC had a significantly worse prognosis than ccRCC in patients aged ≤ 45 years (24). Su et al. collected the SEER database of pRCC patients who underwent nephrectomy from 2010 to 2016 for analysis. They confirmed that age is a key factor influencing the all-cause mortality of pRCC (25). The study of Nelson et al. also found that the survival rate of mRCC patients aged ≥ 75 years was significantly lower than that of patients aged < 75 years (26). There is no consensus on defining the age of elderly patients, but more than 60% of initial cancer diagnoses and more than 70% of cancer deaths occur in patients over 65 years old (8). To improve the accuracy and representativeness of the prediction model, pRCC patients over 65 years old were included in this study.

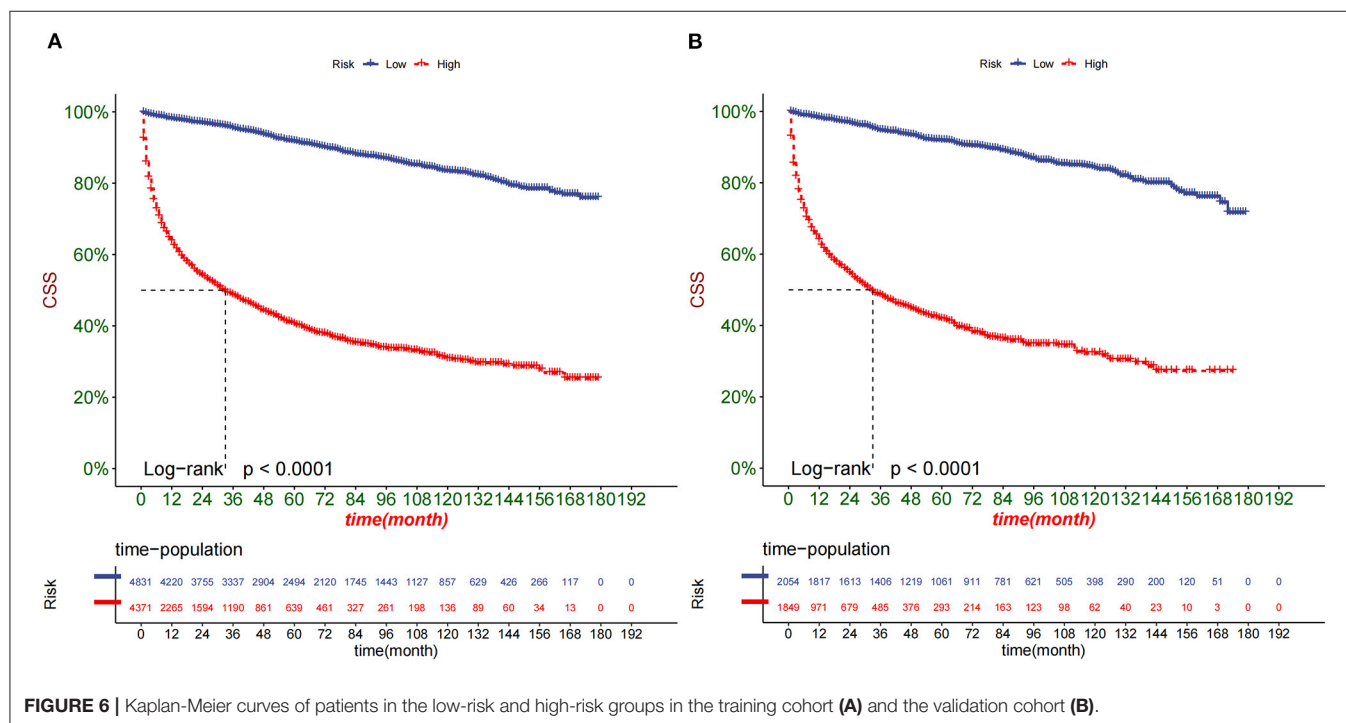
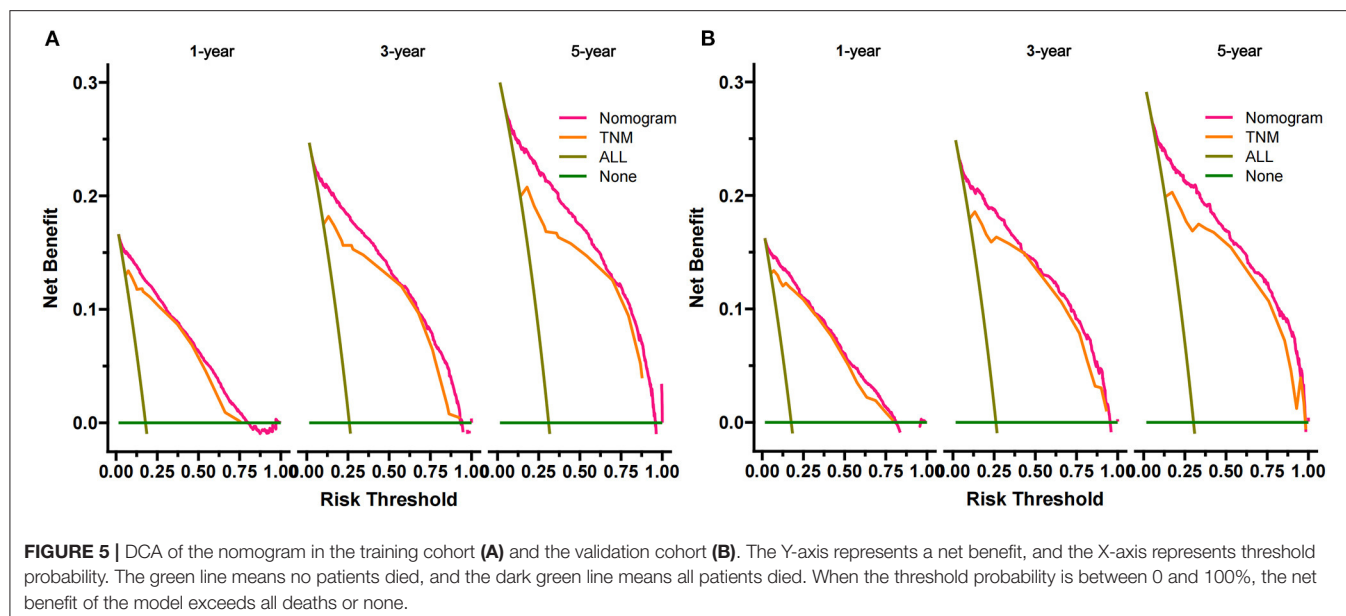
At the same time, we found that tumor size is a major risk factor affecting the prognosis of pRCC in the elderly, and larger tumor occurrence often suggests poor prognosis, which is consistent with the results of previous studies. Hutterer et al. previously established a nomogram to predict the survival rate of



RCC and found that tumor size was an important risk factor (27). Zastrow et al. also found that tumor size was a risk factor for the long-term survival of pRCC (28).

As is known to all, the TNM staging system is a common method for clinical evaluation of various malignant tumors, which helps to judge the prognosis of cancer patients and guide clinicians to take better treatment (29, 30). However, only the size of the tumor, the presence of lymph node metastasis, and distant metastasis were used as criteria. Age, marital status, surgical method, chemotherapy and radiotherapy, and other important factors that have been proven to affect cancer patients' overall survival rate (OS) were ignored (31). Our study found that in

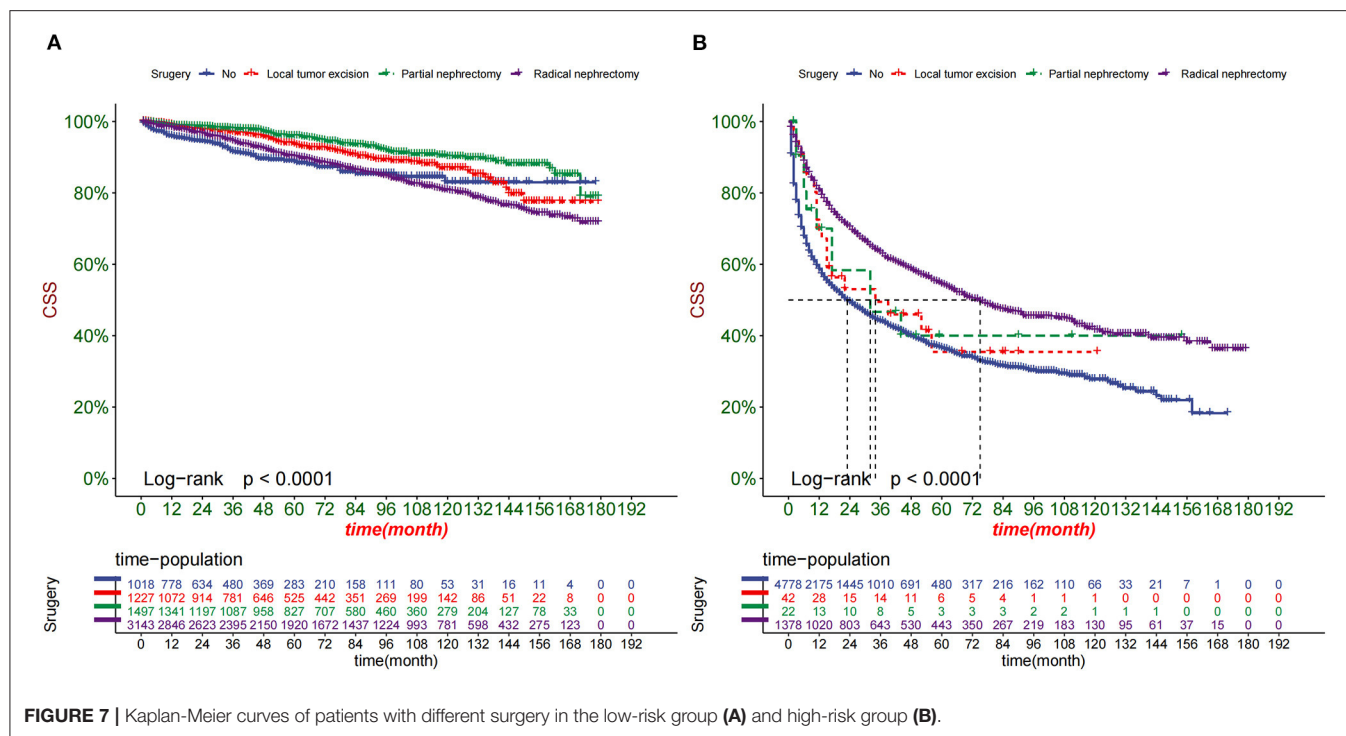
elderly patients undergoing pRCC surgery, partial nephrectomy (PN) had the best prognosis, radical nephrectomy (RN) was intermediate, and local tumor resection had the worst prognosis. It is consistent with most research conclusions. Shum et al. showed that in T2 stage malignancies, the OS of PN was significantly better than that of RN (32). Hellenthal et al. collected RCC patients from 1988 to 2005 in the SEER database. After analysis, it was concluded that PN could still significantly improve OS even with tumor metastasis, benefiting mRCC patients (33). In recent years, postoperative radiotherapy has been gradually included in various cancer guidelines because of its good effect as a key means of postoperative treatment.



RCC is sensitive to radiotherapy, and the strategy has been agreed upon.

Interestingly, we found that postoperative chemotherapy did not improve CSS in elderly patients with pRCC, which is consistent with Tachibana and De Vries-Brilland et al. The former retrospectively analyzed RCC patients who received nivolumab and ipilimumab as a first-line treatment between December 2015 and May 2020 and

found that the chemotherapy regimen achieved good results in ccRCC, but intermediate results in pRCC (34). The latter summarized the treatment methods of pRCC and concluded that the existing chemotherapy regimens were not sensitive to pRCC. The combination of immune checkpoint inhibitors (ICI) and tyrosine kinase inhibitors (MET) may be a new direction for the treatment of pRCC in the future (35).



Finally, the newly constructed nomogram model for predicting CSS in elderly patients with pRCC includes many factors, such as diagnosis age, tumor size, TNM grade, Fuhrman grade, and operation at the primary site, which is convenient for clinical information collection. In summary, the nomograms we developed can accurately predict CSS at 1, 3, and 5 years in patients with pRCC. Furthermore, we used AUC, C-index, and DCA to validate its accuracy and predictive power for elderly papillary renal cell carcinoma.

However, there are still some limitations in this study. First of all, the SEER database does not include BMI, smoking, alcohol consumption, etc. These are important factors affecting patients' survival. However, we included the basic patient information cohort, tumor information, and other key factors. Secondly, because this study is retrospective, there is inevitable selection bias. Finally, the prediction model is only validated internally, and further external validation is necessary to validate the model's accuracy.

CONCLUSION

In this study, we explored the prognostic factors of elderly pRCC patients and the patient's age, histological grade, TNM stage, tumor size, surgery, radiotherapy, and chemotherapy as independent risk factors affecting patients CSS. We constructed a nomogram to predict the CSS of elderly pRCC patients with good accuracy and reliability, which can help doctors and patients make clinical decisions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://seer.Cancer.gov/>.

ETHICS STATEMENT

The data of this study is obtained from the SEER database. The patients' data is public, so this study does not require ethical approval and informed consent.

AUTHOR CONTRIBUTIONS

JW and CZ designed the study. CZ, JW, LL, YX, and HT collected and analyzed the data. JW drafted the initial manuscript. CZ, KZ, and BY revised the article critically. CZ, ZY, and BY reviewed and edited the article. All authors approved the final manuscript.

FUNDING

This study was supported by Yunnan Education Department of Science Research Fund (No. 2020 J0228), Kunming City Health Science and Technology Talent "1000" Training Project (No. 2020- SW (Reserve)-112), Kunming Health and Health Commission Health Research Project (No. 2020-0201-001), and Kunming Medical Joint Project of Yunnan Science and Technology Department (No. 202001 AY070001-271). The funding bodies played no role in the study's design and collection, analysis and interpretation of data, and writing the manuscript.

REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660
- Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. *Nat Rev Dis Primers.* (2017) 3:17009. doi: 10.1038/nrdp.2017.9
- Shuch B, Amin A, Armstrong AJ, Eble JN, Ficarra V, Lopez-Beltran A, et al. Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. *Eur Urol.* (2015) 67:85–97. doi: 10.1016/j.eururo.2014.04.029
- Dralle H, Musholt TJ, Schabram J, Steinmüller T, Frilling A, Simon D, et al. German association of endocrine surgeons practice guideline for the surgical management of malignant thyroid tumors. *Langenbecks Arch Surg.* (2013) 398:347–75. doi: 10.1007/s00423-013-1057-6
- Daugherty M, Sedaghatpour D, Shapiro O, Vourganti S, Kutikov A, Bratslavsky G. The metastatic potential of renal tumors: influence of histologic subtypes on definition of small renal masses, risk stratification, and future active surveillance protocols. *Urol Oncol.* (2017) 35:153. doi: 10.1016/j.urolonc.2016.11.009
- Quivy A, Daste A, Harbaoui A, Duc S, Bernhard JC, Gross-Goupil M, et al. Optimal management of renal cell carcinoma in the elderly: a review. *Clin Interv Aging.* (2013) 8:433–42. doi: 10.2147/CIA.S30765
- González León T, Morera Pérez M. Renal cancer in the elderly. *Curr Urol Rep.* (2016) 17:6. doi: 10.1007/s11934-015-0562-2
- Durinck S, Stawiski EW, Pavia-Jiménez A, Modrusan Z, Kapur P, Jaiswal BS, et al. Spectrum of diverse genomic alterations define non-clear cell renal carcinoma subtypes. *Nat Genet.* (2015) 47:13–21. doi: 10.1038/ng.3146
- Park YH, Lee SJ, Cho EY, La Choi Y, Lee JE, Nam SJ, et al. Clinical relevance of TNM staging system according to breast cancer subtypes. *Ann Oncol.* (2019) 30:2011. Erratum for: *Ann Oncol.* (2011). 22:1554–60. doi: 10.1093/annonc/mdq617
- Capogrosso P, Larcher A, Sjoberg DD, Vertosick EA, Cianflone F, Dell'Oglio P, et al. Risk based surveillance after surgical treatment of renal cell carcinoma. *J Urol.* (2018) 200:61–67. doi: 10.1016/j.juro.2018.01.072
- Parker WP, Cheville JC, Frank I, Zaid HB, Lohse CM, Boorjian SA, et al. Application of the stage, size, grade, and necrosis (ssign) score for clear cell renal cell carcinoma in contemporary patients. *Eur Urol.* (2017) 71:665–73. doi: 10.1016/j.eururo.2016.05.034
- Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. *Lancet Oncol.* (2015) 16:e173–80. doi: 10.1016/S1470-2045(14)71116-7
- Ruddy KJ, Winer EP. Male breast cancer: risk factors, biology, diagnosis, treatment, and survivorship. *Ann Oncol.* (2013) 24:1434–43. doi: 10.1093/annonc/mdt025
- Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, et al. Nivolumab versus everolimus in advanced renal-cell carcinoma. *N Engl J Med.* (2015). 373:1803–13. doi: 10.1056/NEJMoa1510665
- Lu Y, Huang H, Kang M, Yi M, Yang H, Wu S, et al. Combined Ki67 and ERCC1 for prognosis in non-keratinizing nasopharyngeal carcinoma underwent chemoradiotherapy. *Oncotarget.* (2017) 8:8852–562. doi: 10.18632/oncotarget.19158
- Tang J, Wang J, Pan X. A web-based prediction model for overall survival of elderly patients with malignant bone tumors: a population-based study. *Front Public Health.* (2022) 9:812395. doi: 10.3389/fpubh.2021.812395
- Duan J, Xie Y, Qu L, Wang L, Zhou S, Wang Y, et al. A nomogram-based immunoprofile predicts overall survival for previously untreated patients with esophageal squamous cell carcinoma after esophagectomy. *J Immunother Cancer.* (2018) 6:100. doi: 10.1186/s40425-018-0418-7
- Capitanio U, Bensalah K, Bex A, Boorjian SA, Bray F, Coleman J, et al. Epidemiology of Renal Cell Carcinoma. *Eur Urol.* (2019) 75:74–84. doi: 10.1016/j.eururo.2018.08.036
- Morgan TM, Mehra R, Tiemeny P, Wolf JS, Wu S, Sangale Z, et al. A multigene signature based on cell cycle proliferation improves prediction of mortality within 5 yr of radical nephrectomy for renal cell carcinoma. *Eur Urol.* (2018) 73:763–9. doi: 10.1016/j.eururo.2017.12.002
- Mejean A, Hopirtean V, Bazin JP, Larousserie F, Benoit H, Chrétien Y, et al. Prognostic factors for the survival of patients with papillary renal cell carcinoma: meaning of histological typing and multifocality. *J Urol.* (2003) 170:764–7. doi: 10.1097/01.ju.0000081122.57148.ec
- Escudier B, Porta C, Schmidinger M, Rioux-Leclercq N, Bex A, Khoo V, et al. Electronic address: clinicalguidelines@esmo.org. Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up[†]. *Ann Oncol.* (2019) 30:706–20. doi: 10.1093/annonc/mdz056
- Dias-Santos D, Ferrone CR, Zheng H, Lillemoe KD, Fernández-Del Castillo C. The Charlson age comorbidity index predicts early mortality after surgery for pancreatic cancer. *Surgery.* (2015) 157:881–7. doi: 10.1016/j.surg.2014.12.006
- Huang J, Huang D, Yan J, Chen T, Gao Y, Xu D, et al. Comprehensive subgroup analyses of survival outcomes between clear cell renal cell adenocarcinoma and papillary renal cell adenocarcinoma. *Cancer Med.* (2020) 9:9409–418. doi: 10.1002/cam4.3563
- Su X, Hou NN, Yang LJ, Li PX, Yang XJ, Hou GD, et al. The first competing risk survival nomogram in patients with papillary renal cell carcinoma. *Sci Rep.* (2021) 11:11835. doi: 10.1038/s41598-021-91217-z
- Nelson RA, Vogelzang N, Pal SK. A gap in disease-specific survival between younger and older adults with de novo metastatic renal cell carcinoma: results of a SEER database analysis. *Clin Genitourin Cancer.* (2013) 11:303–10. doi: 10.1016/j.clgc.2013.04.011
- Hutterer GC, Patard JJ, Jeldres C, Perrotte P, de La Taille A, Salomon L, et al. Patients with distant metastases from renal cell carcinoma can be accurately identified: external validation of a new nomogram. *BJU Int.* (2008) 101:39–43. doi: 10.1111/j.1464-410X.2007.07170.x
- Zastrow S, Phuong A, von Bar I, Novotny V, Hakenberg OW, Wirth MP. Primary tumor size in renal cell cancer in relation to the occurrence of synchronous metastatic disease. *Urol Int.* (2014) 92:462–7. doi: 10.1159/000356325
- Zhou H, Zhang Y, Song Y, Tan W, Qiu Z, Li S, et al. Marital status is an independent prognostic factor for pancreatic neuroendocrine tumors patients: an analysis of the surveillance, epidemiology, and end results (SEER) database. *Clin Res Hepatol Gastroenterol.* (2017) 41:476–86. doi: 10.1016/j.clinre.2017.02.008
- Zhang G, Wu Y, Zhang J, Fang Z, Liu Z, Xu Z, et al. Nomograms for predicting long-term overall survival and disease-specific survival of patients with clear cell renal cell carcinoma. *Onco Targets Ther.* (2018) 11:5535–5544. doi: 10.2147/OTT.S171881
- Wang J, Liu X, Tang J, Zhang Q, Zhao Y. A web-based prediction model for cancer-specific survival of elderly patients with hypopharyngeal squamous cell carcinomas: a population-based study. *Front Public Health.* (2022) 9:815631. doi: 10.3389/fpubh.2021.815631
- Shum CF, Bahler CD, Sundaram CP. Matched comparison between partial nephrectomy and radical nephrectomy for T2 N0 M0 Tumors, a study based on the national cancer database. *J Endourol.* (2017) 31:800–805. doi: 10.1089/end.2017.0190
- Hellenthal NJ, Mansour AM, Hayn MH, Schwaab T. Is there a role for partial nephrectomy in patients with metastatic renal cell carcinoma? *Urol Oncol.* (2013) 31:36–41. doi: 10.1016/j.urolonc.2010.08.026
- Tachibana H, Kondo T, Ishihara H, Fukuda H, Yoshida K, Takagi T, et al. Modest efficacy of nivolumab plus ipilimumab in patients with papillary renal cell carcinoma. *Jpn J Clin Oncol.* (2021) 51:646–53. doi: 10.1093/jjco/hyaa229

35. de Vries-Brilland M, McDermott DE, Suárez C, Powles T, Gross-Goupil M, Ravaud A, et al. Checkpoint inhibitors in metastatic papillary renal cell carcinoma. *Cancer Treat Rev.* (2021) 99:102228. doi: 10.1016/j.ctrv.2021.102228

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of

the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhanghuang, Wang, Yao, Li, Xie, Tang, Zhang, Wu, Yang and Yan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Microcalcification Discrimination in Mammography Using Deep Convolutional Neural Network: Towards Rapid and Early Breast Cancer Diagnosis

Yew Sum Leong¹, Khairunnisa Hasikin^{1,2*}, Khin Wee Lai¹, Norita Mohd Zain¹ and Muhammad Mokhzaini Azizan^{3*}

¹ Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia, ² Department of Biomedical Engineering, Center for Image and Signal Processing (CISIP), Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia, ³ Department of Electrical and Electronic Engineering, Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, Nilai, Malaysia

OPEN ACCESS

Edited by:

Yu-Hsiu Lin,
National Chung Cheng
University, Taiwan

Reviewed by:

Raffaella Massafra,
National Cancer Institute Foundation
(IRCCS), Italy
Mohammad Kamrul Hasan,
Universiti Kebangsaan
Malaysia, Malaysia

*Correspondence:

Khairunnisa Hasikin
khairunnisa@um.edu.my
Muhammad Mokhzaini Azizan
mokhzainiazizan@usim.edu.my

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 14 February 2022

Accepted: 04 April 2022

Published: 28 April 2022

Citation:

Leong YS, Hasikin K, Lai KW, Mohd
Zain N and Azizan MM (2022)
Microcalcification Discrimination in
Mammography Using Deep
Convolutional Neural Network:
Towards Rapid and Early Breast
Cancer Diagnosis.
Front. Public Health 10:875305.
doi: 10.3389/fpubh.2022.875305

Breast cancer is among the most common types of cancer in women and under the cases of misdiagnosed, or delayed in treatment, the mortality risk is high. The existence of breast microcalcifications is common in breast cancer patients and they are an effective indicator for early sign of breast cancer. However, microcalcifications are often missed and wrongly classified during screening due to their small sizes and indirect scattering in mammogram images. Motivated by this issue, this project proposes an adaptive transfer learning deep convolutional neural network in segmenting breast mammogram images with calcifications cases for early breast cancer diagnosis and intervention. Mammogram images of breast microcalcifications are utilized to train several deep neural network models and their performance is compared. Image filtering of the region of interest images was conducted to remove possible artifacts and noises to enhance the quality of the images before the training. Different hyperparameters such as epoch, batch size, etc were tuned to obtain the best possible result. In addition, the performance of the proposed fine-tuned hyperparameter of ResNet50 is compared with another state-of-the-art machine learning network such as ResNet34, VGG16, and AlexNet. Confusion matrices were utilized for comparison. The result from this study shows that the proposed ResNet50 achieves the highest accuracy with a value of 97.58%, followed by ResNet34 of 97.35%, VGG16 96.97%, and finally AlexNet of 83.06%.

Keywords: transfer learning, region of interest (ROI), intervention, machine learning, artificial intelligence

INTRODUCTION

In 2020, World Health Organization (WHO) reported 2.3 million cases of breast cancer worldwide with over 685,000 fatalities, making it among the highest fatal diseases in the world. Although extensive efforts on breast cancer screening have shown promising results for early intervention, localizing breast lesions has remained a challenge. This is because detection of breast lesions on mammogram images heavily depended on the radiologist's skill (1), which proved to be time consuming, and at times lacked the accuracy and precision. Thus, this factor poses a serious

challenge onto rapid diagnosis process which in the case of breast cancer, late detection may prove terminal. Advancements and involvement of artificial intelligence (AI) in the healthcare sector have improved accuracy and assisted radiologists by minimizing the rates of false positives and false negatives during clinical diagnosis. Deep Convolutional Neural Networks (D-CNN), a subsidiary of AI, have advanced to the point where they can automatically learn from enormous picture data sets and detect abnormalities in mammograms such as mass lesions (2). D-CNN has quickly become the preferred approach for evaluating medical images to aid the early detection of breast cancer diseases, which resulted in a favourable prognosis and a higher percentage of survival (3, 4).

The presence of microcalcification during breast cancer screening is often missed due to its small size which is approximately 0.1–1.0 mm. In addition, it may be scattered and less visible to naked eyes due to the surrounding dense breast tissues. Different from microcalcification, breast lump has a relatively high predictive value for malignancy (5, 6). Calcifications may appear as white dots with specific patterns, size, density, and location on mammogram images, which might signify breast cancer or precancerous alterations in breast tissue (7). Even with visible calcifications, most lesions are not recalled immediately but identified as interval cancer in subsequent screening due to the poor sensitivity of screening for malignant calcifications (8). This is due to the low contrast and unclear boundaries on conventional images of breast mammograms (9). According to WHO, the survival probabilities of breast cancer patients may reach an astonishing number of 90% if the disease is identified and treated effectively in early stages.

Generally, in terms of detection, diagnosis, and treatment, many healthcare providers are faced with problems such as a lack of human resources and technological capabilities to deliver timely care to breast cancer patients (10). This problem worsens in developing and under-developed countries, where inexperienced radiologists are faced with a myriad number of mammogram images during screening. Therefore, the emergence of current computer-aided diagnosis (CAD) systems aids breast cancer diagnosis by allowing more comprehensive and objective analyses to be performed on many mammogram images. However, the CAD system is mostly based on hand-crafted features. The prognostic choice on the categorization of microcalcification clusters is mostly based on extracting useful handmade characteristics and then creating a highly discriminative classifier on top of them, which frequently yielded false results (11). Also, the installation of a sophisticated computer program in healthcare usually necessitates a multi-pronged strategy as it often involves political, economic, and social issues (12, 13).

The use of AI as an automated image classification tool has increased over the years as it allows automated disease diagnosis, characterization of histology, stage, or subtype, and patient classification based on therapeutic outcome or prognosis (14). Many types of diseases have incorporated the use of AI to form an automated prediction system. As such, the use of the Hippocampal Unified Multi-Atlas Network (HUMAN) algorithm to diagnose Alzheimer's disease (AD) (15). Current

algorithms normally utilize transfer learning techniques or pre-trained CNNs to reduce the cost and time of training the network to allow automatic extraction of features at various levels of abstraction, features, and objects from raw images (16).

In the proposed work, we propose an end-to-end machine learning technique for automated breast cancer diagnosis using a pre-trained network to discriminate microcalcification, specifically a novel D-CNN architecture with adaptive transfer learning. In this study, curated Breast Imaging Subset of Digital Database for Screening Mammography (CIBS-DDSM) dataset from The Cancer Imaging Archive (TCIA) data portal which contains ROI images of digital mammography in grayscale will be utilized to facilitate training of the model. Our work utilizes CNN networks to automatically extract features of benign and malignant microcalcification instead of directing the machine to learn from locations identified *via* *.csv files. A series of pre-processing algorithms are introduced to ensure the images were well prepared before beginning the process of feature extraction to enhance the accuracy of the model.

The primary contribution of the work involves; (i) proposing end-to-end machine learning architecture to diagnose breast cancer using microcalcifications' characteristics, (ii) performing pre-processing operations for the collected mammogram images before classification using deep learning algorithms, and (iii) proposing an adaptive transfer learning technique of CNN to build a breast cancer image classifier. The proposed work involves four state-of-the-art deep learning architectures such as ResNet38, ResNet50, VGG16, and AlexNet, and the performance of the models is compared to evaluate their performance.

Related Works

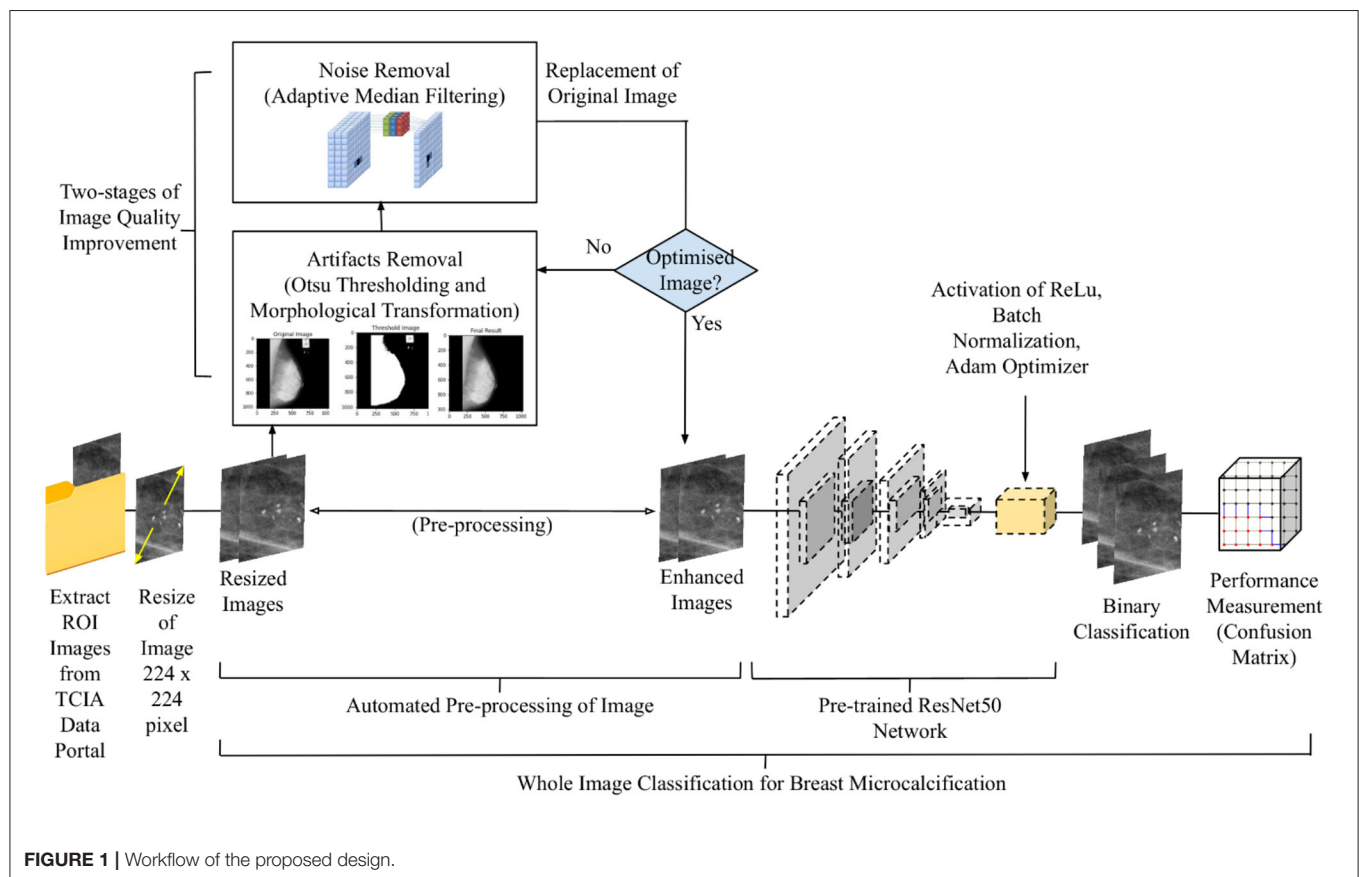
The introduction of digital mammography images has made deep learning approaches for breast cancer diagnosis possible in recent years (17, 18). Significant research which involves the use of machine learning, specifically D-CNN-based supervised machine learning for microcalcification detection has been performed. CNNs are able to achieve higher detection accuracy as compared to CAD models by delivering quantitative analysis of suspicious lesions (19). **Table 1** depicts the examples of studies that involve the classification of microcalcification of the breast into malignant and benign cases in recent years, including the model used and the accuracy achieved. Existing models of breast image classifiers for microcalcification detection are shown in **Table 1**. Based on **Table 1**, the highest accuracy for research breast image classifier involving VGG16 models is 94.3%, AlexNet is 88.6%, Resnet34 is 76.0%, and Resnet50 is 91.0%. Logic-based supervised learning such as Random Forest also managed to achieve an accuracy of 85.0% while Support Vector Machine (SVM) reached 95.8%.

As compared to learning algorithms such as SVM, CNN has gained its popularity due to higher accuracy and greater flexibility when it comes to tuning of hyperparameters. CNNs are feed-forward neural networks that are fully connected and are exceptionally good at lowering the number of parameters without sacrificing model quality. Since images have a high dimensionality as each pixel is considered a feature, it suits the capabilities of CNNs mentioned above.

TABLE 1 | Models of breast image classifier for microcalcification detection.

References	Base model	Type of image	Database	Accuracy
Wang et al. (20)	Support vector machine (SVM)	Histopathology	Private	95.8%
Fadil et al. (21)	Random Forest	Mammography	DDSM	85.0%
Tsochatzidis et al. (22)	AlexNet	Mammography	CBIS-DDSM	75.3%
	VGG16			71.6%
	ResNet50 (training from scratch)			62.7%
	ResNet50 (pre-trained network)			74.9%
Xiao et al. (23)	2D ResNet34 with anisotropic 3D ResNet34	Digital breast Tomosynthesis (DBT)	Private DBT	76.0%
Li (24)	Modified VGG16	Mammography	Private, DDSM	90.0%
Khamparia et al. (25)	Hybrid ImageNet Modified VGG16	Mammography	DDSM	94.3%
	Modified VGG16			89.8%
	ResNet50			85.1%
	AlexNet			83.4%
Heenaye-Mamode Khan et al. (26)	ResNet50	Mammography	CBIS-DDSM, UPMC	88.0%
Cai et al. (27)	AlexNet	Mammography	Private	88.6%
Hekal et al. (28)	Modified AlexNet	Mammography	CBIS-DDSM	84.0%
	Modified ResNet50			91.0%

Private = SunYat-sen University Cancer Center (Guangzhou, China) (SYUCC) and Nanhai Affiliated Hospital of Southern Medical University (NAHSMU) (Foshan, China).



As more models surfaced, accuracy has become one of the main aspects to compare the performance of models. Works of (22) highlighted that the accuracy for a pre-trained model

is higher as compared to the scratch model. The accuracy for ResNet50 has achieved 62.7% for scratch model and 74.9% for pre-trained model respectively with the utilization of dataset

TABLE 2 | Dataset distribution.

Image	Calcified benign ROI	Calcified malignant ROI
Original ROI image	1,077	577
Rotated at 90 degrees	1,077	577
Rotated at 180 degrees	1,077	577
Rotated at 270 degrees	1,077	577
Total number of images	4,958	1,653

```
def psnr(img_gray, img_new1):

mse = np. mean((img_gray - output) ** 2)
if mse == 0:
return 100
PIXEL_MAX = 255.0
return 20 * math.log10(PIXEL_MAX / math. sqrt(mse))

Y = np. square(np.subtract(img_gray,processed_image)).mean()
print ("PSNR value is ",d); print ("MSE value is ",Y)
```

FIGURE 2 | Code section for computing PSNR and MSE values based on filtered image and original image. “output” represents the finalized filtered image that will be used to compare with the original image, in this case is img_new1.

from CBIS-DDSM. Ensemble modelling has also been observed in (24, 25, 27), where fusion or modification of existing models has been performed to produce a better model. For instance, the fusion of Modified VGG and ImageNet is observed in works of Khamparia et al. (25). This hybrid model enhances the performance of the model and achieved an astonishing accuracy of 94.3% in breast image classification (25). On the other hand, AlexNet based CNN model that is modified with multiple layer architecture and drop-out strategy together with the fusion of “off-the-shelf” model from ImageNet observed in (27) has demonstrated the ability to get robust and spatially invariant features, achieving an accuracy of 88.6% for morphologically filtered CNN feature.

Inspired by the promising results produced by the deep learning neural network, our research seeks to propose an end-to-end novel adaptive transfer learning convolutional neural network to discriminate microcalcifications of breast mammograms into benign or malignant cases. Most of the methods used were based on the Mammographic Image Analysis Society (MIAS) and InBreast dataset, which uses handcrafted features for machine learning. This research utilizes the CIBS-DDSM dataset obtained from TCIA data portal, which provides a higher resolution and number of images for machine learning to enhance the accuracy of diagnosis. Instead of training the model using a whole mammogram image, the model in this research is trained by using ROI images of calcifications, allowing the model to extract features from a focused area. The main goal of this research is to detect and categorize microcalcification as accurately as possible to aid radiologists to prepare the

TABLE 3 | Parameters of data augmentation.

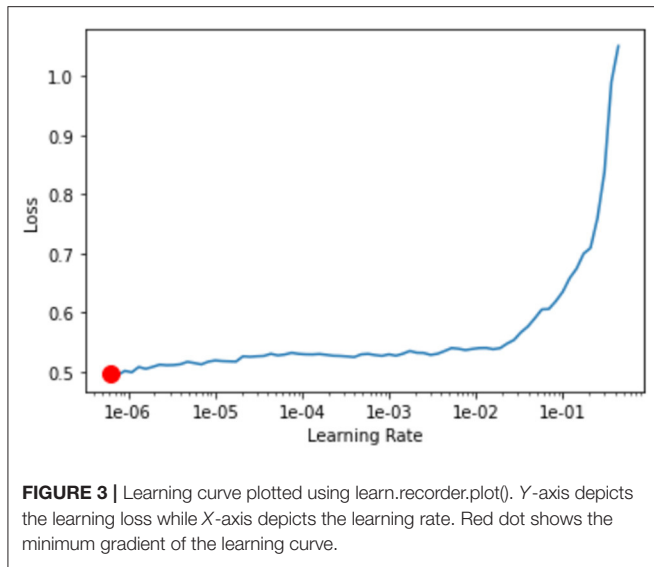
Parameter	Function	Description
Flipping	do_flip (), flip_vert ()	Flips the images at vertical and horizontal axis randomly
Zooming	max_zoom ()	Zooms the images at certain scale randomly
Rotating	max_rotate ()	Rotates the images at certain degree randomly
Lighting	max_lighting (), p_lighting()	Changes the contrast of image randomly controlled by max_lighting () with random probability ()

diagnosis report rapidly. The model is beneficial to be applied in a clinical setting.

MATERIALS AND METHODS

The proposed deep learning model is developed in Google Collab's platform with an OpenCV library of programming functions. Data acquisition is performed by downloading the breast mammography ROI images with microcalcification from the TCIA database. Micro-calcified images of the breast mammography were categorized into benign and malignant cases based on the information given in the *.csv files from TCIA. Moving on, the downloaded images were pre-processed to remove artifacts and noises. Since the size of microcalcification is small and scattered in the mammogram, a conventional D-CNN model often failed to classify and often resulted in false positive or false-negative results. Therefore, we propose an end-to-end machine learning technique, which consists two stages of pre-processing technique, specifically implementation of artifacts removal to remove the existence of artifacts surfaced and filtering of images to lower the noise level of images prior to implementation of machine learning. The focus of work on enhancing quality of images were performed automatically upon identifying threshold value of breast region using Google Colab's platform. This step is crucially important to build and train a model with quality information of features extracted from the image itself.

A D-CNN model is developed with finely tuned hyperparameters. To categorize the mammogram images into benign and malignant cases, a CNN model is utilized as a baseline. Transfer learning is used instead of training CNN from scratch. As such, different CNN models pre-trained with torch vision from the fastai library will be transferred to conduct the classification. To get the best possible result, hyperparameters such as the number of extra layers, learning rate, batch size, and epochs will be tuned. Finally, the confusion matrix will be utilized to assess the performance of the model to get the best possible accuracy. The overall algorithm for automated breast microcalcification classification is presented in **Figure 1**.



Materials and Preparation of Dataset

The following are the materials needed for the work of this research:

1. Intel Core i7-4710 HQ, 3.5 GHz, 1 TB SSD, 4 GB RAM,
2. Google Colaboratory Platform (Python OpenCV language and fastai Library)
3. Breast Image dataset CIBS-DDSM from TCIA.

The CIBS-DDSM dataset of ROI microcalcification images for this research is obtainable from Cancer Imaging Archive (TCIA). The prepared dataset consists of 1,077 benign and 577 malignant ROI images in various sizes in DICOM format. Data Retriever software was installed to download radiological pictures from the TCIA Radiology Portal and was later fed into DICOM software to be saved in *.jpeg format with a size of 224×224 to achieve uniformity in feature learning. The total number of images for benign and malignant as was multiplied by rotation at 90° , 180° , and 270° , resulting in 4,958 mammogram images for calcified benign ROI and 1,653 mammogram images for calcified malignant ROI. **Table 2** shows the distribution of the dataset utilized in this study.

Pre-processing of Dataset

Before any pre-processing work was performed, the notebook on Google Collab was set to be under GPU Runtime to allow heavier computational work. Prior to training CNNs, the images will be pre-processed to remove the artifacts and improve the contrast by removing noise. *Otsu Segmentation Method* and *MorphologicalEx Method* presented by (29) were utilized to remove the artifacts that may be present at the image. *Otsu Segmentation Method* works on grayscale images and involves global thresholding or local thresholding to classify pixels values (30, 31). For instance, we denote mammogram image as function of $G(x, y)$ and intensity value of $I \{I = 0, 1, 2, \dots, I-1\}$. The variance of these two variables can be computed by using Equation (1).

$$\sigma_m^2 = \theta_1^{(th)} \cdot \sigma_1^2(th) + \theta_2^{(th)} \cdot \sigma_2^2(th) \quad (1)$$

whereby,

$$\theta_1(th) = \sum_{i=1}^{th} P(i) \quad (2)$$

$$\theta_2(th) = \sum_{i=th+1}^I P(i) \quad (3)$$

$P(i)$ denotes the probability of gray-level i occurred, given as $P(i) = \frac{n_i}{n}$. In which, the number of pixels with a certain gray-level I is denoted by i . The image's total number of pixels is n . Threshold value th , which determines the class probability of pixels, is denoted as θ_1 and θ_2 , and the mean of the class is calculated as u_1 and u_2 as in Equations (4), (5) below. The threshold value that is predetermined earlier, th , which falls within the range of $0 < th < I$ will be utilized to divide the original mammogram image into two segments according to the intensity, which are $[0, th]$ and $[th + 1, I]$, where I is the maximum pixel value (255).

$$u_1(th) = \sum_{i=1}^{th} \frac{iP(i)}{\theta_1(th)} \quad (4)$$

$$u_2(th) = \sum_{i=th+1}^I \frac{iP(i)}{\theta_2(th)} \quad (5)$$

The value of interclass variance and global mean-variance can then be computed by using Equations (6) and (7), respectively.

$$\sigma_1^2(th) = \sum_{i=1}^{th} [I - u_1(th)]^2 \frac{P(i)}{\theta_1(th)} \quad (6)$$

$$\sigma_2^2(th) = \sum_{i=th+1}^I [I - u_2(th)]^2 \frac{P(i)}{\theta_2(th)} \quad (7)$$

The optimum threshold value is identified to achieve the best performance in distinguishing the target class from the background class, which is mostly utilized in mammography image binarization. Before executing the procedures for breast cancer detection segmentation and feature extraction, this thresholding approach is employed as a pre-processing technique (32, 33).

On the other hand, simple logical operations on local groupings of pixels, which is also defined as morphological operators are utilized in this research. Two of the main morphological operations used are dilation and erosion, which are shown in Equations (8) and (9), respectively (34). The binary image is denoted as X while the structuring element is denoted as B . The term B_x can be understood as translation of B by the vector x . Erosion reduces the size of an image by removing a layer of pixels from the inner and outer boundaries of regions. Dilation, on the other hand, has the reverse effect of erosion in that it adds a layer of pixels to both the inner and outer boundaries of regions. Many functions, such as opening and closing, are derived from these operators. When a picture is opened, it undergoes erosion and then dilation, and when it is closed, it undergoes dilation and then erosion (34).

$$X \ominus B = \{x | B_x^1 \subset X\} \quad (8)$$

$$X \oplus B = \{x | B_x^2 \subset X^c\} \quad (9)$$

Adaptive median filter, mean filter and median filter were included in this research. The performance of filter was assessed according to Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE). PSNR value is closely linked with MSE as it is computed based on MSE values, as in Equation (10).

$$PSNR = 20 \log_{10} \left(\frac{MAX_f}{\sqrt{MSE}} \right) \quad (10)$$

MAX_f is the maximum signal value that exists in the original image. Lower MSE indicates better filtration as MSE is the squared average of the “errors” between the actual image and the noisy image. The best filter will be selected based on the highest PSNR and lowest MSE value. The pseudocode of calculating MSE and PSNR value is shown in **Figure 2**.

Upon identifying the best performing filter, the filter will be applied to the images which has undergone artifacts removal process to further remove the noises of the image for clarity enhancement of the images, therefore completing the two-stages of optimization. The enhanced images will replace the original images to store the image in the same file location for machine learning. Before finalizing the two-stages of optimization process, the enhanced images will be inspected again to make sure the artifacts have been removed completely before proceeding to the next stage.

Deep CNNs Architecture

Prior training, `valid_pct ()` splits the dataset into training and testing sets at a particular ratio of 0.80 testing sets and

0.20 validation set. In total, there are 5,288 training images and 1,323 validation images. Data augmentation technique was implemented on the training set to avoid over-fitting by including `get_transforms ()` function to increase the volume of the dataset by artificially producing new training data from the current data. Parameters of data augmentation is tabulated in **Table 3**.

Hyperparameters were chosen manually in each set of tests to identify the best possible accuracy on binary classification. Hyperparameters that is tuned involves number of layers, learning rate, batch size as well as epoch. ADAM optimization algorithm was included to enhance the effectiveness of the model in to computing adaptive learning rate in complicated network architectures. In addition to that, ReLu is activated to prevent the computation required to run the neural network from growing exponentially. Batch Normalization is also activated to enable each layer of the network to conduct learning more independently by re-centering and re-scaling the layers' inputs to improve the speed and stability of the network.

TABLE 4 | PSNR and MSE values for adaptive median filter, median filter and mean filter.

Parameter	Adaptive median filter	Median filter	Mean filter
PSNR	42.3863	37.5911	36.9511
MSE	3.7536	11.3233	13.1213

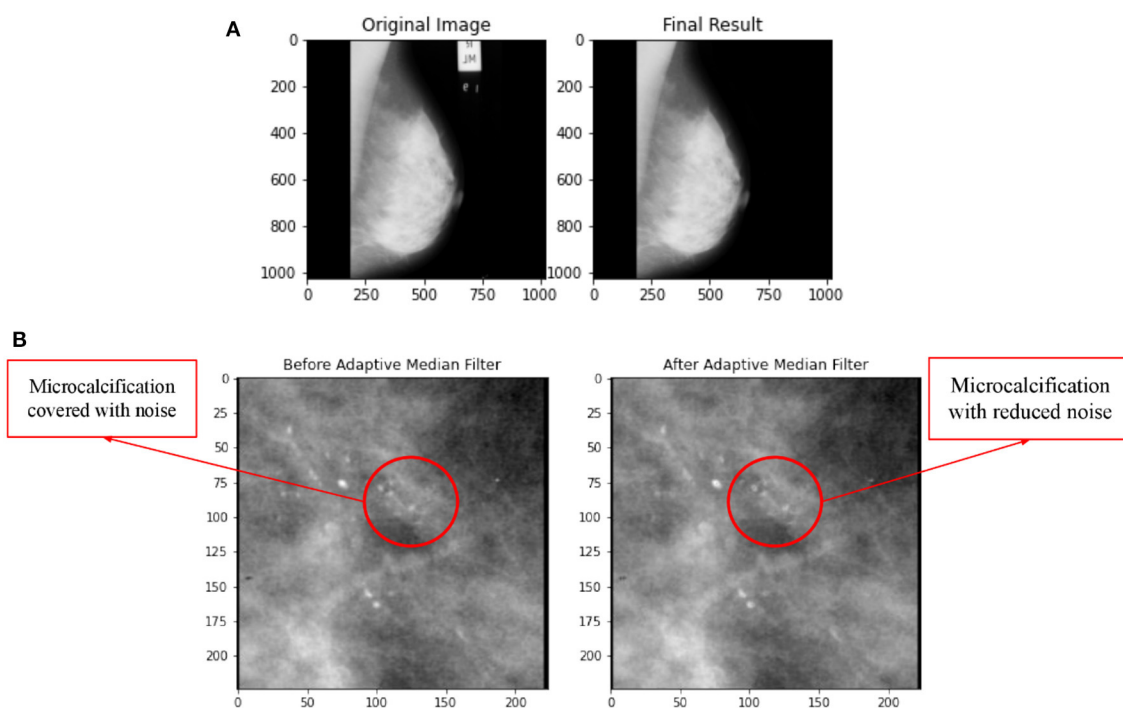


FIGURE 4 | (A) Comparison on application of artifacts removal with implementation of *Otsu Segmentation Method* and *MorphologicalEx Method* for a sample of full breast mammogram image. (B) Comparison on before and after application of adaptive median filter.

TABLE 5 | Output of VGG16 model.

Test	Batch size	Learning rate	Epoch	Training loss	Validation loss	Error rate	Accuracy
1	32	8e-6, 1e-4	15	42.7083	43.9302	23.4493	76.5507
2	64	8e-6, 1e-4	15	76.4934	50.4612	22.4917	77.5083
3	64	8e-6, 1e-4	30	26.2982	45.7910	18.0787	81.9213
4	32	2e-6, 1e-3	15	30.7861	32.0147	16.4522	83.5478
5	64	2e-6, 1e-3	15	25.4205	25.4679	10.9682	89.0318
6	64	2e-6, 1e-3	30	7.5000	8.4696	3.0257	96.9743

TABLE 6 | Output of ResNet34 model.

Test	Batch size	Learning rate	Epoch	Training loss	Validation loss	Error rate	Accuracy
7	32	8e-6, 1e-4	15	42.2252	43.1934	21.4070	78.5930
8	64	8e-6, 1e-4	15	41.1351	42.8464	21.5582	78.4418
9	64	8e-6, 1e-4	30	12.6166	36.0723	16.3888	83.6112
10	32	2e-6, 1e-3	15	35.0093	30.7748	14.2965	85.7035
11	64	2e-6, 1e-3	15	26.2728	26.9305	10.8926	89.1074
12	64	2e-6, 1e-3	30	7.6075	9.5925	2.6475	97.3525

Pretrained network was downloaded from the fastai library using `create_cnn()`. The first layer of the model was trained by using `learn.fit_one_cycle()`. Later, the learning rate for the model was determined with the aid of `learn.lr_find()` and `learn.recorder.plot()`, which illustrates the learning curve of the model after training the first layer and suggests the lowest gradient of the learning curve. The example of learning curve plotted by using `learn.recorder.plot()` is shown in **Figure 3**.

Moving on, all layers of the model were unfreeze using `learn.unfreeze()` to allow more parameters to be trainable. The model undergoes training again with Cylindrical Learning Rate (CLR) using `learn.fit_one_cycle()`, but restrained on a cyclic learning rate using `max_lr()`. CLR enables the learning rate to fluctuate between appropriate minimum and maximum boundaries and is computationally cheap and eliminates the need to identify the ideal learning rate.

Upon running the number of epochs predetermined, the confusion matrix of the model on the validation set was plotted. The top losses of images during training were plotted with labels of “Prediction/Actual/Loss/Probability.” By the end of the training, the value for training loss, validation loss, error rate and accuracy were recorded.

Performance Measurement

When it comes to evaluating the performance of the model, a confusion matrix is utilized. Four main parameters that are presented in a confusion matrix, which are: (i) True positive (TP) which shows the outcome of the model correctly predicts the benign cases, (ii) True negative (TN) which shows the outcome where the model correctly predicts the malignant cases, (iii) False positive (FP) which indicates the number of benign cases that are recognized as malignant cases by the model, and (iv) False negative (FN) which indicates the number of malignant case that are recognized as benign case by the model.

The values obtained from the confusion matrix will be further analyzed to compute additional parameters such as Recall, Precision, Specificity, Accuracy, F-1 score and Matthew Correlation Coefficient (MCC). MCC measures the performance of the parameters in the confusion matrix. The classifier produces a more accurate classifier if the MCC values trend more towards +1, and the reverse situation occurs if the MCC values trend more towards -1.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (13)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$\text{F1Score} = \frac{2 * \text{Recall}}{2 * \text{Recall} + FP + FN} \quad (15)$$

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

RESULTS AND DISCUSSION

Artifacts Removal

Wedges and labels in the raw mammography picture may cause needless disruptions during the mass detection procedure (35). By manually looking at each ROI images of breast calcification downloaded from the TCIA database, the images were found to be free labelling artefacts. However, the algorithms for removal of artifacts were still conducted just in case there is hidden or unobvious artifact. In order to ensure that this section of coding works properly, a sample

TABLE 7 | Output of AlexNet model.

Test	Batch size	Learning rate	Epoch	Training loss	validation loss	Error rate	Accuracy
13	32	8e-6, 1e-4	15	52.147	48.5449	26.0968	73.9032
14	64	8e-6, 1e-4	15	49.9790	46.5579	25.416	74.5840
15	64	8e-6, 1e-4	30	42.8953	44.6564	24.2814	75.7186
16	32	2e-6, 1e-3	15	46.3035	42.8736	22.1044	77.8956
17	64	2e-6, 1e-3	15	44.2203	42.6651	22.0121	77.9879
18	64	2e-6, 1e-3	30	39.0666	35.3782	16.9440	83.0560

TABLE 8 | Output of ResNet50 model.

Test	Batch size	Learning rate	Epoch	training loss	Validation loss	Error rate	Accuracy
19	32	8e-6, 1e-4	15	39.0362	41.5517	20.5749	79.4251
20	64	8e-6, 1e-4	15	35.1833	40.6826	19.5159	80.4841
21	64	8e-6, 1e-4	30	21.1929	36.9642	14.2965	85.7035
22	32	2e-6, 1e-3	15	20.5363	37.8652	15.5068	84.4932
23	64	2e-6, 1e-3	15	29.6796	24.4782	10.6657	89.3343
24	64	2e-6, 1e-3	30	10.8362	5.8117	2.4206	97.5794

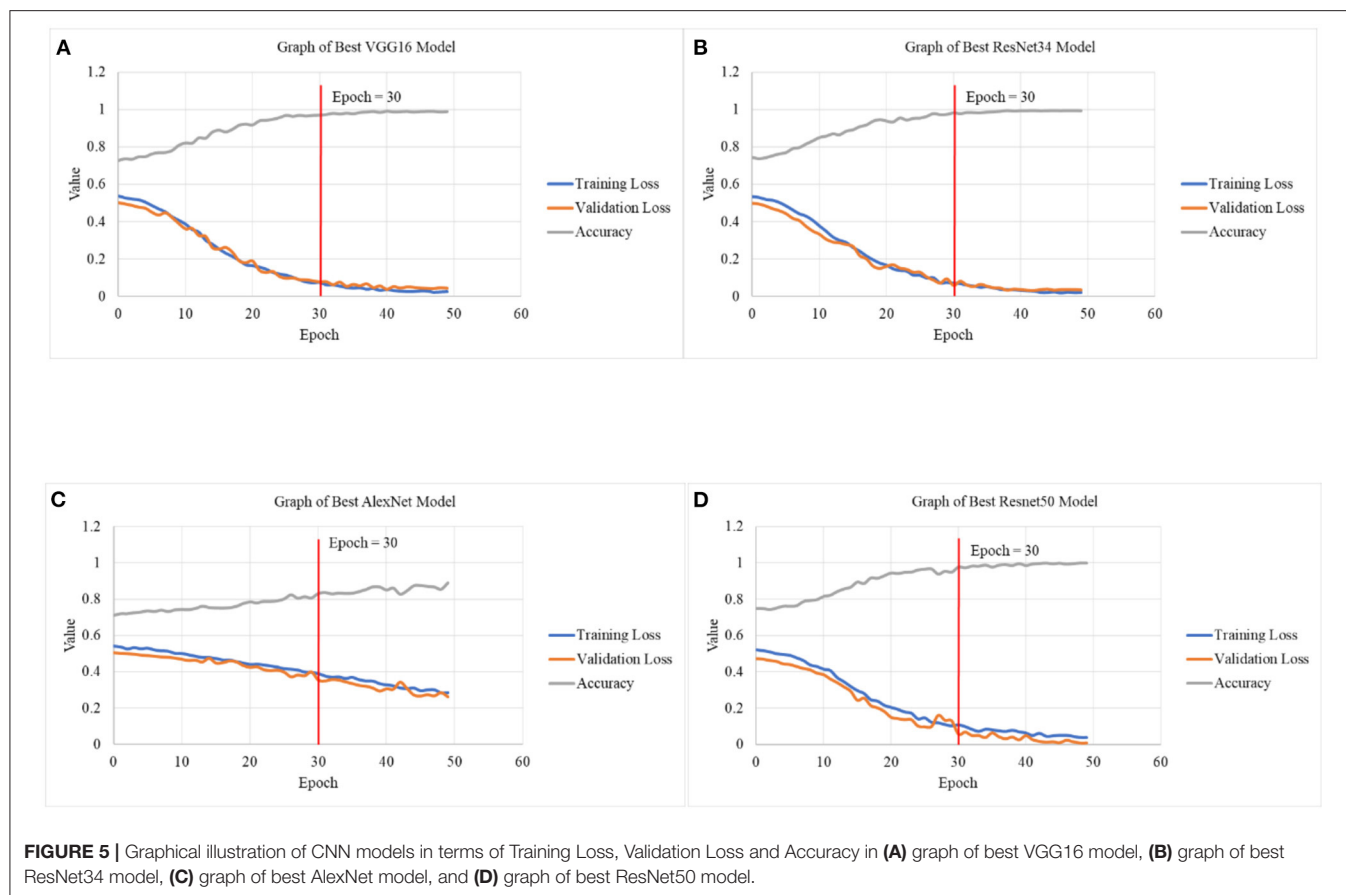


image of whole breast mammogram with obvious artifacts were imported and tested. The test result in **Figure 4** shows successful removal of labelling artifacts with the whole breast

mammogram image. Upon confirming the workability of the coding, the algorithm is then implemented to the ROI images in this study.

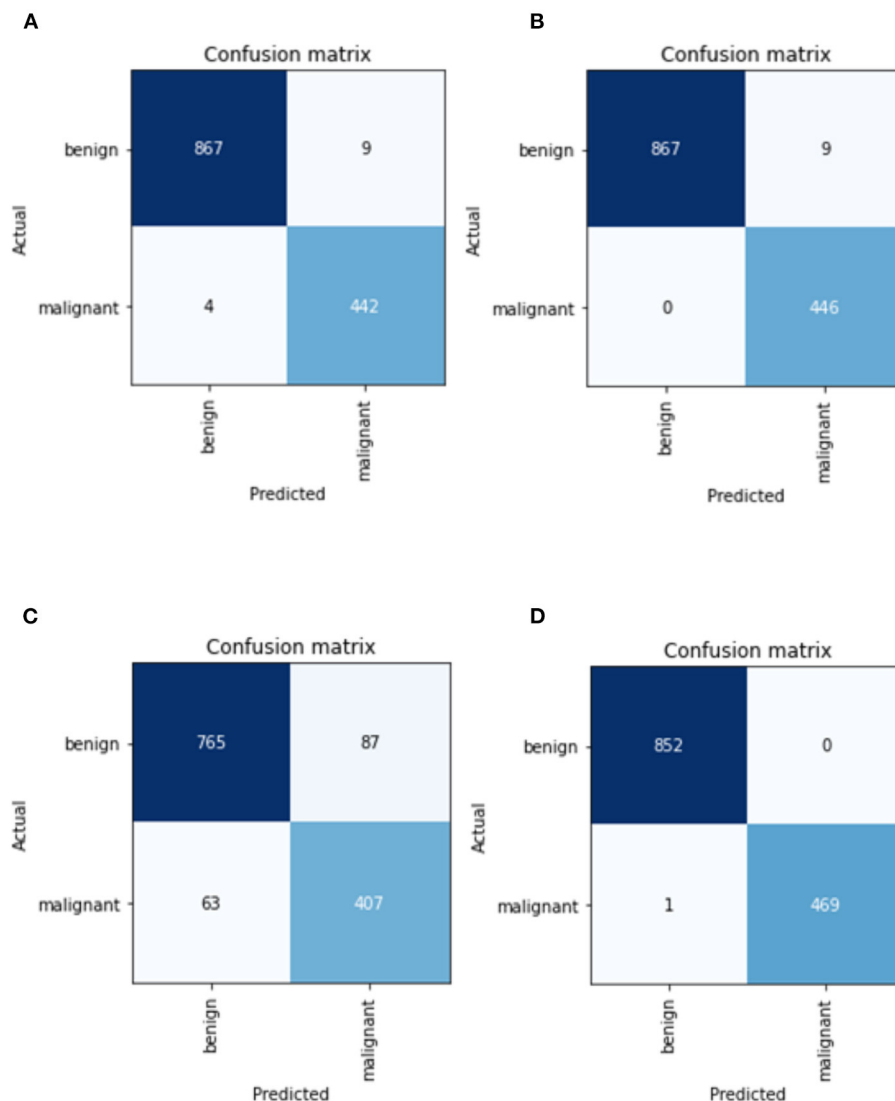


FIGURE 6 | Confusion matrix of CNN models in (A) best VGG16 model, (B) best ResNet34 model, (C) best AlexNet model, and (D) best ResNet50 model.

Image Enhancement

In this research, three types of filters, namely adaptive median filter, mean filter, and median filter were applied on the same image and the MSE and PSNR value for each filter was computed to identify the best filter. **Figure 4B** shows a comparison of before and after application of adaptive median filter on breast mammogram image. The PSNR and MSE values for adaptive median filter, median filter and mean filter is tabulated in **Table 4**.

By referring to **Table 4**, value for MSE is lowest for adaptive median filter, indicating that the error difference between the original image's values and the degraded image's values for adaptive median filter is the least among all three types of filters. Similar to (36, 37), comparison for adaptive median, mean and median filter for breast mammogram images were reported and the authors had concluded that adaptive median filter is the best

filter for noise reduction since the quality of the image produced is much superior. Hence, this research utilizes adaptive median filter for image enhancement of breast microcalcification images.

CNN Model Architecture

Tables 5–8 show the output of VGG16, ResNet34, AlexNet and Resnet50 models respectively. Identifying ideal batch size for CNNs is important as it helps the network to reach maximum accuracy in the quickest possible time, particularly for complicated datasets, such as a medical picture dataset (38). Results obtained from this study demonstrates that with learning rate and epochs remains, the accuracy of the model increases when the number of batch sizes increases from 32 to 64. In **Table 7**, the increase in batch size from 32 to 64 in Test 10 and Test 11 has resulted in increase in accuracy with an additional

TABLE 9 | Additional performance measurement for best Resnet34, Resnet50, VGG16 and AlexNet model.

Architecture	Recall	Precision	Specificity	Accuracy	F-1 Score	MCC
Resnet34	1.0000	0.9897	0.9802	0.9932	0.1818	0.8950
Resnet50	0.9988	1.0000	1.0000	0.9992	0.6664	0.9983
VGG16	0.9954	0.9897	0.9800	0.9902	0.1328	0.9781
AlexNet	0.9239	0.8979	0.8239	0.8865	0.0122	0.7558

Bold values indicates the model with the best performance.

value of 4.67%. Findings from this research also implies that the larger the batch size, the greater the network accuracy, implying that batch size has a significant influence on CNN performance.

Figure 5 depicts the graphical illustration of CNN models in terms of Training Loss, Validation Loss and Accuracy for different models. Graphs obtained from this study suggests better accuracy was achieved with smaller learning rates of $2e-6, 1e-3$ as compared to $8e-6, 1e-4$. With number of epochs increases, the accuracy tends to increase as well. In **Table 5** Test 5 and 6, with learning rate of $2e-6, 1e-3$, the accuracy of VGG16 has managed to reach 96.9743% for 30 epochs as compared to 89.0318% for 15 epochs. Test 17 and 18 also demonstrates the same characteristic with an increase of accuracy from 77.99 to 83.06%, about a 6.50% difference with increase of 15 to 30 epochs.

By referring to **Figure 5**, upon reaching 30 epochs, the losses and accuracy starts to flatten out, suggesting overfitting. Overfitting occurs when the network begins to overfit the data and the error on the validation set will soon begin to rise on a regular basis. This is where training should be terminated (39, 40). Therefore, the number of epochs for all the models is fixed at 30. In addition to that, the training and validation loss at 30 epochs is not increasing nor achieving linearity before minimal loss is achieved, suggesting that the result is not overfitting.

Comparison of Models With Existing Work

As deep learning becomes more popular, more researchers created new architectures with deeper CNN in radiomics of mammographic imaging to improve breast cancer diagnosis (41). VGG net requires much more parameters to thoroughly evaluate its performance. In (30, 31), the use of VGG16 was modified to classify microcalcification images into benign or malignant cases from the DDSM database and obtained accuracy of 94.3 and 87.0%, respectively. Study of (33) utilized AlexNet and managed to achieve an accuracy of 79.1% upon utilizing 10-fold cross validation technique with 300 epochs and learning rate of 0.01 based on 900 images from SYUCC and NAHSMU database. In this research, the technique of cross validation was not performed, but the accuracy achieved in AlexNet is much higher, reaching 83.1% with just 30 epochs. The difference in the result might be due to the different database of images that was used. For instance, this research utilizes ROI calcification images of CIBS-DDSM database which provides higher resolution. Also, the learning rate that was used in this study is much smaller. Study of (42) highlights that smaller learning rate can frequently increase generalization of accuracy substantially. A slower learning rate may allow the model to learn a set of weights

that is more optimum or even globally optimal. This might explain why smaller learning rates may also be able to produce models with higher accuracy.

Study of (34) classified 1,852 calcification images of CIBD-DDSM database into CNN pretrained models of modified AlexNet and ResNet50, of which the FC8 layer in AlexNet or FC1000 layer in ResNet50 is replaced with a shallow classifier (SVM). With 20 epochs, the accuracy for breast microcalcification for Resnet50 has managed to reach 91% while AlexNet has reached 90%. Although the accuracy for the AlexNet model in this study was lower (83.1%), the accuracy for Resnet50 managed surpassed with a value of 97.6%. Modified ResNet50 was also observed in (26, 32, 43), with (43) achieving the highest accuracy of 90.3% upon utilizing 354 images from Inbreast dataset. The Resnet50 model in this study is able to surpass existing work with accuracy value of 97.6%. The main difference between the models is the image that is fed to the machine for training. For instance, this research directly utilizes ROI calcification images of CIBS-DDSM database, which enables the machine to learn the features of malignant and benign calcified cases accurately.

The use of Resnet34 in breast microcalcification can be observed in the study of (23), where the authors utilized 2D Resnet34 together with anisotropic 3D Resnet to classify 495 Digital Breast Tomosynthesis (DBT) microcalcification images and reached an accuracy value of 76.0%. The model of Resnet34 in this study is able to provide a significantly higher accuracy value, which is 97.4%, probably due to the large number of images (6,611 images) utilized for machine learning, of which is 13 times larger than the study of (23).

Figure 6 depicts the confusion matrix of CNN models. Overall, the AlexNet model has the highest percentage of both falsely classified benign and falsely classified malignant cases, which is 11.37% and 15.48%, respectively. The performance of the Resnet50 is considered as the best because it only has 1 misclassified image over 1,322 images, while Resnet34 has a total of nine misclassified images. For the case of VGG, it has a total of 13 misclassified images. Based on the values obtained in the confusion matrix, calculation for additional performance measurement was performed and tabulated in **Table 9**.

Based on **Table 9**, Resnet50 has the highest precision, specificity, and accuracy, while ResNet 34 model has the highest Recall, which is also referred to as True Positive Rate or Sensitivity. Result from this study suggests that the performance by ResNet model outperforms VGG and AlexNet models. ResNet50 also has the highest F1-score (0.6664), which indicates

how accurate a model is on a given dataset. MCC, can be considered as the most credible statistical metric since it is only high if all four confusion matrix categories are correctly predicted. From this study, Resnet50 is able to achieve the highest score of MCC with a value of 0.9983.

In a summary, an automated microcalcification detection in mammography for early breast cancer diagnosis using deep learning techniques has been successfully developed. Collected greyscale mammogram images had undergone pre-processing operations which includes conversion of images from DICOM to *.jpeg format, resizing to 224×224 pixels, removal of artifacts, and image enhancement by application of adaptive median filter. Transfer learning technique for CNN architectures was employed and result shows that ResNet50 achieves the highest accuracy with a value of 97.58%, followed by ResNet34 of 97.35%, VGG16 of 96.97% and finally AlexNet of 83.06%. The main limitation with current work is the possibility of the machine to remember the repeated patterning of the dataset for classification into benign or malignant cases *via* implementation of data augmentation. Resizing of ROI images might also result in data compression and loss of useful features or information of the image.

CONCLUSIONS

Our proposed work has built an end-to-end novel adaptive transfer learning convolutional neural network that has shown ability to discriminate microcalcifications of breast mammograms into benign or malignant cases. ROI breast images were acquired from CIBS-DDSM database to obtain a higher resolution image of breast mammogram. The selection of quality datasets, abundance of images for training, as well as tuning of hyperparameters are all important in improving the accuracy of the models. We have also shown a quantitative analysis on the effectiveness of three filters, namely adaptive median, median and mean filter in noise removal of breast microcalcification mammogram images by calculating the MSE and PSNR value. As compared to traditional method of feature extraction which uses coordinates to identify the location of microcalcification, we have successfully automate the model to identify the characterization of benign and malignant microcalcification patterns. All CNN models that were trained

shows powerful ability to discriminate benign and malignant microcalcification, with ResNet50 achieving the highest accuracy of 97.58%.

Breast cancer is a significant threat to women or men all over the world and improving the existing state of breast cancer detection systems is definitely a critical challenge. Findings from this study will be able narrow the gap of findings for CNNs models which were mostly tailored for binary classifier that focuses solely on breast microcalcification classification by providing a comparative comparison beginning from datasets that is utilized, pre-processing algorithms that are included, up to the algorithms utilized during machine learning. In addition, this study will also be able to aid research in developing a competent binary classification model by providing a comprehensive approach to the recent results on different CNN models in breast microcalcification detection. In future, different sources of breast images could be incorporated, such as 3D mammogram images, in order to identify and compare the effectiveness of the model in classifying different sources of microcalcification images. K-fold cross validation could also be incorporated in the algorithm to combine metrics of prediction fitness to get a more accurate estimate of model prediction performance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: <https://www.cancerimagingarchive.net>.

AUTHOR CONTRIBUTIONS

YL and KH designed and developed the algorithm as well as major contributors to the article writing. KL, NM, and MA performed the comparison analysis and checked all the synthesized data. All authors approved the final version to be submitted for publication.

FUNDING

The project was funded by University Malaya Research Grant Faculty Programme (RF010-2018A).

REFERENCES

- Kashif M, Malik KR, Jabbar S, Chaudhry J. Application of machine learning and image processing for detection of breast cancer. In: Lytras MD, Sariete A, editors. *Innovation in Health Informatics*. Cambridge, MA: Academic Press (2020), p. 145–62. doi: 10.1016/B978-0-12-819043-2.00006-X
- Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. survey on deep learning in medical image analysis. *Med Image Anal.* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
- Nelson HD, O'meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors associated with rates of false-positive and false-negative results from digital mammography screening: an analysis of registry data. *Ann Intern Med.* (2016) 164:226–35. doi: 10.7326/M15-0971
- Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk factors and preventions of breast cancer. *Int J Biol Sci.* (2017) 13:1387. doi: 10.7150/ijbs.21635
- Koo MM, von Wagner C, Abel GA, McPhail S, Rubin GP, Lyratzopoulos G. Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: evidence from a national audit of cancer diagnosis. *Cancer Epidemiol.* (2017) 48:140–6. doi: 10.1016/j.canep.2017.04.010
- Redaniel MT, Martin RM, Ridd MJ, Wade J, Jeffreys M. Diagnostic intervals and its association with breast, prostate, lung and colorectal cancer survival in England: historical cohort study using the Clinical Practice Research Datalink. *PLoS ONE.* (2015) 10:e0126608. doi: 10.1371/journal.pone.0126608

7. Baker R, Rogers KD, Shepherd N, Stone N. New relationships between breast microcalcifications and cancer. *Br J Cancer*. (2010) 103:1034–9. doi: 10.1038/sj.bjc.6605873
8. Mordang JJ, Gubern-Mérida A, Bria A, Tortorella F, Mann RM, Broeders MJ, et al. The importance of early detection of calcifications associated with breast cancer in screening. *Breast Cancer Res Treat*. (2018) 167:451–8. doi: 10.1007/s10549-017-4527-7
9. Nayak T, Bhat N, Bhat V, Shetty S, Javed M, Nagabhushan P. Automatic segmentation and breast density estimation for cancer detection using an efficient watershed algorithm. In: Nagabhushan P, Guru DS, Shekar BH, Sharath Kuma, YH, editors. *Data Analytics and Learning*. Singapore: Springer (2019), p. 347–58. doi: 10.1007/978-981-13-2514-4_29
10. Beeravolu AR, Azam S, Jonkman M, Shanmugam B, Kannoorpatti K, Anwar A. Preprocessing of breast cancer images to create datasets for deep-cnn. *IEEE Access*. (2021) 9:33438–63. doi: 10.1109/ACCESS.2021.3058773
11. Alkhaleefah M, Wu CC. A hybrid CNN and RBF-based SVM approach for breast cancer classification in mammograms. In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Miyazaki: IEEE (2018), p. 894–9. doi: 10.1109/SMC.2018.00159
12. Spruit M, Lytras M. Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients. *Telemat Inform*. (2018) 35:643–53. doi: 10.1016/j.tele.2018.04.002
13. Jusman Y, Ng SC, Hasikin K, Kurnia R, Osman NAA, Teoh KH. A system for detection of cervical precancerous in field emission scanning electron microscope images using texture features. *J Innov Opt Health Sci*. (2017) 10:1650045. doi: 10.1142/S1793545816500450
14. Avanzo M, Porzio M, Lorenzon L, Milan L, Sghedoni R, Russo G, et al. Artificial intelligence applications in medical imaging: A review of the medical physics research in Italy. *Physica Medica*. (2021) 83:221–41. doi: 10.1016/j.ejmp.2021.04.010
15. Amoroso N, Rocca ML, Bellotti R, Fanizzi A, Monaco A, Tangaro S. Alzheimer's disease diagnosis based on the hippocampal unified multi-atlas network (HUMAN) algorithm. *Biomed Eng Online*. (2018) 17:6. doi: 10.1186/s12938-018-0439-y
16. Comes MC, Fanizzi A, Bove S, Didonna V, Diotaiuti S, La Forgia D, et al. Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs. *Sci Rep*. (2021) 11:1–2. doi: 10.1038/s41598-021-93592-z
17. Abdelhafiz D, Yang C, Ammar R, Nabavi S. Deep convolutional neural networks for mammography: advances, challenges and applications. *BMC Bioinformatics*. (2019) 20:1–20. doi: 10.1186/s12859-019-2823-4
18. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. *Korean J Radiol*. (2017) 18:570–84. doi: 10.3348/kjr.2017.18.4.570
19. Kooi T, Gubern-Merida A, Mordang JJ, Mann R, Pijnappel R, Schuur K, Heeten AD, Karssemeijer N. A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In: Tingberg A, Lång K, Timberg P, editors. *International Workshop on Breast Imaging*. Cham: Springer (2016), p. 51–6. doi: 10.1007/978-3-319-41546-8_7
20. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep*. (2016) 6:1–9. doi: 10.1038/srep27327
21. Fadil R, Jackson A, Abou El Majd B, El Ghazi H, Kaabouch N. Classification of microcalcifications in mammograms using 2D discrete wavelet transform and random forest. In: *2020 IEEE International Conference on Electro Information Technology (EIT)*. Chicago, IL: IEEE (2020), p. 353–9. doi: 10.1109/EIT48999.2020.9208290
22. Tsochatzidis L, Costaridou L, Pratikakis I. Deep learning for breast cancer diagnosis from mammograms—a comparative study. *J Imaging*. (2019) 5:37. doi: 10.3390/jimaging5030037
23. Xiao B, Sun H, Meng Y, Peng Y, Yang X, Chen S, et al. Classification of microcalcification clusters in digital breast tomosynthesis using ensemble convolutional neural network. *Biomed Eng Online*. (2021) 20:1–20. doi: 10.1186/s12938-021-00908-1
24. Li J, Pei Y, Yasin A, Ali S, Mahmood T. Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable Convolutional Neural Network. *Sensors*. (2021) 21:4854. doi: 10.3390/s21144854
25. Khamparia A, Bharati S, Podder P, Gupta D, Khanna A, Phung TK, et al. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimens Syst Signal Process*. (2021) 32:747–65. doi: 10.1007/s11045-020-00756-7
26. Heenaye-Mamode Khan M, Boodoo-Jahangeer N, Dullull W, Nathire S, Gao X, Sinha GR, et al. Multi-class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). *PLoS ONE*. (2021) 16:e0256500. doi: 10.1371/journal.pone.0256500
27. Cai H, Huang Q, Rong W, Song Y, Li J, Wang J, Chen J, Li L. Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. *Comput Math Methods Med*. (2019) 2019:217454. doi: 10.1155/2019/217454
28. Hekal AA, Elnakib A, Moustafa HE. Automated early breast cancer detection and classification system. *Signal Image Video Process*. (2021) 15:1497–505. doi: 10.1007/s11760-021-01882-w
29. Xi P, Shu C, Goubran R. Abnormality detection in mammography using deep convolutional neural networks. In: *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. Istanbul: IEEE (2018), p. 1–6.
30. Thanh DN, Erkan U, Prasath VS, Kumar V, Hien NN. A skin lesion segmentation method for dermoscopic images based on adaptive thresholding with normalization of color models. In: *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*. Patna: IEEE (2019), p. 116–20. doi: 10.1109/ICEEE2019.2019.00030
31. Suradi SH, Abdullah KA, Isa NA. Breast lesions detection using FADHECAL and Multilevel Otsu Thresholding Segmentation in digital mammograms. In: Badnjevic A, Pokvić LG, editors. *International Conference on Medical and Biological Engineering*. Cham: Springer (2021), p. 751–9. doi: 10.1007/978-3-030-73909-6_85
32. Bhandari AK, Maurya S, Meena AK. Social spider optimization based optimally weighted Otsu thresholding for image enhancement. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. Washington, DC (2018). doi: 10.1109/JSTARS.2018.2870157
33. Khairnar S, Thepade SD, Gite S. Effect of image binarization thresholds on breast cancer identification in mammography images using OTSU, Niblack, Burns, Thepade's SBTC. *Intell Syst Appl*. (2021) 10:200046. doi: 10.1016/j.iswa.2021.200046
34. Nixon M, Aguado A. *Feature Extraction and Image Processing for Computer Vision*. Cambridge, MA: Academic Press (2019). doi: 10.1016/B978-0-12-814976-8.00003-8
35. Boss R, Thangavel K, Daniel D. Automatic mammogram image breast region extraction and removal of pectoral muscle. *arXiv preprint arXiv:1307.7474*. (2013). doi: 10.48550/arXiv.1307.7474
36. Cashmi AJ, Chehelamirani MAJ. Using adaptive median filter for noise removal from image to diagnose breast cancer. *Merit Res J Eng Pure Appl Sci*. (2019) 5:14–8. doi: 10.5281/zenodo.3374916
37. Ramani R, Vanitha NS, Valarmathy S. The pre-processing techniques for breast cancer detection in mammography images. *Int J Image Graph Signal Process*. (2013) 5:47. doi: 10.5815/ijigsp.2013.05.06
38. Kandel I, Castelli M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*. (2020) 6:312–5. doi: 10.1016/j.icte.2020.04.010
39. Afaq S, Rao S. Significance of epochs on training a neural network. *Int J Sci Technol Res*. (2020) 19:485–8. Available online at: <https://www.ijstr.org/final-print/jun2020/Significance-Of-Epochs-On-Training-A-Neural-Network.pdf>
40. Swathi P. Analysis on solutions for over-fitting and under-fitting in machine learning algorithms. *Int J Innov Res Sci Eng Technol*. (2018). 7:12404. doi: 10.15680/IJIRSET.2018.0712086
41. Pang T, Wong JH, Ng WL, Chan CS. Deep learning radiomics in breast cancer with different modalities: overview and future. *Expert Syst Appl*. (2020) 158:113501. doi: 10.1016/j.eswa.2020.113501
42. Wilson DR, Martinez TR. The need for small learning rates on large problems. In: *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*. Vol. 1. IEEE (2001), p. 115–9.

43. Hakim AN, Prajitno P, Soejoko DS. Microcalcification detection in mammography image using computer-aided detection based on convolutional neural network. In: *AIP Conference Proceedings*. Vol. 2346, No. 1. AIP Publishing LLC (2021), p. 040001. doi: 10.1063/5.0047828

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of

the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Leong, Hasikin, Lai, Mohd Zain and Azizan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



m6A Regulator-Mediated Methylation Modification Patterns and Characteristics in COVID-19 Patients

Xin Qing^{1,2}, Qian Chen³ and Ke Wang^{2*}

¹ School of Medicine, Southeast University, Nanjing, China, ² Clinical Laboratory, Boai Hospital of Zhongshan Affiliated to Southern Medical University, Zhongshan, China, ³ Department of Pediatrics, The Affiliated Hospital of Southwest Medical University, Luzhou, China

OPEN ACCESS

Edited by:

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

Reviewed by:

Keyang Xu,
Zhejiang Chinese Medical
University, China
Qi Tian,
Hunan Provincial Maternal and Child
Health Care Hospital, China

*Correspondence:

Ke Wang
bayywk2022@163.com

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 06 April 2022

Accepted: 25 April 2022

Published: 17 May 2022

Citation:

Qing X, Chen Q and Wang K (2022)
m6A Regulator-Mediated Methylation
Modification Patterns and
Characteristics in COVID-19 Patients.
Front. Public Health 10:914193.
doi: 10.3389/fpubh.2022.914193

Background: RNA N6-methyladenosine (m6A) regulators may be necessary for diverse viral infectious diseases, and serve pivotal roles in various physiological functions. However, the potential roles of m6A regulators in coronavirus disease 2019 (COVID-19) remain unclear.

Methods: The gene expression profile of patients with or without COVID-19 was acquired from Gene Expression Omnibus (GEO) database, and bioinformatics analysis of differentially expressed genes was conducted. Random forest model and nomogram were established to predict the occurrence of COVID-19. Afterward, the consensus clustering method was utilized to establish two different m6A subtypes, and associations between subtypes and immunity were explored.

Results: Based on the transcriptional data from GSE157103, we observed that the m6A modification level was markedly enriched in the COVID-19 patients than those in the non-COVID-19 patients. And 18 essential m6A regulators were identified with differential analysis between patients with or without COVID-19. The random forest model was utilized to determine 8 optimal m6A regulators for predicting the emergence of COVID-19. We then established a nomogram based on these regulators, and its predictive reliability was validated by decision curve analysis. The consensus clustering algorithm was conducted to categorize COVID-19 patients into two m6A subtypes from the identified m6A regulators. The patients in cluster A were correlated with activated T-cell functions and may have a superior prognosis.

Conclusions: Collectively, m6A regulators may be involved in the prevalence of COVID-19 patients. Our exploration of m6A subtypes may benefit the development of subsequent treatment modalities for COVID-19.

Keywords: COVID-19, m6A methylation modification, m6A regulators, diagnostic biomarkers, consensus clustering

INTRODUCTION

Coronavirus disease 2019 (COVID-19) derived from severe acute respiratory syndrome coronavirus clade 2 (SARS-CoV-2) has evolved as a significant challenge to the public health of global populations (1). Although various vaccines and antiviral agents are now being developed to reduce virus infection and combat this epidemic, little is known about how viruses interact with their hosts (2, 3). Recent studies have demonstrated a clear genetic link between SARS-CoV-2 infection and COVID-19 severity, and have identified multiple human genomic regions that are linked to disease severity (4, 5). Moreover, COVID-19 patients displayed obvious variations in the immune system, including immune cells, immune checkpoint, and cytokines (6–8). A deeper understanding of the pathogenesis of COVID-19 will facilitate better management of it, and determination of susceptible populations benefit for rationalizing the allocation of medical resources. It is critical and urgent to identify the association between patients' genomes and immune function during viral infections. Accordingly, early detection and appropriate intervention of high-risk patients from a genomic perspective will provide a significant benefit to managing the prevalence of COVID-19.

The N6-methyladenosine (m6A), an innate modification of mRNA and lncRNA, is a reversible procedure regulated by “writers,” “readers,” and “erasers” (9). For its biological characteristics, m6A can regulate carcinogenesis, immunity, stemness, and so on (10–12). Numerous reports have demonstrated that m6A modification serves a prominent part in tumorigenesis through modulating the activity of tumor-associated genes (13, 14). Similarly, m6A is observed and widely studied in diverse virus infections (15, 16), and existing studies have proven the significant role of m6A in the occurrence and progression of COVID-19 (17, 18). However, these researches concentrated predominantly on several m6A-related genes, and a majority of these models were constructed based on non-virally infected cells, which may not fully reveal the authentic status of m6A methylome modifications in immune cells of COVID-19 patients. Therefore, the function of m6A regulators in COVID-19 remain to be further investigated.

In this research, we systematically explored the roles of m6A regulators in the management and categorization of COVID-19. We constructed a gene signature to predict the occurrence of COVID-19 based on 8 selected m6A regulators and observed that patients could benefit from clinical decisions from this signature. Additionally, we identified two m6A subtypes that were closely associated with T-cell activation, indicating that m6A subtypes may distinguish COVID-19 and non-COVID-19 and provide reliable options for clinical treatment.

MATERIALS AND METHODS

Data Collection and Processing

The GSE157103 dataset, composed of 100 COVID-19 patients and 26 non-COVID-19 patients, was acquired from the GEO database (19). This dataset was selected based on some characteristics: sample size >100, diverse disease status, and

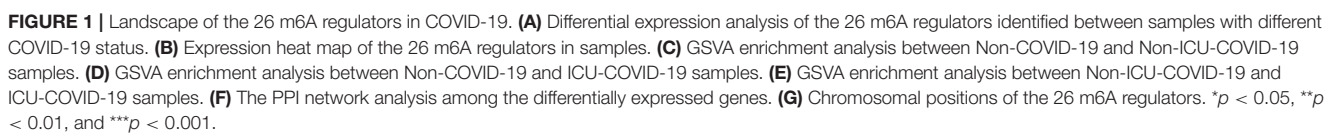
TABLE 1 | m6A modification regulators and their major biological functions.

Type	m6A regulator	Function
Writer	METTL3	Catalyze m6A modification
	METTL14	Facilitate METTL3 recognition of subunits
	METTL16	Catalyze m6A modification
	WTAP	Facilitate METTL3-METTL14 heterodimer to the nuclear speckle
	VIRMA	Bind the m6A complex and mobilize it to specific site
	RBM15	Bind the m6A complex and mobilize it to specific site
	RBM15B	Bind target RNAs and recruiting the WMM complex
	CBLL1	Regulate mRNA splicing and RNA processing
	ZC3H13	Bridge WTAP to the mRNA-binding factor Nito
	YTHDC1	Promote RNA splicing and translocation
Reader	YTHDC2	Promote target RNA translocation
	YTHDF1	Promote RNA translocation
	YTHDF2	Decrease mRNA stability
	YTHDF3	Regulate the translation or degradation
	HNRNPC	Regulate mRNA splicing
	FMR1	Regulate mRNA splicing, stability, dendritic transport and postsynaptic local protein synthesis
	LRPPRC	Regulate nuclear mRNA exportation
	HNRNPA2B1	Promote primary microRNA processing
	IGFBP1/2/3	Recruiting RNA stabilizer
	IGF2BP1	Improve mRNA stability
Eraser	ELAVL1	Improve mRNA stability
	RBMX	Regulate gene transcription and pre-mRNAs splicing
	ALKBH	Regulate mRNA intranuclear transport
	FTO	Catalyze the demethylation of m6A

publicly available data. And all samples are extracted from plasma and leukocyte samples of hospitalized patients. Normalization of the read count values was completed with the limma package (20). A total of 26 m6A regulators was collected from previous studies, and these regulators contain 9 writers, 15 readers, and 2 erasers (Table 1). Differently expressed analysis of these regulators based on limma package was performed between patients with or without COVID-19 to subsequent exploration. A protein-protein interaction (PPI) analysis of differentially expressed genes (DEGs) was performed through the string website (<https://cn.string-db.org>), and we exhibited gene set variation analysis (GSVA) with the “GSVA” package (21), thus matching the biological function between patients with or without SARS-COV-2 infection.

Establishment of a Random Forest Model and Support Vector Machine Model

Random forest (RF) and support vector machine (SVM) model was established to predict the prevalence of COVID-19 patients. Several methods, including “Reverse cumulative distribution



of residual,” “Boxplots of residual” and receiver operating characteristic (ROC) curve was conducted to validate these models. “RandomForest” package was applied to construct an RF model to identify optimal m6A regulators within the 26

m6A regulators for predicting the prevalence of COVID-19 (22). In this study, to identify optimal RF model, mtry and ntree were given as 3 and 500 after multiple adjustment. We also discussed the relevance of the 26 m6A regulators and determined

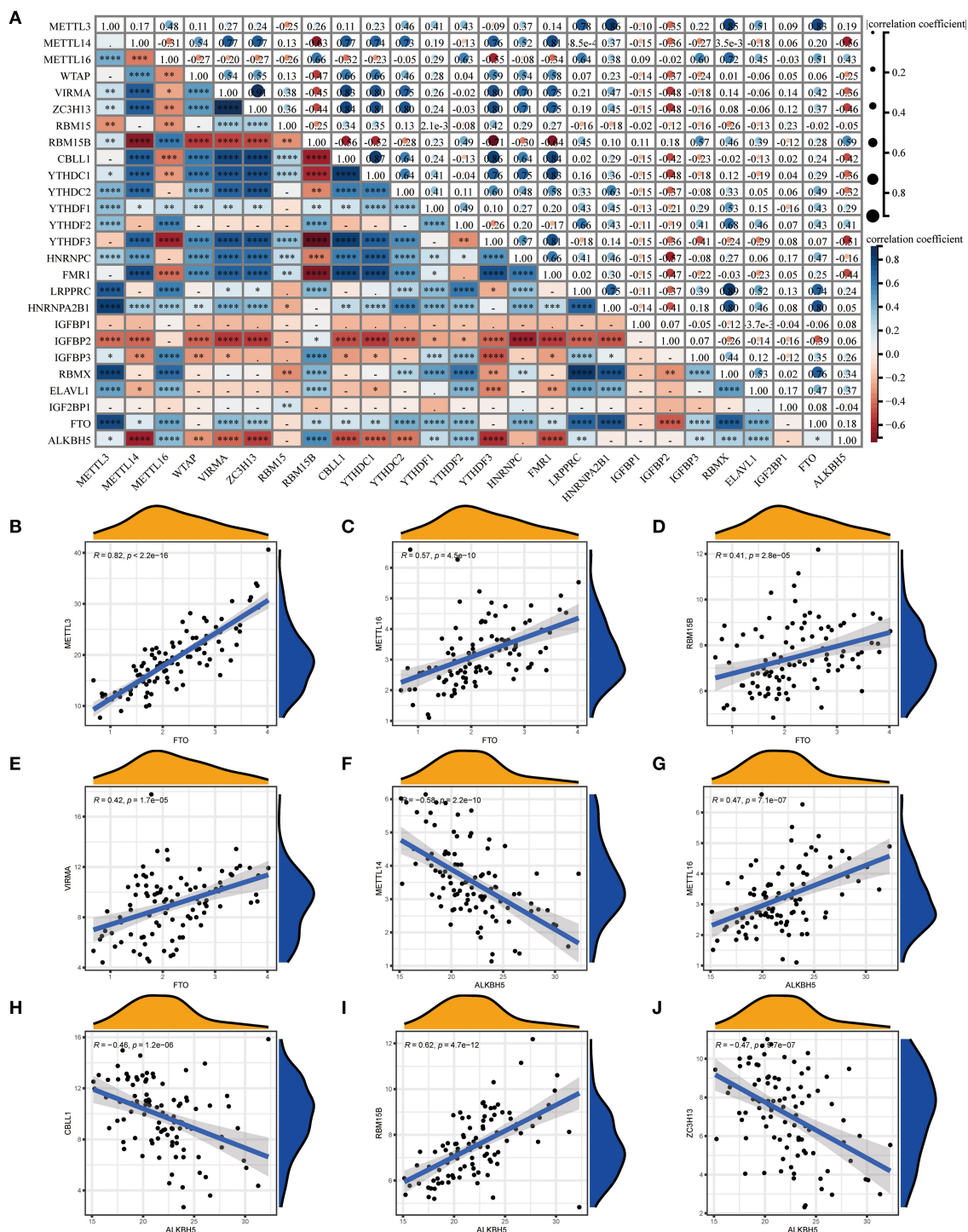
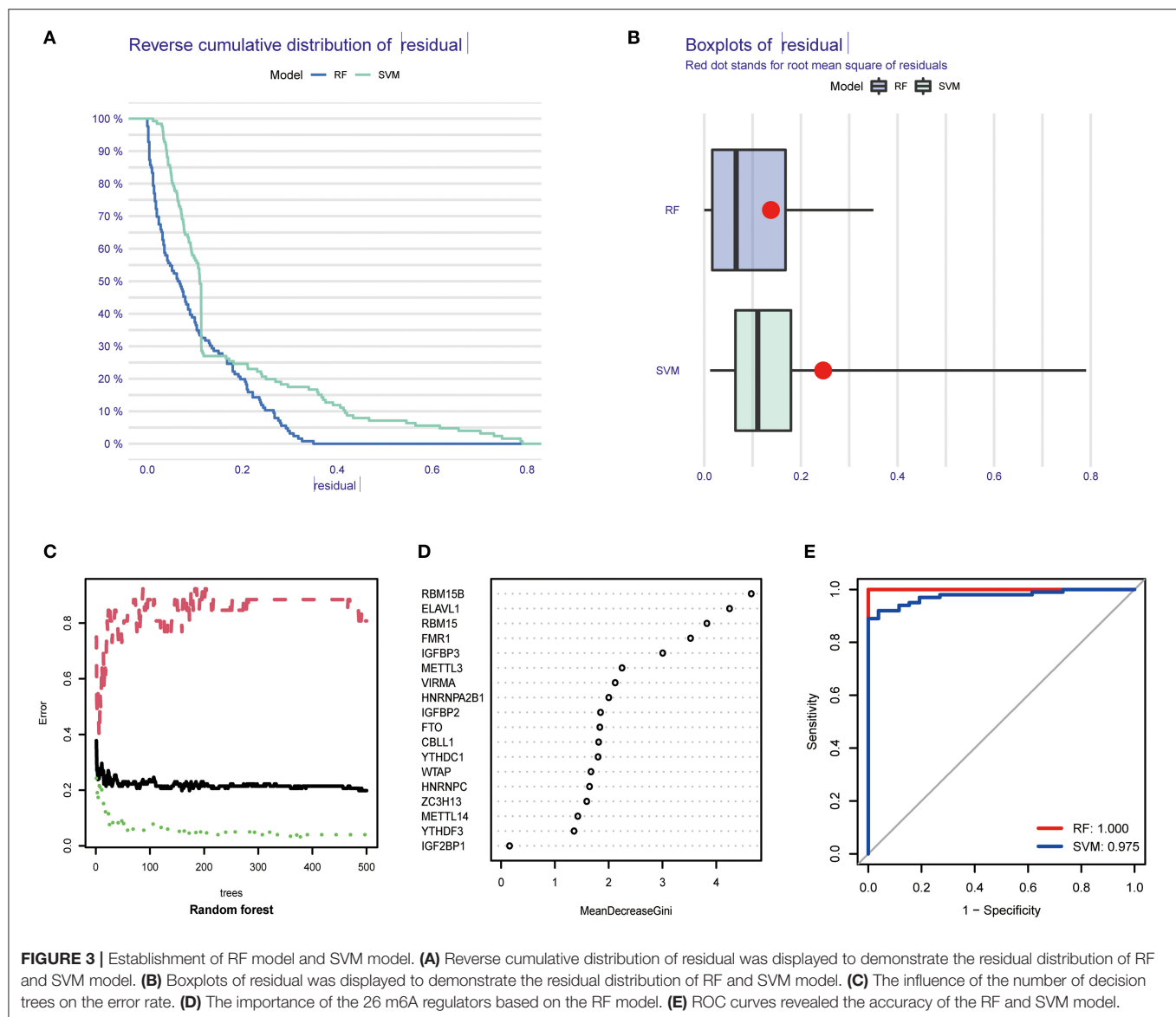


FIGURE 2 | Correlation between m6A regulators in COVID-19. **(A)** Correlation plot of 26 m6A regulators. **(B–J)** Correlation between writers and erasers in COVID-19. Writer genes: METTL3, METTL14, METTL16, RBM15B, VIRMA, CBLL1, and ZC3H13; eraser genes: ALKBH5 and FTO. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

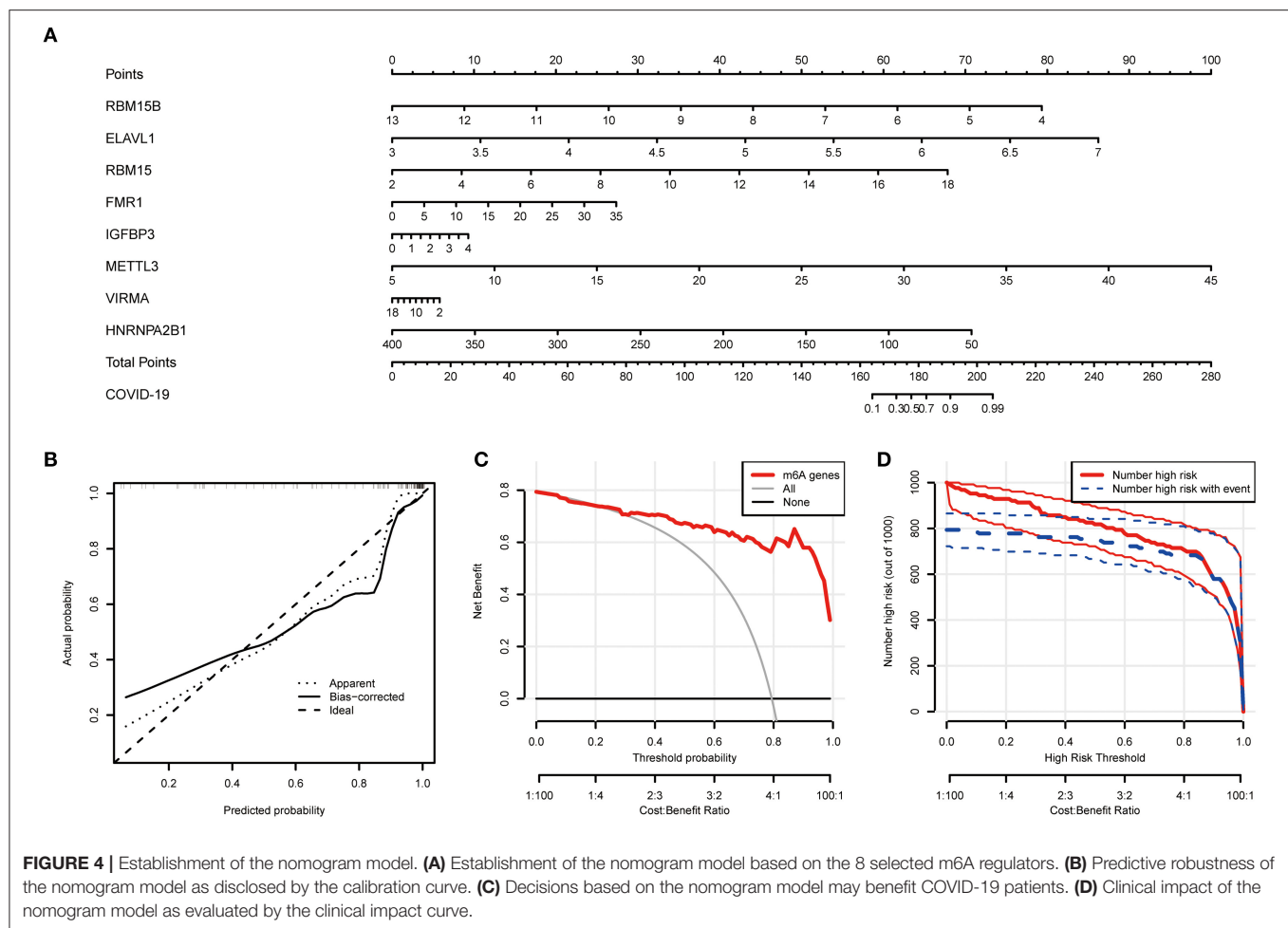


the candidate m6A regulators based on 10-fold cross-validation. The Y-axis of the 10-fold cross-validation curve represents the precision of the model when identifying different numbers of m6A regulators. The genes with an importance value over 2 were considered as the disease specific genes for the further analysis. SVM can minimize structural risk, thus enabling classification and regression analysis (23). In SVM model, the expression level of m6A regulators was regarded as the continuous predictive parameter and the sample type was regarded as the categorical variable. The “caret” package was applied to conduct a grid search for the determination of the reasonable hyperparameters for the SVM model with a 5-fold cross-validation (24). Each data is considered as a point in the n-dimensional space (n is 26 in this study), and an appropriate plane was found to distinguish well between the two categories (COVID-19 and non-COVID-19). A repeated 10-fold cross-validation was utilized to

tune and evaluate the models. The sample was split into 70% training and 30% test sets. We randomly split the training-test dataset 500 times and used 10-fold repeated 10 times cross-validation approach to optimize the model factors of each round of evaluation. The robustness of these model was assessed based on the area under curve (AUC) value of the receiver operating characteristics (ROC) curve.

Establishment of the Nomogram

Based on the abovementioned m6A regulators, a nomogram was developed to predict the occurrence of COVID-19 (25). Then, the reliability of this nomogram was assessed by the calibration curve, and decision curve analysis (DCA) was also constructed (26). Moreover, a clinical impact curve was established to evaluate the rationality and benefit of decisions from this nomogram (25).



Identification of Molecular Subtypes From m6A Regulators

Consensus clustering with K-means algorithms was applied to identify m6A regulators-related subtypes correlated with gene expression (27). The quantity and robustness of clusters were determined with a consensus clustering algorithm realized in the “ConsensuClusterPlus” package (28).

Identification and Functional Enrichment Analysis of Differentially Expressed Genes

The “limma” package was applied to identify DEGs between different m6A subtypes with the criterion of $p < 0.001$ (29). GO enrichment analysis was utilized to investigate the potential function of the DEGs responsible for COVID-19 with the “clusterProfiler” package (30).

Establishment of the m6A Gene Signature

Principal component analysis (PCA) was conducted to obtain the m6A score for individual specimens, thus quantifying the m6A subtypes (31). We exhibited the PCA method to identify the m6A subgroups, and the m6A score was acquired based on the following method: $m6A \text{ score} = PC1i$, of

which PC1 refers to principal component 1 and i to DEG expression (32).

Exploration of Infiltrating Immune Cell

Single sample gene set enrichment analysis (ssGSEA) was applied to assess the infiltration of immune cells in COVID-19 specimens (33). The gene expression levels in the specimens were sequenced with ssGSEA to acquire an individual grade. We then summarized the expression data of these genes for immunological analysis. Consequently, we gained the enrichment of immune cells in the individual specimen.

Statistics Analysis

Linear regression analyses were applied to determine the relationship between m6A regulators. Kruskal-Wallis tests were utilized to identify a discrepancy between clusters. All statistical analyses were carried out with two-tailed tests, and the significant value was considered $p < 0.05$. The R software was utilized to perform relevant analysis.

RESULTS

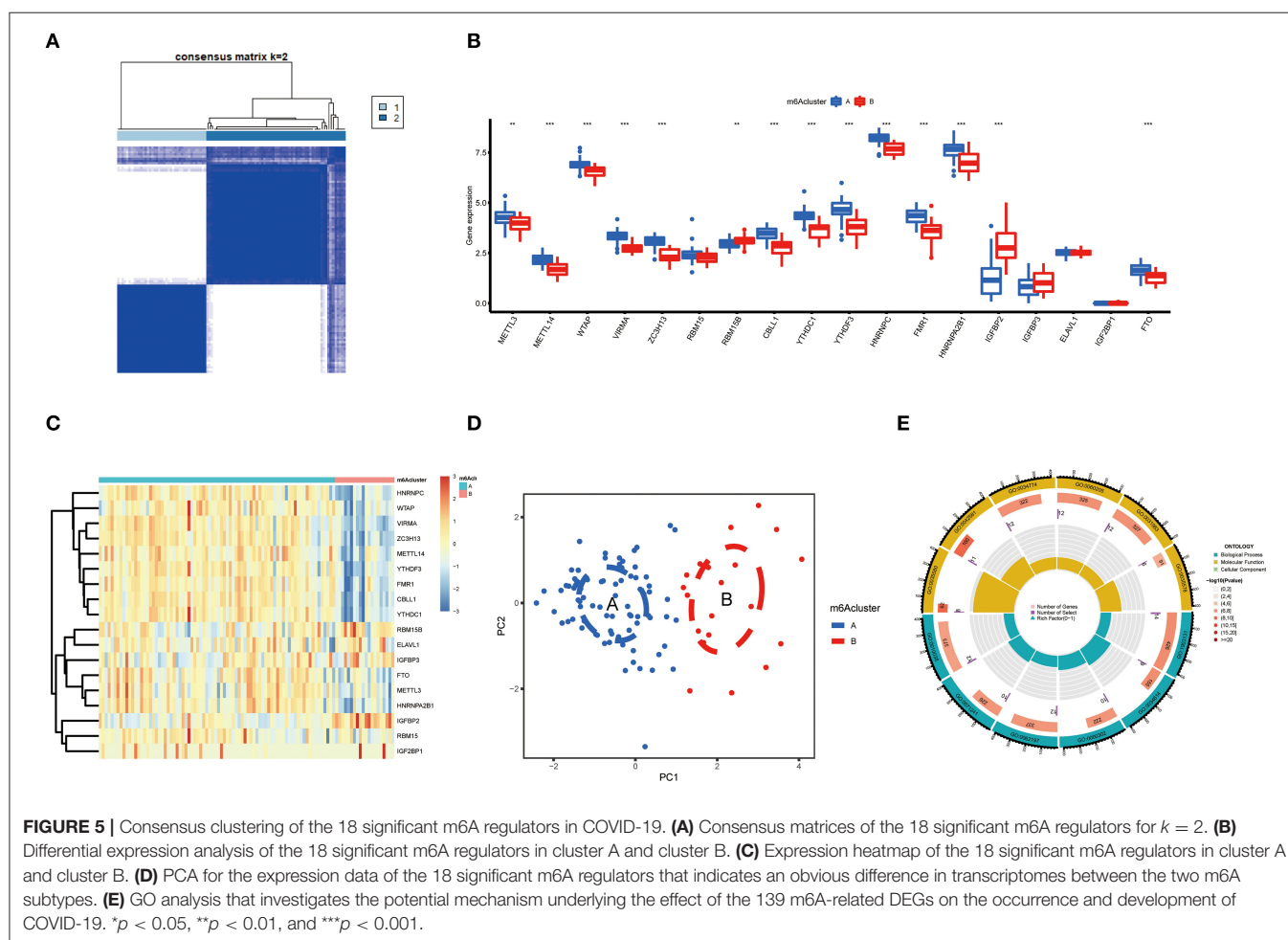
Landscape of the 26 m6A Regulators in COVID-19

Based on the GSE157103 dataset, all samples were divided into three groups (Non-COVID-19, ICU-COVID-19, and Non-ICU-COVID-19). We identified the expression levels of 26 m6A regulators in these groups, of which 22 regulators were differentially expressed in these samples. The expression landscape and heatmap of these differentially expressed genes (DEGs) were presented in **Figures 1A,B**. According to differentially expressed analysis of m6A regulators between COVID-19 samples and Non-COVID-19 samples, 18 DEGs were subsequently observed. Most of DEGs were overexpressed in COVID-19 patients compared to non-COVID-19 patients, including METTL3, METTL14, WTAP, VIRMA, ZC3H13, RBM15, CBLL1, YTHDC1, YTHDF3, HNRNPC, HNRNPA2B1, FMR1, ELAVL1, and FTO, and several DEGs, such as RBM15B, IGFBP2, and IGFBP3 were downregulated in COVID-19 patients. Some of DEGs may be associated with the varying severity of COVID-19, such as METTL3, FTO, and RBM15. The finding was consistent with previous reports (17, 34, 35). We further conducted

GSVA analysis to explore the biological difference between different groups. Compared to samples without COVID-19, p53 signaling pathway, cell cycle, oocyte meiosis, and olfactory transduction were obviously enriched in COVID-19 samples (**Figures 1C,D**). Similarly, we observed that diverse signaling pathways were more enriched in the ICU-COVID-19 samples than Non-ICU-COVID-19 samples, such as oocyte meiosis, ERBB signaling pathway, and TGF- β signaling pathway (**Figure 1E**). These results demonstrated that identified signaling pathways were potentially associated with the occurrence and severity of COVID-19. A protein-protein interaction (PPI) analysis was also performed to show the interactivity of DEGs, which demonstrated that METTL3 and YTHDF3 were hub genes (**Figure 1F**). Additionally, the location of m6A regulators on the chromosome was discussed and displayed in **Figure 1G**.

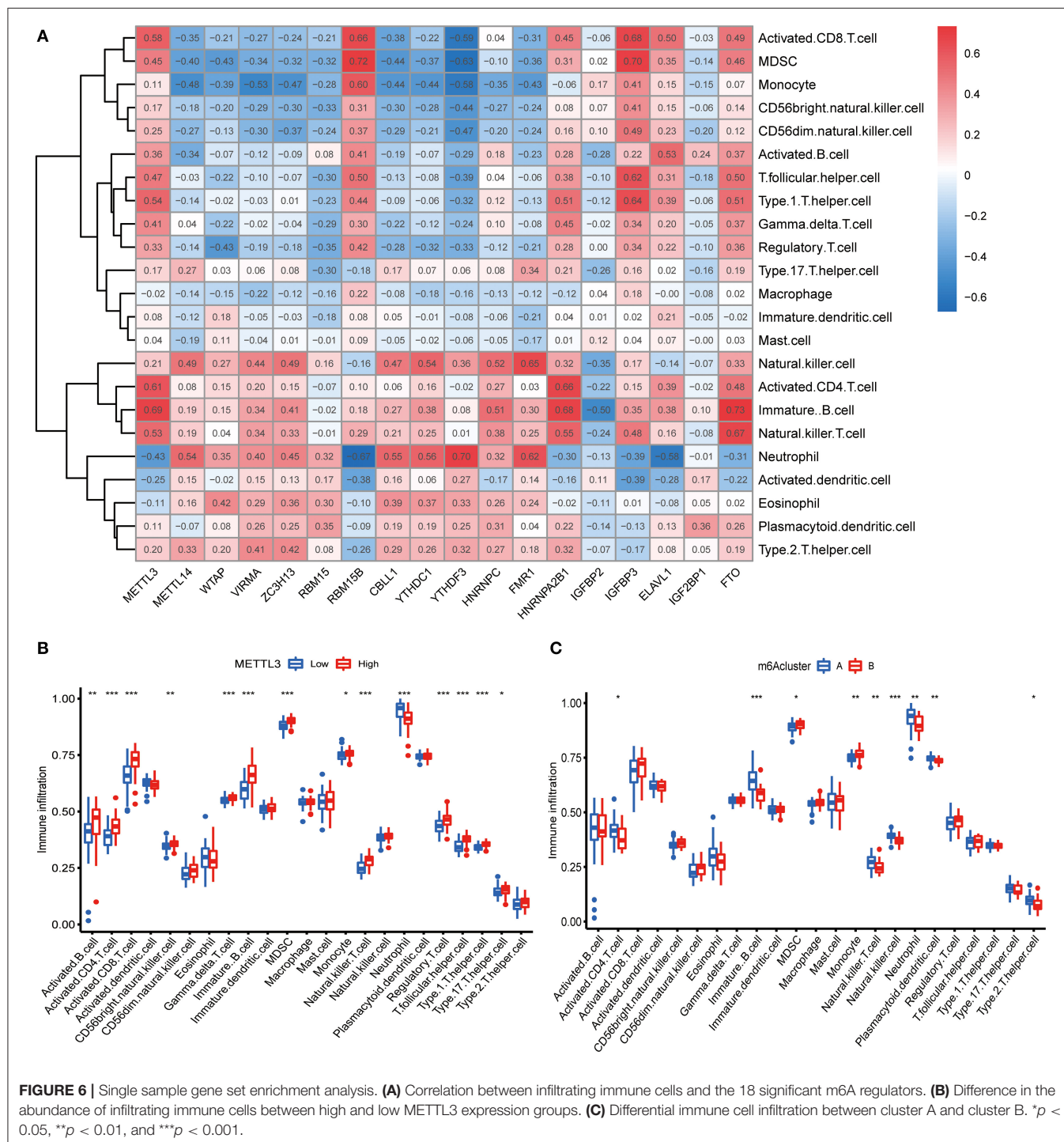
Association Between Writers and Erasers in COVID-19

We investigated the correlation between three types of m6A modification, and the result was presented in **Figure 2A**. Interestingly, m6A regulators of a different type, such as



METTL3 and HNRNPA2B1, can display cooperative activities (coefficient = 0.86). We also discussed the possibility of regulators co-expression, and observed a clear relationship between FTO and additional regulators, with the greatest relevance for METTL3 and FTO (correlation coefficient = 0.83). This finding is consistent with PPI analysis and

provides a possible explanation for the regulation mechanism of m6A regulators. To further investigate the relationship between writers and erasers in COVID-19, we discussed the expression levels of these regulators with linear regression analyses. Significant positive correlations were observed between METTL3, METTL16, RBM15B, VIRMA, and FTO in COVID-19



patients. COVID-19 patients with high expression levels of FTO tend to display high levels of METTL3, METTL16, RBM15B, or VIRMA (Figures 2B–E). Similarly, we also found a close association between CBLL1, METTL14, METLL16, RBM15B, ZC3H13, and ALKBH5. COVID-19 patients with elevated expression levels of CBLL1, METLL16, and RBM15B presented elevated expression levels of ALKBH5 while elevated METTL14 and ZC3H13 expression demonstrated a negative association with ALKBH5 (Figures 2F–J). Consequently, we proved a clear association between diverse writers and erasers.

Evaluation of the RF Model and SVM Model

We next constructed an RF and SVM model to identify optimal m6A regulators from abovementioned DEGs to predict the occurrence of COVID-19. Based on “Reverse cumulative distribution of residual” and “Boxplots of residual” (Figures 3A,B), the RF model with the least residuals were established. As a majority of the specimens in this model retained only small residuals, the predictive performance of the RF model is extremely excellent. Then, we chose 500 trees as the variables of the current model based on the relationship overview between the model error and the number of decision trees, and this model presented a stable error possibility (Figure 3C). We also ranked 18 DEGs depending on their respective gene importance based on RF model, and this result demonstrated that RBM15B and ELAVL1 had a high priority in this model (Figure 3D). Additionally, the ROC curves were established to assess the accuracy of these models, and the AUC value also demonstrated that the RF model has superior performance compared to the SVM model (Figure 3E).

Evaluation of a Predictive Nomogram

Based on the abovementioned findings, 8 recommended m6A regulators were utilized to develop a predictive nomogram for predicting the incidence of COVID-19 (Figure 4A). Interestingly, we observed that the expression level of RBM15B was negatively correlated with the patients’ risk score, and RBM15B may be a protective factor for COVID-19 patients. This result was consistent with abovementioned analysis based on the expression difference in the patients with different disease status. Calibration curves proved the predictive accuracy of the nomogram (Figure 4B). The model developed by the m6A regulator is always at the top of the DCA curve (Figure 4C), indicating that COVID-19 patients were clearly benefited from the decisions based on this nomogram. Furthermore, the clinical impact curve also demonstrated that the predictive robustness of this nomogram was reliable (Figure 4D).

Analysis of Specific Subtypes Based on m6A Regulators

Based on differently expressed m6A regulators, we performed a consensus clustering algorithm to identify different subtypes (Figure 5A), and COVID-19 patients were well-categorized into two clusters when the cluster variable is 2. Cluster A

consisted of 80 cases, and cluster B consisted of 20 cases. Subsequently, we detected the expression of these m6A regulators in cluster A and Cluster B. METTL3, METTL14, WTAP, VIRMA, ZC3H13, CBLL1, YTHDC1, YTHDF3, HNRNPC, FMR1, HNRNPA2B1, and FTO presented increased expression in cluster A compared to those in the cluster B, while the opposite performance was observed in IGF2BP2. Meanwhile, RBM15, RBM15B, IGF2BP3, ELAVL1, and IGF2BP1 displayed no significant differences between these clusters (Figures 5B,C). PCA revealed that the 18 m6A regulators could exactly classify the two m6A subtypes (Figure 5D). Totally, 139 m6A-related DEGs were identified between the two m6A subtypes. To explore the potential role of these DEGs in COVID-19, the findings from GO enrichment analysis revealed that the DEGs were particularly abundant in cellular response and cell differentiation-related pathways (Figure 5E).

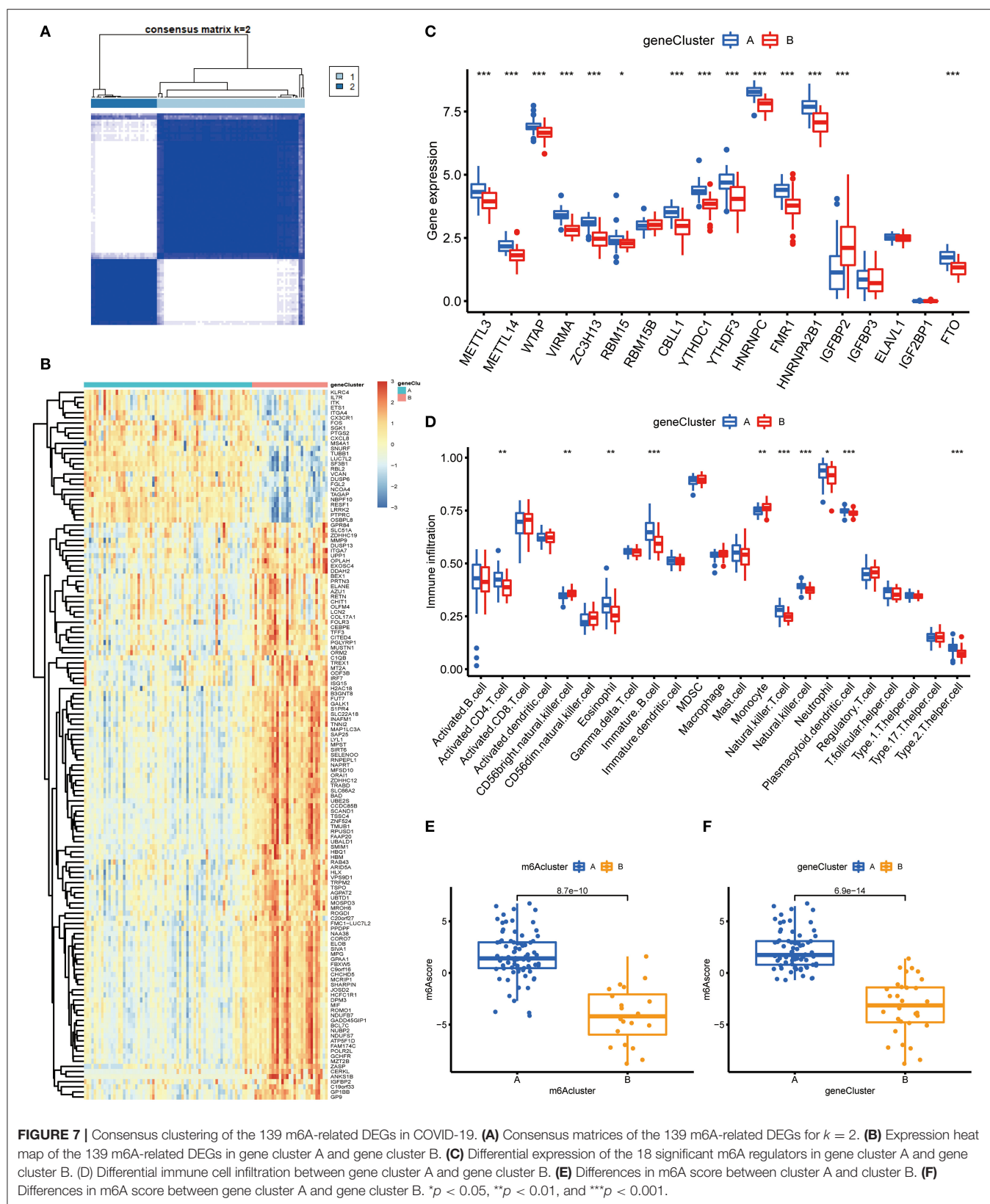
We further conducted ssGSEA to assess the enrichment of immune cells in COVID-19 specimens and discussed the relationship between the m6A regulators and immune cells (Figure 6A). METTL3 had positive associations with various immune cells. Afterward, we investigated the distinct enrichment of immune cells in patients with high- or low-METTL3 (Figure 6B). The findings demonstrated that patients with high METTL3 expression had obviously enriched immune cells. Ultimately, we also discussed the differential immune cell enrichment between the m6A subtypes. We observed that cluster A displayed higher infiltrating levels of immune cells, particularly T helper cells (Th1 and Th2), than cluster B (Figure 6C), which indicated that patients in cluster A may have a positive immune response for COVID-19.

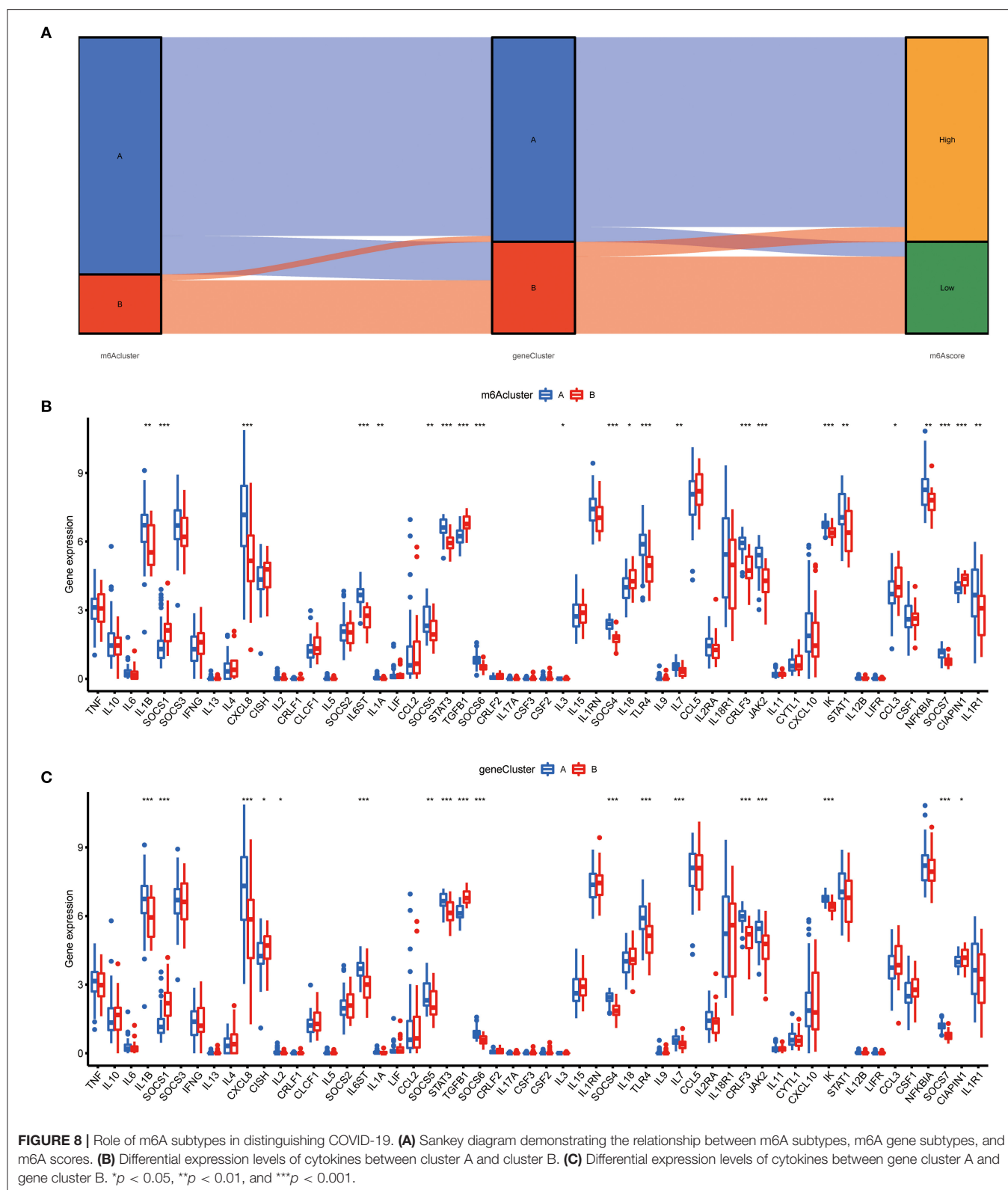
Evaluation of the m6A Gene Signature

To prove the m6A subtypes, we performed the consensus clustering algorithm to categorize the COVID-19 patients into distinct gene subgroups based on 139 m6A-related DEGs (Figure 7A). We observed that these genomic subtypes were in accordance with m6A subtypes, and Figure 7B displayed the differential expression of the 139 DEG. Afterward, the differential expression of the 18 m6A regulators and infiltrating immune cells between different gene clusters were also similar to those in the m6A subtypes (Figures 7C,D). This result demonstrated the rationality of the clustering algorithm. Moreover, PCA was utilized to obtain m6A scores for individual specimens, thus quantifying the m6A subtype. We also compared the m6A score in the m6A clusters or gene clusters, and the finding revealed the m6A score in cluster A or gene cluster A was greater than that in cluster B or gene cluster B (Figures 7E,F). Additionally, the correlation between the m6A cluster, m6A gene clusters, and m6A scores were displayed in Figure 8A.

Relationship Between m6A Subtypes and Cytokines

The “cytokine storm” is an inappropriate immune response that is the main cause of death in COVID-19, and many





cytokines and their inhibitors are now used in the clinical treatment of COVID-19. To further determine the correlation

between m6A subtypes and COVID-19, we comprehensively discussed the association between m6A subtypes and various

cytokines. As displayed in **Figures 8B,C**, diverse cytokines presented significant discrepancies in the m6A clusters and genomic clusters. It is noteworthy that IL1B, IL7, IL8, and IL6ST were overexpressed in the cluster A and gene cluster A compared to cluster B and gene cluster B, consistent with existing reports. This finding revealed that cluster A or gene cluster A is closely correlated with COVID-19 characterized by multiple cytokines.

DISCUSSION

COVID-19 is an infectious respiratory disease with general susceptibility in the population, and there are limited treatment strategies for COVID-19 at present (36). To improve the management and recovery of patients with limited medical facilities, it is essential to clarify the pathogenesis of COVID-19 and the associated susceptible population. Emerging evidence demonstrated that m6A regulators participate in the diverse biological behavior of SARS-CoV-2 (18, 37). However, the potential role of m6A regulators in the COVID-19 is still unclear.

In the present research, we comprehensively explored the basic elements of m6A modification in COVID-19 patients. The expression levels of m6A regulators were obviously overexpressed in COVID-19 patients compared to in non-COVID-19 patients. This different expression of m6A regulators was also observed between COVID-19 patients with ICU status and non-ICU status. These results indicated that m6A modification may have a close correlation with development and severity of COVID-19. We also performed GSVA to identify COVID-19-related pathways and found diverse signaling pathways may serve a critical role in the development of COVID-19, and the exploration of these pathways may be beneficial for clarifying the special mechanism of COVID-19. We further discussed the intrinsic relevance of m6A regulators in the patients with or without COVID-19, and a significant association between m6A regulators in COVID-19 was observed. Moreover, an RF model was constructed to identify 8 regulators from differential expressed m6A regulators and thus predict the occurrence of COVID-19. However, this model cannot yet be validated in the absence of adequate information of m6A regulators in the public databases. Additionally, univariate analysis for feature selection had a possibility to ignore the multivariate association in the feature selection process, and multivariate analysis was further considered to identify optimal DEGs. Previous reports have demonstrated that the selected m6A regulators are responsible for the initiation and progression of tumors, such as hepatocellular carcinoma, lung cancer, and gastric cancer (32, 38, 39). Currently, there are few studies on the correlation between these selected regulators and COVID-19. This study provides a novel option for further genomic analysis on these m6A regulators in the COVID-19 patients.

A multicomponent m6A methyltransferase complex (MTC) consisted of a METTL3-METTL14 heterodimer core and additional binding elements (40). MTC can promote m6A modification to regulate the disease processes. A nomogram based on 8 candidate m6A regulators was constructed to guide clinical treatment for COVID-19 patients, and the

DCA curve demonstrated that COVID-19 patients may benefit from the decisions based on this nomogram. We observed that RBM15B, HNRNPA2B1, and VIRMA may be protective factors in the development of COVID-19, and the opposite performance was found in ELAVL1, RBM15, FMR1, IGFBP3, and METTL3. RBM15 and its paralogue RBM15B bind the m6A-methylation compound and mobilize it to appropriate sites in RNA (41). RBM15 was markedly upregulated in laryngeal squamous cell carcinoma and correlated with a worse prognosis (42). METTL3 serves a critical role in various cellular biological processes, such as promoting the anti-tumor immunity of natural killer cells (43). As a prominent subunit of the MTC, METTL3 facilitates the generation of m6A. It is reported that METTL3 and RBM15 can modulate intrinsic immune responses of the host cell during SARS-CoV-2 infection in diverse cells (18). Similarly, the specific role of VIRMA, ELAVL1, and FMR1 in COVID-19 was mentioned in several studies (44–46). Numerous studies demonstrated that the 8 selected m6A regulators may be involved in the emergence and lymphocyte responses of COVID-19 patients.

At present, the immune response activated by T cells may benefit COVID-19 patients, and reduce the damage caused by cytokine storms (47, 48). Based on DEGs between COVID-19 and non-COVID-19, we found 18 m6A regulators for subsequent analysis. Unsupervised cluster analysis of differential expressed m6A regulators was performed to identify two distinct modification subtypes in COVID-19 patients. m6A cluster A presented activated T cell behaviors, while m6A cluster B was marked by monocyte-related activity. Similar to the m6A categorization, two genomic subtypes were established based on DEGs between cluster A and cluster B, and we found that gene cluster A displayed higher infiltrating levels of T cells than gene cluster B, such as CD4+ T cells and natural killer T cells. JAK-STAT pathway may participate in T cell differentiation (49), and we observed that components in the JAK-STAT pathway were more enriched in cluster A or gene cluster A than those in cluster B or gene cluster B. Consequently, these findings demonstrated that m6A cluster A and gene cluster A with positive T cell activity to defend against SARS-CoV-2 could present a superior clinical performance. Furthermore, the m6A score was identified to quantify the m6A subtype for individual COVID-19 patients. Consistent with the above results, patients in m6A cluster A or gene cluster A displayed higher m6A scores compared to m6A cluster B or gene cluster B.

Nonetheless, there are some limitations in the present research. Since our findings have not been supported by clinical specimens, the specific relationship between m6A regulator and COVID-19 remains to be further confirmed. And this signature will be evaluated and validated in future experimental studies.

CONCLUSION

Briefly, this research identified 8 recommended m6A regulators and constructed a nomogram that predicts

the susceptibility of COVID-19. Based on differently expressed m6A regulators, we then determined two m6A subtypes, and cluster B may be clearly associated with COVID-19.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary

material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XQ performed data collection and analysis. XQ and QC wrote the manuscript. KW polished and revised the manuscript. All authors contributed to the study's conception and design. All authors commented on previous versions of the manuscript, read and approved the final manuscript.

REFERENCES

- Zhong P, Xu J, Yang D, Shen Y, Wang L, Feng Y, et al. COVID-19-associated gastrointestinal and liver injury: clinical features and potential mechanisms. *Signal Transduct Target Ther.* (2020) 5:256. doi: 10.1038/s41392-020-00373-7
- Di Maria E, Latini A, Borgiani P, Novelli G. Genetic variants of the human host influencing the coronavirus-associated phenotypes (SARS, MERS and COVID-19): rapid systematic review and field synopsis. *Hum Genomics.* (2020) 14:30. doi: 10.1186/s40246-020-00280-6
- Rawat K, Kumari P, Saha L. COVID-19 vaccine: a recent update in pipeline vaccines, their design and development strategies. *Eur J Pharmacol.* (2021) 892:173751. doi: 10.1016/j.ejphar.2020.173751
- Wang H, Li X, Li T, Zhang S, Wang L, Wu X, et al. The genetic sequence, origin, and diagnosis of SARS-CoV-2. *Eur J Clin Microbiol Infect Dis.* (2020) 39:1629–35. doi: 10.1007/s10096-020-03899-4
- Fricke-Galindo I, Falfan-Valencia R. Genetics insight for COVID-19 susceptibility and severity: a review. *Front Immunol.* (2021) 12:622176. doi: 10.3389/fimmu.2021.622176
- Ramasamy S, Subbian S. Critical determinants of cytokine storm and type I interferon response in COVID-19 pathogenesis. *Clin Microbiol Rev.* (2021) 34:e00299–20. doi: 10.1128/CMR.00299-20
- Sette A, Crotty S. Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell.* (2021) 184:861–80. doi: 10.1016/j.cell.2021.01.007
- Teijaro JR, Farber DL. COVID-19 vaccines: modes of immune activation and future challenges. *Nat Rev Immunol.* (2021) 21:195–7. doi: 10.1038/s41577-021-00526-x
- Zhang Y, Geng X, Li Q, Xu J, Tan Y, Xiao M, et al. m6A modification in RNA: biogenesis, functions and roles in gliomas. *J Exp Clin Cancer Res.* (2020) 39:192. doi: 10.1186/s13046-020-01706-8
- Sun T, Wu R, Ming L. The role of m6A RNA methylation in cancer. *Biomed Pharmacother.* (2019) 112:108613. doi: 10.1016/j.biopha.2019.108613
- Ma, Z., and Ji, J. (2020). N6-methyladenosine (m6A) RNA modification in cancer stem cells. *Stem Cells.* 38:1511–1519. doi: 10.1002/stem.3279
- Deng LJ, Deng WQ, Fan SR, Chen MF, Qi M, Lyu WY, et al. m6A modification: recent advances, anticancer targeted drug discovery and beyond. *Mol Cancer.* (2022) 21:52. doi: 10.1186/s12943-022-01510-2
- Ma S, Chen C, Ji X, Liu J, Zhou Q, Wang G, et al. The interplay between m6A RNA methylation and noncoding RNA in cancer. *J Hematol Oncol.* (2019) 12:121. doi: 10.1186/s13045-019-0805-7
- Chen M, Wong CM. The emerging roles of N6-methyladenosine (m6A) deregulation in liver carcinogenesis. *Mol Cancer.* (2020) 19:44. doi: 10.1186/s12943-020-01172-y
- Brocard M, Ruggieri A, Locker N. m6A RNA methylation, a new hallmark in virus-host interactions. *J Gen Virol.* (2017) 98:2207–14. doi: 10.1099/jgv.0.000910
- Kim GW, Imam H, Khan M, Mir SA, Kim SJ, Yoon SK, et al. HBV-induced increased N6 methyladenosine modification of PTEN RNA affects innate immunity and contributes to HCC. *Hepatology.* (2021) 73:533–47. doi: 10.1002/hep.31313
- Burgess HM, Depledge DP, Thompson L, Srinivas KP, Grande RC, Vink EI, et al. Targeting the m(6)A RNA modification pathway blocks SARS-CoV-2 and HCoV-OC43 replication. *Genes Dev.* (2021) 35:1005–19. doi: 10.1101/gad.348320.121
- Liu J, Xu YP, Li K, Ye Q, Zhou HY, Sun H, et al. The m(6)A methylome of SARS-CoV-2 in host cells. *Cell Res.* (2021) 31:404–14. doi: 10.1038/s41422-020-00465-7
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* (2013) 41:D991–995. doi: 10.1093/nar/gks1193
- Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. *J Vis Exp.* (2021) 175. doi: 10.3791/62528
- Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics.* (2013) 14:7. doi: 10.1186/1471-2105-14-7
- Wang H, Zhou L. Random survival forest with space extensions for censored data. *Artif Intell Med.* (2017) 79:52–61. doi: 10.1016/j.artmed.2017.06.005
- Winters-Hilt S, Merat S. SVM clustering. *BMC Bioinformatics.* (2007) 8(Suppl. 7):S18. doi: 10.1186/1471-2105-8-S7-S18
- Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O, written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med.* (2019) 7:152. doi: 10.21037/atm.2019.03.29
- Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol.* (2008) 26:1364–70. doi: 10.1200/JCO.2007.12.9791
- Le Thi HA, Le HM, Phan DN, Tran B. Stochastic DCA for minimizing a large sum of DC functions with application to multi-class logistic regression. *Neural Netw.* (2020) 132:220–31. doi: 10.1016/j.neunet.2020.08.024
- Briere G, Darbo E, Thebault P, Uricaru R. Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics.* (2021) 22:361. doi: 10.1186/s12859-021-04279-1
- Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics.* (2010) 26:1572–3. doi: 10.1093/bioinformatics/btq170
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007
- Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* (2012) 16:284–7. doi: 10.1089/omi.2011.0118
- David CC, Jacobs DJ. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol.* (2014) 1084:193–226. doi: 10.1007/978-1-62703-658-0_11
- Zhang B, Wu Q, Li B, Wang D, Wang L, Zhou YL. m(6)A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. *Mol Cancer.* (2020) 19:53. doi: 10.1186/s12943-020-01170-0
- Xiao B, Liu L, Li A, Xiang C, Wang P, Li H, et al. Identification and verification of immune-related gene prognostic signature based on ssGSEA for osteosarcoma. *Front Oncol.* (2020) 10:607622. doi: 10.3389/fonc.2020.607622
- Li N, Hui H, Bray B, Gonzalez GM, Zeller M, Anderson KG, et al. METTL3 regulates viral m6A RNA modification and host cell innate immune responses during SARS-CoV-2 infection. *Cell Rep.* (2021) 35:109091. doi: 10.1016/j.celrep.2021.109091

35. Meng Y, Zhang Q, Wang K, Zhang X, Yang R, Bi K, et al. RBM15-mediated N6-methyladenosine modification affects COVID-19 severity by regulating the expression of multitarget genes. *Cell Death Dis.* (2021) 12:732. doi: 10.1038/s41419-021-04012-z
36. Jacob-Dolan C, Barouch DH. COVID-19 vaccines: adenoviral vectors. *Annu Rev Med.* (2022) 73:41–54. doi: 10.1146/annurev-med-012621-102252
37. Liu R, Ou L, Sheng B, Hao P, Li P, Yang X, et al. Mixed-weight Neural Bagging for Detecting m6A Modifications in SARS-CoV-2 RNA Sequencing. *IEEE Trans Biomed Eng.* (2022). doi: 10.1109/TBME.2022.3150420
38. Zhou T, Li S, Xiang D, Liu J, Sun W, Cui X, et al. m6A RNA methylation-mediated HNF3gamma reduction renders hepatocellular carcinoma dedifferentiation and sorafenib resistance. *Signal Transduct Target Ther.* (2020) 5:296. doi: 10.1038/s41392-020-00299-0
39. Yin H, Chen L, Piao S, Wang Y, Li Z, Lin Y, et al. M6A RNA methylation-mediated RMRP stability renders proliferation and progression of non-small cell lung cancer through regulating TGFBR1/SMAD2/SMAD3 pathway. *Cell Death Differ.* (2021). doi: 10.1038/s41418-021-00888-8 [Online ahead of print].
40. Shen L, Liang Z, Gu X, Chen Y, Teo ZW, Hou X, et al. N(6)-methyladenosine RNA modification regulates shoot stem cell fate in arabidopsis. *Dev Cell.* (2016) 38:186–200. doi: 10.1016/j.devcel.2016.06.008
41. Zhang L, Tran NT, Su H, Wang R, Lu Y, Tang H, et al. Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing. *Elife.* (2015) 4:e07938. doi: 10.7554/eLife.07938.036
42. Wang X, Tian L, Li Y, Wang J, Yan B, Yang L, et al. RBM15 facilitates laryngeal squamous cell carcinoma progression by regulating TMBIM6 stability through IGF2BP3 dependent. *J Exp Clin Cancer Res.* (2021) 40:80. doi: 10.1186/s13046-021-01871-4
43. Lin S, Choe J, Du P, Triboulet R, Gregory RI. The m(6)A methyltransferase METTL3 promotes translation in human cancer cells. *Mol Cell.* (2016) 62:335–45. doi: 10.1016/j.molcel.2016.03.021
44. Yue Y, Liu J, Cui X, Cao J, Luo G, Zhang Z, et al. VIRMA mediates preferential m(6)A mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov.* (2018) 4:10. doi: 10.1038/s41421-018-0019-0
45. Aloufi N, Haidar Z, Ding J, Nair P, Benedetti A, Eidelman DH, et al. Role of human antigen R (HuR) in the regulation of pulmonary ACE2 expression. *Cells.* (2021) 11:22. doi: 10.3390/cells11010022
46. Jiang X, Liu B, Nie Z, Duan L, Xiong Q, Jin Z, et al. The role of m6A modification in the biological functions and diseases. *Signal Transduct Target Ther.* (2021) 6:74. doi: 10.1038/s41392-020-00450-x
47. Azkur AK, Akdis M, Azkur D, Sokolowska M, van de Veen W, Bruggen MC, et al. Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. *Allergy.* (2020) 75:1564–81. doi: 10.1111/all.14364
48. Toor SM, Saleh R, Sasidharan Nair V, Taha RZ, Elkord E. T-cell responses and therapies against SARS-CoV-2 infection. *Immunology.* (2021) 162:30–43. doi: 10.1111/imm.13262
49. Meyer LK, Verbist KC, Albeituni S, Scull BP, Bassett RC, Strohm AN, et al. JAK/STAT pathway inhibition sensitizes CD8 T cells to dexamethasone-induced apoptosis in hyperinflammation. *Blood.* (2020) 136:657–68. doi: 10.1182/blood.2020006075

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Qing, Chen and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



An Explainable AI Approach for the Rapid Diagnosis of COVID-19 Using Ensemble Learning Algorithms

Houwu Gong^{1,2†}, Miye Wang^{3,4†}, Hanxue Zhang¹, Md Fazla Elahe¹ and Min Jin^{1*}

¹ Department of Software Engineering, College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, ² Academy of Military Sciences, Beijing, China, ³ Engineering Research Center of Medical Information Technology, Ministry of Education, West China Hospital, Chengdu, China, ⁴ Information Center, West China Hospital, Chengdu, China

OPEN ACCESS

Edited by:

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

Reviewed by:

Wellington Pinheiro dos Santos,
Federal University of
Pernambuco, Brazil
Joseph Bamidele Awotunde,
University of Ilorin, Nigeria

*Correspondence:

Min Jin
jinmin@hnu.edu.cn

[†]These authors share first authorship

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 12 February 2022

Accepted: 19 May 2022

Published: 21 June 2022

Citation:

Gong H, Wang M, Zhang H, Elahe MF
and Jin M (2022) An Explainable AI
Approach for the Rapid Diagnosis of
COVID-19 Using Ensemble Learning
Algorithms.
Front. Public Health 10:874455.
doi: 10.3389/fpubh.2022.874455

Background: Artificial intelligence-based disease prediction models have a greater potential to screen COVID-19 patients than conventional methods. However, their application has been restricted because of their underlying black-box nature.

Objective: To address this issue, an explainable artificial intelligence (XAI) approach was developed to screen patients for COVID-19.

Methods: A retrospective study consisting of 1,737 participants (759 COVID-19 patients and 978 controls) admitted to San Raphael Hospital (OSR) from February to May 2020 was used to construct a diagnosis model. Finally, 32 key blood test indices from 1,374 participants were used for screening patients for COVID-19. Four ensemble learning algorithms were used: random forest (RF), adaptive boosting (AdaBoost), gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost). Feature importance from the perspective of the clinical domain and visualized interpretations were illustrated by using local interpretable model-agnostic explanations (LIME) plots.

Results: The GBDT model [area under the curve (AUC): 86.4%; 95% confidence interval (CI) 0.821–0.907] outperformed the RF model (AUC: 85.7%; 95% CI 0.813–0.902), AdaBoost model (AUC: 85.4%; 95% CI 0.810–0.899), and XGBoost model (AUC: 84.9%; 95% CI 0.803–0.894) in distinguishing patients with COVID-19 from those without. The cumulative feature importance of lactate dehydrogenase, white blood cells, and eosinophil counts was 0.145, 0.130, and 0.128, respectively.

Conclusions: Ensemble machine learning (ML) approaches, mainly GBDT and LIME plots, are efficient for screening patients with COVID-19 and might serve as a potential tool in the auxiliary diagnosis of COVID-19. Patients with higher WBC count, higher LDH level, or higher EOT count, were more likely to have COVID-19.

Keywords: artificial intelligence, ensemble learning, explainable, disease prediction, COVID-19

INTRODUCTION

Coronavirus disease 2019 (COVID-19, also called novel coronavirus pneumonia) is characterized by fever, cough, and shortness of breath. COVID-19 spreads rapidly due to its highly infectious nature, and caused huge manpower and material resources losses (1, 2). Early detection, diagnosis, isolation, and treatment are keys to improving the cure and survival rates of COVID-19 patients.

To respond to this unprecedented pandemic emergency, early identification of infected patients is very important. Infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes COVID-19 is typically identified with molecular detection using reverse transcriptase PCR (RT-PCR) as the gold standard (3). However, the test process is time-consuming (no <4 h under ideal conditions) and requires the use of special equipment and reagents and specialized and trained personnel for sample collection. Furthermore, the high cost and slow processing speed of RT-PCR make it less feasible for massive population screening in remote areas or backward countries (4). The development of artificial intelligence (AI) technology has made the mining of medical information and the development of disease prediction models for assisting doctors in disease prediction or diagnosis a popular research subject.

To improve the ability to diagnose COVID-19 and curb the spread of the pandemic, the data science community has proposed several machine learning (ML) models, most of which are based on computed tomography (CT) scans or chest X-rays (5–9). Although promising results have been reported, some concerns have been raised about these efforts, especially the chest X-ray-based solutions, regarding the high incidence of false negative results (10). Additionally, while the CT imaging method is accurate, it is costly and time-consuming and requires specialized equipment. As a result, methods based on this imaging technology are inappropriate for screening. Although various clinical studies (11–15) have emphasized the usefulness of blood test-based diagnoses in providing an effective and low-cost alternative for the early detection of COVID-19, relatively few ML models are based on hematological parameters.

The primary goal of medicine in the 21st century has switched from disease prevention and treatment to health maintenance, and the medical mode has changed from a simple disease treatment mode to the so-called “4P” medical mode: prevention, prediction, personalization, and participation (16). To address issues regarding medical complexity, the methodological system of clinical research is also constantly improving. A disease prediction model is a statistical evaluation method based on disease risk factors that divides scores according to the degree of influence of the underlying factor and calculates the probability of a certain event in the future by a mathematical formula (17). These disease prediction models can enable medical staff to implement targeted intervention measures for patients with different risk probabilities and improve patient care. Due to the powerful ability to mine information and explore the hidden links behind the data, machine learning algorithms have been used in many studies and a wide variety of fields to develop predictive models of disease risk.

As the main caregivers for patients, nurses play a key role in patient condition observation and disease prediction. Compared with traditional risk prediction models or scores, machine learning models are more precise, sensitive, and generalizable, capable of analyzing the deep-seated interaction of multiple factors among data (18) and explore more complex linear or nonlinear correlations. In diverse clinical situations, the capacity to forecast disease risk using the ML technique is greater, which is vital for encouraging medical professionals to intervene early to enhance patient care.

The core of machine learning is the algorithm, which has three main learning patterns: (1) supervised learning, which adjusts the prediction algorithm based on the previous examples to make the prediction results match as close as possible to the output values of the examples when reinput; (2) unsupervised learning, which does not output a value; instead, the training system models the underlying structure of the data; and (3) reinforcement learning, which uses reward/punishment sequences to form strategies for action in a specific problem space through trial and error (19). Machine learning adopts supervised learning algorithms such as support vector machine (SVM), Bayesian learning, decision tree, and regression, and unsupervised learning algorithms such as K-means clustering and association rule learning. Reinforcement learning algorithms (20), such as Q-learning (21) and SARSA, as well as neural networks and other special algorithms, are also implemented in machine learning. At present, the main idea of the quantitative identification technology of disease prediction is to transform the problem of disease risk into a classification problem and then use the corresponding model to perform the classification. According to the literature, the most commonly used and best performing algorithms for disease prediction (22) include SVM, backpropagation (BP) neural network, random forest, and naive Bayes.

However, only single prediction models are implemented in these studies, and the accuracy and stability need to be improved. Ensemble learning is based on the idea of learning from the strengths of others. Constructing and combining multiple machine learning devices to complete the learning task can effectively prevent overfitting and underfitting problems and thus improve the prediction performance (23). In the disease prediction task, there are some problems, such as high feature dimension, multicollinearity between features, and highly noisy physical examination data, that can produce unideal stability in single models. To overcome the above problems and obtain better stability, this paper proposes an ensemble learning method to integrate multiple models to predict disease risk. Bagging and boosting strategies are adopted to evaluate disease prediction based on the ensemble idea.

Prediction models can be coarsely divided into “black-box” and “white-box” models. Most existing prediction models in the medical and health fields are “white-box” models due to the high demands for comprehensibility, interpretability, and transparency. These “white-box” models, which include linear regression and decision tree, have a strong visualization ability but relatively poor prediction precision (24). If the prediction problem is difficult and requires high precision, neural networks, random forests, and other “black-box” models must be used (25).

Abbreviations: AdaBoost, adaptive boosting; AUC, area under the curve; BP, backpropagation; CBC, complete blood count; CI, confidence interval; GBDT, gradient boosting decision tree; GGT, gamma-glutamyl transferases; HCT, hematocrit; HGB, hemoglobin; JMIR, Journal of Medical Internet Research; LDH, lactate dehydrogenase; LIME, local interpretable model-agnostic explanations; ML, machine learning; OSR, San Raphael Hospital; PAC, probably approximately correct; RF, random forest; RCT, randomized controlled trial; XAI, explainable artificial intelligence; XGBoost, extreme gradient boosting.

In recent years, explainable machine learning has become a popular topic in different research fields (26). Explainable machine learning focuses on improving the transparency and credibility of black-box model decision-making. There are two methods for bestowing explicability to a predictive model. First, intrinsically interpretable machine learning methods, such as logistic regression, can be used as the predictive model. Second, postinterpretation methods, such as local interpretable model-agnostic explanations (LIME) (27) and SHapely Additive exPlanations (SHAP) (28), explain complex models through postassisted attribute analysis. This paper improves upon LIME and uses an explainable additive model proposed in recent years to approximate the complex model further to improve the interpretability of the ensemble learning model.

This work aims to overcome the limitations described above by building a COVID-19 diagnostic model based on hematological parameters to provide a new method to screen COVID-19. Different classification models have been developed by applying AI technology to blood test results that can be obtained in a short amount of time (<10 min even in an emergency) and at only a small percentage of the cost of RT-PCR and CT. Our approach can be used to screen COVID-19 patients using regular blood tests in resource-constrained situations, especially during the peak of an outbreak, when RT-PCR reagent shortages become a severe issue. The developed method can also be used as a supplement to RT-PCR tests to increase their sensitivity.

METHODS

Data Sources

COVID-19 spread rapidly throughout many countries worldwide (29, 30). Early identification of COVID-19 patients and SARS-CoV-2-infected persons is very important and can play a key role in epidemic prevention and control. Therefore, the routine blood test data of patients with COVID-19 was used in this study (31). The data were extracted from a database including the hematochemical values from 1,737 patients (47.00% COVID-19 positive) admitted to San Raphael Hospital (OSR) from February to May 2020. Patient age and sex, the presence of COVID-19-related symptoms at admission (dyspnea, pneumonia, pyrexia, sore throat, influenza, cough, pharyngitis, bronchitis, generalized illness), and a set of hematochemical values from laboratory tests (complete blood count and coagulation, biochemical, blood gas analysis and CO-oximetry values) were considered covariate features. The goal of this study is to classify patients as positive or negative for COVID-19.

Feature Selection

First, features with no significant differences between the positive and negative COVID-19 groups were eliminated. Student's *t*-test or the Kruskal–Wallis test were used to compare continuous variables, which are presented as the mean \pm standard deviation. The chi-square test was used to compare categorical variables, which are presented as frequencies and percentages. A two-tailed *p* value of <0.05 was considered statistically significant. Then, feature correlation analysis was performed according to the

Pearson correlation coefficient matrix. Highly correlated features were eliminated to avoid issues related to multicollinearity.

Machine Learning Algorithms

Four ensemble learning algorithms, including random forest (RF), adaptive boosting (AdaBoost), gradient boosting decision tree (GBDT) and eXtreme gradient boosting (XGBoost), are used as representative boosting algorithms to determine the best performing model. The most optimal variables were further validated using the GBDT method.

Compared with single learning models, the advantage of an ensemble learning model is that it can combine multiple single learning models to obtain more accurate, stable, and robust results (32). The principle of ensemble learning came from the probably approximately correct (PAC) learning model (33). Kearns and Valiant first explored the equivalence of weak and strong learning algorithms (34). Bagging and boosting strategies both combine existing classification algorithms or regression algorithms to form a more powerful predictor. In this paper, RF was used as the representative bagging algorithm. AdaBoost, GBDT, and XGBoost are used as representative boosting algorithms.

Bagging

Bagging, also known as bootstrap aggregation, refers to the use of bootstrapping to extract training samples under the same base classifier to train multiple base classifiers and finally obtain the results through a voting method. This approach can help reduce errors caused by random fluctuations in the training data (35). The steps of the bagging process are as follows. The training sets are extracted from the original sample set. In each round, *n* training samples are extracted from the original sample set by bootstrapping, and a total of *k* rounds of extraction are performed to obtain *k* training sets. One training set is used to obtain a model, and so *k* training sets obtain a total of *k* models. [The model can be determined according to the specific situation; it can be a decision tree, K-nearest neighbor (KNN), etc.] The classification results are produced by voting.

Boosting

Boosting transforms weak learners into strong learners through iteration. By increasing the number of iterations, a strong learner with high performance is generated (36); this is considered one of the best performing approaches in machine learning. Boosting increases the weights of samples that were incorrectly classified by the weaker classifier in the previous round and decreases the weights of samples that were correctly classified in the previous round so that the classifier has a better effect on the misclassified data. The final boosting model is obtained according to this rule. The main idea is to combine multiple weak classifiers into one strong classifier. Under the PAC learning framework, the weak classifier must be assembled into a strong classifier.

Model Validation

All patients were randomly divided into training and testing sets at a ratio of 8:2. To minimize the randomness effect

of the training result, 10-fold cross-validation was also adopted. First, the training sets are divided into 10-fold, then the model is trained with nine-fold and verified with the remaining fold. The training is repeated for 10 times, with each a different fold for verification, and the average value of the performance is represented as the generalization performance. Once the models were derived, the performances of the different models were further validated using the receiver operating characteristic (ROC) curve as the evaluation metric. The accuracy, precision, recall, sensitivity, F1 score, youden's index and area under the curve (AUC) were calculated to evaluate the performance of the ML algorithm on testing sets. Finally, the optimal ML algorithm was selected.

Model Interpretation

The local interpretable model-agnostic explanation (LIME) was used to explain the predictions. The rationale by which a model predicts a single sample using a local linear approximation of the model behavior can be better trusted.

LIME, proposed by (27), is a tool that helps explain how a complex black-box model makes decisions. A new dataset is generated by randomly perturbing the samples in LIME. The new dataset is then used to train a linear model, which locally approximates the black-box model. Then, the local decision behavior of the black-box model is obtained according to the interpretable model.

Note that $x \in R^d$ are the samples that need to be interpreted. First, the more important d' dimensional features are selected, and x becomes $x' \in R^{d'}$ after removing the less important features. A new sample z' is generated by perturbing x' , and the all-new samples constitute a new dataset Z' . After adding the removed features to the samples, z' is restored to $Z \in R^d$. $\pi_x(z)$ is defined as the similarity of samples before and after modification and can be calculated as follows:

$$\pi_x(z) = \exp\left(-\frac{D(x, z)^2}{\sigma^2}\right), \quad (1)$$

where $D(x, z)$ is the distance formula, whose definition varies with the sample type. When the sample is an image, for example, $D(x, z)$ is usually the $L2$ norm distance, and when it is text, $D(x, z)$ is usually the cosine similarity function.

If f is the complex model to be explained and g is a simple model, the objective function to measure the difference between the two models is as follows:

$$\xi(x) = \sum_{z, z'} \pi_x(z) f(z) - g(z')^2 + \Omega(g), \quad (2)$$

where $\Omega(g)$ is the complexity of model g . When g is a linear regression model, the number of nonzero weight coefficients determines the model's complexity. The flow of the LIME algorithm is shown in Table 1.

Statistical Analyses

Categorical variables were described as number (%) and compared by Chi-square or Fisher's exact test where appropriate.

TABLE 1 | Algorithm: LIME.

Algorithm: LIME

Input: (1) Complex Model f ; (2) Samples X ; (3) Number of randomly generated samples N

Steps:

1. Through feature screening, the more important d' features are preliminarily obtained, allowing the interpretation version X' of X to be obtained
2. A new sample Z' is generated by randomly perturbing X' ; then, Z' is restored to Z with the same dimensions as X . The complex model is used to predict and obtain the labels
3. The newly generated dataset is fitted with a linear model

Output: The weight of the linear model

Continuous variables that satisfy normal distribution were described as mean [standard deviation (SD)] and compared by the 2-tailed Student's t -test; otherwise, median [interquartile range (IQR)] and Wilcoxon Mann-Whitney U -test were used. A two-sided p -value < 0.05 was considered statistically significant. All statistical analyses were performed with Python (version 3.8.5).

RESULTS

Among 1,736 patients. 362 patients were excluded because they had more than four missing attribute values. After processing, 1,374 patients remained in the database. Two features (CK and UREA) were removed because their missing value was larger than 30% of their overall value; the average value of each feature was used to fill in the remaining missing values. Thirty-two features were selected for screening patients for COVID-19 (Table 2).

Baseline Characteristics

Table 2 presents the characteristics of the positive and negative COVID-19 patients. The chi-square test for sex yielded a Pearson's chi-square value of 14.918, and $p = 0.000$ (close to but not equal to zero) < 0.05 , indicating that the sex differences between the positive and negative COVID-19 groups were significant. In contrast, Student's t -test or the Kruskal-Wallis test showed that there was no difference in age, CREA, KAL, or MCH between the two groups ($p > 0.05$).

Figure 1 shows that Sex ($r = 0.13$), GGT ($r = 0.07$), GLU ($r = 0.11$), AST ($r = 0.22$), ALT ($r = 0.18$), LDH ($r = 0.24$), PCR ($r = 0.23$), RBC ($r = 0.17$), HGB ($r = 0.17$), HCT ($r = 0.16$), MCHC ($r = 0.10$), NE ($r = 0.14$), and Suspect ($r = 0.32$) were positively correlated with the target, while, CA ($r = -0.14$), ALP ($r = -0.09$), NAT ($r = -0.10$), WBC ($r = -0.22$), MCV ($r = -0.06$), PLT1 ($r = -0.11$), LY ($r = -0.09$), MO ($r = -0.05$), EO ($r = -0.31$), BA ($r = -0.31$), NET ($r = -0.14$), LYT ($r = -0.26$), MOT ($r = -0.17$), EOT ($r = -0.31$), and BAT ($r = -0.29$) were negatively correlated with the target. Therefore, we believed that there were no redundant features and selected all of them to develop the model.

TABLE 2 | Characteristics of the positive and negative COVID-19 patients.

	Total (N = 1,374)	COVID-19 negative (N = 615)	COVID-19 positive (N = 759)	p-Value
Age, year	60.40 ± 20.83	60.40 ± 20.83	62.27 ± 15.84	0.066
Female	583 (42.43%)	304 (49.43%)	279 (36.76%)	<0.001
CA, mmol/L	2.20 ± 0.751	2.29 ± 0.74	2.14 ± 0.14	<0.001
CREA, mg/dl	1.18 ± 1.01	1.22 ± 1.20	1.14 ± 0.82	0.180
ALP, U/L	87.74 ± 64.26	94.18 ± 77.16	82.53 ± 50.95	0.001
GGT, U/L	66.12 ± 101.95	58.52 ± 118.90	72.27 ± 85.40	0.013
GLU, mg/dl	119.03 ± 55.85	112.19 ± 49.85	124.58 ± 59.73	<0.001
AST, U/L	47.11 ± 51.37	34.60 ± 33.44	57.25 ± 60.37	<0.001
ALT, U/L	40.15 ± 40.67	32.23 ± 35.22	46.56 ± 43.58	<0.001
LDH, U/L	336.86 ± 210.61	280.76 ± 243.48	382.33 ± 166.44	<0.001
PCR,	72.22 ± 79.59	52.86 ± 70.90	89.72 ± 82.43	<0.001
KAL	4.22 ± 0.51	4.25 ± 0.50	4.20 ± 0.52	0.101
NAT	138.58 ± 4.66	139.10 ± 3.92	138.15 ± 5.15	<0.001
WBC, 10 ⁹ /L	8.56 ± 4.75	9.73 ± 5.45	7.62 ± 3.85	<0.001
RBC, 10 ¹² /L	4.53 ± 0.73	4.40 ± 0.75	4.64 ± 0.69	<0.001
HGB, g/dl	13.18 ± 2.05	12.80 ± 2.13	13.49 ± 1.94	<0.001
HCT, %	39.32 ± 5.64	38.32 ± 5.79	40.14 ± 5.39	<0.001
MCV, fl	87.33 ± 6.93	87.76 ± 7.23	86.97 ± 6.65	<0.001
MCH, pg/cell	29.25 ± 2.69	29.27 ± 2.76	29.23 ± 2.63	0.783
MCHC, g Hb/dl	33.48 ± 1.34	33.34 ± 1.35	33.60 ± 1.32	<0.001
PLT1, 10 ⁹ /L	234.74 ± 95.89	246.55 ± 98.70	225.17 ± 92.51	<0.001
NE, %	72.35 ± 13.26	70.33 ± 13.47	73.98 ± 12.86	<0.001
LY, %	18.58 ± 11.00	19.73 ± 11.37	17.65 ± 10.62	0.001
MO, %	7.83 ± 3.88	8.06 ± 3.61	7.65 ± 4.08	0.045
EO, %	0.88 ± 1.62	1.43 ± 2.02	0.44 ± 1.00	<0.001
BA, %	0.34 ± 0.327	0.43 ± 0.31	0.26 ± 0.21	<0.001
NET, 10 ⁹ /L	6.45 ± 4.48	7.15 ± 5.28	5.88 ± 3.60	<0.001
LYT, 10 ⁹ /L	1.37 ± 0.95	1.64 ± 1.02	1.15 ± 0.83	<0.001
MOT, 10 ⁹ /L	0.62 ± 0.54	0.72 ± 0.45	0.54 ± 0.59	<0.001
EOT, 10 ⁹ /L	0.07 ± 0.14	0.12 ± 0.18	0.03 ± 0.08	<0.001
BAT, 10 ⁹ /L	0.02 ± 0.04	0.03 ± 0.05	0.01 ± 0.02	<0.001
Suspect, %	0.83 ± 0.33	0.71 ± 0.39	0.92 ± 0.23	<0.001

CA, calcium; CREA, creatinine; ALP, alkaline phosphatase; GGT, gamma-glutamyl transferase, an enzyme that converts glutamyl to glutamine; GLU, glucose; AST, aspartate aminotransferase; ALT, alanine aminotransferase; LDH, lactate dehydrogenase, a type of enzyme that breaks down lactate; WBC, white blood cell; RBC, red blood cell; HGB, hemoglobin, a protein that transports oxygen throughout the body; HCT, hematocrit, a metric representing the proportion of RBCs in the blood; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; PLT1, platelets; NE, neutrophil count (%); LY, lymphocyte count (%); MO, monocyte count (%); EO, eosinophil count (%); BA, basophil count (%); NET, neutrophil count; LYT, lymphocyte count; MOT, monocyte count; EOT, eosinophil count; BAT, basophil count; Suspect, suspected COVID-19.

ML Algorithms' Performance Comparison

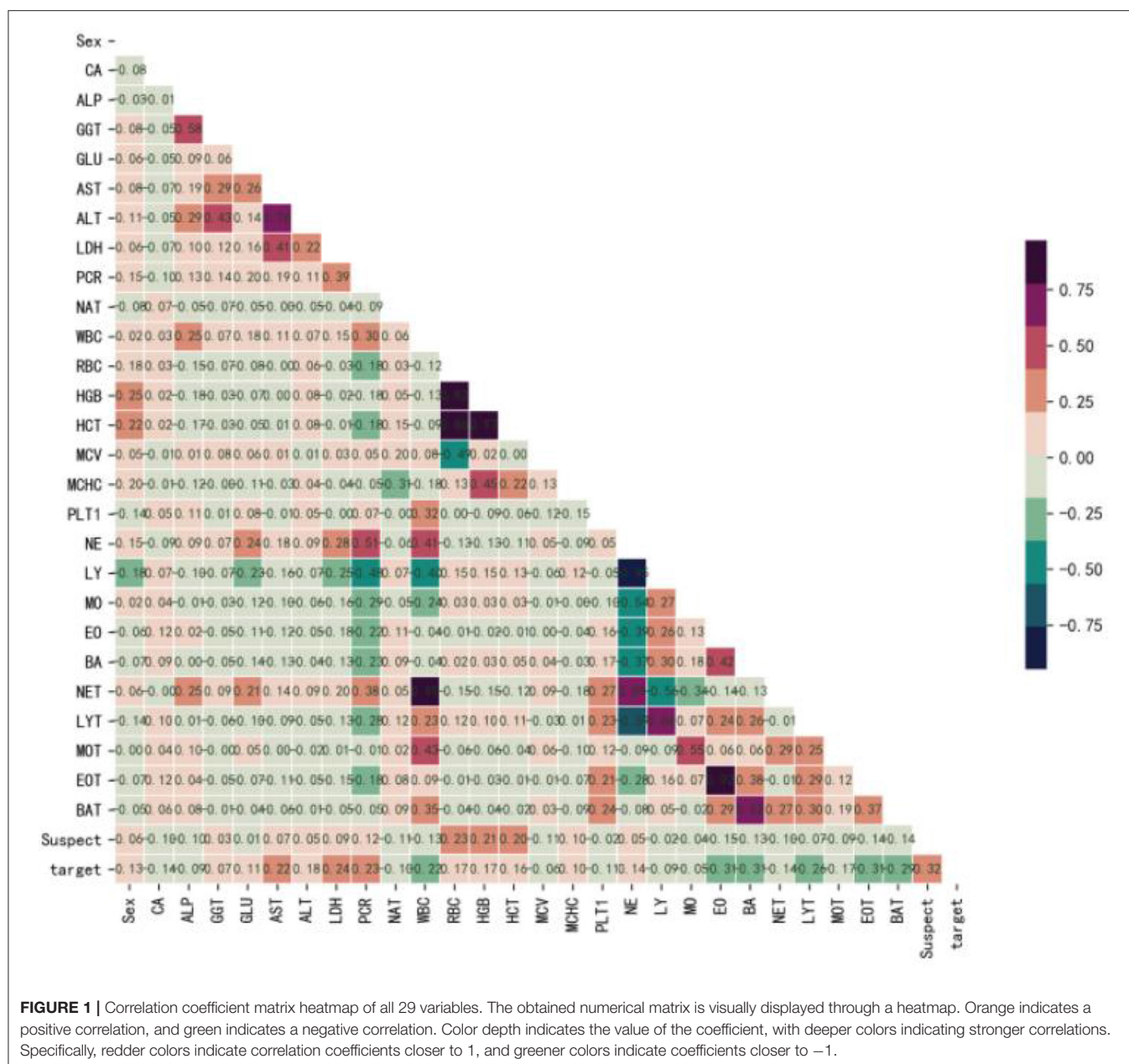
Data from 80% of the 1,374 patients were randomly selected and used as the training set, while the data from the remaining 20% of the patients were used as the testing set. The prediction models were developed with the training set, and their performance was evaluated with the testing set. Random forest, AdaBoost, GBDT, and XGBoost were selected as the typical algorithms of the ensemble learning model. The performance of the ML models was evaluated by using the area under the receiver operating characteristic curve (AUC).

The GBDT algorithm had the best fitting effect on the COVID-19 dataset, with an accuracy of 93.8% and an AUC of 98.4% [95% CI (0.978, 0.990)] on the training set and 80.4 and

86.4% [95% CI (0.821, 0.907)], respectively, on the test set (see **Tables 3, 4** for details on the performance metrics).

As shown in **Figure 2**, the performance of GBDT was better than that of random forest, AdaBoost, and XGBoost. DeLong's test was further used to assess the difference between two AUCs, which confirmed that the AUC of the GBDT model was significantly different from that of the other three models ($p < 0.01$).

A calibration curve was obtained with the bucket method (continuous data discretization) to observe whether the prediction probability of the classification model was close to the actual probability. It is an evaluation index of a probability model. The calibration curve of the GBDT model was drawn



with the predicted probability as the abscissa and the true probability in each bin as the ordinate. As shown in **Figure 3**, the calibration curve was close to the diagonal, indicating that in the model testing experiment, the GBDT model performed well.

Explanation of the Best Model

Feature Importance of GBDT

The meaning of “GradientBoostingClassifier ($n_estimators = 100$, $learning_rate = 1.0$, $max_depth = 1$, $random_state = 0$)” in classifying the patients could not be explained to the doctors sufficiently. In general, the interpretability of GBDT is reflected in its feature importance. The feature importance derived from the XGBoost model is shown in **Figure 4**.

Interpretation by LIME

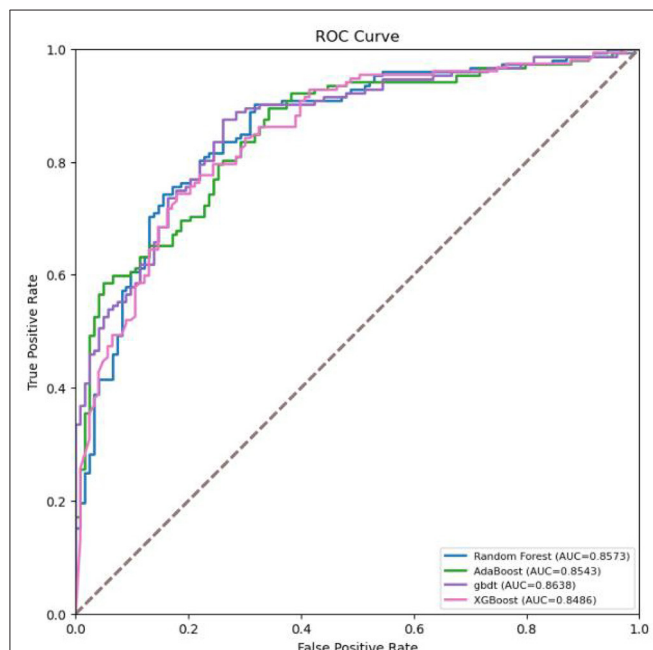
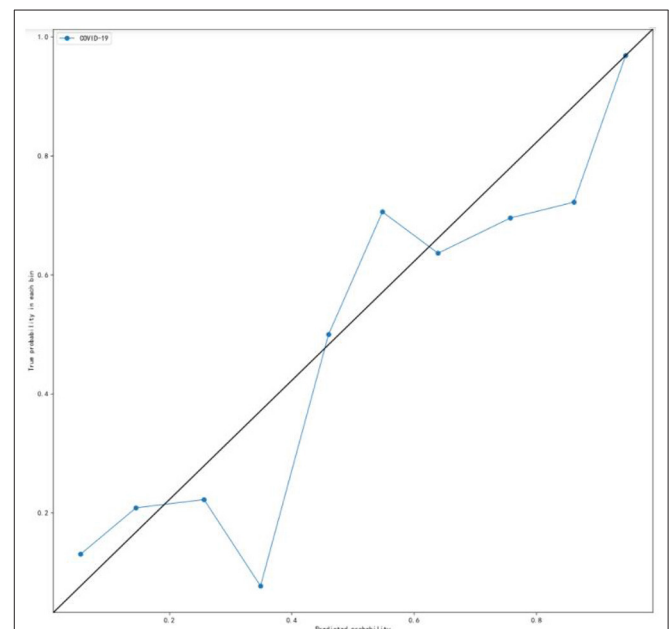
Local interpretable model-agnostic explanations selects a specific sample in the test dataset to obtain the probability value of each class and explains the reason for assigning the probability. **Figure 5** shows the prediction results of the sample. The figure shows which features determined that the sample should be classified as COVID-19 positive (blue) and which determined that the sample should be classified as COVID-19 negative (orange). The values of the features for the sample are listed in the figure to show the contribution of the features. Specifically, CA, PCR, and LDH were important factors for determining positive COVID-19 patients. These three features were further discretized

TABLE 3 | Performance of random forest, AdaBoost, GBDT, and XGBoost models in screening COVID-19.

Model	Accuracy	Precision	Recall	Sensitivity	F1 score	Youden's index
Random forest	74.2%	70.8%	90.8%	53.7%	0.795	0.589
AdaBoost	76.7%	78.2%	80.3%	72.4%	0.792	0.553
GBDT	80.4%	80.3%	85.5%	74.0%	0.828	0.615
XGBoost	75.3%	73.3%	86.8%	61.0%	0.795	0.565

TABLE 4 | Performance of random forest, AdaBoost, GBDT, and XGBoost models to screen COVID-19.

Model	AUC	AUC_95% CI	AUC_SD	AUC_p value	Confusion matrix
Random Forest	85.7%	0.813, 0.902	0.02	<0.001	[66, 57], [14, 138]
AdaBoost	85.4%	0.810, 0.899	0.02	<0.001	[89, 34], [30, 122]
GBDT	86.4%	0.821, 0.907	0.02	<0.001	[91, 32], [22, 130]
XGBoost	84.9%	0.803, 0.894	0.02	<0.001	[75, 48], [20, 132]

**FIGURE 2** | Receiver operating characteristic (ROC) curves for the machine learning models in screening COVID-19.**FIGURE 3** | Calibration curve for the internal validation set. The calibration curve was plotted using the bucket method (continuous data discretization) to observe whether the prediction probability of the classification model is close to the empirical probability (that is, the real probability). Ideally, the calibration curve lies along the diagonal (i.e., the prediction probability is equal to the empirical probability).

and used to develop a simplified decision tree model (Figure 6).

DISCUSSION

The COVID-19 outbreak is currently under control in China and is in a state of normalized prevention and control, but imported cases from other countries occur often, and the number of infections worldwide continues to rise. Virus nucleic acid detection is the “gold standard” for the diagnosis of COVID-19. However, due to premature collection times, nonstandard collection methods, and inaccurate

collection locations, false negative results have occurred many times in virus nucleic acid detection (37). Chest CT plays an important role in the early diagnosis of COVID-19, with a high sensitivity but low specificity (25%) (38). Therefore, developing a new strategy for achieving a rapid and accurate diagnosis for COVID-19 is essential from a clinical perspective.

Since the start of the COVID-19 outbreak, a large number of scholars have been committed to applying AI technology

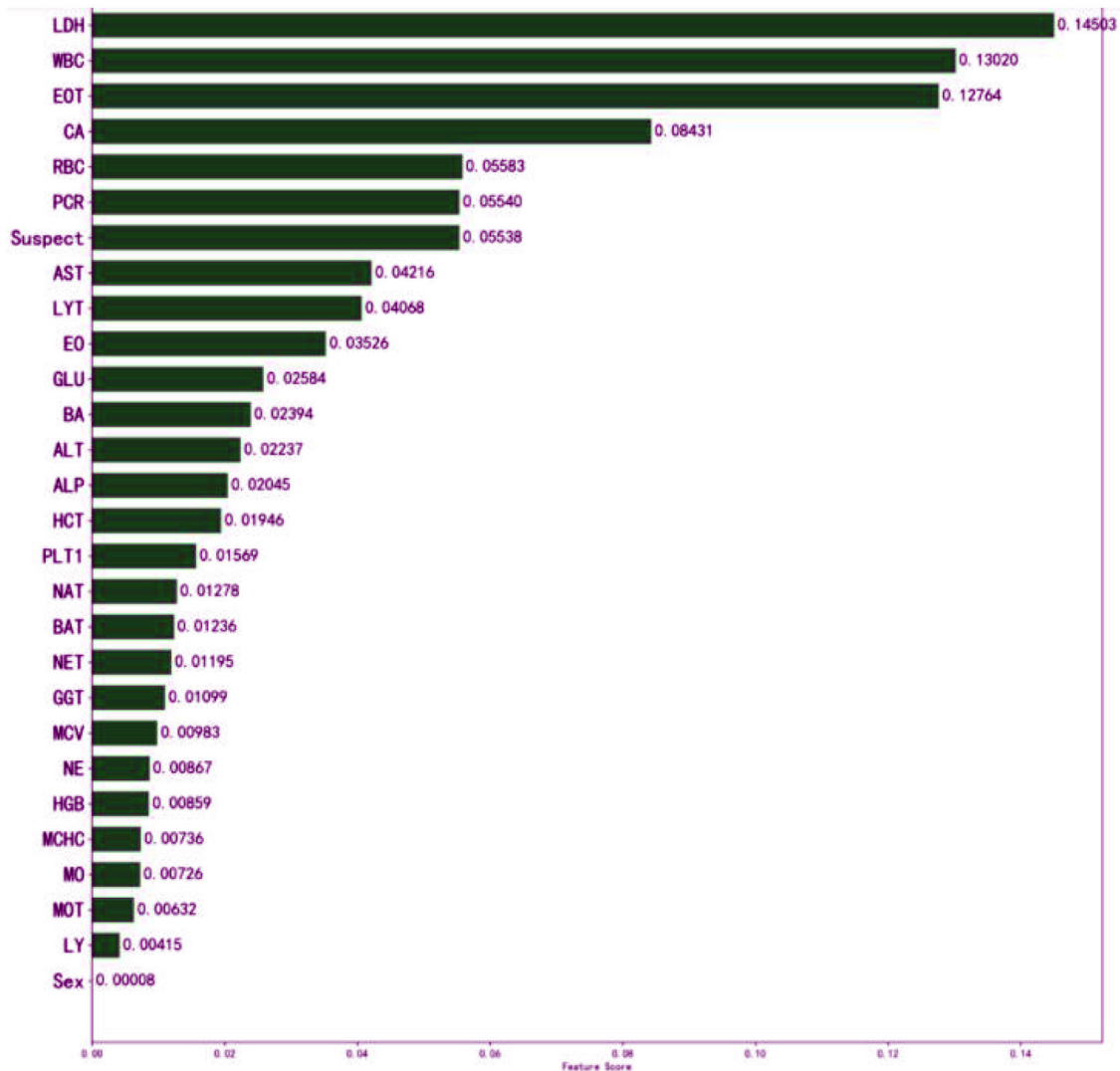
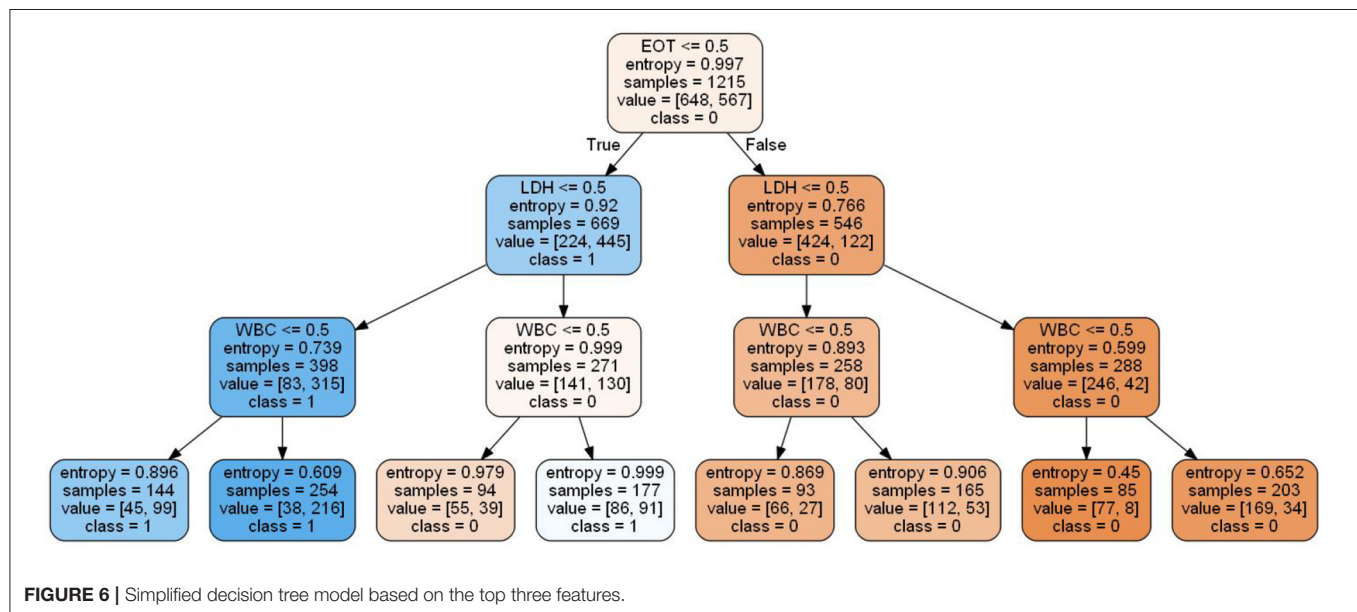
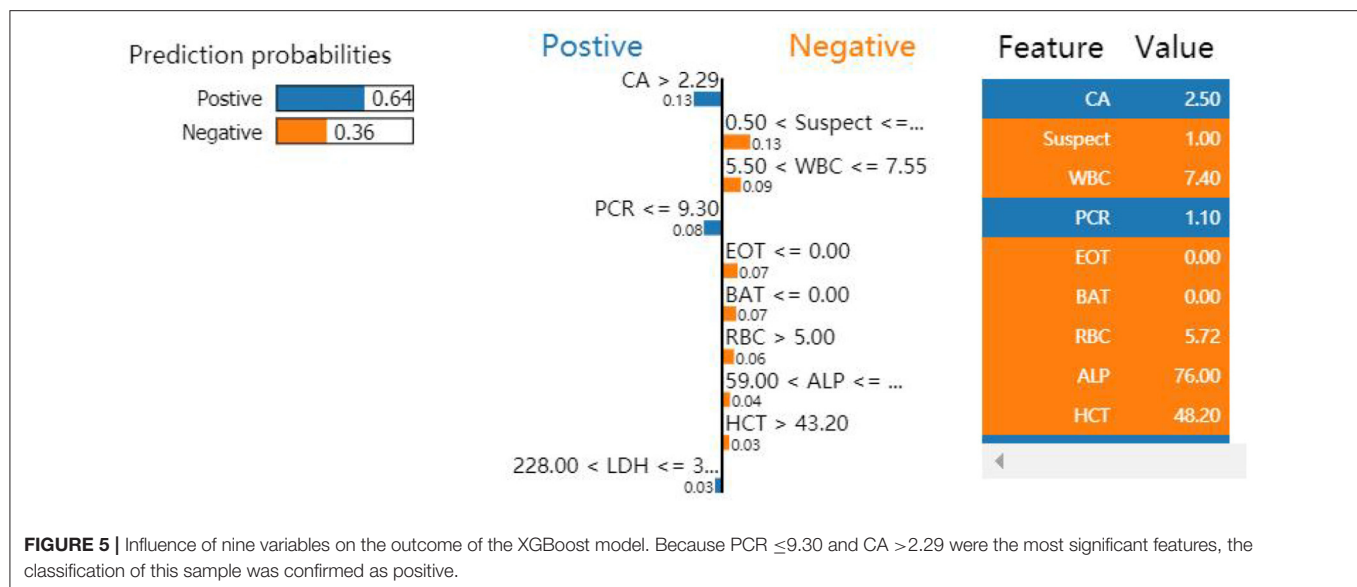


FIGURE 4 | Influence of input features on the outcome of the XGBoost model. The top three features are LDH, WBC, and EOT. It indicates that they have important auxiliary diagnostic significance for COVID-19. The model found that patients with higher WBC count, higher LDH level, or higher EOT count, were more likely to have COVID-19. It might assist physicians to make their decisions.

to rapidly diagnose COVID-19. Wu et al. (39) constructed a COVID-19 differential diagnosis model by mining 11 key blood indices through an ML algorithm and obtained accuracy rates of 0.9795, 0.9697 and 0.9595 with their cross-validation set, test set and external validation set, respectively. Li et al. (40) developed a deep learning model based on CT images to distinguish COVID-19 from community-acquired pneumonia. With the independent validation set, the AUC for identifying COVID-19 was 0.96 and that for identifying community-acquired pneumonia was 0.95. Ozturk et al. (8) constructed a deep learning classification model based on the chest X-ray films of COVID-19 patients. The results showed that the accuracy of the model in performing two-class and multiclass classification were 0.9808 and 0.8702, respectively. All the AI models in the above studies showed good diagnostic performance but only included a single index for

evaluation and analysis and participation in model construction (laboratory examination index or chest image index). Combined with the comprehensive analysis of clinical manifestations, laboratory examination, CT and other indicators, this study jointly constructed a predictive diagnosis model for COVID-19 based on ML that better reflects the real-world COVID-19 situation.

Artificial intelligence technology has an excellent ability to process big data and mine complex medical information. In medical scenarios, the most common problem is binary classification, such as predicting whether a patient has a disease through data analysis and model establishment. Simple models used to solve classification problems include logistic regression, decision tree, and SVM. However, due to the limitations of these simple models, they often cannot achieve optimal prediction



efficiency, so the application of ensemble learning models is becoming more widespread in the machine learning field. AdaBoost was the first boosting model and functions by training different weak models based on the same training dataset and then integrating these weak models to form a stronger classifier with a better effect. XGBoost is a machine learning method focusing on the gradient lifting algorithm. The loss function is expanded as a second-order Taylor expansion, the second derivative of the loss function is used to optimize the loss function, and depending on whether the loss function is reduced, a decision on whether to split nodes is made. The disadvantage of XGBoost is that it is sensitive to outliers.

In GBDT, a tree is trained first by using the training set and the real classification of the samples; then, the tree is used to predict the classification of the training set to obtain the

predicted value of each sample, and the deviation between the predicted value and the true value, that is, the residual, is used as the standard answer to train the next tree. Then, the residual is used to train a third tree, and the final prediction result is obtained. Because the growth process of the decision tree continuously selects and segments features, GBDT composed of a large number of decision trees has inherent advantages and can easily yield the importance ranking of its features. The advantages of the chosen methods over the others are as follows. (1) The prediction accuracy is higher, it is more suitable for low-dimensional data, and it can contend with nonlinear data. (2) It can flexibly handle various types of data, including continuously and discretely valued data. (3) In the case of a relatively short parameter adjustment time, the preparation rate of the prediction can be high relative to that of SVM. (4) Certain robust loss

functions, such as the Huber and quantile loss functions, make the model very robust to outliers.

The model constructed in this study has high clinical application value. The three features identified, LDH, WBC, and EOT, can assist doctors in rapidly and accurately diagnosing COVID-19 patients. Under normal circumstances, LDH is limited to the cytoplasm of tissue cells; it is released only when cell damage and necrosis cause an increase in cell membrane permeability, resulting in a rise in serum LDH concentration. The degree of lung tissue injury is directly proportional to the level of serum LDH, so the level of serum LDH can indirectly reflect the severity of the disease. The sickness is mild when a patient is first infected with SARS-CoV-2. As the disease progresses, the condition gradually worsens, and the LDH level gradually increases (41, 42). The number of white blood cells in a unit volume of blood is measured by the white blood cell count (WBC). White blood cells are an important part of the body's defense system and a common marker for identifying infection, with a high specificity in the diagnosis of infectious fever. According to previous research, infection should still be considered first when the WBC rises. SARS-CoV-2 infection stimulates the innate and adaptive immune responses of the infected body, resulting in a series of inflammatory reactions and pathological changes. The excessive immunological response of the body to external stimuli such as viruses and bacteria is referred to as a cytokine storm (43). It can cause the body to quickly produce a large number of cytokines, such as IL-6, IL-12, IL-8, and IFN- α ; this abnormal increase in the number of cytokines can cause aggregation of eosinophils and other infectious lesions. The organs and tissues are also severely damaged in the process of effectively eradicating the infection (44, 45).

The application of AI technology in the medical field has created new opportunities for solving many medical challenges. However, it can be difficult for users to understand the internal working principle and decision-making process of the model due to its inherent inexplicability. This reduces doctors' trust and acceptance of the AI model and limits the development of AI products in the medical field. Therefore, the construction of interpretable AI models has become the focus of research in recent years. The decision tree model can reflect both linear and nonlinear relationships, allowing it not only to make accurate predictions but also to be interpretable (46). The interpretability of the model is reflected in both global interpretability and local interpretability. The global interpretability shows that the decision tree model can visualize the weight of each index variable, allowing it to assess the value of each index in the prediction model. The higher the index weight value is, the greater the importance of the index. In this study, LDH was the most important index in the construction of the GBDT model, with a weight value of 0.145. Local interpretability explains the diagnosis results for a specific case, which can indicate which indicators support the diagnosis of the disease, which indicators deny the diagnosis of the disease, and the basis for the diagnosis, which is helpful in making an individualized prediction for each patient and providing accurate treatment. To determine whether a patient is infected with COVID-19, the patient is selected from

the validation set and input into the LIME model. The results show that although the CA and PCR2 indicators confirm that the model can diagnose COVID-19 patients, all other indicators deny a diagnosis of COVID-19; the overall tendency, however, is toward a positive diagnosis of COVID-19 for the patient, consistent with the actual patient diagnosis (Figure 5).

In the fight against COVID-19, top international journals have published many research results, including epidemiological and clinical feature analysis, epidemic trend prediction, death-related risk factors, prognostic impact of basic diseases, and critical disease prediction models, which provide important scientific support for this fight and play a positive role in guiding epidemic prevention and control. In a study published in the *Lancet*, a susceptible-exposed-infectious-recovered metapopulation model was used to simulate epidemics across all major cities in China. The study suggested that preparedness plans and mitigation interventions should be readied for quick deployment globally (47). In a study published in *JAMA*, Pan et al. (48) applied surveillance data to quantify the temporal evolution of the intensity of COVID-19 transmission across different periods. Their study may have important implications for ongoing and potential future nonpharmaceutical bundles in the US and other nations with respect to daycare for children (49). Liang et al. (50) developed a clinical risk score to predict the occurrence of critical illness in hospitalized patients. The score may help identify patients with COVID-19 who may subsequently develop a critical illness. Vaid et al. (51) developed machine learning models to predict critical illness and mortality in a cohort of patients in New York City. These models identified at-risk patients and uncovered underlying relationships that predicted patient outcomes. In most studies, a kind of model was applied without considering the ensemble learning algorithms.

This study used a small sample of COVID-19 patients, which may affect the accuracy of the results. Additionally, utilizing a deep learning model with such a small sample size is not ideal. The dataset is not sufficiently standardized, resulting in the elimination of several indicators due to the large number of missing values. In future research, the sample size must be further increased, and a more standardized sample set should be selected to confirm the results of this study.

CONCLUSIONS

In this study, random forest, AdaBoost, GBDT, and XGBoost algorithms were used to develop bagging and boosting ensemble learning models to predict disease risk and then compared in terms of the AUC, accuracy, recall, and F score. Finally, the optimal model was explained by way of the LIME algorithm. Taking the COVID-19 data as a case study, the research is summarized as follows.

First, compared with other classifiers, the precision of GBDT was 80.3%, and the recall was 85.6%. The AUC was 86.4% [95% CI (0.821, 0.907)], indicating better performance. Therefore, GBDT was chosen as the prediction model for the early diagnosis of COVID-19. The model, which was developed based on blood tests, can provide an alternative method to rRT-PCR for the fast

and cost-effective identification of COVID-19-positive patients. It is especially effective in places where outbreaks are on the rise.

Second, the risk factors in the prediction model were visualized using the LIME algorithm. CA, PCR, and LDH were revealed as important factors for identifying patients positive for COVID-19. These findings can help doctors control and treat patients in a timely manner. In addition, the same method can be extended to predict other diseases.

Third, in future studies, multiple features will be fused to enhance the richness and effectiveness of the features. In the ensemble strategy, stacking is a hierarchical model integration framework that will be incorporated into an integration model in future studies. Finally, for classification algorithms, the most popular models were tested. To obtain improved precision in early disease risk identification, combinations of models will be investigated, model complexity will be reduced, and graph neural networks will be integrated in future works.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Methodology, software, validation, and visualization: HG. Data curation: MW. Writing—original draft preparation: HZ, ME, and HG. Writing—review and editing: HZ, ME, and MW. Supervision: MW and MJ. Project administration: HG and MJ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the Natural Sciences Foundation of Hunan Province (Grant No. 2021JJ30139), the National Natural Science Foundation of China (Grant No. 61773157), and the Key Project of R & D plan of Changsha (Grant No. kq2004011).

ACKNOWLEDGMENTS

The authors would like to acknowledge the COVID-19 database, which was the source of the data that supported this study. They would also like to thank Kejia Liu and Yansheng Li (DHC Mediway Technology Co. Ltd., Beijing, China) for their suggestions in conducting the data analysis.

REFERENCES

- Nuzzo Jennifer B, Gostin Lawrence O. COVID-19 and lessons to improve preparedness for the next pandemic-reply. *JAMA*. (2022) 327:1823. doi: 10.1001/jama.2022.4169
- Khan M, Khan H, Khan S. Epidemiological and clinical characteristics of coronavirus disease (COVID-19) cases at a screening clinic during the early outbreak period: a single-centre study. *J Med Microbiol*. (2020) 69:1114–23. doi: 10.1099/jmm.0.001231
- Vogels CBE, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT-qPCR primer-probe sets. *Nat Microbiol*. (2020) 5:1299–305. doi: 10.1038/s41564-020-0761-6
- Rózański M, Walczak-Drzewiecka A, Witaszewska J, Wójcik E, Guziński A, Zimoń B. RT-qPCR-based tests for SARS-CoV-2 detection in pooled saliva samples for massive population screening to monitor epidemics. *Sci Rep*. (2022) 12:8082. doi: 10.1038/s41598-022-12179-4
- Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*. (2020) 369:m1328. doi: 10.1136/bmj.m1328
- Li L, Qin L, Xu Z, Yin Y, Wang X, Wang B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology*. (2020) 296:E65–71. doi: 10.1148/radiol.20200905
- Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. *arXiv*. <http://arxiv.org/abs/2003.05037>
- Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med*. (2020) 121:103792. doi: 10.1016/j.combiomed.2020.103792
- Mei X, Lee HC, Diao K, Huang M, Lin B, Liu C. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med*. (2020) 26:1224–8. doi: 10.1038/s41591-020-0931-3
- Weinstock MB, Echenique A, Russell JW, Leib A, Miller J, Cohen DJ. Chest X-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest X-ray is no guarantee. *J Urgent Care Med*. (2020) 10:13–8. Available online at: <https://www.jucm.com/chest-x-ray-findings-in-636-ambulatory-patients-with-covid-19-presenting-to-an-urgent-care-center-a-normal-chest-x-ray-is-no-guarantee/>
- Ferrari D, Motta A, Strollo M, Banfi G, Locatelli M. Routine blood tests as a potential diagnostic tool for COVID-19. *Clin Chem Lab Med*. (2020) 58:1095–9. doi: 10.1515/cclm-2020-0398
- Barbosa VAF, Gomes JC, Santana D, de Lima MA, Calado CL, Bertoldo RB Jr. et al. Covid-19 rapid test by combining a Random Forest-based web system and blood tests. *J Biomol Struct Dyn*. (2021) 2021:1–20. doi: 10.1080/07391102.2021.1966509
- Barbosa VAF, Gomes JC, Santana D, Albuquerque MA, de Souza JEDA, de Souza RGRE et al. HegIA: an intelligent system to support diagnosis of Covid-19 based on blood tests. *Res Biomed Eng*. (2022) 38:99–116. doi: 10.1007/s42600-020-00112-5
- Szklanna PB, Altaie H, Comer SP, Cullivan S, Kelliher S, Weiss L, et al. Routine hematological parameters may be predictors of COVID-19 severity. *Front Med*. (2021) 8:682843. doi: 10.3389/fmed.2021.682843
- Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. *J Med Syst*. (2020) 44:1–12. doi: 10.1007/s10916-020-01597-4
- Bonfiglio R, Pietro Di, The impact ML of oral contraceptive use on breast cancer risk: state of the art and future perspectives in the era of 4P medicine. *Semin Cancer Biol*. (2021) 72:11–8. doi: 10.1016/j.semcancer.2020.10.008
- Lindholm D, Lindbäck J, Armstrong PW, Budaj A, Cannon CP, Granger CB, et al. Biomarker-based risk model to predict cardiovascular mortality

- in patients with stable coronary disease. *J Am Coll Cardiol.* (2017) 70:813–26. doi: 10.1016/j.jacc.2017.06.030
18. Than MP, Pickering JW, Sandoval Y, Shah ASV, Tsanas A, Apple FS, et al. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation.* (2019) 140:899–909. doi: 10.1161/CIRCULATIONAHA.119.041980
 19. Zhou ZH, Washio T. *Advances in Machine Learning.* Berlin, Heidelberg: Springer Berlin Heidelberg (2009). doi: 10.1007/978-3-642-05224-8
 20. Sutton R, Barto A. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press (1998). doi: 10.1109/TNN.1998.712192
 21. Watkins C, Dayan P. Q-learning. *Mach Learn.* (1992) 8:279–92. doi: 10.1007/BF00992698
 22. Wong D, Yip S. Machine learning classifies cancer. *Nature.* (2018) 555:446–7. doi: 10.1038/d41586-018-02881-7
 23. Zhou ZHE. *Foundations and Algorithms.* Abingdon, VA: Taylor & Francis. (2012).
 24. Hutson M. Has artificial intelligence become alchemy. *Science.* (2018) 360:478–478. doi: 10.1126/science.360.6388.478
 25. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept “black box” medicine? *Ann Intern Med.* (2020) 172:59–60.
 26. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—explainable artificial intelligence. *Sci Rob.* (2019) 4:eaay7120. doi: 10.1126/scirobotics.aay7120
 27. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: explaining the predictions of any classifier. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco, CA: ACM. (2016), p. 1135–44. doi: 10.1145/2939672.2939778
 28. Lundberg SM, Lee S-I. *A Unified Approach to Interpreting Model Predictions.* Long Beach, CA: NIPS (2017). p. 4768–77.
 29. Kupferschmidt K. WHO relaunched global drug trial with three new candidates. *Science.* (2021) 373:606–7. doi: 10.1126/science.373.6555.606
 30. Woloshin S, Patel N, Kesselheim AS. False negative tests for SARS-CoV-2 infection — challenges and implications. *N Engl J Med.* (2020) 383:e38. doi: 10.1056/NEJMp2015897
 31. Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clin Chem Lab Med.* (2020) 59:421–31. doi: 10.1515/cclm-2020-1294
 32. Chen T, Tong H, Benesty M. *xgboost: Extreme Gradient Boosting.* San Francisco, CA (2016).
 33. Kearns MJ, Umesh V. *Vazirani An Introduction to Computational Learning Theory.* Cambridge, MA: MIT Press (1994).
 34. Kearns M, Valiant L. Cryptographic limitations on learning Boolean formulae and finite automata. *Symposium on Theory of Computing.* 21. Seattle, WA: ACM (1989). p. 433–444. doi: 10.1145/73007.73049
 35. Breiman, L. Bagging prediction. *Mach Learn.* (1996) 24:123–40. doi: 10.1007/BF00058655
 36. Schapire RE, Freund Y. *Boosting: Foundations and Algorithms.* Cambridge, MA: MIT Press (2013). doi: 10.1108/03684921311295547
 37. Wang CB. Analysis of low positive rate of nucleic acid detection method used for diagnosis of novel coronavirus pneumonia. *Zhonghua Yi Xue Za Zhi.* (2020) 100:961–4. doi: 10.3760/cma.j.cn112137-20200213-00280
 38. Tao Ai, Yang Zhenlu, Hou Hongyan, Zhan Chenao, Chen Chong, Lv Wenzhi, et al. Correlation of chest CT and RTPCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology.* (2020) 296:E32–40. doi: 10.1148/radiol.2020200642
 39. Wu J, Zhang P, Zhang L, Meng W, Li J, Tong C. Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. *medRxiv.* (2020). doi: 10.1101/2020.04.02.20051136
 40. Li Z, Zhong Z, Li Y, Zhang T, Gao L, Jin D. From community-acquired pneumonia to COVID-19: a deep learning-based method for quantitative analysis of COVID-19 on thick-section CT scans. *Eur Radiol.* (2020) 30:6828–37. doi: 10.1007/s00330-020-07042-x
 41. Fattizzo B, Pasquale R, Bellani V, Barcellini W, Kulasekararaj AG. Complement mediated hemolytic anemias in the COVID-19 era: case series and review of the literature. *Front Immunol.* (2021) 12:791429. doi: 10.3389/fimmu.2021.791429
 42. Shcherbak SG, Anisenkova AY, Mosenko SV, Glotov OS, Chernov AN, Apalko SV, et al. Basic predictive risk factors for cytokine storms in COVID-19 Patients. *Front Immunol.* (2021) 12:745515. doi: 10.3389/fimmu.2021.745515
 43. de Oliveira Costa R, Nascimento JS, Reichert CO, da Costa APA, Dos Santos MAP, Soares AM, et al. “H” is not for hydroxychloroquine. “H” is for heparin: lack of efficacy of hydroxychloroquine and the role of heparin in COVID-19-preliminary data of a prospective and interventional study from Brazil. *BMC Infect Dis.* (2022) 22:120. doi: 10.1186/s12879-022-07110-1
 44. Lorenza L, Carter D. Emergency online teaching during COVID-19: a case study of Australian tertiary students in teacher education and creative arts. *Int J Educ Res Open.* (2021) 2:100057. doi: 10.1016/j.ijedro.2021.100057
 45. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug Alcohol Depend.* (2021) 227:108986. doi: 10.1016/j.drugalcdep.2021.108986
 46. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: a unified framework for machine learning interpretability. *arXiv.* (2019) doi: 10.48550/arXiv.1909.09223
 47. Joseph T, Wu, Kathy Leung, Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet.* (2020). 395:689–97. doi: 10.1016/S0140-6736(20)30260-9
 48. Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q. Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China. *JAMA.* (2020) 323:1915–23. doi: 10.1001/jama.2020.6130
 49. Hartley DM, Perencevich EN. Public Health Interventions for COVID-19: emerging evidence and implications for an evolving public health crisis. *JAMA.* (2020) 323:1908–9. doi: 10.1001/jama.2020.5910
 50. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med.* (2020) 180:1081–9. doi: 10.1001/jamainternmed.2020.2033
 51. Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, et al. (2020). Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. *J Medi Internet Res.* 22, e24018. doi: 10.2196/24018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gong, Wang, Zhang, Elahe and Jin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Machine Learning Algorithm for Predicting the Risk of Developing to M1b Stage of Patients With Germ Cell Testicular Cancer

Li Ding¹, Kun Wang¹, Chi Zhang¹, Yang Zhang¹, Kanlirong Wang², Wang Li^{1*} and Junqi Wang^{1*}

OPEN ACCESS

Edited by:

Yu-Hsiu Lin,
National Chung Cheng
University, Taiwan

Reviewed by:

Wenle Li,
Xian Yang Central Hospital, China
Jaya Lakshmi Thangaraj,
University of California, San Diego,
United States
Narit Hnoohom,
Mahidol University, Thailand
Donghui Yan,
University of Massachusetts
Dartmouth, United States

*Correspondence:

Wang Li
lizhixin88mm@163.com
Junqi Wang
wjq68@sina.cn

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 09 April 2022

Accepted: 06 June 2022

Published: 29 June 2022

Citation:

Ding L, Wang K, Zhang C, Zhang Y,
Wang K, Li W and Wang J (2022) A
Machine Learning Algorithm for
Predicting the Risk of Developing to
M1b Stage of Patients With Germ Cell
Testicular Cancer.
Front. Public Health 10:916513.
doi: 10.3389/fpubh.2022.916513

¹ Department of Urology, the Affiliated Hospital of Xuzhou Medical University, Xuzhou, China, ² Nanjing First Hospital, Nanjing, China

Objective: Distant metastasis other than non-regional lymph nodes and lung (i.e., M1b stage) significantly contributes to the poor survival prognosis of patients with germ cell testicular cancer (GCTC). The aim of this study was to develop a machine learning (ML) algorithm model to predict the risk of patients with GCTC developing the M1b stage, which can be used to assist in early intervention of patients.

Methods: The clinical and pathological data of patients with GCTC were obtained from the Surveillance, Epidemiology, and End Results (SEER) database. Combining the patient's characteristic variables, we applied six machine learning (ML) algorithms to develop the predictive models, including logistic regression(LR), eXtreme Gradient Boosting (XGBoost), light Gradient Boosting Machine (lightGBM), random forest (RF), multilayer perceptron (MLP), and k-nearest neighbor (kNN). Model performances were evaluated by 10-fold cross-receiver operating characteristic (ROC) curves, which calculated the area under the curve (AUC) of models for predictive accuracy. A total of 54 patients from our own center (October 2006 to June 2021) were collected as the external validation cohort.

Results: A total of 4,323 patients eligible for inclusion were screened for enrollment from the SEER database, of which 178 (4.12%) developing M1b stage. Multivariate logistic regression showed that lymph node dissection (LND), T stage, N stage, lung metastases, and distant lymph node metastases were the independent predictors of developing M1b stage risk. The models based on both the XGBoost and RF algorithms showed stable and efficient prediction performance in the training and external validation groups.

Conclusion: S-stage is not an independent factor for predicting the risk of developing the M1b stage of patients with GCTC. The ML models based on both XGBoost and RF algorithms have high predictive effectiveness and may be used to predict the risk of developing the M1b stage of patients with GCTC, which is of promising value in clinical decision-making. Models still need to be tested with a larger sample of real-world data.

Keywords: machine learning algorithms, prediction model, germ cell testicular cancer, M1b stage, real-world research

INTRODUCTION

Testicular cancer (TC), as a rare malignant tumor of the genitourinary system, accounts for about 1% of male tumors and about 5% of urogenital tumors. In Occident, the annual rate of new cases is <1 in 10,000 (1). Despite having a relatively low overall incidence rate and a good prognosis, TC is the most common malignancy in men aged 15 to 35 years (2, 3). Germ cell testicular cancer (GCTC) is the most common kind of testicular cancer, accounting for over 95% of all testicular cancer histological types. There are two types of GCTC: seminoma and non-seminomatous germ cell tumors (NSGCTs). The former is the most common type of GCTC, accounting for about one-third of its total, and the latter includes embryonal carcinomas, yolk sac tumors, choriocarcinomas, teratomas, and mixed germ cell cancers (4). Cryptorchidism, family history, Klinefelter's syndrome, androgen insensitivity syndrome (AIS), and industrial exposure may be the major risk factors for testicular cancer (5–8). Serum levels of alphafetoprotein (AFP), human chorionic gonadotropin (hCG), and lactate dehydrogenase (LDH) should be determined before and after orchiectomy, as they can assist in diagnosis and predict prognosis. Genetic studies have shown that TC is associated with ectopic short arms of chromosome 12 (i12p) and that alterations in the P53 gene have a correlation with their occurrence (1, 9). Radical orchiectomy, together with bilateral retroperitoneal lymph node dissection, is the standard surgical management of patients with TC, and radiotherapy and/or chemotherapy is recommended for patients with advanced TC (10, 11).

Germ cell testicular cancer outward invasion includes lymph nodes, lungs, liver, brain, bones, etc. Although distant metastases are more likely to invade the lungs and distant lymph nodes for GCTC, the risk of other atypical metastases (including liver, brain, bones, and other rare organs or tissues), which account for approximately 10% of all patients, cannot be ignored (12–16). The International Germ Cell Cancer Collaborative Classification for Metastatic Testicular Cancer (IGCCCG) identifies non-pulmonary visceral metastases as a strong influence on poor prognosis in metastatic patients with TC (15). A recent study also showed that patients with liver metastases and bone metastases had a significantly poor prognosis compared to distant lymph node and lung metastases (13). Although most metastatic lesions are not palpable, if a patient has supraclavicular lymph node metastases, they may palpate a left cervical mass. Lung metastases may present with the shortness of breath or even rare hemoptysis. If a patient has extensive retroperitoneal metastases, they may present with low back pain due to organ compression. Meanwhile, brain metastases may cause headaches as well as various neurological symptoms (17). Contrast-enhanced computerized tomography (CECT) is the most sensitive method to evaluate patients with TC for tumor invasion in the chest, abdomen, and pelvis (18, 19). Although both CECT and magnetic resonance imaging (MRI) are the key image modalities for detecting brain metastases, MRI is much more sensitive than CECT, and therefore, MRI plays a major role in detecting brain metastases (20). However, imaging scans may not be effective enough in screening out patients with GCTC at high risk for

developing to M1b stage. Therefore, a model to predict the risk of progression to M1b in patients with GCTC can be used for clinical applications to improve patient prognosis.

Machine learning (ML) is an advanced algorithmic model that automatically learns and improves performance by identifying complex non-linear relationships in different patterns and is considered superior to traditional algorithms (21–23). As one of the topics of artificial intelligence (AI), ML has been widely used in clinical practice, such as image recognition, complications prediction, and survival analysis (24, 25). The aim of this study was to establish and validate an ML-based model predicting the risk of progression to the M1b stage in patients with GCTC.

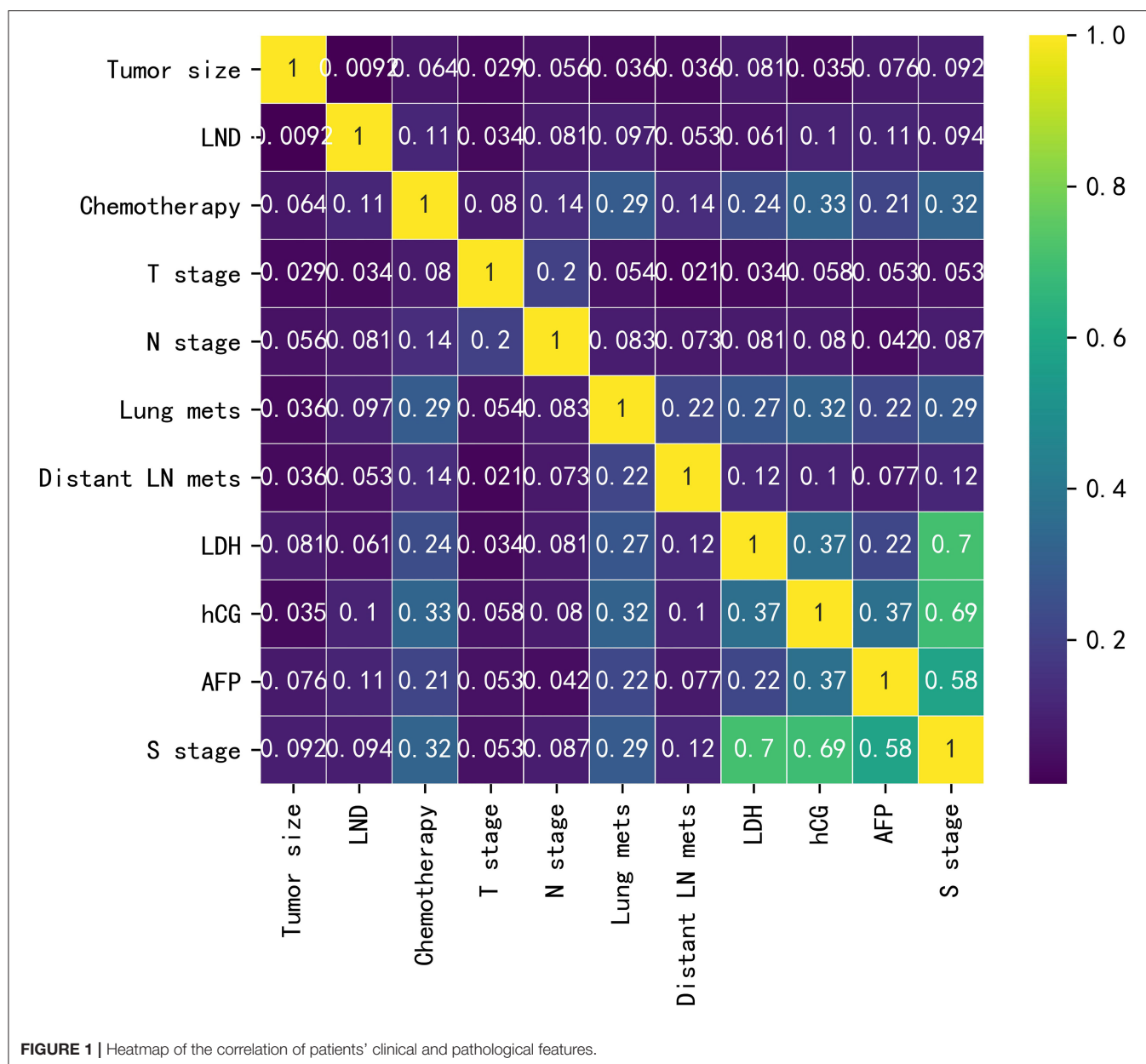
MATERIALS AND METHODS

Data Collection

A retrospective cohort research approach was adopted. The information came from the SEER research database, which covers approximately 27.8% of the US population. We used I CD-O-3 site codes C62.1 and C62.9 and histological codes 9061 to 9064, 9070 to 9071, 9080 to 9085, and 9100 to 9102 to identify patients with GCTC. To develop the ideal ML model, several variables were obtained, including survival data, age, race, marital status at diagnosis, histology type, TNM stage, tumor laterality, radiotherapy documents, chemotherapy documents, LND, lymph-vascular invasion (LVI), metastatic sites, and AFP/hCG/LDH index after orchiectomy. We evaluated the S-stage of patients based on the postoperative serum tumor marker data obtained above. An external validation set was constructed by collecting the same variables from the Affiliated Hospital of Xuzhou Medical University. The flow chart for patient selection of the SEER database is shown in **Supplementary Figure 1**.

Statistical Analysis

For continuous variables, the Student's *t*-test was used for normally distributed data and the Mann-Whitney *U*-test for non-normally distributed data. The chi-square test was used to analyze categorical data. The Kaplan-Meier method was being used to determine the clinical endpoints of the patients, and the log-rank test was used to analyze them. Uni- and multivariate logistic regression analyses were used to calculate the odds ratio (OR) with 95% confidence intervals (Cis). Only two-sided *p*-value < 0.05 was considered statistical significance. We used six different ML algorithms to analyze our data: LR, XGBoost, lightGBM, RF, MLP, and kNN. The model with the highest average AUC was chosen as the best algorithm. Furthermore, the ML-based model was tuned to avoid overfitting, and the accuracy of the algorithm was tested using the 10-fold cross-validation method. R 4.1.2 (<https://www.r-project.org/>), Python 3.10 (<https://www.python.org/>), and SEER*Stat (<https://seer.cancer.gov/seers tat/>) were used in this study. Detailed packages used in the development of our ML models including XGBoost 1.2.1, lightGBM 3.2.1, and sklearn 0.22.1. For the kNN classifier, the number of neighbors is set as 3. For the RF algorithm, we set the "ntree" as 100 and "mtree" as 3. To avoid overfitting and enhance interpretability, the maximum tree depth was set to 8



nodes in the XGBoost algorithm. The hidden layer sizes of MLP algorithm were (10, 10).

RESULTS

Patient's Characteristics

Baseline data for the training cohort and external validation cohort are listed in **Supplementary Table 1**. In the training cohort, the variables with $p < 0.05$ were LND, chemotherapy, T-stage, N-stage, lung metastasis, distant lymph node metastasis, LDH, hCG, AFP, and S-stage. The differences were not statistically significant in age, tumor size, race, histology type, laterality, marital status, radiotherapy, and LVI. The correlations between the variables chosen as predictors were analyzed and

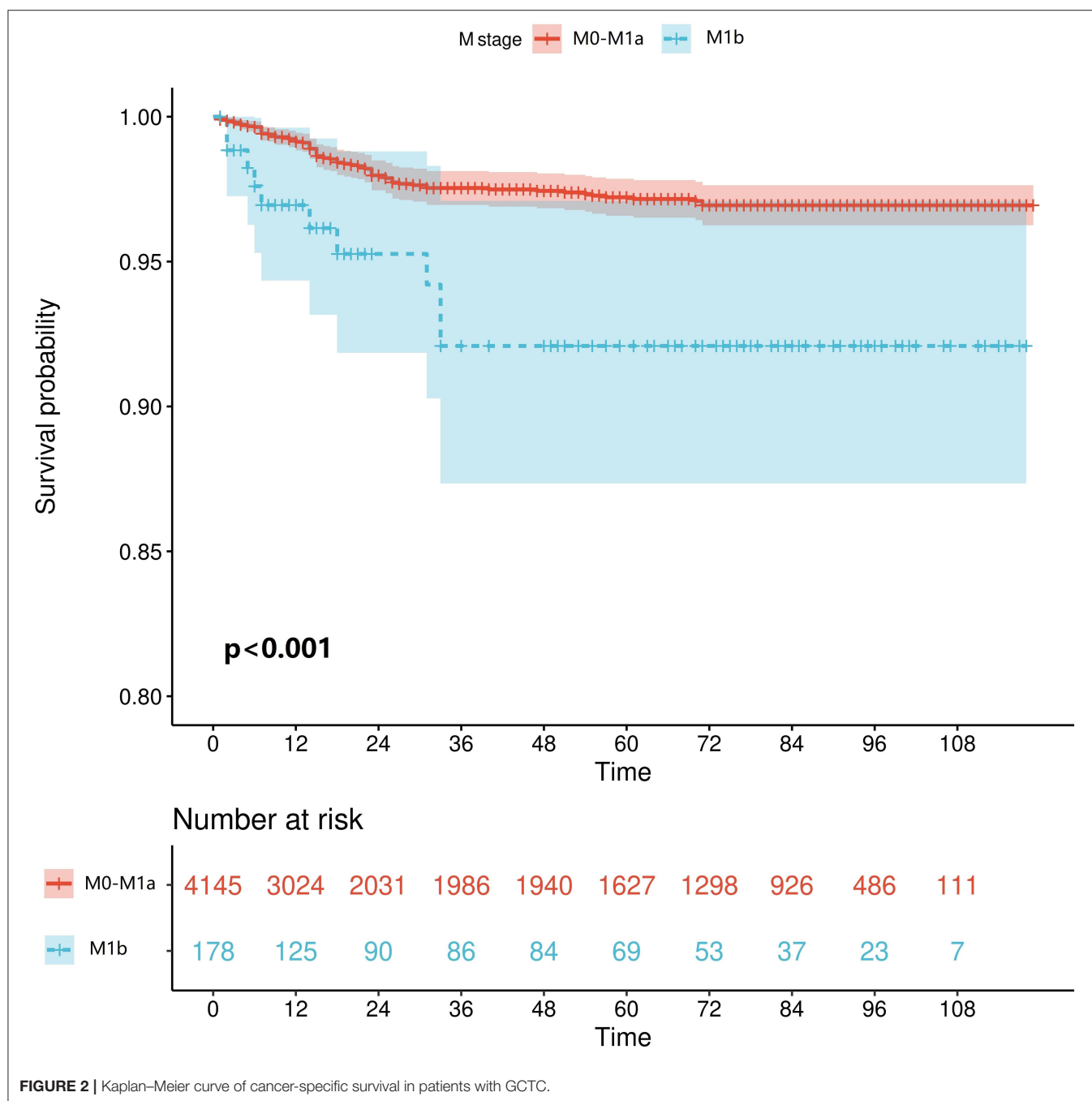
visualized by a heatmap using Spearman's rank correlation coefficient (**Figure 1**).

Survival Analysis

We retrieved patients' survival data from the SEER database, cancer-specific survival (CSS) was considered as the endpoint, and Kaplan–Meier survival analysis was used to estimate the survival. As shown in **Figure 2**, patients who reached the M1b stage had significantly worse CSS ($p < 0.001$).

Univariate and Multivariate Logistic Regression Analyses

As illustrated in **Table 1**, in terms of univariate logistic regression analysis, LND, chemotherapy, T-stage, N-stage, lung metastasis,



distant lymph node metastasis, LDH, hCG, AFP, and S-stage were all significantly associated with the occurrence of developing M1b stage in the overall population ($p < 0.05$). In multivariable logistic regression analysis (Table 2), given the high correlation between serum tumor markers and S-stage as shown by heatmap, two models were carried out to avoid collinearity. Factors with statistical significance were T-stage, N-stage, lung metastasis, and distant lymph node metastasis ($p < 0.001$) in both model 1 (included S-stage) and model 2 (included three serum tumor markers). The p -value of LND was 0.056 in model 1 and 0.049

in model 2. After comprehensively considering the performance of this variable in the two models, we finally incorporated it into the model algorithm of ML.

Performance of ML Algorithms

To compare the predictive efficiency of six ML algorithm models, 10-fold cross-validation was applied in this study (Figure 3). Both the XGBoost model (AUC = 0.814, 95% CI 0.777–0.851) and the RF model (AUC = 0.816, 95% CI 0.779–0.852) performed well in the training cohort. The learning curves of models

TABLE 1 | Univariable logistic regression analysis of the training cohort.

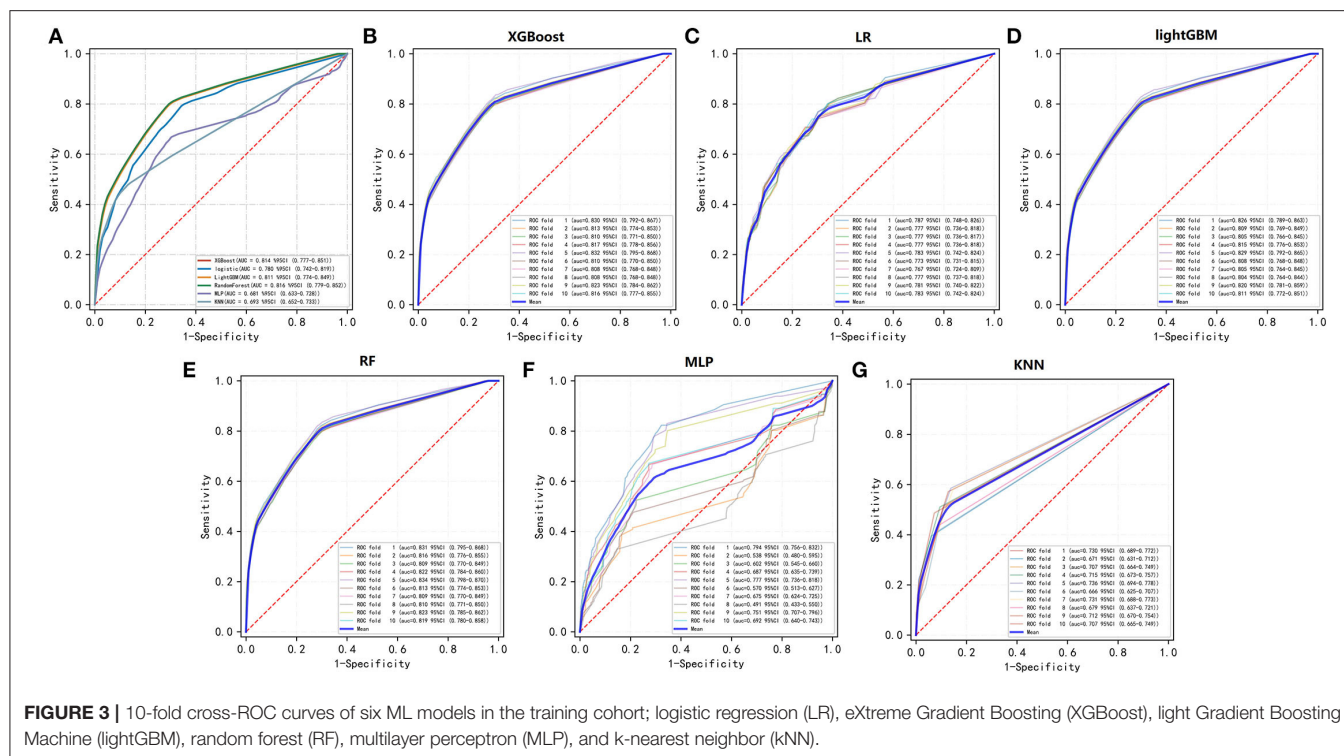
Variables	Level	Univariate OR	95%CI	p-value
Age (year)	NA	1.006	[0.993, 1.019]	0.367
Tumor size (mm)	NA	1.002	[0.999, 1.005]	0.113
Race	White	Ref		0.602
	Black	0.672	[0.211, 2.141]	0.501
	Other	1.191	[0.739, 1.919]	0.473
Histology type	Seminoma	Ref		0.139
	NGSTC	1.257	[0.928, 1.701]	
Laterality	Left	Ref		0.83
	Right	1.033	[0.765, 1.396]	
Marital status	Single	Ref		0.505
	Married	1.205	[0.881, 1.648]	0.242
	Other status	1.08	[0.596, 1.957]	0.799
LND	No/Biopsy only	Ref		<0.001
	Yes	2.309	[1.592, 3.349]	
Radiotherapy	No	Ref		0.984
	Yes	0.993	[0.501, 1.969]	
Chemotherapy	No	Ref		<0.001
	Yes	2.571	[1.854, 3.566]	
LVI	Absent	Ref		0.643
	Present	0.926	[0.668, 1.283]	
T stage	T1	Ref		<0.001
	T2	1.379	[0.973, 1.955]	0.071
	T3	6.214	[4.118, 9.377]	<0.001
	T4	10.848	[3.425, 34.362]	<0.001
N stage	N0	Ref		<0.001
	N1	5.214	[3.485, 7.801]	<0.001
	N2	4.166	[2.622, 6.620]	<0.001
	N3	9.431	[6.300, 14.119]	<0.001
Lung metastasis	No	Ref		<0.001
	Yes	4.648	[3.264, 6.620]	
Distant lymph node metastasis	No	Ref		<0.001
	Yes	9.593	[5.674, 16.218]	
LDH (U/l)	Within normal limits	Ref		0.002
	<1.5 x N	1.5	[1.008, 2.233]	0.045
	1.5–10 x N	2.109	[1.315, 3.383]	0.002
	>10 x N	2.822	[1.268, 6.283]	0.011
	Only know elevated after orchiectomy	0.914	[0.285, 2.931]	0.88
hCG (mIU/ml)	Within normal limits	Ref		<0.001
	<5,000	1.44	[0.967, 2.144]	0.072
	5,000–50,000	2.765	[1.307, 5.849]	0.008
	5,000–50,000	4.814	[2.400, 9.657]	<0.001
	Only know elevated after orchiectomy	1.926	[0.589, 6.297]	0.278
AFP (ng/ml)	Within normal limits	Ref		0.011
	<1,000	1.07	[0.714, 1.603]	0.742
	1,000–9,999	2.88	[1.546, 5.367]	0.001
	≤ 10,000	1.374	[0.327, 5.764]	0.664
S-stage	S0	Ref		<0.001
	S1	1.143	[0.756, 1.729]	0.527
	S2	1.607	[1.104, 2.338]	0.013
	S3	3.262	[1.889, 5.631]	<0.001

OR, odds ratio; CIs, confidence intervals; NSGCT, non-seminomatous germ cell tumor; LND, lymph node dissection; LVI, lymph-vascular invasion; LDH, lactate dehydrogenase; hCG, human chorionic gonadotropin; AFP, alpha-fetoprotein; other marital status includes divorced/widowed/unknown; N indicates the upper limit of normal; serum tumor markers were determined after orchiectomy/before chemotherapy.

TABLE 2 | Multivariate logistic regression analysis of the training cohort.

Variables	Level	Model 1			Model 2		
		Multivariate OR	95%CI	p-value	Multivariate OR	95%CI	p-value
LND	No/Biopsy only	Ref		0.056			0.049
	Yes	1.492	[0.989, 2.250]		1.517	[1.002, 2.295]	
Chemotherapy	No	Ref		0.085			0.117
	Yes	1.397	[0.955, 2.044]		1.358	[0.926, 1.991]	
T stage	T1	Ref		<0.001			<0.001
	T2	1.053	[0.728, 1.523]		1.072	[0.74, 1.554]	
	T3	3.216	[2.054, 5.035]		3.259	[2.074, 5.121]	
	T4	5.6	[1.643, 19.090]		5.079	[1.436, 17.965]	
N stage	N0	Ref		<0.001			<0.001
	N1	4.201	[2.756, 6.404]		4.291	[2.808, 6.559]	
	N2	3.159	[1.945, 5.129]		3.288	[2.019, 5.354]	
	N3	6.148	[3.159, 1.945]		6.416	[4.138, 9.947]	
Lung metastasis	No	Ref		<0.001			0.001
	Yes	2.396	[1.538, 3.734]		2.254	[1.406, 3.613]	
Distant lymph node metastasis	No	Ref		<0.001			<0.001
	Yes	4.288	[2.335, 7.877]		4.588	[2.494, 8.441]	
LDH (U/l)	Within normal limits	/	/	/			0.697
	<1.5 x N	/	/	/	1.014	[0.644, 1.599]	
	1.5–10 x N	/	/	/	0.735	[0.404, 1.339]	
	>10 x N	/	/	/	0.976	[0.376, 2.532]	
	Only know elevated after orchiectomy	/	/	/	0.495	[0.142, 1.721]	
hCG (mIU/ml)	Within normal limits	/	/	/			0.177
	<5,000	/	/	/	1.021	[0.634, 1.645]	
	5,000–50,000	/	/	/	1.368	[0.553, 3.382]	
	5,000–50,000	/	/	/	2.873	[1.196, 6.901]	
	Only know elevated after orchiectomy	/	/	/	1.57	[0.434, 5.689]	
AFP (ng/ml)	Within normal limits	/	/	/			0.396
	<1,000	/	/	/	0.703	[0.442, 1.116]	
	1,000–9,999	/	/	/	1.143	[0.544, 2.403]	
	≤10,000	/	/	/	0.611	[0.123, 3.029]	
S-stage	S0	Ref		0.397	/	/	/
	S1	0.834	[0.534, 1.302]		/	/	/
	S2	0.791	[0.512, 1.221]		/	/	/
	S3	1.299	[0.678, 2.489]		/	/	/

OR, odds ratio; Cis, confidence intervals; LND, lymph node dissection; LVI, lymph-vascular invasion; LDH, lactate dehydrogenase; hCG, human chorionic gonadotropin; AFP, alpha-fetoprotein; N indicates the upper limit of normal; serum tumor markers were determined after orchiectomy/before chemotherapy.



in the training cohort are shown in **Supplementary Figure 2**. In external validation, as shown in **Figure 4**, the XGBoost model (AUC = 0.957, 95% CI 0.904–1.000) showed the best performance in ROC curve analysis among six algorithms, and the RF model also showed great performance (AUC = 0.946, 95% CI 0.886–1.000). Since both the XGBoost model and the RF model were efficient and stable in the training and validation groups, we suggested that both the two algorithmic models can be considered as ideal for predicting the risk of developing M1b stage with patients with GCTC.

Relative Importance of Variables

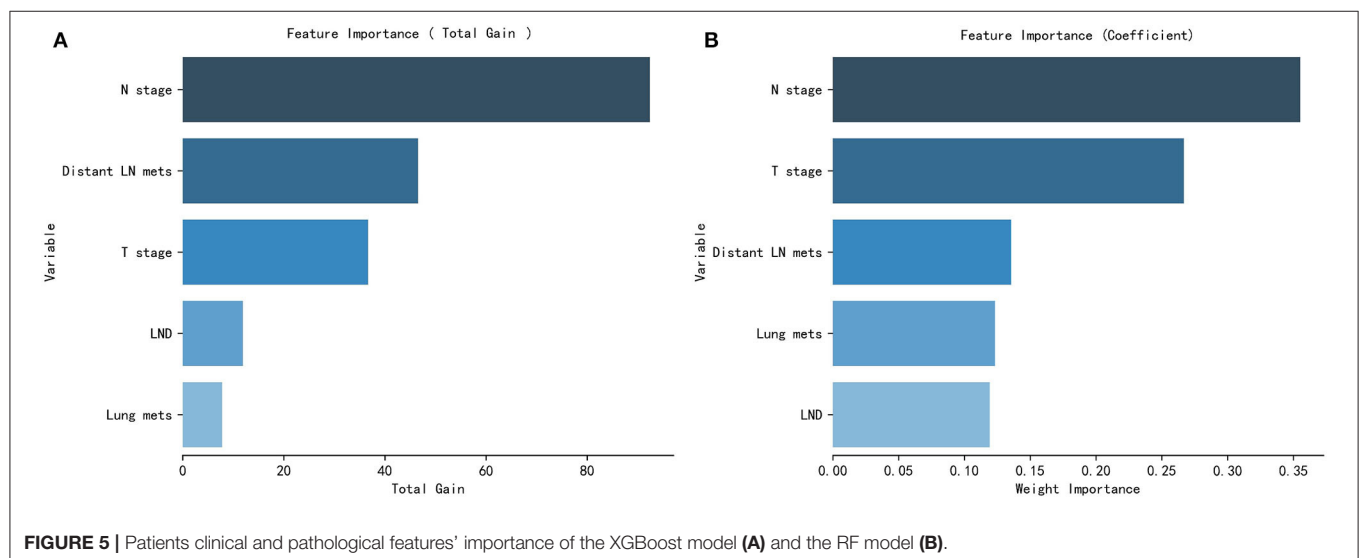
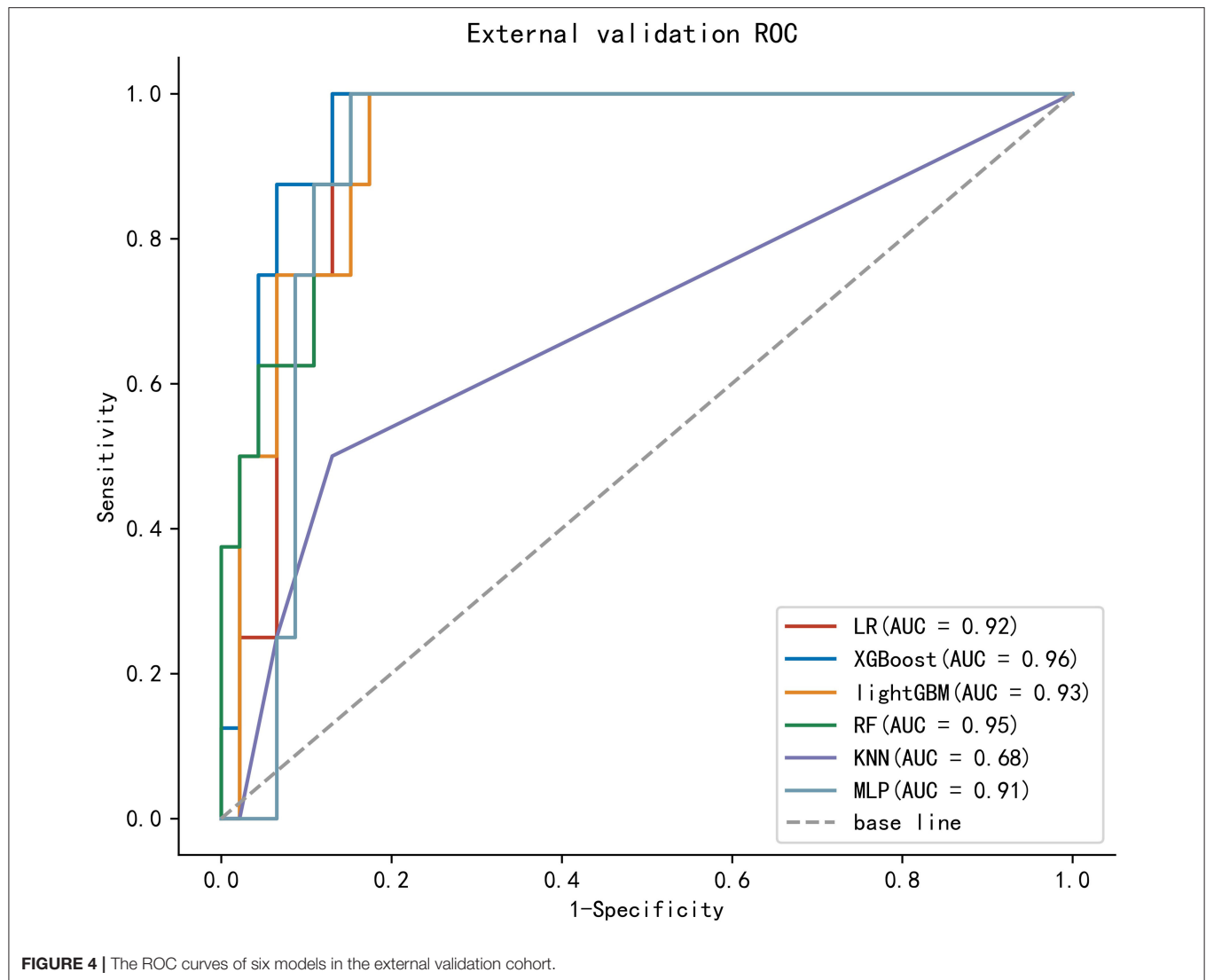
The GCTC patients' clinical feature importance based on the XGBoost and the RF model is shown in **Figure 5**.

DISCUSSION

For patients with undetectable metastatic lesions, early application of systemic chemotherapy and combination therapy may improve the prognosis and increase the median survival rate (26). The IGCCCG-related metastatic germ cell testicular cancer prognostic-based staging system (15) is clinically recognized as an effective system. This system showed that for patients with TC who developed metastases, the prognosis for pulmonary metastases was better, whereas patients with non-pulmonary metastases had a poorer prognosis. A recent study also showed that patients with TC who developed organ metastases, such as bone and liver, had over all poor survival and cancer-specific survival (13). Some patients fail to detect metastatic lesions at the first diagnosis or even at the early postoperative review. Some

patients with early metastatic GCTC (mGCTC) have subclinical metastases (most common in the retroperitoneum) that are not identified by imaging and are identified and diagnosed as clinical M1 at follow-up after orchiectomy (14, 27). The S-stage is a classification based on serum tumor markers (post-orchiectomy and pre-chemotherapy initiation) and is complementary to the TNM stage. Since the serum half-lives of AFP and β -hCG are 5 to 7 days and 1 to 3 days, respectively, it will take several weeks to return to normal levels (28, 29). These tumor markers not only have prognostic predictive value, but also should be continued during follow-up to assist in determining whether postoperative metastases have occurred (30). The BEP-based (bleomycin, etoposide, and cisplatin) chemotherapy regimen is the standard treatment for metastatic patients with TC (31). A randomized phase III trial showed similar relapse-free survival rates and no significant difference in quality of survival between patients who underwent retroperitoneal lymph node dissection and adjuvant BEP (32). Most patients with GCTC are sensitive to radiotherapy as well (33).

Previous studies have shown that patients with metastases to internal organs other than the lungs have a significantly poor prognosis (13, 15). We confirmed this by obtaining GCTC patients' survival indicators from the SEER database, utilizing the Kaplan–Meier method. Since most patients have no conscious symptoms in the early clinical stage of metastasis, and there is a possibility of missing micrometastases on imaging, the construction of an effective model to predict the risk of stage M1b in patients with GCTC is of great value in clinical application. To the best of our knowledge, this study is the first study to develop an accurate predictive model for predicting the risk of developing



the M1b stage in patients with GCTC by incorporating multiple clinical and pathological indicators. In the baseline analysis, we found that the majority of patients received chemotherapy, but only a small percentage of patients received radiotherapy and LND, which is in line with our clinical experience and guideline recommendations. In terms of univariate logistic regression analysis, LND, chemotherapy, T-stage, N-stage, lung metastasis, distant lymph node metastasis, LDH, hCG, AFP, and S-stage were all significantly associated with the occurrence of developing the M1b stage. In the multivariate logistic regression, LND, T-stage, N-stage, lung metastasis, and distant lymph node metastasis were considered significant risk factors. Based on clinical reality, the inclusion of LND in the ML model means that the patient is judged to have an indication for LND by imaging or other assessment modalities preoperatively, rather than receiving LND, which results in an elevated risk of progression to the M1b stage. Unfortunately, in both models of multivariate logistic regression, serum tumor markers were not a predictor of progression to M1b stage in patients with GCTC, which may indicate that serum tumor markers (postoperative LDH, hCG, AFP) are more clinically significant in suggesting metastasis in the lung and distant lymph nodes and have limited predictive value for metastasis in other tissues or organs.

Machine learning is an important branch of AI, which learns the data structure of input data and its intrinsic patterns, selects corresponding learning methods and training methods to construct optimal mathematical models, and continuously adjusts model parameters to seek optimal solutions through mathematical methods to improve generalization ability and effectively prevent the occurrence of overfitting. ML has been widely used in various medical research fields as a powerful algorithm for predictive model building. Compared with traditional statistical methods, ML can better deal with overfitting, unbalanced data distribution and other problems (21, 24, 25). A total of six common ML algorithms were utilized in this study, including LR, XGBoost, lightGBM, RF, MLP, and kNN. The LR algorithm is often thought of as a traditional algorithm, but is essentially a form of machine learning (34). The XGBoost is a ML approach that has the unique ability to integrate missing data quickly and flexibly, as well as to assemble poor prediction models into a more accurate one (35, 36). The RF is a ML classifier that employs multiple trees to train and predict samples. It may be used to reduce training variance and increase integration and generalization (37). The other algorithms included have also shown high prediction accuracy, model stability, and computational efficiency in previous studies (38–40). Integrating the effectiveness and stability of the models in the training and external validation sets, XGBoost and RF were finally identified as two best prediction model algorithms for the risk prediction of progression to M1b in patients with GCTC. We hope to further validate the performance of these two models in the future through collaboration with multicenter medical units, hoping to specify a most efficient algorithmic model and to work with software development experts to develop a mobile program that facilitates clinically friendly applications.

Our study has certain limitations. First, the unavailability of data, including immunohistochemistry, patients' underlying

disease, and hematology index, limits the ability to further optimize the ML model, and we hope to incorporate these metrics at a later stage when a multicenter, real-world database is established. Second, S-stage was assessed by the postoperative serum tumor markers we obtained, which may have some human analysis errors because they are not directly available from the database. Meanwhile, the criteria for whether a patient has an indication for adjuvant therapy or LND are inconsistent from one medical institution to another and may be subjected to some errors in practical application. In addition, the practical value of the model obtained based on a predominantly Caucasian database for application in other centers (including China) is unclear due to the inevitable differences in ethnicity and treatment levels in different countries' or regions' validation. Nevertheless, our study is an important step forward in developing a model to predict the risk of developing the M1b stage in patients with GCTC.

CONCLUSION

We developed and validated ML algorithms for individualized prediction of the risk of progression to M1b stage in patients with GCTC who underwent orchiectomy by utilizing readily available perioperative patient clinical and pathological data. The ML-based prediction models can identify whether patients are at high risk and may assist the clinician in decision-making.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LD, KW, and CZ contributed to the idea and design. KW, CZ, YZ, and KLRW collected and analyzed the data. LD drew the figures and tables. LD and KW wrote the draft. LD, KW, CZ, YZ, KLRW, WL and JW contributed to manuscript writing and revision. All authors contributed to the article and approved the submitted version.

FUNDING

This study was sponsored by the Second Round of Xuzhou Medical Leading Talents Training Project (No. XWRCHT20210027).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.916513/full#supplementary-material>

Supplementary Figure 1 | Flow chart for patients selection of the SEER database.

Supplementary Figure 2 | Learning curves of six ML models in training set and cross-validation set.

Supplementary Figure 3 | The definition of S-stage based on the TNM classification for testicular cancer.

Supplementary Table 1 | Baseline characteristics of patients with GCTC in the SEER database.

Supplementary Table 2 | Original data from SEER database.

Supplementary Table 3 | Original data from the Affiliated Hospital of Xuzhou Medical University.

REFERENCES

- Bosl GJ, Motzer RJ. Testicular germ-cell cancer. *N Engl J Med.* (1997) 337:242–53. doi: 10.1056/NEJM199707243370406
- La Vecchia C, Bosetti C, Lucchini F, Bertuccio P, Negri E, Boyle P, et al. Cancer mortality in Europe, 2000–2004, and an overview of trends since 1975. *Ann Oncol.* (2010) 21:1323–60. doi: 10.1093/annonc/mdp530
- McGlynn KA, Trabert B. Adolescent and adult risk factors for testicular cancer. *Nat Rev Urol.* (2012) 9:339–49. doi: 10.1038/nrurol.2012.61
- Park JS, Kim J, Elghiaty A, Ham WS. Recent global trends in testicular cancer incidence and mortality. *Medicine.* (2018) 97:e12390. doi: 10.1097/MD.00000000000012390
- Hanna NH, Einhorn LH. Testicular cancer—discoveries and updates. *N Engl J Med.* (2014) 371:2005–16. doi: 10.1056/NEJMra1407550
- Dalgaard MD, Weinhold N, Edsgard D, Silver JD, Pers TH, Nielsen JE, et al. A genome-wide association study of men with symptoms of testicular dysgenesis syndrome and its network biology interpretation. *J Med Genet.* (2012) 49:58–65. doi: 10.1136/jmedgenet-2011-100174
- Jorgensen N, Rajpert-De ME, Main KM, Skakkebaek NE. Testicular dysgenesis syndrome comprises some but not all cases of hypospadias and impaired spermatogenesis. *Int J Androl.* (2010) 33:298–303. doi: 10.1111/j.1365-2605.2009.01050.x
- Schaapveld M, van den Belt-Dusebout AW, Gietema JA, de Wit R, Horenblas S, Witjes JA, et al. Risk and prognostic significance of metachronous contralateral testicular germ cell tumours. *Br J Cancer.* (2012) 107:1637–43. doi: 10.1038/bjc.2012.448
- Kuczyk MA, Serth J, Bokemeyer C, Jonassen J, Machtens S, Werner M, et al. Alterations of the p53 tumor suppressor gene in carcinoma in situ of the testis. *Cancer-Am Cancer Soc.* (1996) 78:1958–66. doi: 10.1002/(SICI)1097-0142(19961101)78:9<1958::AID-CNCR17>3.0.CO;2-X
- Hoffmann R, Plug I, McKee M, Khoshiba B, Westerling R, Looman C, et al. Innovations in health care and mortality trends from five cancers in seven European countries between 1970 and 2005. *Int J Public Health.* (2014) 59:341–50. doi: 10.1007/s00038-013-0507-9
- Fossa SD, Horwich A, Russell JM, Roberts JT, Cullen MH, Hodson NJ, et al. Optimal planning target volume for stage I testicular seminoma: a medical research council randomized trial. Medical research council testicular tumor working group. *J Clin Oncol.* (1999) 17:1146. doi: 10.1200/JCO.1999.17.4.1146
- Einhorn LH. Testicular cancer as a model for a curable neoplasm: the Richard and Hinda Rosenthal foundation award lecture. *Cancer Res.* (1981) 41:3275–80.
- Xu P, Wang J, Abudurexiti M, Jin S, Wu J, Shen Y, et al. Prognosis of patients with testicular carcinoma is dependent on metastatic site. *Front Oncol.* (2019) 9:1495. doi: 10.3389/fonc.2019.01495
- Cohn-Cedermark G, Stahl O, Tandstad T. Surveillance vs. adjuvant therapy of clinical stage I testicular tumors—a review and the SWENOTECA experience. *Andrology.* (2015) 3:102–10. doi: 10.1111/andr.280
- International Germ Cell Consensus Classification: A prognostic factor-based staging system for metastatic germ cell cancers. International Germ Cell Cancer Collaborative Group. *J Clin Oncol.* (1997) 15: 594–603. doi: 10.1200/JCO.1997.15.2.594
- Jamal-Hanjani M, Karpathakis A, Kwan A, Mazhar D, Ansell W, Shamash J, et al. Bone metastases in germ cell tumours: lessons learnt from a large retrospective study. *BJU Int.* (2013) 112:176–81. doi: 10.1111/bju.12218
- Smith ZL, Wernitz RP, Eggener SE. Testicular cancer: epidemiology, diagnosis, and management. *Med Clin North Am.* (2018) 102:251–64. doi: 10.1016/j.mcna.2017.10.003
- Isidori AM, Pozza C, Gianfrilli D, Giannetta E, Lemma A, Pofi R, et al. Differential diagnosis of nonpalpable testicular lesions: qualitative and quantitative contrast-enhanced US of benign and malignant testicular tumors. *Radiology.* (2014) 273:606–18. doi: 10.1148/radiol.14132718
- Pierorazio PM, Cheaib JG, Tema G, Patel HD, Gupta M, Sharma R, et al. Performance characteristics of clinical staging modalities for early stage testicular germ cell tumors: a systematic review. *J Urol.* (2020) 203:894–901. doi: 10.1097/JU.0000000000000594
- Li W, Zhou Q, Liu W, Xu C, Tang ZR, Dong S, et al. A machine learning-based predictive model for predicting lymph node metastasis in patients with ewing's sarcoma. *Front Med.* (2022) 9:832108. doi: 10.3389/fmed.2022.832108
- Li W, Wang J, Liu W, Xu C, Li W, Zhang K, et al. Machine learning applications in the prediction of bone cement leakage in percutaneous vertebroplasty. *Front Public Health.* (2021) 9:812023. doi: 10.3389/fpubh.2021.812023
- Deo RC. Machine learning in medicine. *Circulation.* (2015) 132:1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593
- Vougas K, Sakellaropoulos T, Kotsinas A, Foukas GP, Ntargaras A, Koinis F, et al. Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on association rule mining. *Pharmacol Ther.* (2019) 203:107395. doi: 10.1016/j.pharmthera.2019.107395
- Gallagher DJ, Kemeny N. Metastatic colorectal cancer: from improved survival to potential cure. *Oncology.* (2010) 78:237–48. doi: 10.1159/000315730
- Chung P, Daugaard G, Tyldesley S, Atenafu EG, Panzarella T, Kollmannsberger C, et al. Evaluation of a prognostic model for risk of relapse in stage I seminoma surveillance. *Cancer Med.* (2015) 4:155–60. doi: 10.1002/cam4.324
- Barlow LJ, Badalato GM, McKiernan JM. Serum tumor markers in the evaluation of male germ cell tumors. *Nat Rev Urol.* (2010) 7:610–7. doi: 10.1038/nrurol.2010.166
- Gilligan TD, Seidenfeld J, Basch EM, Einhorn LH, Fancher T, Smith DC, et al. American society of clinical oncology clinical practice guideline on uses of serum tumor markers in adult males with germ cell tumors. *J Clin Oncol.* (2010) 28:3388–404. doi: 10.1200/JCO.2009.26.4481
- Nicholson BD, Jones NR, Protheroe A, Joseph J, Roberts NW, Van den Bruel A, et al. The diagnostic performance of current tumour markers in surveillance for recurrent testicular cancer: a diagnostic test accuracy systematic review. *Cancer Epidemiol.* (2019) 59:15–21. doi: 10.1016/j.canep.2019.01.001
- van Dijk MR, Steyerberg EW, Habbema JD. Survival of non-seminomatous germ cell cancer patients according to the IGCC

- classification: an update based on meta-analysis. *Eur J Cancer*. (2006) 42:820–6. doi: 10.1016/j.ejca.2005.08.043
32. Flechtner HH, Fischer F, Albers P, Hartmann M, Siener R. Quality-of-life analysis of the german prospective multicentre trial of single-cycle adjuvant BEP versus retroperitoneal lymph node dissection in clinical stage i nonseminomatous germ cell tumours. *Eur Urol*. (2016) 69:518–25. doi: 10.1016/j.eururo.2015.11.007
 33. Melchior D, Hammer P, Fimmers R, Schuller H, Albers P. Long term results and morbidity of paraaortic compared with paraaortic and iliac adjuvant radiation in clinical stage I seminoma. *Anticancer Res*. (2001) 21:2989–93.
 34. Choi Y, Boo Y. Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality. *Int J Environ Res Public Health*. (2020) 17:897. doi: 10.3390/ijerph17030897
 35. Davagdorj K, Pham VH, Theera-Umpon N, Ryu KH. XGBoost-based framework for smoking-induced non-communicable disease prediction. *Int J Environ Res Public Health*. (2020) 17:6513. doi: 10.3390/ijerph17186513
 36. Liu Y, Wang H, Fei Y, Liu Y, Shen L, Zhuang Z, et al. Research on the prediction of green plum acidity based on improved XGBoost. *Sensors*. (2021) 21:930. doi: 10.3390/s21030930
 37. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep*. (2020) 10:5245. doi: 10.1038/s41598-020-62133-5
 38. Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, Xiao Y, et al. LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biol*. (2021) 22:271. doi: 10.1186/s13059-021-02492-y
 39. Ehsani R, Drablos F. Robust distance measures for kNN classification of cancer data. *Cancer Inform*. (2020) 19:1882255450. doi: 10.1177/1176935120965542
 40. Haghighat F. Predicting the trend of indicators related to Covid-19 using the combined MLP-MC model. *Chaos Solitons Fractals*. (2021) 152:111399. doi: 10.1016/j.chaos.2021.111399

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ding, Wang, Zhang, Zhang, Wang, Li and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY

Yu-Hsiu Lin,
National Chung Cheng
University, Taiwan

REVIEWED BY

Saptarshi Bej,
University of Rostock, Germany
Yi Han,
Nanjing Medical University, China

*CORRESPONDENCE

Yongjun Wu
wuyongjun@zzu.edu.cn
Lijun Miao
miaolily@126.com
Yanbin Wang
1611428792@qq.com

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 08 May 2022

ACCEPTED 23 June 2022

PUBLISHED 28 July 2022

CITATION

Effah CY, Miao R, Drokow EK,
Agboyibor C, Qiao R, Wu Y, Miao L and
Wang Y (2022) Machine
learning-assisted prediction of
pneumonia based on non-invasive
measures.
Front. Public Health 10:938801.
doi: 10.3389/fpubh.2022.938801

COPYRIGHT

© 2022 Effah, Miao, Drokow,
Agboyibor, Qiao, Wu, Miao and Wang.
This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Machine learning-assisted prediction of pneumonia based on non-invasive measures

Clement Yaw Effah¹, Ruoqi Miao¹,
Emmanuel Kwateng Drokow², Clement Agboyibor³,
Ruiping Qiao⁴, Yongjun Wu^{1*}, Lijun Miao^{4*} and Yanbin Wang^{5*}

¹College of Public Health, Zhengzhou University, Zhengzhou, China, ²Department of Radiation Oncology, Zhengzhou University People's Hospital, Henan Provincial People's Hospital, Zhengzhou, China, ³School of Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, China, ⁴Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, ⁵Center of Health Management, General Hospital of Anyang Iron and Steel Group Co., Ltd, Anyang, China

Background: Pneumonia is an infection of the lungs that is characterized by high morbidity and mortality. The use of machine learning systems to detect respiratory diseases via non-invasive measures such as physical and laboratory parameters is gaining momentum and has been proposed to decrease diagnostic uncertainty associated with bacterial pneumonia. Herein, this study conducted several experiments using eight machine learning models to predict pneumonia based on biomarkers, laboratory parameters, and physical features.

Methods: We perform machine-learning analysis on 535 different patients, each with 45 features. Data normalization to rescale all real-valued features was performed. Since it is a binary problem, we categorized each patient into one class at a time. We designed three experiments to evaluate the models: (1) feature selection techniques to select appropriate features for the models, (2) experiments on the imbalanced original dataset, and (3) experiments on the SMOTE data. We then compared eight machine learning models to evaluate their effectiveness in predicting pneumonia

Results: Biomarkers such as C-reactive protein and procalcitonin demonstrated the most significant discriminating power. Ensemble machine learning models such as RF (accuracy = 92.0%, precision = 91.3%, recall = 96.0%, f1-Score = 93.6%) and XGBoost (accuracy = 90.8%, precision = 92.6%, recall = 92.3%, f1-score = 92.4%) achieved the highest performance accuracy on the original dataset with AUCs of 0.96 and 0.97, respectively. On the SMOTE dataset, RF and XGBoost achieved the highest prediction results with f1-scores of 92.0 and 91.2%, respectively. Also, AUC of 0.97 was achieved for both RF and XGBoost models.

Conclusions: Our models showed that in the diagnosis of pneumonia, individual clinical history, laboratory indicators, and symptoms do not have adequate discriminatory power. We can also conclude that the ensemble ML models performed better in this study.

KEYWORDS

pneumonia, machine learning, non-invasive measures, electronic health records (EHR), decision support system (DSS)

Introduction

Pneumonia has been a major cause of morbidity and mortality in both developed and developing countries, especially among patients who are diagnosed and treated at a later stage (1, 2). Specific symptoms such as cough with sputum production, fever, chest pain, shortness of breath, and chills are the main characteristics associated with pneumonia (3). Because of several reasons such as difficulty in identifying the etiological agents in individuals, low specificity of symptoms of lower respiratory tract infections, and lack of widespread availability of laboratory tests and imaging, the accurate definition and diagnosis of pneumonia are still debatable (4). Diagnostic findings such as decreased breathing sounds, crackles, bronchial breath sounds, egophony, along with a sharp increase in body temperature, tachypnea, hypoxia, tachycardia, and dyspnea, should suggest pneumonia (either broncho- or lobar). Pneumonia benefits from antibiotics. So, to prevent unnecessary administration of antibiotics that might ultimately create multi-drug-resistant “superbugs” - as has already happened - procalcitonin levels are monitored along with clinical symptoms. Procalcitonin is released from lung neuroendocrine cells after exposure to bacterial endotoxin and lipopolysaccharides which typically increases the production of procalcitonin. The appearance of pneumonia symptomatology coupled by a rise in procalcitonin levels would trigger antibiotic administration.

Although chest radiography is the recommended technique for pneumonia diagnosis, factors such as lack of standardized interpretation (5), inter-rater variability (6), absence of abnormalities in the chest radiographs of children (7), low sensitivity to early-stage pneumonia, and potential harm due to exposure to x-rays hinder their use. Most importantly, radiography is usually not available in areas with the highest disease burden such as those in low-income settings. Consequently, general practitioners mainly rely on non-invasive measures such as the use of signs, symptoms, and simple laboratory tests as tools to diagnose pneumonia. To improve diagnostic accuracy and enhance various treatment strategies for pneumonia, prediction models based on non-invasive measures have been proposed.

Machine learning (ML), a powerful computer-based method that has the capacity to learn, reason, and self-correct without explicit programming, has the potential to provide solutions to the above problems. In recent years, the use of ML has achieved great advances and major benefits in medicine. Researchers have used large clinical databases to answer previously unanswerable questions and create systems that facilitate human decision-making (8, 9). Over the years, enthusiasm and optimism have been alternated with skepticism and pessimism in this fascinating field of research. Although some claims associated with this kind of research are currently being made with great grandiose claims (10), ML-based models have already proven to be useful in some clinical applications (11). ML has been shown to improve diagnostic accuracy for pneumonia when applied to hospitalized patients (12). The use of machine learning techniques to detect pneumonia *via* non-invasive measures such as signs and symptoms is gaining much attention. In several clinical studies, clinical history and physical examination parameters have been evaluated for their diagnostic value in predicting pneumonia (13, 14).

Based on the above, this study conducted several experiments on various ML models to predict pneumonia based on biomarkers, laboratory, and physical features.

Methods

Data collection and preprocessing

We retrospectively recruited patients aged at least 18 with confirmed acute lower respiratory illness and treated at the First Affiliated Hospital of Zhengzhou University in Henan Province between October 29, 2019, and May 21, 2021. The First Affiliated Hospital of Zhengzhou University is one of the largest hospitals in central China, with an ~13,000-bed capacity. We extracted patient demographic information (including age, sex, and comorbidities), physical parameters (tachycardia, tracheal secretion, pleural effusion, mean arterial pressure, heart rates, breathing rates, and systolic blood pressure), and hematological parameters. Hematological parameters included serum sodium, serum potassium, serum creatinine, hematocrit, WBC count, platelet, total bilirubin, hemoglobin, C-reactive protein (CRP),

and procalcitonin (PCT). Unfortunately, some patients had some missing data. As a result, we later addressed some of these missing values in the dataset (data preprocessing). Typically, real-world data contains multiple errors, incompleteness, and incoherence, requiring data preprocessing. Because of this, we preprocessed the data following these four steps:

Missing values

Missing data causes problems when a ML model is applied to the dataset. Mostly, ML models don't process data with missing values. In this study, some variables had several missing values of about 15% of that variable data. We used the median and mode of the corresponding columns to fill in the missing values of numerical attributes and categorical attributes, respectively. Median is the centrally located value of the dataset in ascending order. We filled missing numerical attribute values with the median value. Mode is the most repeated value in the given categorical observations. We filled missing entries with the mode observations.

Imbalance data

The dataset was unbalanced. A classification dataset with skewed class proportions is called imbalanced. Classes that make up a large proportion of the dataset are called majority classes. Those that make up a smaller proportion are minority classes. The degree of imbalance in the minority class can be mild (20–40% of the dataset), moderate (1–20% of the dataset), and extreme (<1% of the dataset). In this study, the minority class was 22% lesser than the majority class. Therefore, we needed to resolve the issue before applying machine learning in order to reduce data bias. One of the over-sampling approaches to fix imbalanced data is the synthetic minority over-sampling technique (SMOTE) (15). It manages overfitting induced by a limited decision interval by controlling the generation and distribution of manual samples using the minority class sample. Specifically, SMOTE is based on selecting a random minority class as the last sample. Then it finds the k nearest neighbors (normally $k = 5$) of the selected prior sample. Finally, it selects a random neighbor and creates a synthetic sample between the two samples (prior sample and selected neighbor) in the feature space at a randomly selected point. We can express SMOTE as

$$SMOTE(x_{syn}) = x_p + (x_{knn} - x_p) \times \alpha,$$

where x_p denotes feature vector of a prior sample, x_{knn} represents the k nearest neighbors, and α is the randomly selected point.

Data rescaling

Before applying ML algorithms, one important step required in data preprocessing is data rescaling. This makes the various

ML models more effective. The dataset contained various scales for various quantities (e.g., age, mean arterial pressure, heart rate, C-reactive protein, and procalcitonin). Therefore, we perform data normalization to rescale all real-valued features:

$$\tilde{x} = \frac{x - avg}{std},$$

where x denotes the value, avg is the average of the values, and std is the standard deviation. For models like logistic regression, which rely on the magnitude of values to determine coefficients, this step is highly important.

Feature selection

Some features contribute to predicting a variable of interest than others. Feature selection is, therefore, performed to automatically select those features. By doing this, the accuracy is improved, overfitting is reduced, and most importantly, the time required for training is reduced. Irrelevant features can reduce the performance of several machine learning models. We investigated six techniques of feature selection: Lower variance, L1 regularization-based feature selection, L2 regularization-based feature selection, Univariate feature selection, Tree-based feature selection, and Principal Component Analysis (PCA).

- Eliminate lower variance (LV): Variance quantifies how widely apart a collection of data is. When the variance is 0, all of the data values are the same and vice-versa. The formula to compute variance is given as

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

where n is the number of pieces of data, x_i is each of the values in the data, and \bar{x} is the mean (average) of the data. If the variance is low or near zero, that feature is relatively constant and will not increase the model's performance. Hence, it should be removed.

- Univariate feature selection: The univariate feature selection (UFS) selects the best features by applying univariate statistical tests. Specifically, UFS examines each feature exclusively to determine the strength of the feature's relationship with the response variable using the Chi-Squared Test. Given the data of two variables, the Chi-Squared Test observes count O and expected count E . Chi-Square measures how expected count E and observed count O deviate from each other. The formula for chi-square is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where c is the degree of freedom, O denotes observed values(s), and E denotes expected values(s).

- L1/L2 regularization-based feature selection: The solutions to linear models penalized with the L1 norm are sparse: many estimated coefficients are 0. L1/L2 regularization-based feature selection can reduce the dimensionality of the data by selecting features with non-zero coefficients. The L2 norm adds “squared magnitude” of coefficient as a penalty term to the loss function as

$$\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

While the L1 norm adds an absolute value of the magnitude of coefficient as a penalty term to the loss function as

$$\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- Tree-based feature selection: We used tree-based estimators to calculate the impurity-based feature importance; this can be used to remove irrelevant features. We used a Random Forest algorithm. We selected 50 decision trees, each constructed using a random extraction of observations from the dataset and features. Because most data characteristics are not seen by some trees, they (the trees) are de-correlated which makes them less prone to over-fitting. Each tree is also a series of yes-no questions based on a single or many characteristics. The tree splits the dataset into two buckets at each node, each containing more similar observations and distinct from those in the other bucket. As a result, the significance of each attribute is determined by how “pure” each of the buckets is.
- Principal Component Analysis (PCA): We utilized PCA to reduce the dimensions of our larger dataset. Essentially, the reduced dataset still contains much of the information in the large set. It is accomplished by evaluating the correlation between features in order to find the most important principal components. Although it is clear that there are other better options such as t-SNE and UMAP for dimension reduction, these reasons were considered before choosing PCA for this task. t-SNE involves a lot of calculations and computations because it computes pairwise conditional probabilities for each data point and tries to minimize the sum of the difference of the probabilities in higher and lower dimensions. t-SNE has a quadratic time and space complexity in the number of data points. This makes it particularly slow, computationally quite heavy and resource draining. Also,

the main disadvantage of UMAP is its lack of maturity. It is a very new technique, so the libraries and best practices are not yet firmly established or robust. The short summary is that PCA is far and away from the fastest option, it is deterministic and linear. However, we potentially gave up a lot for that speed. We set the principal components to 26.

Experimental setup

We perform machine-learning analysis on 535 different patients, each with 45 features. Since it is a binary problem, we categorized each patient into one class at a time. We designed three experiments to evaluate the models: (1) feature selection techniques to select appropriate features for the models, (2) experiments on the imbalanced original dataset, and (3) experiments on the balanced data via SMOTE.

Prediction models

We compared several models to evaluate their effectiveness in predicting pneumonia: Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Adaboost Decision Tree (ADT), K-Nearest Neighbor (KNN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP). These models have been extensively used for many classification tasks.

Evaluation metrics

Following previous works (16, 17), and considering that machine learning models have multiple tuning parameters, which are essential for model performance, we adopted 5-fold cross-validation (CV) to evaluate all the classification models using confusion matrices (Figure 1) and ROCs. CV is a resampling technique used for evaluating and validating ML algorithms based on a small dataset sample. The dataset is randomly divided into k equal portions (number of folds). In training the model, the residual $k-1$ dataset is used, while the remaining dataset (validation dataset) is used to test the model. The CV procedure is then replicated k times with different folds as the test set each time. In order to achieve a specific outcome, all k outcomes from k -folds are summed and the average is then calculated (18, 19). In the 5-fold cross-validation, we randomly partition the dataset into five equal subsamples. One subsample was used as the validation set and the

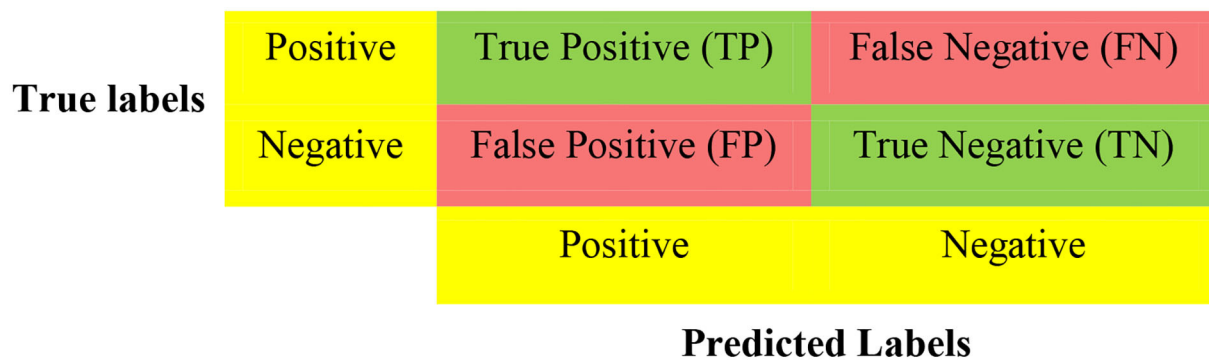


FIGURE 1
Confusion matrix.

TABLE 1 Performance evaluation metrics equations.

Metric	Equation
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F-measure	$2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$

remaining four subsamples were used as the training set. We divided all datasets into 80% training and 20% testing. We used the training data during the feature selection and training. However, the test data was used for model selection.

For binary classification, multiple criteria are needed in evaluating the performance of the models. As such, we evaluate the performance of the various models based on f-measure, Area Under the Curve of the Receiver Operator Characteristic (AUC-ROC), accuracy, recall, and precision. These performance metrics can be determined using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) as seen in Figure 1. The accuracy is the proportion of all correctly predicted samples to the total samples. The recall rate is the proportion of positive samples correctly identified as positive to the total number of positive samples. This metric is critical for our work since prediction models want to identify as many positive samples as possible. The precision defines the ratio of the number of positive samples accurately predicted as positive to the number of positive examples. Naturally, an excellent predictive model seeks a high recall rate and precision. There is, however, a trade-off between recall rate and accuracy; the F-measure provides a thorough assessment by computing the harmonic mean of recall and precision. Table 1 shows the equations used for calculating the desired performance metrics: accuracy, precision, recall, and f-measure.

TABLE 2 LR prediction result of feature selection methods on original dataset.

Feature selection	Accuracy	Precision	Recall	F1-score
LV	80.4	83.7	84.4	84.0
UFS	82.6	85.8	85.9	85.8
L1	75.9	79.0	82.5	80.7
L2	77.9	82.3	81.6	81.8
Tree-based	83.0	85.7	86.8	86.2
PCA	81.1	84.5	84.7	84.6

Results

Data balancing, rescaling, and feature selection

The dominant class of the dataset had 22% more samples (Figure 2). After SMOTE, we obtain two types of datasets: the original imbalanced dataset and the SMOTE dataset.

We then used Logistic Regression as the baseline model to choose the appropriate feature selection methods. The results show that the Tree-based is most effective on the original data followed by UFS (Table 2). In the SMOTE dataset, PCA is most effective, followed by LV (Table 3). We used Tree-based and UFS in subsequent experiments on the original dataset and reported the best results. Likewise, we used PCA and LV in subsequent experiments on the balanced SMOTE dataset and reported the best results.

Evaluation of the performance of the machine learning models on the original dataset

We conducted experiments to acquire empirical evidence on the original imbalanced dataset using the ML models listed

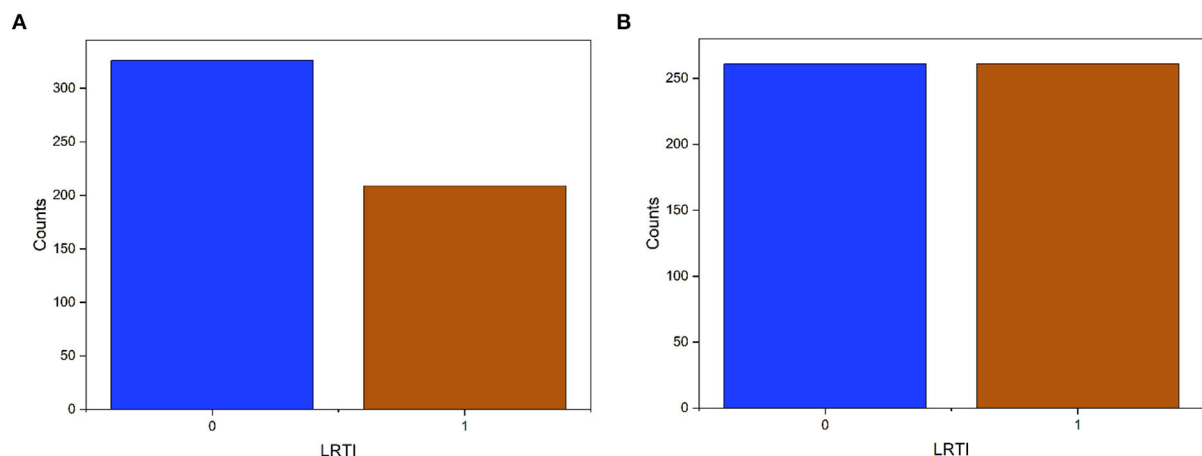


FIGURE 2
Target data (LRTI) distribution before and after applying SMOTE. The label '0' is pneumonia and "1" for bronchitis. (A) Imbalanced data. (B) Balance data.

TABLE 3 LR prediction result of feature selection method on balanced dataset.

Feature selection	Accuracy	Precision	Recall	F1-score
LV	83.6	85.4	81.3	83.4
UFS	82.2	83.3	80.9	82.0
L1	77.3	78.2	75.4	77.1
L2	79.1	81.5	75.2	78.0
Tree-based	82.0	83.1	80.3	81.6
PCA	85.4	86.6	83.0	84.7

TABLE 4 Machine learning model prediction results on the original dataset.

Model	Accuracy	Precision	Recall	F1-score
LR	81.4	82.7	84.2	84.3
NB	59.8	89.6	39.2	53.7
SVM	80.7	82.8	86.5	84.5
ADT	90.1	91.3	92.7	91.9
KNN	72.1	87.3	63.8	73.5
RF	92.0	91.3	96.0	93.6
XGBoost	90.8	92.6	92.3	92.4
MLP	79.4	83.7	82.5	82.9

above. From [Table 4](#), the Ensemble machine learning models such as RF (accuracy = 92.0%, precision = 91.3%, recall = 96.0%, f1-Score = 93.6%) and XGBoost (accuracy = 90.8%, precision = 92.6%, recall = 92.3%, f1-score = 92.4%) achieved the highest performance accuracy while NB achieved the lowest performance accuracy on the original imbalanced dataset. Also,

ADT (accuracy = 90.1%, precision = 91.3%, recall = 92.7%, F1-Score = 91.9%) had a performance which was almost similar to that of XGBoost.

We also visualize the confusion matrix of RF and XGBoost in [Figure 3](#). We observe that the XGBoost model wrongly predicted more pneumonia samples (25) than the RF model (13). However, XGBoost performed better than RF when predicting other LRTIs other than pneumonia. Generally, it can be deduced that the models could learn from the data.

The ROC curves of the XGBoost and RF are shown in [Figure 4](#). We observe that both the XGBoost and RF models achieve a similar performance of 0.97 and 0.96, respectively. Also, the “steepness” of the ROC shows that the XGBoost model has a slightly high positive rate than the RF model.

[Figures 5, 6](#) show the essential features that XGBoost and RF models consider essential for prediction. Both XGBoost and RF models consider hemoglobin, C-reactive protein, and procalcitonin features very notably. Tracheal secretion, antibiotics taken within the last 90 days, total bilirubin and hematocrit features are also considered necessary by both models, but their importance is relatively low compared with those listed earlier. However, XGBoost does not consider some features necessary (e.g., age, years of smoking, years of drinking, dyspnea, tachycardia) compared to the RF model.

Evaluation performance of the machine learning models on the SMOTE dataset

We also conducted experiments to acquire empirical evidence on the SMOTE dataset using similar machine learning models listed above.

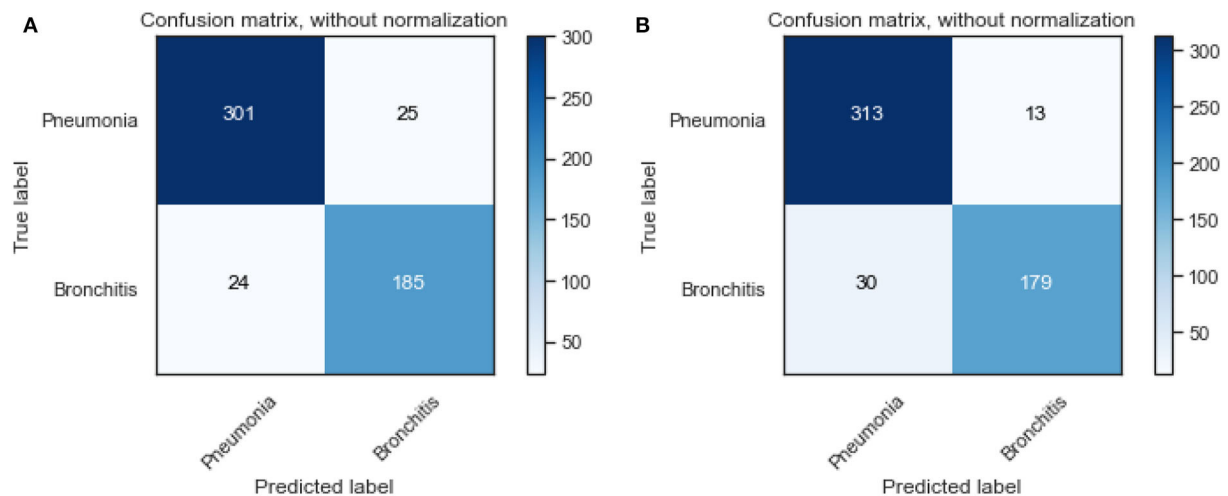


FIGURE 3
Confusion matrix of XGBoost and random forest on the original dataset. (A) XGBoost. (B) RF.

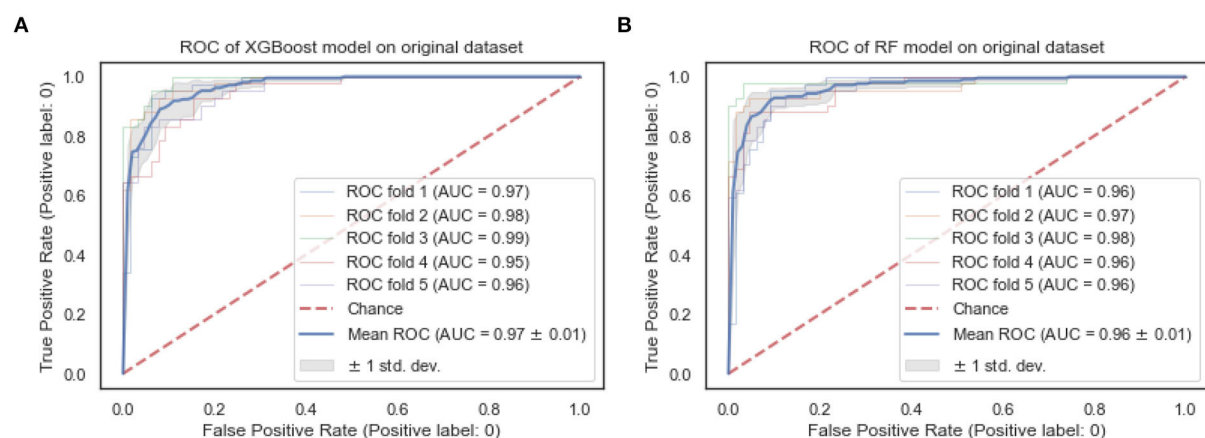


FIGURE 4
ROC curves of XGBoost and random forest on the original dataset. (A) XGBoost. (B) RF.

From Table 5, the RF model achieved the highest performance followed by XGBoost and ADT, while NB achieved the lowest prediction performance. The f1-scores of RF and XGBoost are 92.0 and 91.2%, respectively, which indicates how robust the models are.

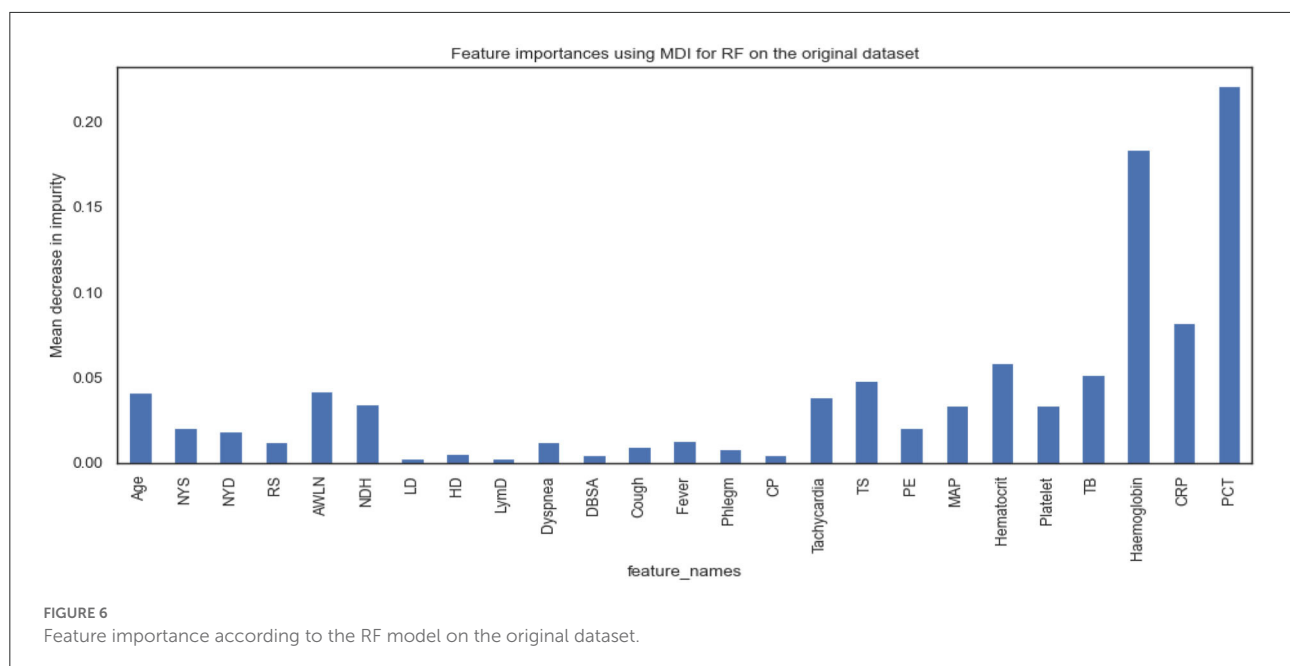
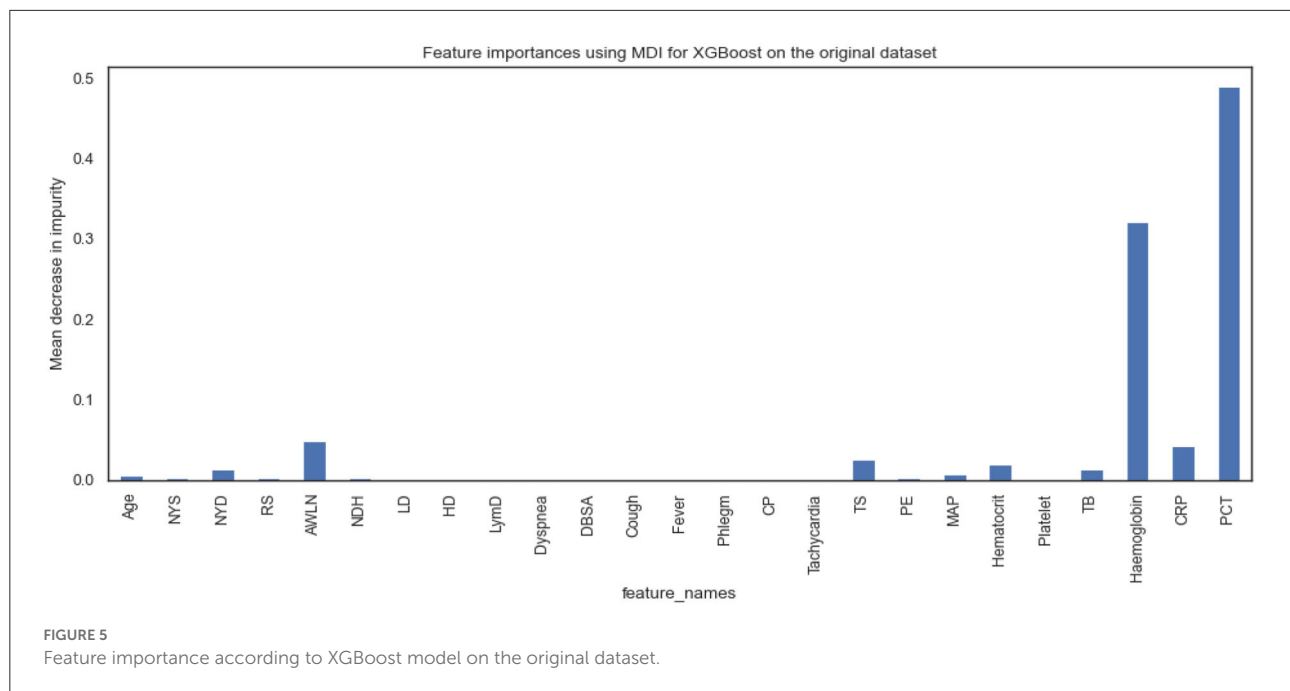
We also visualized the confusion matrix of XGBoost and RF in Figure 7 and made the following observations. The XGBoost model wrongly predicted more pneumonia samples (24) than the RF model (18). Generally, it was observed that the models could learn significantly from the data.

The ROC curves of the XGBoost and RF are shown in Figure 8. We observe that RF models achieve the same superior performance as the XGBoost model. Also, the “steepness” of the ROC shows that the

RF model has a slightly high positive rate than the XGBoost model.

Figures 9, 10 show the features the XGBoost and RF model considers vital for prediction. XGBoost and RF models consider hemoglobin, hematocrit, drinking, mean arterial pressure, plural effusion, tracheal secretion, tachycardia, years of smoking, C-reactive protein, antibiotics taken within the last 90 days, procalcitonin, and total bilirubin features significantly in the prediction.

Because we performed machine learning experiments on both the original and the SMOTE data, we run ANOVA to compare whether there are statistical differences in the prediction performances of the models before and after SMOTE.



We did this by comparing their AUCs. AUC is a measure of the accuracy of a quantitative diagnostic test. It is the average value of specificity for all values of sensitivity. Table 6 shows the AUCs of the models for the original and balanced datasets. We observed that the AUCs of some models (LR, MLP, KNN, NB) differ significantly in the two datasets while others (SVM, XGBoost, ADT, RF) achieved similar or showed no significant difference in their before and after SMOTE AUCs.

Decision boundaries of the models

Decisions, or boundaries, are lines drawn using the best decisions (for our purposes, binary classifiers) that separate samples of one class from the other class. All instances of one class and the opposing class are found on each side. The decision boundaries of the models show that the RF and XGBoost models learn a robust decision boundary (Figure 11). RF and XGBoost models can learn and correctly classify the samples at the bottom

compared to the other models. This observation is expected because the two models (RF and XGBoost) achieved the best performance on the original dataset.

Based on the balanced dataset (Figure 12), the ADT, RF, and XGBoost models demonstrate a well-bodied boundary while LR and SVM show poor boundaries.

External validation of the models

To validate our models for generalizability, we externally collected data from 77 patients with lower respiratory tract infections (either pneumonia or bronchitis). The two best models, RF and XGBoost, were chosen for the external validation. Although the data used for this experiment was limited, the models were still robust in the prediction of pneumonia (Table 7). The ROCs values (Figure 13) show AUCs

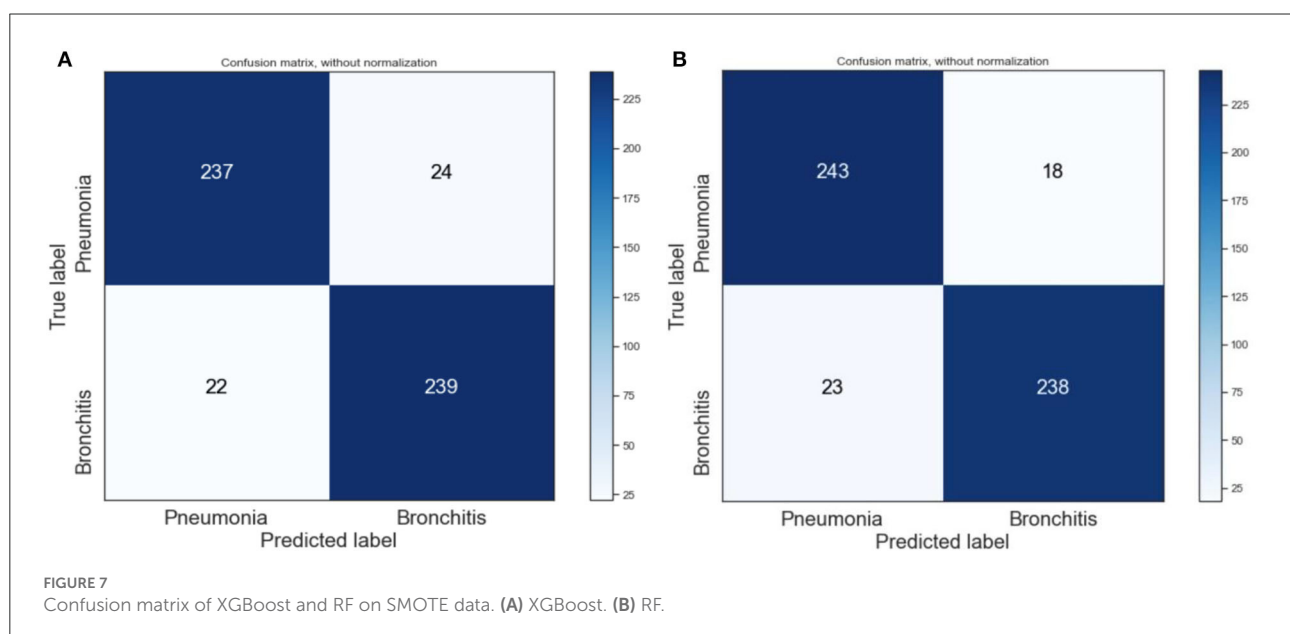
of 95 and 96% for XGBoost and RF models confirming that our models have good generality.

Discussion

Laboratory tests, blood culture, C-reactive protein, serology, and procalcitonin are diagnostic tests with varying rates of accuracy (20). Our models showed that individual clinical history and symptoms do not have adequate discriminatory power except dyspnea, diminishing breath sound on auscultation, cough, fever, and phlegm to diagnose pneumonia. Earlier studies have shown that radiographic pneumonia cannot be diagnosed by a single clinical symptom and this was consistent with our study (21). Fever, tachycardia, and breathing rate were among the most useful predictors of the clinical signs. Evidence suggests that adults whose respiration rates exceed 20 breaths per minute are probably unwell, and those whose respiration rates exceed 24 breaths per minute are deemed to be critically ill (22). The findings of this study are similar to previously published study (23). Similar to other studies (24), diminishing sound on auscultation was shown to be an important predictor of pneumonia in our models. As part of externally validated prediction models for pneumonia, diminishing sound on auscultation, tachycardia, and fever were found to be useful predictors (25). In a study by Niederman et al., it was postulated that patients with symptoms such as cough, sputum production, and/or dyspnea, in addition to other indicators like fever and auscultatory findings of abnormal breath sounds may have a higher risk of developing pneumonia (26). Tracheal secretion, antibiotics taken within the last 90 days, total bilirubin, and hematocrit were all features considered

TABLE 5 Machine learning model prediction results in the balanced dataset.

Model	Accuracy	Precision	Recall	F1-score
LR	83.6	84.9	81.2	83.1
NB	68.4	75.8	54.4	62.7
SVM	81.1	83.0	77.2	80.1
ADT	91.0	91.2	90.1	90.9
KNN	75.0	91.9	54.8	68.4
RF	92.2	93.0	91.2	92.0
XGBoost	91.2	91.1	91.6	91.2
MLP	81.4	81.9	83.2	82.4



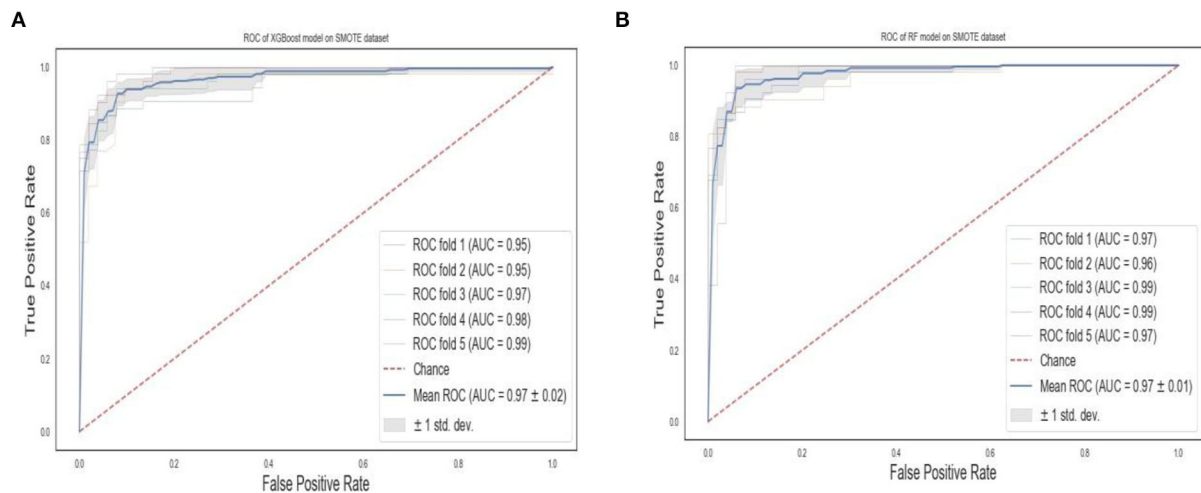


FIGURE 8
ROC curves of XGBoost and random forest on the SMOTE dataset. (A) XGBoost. (B) RF.

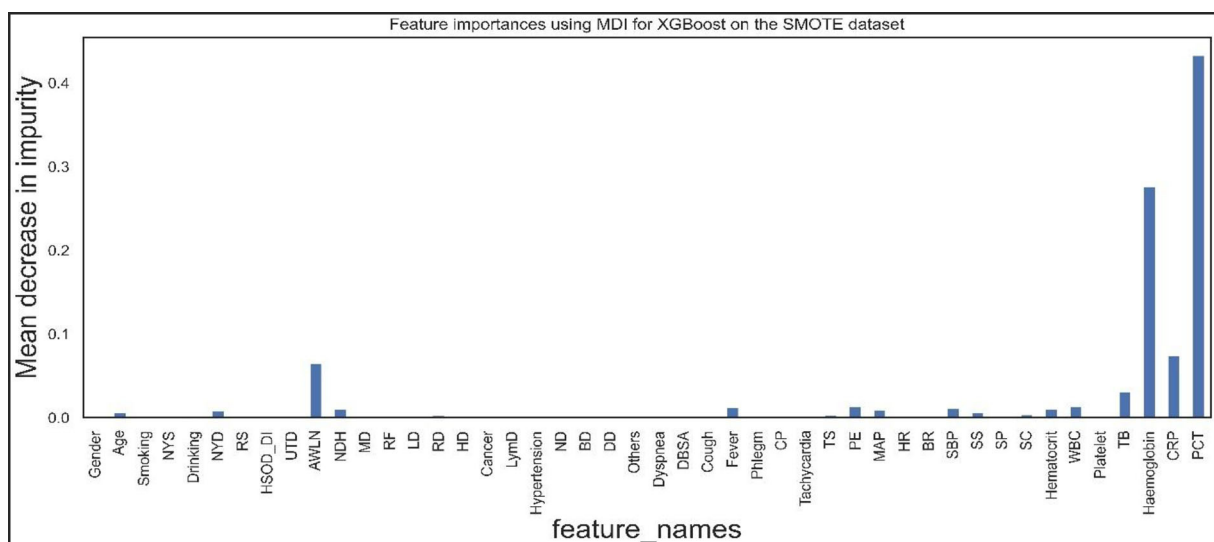


FIGURE 9
Feature importance according to the XGBoost model on the SMOTE dataset.

important for pneumonia prediction in our models. Tracheal secretion has been noted by several authors as an important diagnostic tool for pneumonia (27, 28).

Biomarkers can support clinicians in the differentiation of bacterial pneumonia from other disorders. Among all the variables tested in our prediction models, biomarkers such as CRP and PCT demonstrated the most significant discriminating power in the prediction of pneumonia. CRP and PCT, are extensively used in the monitoring of treatment of severe infections in the ICU. PCT is a marker that is strongly correlated with bacteria load and severity of infection

(29). Also, a high PCT level indicates a bacterial infection rather than a viral infection. A meta-analysis reported that the use of PCT to guide antibiotic treatment in pneumonia resulted in a reduction in exposure to antibiotics (30). A PCT level of >0.25 ng/ml is indicative of an underlying bacterial infection (31). This evidence supports our results that, PCT can accurately predict pneumonia. Among patients with pneumonia, the prognostic value of PCT and its correlation with disease severity has been exclusively studied (31). In ambulatory care, CRP has been widely used as a point of care test. Researchers have examined CRP as a diagnostic

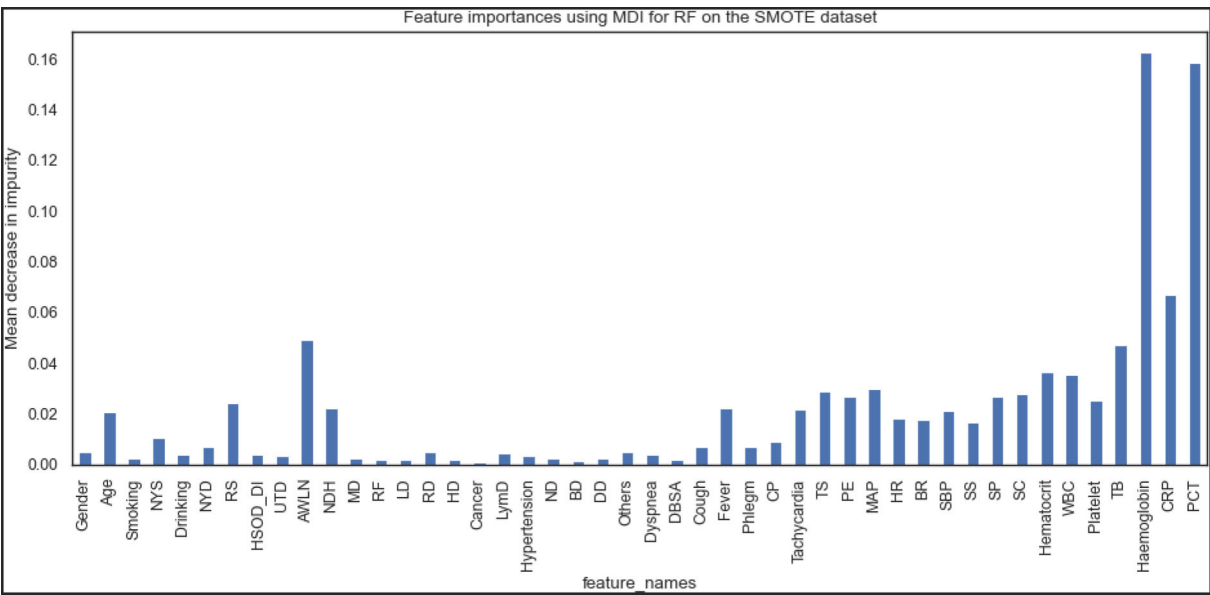


FIGURE 10 Feature importance according to the RF model on the SMOTE dataset.

TABLE 6 AUCs of the various models before and after SMOTE.

Model	Original dataset	Balanced dataset	P value
LR	89	91	0.032
NB	82	76	0.019
SVM	89	86	0.221
ADT	91	94	0.071
KNN	79	84	0.016
RF	96	97	0.050
XGBoost	97	97	0.314
MLP	80	86	0.005

tool in screening for inflammation and detecting bacterial infections (32). Through the use of CRP in primary care, antibiotic exposure can be reduced in suspected LRTI (risk ratio [RR] 0.78 [95% CI 0.66–0.92]) (33). According to the NICE’s guidelines, antibiotics should not be given to patients without a convincing clinical diagnosis of pneumonia, when their CRP is <20 mg/L (34). Our results showed that CRP is a useful diagnostic tool to predict pneumonia. This finding is similar to previous studies (32). CRP has been shown to improve the diagnostic discriminatory power of models built on basic signs and symptoms during the prediction of patients with pneumonia (35).

From our machine learning models, RF and XGBoost were considered the best models on both the original dataset and the

SMOTE balanced data. RF model has demonstrated superiority and stability in numerous medical studies (36, 37). Because of the extensive application of integrated algorithms, the RF model has become a well-established technology (38). RF uses the bagging ensemble technique for classification. Decision trees (DTs) are the building blocks of the RF classifier. In order to train uncorrelated decision trees, each tree is trained with a random sample selected from the dataset. Then, final decisions are made by combining the outputs from all the DTs. Because the forest is randomized, it slightly increases the biasness of the forest. However, due to the averaging of the outputs, its variance decreases, hence yielding an overall better model. As an efficient and scalable tree boosting system (39), the XGBoost model has shown excellent performance in several ML competitions, primarily due to its simplicity and accuracy in prediction (40). Our study showed that the XGBoost model had a good performance, with an F1-score of 92.4% and an accuracy of 90.8%. Because ensemble ML models (RF and XGBoost) integrate multiple base learners or classifiers, they are robust and have high accuracy which was confirmed in this study. All models on the original data had AUC values lower than those observed in the ensemble ML models. However, comparing XGBoost, a boosting ensemble method to RF which is a bagging ensemble method, RF needs to train a large amount of decision trees and aggregate them, thereby requiring longer time to trade numerous random computations for high accuracy. Moreover, XGBoost leverages second order derivative and implements sampling method in

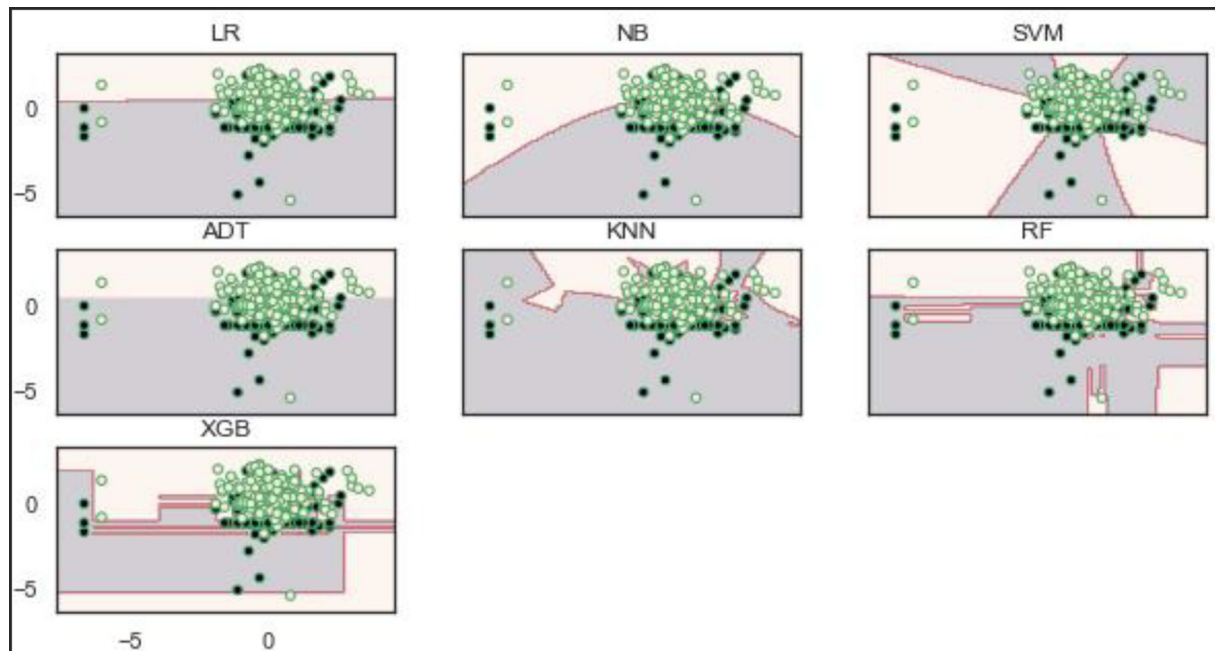


FIGURE 11
Decision boundaries of the models on the original dataset.

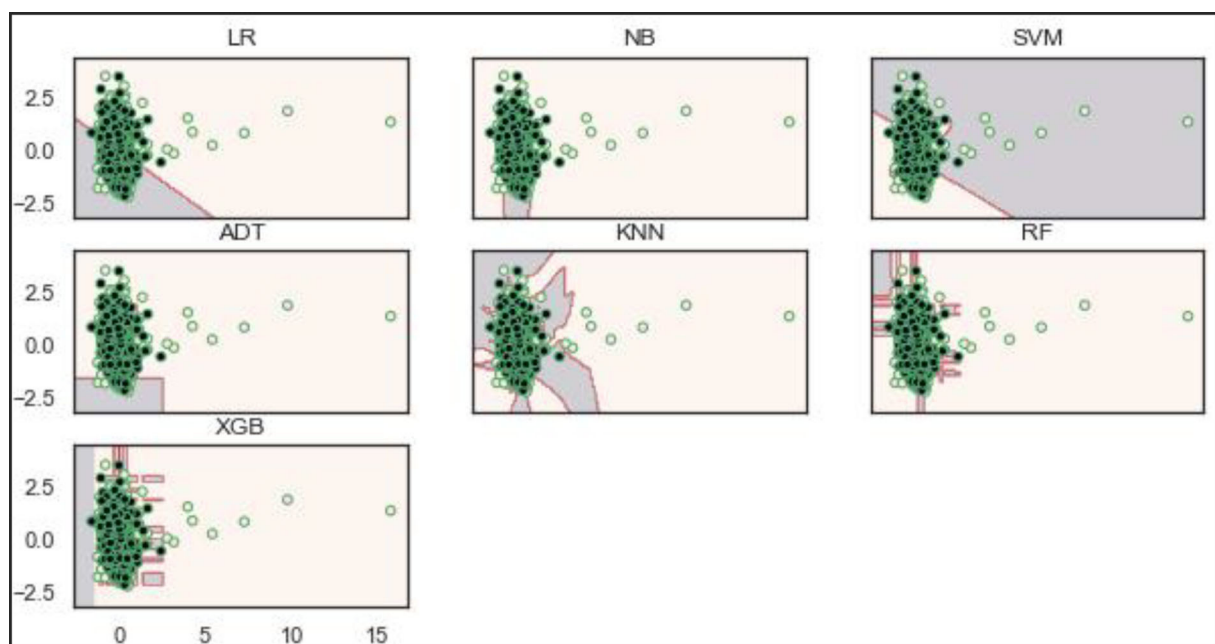


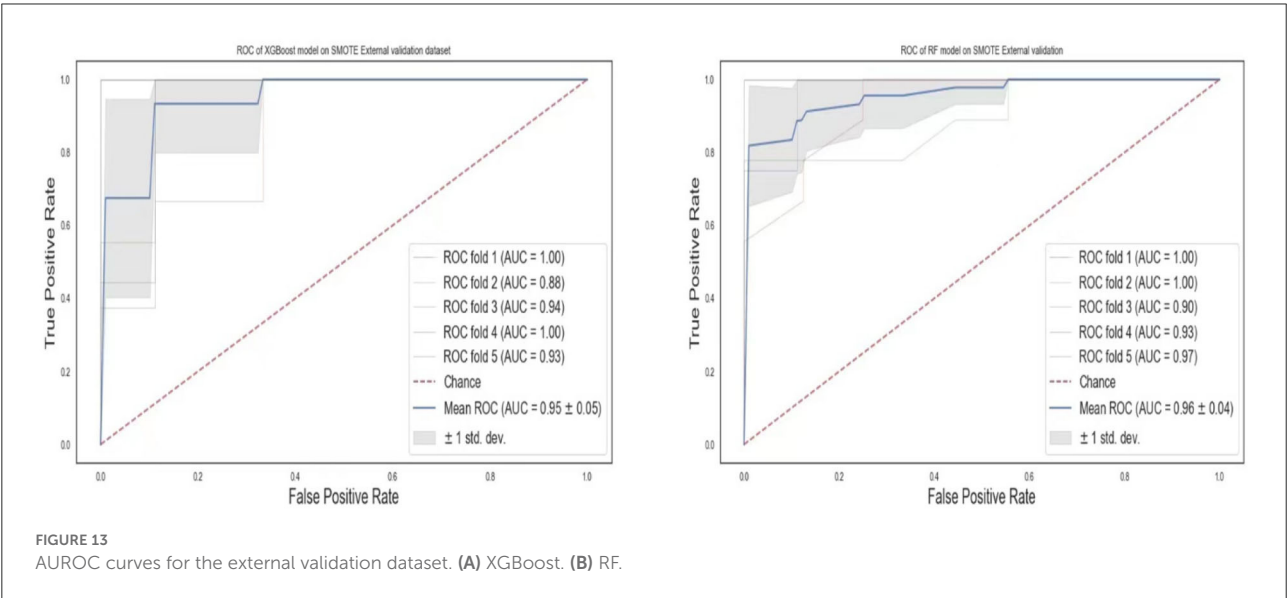
FIGURE 12
Decision boundaries of the models on the balanced dataset.

each iteration to alleviate overfitting and speed up computation. In addition to the RF and XGBoost models, ADT also achieved better performance on the SMOTE balanced data.

The strength of AdaBoost resides in combining weak learners with a powerful learner with a high prediction accuracy based on the adjustments of weights (41). These weights are

TABLE 7 External validation results from the best models.

Model	Accuracy	Precision	Recall	F1-score
RF	88.6	84.8	95.6	89.7
XGBoost	88.7	86.4	93.1	89.3



mainly related to samples that are used by the learner in the training phase. The learners in this phase can generate a set of misclassified samples. AdaBoost tries to resolve this issue by providing appropriate weights for samples that have been wrongly classified. Those samples that are misclassified are assigned a larger weight while samples that are already well classified receive a smaller weight. The unique ability of AdaBoost to spot the misclassified samples, correct them, and re-feed them to the next learner until an accurate predictor model is constructed, makes it one of the best powerful binary classification models. Comparing the results of this study with other studies that used non-invasive measure to build algorithms for disease predictions, we realized that our results were comparable to these studies or even performed better than most studies (Table 8).

Conclusions

This study predicted pneumonia from other LRTIs such as bronchitis using biomarkers, physical indicators, and laboratory parameters. Individual clinical history and symptoms do not have adequate discriminatory power, hence should not be considered in unison during the diagnosis of pneumonia. Two biomarkers, C-reactive protein and procalcitonin, in

TABLE 8 Comparing prediction performance from various studies that used non-invasive measures.

Models	Predicted disease	Performance evaluation	Ref
DT, SVM, LR	Pneumonia	Accuracy-84, 82, 83	(42)
RF, LightGBM, SVM, DT	COVID-19	Accuracy-89, 88, 84, 82	(43)
LogitBoost, RF, DT	Blood diseases	Accuracy-98.2, 97.1, 97	(44)
XGBoost, LightGBM		Accuracy-93, 91	(45)
LR	COVID-19	Specificity-0.95; AUC-0.971; Sensitivity-0.82	(46)
RF, XGBoost	Pneumonia	Accuracy-92, 90.8; AUCs-0.96, 0.97	This study

addition to other features, were considered very important in the prediction of pneumonia. Compared to the SMOTE balanced data, using the original data achieved a higher prediction performance. Therefore, we can conclude that the original dataset was sufficient to predict pneumonia without balancing. RF and XGBoost were considered the best models on both the original dataset and the

SMOTE balanced data. From this, we can conclude that the ensemble ML models performed better in the prediction of pneumonia.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

CE contributed to the conceptualization, study design, data collection, interpretation, and writing of manuscript. RM, ED, RQ, and CA contributed to data collection, literature search, data analysis, and interpretation. YoW contributed to the conceptualization, data analysis, interpretation, writing of the manuscript, fund sourcing, and supervision. LM and YaW contributed to data analysis, interpretation, writing of the manuscript, and supervision. All authors contributed to the article and approved the submitted version.

References

- O'Brien KL, Baggett HC, Brooks WA, Feikin DR, Hammit LL, Higdon MM, et al. Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *Lancet*. (2019) 394:757–79. doi: 10.1016/S0140-6736(19)30721-4
- Peyrani P, Mandell L, Torres A, Tillotson GS. The burden of community-acquired bacterial pneumonia in the era of antibiotic resistance. *Expert Rev Respir Med*. (2019) 13:139–52. doi: 10.1080/17476348.2019.1562339
- Biscevic-Tokic J, Tokic N, Musanovic A. Pneumonia as the most common lower respiratory tract infection. *Med Arch*. (2013) 67:442. doi: 10.5455/medarch.2013.67.442-445
- Zanfardino M, Pane K, Mirabelli P, Salvatore M, Franzese M, TCGA-TCIA. impact on radiogenomics cancer research: a systematic review. *Int J Mol Sci*. (2019) 20:6033. doi: 10.3390/ijms20236033
- World Health Organization pneumonia vaccine trial investigator' group. *Standardization of Interpretation of Chest Radiographs for the Diagnosis of Pneumonia in Children*. (2001). p. 1–39.
- Elemraid MA, Muller M, Spencer DA, Rushton SP, Gorton R, Thomas MF, et al. Accuracy of the interpretation of chest radiographs for the diagnosis of paediatric pneumonia. *PLoS ONE*. (2014) 9:e106051. doi: 10.1371/journal.pone.0106051
- Garber MD, Quinonez RA. Chest radiograph for childhood pneumonia: good, but not good enough. *Pediatrics*. (2018) 142:e20182025. doi: 10.1542/peds.2018-2025
- Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. (2018) 319:1317–8. doi: 10.1001/jama.2017.18391
- Deo RC. Machine learning in medicine. *Circulation*. (2015) 132:1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593
- Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. *NPJ Digit Med*. (2019) 2:1–3. doi: 10.1038/s41746-019-0155-4
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7
- Naydenova E, Tsanas A, Howie S, Casals-Pascual C, De Vos M. The power of data mining in diagnosis of childhood pneumonia. *J R Soc Interface*. (2016) 13:20160266. doi: 10.1098/rsif.2016.0266
- Hao B, Sotudian S, Wang T, Xu T, Hu Y, Gaitanidis A, et al. Early prediction of level-of-care requirements in patients with COVID-19. *Elife*. (2020) 9:e60519. doi: 10.7554/eLife.60519.sa2
- Zhang ZX, Yong Y, Tan WC, Shen L, Ng HS, Fong KY. Prognostic factors for mortality due to pneumonia among adults from different age groups in Singapore and mortality predictions based on PSI and CURB-65. *Singapore Med J*. (2018) 59:190. doi: 10.11622/smedj.2017079
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE synthetic minority over-sampling technique. *J Artif Intell Res*. (2002) 16:321–57. doi: 10.1613/jair.953
- Cheng M, Zhao X, Ding X, Gao J, Xiong S, Ren Y. Prediction of blood culture outcome using hybrid neural network model based on electronic health records. *BMC Med Inform Decis Mak*. (2020) 20:1–0. doi: 10.1186/s12911-020-1113-4
- Ling W, Dyer C, Black AW, Trancoso I. Two/too simple adaptations of word2vec for syntax problems. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver. (2015). pp. 1299–304.
- Bakator M, Radosav D. Deep learning and medical diagnosis: a review of literature. *Multimodal Technol Interact*. (2018) 2:47. doi: 10.3390/mti2030047
- Chen M, Yang J, Zhou J, Hao Y, Zhang J, Youn CH. 5G-smart diabetes: toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Commun Mag*. (2018) 56:16–23. doi: 10.1109/MCOM.2018.1700788

Funding

This work was supported by the National Natural Science Foundation of China (No. 81973099).

Conflict of interest

Author YaW was employed by Center of Health Management, General Hospital of Anyang Iron and Steel Group Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

20. Niederman MS. Imaging for the management of community-acquired pneumonia: what to do if the chest radiograph is clear. *Chest*. (2018) 153:583–5. doi: 10.1016/j.chest.2017.09.045
21. Rambaud-Althaus C, Althaus F, Genton B, D'Acremont V. Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. *Lancet Infect Dis*. (2015) 15:439–50. doi: 10.1016/S1473-3099(15)70017-4
22. Cretikos MA, Bellomo R, Hillman K, Chen J, Finfer S, Flabouris A. Respiratory rate: the neglected vital sign. *Med J Aust*. (2008) 188:657–9. doi: 10.5694/j.1326-5377.2008.tb01825.x
23. Garin N, Marti C, Scheffler M, Stirnemann J, Prendki V. Computed tomography scan contribution to the diagnosis of community-acquired pneumonia. *Curr Opin Pulm Med*. (2019) 25:242. doi: 10.1097/MCP.0000000000000567
24. van Vugt SF, Broekhuizen BD, Lammens C, Zuithoff NP, de Jong PA, Coenen S, et al. Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study. *BMJ*. (2013) 346:f2450. doi: 10.1136/bmj.f2450
25. Schierenberg A, Minnaard MC, Hopstaken RM, Van De Pol AC, Broekhuizen BD, De Wit NJ, et al. External validation of prediction models for pneumonia in primary care patients with lower respiratory tract infection: an individual patient data meta-analysis. *PLoS ONE*. (2016) 11:e0149895. doi: 10.1371/journal.pone.0149895
26. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. *Am J Respir Crit Care Med*. (2019) 200:e45–67. doi: 10.1164/rccm.201908-1581ST
27. Kollef MH. What is ventilator-associated pneumonia and why is it important? *Respir Care*. (2005) 50:714–24. Available online at: <https://rc.rcjournal.com/content/50/6/714>
28. American Thoracic Society, Infectious Diseases Society of America. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. *Am J Respir Crit Care Med*. (2005) 171:388. doi: 10.1164/rccm.200405-644ST
29. Shilpakar R, Paudel BD, Neupane P, Shah A, Acharya B, Dulal S, et al. Procalcitonin and c-reactive protein as markers of bacteremia in patients with febrile neutropenia who receive chemotherapy for acute leukemia: a prospective study from nepal. *J Glob Oncol*. (2019) 5:1–6. doi: 10.1200/JGO.19.00147
30. Schuetz P, Muller B, Christ-Crain M, Stolz D, Tamm M, Bouadma L, et al. Procalcitonin to initiate or discontinue antibiotics in acute respiratory tract infections. *Cochrane Rev J*. (2013) 8:1297–371. doi: 10.1002/ebch.1927
31. Berg P, Lindhardt BØ. The role of procalcitonin in adult patients with community-acquired pneumonia—a systematic review. *Dan Med J*. (2012) 59:A4357.
32. Meili M, Mueller B, Kulkarni P, Schuetz P. Management of patients with respiratory infections in primary care: procalcitonin, C-reactive protein or both? *Expert Rev Respir Med*. (2015) 9:587–601. doi: 10.1586/17476348.2015.1081063
33. Aabenhus R, Jensen JU, Jørgensen KJ, Hróbjartsson A, Bjerrum L. Biomarkers as point-of-care tests to guide prescription of antibiotics in patients with acute respiratory infections in primary care. *Cochrane Database of Syst Rev*. (2014) 11:CD010130. doi: 10.1002/14651858.CD010130.pub2
34. Eccles S, Pincus C, Higgins B, Woodhead M. Diagnosis and management of community and hospital acquired pneumonia in adults: summary of NICE guidance. *Bmj*. (2014) 349:g6722. doi: 10.1136/bmj.g6722
35. Minnaard MC, Van De Pol AC, De Groot JA, De Wit NJ, Hopstaken RM, Van Delft S, et al. The added diagnostic value of five different C-reactive protein point-of-care test devices in detecting pneumonia in primary care: a nested case-control study. *Scand J Clin Lab Invest*. (2015) 75:291–5. doi: 10.3109/00365513.2015.1006136
36. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Ann Surg*. (2020) 272:1133. doi: 10.1097/SLA.0000000000003297
37. Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, et al. Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation*. (2017) 101:e125. doi: 10.1097/TP.0000000000001600
38. Marchese Robinson RL, Palczewska A, Palczewski J, Kidley N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J Chem Inf Model*. (2017) 57:1773–92. doi: 10.1021/acs.jcim.6b00753
39. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451
40. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme gradient boosting as a method for quantitative structure–activity relationships. *J Chem Inf Model*. (2016) 56:2353–60. doi: 10.1021/acs.jcim.6b00591
41. Zhang PB, Yang ZX. A novel AdaBoost framework with robust threshold and structural optimization. *IEEE Trans Cybern*. (2016) 48:64–76. doi: 10.1109/TCYB.2016.2623900
42. Stokes K, Castaldo R, Franzese M, Salvatore M, Fico G, Pokvic LG, et al. A machine learning model for supporting symptom-based referral and diagnosis of bronchitis and pneumonia in limited resource settings. *Biocybernet Biomed Eng*. (2021) 41:1288–302. doi: 10.1016/j.bbe.2021.09.002
43. Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad AK, Uddin S, Kamal AH, et al. Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development. *JMIR Med Inf*. (2021) 9:e25884. doi: 10.2196/25884
44. Alsheref FK, Gomaa WH. Blood diseases detection using classical machine learning algorithms. *Int J Adv Comput Sci Appl*. (2019) 10:77–81. doi: 10.14569/IJACSA.2019.0100712
45. Park DJ, Park MW, Lee H, Kim YJ, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Sci Rep*. (2021) 11:1–1. doi: 10.1038/s41598-021-87171-5
46. Sun NN, Yang Y, Tang LL, Dai YN, Gao HN, Pan HY, et al. A prediction model based on machine learning for diagnosing the early COVID-19 patients. *MedRxiv*. (2020) [preprint]. doi: 10.1101/2020.06.03.20120881



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Thippa Reddy Gadekallu,
VIT University, India
Tae Keun Yoo,
B&VIIT Eye Center, South Korea

*CORRESPONDENCE

Musatafa Abbas Abboud Albadr
mustafa_abbas1988@yahoo.com

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 22 April 2022

ACCEPTED 01 June 2022

PUBLISHED 01 August 2022

CITATION

Albadr MAA, Ayob M, Tiun S,
AL-Dhief FT and Hasan MK (2022) Gray
wolf optimization-extreme learning
machine approach for diabetic
retinopathy detection.
Front. Public Health 10:925901.
doi: 10.3389/fpubh.2022.925901

COPYRIGHT

© 2022 Albadr, Ayob, Tiun, AL-Dhief
and Hasan. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Gray wolf optimization-extreme learning machine approach for diabetic retinopathy detection

Musatafa Abbas Abboud Albadr^{1*}, Masri Ayob¹, Sabrina Tiun¹,
Fahad Taha AL-Dhief² and Mohammad Kamrul Hasan³

¹Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia, ²Department of Communication Engineering, School of Electrical Engineering, Universiti Teknologi Malaysia (UTM) Johor, Bahru, Malaysia, ³Faculty of Information Science and Technology, Center for Cyber Security, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Many works have employed Machine Learning (ML) techniques in the detection of Diabetic Retinopathy (DR), a disease that affects the human eye. However, the accuracy of most DR detection methods still need improvement. Gray Wolf Optimization-Extreme Learning Machine (GWO-ELM) is one of the most popular ML algorithms, and can be considered as an accurate algorithm in the process of classification, but has not been used in solving DR detection. Therefore, this work aims to apply the GWO-ELM classifier and employ one of the most popular features extractions, Histogram of Oriented Gradients-Principal Component Analysis (HOG-PCA), to increase the accuracy of DR detection system. Although the HOG-PCA has been tested in many image processing domains including medical domains, it has not yet been tested in DR. The GWO-ELM can prevent overfitting, solve multi and binary classifications problems, and it performs like a kernel-based Support Vector Machine with a Neural Network structure, whilst the HOG-PCA has the ability to extract the most relevant features with low dimensionality. Therefore, the combination of the GWO-ELM classifier and HOG-PCA features might produce an effective technique for DR classification and features extraction. The proposed GWO-ELM is evaluated based on two different datasets, namely APTOS-2019 and Indian Diabetic Retinopathy Image Dataset (IDRiD), in both binary and multi-class classification. The experiment results have shown an excellent performance of the proposed GWO-ELM model where it achieved an accuracy of 96.21% for multi-class and 99.47% for binary using APTOS-2019 dataset as well as 96.15% for multi-class and 99.04% for binary using IDRiD dataset. This demonstrates that the combination of the GWO-ELM and HOG-PCA is an effective classifier for detecting DR and might be applicable in solving other image data types.

KEYWORDS

gray wolf optimization, extreme learning machine, Histogram of Oriented Gradients, Principal Component Analysis, Diabetic Retinopathy

Introduction

Diabetic Retinopathy (DR) is a condition of the eye that can cause blindness and vision loss in individuals who have diabetes. Regular examination of the eyes is essential for early retinopathy detection in order to decrease the blindness and vision loss caused by DR (1). The core objective of the DR examination is to reveal whether further treatments are required or not (2). Therefore, a robust and accurate retinal examination system is desired to help the screeners to classify the retinal images effectively as well as with high confidence.

Nowadays, Artificial Intelligence (AI) and Machine Learning (ML) techniques are playing significant roles in aiding medical experts with early illness diagnosis (3–6). Therefore, recently, research has been conducted using various AI and ML techniques in order to automatically detect the DR using images (7–9). One of the well-known feature extraction techniques is Histogram of Oriented Gradients (HOG) and has been widely utilized in many image processing fields, including medical fields (10–12). Moreover, the Principal Component Analysis (PCA) is considered one of the most recognized dimensionality reduction techniques (13), where it condenses most of the information in the database into a small dimensions' number. In addition, recently, the Gray Wolf Optimization-Extreme Learning Machine (GWO-ELM) has been considered one of the most popular ML algorithms (14). Therefore, the major aims of this study are as follows:

- Propose a new DR detection approach based on a GWO-ELM classifier and Histogram of Oriented Gradients-Principal Component Analysis (HOG-PCA) features using image data.
- Test the proposed approach based on two different DR image datasets [i.e., APTOS-2019 and Indian Diabetic Retinopathy Image Dataset (IDRiD)] in both binary and multi-class classifications.
- The NN, SVM, Random Forest (RF), and basic ELM classifiers are also implemented in both binary and multi-class classifications using APTOS-2019 and IDRiD datasets.
- Evaluate the performance of the proposed DR detection approach based on several evaluation measures such as accuracy, recall, precision, specificity, F-measure, G-mean, and Matthews Correlation Coefficient (MCC).
- Compare the proposed DR approach against the most recent studies that have used the same datasets in terms of accuracy for the binary and multi-class classifications.

This research is organized as follows: Section 2 presents the related work of this study. Section 3 provides a deep explanation and description of the materials and proposed method. Section 4 discusses the experiments and their outcomes. Section 5 presents

the conclusion of this research as well as recommendations for future research.

Related work

The authors in Sridhar et al. (15) have proposed an automatic system for detecting DR by using Convolutional Neural Network (CNN). The proposed system was tested based on binary classification and used an image dataset that is available on the Kaggle website. The experiments' outcomes have shown that the highest performance of their proposed CNN was achieved with an accuracy of 86%. However, they have tested the proposed system based on binary classification only and ignored the multi-class classification. In addition, the accuracy rate is still not encouraging and needs more enhancement.

Another attempt has been conducted in Gangwar and Ravi (16). They proposed a hybrid architecture of inception-ResNet-v2 and custom CNN layers for the detection of DR. The proposed model was evaluated based on the multi-class classification using APTOS-19 and Messidor-1 dataset. Results showed that the highest accuracy achieved by the proposed model is 72.33% on the Messidor-1 dataset and 82.18% on the APTOS-19 dataset.

One of the most popular ML algorithms is Extreme Learning Machine (ELM); ELM is a single hidden layer feed-forward neural network that consists of three layers (i.e., input, hidden, and output layers) (17, 18). The neurons of the input layer are connected to the neurons of the hidden layer by randomly generated input weights and biases. The neurons of the hidden layer are connected to the neurons of the output layer by output weights. The output weights are calculated based on discovering the least-squares solution (19, 20). ELM is preferred by researchers as it is superior to traditional Support Vector Machine (SVM) and Back Propagation Neural Network (BPNN) (21, 22) specifically in: (1) preventing overfitting, (2) its implementation on multi and binary classifications, and (3) its similar kernel-based capability SVM and working with a Neural Network (NN) structure. These factors make the ELM more efficient in accomplishing a better learning performance. Therefore, some researchers have implemented the ELM algorithm in DR detection. For example, the authors in Asha and Karpagavalli (23) have proposed a DR detection system. The system is based on combining several extracted features such as standard deviation, mean, edge strength, and centroid as well as using the ELM classifier. The system was evaluated based on a binary classification by using the DIARETDB1 dataset which contains 100 images in total. The experiment results showed that the performance of the ELM outperformed both Naive Bayes (NB) and Multilayer Perceptron (MLP) with the highest achieved accuracy reaching up to 90%.

In addition, the authors in Zhang and An (24) have proposed an automatic DR detection system. The proposed system uses

two features extraction methods (i.e., lesion detection and anatomical part recognition) and Kernel Extreme Learning Machine (KELM) with an active learning technique for the classification process. The evaluation of the proposed system has been conducted based on binary classification using the Messidor dataset. The results have shown that the highest performance of the proposed system was achieved with an accuracy of 88.60%.

Further, Punithavathi and Kumar (25) used four different feature extraction techniques (i.e., mean, standard deviation, entropy, and third momentum) and the ELM classifier in order to detect DR. The proposed DR detection system was tested based on a multi-class classification problem using the DIARETDB0 dataset with four different classes. The outcomes of the experiments have proved the superiority of the ELM performance over both BPNN and SVM with the highest achieved accuracy of 95.40%.

Additionally, Deepa et al. (26) proposed a DR detection system that has three different phases. The first phase is to use several micro-macro feature extraction algorithms. The second phase is to apply the Principal Component Analysis (PCA) on the extracted features in order to reduce the dimensionality. Finally, the third phase is to implement the KELM on the extracted features with low dimensions for classification purposes. The proposed system was tested based on a dataset with four classes, which has been collected by the department of medical retina at Bharath hospital in Kottayam. The outcomes of the experiments have demonstrated that the highest achieved accuracy rate of the proposed system reached up to 93.20%.

Although (23–26) showed that the ELM and KELM outperformed their comparatives, these studies have ignored the fact that the random generated input weights and biases of the ELM and KELM need to be optimized. In other words, there is no guarantee that the trained ELM/KELM is the best for carrying out the classification. This drawback can be resolved by integrating the ELM/KELM with an optimisation approach to achieve the optimal input weights and hidden layer biases that guarantee the best ELM/KELM performance (27). Therefore, one of the most popular improvements of the ELM is the Gray Wolf Optimization-Extreme Learning Machine (GWO-ELM), where the GWO is integrated into ELM in order to obtain the best input weights and biases (14). GWO was established by studying the hunting behavior of gray wolves (28). It has a simple concept with easy implementation, requiring very few coding lines, allowing many to leverage from it. In comparison to other evolutionary algorithms, GWO is highly robust in regulating parameters with greater computational efficacy (29, 30). The effectiveness of this integration (GWO-ELM) has been proven in many domains including breast cancer diagnosis (31), poison diagnosis (32), lung cancer classification (33), identification of cardiovascular disease (34), electricity load projections (35), bankruptcy predictions (36), and paraquat poisoned patients diagnosis (37). However, to the best of our

knowledge, no research has used the GWO-ELM classifier in the detection of DR. Therefore, this study aims to employ the GWO-ELM classifier for detecting DR. Table 1 provides a summary of the previous DR detection works using ML and deep learning techniques.

Materials and proposed method

The general diagram of the proposed work using the GWO-ELM method is demonstrated in Figure 1. The diagram consists of various stages which will be used to create the DR detection approach based on images. The first stage refers to the image dataset that contains five categorizations (i.e., no DR, mild, moderate, severe, and proliferative DR). While, in the second stage, the pre-processing operation will be used in order to prepare the images for the next stage, which is the features extraction stage. In addition, in the third stage, the HOG-PCA method will be utilized in order to extract the needed features from images. Lastly, in the fourth stage, the HOG-PCA extracted features will be fed into the GWO-ELM classifier in order to detect DR based on images. These fourth stages of the proposed DR detection approach will be deliberated as sub-sections, respectively.

Image dataset

In this study, two different datasets will be used in order to evaluate the proposed DR detection approach. The first dataset is APTOS-2019 while the second dataset is IDRiD. The description of both datasets APTOS-2019 and IDRiD are provided as follow:

- APTOS-2019 Dataset has been provided by an Indian hospital, Aravind Eye Hospital. The APTOS-2019 dataset is available online in Hospital (39). In this study, the dataset consists of five main classes, which are no DR, mild, moderate, severe, and proliferative DR, and each class contains 190 images. Thus, 950 is the total number of images in the whole dataset. In this study, 80% of the dataset, which equals 760 images, were used for training purposes, whilst 20% of the dataset, which equals 190 images, were used for testing purposes. In other words, 152 images from each class were used for training purposes whilst the remaining 38 images were used for testing purposes. The description of the APTOS-2019 dataset which is used in this study is provided in Table 2.
- IDRiD is a DR image dataset that is available online at (40). The IDRiD dataset consists of five main classes, which are no DR, mild, moderate, severe, and proliferative DR. In addition, the IDRiD dataset has a total number of images equal to 516 and each class contains a different number of images. In this study, 80% of the dataset that equals

TABLE 1 Illustrates the previous works of DR detection using ML and deep learning techniques.

References	Dataset	Classification mode	Classifier	Results	Disadvantages
Sridhar et al. (15)	Kaggle dataset	Binary classification	CNN	86% Accuracy	<ul style="list-style-type: none"> • The proposed system tested based on binary classification only and ignored the multi-class classification. • The accuracy rate is still not encouraging and needs more enhancement.
Gangwar and Ravi. (16)	APTOS-19 and Messidor-1	Multi-class classification	Hybrid CNN	72.33% Accuracy on the Messidor-1 dataset and 82.18% accuracy on the APTOS-19 dataset.	<ul style="list-style-type: none"> • The evaluation of both systems considered only the multi-class classification and ignored the binary classification. • The accuracies of both systems are still not promising and need more improvement.
Reddy et al. (38)	Messidor	Multi-class classification	SVM	69.09% Accuracy	
Asha and Karpagavalli. (23)	DIARETDB1	Binary classification	ELM	90% Accuracy	<ul style="list-style-type: none"> • The proposed system tested based on binary classification only and ignored the multi-class classification. • The accuracy rates are still not encouraging and need more enhancement. • These studies have ignored the fact that the random generated input weights and biases of the ELM and KELM need to be optimized.
Zhang and An (24)	Messidor	Binary classification	KELM	88.60% Accuracy	
Punithavathi and Kumar (25)	DIARETDB0	Multi-class classification	ELM	95.40% Accuracy	<ul style="list-style-type: none"> • The evaluation of both systems considered only the multi-class classification and ignored the binary classification. • The accuracy rates are still not encouraging and need more enhancement. • These studies have ignored the fact that the random generated input weights and biases of the ELM and KELM need to be optimized.
Deepa et al. (26)	4 classes dataset	Multi-class classification	KELM	93.20%	

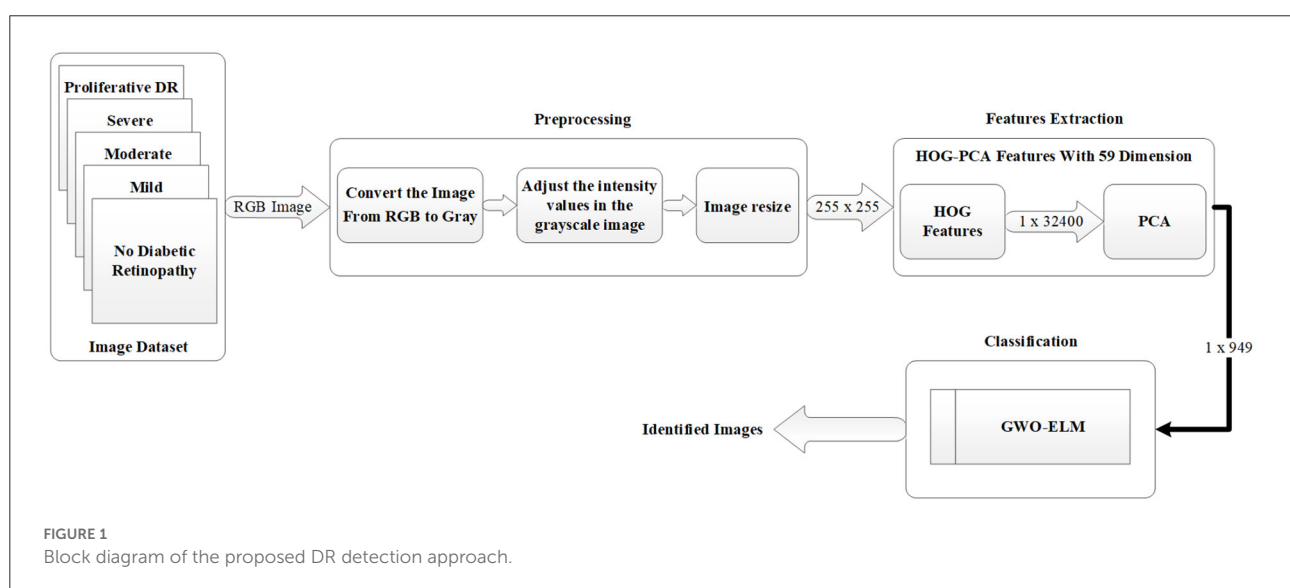


TABLE 2 The description of the APTOS-2019 dataset.

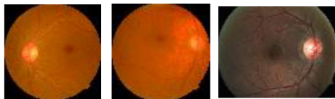
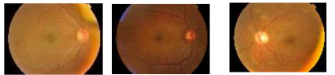
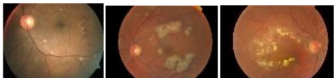
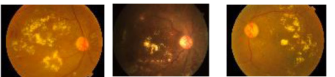

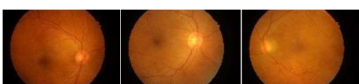

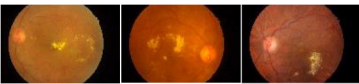
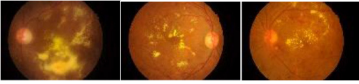
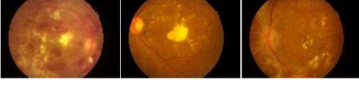
Class	Number of image	Samples of the dataset	Class label
No DR	190		1
Mild	190		2
Moderate	190		3
Severe	190		4
Proliferative DR (PDR)	190		5

TABLE 3 The description of the IDRiD dataset.

Class	Number of image	Samples of the dataset	Class label
No DR	168		1
Mild	25		2
Moderate	168		3
Severe	93		4
Proliferative DR (PDR)	62		5

412 images were used for training purposes, whilst the remaining 20% of the dataset which equals 104 images were used for testing purposes. The description of the IDRiD dataset which is used in this study is provided in Table 3.

Pre-processing

This section discusses the pre-processing of this study, which consists of four steps. The first step is to read the RGB image that will be as an array with three dimensions. The second step

is to convert the image from RGB to Grayscale, which will lead to making it an array with two dimensions. The third step is to adjust the intensity values in the grayscale image which leads to an increase in the contrast of the output image. Finally, the fourth step is to re-size the dimensionality of the image to (255×255) which will be as an input into the features extraction approach. Figure 2 depicts an example of the pre-processing steps which are used in this study.

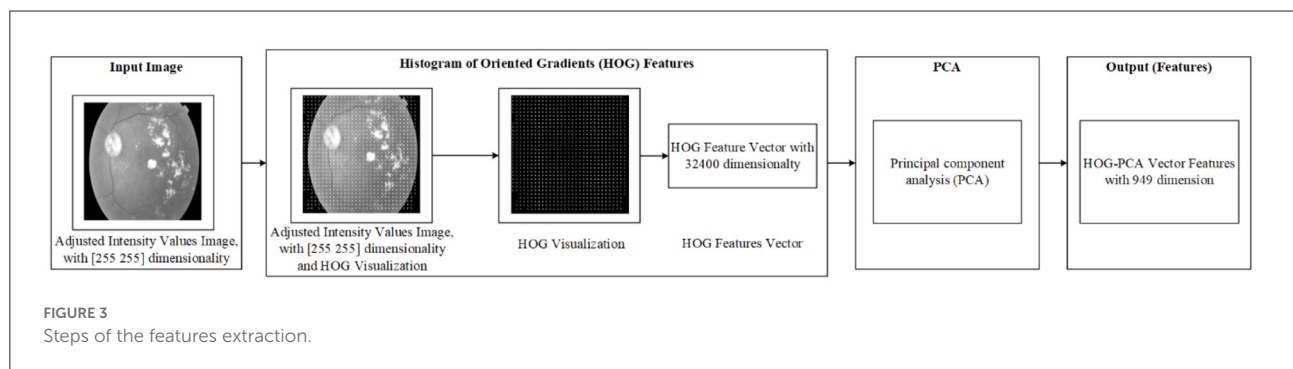
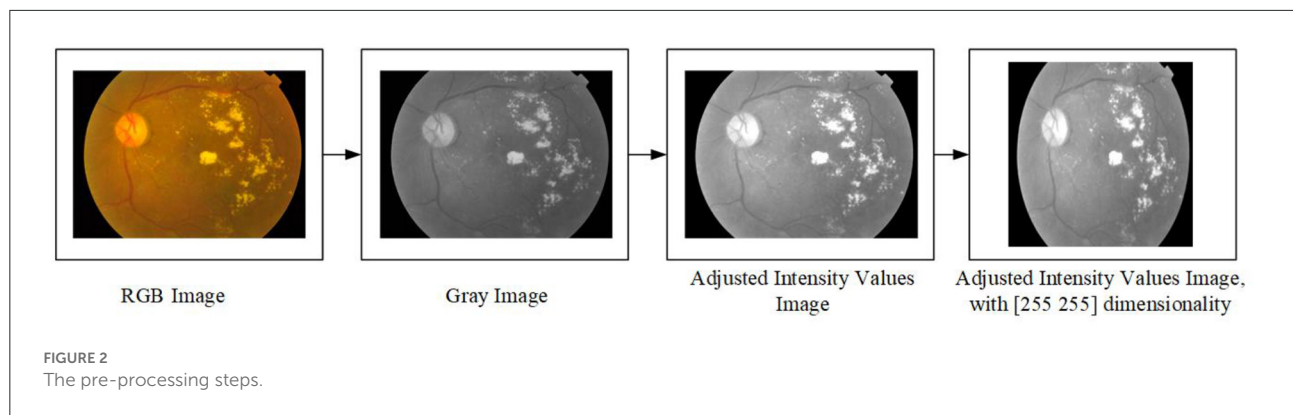
Features extraction

In this work, the required features were calculated in two steps. The first step is to use the output of the pre-processing as an input into the Histogram of Oriented Gradients (HOG) features extraction technique, which begins after the pre-processing phase. HOG is considered as one of the most popular features extraction techniques that has been widely utilized in many image processing fields, including medical fields (10–12). The output of the HOG features extraction approach is a vector with the dimensionality of $(1 \times 32,400)$ per image, and $(950 \times 32,400)$ and $(516 \times 32,400)$ for whole APTOS-2019 and IDRiD dataset, respectively.

Whilst the second stage is to reduce the dimensionality of HOG features using Principal Component Analysis (PCA). PCA is considered one of the most recognized dimensionality reduction techniques (13), where it condenses most of the information in the database into a small dimensions' number. The aim of that is to reduce the high dimensionality of the HOG features from $(950 \times 32,400)$ to (950×949) for whole APTOS-2019 dataset and from $(516 \times 32,400)$ to (516×515) for whole IDRiD dataset. This enables the issue of limited resources (i.e., requiring a large memory space) to be overcome. Literature has addressed the issue that the required memory space is affected by the dimensionality of the features (i.e., number of features). In other words, the higher dimensionality requires a large memory space (41–43). The final output of the features extraction is the HOG-PCA features with (950×949) dimensionality for whole APTOS-2019 dataset and (516×515) for whole IDRiD dataset, both of which will be used as inputs into the classification step. Figure 3 demonstrates the steps of the features extraction in more detail. Further, Table 4 demonstrates the dimensionality of the features extraction steps for a single image and whole dataset images.

Classification

This section provides a deep explanation of both GWO and GWO-ELM approaches separately. The explanation of the GWO approach is delivered in Section 2.4.1, while the explanation of the GWO-ELM approach is presented in Section 2.4.2.



Gray wolf optimization

In recent years, GWO has emerged as a prominent new nature-based metaheuristic algorithm and population-oriented metaheuristic (30). GWO is based on the natural behaviors of the gray wolf (28). The algorithm fundamentally simulates the wolf's social behavior and hunting mechanisms. In GWO, the wolves (search agents) are classified as alpha (α), beta (β), delta (δ), and omega (ω). α is the fittest wolf or the best solution. β and δ each denote the second and third best wolves. Meanwhile, ω denotes the other wolves in the population. Finding the prey (process of optimization) is spearheaded by δ , β , and α whilst the wolves (ω) are the followers. When surrounding the prey, wolves inform about their positions based on δ , β , or α using the following equations (28):

$$D = |C \cdot X_p(it) - X(it)| \quad (1)$$

and

$$X(it+1) = X_p(it) - A \cdot D \quad (2)$$

Where, it denotes the present iteration number. $X_p(it)$ denotes the present position of the prey. $X(it)$ denotes the wolf's present position. D denotes the distance between the prey and wolf. Below are the mathematical formulas for coefficient vectors (A and C):

$$A = 2a \cdot r_1 - a \quad (3)$$

and

$$C = 2 \cdot r_2 \quad (4)$$

Where r_1 and r_2 are the two vectors that are randomly generated between 1 and 0. " a " denotes linear decrement from 2 to 0 as the iterations number increase. The simulation of the wolves' hunting behaviors results in the saving of the first three top values as α , β , and δ . Below is the formula for updating the position of the gray wolf population:

$$\begin{cases} D_\alpha = |C_1 \cdot X_\alpha - X| \\ D_\beta = |C_2 \cdot X_\beta - X| \\ D_\delta = |C_3 \cdot X_\delta - X| \end{cases} \quad (5)$$

$$\begin{cases} X_1 = X_\alpha(it) - A_1 \cdot D_\alpha \\ X_2 = X_\beta(it) - A_2 \cdot D_\beta \\ X_3 = X_\delta(it) - A_3 \cdot D_\delta \end{cases} \quad (6)$$

and

$$X(t+1) = \frac{X_1 + X_2 + X_3}{3} \quad (7)$$

Where X_α , X_β , and X_δ denote the positions of α , β , and δ , respectively. X denotes the current wolf position. C_1 , C_2 , and C_3 are vectors that are randomly generated based on Equation (4). Equation (5) is used to calculate the estimated distances among the current wolf and α , β , and δ , whilst Equations (6) and (7) are used to determine the current wolf's final position. A_1 , A_2 ,

TABLE 4 Elaborate the features extraction step dimensionality for single image and whole dataset images.

APTOS-2019 Dataset		
Features Extraction	Dimensionality of a single image	Dimensionality of the whole dataset
First Step: HOG Features	(1 x 32400)	(950 x 23400)
Second Step: HOG-PCA Features	(1 x 949)	(950 x 949)
IDRiD Dataset		
Features extraction	Dimensionality of a single image	Dimensionality of the whole dataset
First Step: HOG Features	(1 x 32400)	(516 x 32400)
Second Step: HOG-PCA Features	(1 x 515)	(516 x 515)

and A_3 are vectors that randomly generated using Equation (3). it represents the iterations number.

This updating mechanism facilitates the omega wolves in reaching new stochastic places (presumed to be nearer to the prey) in the circle delineated by the leading wolves' positions. GWO is distinguished by its strategy in managing the explorations and exploitations in the search process. With a decrease from 2 to 0 during the iterations, the algorithm progressively moves on from underlining the process of exploration to the process of exploitation (30). Figure 4 shows the GWO algorithm flowchart. Below are the general processing steps of the GWO algorithm (28):

- Parameters of the gray wolf, such as population size or the number of search agents (NSA), are initialized. For the following steps, the search agent term refers to a wolf, position of each wolf (search agent), maximum number of iteration (it_{max}), and upper and lower bound of search.
- Set the iteration counter $it = 0$.
- Initialize the coefficient vectors "A, and C" using Equations (3 and 4) while the initialization of "a", which is the linear decrements from 2 to 0 as the iterations number increase, uses $a = 2 - it * ((2) / it_{max})$.
- Calculate the fitness for all search agents and set the first three best search agents as X_α , X_β , and X_δ where X_α denotes the first best search agent whilst X_β denotes the second best search agent, and X_δ denotes the third best search agent.

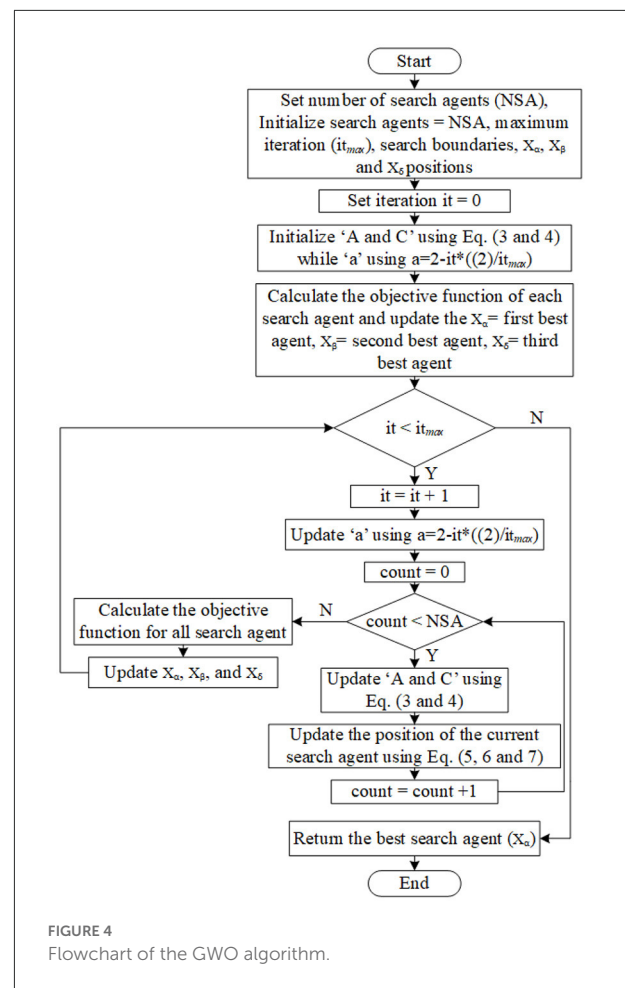
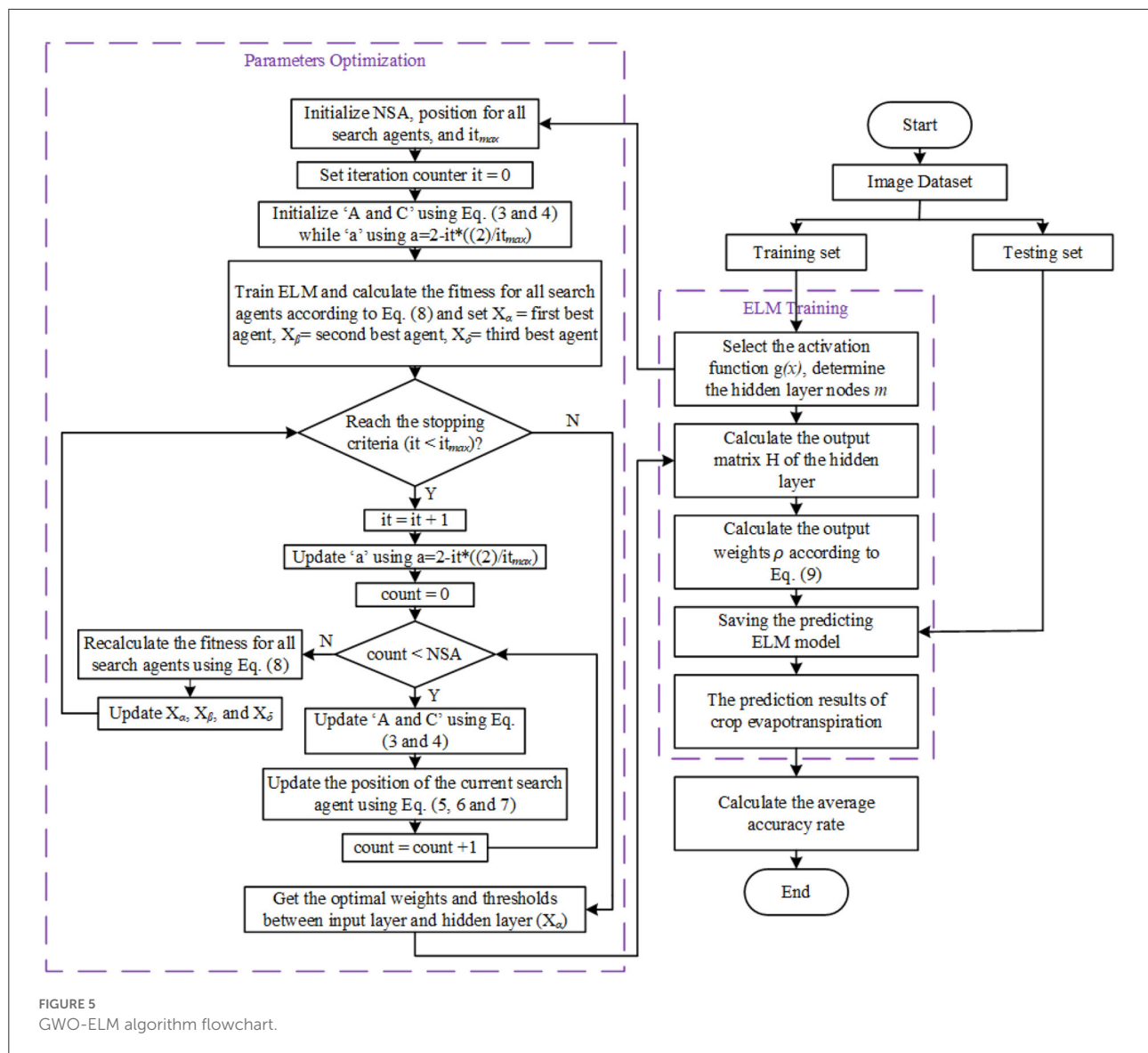


FIGURE 4
Flowchart of the GWO algorithm.

- Increase the iteration counter $it = it + 1$.
- Update "A, and C" using Equations (3 and 4) while "a" using $a = 2 - it * ((2) / it_{max})$.
- Update the position of all current search agents using Equations (5 and 6).
- Recalculate the fitness for all search agents.
- If any better search agent is found, then update the best agents X_α , X_β , X_δ .
- Repeat steps from "e" if the stopping criteria are not satisfied.
- The best-calculated optimum (best search agent) will be returned as X_α .

GWO-ELM

The GWO-ELM follows the GWO concept in Mirjalili et al. (28). It adjusts the input weight values and the biases of the hidden nodes by updating the GWO parameters toward achieving greater accuracy. The GWO-ELM steps are presented below while the flowchart is illustrated in Figure 5. Table 5 shows the ELM-GWO parameter settings.



Let N be the number of training samples and (X_j, t_j) refer to a single sample of the training samples.,

Where:

X_j is the input extracted from HOG-PCA features where $X_j = [x_{j1}, x_{j2}, \dots, x_{jn}]^T \in R^n$,

t_j is the expected output (true value) where $t_j = [t_{j1}, t_{j2}, \dots, t_{jm}]^T \in R^m$.

Step 1: Random initialization of the gray wolf population (position of all search agents) within the range of $[-1, 1]$ for the values of the input weights, and $[0, 1]$ for the hidden nodes' bias. Ascertaining the initial GWO parameters entails: 1) the population size or number of search agents (NSA), 2) the maximum number of iterations (it_{max}), and 3) the iteration counter $it = 0$. Each wolf (search agent) in the population is reshaped using the following form:

$$SA_i = \begin{Bmatrix} w_{11}, w_{12}, \dots, w_{1n}, w_{21}, w_{22}, \dots \\ w_{2n}, w_{L1}, w_{L2}, \dots, w_{Ln}, b_1, \dots, b_L \end{Bmatrix}$$

Where:

w_{ij} = value of input-weights which connect between the i_{th} hidden node and j_{th} input node, $w_{ij} \in [-1, 1]$.

$b_i = i_{th}$ hidden node's bias, $b_i \in [0, 1]$.

n = number of the input-nodes.

L = number of the hidden nodes.

$L \times (1+n)$ denotes the dimension of the search agent, which therefore requires optimization of its parameters.

Step 2: Initialization of the coefficient vectors 'A, and C' using Equations (3 and 4) while the initialization of the 'a' which is the linear decrements from 2 to 0 as the

TABLE 5 The parameters settings for the ELM and GWO.

ELM		GWO	
Parameter	Value	Parameter	Value
AS	assemble of the biases and input weights	Population (wolves or search agents)	Consists of the position of all search agents
ρ	Output-weights matrix	Position	Start stochastically generated within the range of $[-1, 1]$ for the input-weights and $[0, 1]$ for the biases
Input-weights (w)	-1 to 1	Population size or number of search agents (NSA)	50
Bias values (b)	0 to 1	r_1 and r_2	Stochastically generated with the range of $[0, 1]$
Input-nodes number (n)	Input attributes	Number of iterations it_{max}	100
Hidden-nodes number (L)	$[100-300]$; with a 25 increment step	C_1, C_2 , and C_3	Randomly generated vectors based on Equation (4)
Output neurones number (m)	Number of classes	A_1, A_2 , and A_3	Randomly generated vectors using Equation (3)
Activation function	Sigmoid	X_α	Best position of all search agents.

iterations number increase, using $a = 2 - it^* ((2) / it_{max})$.

Step 3: Division of the dataset into training and testing sets
Set the hidden layer nodes as m , and choose a suitable activation function $g(x)$ for ELM;

$$f(X) = \sqrt{\frac{\sum_{j=1}^N \|\sum_{i=1}^L \rho_i g(w_i x_j + b_i) - t_j\|_2^2}{N}} \quad (8)$$

Where:

ρ = output weight matrix;
 t_j = true value; and
 N = number of training samples.

Where:

$$\rho = H^\dagger T \quad (9)$$

$$H = \begin{bmatrix} g(w_1.X_1 + b_1) & \cdots & g(w_L.X_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(w_1.X_N + b_1) & \cdots & g(w_L.X_N + b_L) \end{bmatrix}_{N \times L} \quad (10)$$

$$\rho = \begin{bmatrix} \rho_1^T \\ \vdots \\ \rho_L^T \end{bmatrix}_{L \times m} \quad \text{and} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m}$$

H in Equation (10) is the hidden layer output matrix of the ELM network; in H , the i_{th} column is indicated to the i_{th} hidden layer neuron on the input neurons. While the H^\dagger is the Moore–Penrose generalized inverse of H . The activation function g is infinitely distinguishable when the desired number of hidden neurons is $L \leq N$.

Step 4: Train the ELM and evaluate the fitness value of each search agent according to the accuracy of the classification.

Step 5: Based on the fitness values of each search agent, set the first three best search agents as X_α , X_β , and X_δ , where X_α denotes the first best search agent whilst X_β denotes the second best search agent, and X_δ denotes the third best search agent.

Step 6: Increase the iteration counter $it = it + 1$.

Step 7: Update ‘A, and C’ using Equations (3 and 4) while ‘a’ using $a = 2 - it^* ((2) / it_{max})$.

Step 8: Update the position of all current search agents using Equations (5–7).

Step 9: Recalculate the fitness for all search agents using Equation (8).

Step 10: If any better search agent is found, then update the best agents X_α , X_β , X_δ .

Step 11: Repeat steps from step 6 if the stopping criteria are not satisfied, or else save the optimal weights and thresholds between input layers and hidden layers (X_α).

Step 12: The results of GWO are utilized as input-weights and hidden-layer biases of the ELM, calculating the hidden layer output matrix (H) via the activation function $g(x)$;

Step 13: Calculate the output-weights (ρ) according to Equation (9) and save the forecasting ELM model for testing.

Experiments and results

The proposed GWO-ELM approach was utilized in both binary and multi-class classification experiments with a hidden neurons number in a range of $[100-300]$ and increment steps of 25. In the multi-class classification experiments, we have used both APTOS-2019 and IDRiD datasets in order to classify five

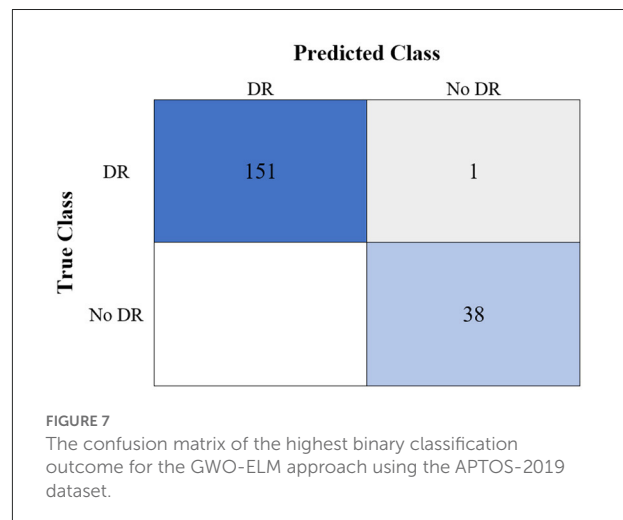
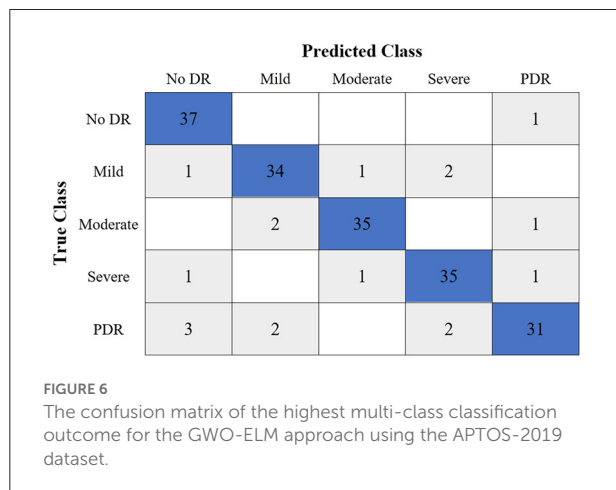
TABLE 6 The highest experiment outcomes of the binary and multi-class classifications for GWO-ELM approach using APTOS-2019 and IDRiD datasets.

APTOS-2019 dataset

Number of class	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
5	96.21	90.53	90.53	97.63	88.16	90.53	90.53
2	99.47	99.34	100.00	97.44	98.38	99.67	99.67

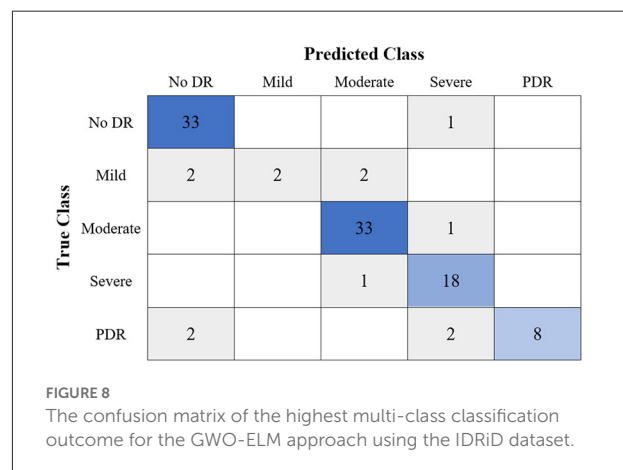
IDRiD dataset

Number of class	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
5	96.15	90.38	90.38	97.60	87.98	90.38	90.38
2	99.04	100.00	98.59	100.00	97.82	99.29	99.29

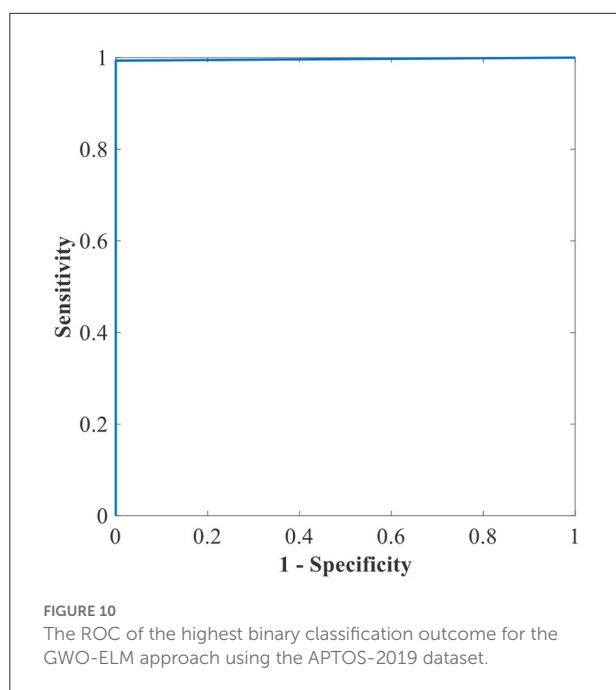
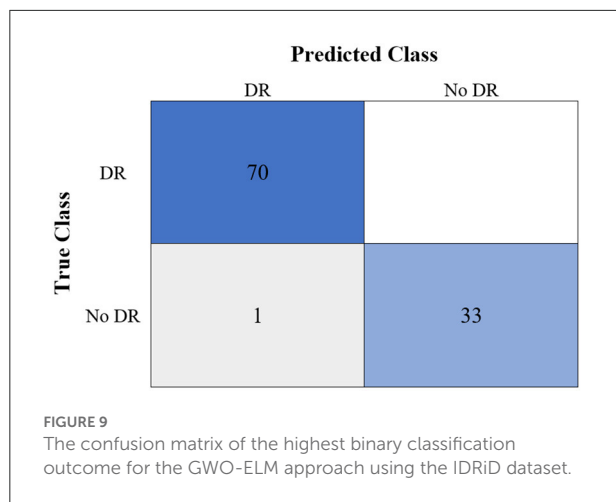


different classes, namely no DR, mild, moderate, severe, and proliferative DR. In the binary classification experiments, we have used both APTOS-2019 and IDRiD datasets in order to classify two different classes (i.e., no DR and DR). The class of DR was obtained by combining mild, moderate, severe, and proliferative DR classes. Hence, the total number of both binary and multi-class classification experiments for the GWO-ELM approach is 36, and each experiment has 100 iterations. All the experiments have been applied based on using 80% of the dataset as a training dataset and the remaining 20% as a testing dataset. In addition, it is worth mentioning that all the experiments have been implemented in MATLAB R2019a programming language over a PC Core i7 of 3.20 GHz with 16 GB RAM and SSD 1 TB (Windows 10).

In this study, numerous evaluation measurements were utilized to evaluate the proposed approach GWO-ELM. The evaluation measurements rely on the ground truth, which entails the application of the model to expect the answer on the evaluation dataset followed by a comparison between the predicted target and the actual answer. The evaluation



measurements have been used in order to evaluate the proposed GWO-ELM approach regarding True Positive (TP), True Negative (TN), False Positive (FP), False Negative



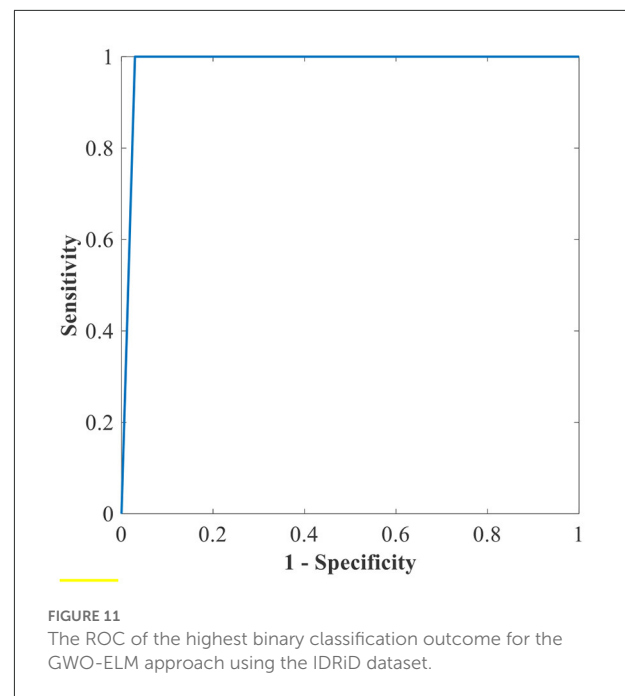
(FN), recall, accuracy, specificity, G-mean, precision, F-measure, and MCC. Equations (11–17) (44–46) depict these evaluation measurements.

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (11)$$

$$precision = \frac{TP}{TP + FP} \quad (12)$$

$$recall = \frac{TP}{TP + FN} \quad (13)$$

$$F - Measure = \frac{(2 \times precision \times recall)}{(precision + recall)} \quad (14)$$



$$G - Mean = \sqrt[3]{recall \times precision} \quad (15)$$

$$Specificity = \frac{TN}{TN + FP} \quad (16)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (17)$$

Table 6 shows the highest outcomes of the binary and multi-class classification experiments that have been conducted using the proposed GWO-ELM approach based on both datasets APTOS-2019 and IDRiD. Table 6 presents the evaluation outcomes of the GWO-ELM in terms of recall, accuracy, specificity, G-mean, precision, F-measure, and MCC. The highest achieved multi-class classification accuracies of the GWO-ELM approach were 96.21% and 96.15% using APTOS-2019 and IDRiD datasets, respectively. Whilst the highest achieved binary classification accuracies of the GWO-ELM approach were 99.47% using the APTOS-2019 dataset and 99.04% using the IDRiD dataset. In addition, Figures 6–10 show the confusion matrices for the highest outcomes of the binary and multi-class classification using the GWO-ELM approach based on both datasets APTOS-2019 and IDRiD. Further, Figures 10, 11 present the ROC of the best binary classification outcome for the GWO-ELM approach using the APTOS-2019 and IDRiD datasets.

Further, additional experiments have been implemented utilizing feedforward NN and basic ELM as classifiers and HOG-PCA features to perform binary and multi-class classification of the DR. Both classifiers NN and basic ELM were implemented in binary and multi-class classifications when varying the number of the hidden nodes in the range of [100–300] and

TABLE 7 The highest experiment outcomes of the binary and multi-class classifications for ELM approach using APTOS-2019 and IDRiD datasets.

APTOS-2019 dataset							
Number of class	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
5	80.21	50.53	50.53	87.63	38.16	50.53	50.53
2	92.63	93.42	93.42	77.27	78.60	95.30	95.32
IDRiD dataset							
Number of class	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
5	74.62	36.54	36.54	84.13	20.67	36.54	36.54
2	72.12	85.71	75.95	60.00	32.75	80.54	80.68

TABLE 8 The highest experiments outcomes of the classification and detection for NN approach using APTOS-2019 and IDRiD datasets.

APTOS-2019 dataset							
Number of class	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
5	78.53	46.32	46.32	86.58	32.89	46.32	46.32
2	90.53	98.68	90.36	91.67	68.13	94.34	94.43
IDRiD dataset							
Number of class	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
5	72.31	30.77	30.77	82.69	13.46	30.77	30.77
2	71.15	97.14	70.83	75.00	26.04	81.93	82.95

increment steps of 25. Tables 7, 8 provide the highest binary and multi-class classification experiments outcomes of the NN and ELM classifiers using both APTOS-2019 and IDRiD datasets. The best performance of the basic ELM in multi-class classification has been obtained with an accuracy of 80.21% and 74.62% for APTOS-2019 and IDRiD datasets, respectively. While the highest performance of the basic ELM in binary classification has acquired an accuracy of 92.63% using APTOS-2019 dataset and 72.12% using IDRiD dataset. Furthermore, the best achieved multi-class classification accuracies of the NN approach were 78.53% and 72.31% using APTOS-2019 and IDRiD datasets, respectively. The highest achieved binary classification accuracies of the NN approach were 90.53% using the APTOS-2019 dataset and 71.15% using the IDRiD dataset.

Moreover, further experiments have been conducted utilizing SVM (linear kernel), SVM (precomputed kernel), and RF as classifiers and HOG-PCA features to perform binary and multi-class classifications of the DR. Table 9 provides the outcomes of the binary and multi-class classification experiments for the SVM (linear kernel), SVM (precomputed

kernel), and RF classifiers using both APTOS-2019 and IDRiD datasets. In multi-class classification and when using APTOS-2019 dataset, the best performance of the SVM (linear) was achieved with an accuracy of 79.58% while the highest performance of the SVM (precomputed kernel) and RF classifiers was equal with an accuracy of 79.37%. Moreover, in binary classification and when using APTOS-2019 dataset, the best performance of the SVM (linear) and SVM (precomputed kernel) was equal and achieved an accuracy of 88.95% while the highest performance of RF classifier was achieved with an accuracy of 91.58%. Additionally, in multi-class classification and using IDRiD dataset, the best performance of the SVM (linear), SVM (precomputed kernel), and RF classifiers was achieved with an accuracy of 73.85, 73.08, and 74.23%, respectively. In binary classification and using IDRiD dataset, the highest performance of the SVM (linear), SVM (precomputed kernel), and RF classifiers was achieved with an accuracy of 68.27, 67.31, and 69.23%, respectively.

The outcomes for binary and multi-class classification are shown in Tables 6–9. The performance of the GWO-ELM

TABLE 9 The experiments outcomes of the binary and multi-class classification for SVM (linear kernel), SVM (precomputed kernel), and RF approaches using APTOS-2019 and IDRiD datasets.

APTOS-2019 dataset with 5 classes							
Classifier	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
SVM (linear)	79.58	48.95	48.95	87.24	36.18	48.95	48.95
SVM (Precomputed Kernel)	79.37	48.42	48.42	87.11	35.53	48.42	48.42
RF	79.37	48.42	48.42	87.11	35.53	48.42	48.42
APTOS-2019 dataset with 2 classes							
Classifier	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
SVM (linear)	88.95	100.00	87.86	100.00	62.69	93.54	93.73
SVM (Precomputed Kernel)	88.95	100.00	87.86	100.00	62.69	93.54	93.73
RF	91.58	100.00	90.48	100.00	72.37	95.00	95.12
IDRiD dataset with 5 classes							
Classifier	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
SVM (linear)	73.85	34.62	34.62	83.65	18.27	34.62	34.62
SVM (Precomputed Kernel)	73.08	32.69	32.69	83.17	15.87	32.69	32.69
RF	74.23	35.58	35.58	83.89	19.47	35.58	35.58
IDRiD dataset with 2 classes							
Classifier	Accuracy	Precision	Recall	Specificity	MCC	F-measure	G-mean
SVM (linear)	68.27	98.57	68.32	66.67	12.48	80.70	82.06
SVM (Precomputed Kernel)	67.31	84.29	71.95	50.00	19.11	77.63	77.87
RF	69.23	100.00	68.63	100.00	20.09	81.40	82.84

approach outperformed the NN, ELM, SVM (linear kernel), SVM (precomputed kernel), and RF in all experiments. This discovery confirms that generating the appropriate weights and biases for the ELM's single hidden layer decreases classification errors. In other words, avoiding inappropriate weights and biases prevents the ELM algorithm from becoming stuck in the local maxima of the weights and biases. Consequently, the performances of the proposed GWO-ELM approach in the multi-class and binary classification were impressive and achieved an accuracy of 96.21, 99.47, 96.15, and 99.04% using APTOS-2019 and IDRiD datasets, respectively. This research confirms that the combination of the GWO-ELM classifier with HOG-PCA features is an effective approach for detecting the DR using retinal images which could help physicians in easily screening for DR.

Furthermore, the proposed GWO-ELM technique is compared with some recent works (47–65) in terms of accuracy

based on binary and multi-class classifications using APTOS-2019 and IDRiD datasets. Table 10 exhibits the comparison accuracy results of the proposed GWO-ELM and some other previous works.

Based on all the results in Table 10, it is clear that the performance of the GWO-ELM outperformed all the other previous works in binary and multi-class classifications using both datasets APTOS-2019 and IDRiD. This suggests that the proposed GWO-ELM is a reliable technique for the detection of DR when using image data. Although the proposed method has shown a good performance, there are some limitations which are provided as follows:

- The image datasets which have been used in this study for the training and testing purposes are small.
- The evaluations of this study did not consider the execution time measurement of the proposed GWO-ELM approach.

TABLE 10 The comparison of accuracy between the proposed GWO-ELM and other previous works.

Accuracy results based on APTOS-2019 dataset with 5 classes		Accuracy results based on APTOS-2019 dataset with 2 classes	
Method	Accuracy	Method	Accuracy
DNN (50)	81.70	DNN (50)	97.41
Hybrid model (56)	86.34	DNN (51)	98.00
DNN (51)	82.54	Hybrid CNN-SVD and ELM (57)	99.32
V-SVM (52)	77.90	Ensemble (trimmed mean) (61)	98.60
MLP (55)	83.09	ResNet34 (47)	96.35
CNN512 (48)	89.00	CNN (62)	91.00
Tuned XGBoost (59)	94.20	RA-EfficientNet (64)	98.36
Proposed GWO-ELM	96.21	Proposed GWO-ELM	99.47

Accuracy results based on IDRiD dataset with 5 classes		Accuracy results based on IDRiD dataset with 2 classes	
Method	Accuracy	Method	Accuracy
MLP (53)	92.01	MLP (53)	98.87
ResNet50 + J48 (54)	92.46	CNN (58)	90.29
XG-Boost (49)	88.20	Coarse Network (63)	80.00
Lesion(Semi + Adv) (65)	91.34	HE-CNN (60)	96.76
Proposed GWO-ELM	96.15	Proposed GWO-ELM	99.04

- The current study has considered only the off-line aspect for detecting DR.

Conclusion

In this study, we have proposed a DR detection approach based on HOG-PCA features and GWO-ELM classifier. The GWO-ELM classifier underwent evaluations using the APTOS-2019 and IDRiD datasets. The outcomes indicated the superiority of the GWO-ELM over the existing methods [i.e., NN, ELM, SVM (linear kernel), SVM (precomputed kernel), and RF] (see Tables 6–10) in all experiments. In addition, the performance of the GWO-ELM classifier has been proven to outperform some recent studies (see Table 10) in both binary and multi-class classifications. The maximum multi-class classification performance of the GWO-ELM classifier was achieved with an accuracy reaching up to 96.21%. Further, the maximum binary classification performance of the GWO-ELM classifier was achieved with an accuracy of 99.47%. This demonstrates that the combination of the GWO-ELM and HOG-PCA is an effective classifier for detecting DR and might be applicable in solving other image data type. However, the current research has taken into account only the off-line aspect for detecting DR. Therefore, the future plan of the current research is to establish an approach to detect DR, which can

handle the online execution for both classification and feature extraction in order to meet the real-time aspects. The proposed DR detection approach will be tested under adversarial attacks. Additionally, other optimization methods for ELM will be further explored in order to generate the most suitable weights and biases for the ELM which leads to minimizing classification process errors.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: APTOS-2019: <https://www.kaggle.com/competitions/aptos2019-blindness-detection/data>; IDRiD: <https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid>.

Author contributions

MAA: conceptualization, methodology, writing—original draft, software, writing review, and editing. MA: supervision, funding acquisition, and project administration. ST: supervision. FA-D: writing review and editing. MH: investigation. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the Universiti Kebangsaan Malaysia under Grant DIP-2019-013.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Akram MU, Khalid S, Khan S. Identification A. and classification of microaneurysms for early detection of diabetic retinopathy. *Pattern Recognit.* (2013) 46:107–16. doi: 10.1016/j.patcog.2012.07.002
- Taylor R, Batey D. *Handbook of Retinal Screening in Diabetes: Diagnosis and Management*. Hoboken, NJ: John Wiley and Sons (2012).
- AL-Dhief FT, Latiff ANM, Baki MM, Malik NNA N, Sabri N, Albadr MAA. Voice pathology detection using support vector machine based on different number of voice signals. In: *2021 26th IEEE Asia-Pacific Conference on Communications (APCC)*. Kuala Lumpur: IEEE (2021). p. 1–6.
- AL-Dhief FT, Latiff ANM, Malik NNA N, Sabri N, Baki MM, Albadr MAA, et al. Voice pathology detection using machine learning technique. In: *2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT)*. Shah Alam: IEEE (2020). p. 99–104.
- Al-Dhief FT, Latiff ANM, Malik NNA N, Salim NS, Baki MM, Albadr MAA, et al. A survey of voice pathology surveillance systems based on Internet of Things and machine learning algorithms. *IEEE Access.* (2020) 8:64514–33. doi: 10.1109/ACCESS.2020.2984925
- Albadr MAA, Tiun S, Ayob M, Al-Dhief FT, Omar K, Hamzah FA. Optimised genetic algorithm-extreme learning machine approach for automatic COVID-19 detection. *PLoS ONE.* (2020) 15:e0242899. doi: 10.1371/journal.pone.0242899
- Gadekallu TR, Khare N, Bhattacharya S, Singh S, Maddikunta PKR, Ra I-H, et al. Early detection of diabetic retinopathy using PCA-firefly based deep learning model. *Electronics.* (2020) 9:274. doi: 10.3390/electronics9020274
- Gadekallu TR, Khare N, Bhattacharya S, Singh S, Maddikunta PKR, Srivastava G. Deep neural networks to predict diabetic retinopathy. *J Ambient Intell Hum Comput.* (2020) 2020:1–14. doi: 10.1007/s12652-020-01963-7
- Reddy GT, Bhattacharya S, Ramakrishnan SS, Chowdhary CL, Hakak S, Kaluri R, et al. An ensemble based machine learning model for diabetic retinopathy classification. In: *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. Vellore: IEEE (2020). p. 1–6.
- Abdel-Nasser M, Moreno A, Puig D. Breast cancer detection in thermal infrared images using representation learning and texture analysis methods. *Electronics.* (2019) 8:100. doi: 10.3390/electronics8010100
- Abdel-Nasser M, Saleh A, Moreno A, Puig D. Automatic nipple detection in breast thermograms. *Expert Syst Appl.* (2016) 64:365–74. doi: 10.1016/j.eswa.2016.08.026
- Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. San Diego, CA: IEEE (2005). p. 886–93.
- Marsboom C, Vrebois D, Staes J. Meire Using dimension reduction PCA to identify ecosystem service bundles. *Ecol Indic.* (2018) 87:209–60. doi: 10.1016/j.ecolind.2017.10.049
- Shi-fan Q, Jun-kun T, Yong-gang Z, Li-jun W, Ming-fei Z, Jun T, et al. Settlement prediction of foundation pit excavation based on the GWO-ELM model considering different states of influence. *Adv Civil Eng.* (2021) 2021:8896210. doi: 10.1155/2021/8896210
- Sridhar S, PradeepKandhasamy J, Sinthuja M, Minish TS. Diabetic retinopathy detection using convolutional neural networks algorithm. *Mater Today: Proc.* (2021) 206:106094. doi: 10.1016/j.matpr.2021.01.059
- Gangwar AK, Ravi V. Diabetic retinopathy detection using transfer learning and deep learning. In: *Evolution in Computational Intelligence*. Singapore: Springer (2021). p. 679–89.
- Albadr MAA, Tiun S, Ayob M, Al-Dhief FT, T-Abdali AN, Abbas AF. Extreme learning machine for automatic language identification utilizing emotion speech data. In: *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. Kuala Lumpur: IEEE (2021). p. 1–6.
- Albadr MAA, Tiuna S. Extreme learning machine: a review. *Int J Appl Eng Res.* (2017) 12:4610–23. doi: 10.37622/000000
- Albadr MAA, Tiun S, Ayob M, AL-Dhief FT, Omar K, Maen MK. Speech emotion recognition using optimized genetic algorithm-extreme learning machine. *Multimedia Tools Appl.* (2022) 81:1–27. doi: 10.1007/s11042-022-12747-w
- Huang G, Huang G-B, Song S, You K. Trends in extreme learning machines: a review. *Neural Networks.* (2015) 61:32–48. doi: 10.1016/j.neunet.2014.10.001
- Albadr MAA, Tiun S, Al-Dhief FT, Sammour MA. Spoken language identification based on the enhanced self-adjusting extreme learning machine approach. *PLoS ONE.* (2018) 13:e0194770. doi: 10.1371/journal.pone.0194770
- Albadr MAA, Tiun S, Ayob M, Mohammed M, AL-Dhief FT. Mel-frequency cepstral coefficient features based on standard deviation and principal component analysis for language identification systems. *Cognit Comput.* (2021) 13:1136–53. doi: 10.1007/s12559-021-09914-w
- Asha P, Karpagavalli S. Diabetic retinal exudates detection using machine learning techniques. In: *2015 International Conference on Advanced Computing and Communication Systems*. Coimbatore: IEEE (2015). p. 1–5.
- Zhang Y, An M. An active learning classifier for further reducing diabetic retinopathy screening system cost. *Comput Math Methods Med.* (2016) 2016:4345936. doi: 10.1155/2016/4345936
- Punithavathi IH, Kumar PG. Severity grading of diabetic retinopathy using extreme learning machine. In: *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. Srivilliputtur: IEEE (2017). p. 1–6.
- Deepa V, Kumar CS, Andrews SS. Fusing dual-tree quaternion wavelet transform and local mesh based features for grading of diabetic retinopathy using extreme learning machine classifier. *Int J Imaging Syst Technol.* (2021) 31:1625–37. doi: 10.1002/ima.22573
- Albadr MAA, Tiun S, Ayob M, AL-Dhief FT. Spoken language identification based on optimised genetic algorithm-extreme learning machine approach. *Int J Speech Technol.* (2019) 22:711–27. doi: 10.1007/s10772-019-09621-w
- Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. *Adv Eng Software.* (2014) 69:46–61. doi: 10.1016/j.advengsoft.2013.12.007
- Faris H, Aljarah I, Al-Betar MA, Mirjalili S. Grey wolf optimizer: a review of recent variants and applications. *Neural Comput Appl.* (2018) 30:413–35. doi: 10.1007/s00521-017-3272-5
- Faris H, Mirjalili S, Aljarah I. Automatic selection of hidden neurons and weights in neural networks using grey wolf optimizer based on a hybrid encoding scheme. *Int J Mach Learn Cybern.* (2019) 10:2901–20. doi: 10.1007/s13042-018-00913-2
- Li Q, Chen H, Huang H, Zhao X, Cai Z, Tong C, et al. An enhanced grey wolf optimization based feature selection wrapped kernel extreme

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

learning machine for medical diagnosis. *Comput Math Methods Med.* (2017) 2017. doi: 10.1155/2017/9512741

32. Hu L, Li H, Cai Z, Lin F, Hong G, Chen H, et al. A new machine-learning method to prognosticate paraquat poisoned patients by combining coagulation, liver, kidney indices. *PLoS ONE.* (2017) 12:e0186427. doi: 10.1371/journal.pone.0186427

33. Jasmine Selvakumari Jeya I, Revathi M, Uma Priya M. Proposed self-regulated gray wolf optimizer based extreme learning machine neural network classifier for lung cancer classification. *Int J Recent Technol Eng.* (2019) 8:2S119. doi: 10.35940/ijrte.B1064.0982S1119

34. SharmilaSM, Indra Gandhi MP. A novel method for identification of cardiovascular disease using KELM optimized by grey wolf algorithm. *Int J Innovat Technol Exploring Eng.* (2019) 8:8919. doi: 10.35940/ijitee.19006.078919

35. Naz A, Javaid N, Javaid S. Enhanced recurrent extreme learning machine using gray wolf optimization for load forecasting. In: *2018 IEEE 21st International Multi-Topic Conference (INMIC)*. Karachi: IEEE (2018). p. 1–5.

36. Wang M, Chen H, Li H, Cai Z, Zhao X, Tong C, et al. Grey wolf optimization evolving kernel extreme learning machine: application to bankruptcy prediction. *Eng Appl Artif Intell.* (2017) 63:54–68. doi: 10.1016/j.engappai.2017.05.003

37. Zhao X, Zhang X, Cai Z, Tian X, Wang X, Huang Y, et al. Chaos enhanced grey wolf optimization wrapped ELM for diagnosis of paraquat-poisoned patients. *Comput Biol Chem.* (2019) 78:481–90. doi: 10.1016/j.compbiolchem.2018.11.017

38. Reddy GSV, Das D, Biswas SK, Prashanth BS, Bhargav BP, Kumar TV, et al. Comparative analysis of intelligent systems using support vector machine for the detection of diabetic retinopathy. In: *Intelligent Computing and Communication Systems*. Singapore: Springer (2021). p. 245–57.

39. Hospital E. *APTOS-2019*. Kaggle. (2021). Available online at: <https://www.kaggle.com/c/aptos2019-blindness-detection/data> (accessed January 1, 2019).

40. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabudhe V, et al. Indian diabetic retinopathy image dataset (IDRID): a database for diabetic retinopathy screening research. *Data.* (2018) 3:25. doi: 10.3390/data3030025

41. Mu Y, Liu X, Wang L. A Pearson's correlation coefficient based decision tree and its parallel implementation. *Inf Sci.* (2018) 435:40–58. doi: 10.1016/j.ins.2017.12.059

42. Xu M, Li T, Wang Z, Deng X, Yang R, Guan Z. Reducing complexity of HEVC: a deep learning approach. *IEEE Trans Image Process.* (2018) 27:5044–59. doi: 10.1109/TIP.2018.2847035

43. Yu T, Zhang J, Cai W, Qi F. Toward real-time volumetric tomography for combustion diagnostics via dimension reduction. *Opt Lett.* (2018) 43:1107–10. doi: 10.1364/OL.43.001107

44. Al-Dhief FT, Baki MM, Latiff ANM, Malik NNA N, Salim NS, Albader MAA, et al. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access.* (2021) 9:77293–306. doi: 10.1109/ACCESS.2021.3082565

45. Albadr MA, Tiun S, Ayob M, Al-Dhief F. Genetic algorithm based on natural selection theory for optimization problems. *Symmetry.* (2020) 12:1758. doi: 10.3390/sym12111758

46. Albadr MAA, Tiun S. Spoken language identification based on particle swarm optimisation-extreme learning machine approach. *Circ Syst Signal Process.* (2020) 39:1–27. doi: 10.1007/s00034-020-01388-9

47. Adriman R, Muchtar K, Maulina N. Performance evaluation of binary classification of diabetic retinopathy through deep learning techniques using texture feature. *Procedia Comput Sci.* (2021) 179:88–94. doi: 10.1016/j.procs.2020.12.012

48. Alyoubi WL, Abulkhair MF, Shalash WM. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. *Sensors.* (2021) 21:3704. doi: 10.3390/s21113704

49. Alzami F, Megantara RA, Abdussalam P, Fanani AZ, Andono PN, Soeleman MA. *Exudates detection for multiclass diabetic retinopathy grade detection using ensemble*. Technology Reports of Kansai University (2020).

50. Bodapati JD, Naralasetti V, Shareef SN, Hakak S, Bilal M, Maddikunta PKR, et al. Blended multi-modal deep convnet features for diabetic retinopathy severity prediction. *Electronics.* (2020) 9:914. doi: 10.3390/electronics9060914

51. Bodapati JD, Shaik NS, Naralasetti V. Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. *J Ambient Intell Hum Comput.* (2021) 12:1–15. doi: 10.1007/s12652-020-02727-z

52. Dondeti V, Bodapati JD, Shareef SN, Veeranjanyulu N. Deep convolution features in non-linear embedding space for fundus image classification. *Rev d'Intelligence Artif.* (2020) 34:307–13. doi: 10.18280/ria.340308

53. Gayathri S, Gopi VP. Palanisamy automated classification of diabetic retinopathy through reliable feature selection. *Phys Eng Sci Med.* (2020) 43:927–45. doi: 10.1007/s13246-020-00890-3

54. Gayathri S, Gopi VP, Palanisamy A. lightweight CNN for Diabetic Retinopathy classification from fundus images. *Biomed Signal Process Control.* (2020) 62:102115. doi: 10.1016/j.bspc.2020.102115

55. Kassani SH, Kassani PH, Khazaeinezhad R, Wesolowski MJ, Schneider KA, Deters R. Diabetic retinopathy classification using a modified xception architecture. In: *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. Ajman: IEEE (2019). p. 1–6.

56. Liu H, Yue K, Cheng S, Pan C, Sun J, Li W. Hybrid model structure for diabetic retinopathy classification. *J Healthc Eng.* (2020) 2020:8840174. doi: 10.1155/2020/8840174

57. Nahiduzzaman M, Islam MR, Islam SR, Goni MOF, Anower MS, Kwak K-S. Hybrid CNN-SVD based prominent feature extraction and selection for grading diabetic retinopathy using extreme learning machine algorithm. *IEEE Access.* (2021) 9:152261–74 doi: 10.1109/ACCESS.2021.3125791

58. Saranya P, Prabakaran S. Automatic detection of non-proliferative diabetic retinopathy in retinal fundus images using convolution neural network. *J Ambient Intell Hum Comput.* (2020) 2020:1–10. doi: 10.1007/s12652-020-02518-6

59. Sikder N, Masud M, Bairagi AK, Arif ASM, Nahid A-A, Alhumyani HA. Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. *Symmetry.* (2021) 13:670. doi: 10.3390/sym13040670

60. Singh RK, Gorantla R. DMENet: diabetic macular edema diagnosis using hierarchical ensemble of CNNs. *PLoS ONE.* (2020) 15:e0220677. doi: 10.1371/journal.pone.0220677

61. Tymchenko B, Marchenko P, Spodarets D. Deep learning approach to diabetic retinopathy detection. *arXiv preprint arXiv:2003.02261.* (2020) doi: 10.5220/0008970805010509

62. Vaibhavi P, Manjesh R. Binary classification of diabetic retinopathy detection and web application. *Int J Res Eng Sci Manag.* (2021) 4:142–5.

63. Wu Z, Shi G, Chen Y, Shi F, Chen X, Coatrieux G, et al. Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network. *Artif Intell Med.* (2020) 108:101936. doi: 10.1016/j.artmed.2020.101936

64. Yi S-L, Yang X-L, Wang T-W, She F-R, Xiong X, He J-F. Diabetic retinopathy diagnosis based on RA-EfficientNet. *Appl Sci.* (2021) 11:11035. doi: 10.3390/app112211035

65. Zhou Y, He X, Huang L, Liu L, Zhu F, Cui S, et al. Collaborative learning of semi-supervised segmentation and classification for medical images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA: IEEE (2019). p. 2079–88.



OPEN ACCESS

EDITED BY

Yu-Hsiu Lin,
National Chung Cheng
University, Taiwan

REVIEWED BY

Bin Peng,
Chongqing Medical University, China
Hakseung Kim,
Korea University, South Korea

*CORRESPONDENCE

Chengliang Chai
chlchai@cdc.zj.cn
Chao Wu
chao.wu@zju.edu.cn

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 13 June 2022

ACCEPTED 02 August 2022

PUBLISHED 25 August 2022

CITATION

He J, Li J, Jiang S, Cheng W, Jiang J,
Xu Y, Yang J, Zhou X, Chai C and Wu C
(2022) Application of machine learning
algorithms in predicting HIV infection
among men who have sex with men:
Model development and validation.
Front. Public Health 10:967681.
doi: 10.3389/fpubh.2022.967681

COPYRIGHT

© 2022 He, Li, Jiang, Cheng, Jiang, Xu,
Yang, Zhou, Chai and Wu. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation

Jiajin He¹, Jinhua Li², Siqing Jiang¹, Wei Cheng³, Jun Jiang³,
Yun Xu³, Jieze Yang³, Xin Zhou³, Chengliang Chai^{3*} and
Chao Wu^{4*}

¹School of Public Health, Zhejiang University School of Medicine, Hangzhou, China, ²School of Software Technology, Zhejiang University, Ningbo, China, ³Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China, ⁴School of Public Affairs, Zhejiang University, Hangzhou, China

Background: Continuously growing of HIV incidence among men who have sex with men (MSM), as well as the low rate of HIV testing of MSM in China, demonstrates a need for innovative strategies to improve the implementation of HIV prevention. The use of machine learning algorithms is an increasing tendency in disease diagnosis prediction. We aimed to develop and validate machine learning models in predicting HIV infection among MSM that can identify individuals at increased risk of HIV acquisition for transmission-reduction interventions.

Methods: We extracted data from MSM sentinel surveillance in Zhejiang province from 2018 to 2020. Univariate logistic regression was used to select significant variables in 2018–2019 data ($P < 0.05$). After data processing and feature selection, we divided the model development data into two groups by stratified random sampling: training data (70%) and testing data (30%). The Synthetic Minority Oversampling Technique (SMOTE) was applied to solve the problem of unbalanced data. The evaluation metrics of model performance were comprised of accuracy, precision, recall, F-measure, and the area under the receiver operating characteristic curve (AUC). Then, we explored three commonly-used machine learning algorithms to compare with logistic regression (LR), including decision tree (DT), support vector machines (SVM), and random forest (RF). Finally, the four models were validated prospectively with 2020 data from Zhejiang province.

Results: A total of 6,346 MSM were included in model development data, 372 of whom were diagnosed with HIV. In feature selection, 12 variables were selected as model predicting indicators. Compared with LR, the algorithms of DT, SVM, and RF improved the classification prediction performance in SMOTE-processed data, with the AUC of 0.778, 0.856, 0.887, and 0.942, respectively. RF was the best-performing algorithm (accuracy = 0.871, precision = 0.960, recall = 0.775, F-measure = 0.858, and AUC = 0.942). And the RF model still performed well on prospective validation (AUC = 0.846).

Conclusion: Machine learning models are substantially better than conventional LR model and RF should be considered in prediction tools of HIV infection in Chinese MSM. Further studies are needed to optimize and promote these algorithms and evaluate their impact on HIV prevention of MSM.

KEYWORDS

machine learning, HIV, MSM, prediction, models

Introduction

Acquired immune deficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV) is a global health crisis, which destroys the human immune system and gives rise to a variety of opportunistic infections and death (1, 2). Men who have sex with men (MSM) are one of the highest-risk populations for HIV acquisition because of their tendency to have multiple sexual partners and unprotected anal intercourse (3). Therefore, this group has now received special attention from society.

In China, an increasing body of evidence from different periods and locations has suggested that MSM play an important role in the HIV epidemic. A large-scale systematic analysis, in which data were extracted from 355 cross-sectional studies covered 59 cities from 2001 to 2018, found that the overall national prevalence of HIV among MSM was estimated to be 5.7% (95% CI: 5.4–6.1%), exceeding the WHO 5% AIDS epidemic warning threshold (4, 5). And two reports by the Chinese Center for Disease Control and Prevention (CDC) showed that the proportion of newly identified HIV/AIDS cases due to male-to-male intercourse has increased rapidly, from 13.7% in 2011 to 25.5% in 2017 (6, 7). To improve the status quo, the Chinese government has taken actions to promote HIV testing by MSM, such as Pilot Program for HIV/AIDS Comprehensive Intervention, peer education, and free HIV voluntary counseling and testing (8). However, only 47% of Chinese MSM had ever tested for HIV in 2016, and only 38% had tested for HIV in the last 12 months (9). This is still far behind the target of the Joint United Nations Programme on HIV/AIDS (UNAIDS) for 90% testing among infected individuals (10). It is urgent to develop a reliable model to identify early infected MSM in order to reduce the transmission of virus in this group, which can make up for the defect of incomplete coverage of HIV testing to a certain extent.

Previous studies have used logistic regression or Cox proportional hazards regression models to establish the prediction tool of HIV infection among MSM, but performance is not great due to the problems of data structure which are often non-linear, abnormal, and heterogeneous (11–14). Compared to the above traditional models, the machine learning

algorithm provides a new method to construct models, since it can balance the deviation and variance of data (15). Nowadays, machine learning has been widely applied in the medical field, mainly reflected in medical auxiliary diagnosis and classification prediction, such as image-based cancer diagnostics (16, 17). However, machine learning algorithms have not been used to predict HIV infection among MSM, especially in China. In the present study, we focused on the Chinese MSM population and aimed to develop prediction models for HIV acquisition using logistic regression and several machine learning approaches. The processing of imbalance data by SMOTE before modeling is different from previous related studies. The predictive performance of these models is tested to determine the one that can most accurately identify high-risk MSM individuals with HIV, thus providing a basis for timely intervention and treatment of this population.

Materials and methods

Study population and data collection

MSM Sentinel Surveillance is a national government public health activity. The survey subjects were recruited by Non-Governmental Organizations and local CDC using snowball sampling at MSM event venues or online, with one-on-one questionnaires administered by trained enumerators and 5 ml of venous blood collected. Verbal consent was obtained from all study participants before survey and collection of specimens. Therefore, institutional review board approval was not required for analysis using sentinel surveillance data in China.

In this study, the cross-sectional data was derived from the questionnaire records which were collected from the MSM sentinel surveillance in Zhejiang province between 2018 and 2020. We included MSM that: (i) had oral or anal sex with other men within the past year, (ii) currently resided in Zhejiang Province, and (iii) were aged ≥ 15 years at the time of the survey. We excluded MSM that: (i) had already tested positive for HIV every year, (ii) disagreed to be blood collected. The main content of the questionnaire included five parts: general demographic information, AIDS-related knowledge, the occurrence of sexual behaviors, prevention services, and HIV antibodies testing. HIV

antibodies testing used ELISA reagents for initial screening and retesting, and Western Blot was used for confirmatory testing when the results of both tests were positive.

Data processing

Some samples may exist with missed or abnormal values, so we performed data cleaning to delete them. In addition, we also performed data transformation on the several features: “age” was divided into four classes according to Chinese age group classification (<18, 18–40, 41–65, >65); as for “AIDS-related knowledge”, if the results of 8 questions turn out to be all right, we would give a value of 1, otherwise, value of 0 will be given; “time of last HIV test” can be converted to dichotomous variables that whether had been tested for HIV in the past year or not. After data processing, continuous variables were presented as mean \pm standard deviation or median [interquartile range (IQR)], and categorical variables were presented as the frequency number (percentage).

Feature selection

The purpose of the feature selection was to eliminate redundant and irrelevant variables. Potential features can be selected by traditional statistical methods (15). We applied the filter method of univariate logistic regression to choose the feature subsets in which the independent variables are correlated with the dependent variable in the original data structure. Variables with statistical significance (p -value < 0.05) were selected as predicting features. As an estimate of effect size and variability, we have reported the odds ratio (OR) with a 95% confidence interval (CI).

Data balancing

As the proportion of MSM infected with HIV was imbalanced in this study, we can apply resampling method to handle the disproportionate ratio of observations in each class. The technology of resampling consisted of random under-sampling (RUS) and random over-sampling (ROS). However, RUS removed a number of samples of the majority class so that lost some information. In our experiments, we performed the Synthetic Minority Over-sampling Technique (SMOTE) to balance data (18). SMOTE could generate synthetic data to increase the number in the smaller class by using the nearest neighbor's algorithm (19).

TABLE 1 The confusion matrix.

	Positive actual case	Negative actual case
Positive prediction	True positive (TP)	False positive (FP)
Negative prediction	False negative (FN)	True negative (TN)

Model establishing

We explored three classic machine learning algorithms for predicting HIV infection in MSM compared with Logistic Regression (LR), including Decision Tree (DT), Support Vector Machines (SVM), and Random Forest (RF). These algorithms are widely used for classification problems, and each has its unique features and advantages. LR is a generalized linear regression model that can apply a non-linear sigmoid function to predict the results of two sets of classifications through a series of continuous or categorical variables (20). DT uses tree structure to classify data in a hierarchical fashion and is recommended for problems in which input variables are discrete and final classification is binary (21). SVM employs the “max-margin principle” to create a decision boundary that is as far as possible from the closest data points from each of the classes (22). RF is an ensemble version of decision tree by aggregating predictions from multiple decision trees for a better model, which is more robust against overfitting (23).

Model evaluation

Several standard indicators can be adopted to evaluate models' performance: accuracy, precision, recall, F-measure, and the area under the receiver operating characteristic curve (AUC). The result of a classification job can be classified into four categories in a confusion matrix that can explain the evaluation metrics for better understanding (24), and the tabular form of output is shown in Table 1. Relative concepts were shown as follows (15, 20).

Accuracy measures the ratio of correct classification.

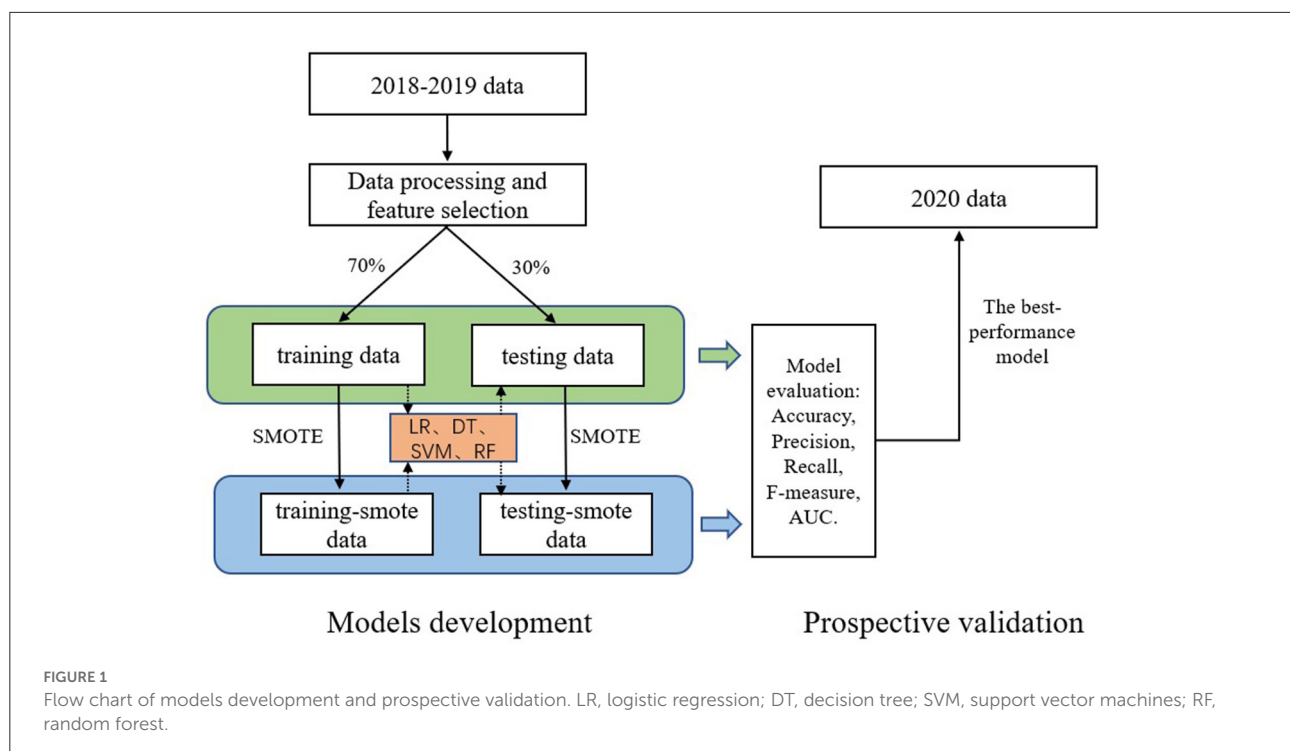
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision means the proportion of positive prediction that are positive actual cases.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall represents the fraction of positive actual cases that are correctly predicted.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



F-measure is the weighted harmonic mean of precision and recall. The higher the F-measure, the better predictive power of the model.

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC is a measure of the discriminative ability of prediction models which represents the area under the receiver operating characteristic curve (ROC). The vertical coordinate of the ROC curve is the true positive rate, and the horizontal coordinate of the ROC curve is the false positive rate. The value of AUC is between 0.5 and 1, and the closer to 1 indicates the better performance of the model (25).

In this study, we built our models according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis statement for prediction models (26) (see Figure 1 for flow chart). After data processing and feature selection, we divided the 2018–2019 data into two groups by stratified random sampling based on the layer of “HIV-positive or negative”: training data (70%) and testing data (30%). Secondly, both training data and testing data were balanced by SMOTE. Then, the hyperparameters optimization of each model was obtained by grid-search and 5-fold cross-validation on the training data (see Supplementary Table 1). Finally, these models were verified on the testing data. We also conducted prospective validation of four models in the 2020 data. All data analyses were carried out with R software 4.1.2 version and Python software 3.9 version.

Results

Demographic characteristics

After data processing in 2018–2019 data, we included 6,346 MSM into this study, 372 of whom were infected with HIV (5.86%). The median age of them was 30.0 (IQR: 25.0–39.0) years, with 72 (1.13%) younger than 18 years, 4,993 (78.68%) aged 18–40 years, 1,245 (19.62%) aged 41–65 years, and 36 (0.57%) older than 65 years. And among them, 3,821 (60.21%) were unmarried; 6,557 (98.12%) identified as the Han ethnicity; 3,836 (60.45%) were census register of Zhejiang; 2,207 (34.78%) had obtained a college degree or above of education background.

Feature selection

Univariate logistic regression analysis was performed to search the possible predictors and their associations with HIV infection. Descriptive summaries were shown in Table 2. Of all 27 potential predictors, age, marital status, census register, ethnicity, years of living in Zhejiang, AIDS-related knowledge, Condom use in the latest homosexual sex, frequency of condom use during homosexual sex in the past 6 months, diagnosed with sexually transmitted diseases, condom promotion/AIDS counseling and testing, AIDS peer education and HIV test in the past year were associated with HIV acquisition (*P*

TABLE 2 Basic characteristics of variables in both 2018–2019 and 2020 MSM and univariate associations of potential predictors with HIV infection in 2018–2019 MSM.

Variables	2018–2019 MSM		Odd ratio (95% confidence interval)	P-value	2020 MSM	
	No-HIV	HIV			No-HIV	HIV
	N = 5,974 (94.14%)	N = 372 (5.86%)			N = 3,219 (95.72%)	N = 145 (4.28%)
Age (years)						
<18	66	6	Ref.		27	2
18–40	4,720	273	0.64 (0.27, 1.48)	0.294	2,470	108
41–65	1,161	84	0.80 (0.34, 1.89)	0.605	698	34
>65	27	9	3.67 (1.19, 11.30)	0.024	24	1
Marital status						
Unmarried	3,574	247	Ref.		1,974	92
Married	1,990	102	0.74 (0.59, 0.94)	0.013	992	44
Cohabiting	49	2	0.59 (0.14, 2.44)	0.467	21	0
Divorced or widowed	361	21	0.84 (0.53, 1.33)	0.461	232	9
Census register of Zhejiang						
No	2,289	221	Ref.		1,260	86
Yes	3,685	151	0.42 (0.34, 0.53)	<0.001	1,929	59
Ethnicity						
Han	5,878	349	Ref.		3,161	139
Others	96	23	4.04 (2.53, 6.44)	<0.001	58	6
Years of living in Zhejiang						
<3 months	280	28	Ref.		155	10
3–6 months	269	16	0.59 (0.31, 1.12)	0.110	107	9
7–12 months	467	24	0.51 (0.29, 0.90)	0.021	197	4
1–2 years	849	51	0.60 (0.37, 0.97)	0.037	633	21
>2 years	4,109	253	0.62 (0.41, 0.93)	0.020	2,127	101
Education background						
Illiteracy	35	4	Ref.		6	0
Primary school	250	18	0.63 (0.20, 1.97)	0.427	107	8
Junior high school	1,553	121	0.68 (0.24, 1.95)	0.475	838	41
Senior high school	2,047	111	0.47 (0.16, 1.36)	0.165	1,153	38
College degree or above	2,089	118	0.49 (0.17, 1.41)	0.189	1,115	58
Sexual orientation						
Homosexuality	3,994	249	Ref.		2,289	102
Heterosexuality	48	2	0.67 (0.16, 2.76)	0.578	36	1
Bisexuality	1,743	109	1.00 (0.79, 1.27)	0.979	774	40
Unascertained	189	12	1.02 (0.17, 1.41)	0.952	120	2
Places of seeking sex partners						
Bar/dance hall	339	13	Ref.		260	0
Tearoom/clubhouse	157	10	1.66 (0.71, 3.87)	0.240	143	8
Public bath	329	20	1.58 (0.78, 3.24)	0.206	175	5
Park	257	6	0.61 (0.23, 1.62)	0.321	72	1
Internet	4,761	315	1.73 (0.98, 3.04)	0.059	2,527	179
Others	131	8	1.59 (0.65, 3.93)	0.313	42	2
AIDS-related knowledge						
No	2,529	203	Ref.		1,143	71
Yes	3,445	169	0.61 (0.50, 0.75)	<0.001	2,076	74

(Continued)

TABLE 2 (Continued)

Variables	2018–2019 MSM		Odd ratio (95% confidence interval)	P-value	2020 MSM	
	No-HIV	HIV			No-HIV	HIV
	N = 5,974 (94.14%)	N = 372 (5.86%)			N = 3,219 (95.72%)	N = 145 (4.28%)
Homosexual sex in the past week						
No	3,062	209	Ref.		1,630	91
Yes	2,912	163	0.82 (0.66, 1.01)	0.065	1,589	54
Condom use in the latest homosexual anal sex						
No	1,072	158	Ref.		363	39
Yes	4,902	214	0.30 (0.24, 0.37)	<0.001	2,856	106
Frequency of condom use during homosexual sex in the past 6 months						
Never	249	39	Ref.		105	10
Sometimes	2,430	233	0.61 (0.43, 0.88)	<0.001	817	82
Every time	3,295	100	0.19 (0.13, 0.29)	<0.001	2,297	53
Commercial sex in the past 6 months						
No	5,730	362	Ref.		3,105	139
Yes	244	10	0.65 (0.34, 1.23)	0.186	114	6
Heterosexual sex in the past 6 months						
No	4,602	298	Ref.		2,636	122
Yes	1,372	74	0.83 (0.64, 1.08)	0.171	583	23
Drug-taking						
No	5,912	366	Ref.		3,213	144
Yes	62	6	1.56 (0.67, 3.64)	0.300	6	1
Diagnosed with sexually transmitted diseases						
No	5,725	350	Ref.		3,116	137
Yes	249	22	1.45 (1.02, 2.26)	0.008	103	8
Condom promotion/AIDS counseling and testing						
No	1,406	124	Ref.		572	45
Yes	4,514	248	0.65 (0.52, 0.81)	<0.001	2,647	100
Community drug maintenance therapy/cleaning needle provision						
No	5,585	344	Ref.		3,038	138
Yes	389	28	1.17 (0.78, 1.74)	0.444	181	7
AIDS peer education						
No	3,059	225	Ref.		1,697	77
Yes	2,915	147	0.69 (0.55, 0.85)	0.001	1,522	68
HIV test in the past year						
No	2,691	212	Ref.		1,278	67
Yes	3,283	160	0.62 (0.50, 0.76)	<0.001	1,941	78

< 0.05), suggesting that these 12 variables can be used as predicting features.

Performance comparison of the models

After data processing and feature extraction, there are three stages in our approach of model construction. The first stage is stratified random sampling on the whole model development

dataset: training data ($n = 4,442$) and testing data ($n = 1,904$). We also implemented the resampling techniques of SMOTE in the original training data and testing data separately: training-smote data ($n = 8,364$) and testing-smote data ($n = 3,584$). Details showed in Table 3. Take the original data and SMOTE-processed data as input for the next stage.

In the second stage, we developed models by using the training data. In reference to Table 4, we summarized the performance of prediction models by validating in testing data.

TABLE 3 Description of original data and SMOTE-processed data.

Dataset	Minority class	Majority class	Samples in total
Training	260	4,182	4,442
Training-smote	4,182	4,182	8,364
Testing	112	1,792	1,904
Testing-smote	1,792	1,792	3,584

TABLE 4 Results of classification models in original unbalanced data.

Models	Accuracy	Precision	Recall	F-measure	AUC
LR	0.941	0.500	0.009	0.018	0.764
DT	0.934	0.208	0.045	0.074	0.549
SVM	0.935	0.071	0.009	0.016	0.632
RF	0.934	0.118	0.018	0.031	0.667

TABLE 5 Results of classification models in SMOTE-processed data.

Models	Accuracy	Precision	Recall	F-measure	AUC
LR	0.702	0.690	0.733	0.711	0.778
DT	0.852	0.954	0.741	0.834	0.853
SVM	0.811	0.906	0.695	0.787	0.887
RF	0.871	0.960	0.775	0.858	0.942

We can see that the only advisable indicator of these models was accuracy (>0.934). However, the results of other indicators were not great that recall ranged from a low of 0.009 to a high of 0.045 and F-1 ranged from a low of 0.016 to a high of 0.074. In the third stage, we also developed models by taking the training-smote data. Compared to the prediction effects of models in the original dataset, the performances of four models in SMOTE-processed data were much better, as shown in Table 5. The accuracy calculated by LR, DT, SVM, and RF was 0.702, 0.852, 0.811, and 0.871, respectively; the precision was 0.690, 0.954, 0.906, and 0.960, respectively; the recall was 0.733, 0.741, 0.695, and 0.755, respectively; the F-measure was 0.711, 0.834, 0.787, and 0.858, respectively; the ROC value was 0.778, 0.853, 0.887, and 0.942, respectively. ROC curves of four algorithms in two situations were shown in Figure 2.

Prospective validation

According to the results of the models above, we used the prospective validation data to further verify the extensibility of models in this stage. The basic characteristics of variables in 2020 data were shown in Table 2. ROC curves of four algorithms were shown in Figure 3. The final results showed that RF model also exhibits better performance compared with LR, DT, SVM (with the AUC of 0.596, 0.812, 0.823, and 0.846, respectively).

Compared with the AUC of the RF model in the internal testing set, we found that the AUC of the RF model in the prospective validation set decreased by 0.096.

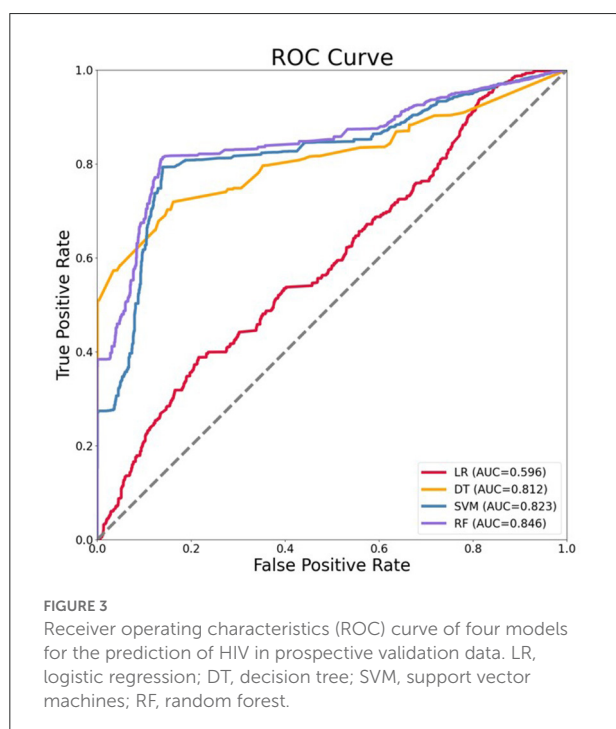
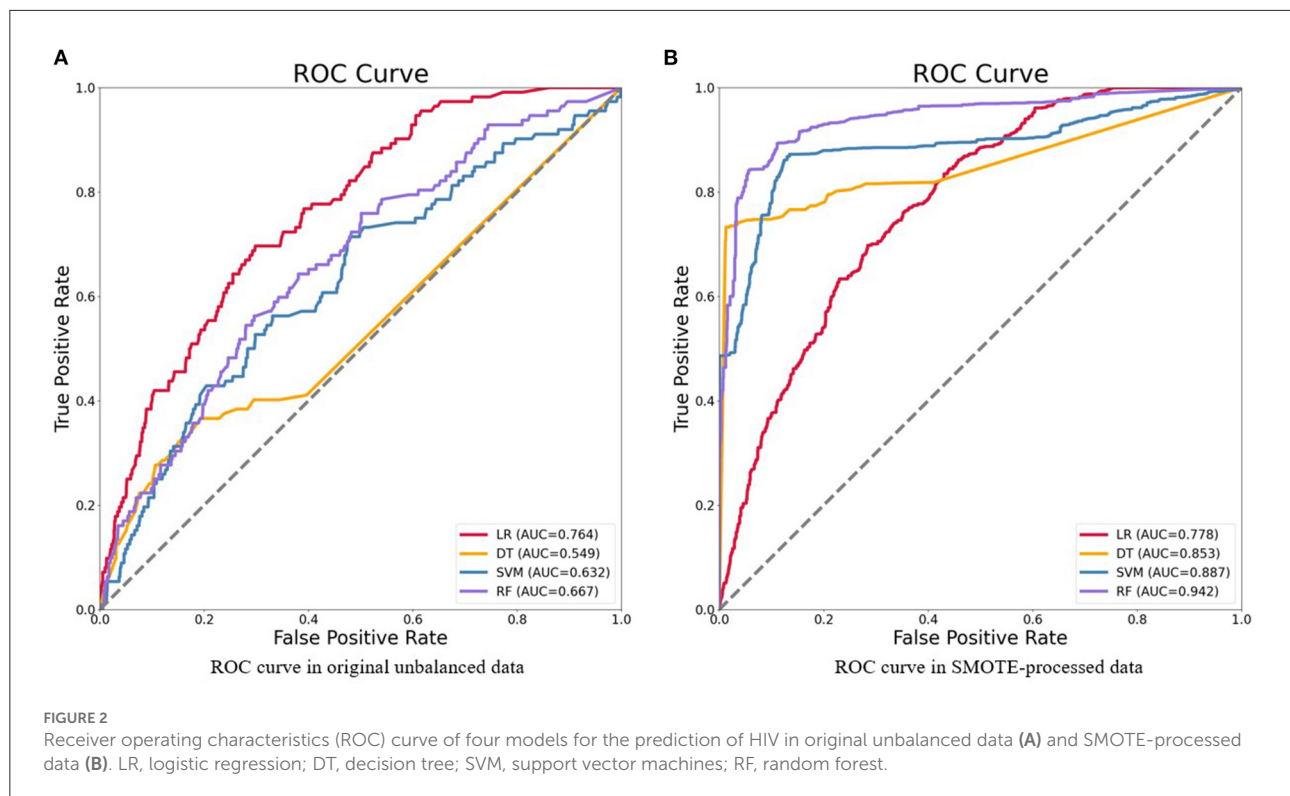
Discussion

MSM are one of the high-risk groups because they are susceptible to infection after engaging in unprotected anal sex (27). An updated systematic review and meta-analysis revealed that the overall HIV incidence for multiple periods among MSM in China was a rising trend, which pooled separately from 2005 to 2008 (3.24/100 PY), 2009 to 2011 (5.29/100 PY), and 2012 to 2014 (5.50/100 PY) (28). Failure to test and receive antiretroviral treatment in time will lead to the progression of diseases and ultimately to the development of AIDS, so MSM has been identified as a priority population for HIV prevention and control interventions in China (29). In response to the fact that the detection rate of MSM population is lower than the first of UNAIDS' 90-90-90 targets, we need to find a fairly high accuracy model for the prediction of HIV status.

To our knowledge, this is the first study to apply machine learning on AIDS sentinel surveillance data to predict HIV infection among MSM in China. The predictive model by machine learning can distinguish between high and low risks. As long as an individual has a predictive value of one, he or she is considered to be at high risk for HIV infection and could benefit from early additional screening and diagnosis (30). In the present study, we examined whether machine learning algorithms provide more accurate prediction models for HIV infection in MSM than the conventional logistic regression model.

In the beginning, the predictors selected for this study were independently associated with HIV infection, including important sexual behavior factors, such as condom use in the latest homosexual anal sex (OR = 0.30, 95% CI: 0.24–0.37), frequency of condom use was every time (OR = 0.19, 95% CI: 0.13–0.29), and diagnosed with sexually transmitted diseases (OR = 1.45, 95% CI: 1.02–2.26). These above variables were generally reported in recent HIV-related studies of behavioral risk factors (31, 32).

Our study shows that the approach of machine learning is feasible and fairly high accuracy. We compared LR, DT, SVM, and RF, and accuracy, precision, recall, F-measure, and AUC value of each model were analyzed. In unbalanced original data, we found that only the indicator of accuracy was acceptable and the other indicators were poor. However, using this metric alone is not meaningful because class distributions that are highly skewed tend to bias the results of machine learning algorithms (33). Even if all cases are predicted to be negative, the accuracy of the model is also more than 90%, but the precision and recall are both 0 (15). Therefore, it is not enough to represent great classifiers in terms of high accuracy value. Then, the comprehensive evaluation indices of machine learning



models in SMOTE-processed data were better than traditional logistic regression model, in which the RF model performed best (accuracy = 0.871, precision = 0.960, recall = 0.775,

F-measure = 0.858, AUC = 0.942). In addition, the RF model also performed well when the optimal model was prospectively validated with 2020 data (AUC = 0.846). The above results indicate that advanced methods of machine learning can be used to develop models with higher prediction accuracy, where the performance of RF is satisfactory to predict HIV status among MSM in China.

Previous studies have also provided evidence of using machine learning algorithms in predicting HIV infection. Krakower et al. (34) developed and validated multiple machine learning models to identify potential HIV pre-exposure prophylaxis (PrEP) candidates by using electronic health records containing 180 potential predictors from an ambulatory practice in Massachusetts in America, found that the best-performing algorithm was obtained with the least absolute shrinkage and selection operator (LASSO) (AUC = 0.86). In a similar setting in California, Marcus et al. (35) used 81 electronic health record variables to identify PrEP candidates by machine learning and demonstrated improved ability to predict incident HIV with inclusion of multiple data domains compared with simpler algorithms that based on MSM status and STI positivity (AUC = 0.86). In Denmark, Ahlstrom et al. (36) applied various machine learning methods in electronic registry data to predict HIV status and found that the RF algorithm also performed slightly better (AUC = 0.89). More recently, Bao et al. (37) developed four machine learning models and evaluated their performance in predicting HIV diagnosis based on a cohort

of MSM in Australia, and he proposed that Machine learning approaches outperformed the multivariable logistic regression model, with the gradient boosting machine achieving the highest performance ($AUC = 0.76$). Our study complements these machine learning studies applied to HIV infection prediction, all of which effectively illustrate that machine learning can be used as an effective method for detecting HIV infection among MSM.

There were several limitations to this study. First, although the questionnaire information was collected through individual interviews between survey subjects and health professionals, some of this occurred in the past, which is subjected to the recall bias. Moreover, the questionnaire needs to be further supplemented due to the absence of some behavioral characteristics (e.g., the number of sexual partners, sex role of accessor/recipient) (38). Second, we only employed the three most commonly-used machine learning algorithms for classification results prediction, so other useful models and methods can be explored in future research, including natural language processing in unstructured data (39). Third, since the research subjects selected for models building came from only Zhejiang province, further exploration is needed in generalizing the optimal model to the whole country and making it universally applicable. Fourth, machine learning for effectively avoiding overfitting is a crucial strategy (40). Our models may have the problem of overfitting and should address it by regularization and penalization of model complexity (41).

In conclusion, the study shows that machine learning has an advantage over traditional models in predicting HIV infection among MSM and the RF has a superior performance. In particular, SMOTE technology helps models to achieve better performance when facing unbalanced data. Within an increase in HIV incidence among MSM, even other high-risk populations, it is expected that prediction models based on machine learning for HIV infection can be an important direction to discriminate whether they are at high-risk for HIV acquisition to be provided with timely interventional treatment. Furthermore, additional researches are needed to further optimize these algorithms, expand useful models to the entire country, and evaluate their usefulness and effects of them on HIV prevention.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author/s.

Author contributions

JH, SJ, and CW came up with the original idea. JH and JL participated in the research design and provided

research methods. WC, JJ, YX, JY, XZ, and CC completed the data collection and improved the manuscript. JH performed the data analysis and drafted the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Key Research and Development Project of China (2021ZD0110400 and 2018AAA0101900), National Natural Science Foundation of China (U19B2042), Program of Zhejiang Province Science and Technology (2022C01044), The University Synergy Innovation Program of Anhui Province (GXXT-2021-004), Zhejiang Lab (2021KE0AC02), Academy of Social Governance Zhejiang University, Fundamental Research Funds for the Central Universities (226-2022-00064), Artificial Intelligence Research Foundation of Baidu Inc., Program of ZJU and Tongdun Joint Research Lab.

Acknowledgments

We are grateful to all the participants in this study. The authors also thank Center for Disease Control and Prevention of Zhejiang Province for their help.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.967681/full#supplementary-material>

References

- Liu H, Zhao M, Ren J, Qi X, Sun H, Qu L, et al. Identifying factors associated with depression among men living with HIV/AIDS and undergoing antiretroviral therapy: a cross-sectional study in Heilongjiang, China. *Health Qual Life Outcomes*. (2018) 16:190. doi: 10.1186/s12955-018-1020-x
- Silva LRD, Araújo ETH, Carvalho ML, Almeida CAPL, Oliveira ADDS, Carvalho PMG, et al. Epidemiological situation of acquired immunodeficiency syndrome (AIDS)-related mortality in a municipality in northeastern Brazil. A retrospective cross-sectional study. *São Paulo Med J*. (2018) 136:37–43. doi: 10.1590/1516-3180.2017.0130100917
- Zhang J, Xu JJ, Song W, Pan S, Chu ZX, Hu QH, et al. HIV Incidence and Care Linkage among MSM First-Time-Testers in Shenyang, China 2012–2014. *AIDS Behav*. (2018) 22:711–21. doi: 10.1007/s10461-017-1840-4
- Dong MJ, Peng B, Liu ZF, Ye QN, Liu H, Lu XL, et al. The prevalence of HIV among MSM in China: a large-scale systematic analysis. *BMC Infect Dis*. (2019) 19:1000. doi: 10.1186/s12879-019-4559-1
- Walker N, Stanecki KA, Brown T, Stover J, Lazzari S, Garcia-Calleja JM, et al. Methods and procedures for estimating HIV/AIDS and its impact: the UNAIDS/WHO estimates for the end of 2001. *AIDS*. (2003) 17:2215–25. doi: 10.1097/00002030-200310170-00010
- NCAIDS, NCSTD, China CDC. Update on the AIDS/STD epidemic in China in December 2017. *Chin J AIDS STD*. (2018) 24:111. doi: 10.13419/j.cnki.aids.2018.02.01
- NCAIDS, NCSTD, China CDC. Update on the AIDS/STD epidemic in China in 2011. *Chin J AIDS STD*. (2012) 18:64. doi: 10.13419/j.cnki.aids.2012.02.007
- Zhou J, Chen J, Goldsamt L, Wang H, Zhang C, Li X, et al. Testing and associated factors among men who have sex with men in Changsha, China. *J Assoc Nurses AIDS Care*. (2018) 29:932–41. doi: 10.1016/j.jana.2018.05.003
- Cao B, Liu C, Durvasula M, Tang W, Pan S, Saffer AJ, et al. Social media engagement and HIV testing among men who have sex with men in China: a nationwide cross-sectional survey. *J Med Internet Res*. (2017) 19:e251. doi: 10.2196/jmir.7251
- Sidibé M, Loures L, Samb B. The UNAIDS 90–90–90 target: a clear choice for ending AIDS and for sustainable health and development. *J Int AIDS Soc*. (2016) 19:21133. doi: 10.7448/IAS.19.1.21133
- Menza TW, Hughes JP, Celum CL, Golden MR. Prediction of HIV acquisition among men who have sex with men. *Sex Transm Dis*. (2009) 36:547–55. doi: 10.1097/OLQ.0b013e3181a9cc41
- Hoenigl M, Weibel N, Mehta SR, Anderson CM, Jenks J, Green N, et al. Development and validation of the San Diego Early Test Score to predict acute and early HIV infection risk in men who have sex with men. *Clin Infect Dis*. (2015) 61:468–75. doi: 10.1093/cid/civ335
- Yin L, Zhao Y, Peratikos MB, Song L, Zhang X, Xin R, et al. Risk prediction score for HIV infection: development and internal validation with cross-sectional data from men who have sex with men in China. *AIDS Behav*. (2018) 22:2267–76. doi: 10.1007/s10461-018-2129-y
- Xue M, Su Y, Li C, Wang S, Yao H. Identification of potential type II diabetes in a large-scale Chinese population using a systematic machine learning framework. *J Diabetes Res*. (2020) 2020:6873891. doi: 10.1155/2020/6873891
- Yin Y, Xue M, Shi L, Qiu T, Xia D, Fu G, et al. A noninvasive prediction model for hepatitis B virus disease in patients with HIV: based on the population of Jiangsu, China. *Biomed Res Int*. (2021) 2021:6696041. doi: 10.1155/2021/6696041
- Deo RC. Machine learning in medicine. *Circulation*. (2015) 132:1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593
- Thomasian NM, Kamel IR, Bai HX. Machine intelligence in non-invasive endocrine cancer diagnostics. *Nat Rev Endocrinol*. (2022) 18:81–95. doi: 10.1038/s41574-021-00543-9
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE. synthetic minority over-sampling technique. *J Artif Intell Res*. (2002) 16:321–57. doi: 10.1613/jair.953
- Yang PT, Wu WS, Wu CC, Shih YN, Hsieh CH, Hsu JL. Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. *Open Med*. (2021) 16:754–68. doi: 10.1515/med-2021-0282
- Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol*. (2001) 54:979–85. doi: 10.1016/S0895-4356(01)00372-9
- Nascimento PM, Medeiros IG, Falcão RM, Stransky B, de Souza JES, A. decision tree to improve identification of pathogenic mutations in clinical practice. *BMC Med Inform Decis Mak*. (2020) 20:52. doi: 10.1186/s12911-020-1060-0
- Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*. (2018) 15:41–51. doi: 10.21873/cgp.20063
- Wang B, Liu F, Deveaux L, Ash A, Gosh S, Li X, et al. Adolescent HIV-related behavioural prediction using machine learning: a foundation for precision HIV prevention. *AIDS*. (2021) 35:S75–84. doi: 10.1097/QAD.0000000000002867
- Shamsara J. Evaluation of the performance of various machine learning methods on the discrimination of the active compounds. *Chem Biol Drug Des*. (2021) 97:930–43. doi: 10.1111/cbdd.13819
- Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC Curve. *Int J Epidemiol*. (2020) 49:1397–403. doi: 10.1093/ije/dy274
- Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. (2015) 350:g7594. doi: 10.1136/bmj.g7594
- Stephenson R, White D, Darbes L, Hoff C, Sullivan P. HIV. testing behaviors and perceptions of risk of HIV infection among MSM with main partners. *AIDS Behav*. (2015) 19:553–60. doi: 10.1007/s10461-014-0862-4
- Zhang W, Xu JJ, Zou H, Zhang J, Wang N, Shang H. HIV incidence and associated risk factors in men who have sex with men in Mainland China: an updated systematic review and meta-analysis. *Sex Health*. (2016) 13:373–82. doi: 10.1071/SH16001
- Zhang BC, Chu QS. MSM and HIV/AIDS in China. *Cell Res*. (2005) 15:858–64. doi: 10.1038/sj.cr.7290359
- Yang H, Li X, Cao H, Cui Y, Luo Y, Liu J, et al. Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. *Comput Methods Programs Biomed*. (2021) 211:106420. doi: 10.1016/j.cmpb.2021.106420
- Wang L, Santella AJ, Wei X, Zhuang G, Li H, Zhang H, et al. Prevalence and protective factors of HIV and syphilis infection among men who have sex with men in Northwest China. *J Med Virol*. (2020) 92:1141–7. doi: 10.1002/jmv.25622
- Guanghua L, Yi C, Shuai T, Zhiyong S, Zhenzhu T, Yuhua R, et al. HIV, syphilis and behavioral risk factors among men who have sex with men in a drug-using area of southwestern China: results of 3 cross-sectional surveys from 2013 to 2015. *Medicine*. (2018) 97:e0404. doi: 10.1097/MD.00000000000010404
- Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Sci Rep*. (2021) 11:24039. doi: 10.1038/s41598-021-03430-5
- Krakower DS, Gruber S, Hsu K, Menchaca JT, Maro JC, Kruskal BA, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. *Lancet HIV*. (2019) 6:e696–704. doi: 10.1016/S2352-3018(19)30139-0
- Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *Lancet HIV*. (2019) 6:e688–95. doi: 10.1016/S2352-3018(19)30137-7
- Ahlström MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine*. (2019) 17:100203. doi: 10.1016/j.eclinm.2019.10.016
- Bao Y, Medland NA, Fairley CK, Wu J, Shang X, Chow EPF, et al. Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *J Infect*. (2021) 82:48–59. doi: 10.1016/j.jinf.2020.11.007
- Chen L, Chen W, Jiang T, Ni Z, Ma Q, Pan X. The characteristics and risk factors of web-based sexual behaviors among men who have sex with men in eastern China: cross-sectional study. *JMIR Public Health Surveill*. (2021) 7:e25360. doi: 10.2196/25360
- Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. *J Acquir Immune Defic Syndr*. (2018) 77:160–6. doi: 10.1097/QAI.0000000000001580
- Takahashi Y, Ueki M, Tamiya G, Ogishima S, Kinoshita K, Hozawa A. Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes. *Transl Psychiatry*. (2020) 10:294. doi: 10.1038/s41398-020-00957-5
- Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: part II-generalization and overfitting. *Acta Neurochir Suppl*. (2022) 134:15–21. doi: 10.1007/978-3-030-85292-4_3



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Yuliuming Wang,
The Second Affiliated Hospital of
Harbin Medical University, China
Marla Weetall,
PTC Therapeutics, United States
Shigao Huang,
Air Force Medical University, China
Shiva Basnet,
Tongji University, China

*CORRESPONDENCE

Ming Chen
mingchenseu@126.com
Jianping Wu
doctorwujianping@126.com
Yuan Meng
lishumy@live.cn

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 12 May 2022

ACCEPTED 22 August 2022

PUBLISHED 12 September 2022

CITATION

Zhu Y, Mao W, Zhang G, Sun S, Tao S,
Jiang T, Wang Q, Meng Y, Wu J and
Chen M (2022) Development and
validation of a prognostic nomogram
for adult patients with renal sarcoma:
A retrospective study based on the
SEER database.
Front. Public Health 10:942608.
doi: 10.3389/fpubh.2022.942608

COPYRIGHT

© 2022 Zhu, Mao, Zhang, Sun, Tao,
Jiang, Wang, Meng, Wu and Chen. This
is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction
in other forums is permitted, provided
the original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Development and validation of a prognostic nomogram for adult patients with renal sarcoma: A retrospective study based on the SEER database

Yongkun Zhu^{1,2†}, Weipu Mao^{1†}, Guangyuan Zhang^{1†}, Si Sun^{1,2},
Shuchun Tao², Tiancheng Jiang², Qingbo Wang³,
Yuan Meng^{4*}, Jianping Wu^{1*} and Ming Chen^{1*}

¹Department of Urology, Affiliated Zhongda Hospital of Southeast University, Nanjing, China,

²Department of Medical College, Southeast University, Nanjing, China, ³Department of
Chemotherapy, Affiliated the Second Hospital of Nanjing, Nanjing University of Chinese Medicine,
Nanjing, China, ⁴Department of Urology, Nanjing Lishui People's Hospital, Zhongda Hospital Lishui
Branch of Southeast University, Nanjing, China

Background: Renal sarcoma (RS) is rarely seen in clinical practice. The purpose of this study was to develop a prognostic nomogram model, which could predict the probability of overall survival (OS) and cancer-specific survival (CSS) in adult patients with RS.

Methods: Patients diagnosed with RS were recruited from the SEER database between 2004 and 2015, and randomized to two cohorts: the training cohort and the validation cohort. Uni- and multivariate Cox regression analyses in the training cohort were used to screen independent prognostic factors for OS and CSS. Prognostic nomograms for OS and CSS were created separately for adult RS patients based on independent risk factors. The area under the receiver operating characteristic (ROC) curves, calibration curves, and decision curve analysis (DCA) were used to validate the nomograms.

Results: A total of 232 eligible patients were recruited, including 162 in the training cohort and 70 in the validation cohort. Sex, histological type, SEER stage, and surgery were independent prognostic factors for OS, while histological type, SEER stage, surgery, chemotherapy were independent prognostic factors for CSS. Based on the above independent prognostic factors, prognostic nomograms for OS and CSS were created respectively. In the training cohort, the AUCs of the nomograms for OS and CSS were 0.742 and 0.733, respectively. In the validation cohort, the AUCs of the nomograms for OS and CSS were 0.837 and 0.758, respectively. The calibration curves of the nomograms showed high consistencies between the predicted and actual survival rates. Finally, the DCA demonstrated that the nomograms in the wide high-risk threshold had a higher net benefit than the SEER stage.

Conclusion: A prognostic nomogram for renal sarcoma was created and validated for reliability and usefulness in our study, which assisted urologists in accurately assessing the prognosis of adult RS patients.

KEYWORDS

adult patients, renal sarcoma, nomogram, SEER, prognosis

Introduction

Sarcomas are a heterogeneous group of tumors arising in the embryonic mesoderm, accounting for approximately 1% of all malignant tumors, of which <5% occur in the urogenital tract (1). Primary renal sarcoma (RS) accounts for around 24.6% of all genitourinary sarcomas and <1% of all primary kidney tumors (1, 2). Renal sarcoma is not only very rare but also leads to a poor prognosis: the overall 1-, 3-, and 5-year survival rate was 86.3, 40.7, and 14.5%, respectively, and the median survival was 28 months (3). According to previous reviews and case reports, renal sarcoma could be classified into the following pathological types: liposarcoma (4), leiomyosarcoma (5), carcinosarcoma (6), rhabdomyosarcoma (7), clear cell sarcoma (8), fibrosarcoma (9) and others, and different pathological types predict distinct prognosis.

RS is currently poorly studied as it is such a rare malignancy. As a result, an accurate prognostic model for RS is essential for both urologists and patients. In fact, the SEER stage grading system was employed by urologists to measure the progression of RS, which includes localized, regional, distant, and unstaged (10, 11). However, other factors including sex, age, year of diagnosis, race, marital status, radiation, chemotherapy,

surgery, etc. may also have an impact on prognosis due to individual variances. In recent years, nomograms have been increasingly employed in clinical practice for cancer prognosis. It has been regarded as a useful statistical prediction tool for benefiting both clinicians and patients (12, 13). So far, there is no report on the application of nomograms in predicting the prognosis of renal sarcoma in adults. In the present study, based on data from the SEER database between 2004 and 2015, nomograms were set up to predict survival outcomes for adult patients with RS and their reliability was also validated.

Materials and methods

Data sources

Data were extracted from the Surveillance Epidemiology and End Results (SEER) database (<https://seer.cancer.gov/>), which is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS). SEER statistics are collected on a national scale, with information from 18 states that represent all regions of the country covering 28% of the US population, including sociodemographic factors, geographic variables, clinical factors, cancer-specific factors, pathologic variables, treatment factors, and outcomes (14). The SEER database is openly accessed, and all authors have obtained permission. SEER*Stat software [Version 8.3.9.2 - August 20, 2021, SEER*Stat Software ([cancer.gov](https://seer.cancer.gov/))] was used to extract the data.

Abbreviations: RS, Renal sarcoma; SEER, Surveillance, Epidemiology, and End Results; OS, Overall survival; CSS, Cancer-specific survival; DCA, Decision curve analysis; ICD-O, The International Classification of Diseases for Oncology; ROC, Receiver operating characteristic; AUC, Area under the curve; HR, Hazard ratios; CI, Confidence intervals.

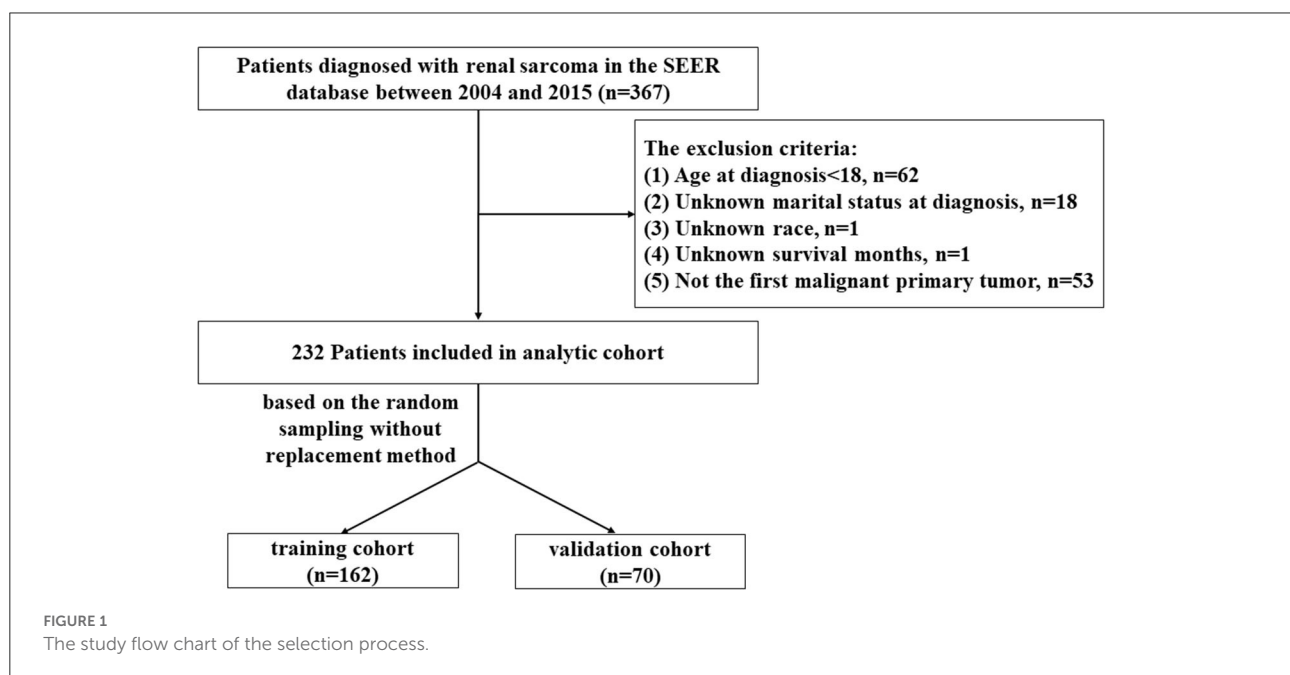


TABLE 1 Baseline demographic and clinical characteristics with adult renal sarcoma patients in our study.

Characteristic	Total no. (%)	The training cohort No. (%)	The validation cohort No. (%)	<i>P</i> value
Total	232 (100)	162 (70.0)	70 (30.0)	
Age, years				0.219
≤60	120 (51.7)	79 (48.8)	41 (58.6)	
>60	112(48.3)	83 (51.2)	29 (41.4)	
Year of diagnosis				0.627
2004–2009	115 (49.6)	82 (50.6)	33 (47.1)	
2010–2015	117 (50.4)	80 (49.4)	37 (52.9)	
Sex				0.814
Male	105 (45.3)	72 (44.4)	33 (47.1)	
Female	127 (54.7)	90 (55.6)	37 (52.9)	
Marital status				0.330
Married	133 (57.3)	89 (54.9)	44 (62.9)	
Unmarried	99 (42.7)	73 (45.1)	26 (37.1)	
Race				0.606
White	188 (81.1)	134 (82.7)	54 (77.1)	
Black	27 (11.6)	17 (10.5)	10 (14.3)	
Others	17 (7.3)	11 (6.8)	6 (8.6)	
Grade				0.878
Grade I	22 (9.5)	14 (8.5)	8 (11.4)	
Grade II	22 (9.5)	16 (9.9)	6 (8.6)	
Grade III	39 (16.8)	28 (17.3)	11 (15.7)	
Grade IV	67 (28.9)	49 (30.2)	18 (25.7)	
Unknown	82 (35.3)	55 (34.0)	27 (38.6)	
Histological type				0.308
Liposarcoma	69 (29.7)	52 (32.1)	17 (24.3)	
Leiomyosarcoma	95 (40.9)	60 (37.0)	35 (50.0)	
Carcinosarcoma	10 (4.3)	6 (3.7)	4 (5.7)	
Rhabdomyosarcoma	4 (1.7)	3 (1.9)	1 (1.4)	
Clear cell sarcoma	19 (8.3)	12 (7.4)	7 (10.0)	
Fibrosarcoma	2 (0.9)	2 (1.2)	0	
Sarcoma, NOS	33 (14.2)	27 (16.7)	6 (8.6)	
SEER stage				0.178
Localized	79 (34.1)	56 (34.6)	23 (32.9)	
Regional	70 (30.2)	43 (26.5)	27 (38.6)	
Distant	73 (31.5)	54 (33.3)	19 (27.1)	
Unstaged	10 (4.2)	9 (5.6)	1 (1.4)	
Surgery				0.274
Yes	52 (22.4)	40 (24.7)	12 (17.1)	
No/Unknown	180 (77.6)	122 (75.3)	58 (82.9)	
Radiotherapy				1.000
Yes	198 (85.3)	138 (85.2)	60 (85.7)	
No/Unknown	34 (14.7)	24 (14.8)	10 (14.3)	
Chemotherapy				0.788
Yes	178 (76.7)	123 (75.9)	55 (78.6)	
No/Unknown	54 (23.3)	39 (24.1)	15 (21.4)	

SEER, Surveillance, Epidemiology, and End Results. Percentages may not total 100 because of rounding.

TABLE 2 Univariate and multivariate analysis of overall survival (OS) rates in the training cohort.

Characteristic	Univariate analysis		Multivariate analysis	
	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Age, years				
≤60	Reference		Reference	
>60	1.469 (1.000–2.159)	0.050	-	0.098
Year of diagnosis				
2004–2009	Reference			
2010–2015	0.652 (0.418–1.015)	0.058		
Sex				
Male	Reference		Reference	
Female	0.601 (0.409–0.881)	0.009	0.498 (0.328–0.756)	0.001
Marital status				
Married	Reference		Reference	
Unmarried	0.869 (0.592–1.277)	0.475	-	0.745
Race				
White	Reference		Reference	
Black	0.453 (0.219–0.935)	0.032	-	0.038
Others	1.217 (0.561–2.641)	0.619	-	0.868
Grade				
Grade I	Reference		Reference	
Grade II	1.489 (0.355–6.240)	0.586	-	0.091
Grade III	3.301 (0.960–11.349)	0.058	-	0.752
Grade IV	5.251 (1.619–17.025)	0.006	-	0.205
Unknown	5.779 (1.791–18.647)	0.003	-	0.498
Histological type				
Liposarcoma	Reference		Reference	
Leiomyosarcoma	1.364 (0.839–2.219)	0.210	1.406 (0.854–2.315)	0.181
Carcinosarcoma	7.253 (2.936–17.919)	<0.001	6.996 (2.703–18.107)	<0.001
Rhabdomyosarcoma	2.590 (0.783–8.562)	0.119	3.797 (1.127–12.789)	0.031
Clear cell sarcoma	1.063 (0.480–2.353)	0.880	0.542 (0.232–1.266)	0.157
Fibrosarcoma	0.701 (0.095–5.170)	0.728	0.374 (0.050–2.809)	0.339
Sarcoma, NOS	1.910 (1.061–3.437)	0.031	1.563 (0.843–2.898)	0.157
SEER stage				
Localized	Reference		Reference	
Regional	2.769 (1.599–4.794)	<0.001	3.623 (2.047–6.410)	<0.001
Distant	4.793 (2.861–8.029)	<0.001	4.317 (2.487–7.494)	<0.001
Unstaged	2.444 (0.924–6.462)	0.072	1.936 (0.645–5.805)	0.239
Surgery				
No/Unknown	Reference		Reference	
Yes	0.478 (0.313–0.728)	0.001	0.515 (0.313–0.847)	0.009
Radiotherapy				
Yes	Reference		Reference	
No/Unknown	0.875 (0.521–1.471)	0.615	-	0.771
Chemotherapy				
Yes	Reference		Reference	
No/Unknown	0.679 (0.444–1.038)	0.074	-	0.348

CSS, Cancer-specific survival; SEER, Surveillance, Epidemiology, and End Results; HR, hazard ratio; CI, confidence interval.

TABLE 3 Univariate and multivariate analysis of cancer-specific survival (CSS) rates in the training cohort.

Characteristic	Univariate analysis		Multivariate analysis	
	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Age, years				
≤60	Reference		Reference	
>60	2.101 (1.144–3.859)	0.017	-	0.083
Year of diagnosis				
2004–2009	Reference			
2010–2015	0.761 (0.557–1.040)	0.087		
Sex				
Male	Reference		Reference	
Female	0.838 (0.466–1.507)	0.556	-	0.767
Marital status				
Married	Reference		Reference	
Unmarried	0.714 (0.394–1.294)	0.267	-	0.256
Race				
White	Reference		Reference	
Black	0.538 (0.192–1.508)	0.238	-	0.433
Others	0.749 (0.180–3.117)	0.691	-	0.652
Grade				
Grade I	-		-	
Grade II	Reference		Reference	
Grade III	2.378 (0.493–11.461)	0.280	-	0.919
Grade IV	4.481 (1.044–19.228)	0.044	-	0.153
Unknown	3.624 (0.836–15.713)	0.085	-	0.698
Histological type				
Liposarcoma	Reference		Reference	
Leiomyosarcoma	2.088 (0.802–5.437)	0.132	2.225 (0.839–5.901)	0.108
Carcinosarcoma	24.382 (7.227–82.262)	<0.001	23.815 (6.516–87.039)	<0.001
Rhabdomyosarcoma	3.799 (0.456–31.661)	0.217	9.022 (0.995–81.826)	0.051
Clear cell sarcoma	3.740 (1.198–11.676)	0.023	2.686 (0.825–8.740)	0.101
Fibrosarcoma	3.026 (0.363–25.224)	0.306	4.303 (0.446–41.551)	0.207
Sarcoma, NOS	5.748 (2.165–15.262)	<0.001	4.816 (1.712–13.547)	<0.001
SEER stage				
Localized	Reference		Reference	
Regional	3.106 (1.214–7.948)	0.018	3.926 (1.492–10.328)	0.006
Distant	7.031 (2.988–16.547)	<0.001	5.867 (2.301–14.962)	<0.001
Unstaged	4.656 (1.201–18.049)	0.026	1.800 (0.379–8.557)	0.460
Surgery				
No/Unknown	Reference		Reference	
Yes	0.350 (0.191–0.639)	0.001	0.352 (0.168–0.739)	0.006
Radiotherapy				
Yes	Reference		Reference	
No/Unknown	0.950 (0.425–2.127)	0.901	-	0.518
Chemotherapy				
Yes	Reference		Reference	
No/Unknown	0.889 (0.450–1.756)	0.735	2.315 (1.065–5.033)	0.034

CSS, Cancer-specific survival; SEER, Surveillance, Epidemiology, and End Results; HR, hazard ratio; CI, confidence interval.

Patients

A total of 367 patients diagnosed with RS between 2004 and 2015 were established according to the International Classification of Disease for Oncology, Third Edition [ICD-O-3] site codes, including liposarcoma (8850/3, 8851/3, 8852/3, 8853/3, 8858/3, 8860/3), leiomyosarcoma (8890/3, 8891/3, 8896/3), carcinosarcoma (8980/3), rhabdomyosarcoma (8900/3, 8901/3, 8910/3), clear cell sarcoma (8964/3), fibrosarcoma (8810/3), sarcoma, NOS (8800/3). The exclusion criteria are based on the following principles: (1) age at diagnosis is below

18 years old, $n = 62$; (2) unknown marital status at diagnosis, $n = 18$; (3) unknown Race, $n = 1$; (4) unknown Survival months, $n = 1$; (5) not the first malignant primary tumor, $n = 53$. Finally, 232 eligible patients were included in the analytic cohort. The flow chart of the selection process was presented in Figure 1.

Variables and endpoints

The following variables were filtered from the SEER database: age, year of diagnosis, sex, marital status, race,

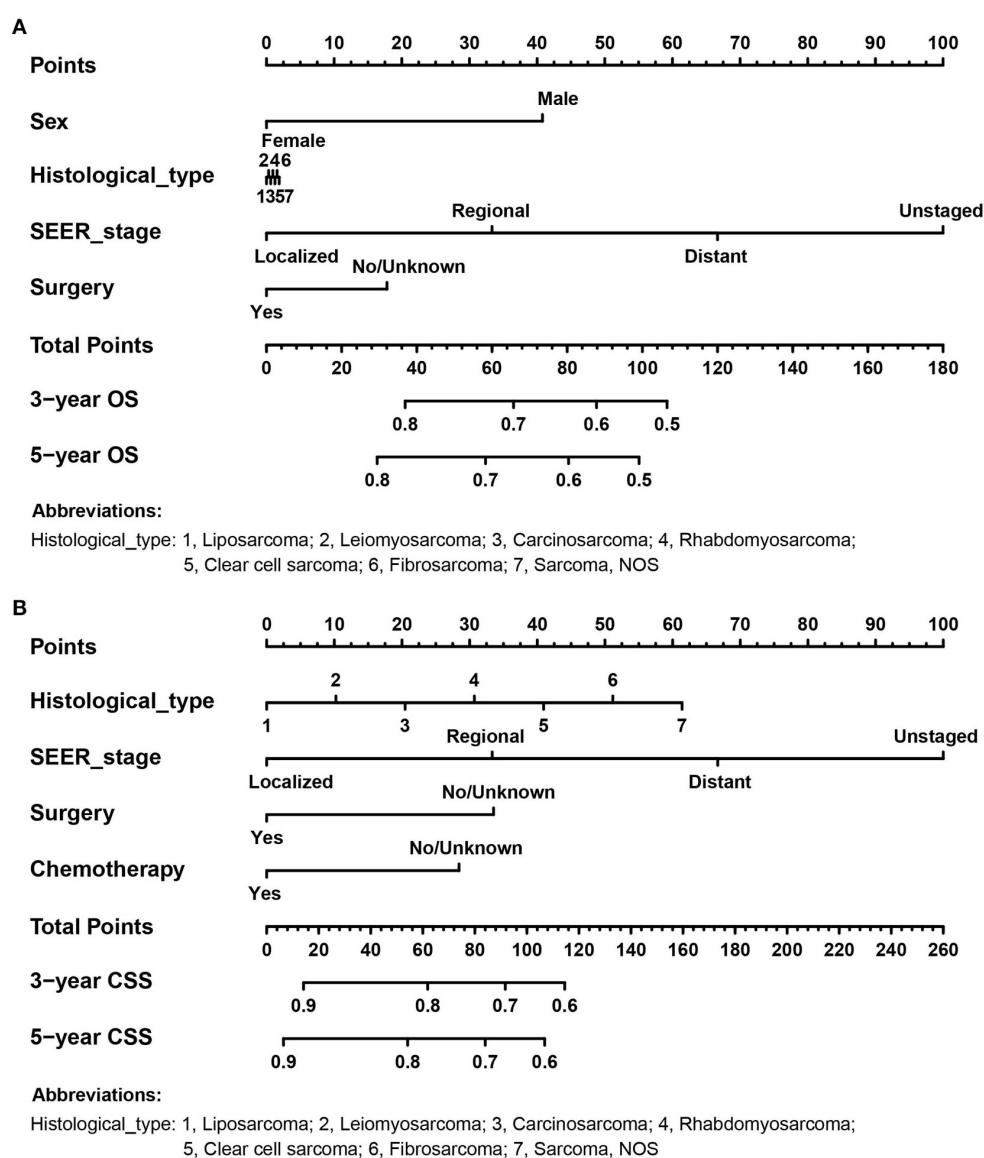


FIGURE 2
The prognostic nomograms for predicting 3- and 5- OS and CSS probabilities of adult RS patients in the training cohort. (A) OS nomogram; (B) CSS nomogram.

grade, histological type, SEER stage, surgery, radiotherapy, chemotherapy. To facilitate the next step of data analysis, the categorical variables were coded directly, and for continuous variables, they were first converted to categorical variables before coding. Some of the variables are explained below:

1. Regarding age, patients were divided into two categories: older than 60 years and ≤ 60 .
2. Regarding year of diagnosis, it was divided into two phases: 2004–2009, 2010–2015.
3. Regarding grade, it was defined as follows: well-differentiated (Grade I); moderately differentiated (Grade II); poorly differentiated (Grade III); undifferentiated (Grade IV); and unknown grade.
4. Regarding the stage of SEER, patients were classified into four subgroups according to the progression of

the sarcoma, including localized, regional, distant, and unstaged.

The death and RS-specific death were regarded as observed endpoints. OS refers to the period between the start of the study and death from any cause, and survivors are censored as of the last follow-up. CSS refers to the period between the commencement of the study and the death due to RS, with deaths due to other causes or survivors omitted.

Statistical methods

Categorical data were described as numbers (n) and percentages (%), and chi-square tests were used to assess differences in categorical variables. The sample was divided

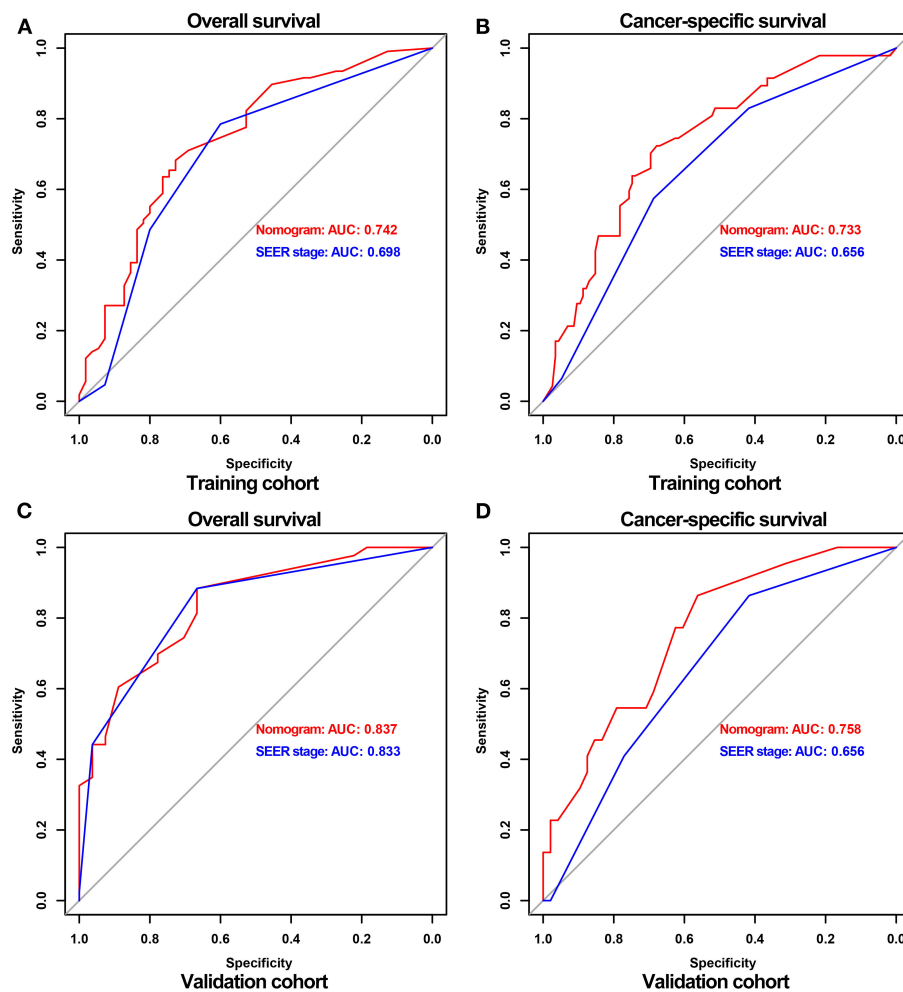


FIGURE 3

ROC curves of nomograms and the SEER stage for predicting OS and CSS probabilities in the training and validation cohort. ROC for OS (A) and CSS (B) in the training cohort, respectively; ROC for OS (C) and CSS (D) in the validation cohort, respectively.

into a training cohort and a validation cohort (in a ratio of 7:3) using a no-replacement random sampling method. The training cohort was used to create nomograms and filter factors for nomograms, while the validation cohort was used to validate the results of the training cohort. Univariate Cox regression was used to identify factors associated with OS and CSS, and multivariate Cox regression to identify associated independent risk factors. Variables with P values <0.05 in univariate Cox regression analysis were included in multivariate Cox regression analysis, and associated hazard ratios (HR) and 95% confidence intervals (CI) were calculated. Based on the results of multivariate Cox regression analysis, independent risk factors were used to create prognostic nomograms to predict the probability of OS and CSS at 3 and 5 years. In addition, receiver operating characteristic (ROC) curves, decision curve analysis (DCA), and calibration curves were used to assess the predictive performance of the nomogram and SEER stage.

A vertical line was drawn on the scale for each variable for a given adult RS patient, and the intersection with the “dot” line represented the score for that variable. The total score is calculated by adding up the scores for each variable. Matching scores were found on the “total score” line and projected onto the OS and CSS lines below, resulting in 3- and 5-OS and CSS probabilities for that individual.

In ROC curve analysis, the area under the curve (AUC) is defined as the area enclosed by the ROC curve and the coordinate axes. The value of the AUC usually ranges between 0.5 and 1, and the diagnostic value of the nomogram is represented by the AUC. In the calibration curve analysis, a bootstrap method with 1,000 resamples was used for testing.

SPSS 26.0 (IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.) was applied to conduct statistical analysis for univariate and multivariate Cox regression. The nomograms were developed and validated by exerting the rms, hmisc, lattice, survival, formula, ggplot2, pROC, timeROC, and rmda packages in R version 4.1.2 (<http://www.r-project.org/>). $P < 0.05$ (two-sided) was considered statistically significant.

Results

Baseline demographic and clinical characteristics

A total of 232 eligible patients diagnosed with RS between 2004 and 2015 were included in our study, which were divided into two cohorts randomly: the training cohort (162, 70.0%) and the validation cohort (70, 30.0%). The number of RS patients aged over or equal to 60 and under 60 was similar in the total cohort. Similarly, the number of patients with the year of diagnosis in 2004–2009 and in 2010–2015 was approximately equal. Most RS patients were female (54.7%), married (57.3%),

and white (81.1%). Grade IV accounts for the largest proportion of known grades. Of the other general type, the majority were leiomyosarcoma (40.9%) and localized (34.1%). Most RS adult patients received radiotherapy (85.3%) and chemotherapy (76.7%), but only a few had undergone surgery (22.4%). Specific baseline demographic and clinical characteristics information are represented in Table 1.

Univariate and multivariate analysis of OS and CSS

The univariate and multivariate Cox regression analysis of OS and CSS rates in the training cohort was carried out for screening independent prognostic variables. Age, year of diagnosis, sex, marital status, race, grade, histological type, SEER stage, surgery, radiotherapy, and chemotherapy were included in our analysis. By univariate regression analysis, it was shown that all variables mentioned above might be substantially linked with OS and CSS. Meanwhile, it was also shown that sex, histological type, SEER stage, and surgery were independent predictive variables for OS by multivariate analysis, while histological type, SEER stage, surgery, and chemotherapy were independent prognostic variables for CSS. Confidence intervals (CI) and corresponding p -values for specific variables in the univariate and multivariate analyses of OS and CSS were summarized in Tables 2, 3, respectively.

Nomogram development and validation

According to the independent prognostic variables of OS and CSS, the nomograms were established, respectively (Figure 2). In the OS nomogram, the SEER stage contributed the most to survival outcome, while the histological type contributed the least. In the CSS nomogram, the SEER stage was the most significant predictor of survival, followed by histological type.

As shown in Figure 3, the ROC curves were drawn, and the AUC of the OS nomogram was significantly greater than that of the SEER stage in the training cohort (nomogram 0.742, SEER stage 0.698), while in the validation cohort the AUC of the OS nomogram was similar to SEER stage (nomogram 0.837, SEER 0.833). However, the AUCs of the nomograms for CSS were considerably higher than those of the SEER stage both in the training cohort (nomogram 0.733, SEER stage 0.656) and validation cohort (nomogram 0.758, SEER stage 0.656). By comprising of the above ROC curves, it was demonstrated that the nomogram had more diagnostic value than the SEER stage to discriminate the survival probability of adult RS patients.

The AUCs for 3- and 5-OS were 0.751 and 0.757, respectively, and 0.779 and 0.750 for 3- and 5-CSS, respectively, in the training cohort. The validation cohort AUCs for 3- and 5-OS were 0.775 and 0.829, respectively, and 0.807 and

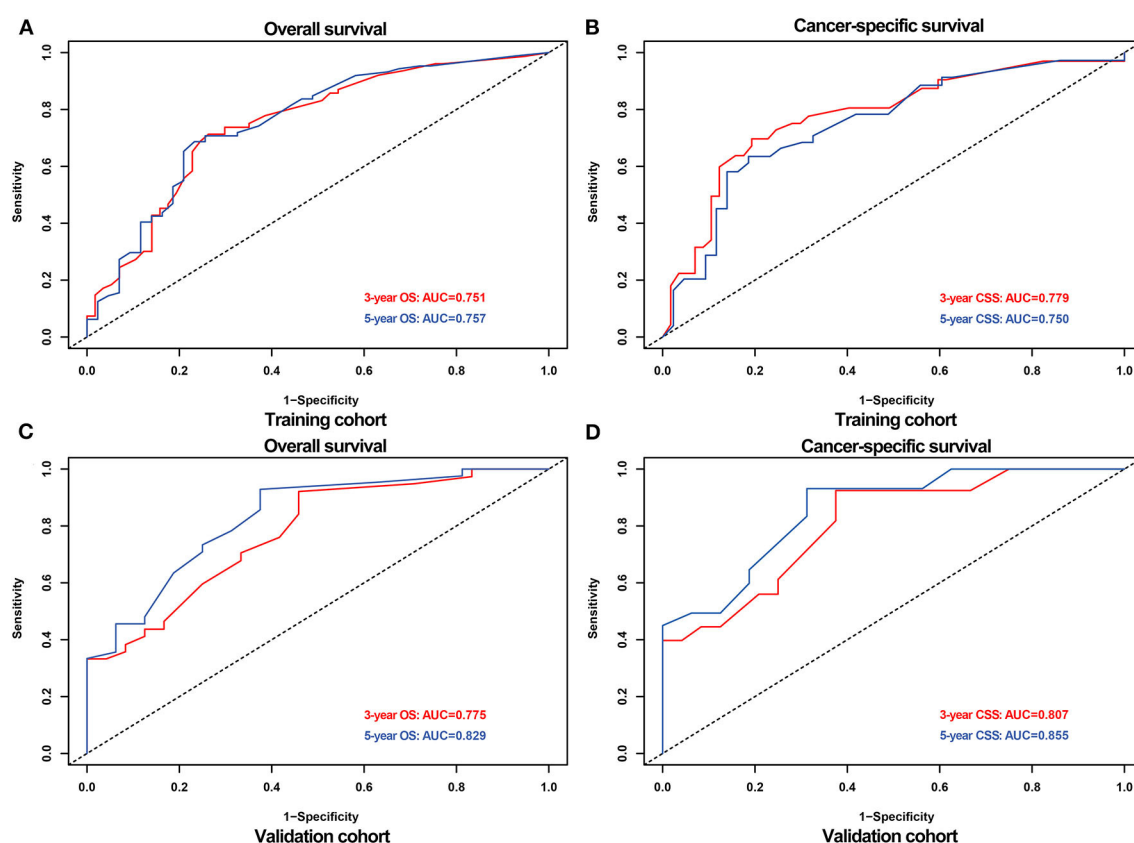


FIGURE 4
ROC curves for predicting 3-, 5- OS and CSS probabilities in the training and validation cohort. ROC for 3-, 5- OS (A) and CSS (B) in the training cohort; ROC for 3-, 5- OS (C) and CSS (D) in the validation cohort.

0.855, respectively, for 3- and 5-CSS. As shown in Figure 4, the nomograms accurately predict the probability of 3- and 5- OS and CSS for adult RS patients.

The calibration curves of the nomograms showed high consistencies between the predicted and actual survival rates both in the training and validation cohorts, illustrated in Figure 5 and Supplementary Figure 1. The gray line in the calibration curves represents the ideal reference line, where the predicted survival probability matches the actual survival probability. The presentation of the nomograms was represented by red dots. The DCA demonstrated that the nomograms in the wide high-risk threshold had a higher net benefit than the SEER stage (Figure 6), which validated the superiority of the nomogram utility over the SEER stage in clinical practice.

Discussion

As mentioned above, adult renal sarcomas are an extremely rare group of tumors, accounting for only 0.8% of primary renal tumors (3). The SEER stage grading system was used

by urologists to evaluate the progression of renal sarcomas. Sarcomas are classified into different grades based on the location and the extent to which it invades organs, blood vessels, and lymph nodes, including localized, regional, distant, and unstaged. However, due to the influence of individual differences, such as sex, age, race, marital status, radiation, chemotherapy, surgery, etc., it is not comprehensive enough to use the extent of tumor invasion alone to evaluate the prognosis for adult RS patients.

The nomogram is a graphical representation of a clinical prediction model that calculates a total score based on the values of individual predictor variables, and then predicts the risk of an event or the probability of survival based on the total score (15, 16). It is a novel prediction model that is gradually sought after by clinicians. In recent years, predictions for the prognosis of various urinary cancer with nomograms have been reported more and more. For instance, Wu et al. employed a genomic-clinicopathologic nomogram to predict preoperative lymph node metastasis in bladder cancer (17); A nomogram was conducted by Mao et al. to predict prognosis in patients with lung metastatic renal cell carcinoma (18). Zhang et al.

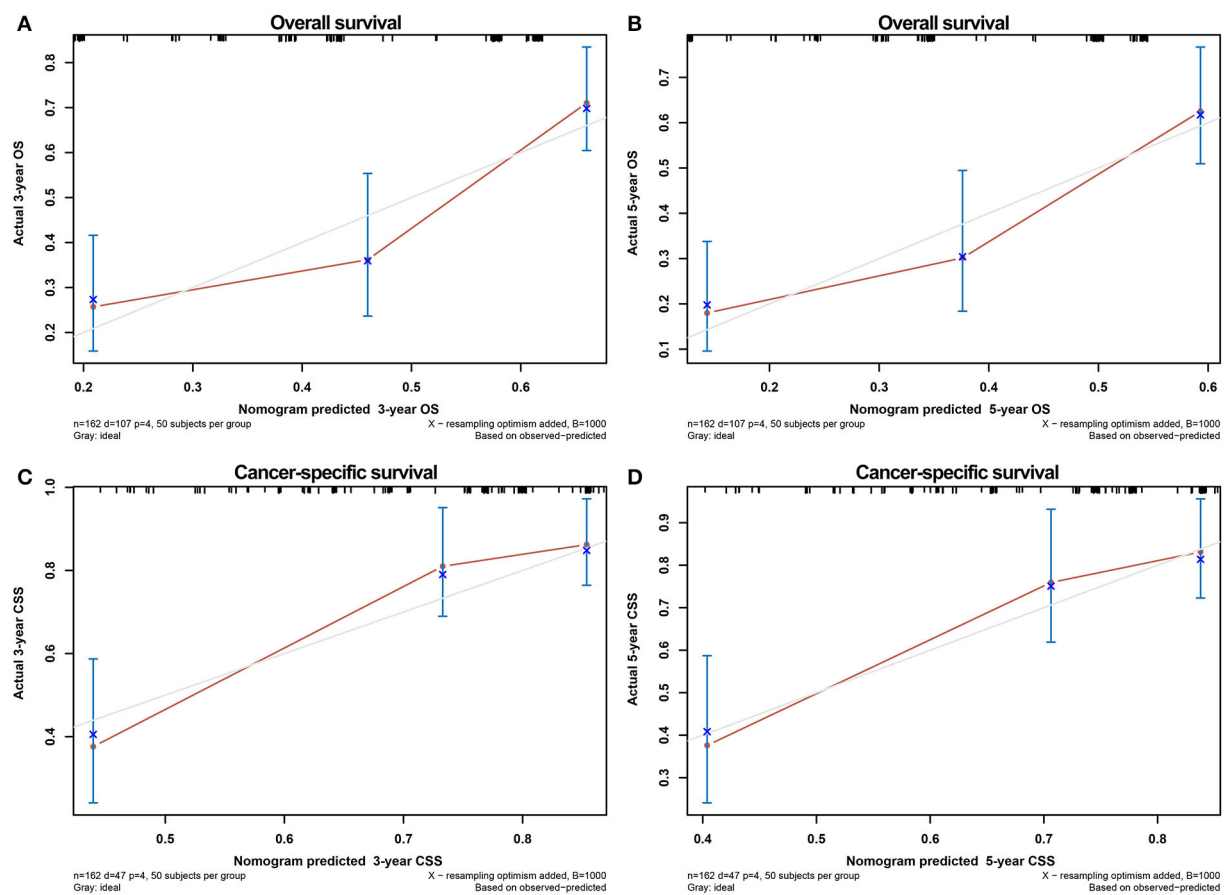


FIGURE 5

Calibration curves for verifying the consistency between predicted 3-, 5- OS and CSS and actual 3-, 5- OS and CSS in the training cohort. 3- OS (A) and 5- OS (B) calibration curves; 3- CSS (C) and 5- CSS (D) calibration curves.

established a radiomics nomogram to predict bone metastasis in newly diagnosed prostate cancer patients (19). The nomogram and Aggtrmmns scoring system were utilized by Zhou et al. for predicting overall survival and cancer-specific survival of kidney cancer patients (20).

As it is known that compared with the SEER stage, nomogram has the following advantages: (1) By combining various independent risk factors according to the patient's condition, it allows for a more intuitive assessment and individualization of the patient's prognosis (21). (2) It quantifies the possibility of OS and CSS in patients, permitting a more precise prognostic evaluation (22). Therefore, for the first time, the prognostic nomograms were developed for adult RS patients to obtain personalized and accurate prognostic predictions in this study.

We extracted data from the SEER database for adult RS patients and used COX univariate and subsequent multivariate regression analysis to conclude that histological type, SEER stage, surgery were independent risk factors for OS and CSS. Based on the multivariate regression analysis, the OS and CSS

nomograms were constructed, respectively. Subsequently, we validated the nomograms. The area under the ROC curves for 3-, 5- OS were 0.775 and 0.829, respectively, and 0.807 and 0.855 for 3-, 5- CSS, respectively, which depicted that the nomograms accurately predict the probability of 3- and 5- OS and CSS for adult RS patients. The calibration curves showed high consistencies between the predicted and actual survival rates.

From the nomograms, it was suggested that RS patients without surgery, with distant SEER stage grade, and histological type of carcinosarcoma had the poorest prognosis. According to the Kaplan-Meier overall and disease-specific survival analysis of patients with RS established by Nazemi et al. (1), liposarcoma had the greatest prognosis, followed by leiomyosarcoma and clear cell sarcoma, while carcinosarcoma had the worst prognosis. Some studies have shown that carcinosarcoma had the worst prognosis, which was consistent with our analysis. In addition, in our study, univariate Cox regression analysis found that Sarcoma, NOC patients had poorer OS compared to liposarcoma patients ($HR = 1.910$, 95% CI 1.061–3.437, $p = 0.031$). However, after multivariate Cox regression analysis,

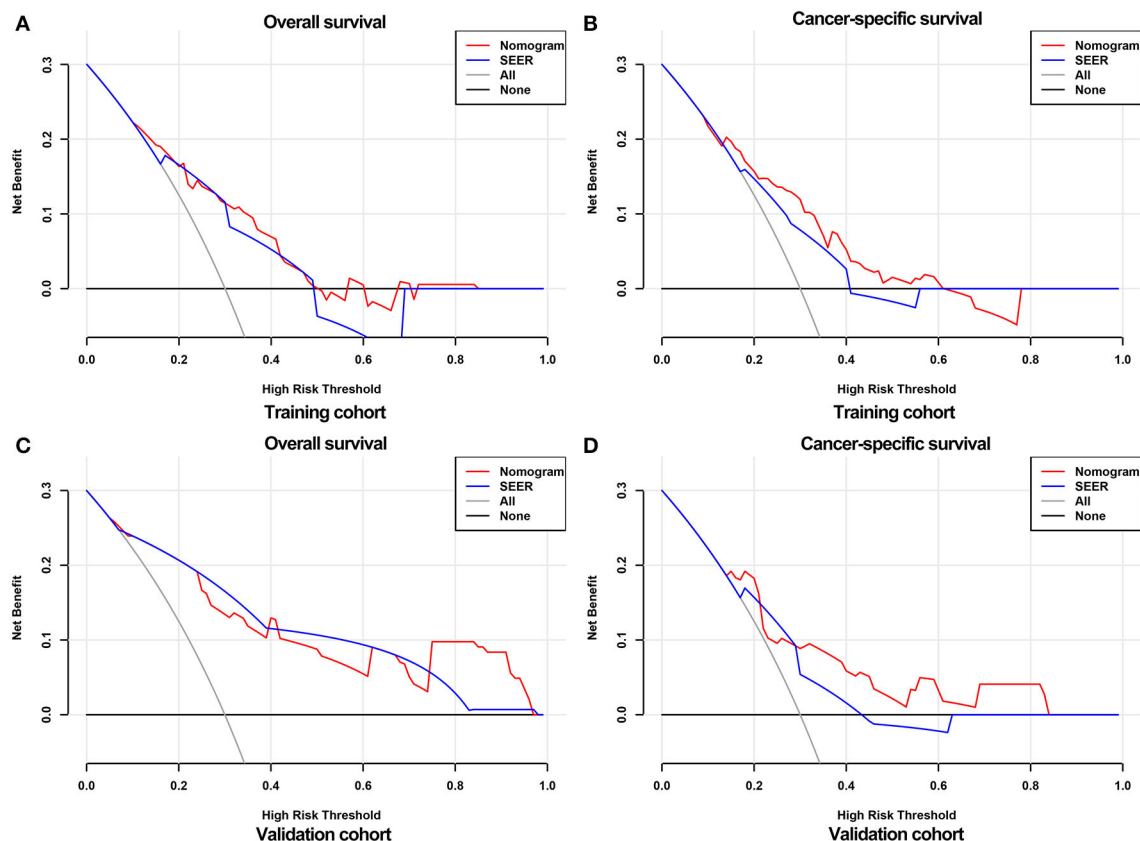


FIGURE 6

DCA curves for validating the clinical utility of the nomograms. DCA curves for OS (A) and CSS (B) in the training cohort. DCA curves for OS (C) and CSS (D) in the validation cohort.

there was no difference between the two groups, which could be explained by the inclusion of other confounding variables, which led to biased results.

In addition, our study demonstrated that surgical treatment for adult RS patients may effectively reduce the risk of death. This is in line with the findings of Moreira et al. (14) and Öztürk (23). Moreover, it is also found that chemotherapy may also improve the prognosis of adult RS patients. Chemotherapy has now been applied clinically to treat advanced or recurrent renal sarcoma, although not standardized (24), and the latest research of Yakirevich et al. suggested that comprehensive genomic analysis of adult RS patients may provide new opportunities for targeted therapy (25).

To our surprise, our data suggested radiotherapy was not an independent prognostic factor for the adult patient with renal sarcoma, which was in accordance with the findings of Li et al. (26). However, Gamboa et al. reported that preoperative radiotherapy may improve the prognosis by making some tumors easier to resect (27). Thus, the prognostic impact of radiotherapy on patients with renal sarcomas should be further explored. The clinical outcome of primary adult renal sarcoma is extremely poor and the optimal treatment remains to be

debated. Further studies are needed to verify whether it is surgery or combination therapy that works best. Furthermore, our data also suggested that female patients had a better prognosis than male patients, which could be attributed to differences in female anatomy or hormone levels.

We appraised the prognosis of adult RS patients with nomograms for the first time, which adds a new dimension to our research. Simultaneously, using the SEER database excluded the influencing factors of single-center. Even so, there are still a few flaws in our study: (1) Because of the rarity of renal sarcoma, limited sample size is inevitable and therefore our findings may not be representative; (2) As our study is retrospective, there is a lack of multicenter data for external validation. (3) Due to the lack of data in the SEER database, genetic factors, laboratory findings, and medication history were not included in our study.

Conclusions

In conclusion, a prognostic nomogram was created to predict overall survival (OS) and cancer-specific survival (CSS) for adult patients with RS, and their reliability and usefulness

were also validated in our study. We anticipate that our study will facilitate urologists in accurately assessing the prognosis of adult RS patients and provide support for further clinical trials.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/[Supplementary material](#).

Author contributions

YZ, MC, JW, YM, and QW: study designing. MC, JW, YM, and QW: administrative support. YZ, WM, GZ, SS, ST, and TJ: data collection. YZ, WM, GZ, and SS: statistical analysis and graphs production. YZ, WM, and GZ: draft writing. YZ, WM, GZ, SS, ST, TJ, QW, YM, JW, and MC: final revision. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by Natural Science Foundation of China (82170703 to GZ and 82100732), Natural Science Foundation of Jiangsu Province (BK20200360), and Excellent Youth Development Fund of Zhongda Hospital, SEU (2021ZDYYYQPY04).

References

- Nazemi A, Daneshmand S. Adult genitourinary sarcoma: a population-based analysis of clinical characteristics and survival. *Urol Oncol.* (2020) 38:334–43. doi: 10.1016/j.urolonc.2019.12.004
- Vogelzang NJ, Fremgen AM, Guinan PD, Chmiel JS, Sylvester JL, Sener SF. Primary renal sarcoma in adults. A natural history and management study by the American Cancer Society, Illinois Division. *Cancer.* (1993) 71:804–10. doi: 10.1002/1097-0142(19930201)71:3<804::aid-cnrcr2820710324>3.0.co;2-a
- Wang X, Xu R, Yan L, Zhuang J, Wei B, Kang D, et al. Adult renal sarcoma: clinical features and survival in a series of patients treated at a high-volume institution. *Urology.* (2011) 77:836–41. doi: 10.1016/j.urol.2010.09.028
- Mayes DC, Fechner RE, Gillenwater JY. Renal liposarcoma. *Am J Surg Pathol.* (1990) 14:268–73. doi: 10.1097/0000478-199003000-00008
- Ding Y, Quan C. Primary renal leiomyosarcoma. *J Coll Physicians Surg Pak.* (2021) 30:725–7. doi: 10.29271/jcpsp.2021.06.725
- Chiu KC, Lin MC, Liang YC, Chen CY. Renal carcinosarcoma: case report and review of literature. *Ren Fail.* (2008) 30:1034–9. doi: 10.1080/08860220802403192
- Sola JE, Cova D, Casillas J, Alvarez OA, Qualman S, Rodriguez MM. Primary renal botryoid rhabdomyosarcoma: diagnosis and outcome. *J Pediatr Surg.* (2007) 42:e17–20. doi: 10.1016/j.jpedsurg.2007.08.011
- Aw SJ, Chang KTE. Clear cell sarcoma of the kidney. *Arch Pathol Lab Med.* (2019) 143:1022–6. doi: 10.5858/arpa.2018-0045-RS
- Agarwal K, Singh S, Pathania OP. Primary renal fibrosarcoma: a rare case report and review of literature. *Indian J Pathol Microbiol.* (2008) 51:409–10. doi: 10.4103/0377-4929.42541
- Chen C, Jiang X, Xia F, Chen X, Wang W. Clinicopathological Characteristics and Survival Outcomes of Primary Renal Leiomyosarcoma. *Front Surg.* (2021) 8:704221. doi: 10.3389/fsurg.2021.704221
- Zhuge Y, Cheung MC, Yang R, Perez EA, Koniaris LG, Sola JE. Pediatric non-Wilms renal tumors: subtypes, survival, and prognostic indicators. *J Surg Res.* (2010) 163:257–63. doi: 10.1016/j.jss.2010.03.061
- Chen JY, Chen JJ, Yang BL, Liu ZB, Huang XY, Liu GY, et al. Predicting sentinel lymph node metastasis in a Chinese breast cancer population: assessment of an existing nomogram and a new predictive nomogram. *Breast Cancer Res Treat.* (2012) 135:839–48. doi: 10.1007/s10549-012-2219-x
- Lee CK, Asher R, Friedlander M, Gebisi V, Gonzalez-Martin A, Lortholary A, et al. Development and validation of a prognostic nomogram for overall survival in patients with platinum-resistant ovarian cancer treated with chemotherapy. *Eur J Cancer.* (2019) 117:99–106. doi: 10.1016/j.ejca.2019.05.029
- Moreira DM, Gershman B, Thompson RH, Okuno SH, Robinson SI, Leibovich BC, et al. Clinicopathologic characteristics and survival for adult renal sarcoma: a population-based study. *Urol Oncol.* (2015) 33:505 e15–20. doi: 10.1016/j.urolonc.2015.07.022
- Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol.* (2008) 26:1364–70. doi: 10.1200/JCO.2007.12.9791

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.942608/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Calibration curves for verifying the consistency between predicted 3-, 5- OS and CSS and actual 3-, 5- OS and CSS in the validation cohort. 3- OS (A) and 5- OS (B) calibration curves; 3- CSS (C) and 5- CSS (D) calibration curves.

16. Mao W, Wang K, Xu B, Zhang H, Sun S, Hu Q, et al. ciRS-7 is a prognostic biomarker and potential gene therapy target for renal cell carcinoma. *Mol Cancer*. (2021) 20:142. doi: 10.1186/s12943-021-01443-2
17. Wu SX, Huang J, Liu ZW, Chen HG, Guo P, Cai QQ, et al. A genomic-clinicopathologic nomogram for the preoperative prediction of lymph node metastasis in bladder cancer. *Ebiomedicine*. (2018) 31:54–65. doi: 10.1016/j.ebiom.2018.03.034
18. Mao W, Fu Z, Wang K, Wu J, Xu B, Chen M. Prognostic nomogram for patients with lung metastatic renal cell carcinoma: a SEER-based study. *Ann Palliat Med*. (2021) 10:2791–804. doi: 10.21037/apm-20-1488
19. Zhang W, Mao N, Wang Y, Xie H, Duan S, Zhang X, et al. A Radiomics nomogram for predicting bone metastasis in newly diagnosed prostate cancer patients. *Eur J Radiol*. (2020) 128:109020. doi: 10.1016/j.ejrad.2020.109020
20. Zhou Y, Zhang R, Ding Y, Wang Z, Yang C, Tao S, et al. Prognostic nomograms and Aggtrmmns scoring system for predicting overall survival and cancer-specific survival of patients with kidney cancer. *Cancer Med*. (2020) 9:2710–22. doi: 10.1002/cam4.2916
21. Hou G, Zheng Y, Wei D, Li X, Wang F, Tian J, et al. Development and validation of a SEER-based prognostic nomogram for patients with bone metastatic prostate cancer. *Medicine (Baltimore)*. (2019) 98:e17197. doi: 10.1097/MD.00000000000017197
22. Wang K, Xu X, Xiao R, Du D, Wang L, Zhang H, et al. Development and validation of a nomogram to predict cancer-specific survival in patients with hypopharyngeal squamous cell carcinoma treated with primary surgery. *J Int Med Res*. (2021) 49:3000605211067414. doi: 10.1177/03000605211067414
23. Ozturk H. Prognostic features of renal sarcomas (Review). *Oncol Lett*. (2015) 9:1034–8. doi: 10.3892/ol.2014.2838
24. Blas L, Roberti J. Primary renal synovial sarcoma and clinical and pathological findings: a systematic review. *Curr Urol Rep*. (2021) 22:25. doi: 10.1007/s11934-021-01038-w
25. Yakirevich E, Madison R, Fridman E, Mangray S, Carneiro BA, Lu S, et al. Comprehensive genomic profiling of adult renal sarcomas provides insight into disease biology and opportunities for targeted therapies. *Eur Urol Oncol*. (2021) 4:282–8. doi: 10.1016/j.euo.2019.04.002
26. Li L, Liang J, Song T, Yin S, Zeng J, Zhong Q, et al. A nomogram model to predict prognosis of patients with genitourinary sarcoma. *Front Oncol*. (2021) 11:656325. doi: 10.3389/fonc.2021.656325
27. Gamboa AC, Gronchi A, Cardona K. Soft-tissue sarcoma in adults: an update on the current state of histiotype-specific management in an era of personalized medicine. *CA Cancer J Clin*. (2020) 70:200–29. doi: 10.3322/caac.21605



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Jaydip Sen,
Praxis Business School, India
Tae Keun Yoo,
B&VIIT Eye Center, South Korea
Shaik Razia,
K L University, India

*CORRESPONDENCE

Pu Liao
liaopu@ucas.ac.cn

†These authors have contributed
equally to this work and share first
authorship

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 03 June 2022

ACCEPTED 23 August 2022

PUBLISHED 14 September 2022

CITATION

Guo Y-y, Li Z-j, Du C, Gong J, Liao P,
Zhang J-x and Shao C (2022) Machine
learning for identifying benign and
malignant of thyroid tumors: A
retrospective study of 2,423 patients.
Front. Public Health 10:960740.
doi: 10.3389/fpubh.2022.960740

COPYRIGHT

© 2022 Guo, Li, Du, Gong, Liao, Zhang
and Shao. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Machine learning for identifying benign and malignant of thyroid tumors: A retrospective study of 2,423 patients

Yuan-yuan Guo^{1†}, Zhi-jie Li^{1†}, Chao Du², Jun Gong³,
Pu Liao^{1*}, Jia-xing Zhang¹ and Cong Shao⁴

¹Department of Laboratory Medicine, Chongqing General Hospital, Chongqing, China, ²Department of Laboratory Medicine, Fuling Center Hospital of Chongqing City, Chongqing, China, ³Department of Information Center, University-Town Hospital of Chongqing Medical University, Chongqing, China, ⁴Department of Breast and Thyroid Surgery, Chongqing General Hospital, Chongqing, China

Thyroid tumors, one of the common tumors in the endocrine system, while the discrimination between benign and malignant thyroid tumors remains insufficient. The aim of this study is to construct a diagnostic model of benign and malignant thyroid tumors, in order to provide an emerging auxiliary diagnostic method for patients with thyroid tumors. The patients were selected from the Chongqing General Hospital (Chongqing, China) from July 2020 to September 2021. And peripheral blood, *BRAFV600E* gene, and demographic indicators were selected, including sex, age, *BRAFV600E* gene, lymphocyte count (Lymph#), neutrophil count (Neu#), neutrophil/lymphocyte ratio (NLR), platelet/lymphocyte ratio (PLR), red blood cell distribution width (RDW), platelets count (PLT), red blood cell distribution width—coefficient of variation (RDW—CV), alkaline phosphatase (ALP), and parathyroid hormone (PTH). First, feature selection was executed by univariate analysis combined with least absolute shrinkage and selection operator (LASSO) analysis. Afterward, we used machine learning algorithms to establish three types of models. The first model contains all predictors, the second model contains indicators after feature selection, and the third model contains patient peripheral blood indicators. The four machine learning algorithms include extreme gradient boosting (XGBoost), random forest (RF), light gradient boosting machine (LightGBM), and adaptive boosting (AdaBoost) which were used to build predictive models. A grid search algorithm was used to find the optimal parameters of the machine learning algorithms. A series of indicators, such as the area under the curve (AUC), were intended to determine the model performance. A total of 2,042 patients met the criteria and were enrolled in this study, and 12 variables were included. Sex, age, Lymph#, PLR, RDW, and *BRAFV600E* were identified as statistically significant indicators by univariate and LASSO analysis. Among the model we constructed, RF, XGBoost, LightGBM and AdaBoost with the AUC of 0.874 (95% CI, 0.841–0.906), 0.868 (95% CI, 0.834–0.901), 0.861 (95% CI, 0.826–0.895), and 0.837 (95% CI, 0.802–0.873) in the first model. With the AUC of 0.853 (95% CI, 0.818–0.888), 0.853 (95% CI, 0.818–0.889), 0.837 (95% CI, 0.800–0.873), and 0.832 (95% CI, 0.797–0.867) in the second model. With the AUC of 0.698 (95% CI, 0.651–0.745), 0.688 (95% CI, 0.639–0.736), 0.693

(95% CI, 0.645–0.741), and 0.666 (95% CI, 0.618–0.714) in the third model. Compared with the existing models, our study proposes a model incorporating novel biomarkers which could be a powerful and promising tool for predicting benign and malignant thyroid tumors.

KEYWORDS

thyroid tumor, machine learning, predictive model, *BRAFV600E* gene mutation, risk-factors

Introduction

The incidence of thyroid tumors has been increasing over the past 20 years, and it was the eighth most commonly diagnosed tumors in the world among endocrine tumors (1–3). According to the National Cancer Registry, thyroid tumors in China will continue to grow at an annual rate of 20% (4, 5). Therefore, identifying benign and malignant tumors owns great significance for early clinical intervention and treatment. Although ultrasonography and fine needle aspiration biopsy (FNAB) cytology methods can diagnose most thyroid tumors, there were still some patients who were misdiagnosed or overtreated. In addition, the limitations of those examinations included the need for a highly experienced cytopathologist for accurate interpretation, and not suitable for early screening of disease.

At present, many biomarkers of thyroid tumors have been discovered by researchers. Ozmen found that higher NLR and PLR were associated with worse survival in differential thyroid tumors (6). Another study from Turkey suggested that mean platelet volume (MPV) levels can be used as an easily available biomarker for monitoring the risk of papillary thyroid carcinoma (PTC) in patients with thyroid nodules, enabling early diagnosis of PTC (7). And Liu found that lower pretreatment platelet count (PLT) levels may indicate a poor prognosis for PTC (8). In particular, the *BRAFV600E* gene is also an important biomarker for the occurrence and progression of papillary thyroid tumors (9). In addition, the review by Qian and Iryani mentions that many genetic biomarkers can differentiate benign from malignant thyroid tumors (10, 11). However, most studies just investigated the diagnostic performance of individual biomarkers, and few studies have integrated these biomarkers to construct models that can be used to diagnose benign and malignant thyroid tumors in clinical practice. Previous studies have the shortcomings of small sample size and large differences in diagnostic performance between different biomarkers.

Machine learning (ML) is an emerging artificial intelligence discipline that analyzes multiple data types and uses them to explore disease risk factors, accurate diagnosis, and prognosis (12). Moreover, it can integrate multiple clinical indicators,

explore the nonlinear relationship between predictors and clinical outcomes, and solve problems such as poor performance of logistic methods in traditional clinical modeling. Sui developed a deep-learning AI model (ThyNet) using ultrasound images to differentiate between malignant tumors and benign thyroid nodules with an AUC of 0.875 (95% CI, 0.871–0.880) (13). Although there have been some studies using ML algorithms to diagnose benign and malignant thyroid tumors, the data selected are mostly image data, which makes data collection more complicated.

Therefore, this study aims to apply ML algorithms to build a predictive model of thyroid tumors with demographic, peripheral blood laboratory, and genetic biomarkers to provide an accurate and reliable prediction method for the early discrimination of benign and malignant thyroid tumors.

Methods

Study participants

Patients with thyroid tumor included in the current study, were selected from the Chongqing General Hospital (Chongqing, China) from July 2020 to September 2021. According to WHO 2017 classification and the eighth edition of the AJCC/TNM classification (TNM-8) (14), operating records and final pathologic reports were reviewed to ascertain tumor categories, they were divided into benign groups and malignant groups. Benign groups are defined as thyroid follicular nodular disease, follicular adenoma, follicular adenoma with papillary architecture, oncocytic adenoma of the thyroid, and benign thyroid nodules. While, malignant groups are defined as follicular thyroid carcinoma, invasive encapsulated follicular variant papillary carcinoma, papillary thyroid carcinoma, oncocytic carcinoma of the thyroid, follicular-derived carcinomas, high-grade, and anaplastic follicular cell-derived thyroid carcinoma (15).

This study was exempt from ethical review by the Institutional Review of the Chongqing General Hospital. The study methods were carried out in accordance with the relevant guidelines and regulations.

TABLE 1 Clinical characteristics and variables of patients in all cohorts.

Predictors	Benign (N = 561)	Malignant (N = 1,481)	P-value
Sex (%)			
Male	105 (18.7)	357 (24.1)	0.011
Female	456 (81.3)	1,124 (75.9)	
BRAFV600E (%)			
Mutation	76 (13.5)	1,170 (79.0)	<0.001
Wild	485 (86.5)	311 (21.0)	
Age (years)	45.00 [35.00, 52.00]	39.00 [32.00, 50.00]	<0.001
Lymph# ($\times 10^9/L$)	1.64 [1.37, 2.01]	1.58 [1.29, 1.94]	
Neu# ($\times 10^9/L$)	3.64 [2.85, 4.65]	3.60 [2.84, 4.57]	0.991
NLR	2.13 [1.69, 2.85]	2.20 [1.70, 2.96]	
PLR	130.06 [103.38, 157.24]	140.00 [110.36, 172.27]	<0.001
RDW (%)	42.30 [40.60, 43.90]	41.90 [40.50, 43.40]	
PLT ($\times 10^9/L$)	215.00 [184.00, 251.00]	222.00 [187.00, 260.00]	0.061
RDW-CV	12.90 [12.50, 13.40]	12.80 [12.50, 13.30]	
ALP (U/L)	67.00 [59.00, 78.14]	67.00 [56.00, 81.00]	0.395
PTH (ng/ml)	49.20 [43.90, 53.75]	48.50 [37.80, 58.90]	

Candidate predictors

The data was collected from the electronic medical record (EMR) system of the Chongqing General Hospital, which contains laboratory examination records, diagnosis and treatment process records, doctor orders, etc. Patient's peripheral blood indicators, *BRAFV600E* gene, and demographic indicators were selected, including age, sex, lymphocyte count (Lymph#), neutrophil count (Neu#), red blood cell distribution width (RDW), red blood cell distribution width - coefficient of variation (RDW-CV), platelets count (PLT), neutrophil/lymphocyte ratio (NLR), platelet/lymphocyte ratio (PLR), alkaline phosphatase (ALP), parathyroid hormone (PTH), and *BRAFV600E* gene mutation as predictors to build a ML model to identify benign and malignant thyroid tumors. All the peripheral blood tests and *BRAFV600E* gene results were obtained at the first examination after the patient was admitted to the hospital.

The *BRAFV600E* gene mutation was detected by real-time PCR using the ABI QuantStudio[®]5 Real-Time PCR System, according to the manufacturer's instructions (Human BRAFV600E Mutation assay Kit, YZY MED, Wuhan, China). The DNA from FNAB specimen was extracted using a companion kit, which was provided by the same manufacturer. The DNA concentration was quantified in a Nano-300 Micro Spectrophotometer (ALLSHENG Instrument Co., Ltd. Hangzhou, China) as per the manufacturer's instructions. The DNA was immediately used to carry out the test of *BRAFV600E* gene mutation.

Statistical analysis

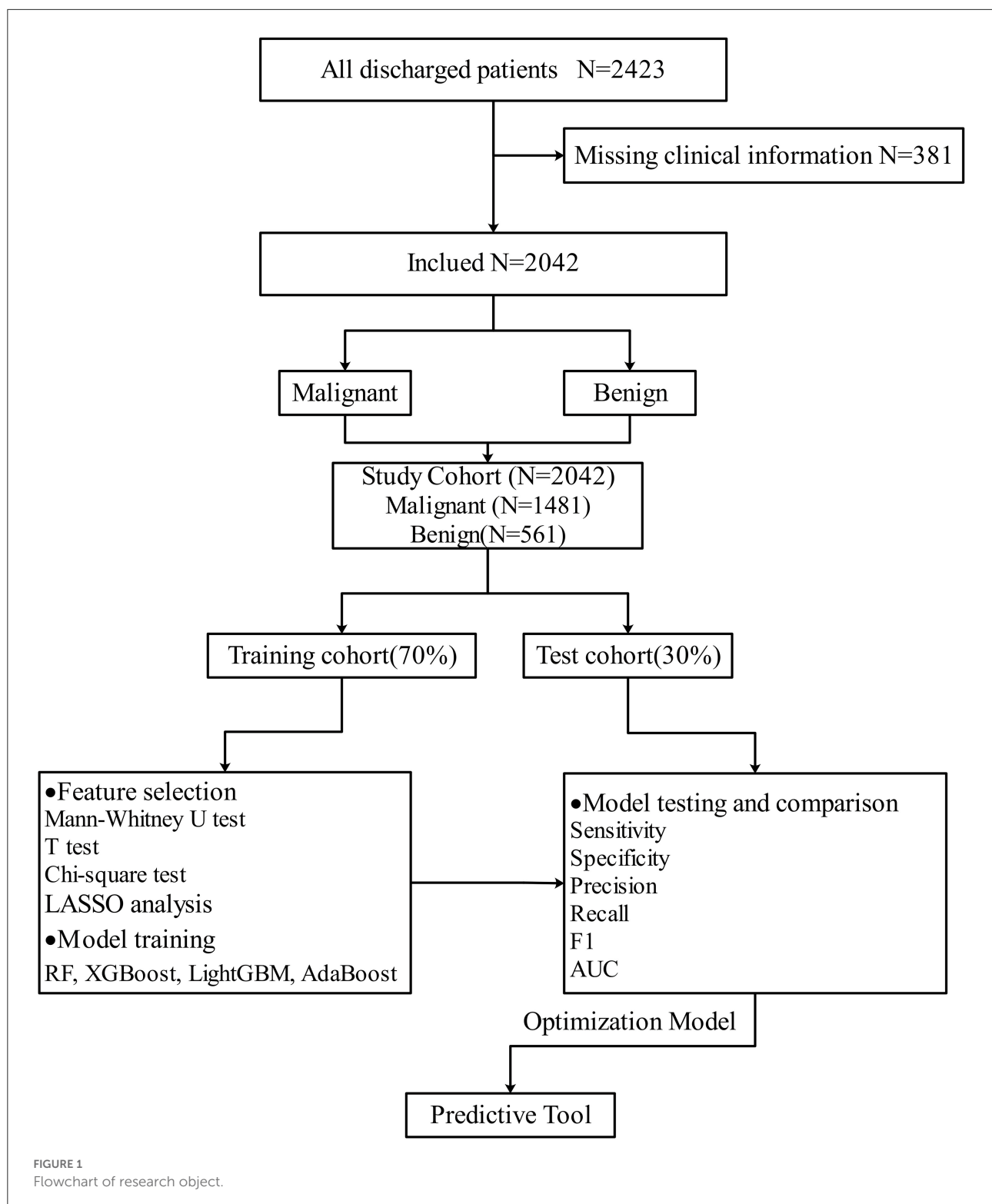
All the statistical analyses and model building were conducted in R for windows (version 4.0.1, <https://www.r-project.org/>). For information on hardware devices in the development environment, please see [Supplementary Table 1](#).

The data were presented as count with percentage for categorical variables, median with interquartile range (IQR), or mean with SD for continuous variables. For the variables with miss rate <30%, missforest algorithm was used to fill. First, the Mann-Whitney *U*-test or *t*-test was performed for the continuous variables, and the chi-square test for categorical variables was carried out used for univariate analysis. The variables after univariate analysis were analyzed by the least absolute shrinkage and selection operator (LASSO). Afterward, random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and adaptive boosting (AdaBoost) were used to establish prediction models. We used the grid search algorithm to find the optimal parameters of each algorithm to optimize the performance of the model. Sensitivity (SEN), specificity (SPE), precision, recall, F1, and the area under the curve (AUC) were intended to determine the model performance.

Result

Sample collection

A total of 2,423 patients met the inclusion criteria and were enrolled in the study. In total, 381 patients were excluded due

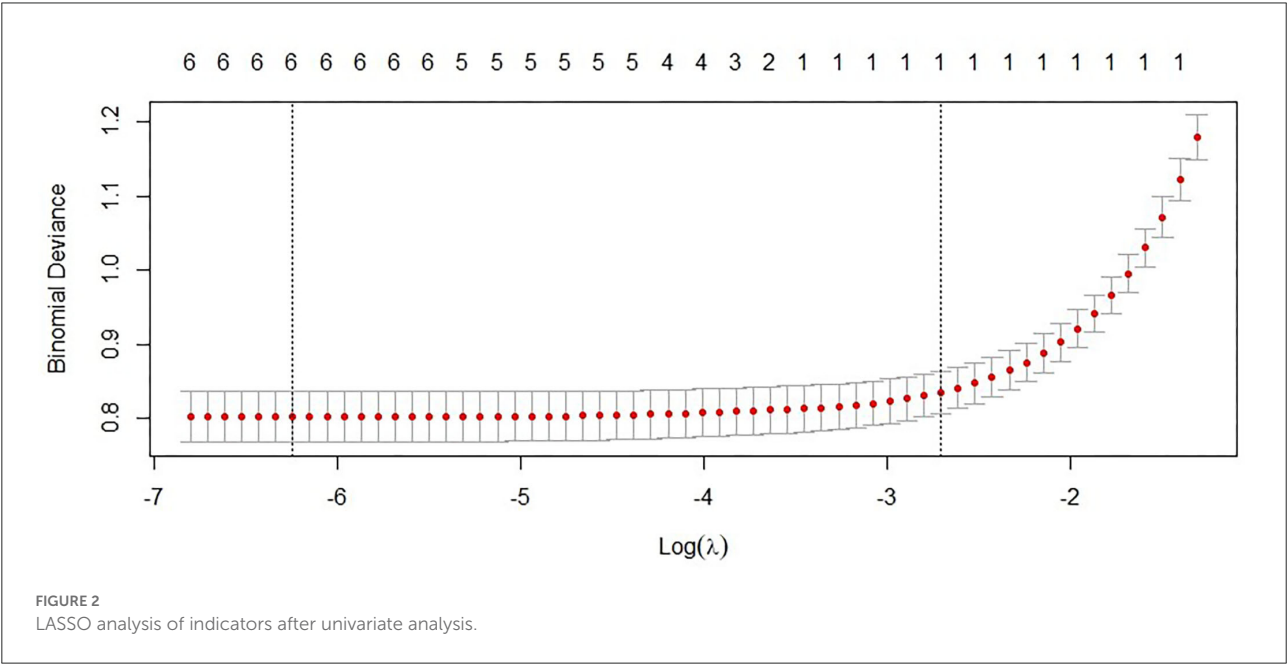


to missing clinical data. At last, a total of 2,042 patients with 12 predictors were included in the final study. [Table 1](#) shows the information of the whole cohort. In the whole cohort, 1,481

malignant patients and 561 benign patients were included. The average age of patients was 42.03 ± 11.30 years, ranging from 14 to 76 years, women accounted for 77.34% (1,580 cases) and men

TABLE 2 Clinical characteristics and variables of patients in training cohort and test cohort.

Predictors	Training cohort			Test cohort		
	Benign (N = 395)	Malignant (N = 1,034)	P-value	Benign (N = 166)	Malignant (N = 447)	P-value
Sex (%)						
Male	70 (17.7)	247 (23.9)	0.015	35 (21.1)	110 (24.6)	0.421
Female	325 (82.3)	787 (76.1)		131 (78.9)	337 (75.4)	
BRAFV600E (%)						
Mutation	55 (13.9)	822 (79.5)	<0.001	21 (12.7)	348 (77.9)	<0.001
Wild	340 (86.1)	212 (20.5)		145 (87.3)	99 (22.1)	
Age (years)	45.00 [36.00, 52.00]	39.00 [33.00, 50.00]	<0.001	44.00 [34.00, 52.00]	38.00 [32.00, 49.00]	0.008
Lymph# (×10 ⁹ /L)	1.64 [1.39, 2.00]	1.56 [1.28, 1.92]	0.001	1.65 [1.35, 2.05]	1.61 [1.31, 1.96]	0.189
Neu# (×10 ⁹ /L)	3.62 [2.83, 4.65]	3.58 [2.83, 4.54]	0.925	3.66 [2.93, 4.66]	3.64 [2.88, 4.63]	0.877
NLR	2.14 [1.69, 2.91]	2.21 [1.71, 2.98]	0.074	2.11 [1.70, 2.73]	2.18 [1.70, 2.95]	0.48
PLR	131.40 [103.99, 160.30]	140.70 [110.93, 173.62]	<0.001	127.47 [101.35, 155.48]	138.33 [109.73, 170.43]	0.013
RDW (%)	42.40 [40.90, 44.00]	41.90 [40.40, 43.48]	<0.001	41.95 [40.30, 43.58]	41.90 [40.50, 43.20]	0.816
PLT (×10 ⁹ /L)	215.00 [185.00, 253.00]	221.00 [186.00, 259.00]	0.222	215.00 [183.00, 248.75]	225.00 [190.00, 261.00]	0.121
RDW-CV	12.90 [12.50, 13.40]	12.80 [12.50, 13.30]	0.387	12.80 [12.40, 13.20]	12.80 [12.50, 13.30]	0.709
ALP (U/L)	67.00 [59.26, 78.28]	66.80 [56.00, 80.89]	0.23	66.44 [58.77, 78.00]	68.00 [56.00, 82.00]	0.791
PTH (ng/ml)	49.03 [43.50, 53.73]	48.70 [37.80, 58.80]	0.925	49.44 [44.19, 53.82]	47.68 [37.90, 59.45]	0.498



22.66% (463 cases). The specific screening process and study protocol are shown in [Figure 1](#).

Model building

The data were split into a training cohort (70%, $N = 1,429$) and a test cohort (30%, $N = 613$) by random number table. In

the training cohort, there were 395 cases of the benign group and 1,034 cases of the malignant group. In the test cohort, there were 166 cases of the benign group and 447 cases of the malignant group. The predictors we collected were used as input variables of ML algorithms. Whether malignancy or benign was regarded as the outcome event (yes = 1, no = 0) to establish prediction model by using training cohort, and the test cohort was used to verify the ability of the established prediction model previously.

TABLE 3 The optimal parameters of the three models.

Categories	Algorithm	Parameter
The first model	RF	mtry = 1, ntree = 60, nodesize = 8
	XGBoost	max_depth = 3, eta = 0.6, nrounds = 5
	LightGBM	nrounds = 20, min_data = 1, learning_rate = 0.1
	AdaBoost	mfinal = 170
The second model	RF	mtry = 6, ntree = 140, nodesize = 12
	XGBoost	max_depth = 4, eta = 0.3, nrounds = 3
	LightGBM	nrounds = 10, min_data = 3, learning_rate = 0.1
	AdaBoost	mfinal = 20
The third model	RF	mtry = 1, ntree = 90, nodesize = 10
	XGBoost	max_depth = 6, eta = 0.7, nrounds = 3
	LightGBM	nrounds = 10, min_data = 3, learning_rate = 0.4
	AdaBoost	mfinal = 5

According to Table 2, univariate analysis results indicated that 6 predictors were statistically significant between the malignant group and benign group in training cohort. We performed the LASSO analysis on the 6 indicators with statistically significant, and the results showed that these 6 indicators were all selected by LASSO (Figure 2). Therefore, our final diagnostic model included the 6 indicators of sex, age, Lymph#, PLR, RDW, and BRAFV600E.

We built 3 ML models with different predictors, the first model included all the predictors we included, the second model included predictors after feature selection, and the third model included patient peripheral blood predictors. For the specific construction steps of the model, please see Supplementary Figure 1, and the detailed description of the three models can be found in Supplementary Table 2. In addition, we also used the grid search algorithm to find the optimal parameters of the ML algorithm. The grid search algorithm permutes and combines each possible parameter value, and then substitutes the results of all possible combinations into the algorithm for model training. The optimal parameter combination was selected from all possible parameter combinations. In our research, we selected the optimal parameters of four ML algorithms: RF, XGBoost, LightGBM, and Adaboost through the grid search algorithm. Please see Table 3 for the optimal parameters of each algorithm.

Performance evaluated in different models

In Table 4, the metrics of three models were compared in terms of SEN, SPE, AUC, etc., in the test cohort. The SEN and precision are indicators to measure the positive predictive performance of the model. In the first and second models, the SEN indicator exceeds 0.7, and the precision indicator reaches 0.9, suggesting that the model we established can well identify malignant patients from thyroid tumor patients. The SPE is an indicator of the model's negative predictive performance, and in our study, the highest SPE was 0.892, indicating that our model could also predict patients with benign thyroid tumor well. The AUC is a comprehensive indicator for comparing prediction performance. Among the three models constructed with different predictors, the first model including all predictors performed best with the highest AUC of 0.874 (95% CI, 0.841, 0.906). The second model had the highest AUC of 0.853 (95% CI, 0.818, 0.889; Figure 3). However, we performed the Delong test on the optimal AUC of the first and second models ($z = 1.65$, $P = 0.099$), and the results showed that the difference was not statistically significant. The third model selects peripheral blood predictors, and the best AUC is 0.698 (95% confidence interval, 0.651, 0.745). In the third model, we selected biomarkers in patients' peripheral blood to establish a prediction model, and the performance of the model is inferior to the first and second models. Biomarkers in peripheral blood are easy to obtain, and the AUC of the model is close to 0.7, suggesting that it also has a certain predictive value.

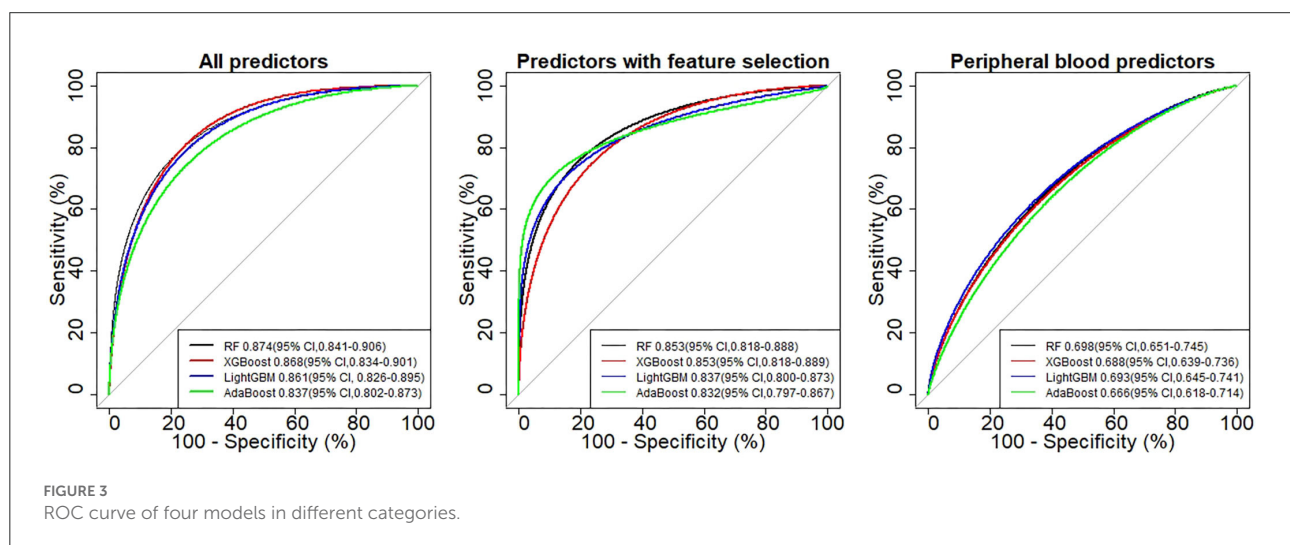
To balance the diagnostic performance and simplicity of the model, according to the comprehensive evaluation of the performance indicators of the model and the Delong test analysis, the second model, using the RF algorithm, was the best at predicting benign and malignant thyroid tumors. The importance ranking of predictors in the RF algorithm is as follows: BRAFV600E, age, PLR, RDW, Lymph#, and sex (Figure 4).

Discussion

In this study, we developed the ML-based predictive models to identify benign and malignant thyroid nodules. The current gold diagnostic standard for thyroid tumors meeting appropriate criteria is a cyto-pathologic assessment of FNAB. However, high operator requirements were needed in FNAB, and the accuracy of diagnosis largely depends on the operator's personal level of experience. Therefore, it is crucial to provide more objective and direct parameters that can help with the identification of benign and malignant thyroid lesions. Thus, predictors including BRAFV600E gene mutation, Lymph#, Neu#, RDW, PLT, NLR, PLR, ALP, PTH, and clinical characters

TABLE 4 Performance evaluation table of three models.

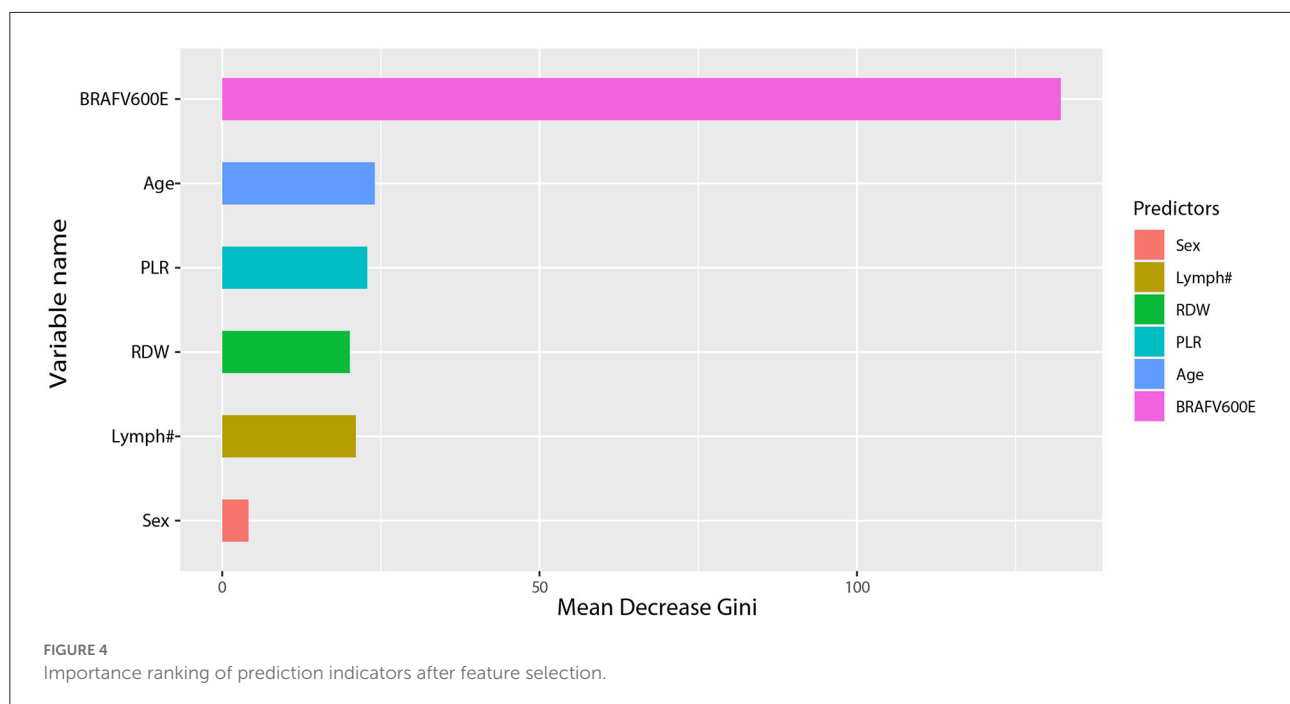
Categories	Algorithm	SEN	SPE	Precision	Recall	F1	AUC (95%CI)
The first model	RF	0.790	0.886	0.949	0.790	0.862	0.874 (0.841–0.906)
	XGBoost	0.790	0.873	0.944	0.790	0.860	0.868 (0.834–0.901)
	LightGBM	0.734	0.892	0.948	0.734	0.827	0.861 (0.826–0.895)
	AdaBoost	0.723	0.855	0.931	0.720	0.812	0.837 (0.802–0.873)
The second model	RF	0.781	0.873	0.943	0.781	0.854	0.853 (0.818–0.888)
	XGBoost	0.754	0.873	0.941	0.754	0.837	0.853 (0.818–0.889)
	LightGBM	0.765	0.873	0.942	0.765	0.844	0.837 (0.800–0.873)
	AdaBoost	0.779	0.880	0.946	0.779	0.854	0.832 (0.797–0.867)
The third model	RF	0.671	0.645	0.836	0.671	0.744	0.698 (0.651–0.745)
	XGBoost	0.781	0.548	0.823	0.781	0.801	0.688 (0.639–0.736)
	LightGBM	0.624	0.705	0.849	0.626	0.721	0.693 (0.645–0.741)
	AdaBoost	0.626	0.651	0.828	0.626	0.713	0.666 (0.618–0.714)



of patients were enrolled and the ML algorithm was used to predict benign and malignant thyroid tumors in our study.

Recent advances in understanding the molecular pathogenesis of thyroid tumors have enabled the application of molecular tests to provide more objective information and play a role in making more personalized clinical treatments (16). A large number of biomarkers such as BRAFV600E, RAS, EIF1AX, PIK3CA, PTEN and AKT1, SWI/SNF, ALK, and CDKN2A, have been excavated, demonstrating the potential of molecular diagnostic detection (17). Nevertheless, the BRAFV600E is the most prevalent mutation detected in PTC, with an average frequency of 60%–70%, and the tests for BRAFV600E mutation are commonly available in the current clinical practice (18). The BRAFV600E protein kinase has received extensive attention because of its function in promoting cell proliferation, growth, and division, and numerous studies have investigated the relationship between the BRAFV600E mutations and various

clinicopathological features. *In vitro* tests have shown a high concordance between the BRAFV600E mutations and the aggressive characteristics of PTC, while clinical trials have shown contrasting results, making it controversial whether the BRAFV600E mutations can be used as an aggressive marker for PTC. Most studies suggest that the BRAFV600E mutations are associated with worse clinical pathology, such as lymph node metastasis, distant metastasis, worse tumor stage, aggressive subtype, tumor size, male, and old age, and therefore, recommend the central lymph node dissection based on total thyroidectomy with more stringent radioiodine therapy and a close follow-up after surgery (19). However, some studies did not find such an association (20). The differences in these studies may be due to the different sample sizes included in the studies, epidemiological characteristics of the patients, papillary carcinoma subtypes, types of specimens used for molecular testing, and testing methods. In this study, the *BRAFV600E*



gene mutation status was important for all algorithms, which is consistent with a recent study. The BRAFV600E mutation has both high specificity and sensitivity to predict thyroid malignancy in the Chinese population. It can accurately complete cytopathology in the guidance of thyroid surgery (21). In our study, the diagnostic performance accuracy of the BRAFV600E gene was 0.810, and the AUC was 0.827, which had a high-diagnostic value.

The peripheral blood routine test and the blood biochemical test have major advantages over the traditional pathological test of tumor lesions in terms of quick and simple sample acquisition, low collection cost, minimal trauma, and preoperative detection, which should be paid more attention to in research (22). Lymph#, Neu#, RDW-CV, PLT, NLR, PLR, ALP, PTH, and other related indicators can quickly and accurately detect the values of blood, in order to effectively indicate abnormalities of infection, anemia, and cruor. In recent years, a wide variety of blood indicators with different changes were concerned and discussed in the study of malignant tumor diseases. The preoperative NLR and RDW-CV are convenient, practical, and easily measured biomarkers for clinical diagnosis and prognostic assessment of patients with esophageal cancer. Moreover, the NLR was more effective than RDW-CV, acting as an independent prognostic biomarker for esophageal cancer (23). On the contrary, the RDW-CV has attracted more attention in cervical, ovarian, and endometrial cancer as studies have shown the hierarchical independent relationship between the RDW and these kinds of cancers (24). The preoperative blood count from peripheral blood

may provide prognostic value in patients with pathologic stage I NSCLC undergoing surgical resection. Of significance in patients with pT1 N0 NSCLC, the high lymphocyte count and high platelet count were associated with higher recurrence (25). Even the NLR, PLR, and LMR, which are the derived indexes of peripheral whole blood cell counts, were developed into new indexes, and have fairly good values of prognostic(26–28). However, the values of NLR and PLR to distinguish between benign and malignant of thyroid nodules is still controversial. Our study found that the Lymph#, RDW-CV, and PLR were statistically different between benign and malignant thyroid nodules ($P < 0.05$).

Recently, the ML algorithms have been extensively used in the medical field, emerging as a powerful tool in dealing with many health care problems. In our study, the ML-based model for diagnosing benign and malignant thyroid tumors showed the highest AUC of 0.874 (95% CI, 0.841, 0.906), which suggests that our model has a high value in diagnosing benign and malignant thyroid tumors. To evaluate the accuracy and simplicity of the model, feature selection is often used to screen indicators with predictive value. We screened out six predictors from 12 predictors by the univariate analysis method. Compared with the inclusion of 12 predictors, the model established by these six predictors also has good predictive performance and was identified as the optimal model. From the perspective of algorithm selection, when the indicators contained in the model are consistent, the performance of the four algorithms is not significantly different. One of the reasons is that if there is a clear correlation between the independent and dependent

TABLE 5 Comparison of the newly created model with the existing model.

Title	Authors	Algorithms	Parameters	AUC
Machine Learning for Identifying Benign and Malignant of Thyroid Tumors: A Retrospective Study of 2,423 Patients (final model)	Yuan-yuan Guo et al	Machine learning (Random forest)	Sex, age, Lymph#, PLR, RDW, BRAFV600E	0.853 (95% CI, 0.818,0.888)
Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study(13)	Sui, Peng. et al	Deep learning (ResNet, ResNeXt, DenseNet)	Ultrasound images	0.875 (95% CI, 0.871–0.880)
Machine learning to identify lymph node metastasis from thyroid cancer in patients undergoing contrast-enhanced CT studies (29)	Masuda et al	machine learning (Support Vector Machines)	CT images	0.86
Deep convolutional neural network for classification of thyroid nodules on ultrasound: Comparison of the diagnostic performance with that of radiologists (30)	Yeonjae et al.	Deep learning	Images of underwent US-guided fine-needle aspiration biopsy	0.83–0.86
Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound (31)	Yeon et al.	Deep learning (Convolutional Neural Network)	Ultrasound image	0.845, 0.835, and 0.850
A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images (32)	Hong-Bo Zhao et al.	Deep learning (Convolutional Neural Network)	CT images	0.901–0.947

variables, then most ML algorithms can handle this nonlinear relationship and have good predictive performance. At present, many scholars have studied the use of artificial intelligence algorithms to accurately identify benign and malignant thyroid tumors (Table 5). The performance of our model is inferior to that of Hong-Bo Zhao, Sui, Peng et al., and similar to that of Masuda, Kim, Su Yeon Ko et al. Current researches mainly use ultrasound or CT images combined with intelligent algorithms to accurately diagnose benign and malignant thyroid tumors, and has excellent performance. In general, CT and ultrasound images have better predictive performance because they contain more information about benign and malignant tumors. However, from the perspective of patient's genetic markers and peripheral blood markers, our predictors are easy to obtain and has good value in identifying benign and malignant thyroid tumors.

In conclusion, the prediction model established in this study can distinguish benign with the risk of identifying malignant thyroid nodules, which could be further developed into a clinical decision support system. Our study also had some limitations. First, all of the data come from southwest China, so there may

be a selection bias. Second, only four algorithms were selected to establish the prediction model, therefore it is still necessary to try whether there are other better predictive algorithms. Third, the missing rate $\geq 30\%$ of the variables were not included in the study. Therefore, further analysis is required to identify these factors related to identifying benign and malignant of thyroid nodules.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Author contributions

Y-yG, Z-jL, and PL took part in the research design and helped to draft the manuscript. J-xZ and CS contributed the acquisition of data. CD and JG performed the statistical analysis. All

authors contributed to the article and approved the final manuscript.

Funding

This work was supported by a grant for the Science and Technology and Health Commission program of Chongqing (2020FYYX157).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Alessandro A, La MC. Novel therapeutic clues in thyroid carcinomas: the role of targeting cancer stem cells. *Med Res Rev.* (2017) 37:1299–317. doi: 10.1002/med.21448
- Ferrari SM, Fallahi P, Elia G, Ragusa F, Ruffilli I, Paparo SR, et al. Thyroid autoimmune disorders and cancer. *Semin Cancer Biol.* (2020) 64:135–46. doi: 10.1016/j.semcancer.2019.05.019
- Lin JS, Bowles EJA, Williams SB, Morrison CC. Screening for thyroid cancer: updated evidence report and systematic review for the US preventive services task force. *JAMA.* (2017) 317:1888–903. doi: 10.1001/jama.2017.0562
- Fang C, Juan X, Chuncheon S, Fengyan H, Lihua W, Yanli J, et al. Burden of thyroid cancer from 1990 to 2019 and projections of incidence and mortality until 2039 in China: findings from global burden of disease study. *Front Endocrinol.* (2021) 12:738213. doi: 10.3389/fendo.2021.738213
- Junyi W, Fangfang Y, Yanna S, Zhiguang P, Li L. Thyroid cancer: incidence and mortality trends in China, 2005–2015. *Endocrine.* (2020) 68:163–73. doi: 10.1007/s12020-020-02207-6
- Ozmen S, Timur O, Calik I, Altinkaynak K, Simsek E, Gozcu H, et al. Neutrophil-lymphocyte ratio (NLR) and platelet-lymphocyte ratio (PLR) may be superior to C-reactive protein (CRP) for predicting the occurrence of differentiated thyroid cancer. *Endocr Regul.* (2017) 51:131–6. doi: 10.1530/endoabs.41.EP1151
- Baldane S, Ipekci SH, Sozen M, Kebapcilar L. Mean platelet volume could be a possible biomarker for papillary thyroid carcinomas. *Asian Pac J Cancer Prev.* (2015) 16:2671–4. doi: 10.7314/APJCP.2015.16.7.2671
- Xiangxiang L, Huang Z, He X, Zheng X, Jia Q, Tan J, et al. Blood prognostic predictors of treatment response for patients with papillary thyroid cancer. *Biosci Rep.* (2020) 40:BSR20202544. doi: 10.1042/BSR20202544
- Shiyang L, Bo J, Shuyu L, Lu Z, Weihong Z, Kun W, et al. Oestrogen receptor alpha in papillary thyroid carcinoma: association with clinical features and BRAFV600E mutation. *Jpn J Clin Oncol.* (2021) 51:1051–8. doi: 10.1093/jjco/hyab058
- Iryani AM, Mat JS, Leong NK, Jacqueline JJ, Barani K, Haji HO, et al. Papillary thyroid cancer: genetic alterations and molecular biomarker investigations. *Int J Med Sci.* (2019) 16:450–60. doi: 10.7150/ijms.29935
- Qian X, Qiang J, Jian T, Zhaowei M. Serum biomarkers for thyroid cancer. *Biomark Med.* (2020) 14:807–15. doi: 10.2217/bmm-2019-0578
- Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* (2019) 20:e262–73. doi: 10.1016/S1470-2045(19)30149-4
- Sui P, Yihao L, Weiming L, Longzhong L, Qian Z, Hong Y, et al. (2021). Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digital Health.* 3:e250–9. doi: 10.1016/S2589-7500(21)00114-X
- Angela B, Ancu?a-Elena Z, Doina P, Elena B, Nicole B, Adela N, et al. A 15 year institutional experience of well-differentiated follicular cell-derived thyroid carcinomas; impact of the new 2017 TNM and WHO Classifications of Tumors of Endocrine Organs on the epidemiological trends and pathological characteristics. *Endocrine.* (2020) 67:630–42. doi: 10.1007/s12020-019-02158-7
- Baloch ZW, Asa SL, Barletta JA, Ghossein RA, Juhlin CC, Jung CK, et al. Overview of the 2022 WHO classification of thyroid neoplasms. *Endocr Pathol.* (2022) 33:27–63. doi: 10.1007/s12022-022-09707-3
- Sik PK, Hoon KS, Hun OJ, Young KS. Highly accurate diagnosis of papillary thyroid carcinomas based on personalized pathways coupled with machine learning. *Brief Bioinform.* (2020) 22:bbaa336. doi: 10.1093/bib/bbaa336
- Ichiro A, Yin LAK. Anaplastic thyroid carcinoma: current issues in genomics and therapeutics. *Curr Oncol Rep.* (2021) 23:31–31. doi: 10.1007/s11912-021-01019-9
- Youn KS, Taeun K, Kwangsoon K, Seong BJ, Soo KJ, Kwon JC, et al. Highly prevalent BRAF V600E and low-frequency TERT promoter mutations underlie papillary thyroid carcinoma in Koreans. *J Pathol Transl Med.* (2020) 54. doi: 10.4132/jptm.2020.05.12
- Chunping L, Tianwen C, Zeming L. Associations between BRAF(V600E) and prognostic factors and poor outcomes in papillary thyroid carcinoma: a meta-analysis. *World J Surg Oncol.* (2016) 14:241. doi: 10.1186/s12957-016-0979-1
- Zhang Q, Liu SZ, Zhang Q, Guan YX, Chen QJ, Zhu QY, et al. Meta-analyses of association between BRAFV600E mutation and clinicopathological features of papillary thyroid carcinoma. *Cell Physiol Biochem.* (2016) 38:763–76. doi: 10.1159/000443032
- Qunzi Z, Yong W, Qin Y, Ping W, Jianyu R. BRAF V600E as an accurate marker to complement fine needle aspiration (FNA) cytology in the guidance of thyroid surgery in the Chinese population: evidence from over 1000 consecutive FNAs with follow-up. *JPN J Clin Oncol.* (2020) 51:590–4. doi: 10.1093/jjco/hyaa209
- Yalun L, Huizhe W, Chengzhong X, Xiaoyun H, Fangxiao Z, Yangjie P, et al. Prognostic evaluation of colorectal cancer using three new comprehensive indexes related to infection, anemia and coagulation derived from peripheral blood. *J Cancer.* (2020) 11:3834–45. doi: 10.7150/jca.42409
- Han F, Liu Y, Cheng S, Sun Z, Sheng C, Sun X, et al. Diagnosis and survival values of neutrophil-lymphocyte ratio (NLR) and red blood cell distribution width (RDW) in esophageal cancer. *Clin Chim Acta.* (2019) 488:150–8. doi: 10.1016/j.cca.2018.10.042
- Lingling Z, Youjun X, Lingling Z. The potential value of red blood cell distribution width in patients with invasive hydatidiform mole. *J Clin Lab Anal.* (2019) 33:e22846. doi: 10.1002/jcla.22846
- Sulibhavi A, Asokan S, Miller MI, Moreira P, Daly BD, Fernando HC, et al. Peripheral blood lymphocytes and platelets are prognostic in

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.960740/full#supplementary-material>

surgical pT1 non-small cell lung cancer. *Ann Thorac Surg.* (2019) 109:337–42. doi: 10.1016/j.athoracsur.2019.09.006

26. Yonatan B, Caitlin O, Wendy H, Julian K, Peter C, Tristan T, et al. Platelet-lymphocyte ratio as a predictor of prognosis in head and neck cancer: a systematic review and meta-analysis. *Oncol Res Treat.* (2019) 42:665–77. doi: 10.1159/000502750

27. Yixi W, Hao Z, Yuhan Y, Tao Z, Xuelei M. Prognostic value of peripheral inflammatory markers in preoperative mucosal melanoma: a multicenter retrospective study. *Front Oncol.* (2019) 9:995. doi: 10.3389/fonc.2019.00995

28. Xinwen Z, Jialin D, Zhenyu W, Hao X, Xiaomin C, Yang L, et al. Are the derived indexes of peripheral whole blood cell counts (NLR, PLR, LMR/MLR) clinically significant prognostic biomarkers in multiple myeloma? a systematic review and meta-analysis. *Front Oncol.* (2021) 11:766672. doi: 10.3389/fonc.2021.766672

29. Masuda T, Nakaura T, Funama Y, Sugino K, Sato T, Yoshiura T, et al. (2021). Machine learning to identify lymph node

metastasis from thyroid cancer in patients undergoing contrast-enhanced CT studies. *Radiography.* 27:920–6. doi: 10.1016/j.radi.2021.03.001

30. Yeonjae K, Yangsean C, Sujin H, Kisun P, Hyunjin K, Minkook S, et al. Deep convolutional neural network for classification of thyroid nodules on ultrasound: comparison of the diagnostic performance with that of radiologists. *Eur J Radiol.* (2022) 152:110335–110335. doi: 10.1016/j.ejrad.2022.110335

31. Yeon KS, Hye LJ, Hyun YJ, Hyesun N, Eunhye H, Kyunghwa H, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head Neck.* (2019) 41:885–91. doi: 10.1002/hed.25415

32. Hongbo Z, Chang L, Jing Y, Lufan C, Qing X, Bowen S, et al. A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images. *Endokrynol Pol.* (2021) 72:217–25. doi: 10.5603/EP.a2021.0015



OPEN ACCESS

EDITED BY
Yu-Hsiu Lin,
National Chung Cheng
University, Taiwan

REVIEWED BY
Weihua Gong,
Zhejiang University, China
Mohammed Ambusaidi,
UTAS, Oman

*CORRESPONDENCE
Xipeng Zhang
zhxp1011@163.com
Liangfu Lu
liangfulv@tju.edu.cn
Mingqing Zhang
zmq@nankai.edu.cn

†These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 02 July 2022

ACCEPTED 25 August 2022

PUBLISHED 20 September 2022

CITATION

Li T, Huang H, Zhang S, Zhang Y,
Jing H, Sun T, Zhang X, Lu L and
Zhang M (2022) Predictive models
based on machine learning for bone
metastasis in patients with diagnosed
colorectal cancer.
Front. Public Health 10:984750.
doi: 10.3389/fpubh.2022.984750

COPYRIGHT

© 2022 Li, Huang, Zhang, Zhang, Jing,
Sun, Zhang, Lu and Zhang. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Predictive models based on machine learning for bone metastasis in patients with diagnosed colorectal cancer

Tianhao Li^{1†}, Honghong Huang^{2†}, Shuocun Zhang³,
Yongdan Zhang^{4,5}, Haoren Jing^{4,5}, Tianwei Sun⁶,
Xipeng Zhang^{4,5,7,8*}, Liangfu Lu^{2*} and Mingqing Zhang^{4,5,7,8*}

¹Tianjin Union Medical Center, Tianjin Medical University, Tianjin, China, ²Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China, ³Department of General Surgery, Tianjin Hongqiao Hospital, Tianjin, China, ⁴Department of Colorectal Surgery, Tianjin Union Medical Center, Tianjin, China, ⁵Tianjin Institute of Coloproctology, Tianjin, China, ⁶Department of Spinal Surgery, Tianjin Union Medical Center, Tianjin, China, ⁷The Institute of Translational Medicine, Tianjin Union Medical Center of Nankai University, Tianjin, China, ⁸Nankai University School of Medicine, Nankai University, Tianjin, China

Background: This study aimed to develop an artificial intelligence predictive model for predicting the probability of developing BM in CRC patients.

Methods: From SEER database, 50,566 CRC patients were identified between January 2015 and December 2019 without missing data. SVM and LR models were trained and tested on the dataset. Accuracy, area under the curve (AUC), and IDI were used to evaluate and compare the models.

Results: For bone metastases in the entire cohort, SVM model with poly as kernel function presents the best performance, whose accuracy is 0.908, recall is 0.838, and AUC is 0.926, outperforming LR model. The top three most important factors affecting the model's prediction of BM include extraosseous metastases (EM), CEA, and size.

Conclusion: Our study developed an SVM model with poly as kernel function for predicting BM in CRC patients. SVM model could improve personalized clinical decision-making, help rationalize the bone metastasis screening process, and reduce the burden on healthcare systems and patients.

KEYWORDS

predictive model, artificial intelligence, colorectal cancer, machine learning, bone metastasis

Introduction

Colorectal cancer (CRC) is a common malignant tumor, ranked the third most malignant tumor worldwide (1, 2). Distant metastasis is the leading cause of death in CRC patients (3), accounting for approximately 50% of patients after CRC surgery (4). The most common metastatic site of CRC is the liver or lung, while bone metastases are rare with an incidence of only 3–7% (5, 6). Patients with bone metastases have a poor prognosis, with a 5-year survival rate of < 5% and a median survival of 5–21 months (7–9).

Due to the low incidence and insignificant initial symptoms, bone metastases of CRC are difficult to diagnose at an early stage. On the one hand, compared with the low incidence of bone metastases in CRC patients, the incidence at autopsy is higher, reaching 10.7–23.7% (10). On the other hand, bone metastases are identified by further imaging or pathological examination after the occurrence of skeletal-related events (SREs) in CRC patients (11), but the median time to SREs is 2 months after the onset of bone metastases (7). Therefore, bone metastases may not be diagnosed on time in many CRC patients. Due to delayed diagnosis, patients may miss the optimal treatment time, leading to further disease progression and poor prognosis. Therefore, it is significant to predict the occurrence of bone metastasis in CRC patients.

Several predictive models for developing bone metastasis in CRC patients have been reported in previous studies (12–14). However, the performance of these models is hardly satisfactory because they are based on simple LR regression models, which may be unsuitable for predicting bone metastases. In addition, these models only identified independent risk factors associated with developing bone metastasis from CRC but did not assess the importance of each factor. Recently, artificial intelligence (AI) models based on machine learning (ML) algorithms have been increasingly used in clinical practice (15, 16). Among them, support vector machine (SVM) and other prediction models based on machine learning are better at predicting the distant metastasis of tumors, such as gastric cancer, thyroid cancer, and prostate cancer (17). SVM used in this study is a binary classification model whose basic model is a linear classifier defined by maximizing the interval on the feature space. SVM can be transformed into a non-linear classifier using the kernel method. SVM learning strategy is to maximize the interval, which can be translated into a convex quadratic programming problem and SVM learning algorithm is the optimization algorithm for solving the convex quadratic programming (18). Notably, SVM has some advantages for solving small sample high-dimensional problems. However, there are remain no studies using artificial intelligence models to predict bone metastasis in CRC patients.

Therefore, this study used population-based data to identify risk factors associated with bone metastasis in CRC patients and then build an artificial intelligence model to predict disease occurrence and help clinicians detect bone metastases in a timely manner. This can provide patients with personalized clinical strategies and promote rational allocation of healthcare resources.

Materials and methods

Study population

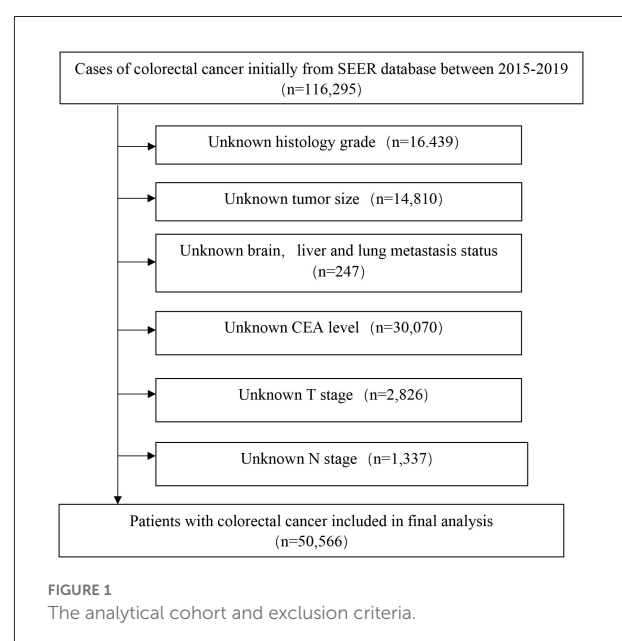
This study was based on SEER database, and patient data were collected from “SEER Research Plus Data, 17 Registries,

Nov 2021 Sub (2000–2019)” using SEER*stat 8.4.0 software and then extracted from the database between January 2015 and December 2019. SEER database covers 28% of the US population and includes information regarding cancer incidence, survival outcome, and treatment strategy from 17 population-based cancer registries. The patient selection procedure is displayed in Figure 1, and informed consent was not required as the patients were anonymized before publication. This study was approved by the Ethics Committee of Tianjin Union Medical Center.

The inclusion criteria were 1) primary CRC cases with histological confirmation, 2) histological classification: adenocarcinoma (icd-o-3:8140 to 8144, 8210 to 8213, 8220 to 8221, 8260 to 8263, 8551–8574) mucinous adenocarcinoma (MC, icd-o-3: 8480, 8481), seal ring cell carcinoma (SRCC, icd-o-3:8490), and 3) with a clear record of bone metastases. The exclusion criteria were (1) unknown information about the size, location, grade, The American Joint Committee on Cancer (AJCC) TNM stage(8th), T stage, N stage, surgery information, extraosseous metastasis, and bone metastatic status, and (2) CRC was not the first tumor.

Data selection

All CRC patients were definitively diagnosed by pathologic examination, and BM was confirmed by imaging examination and/or pathologic examination. A total of 17 population, clinicopathological, serological indicator, extraosseous metastasis, and treatment variables were included. Population variables included age and sex, clinicopathology variables included site, size, grade, histology, AJCC TNM stage, T stage, N stage, and M stage, and serological indicators included



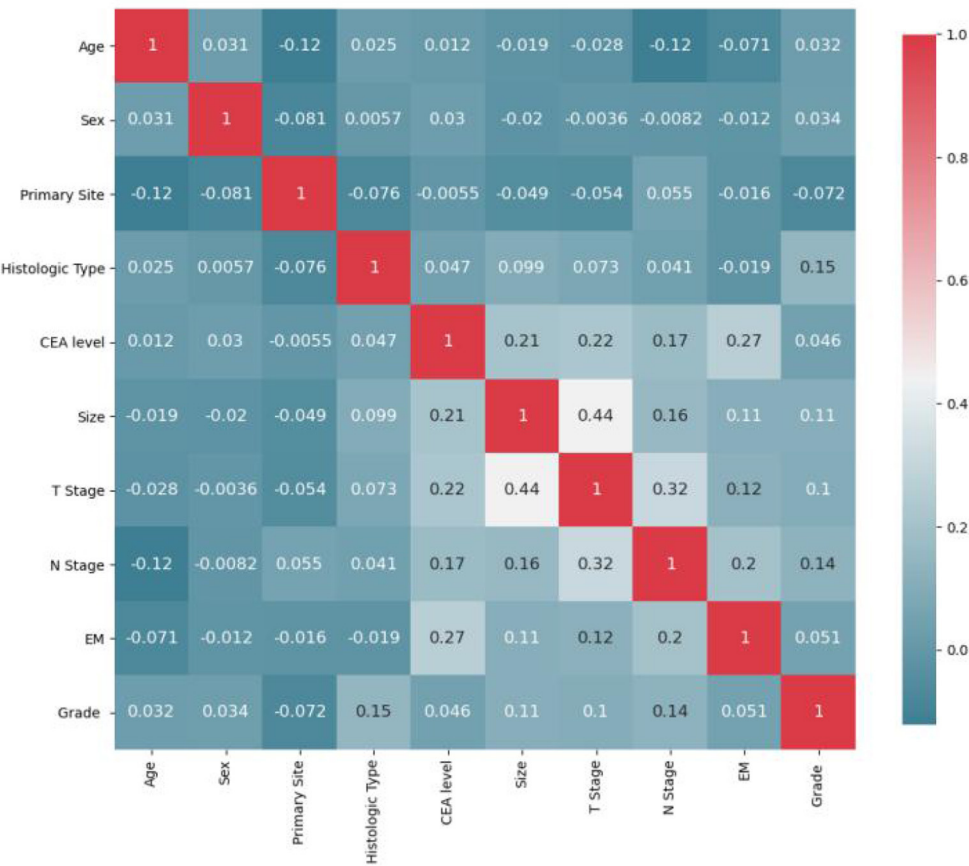


FIGURE 2
Feature correlation heatmap after initial preprocessing.

carcinoembryonic antigen (CEA) levels. Extraosseous metastasis involves bone, brain, liver, and lung metastasis. All methods were conducted according to SEER database relevant guidelines.

Model establishment

All statistics were calculated using python (version 3.8). First, the initial data were preprocessed (12, 19). (1) Continuous variables: “Age” was divided into “>60 years” and “<60 years”; “Size” was divided into “>2 cm,” “2–5 cm” and “>5 cm.” (2) Categorical variables: “Grade” was divided into “Grade I-II,” “Grade III-IV”; “T stage” was divided into “T1/2” and “T3/4”; “N stage” was divided into “N0” and “N1/2.” (3) Due to the small sample size and unbalanced distribution of the original distant metastasis variables (including lung metastasis, liver metastasis and brain metastasis) in SEER database, we added the variable of extraosseous metastasis for later model calculation. Pearson correlations between ten variables were calculated, and heatmaps were drawn. As Figure 2 displays, T stage strongly

correlates with tumor size. For the features involved in the calculation to have low correlation, it is necessary to remove T stage or size, and the principle of feature removal is to remove the feature with less weight in the model calculation. The weight of each feature calculated by the random forest is presented in Figure 3. The figure shows that T stage occupies the smallest weight, implying that it is the least important feature in the model analysis, so it is reasonable to remove T stage feature. To sum, nine features were included: age, sex, primary site, histologic type, CEA, size, N stage, extraosseous metastases (EM), and grade. Considering that the extreme imbalance of this sample (200:1) is likely to affect the model performance, it is necessary to adopt some sampling strategies. SMOTE Tomek was used in the training set as an Integrated Sampling method, and then the dataset was divided into a training set and a test set according to a ratio of 8:2.

SVM, LR, decision tree (DT), random forest (RF), and Extreme Gradient Boosting (XGB) models were used to analyze the data. To select a model with good results, we also include model comparison as part of the study. As a binary classification

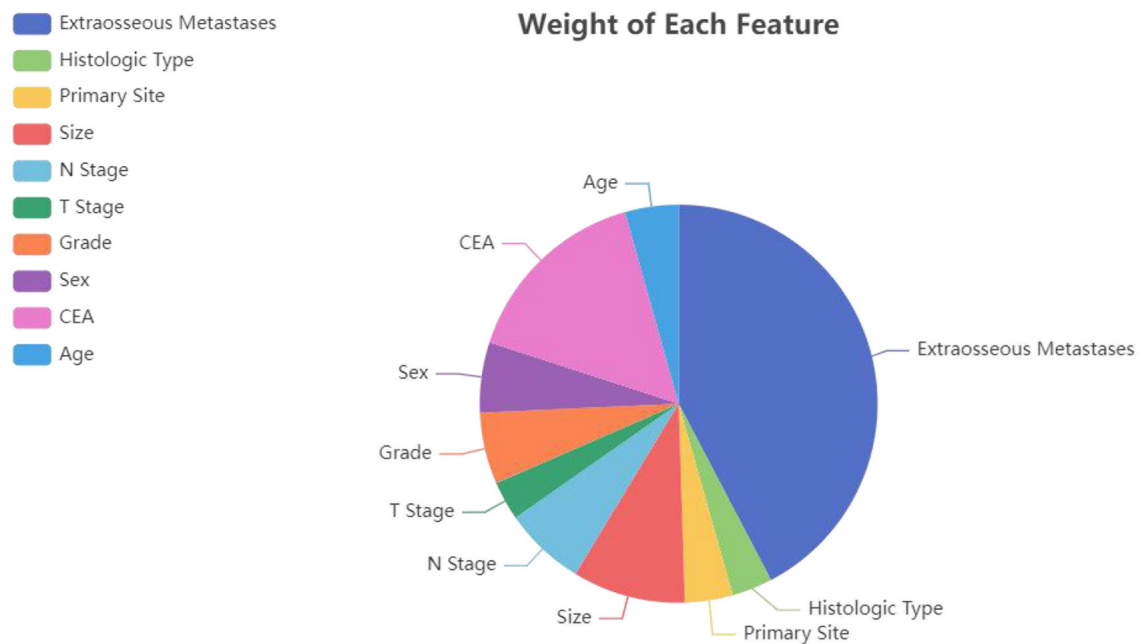


FIGURE 3
The influence weight of each factor calculated by the random forest algorithm.

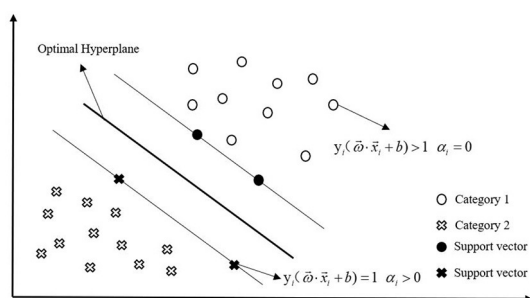


FIGURE 4
Schematic diagram of SVM.

model, SVM aims to find the optimal hyperplane to partition the samples (Figure 4), the learning strategy is to maximize the interval, and the solution of the model must be transformed into a convex optimization problem. The basic principle is to map the sample training data from the low-dimensional space to the high-dimensional space. Consequently, the sample training data is linearly separable and then the boundaries are linearly partitioned. For the sample (x_i, y_i) and the hyperplane $(\vec{\omega}, b)$, the geometric interval is defined as follows.

$$\gamma_i = y_i \left(\frac{\vec{\omega}}{\|\vec{\omega}\|} \cdot \vec{x}_i + \frac{b}{\|\vec{\omega}\|} \right)$$

Under the premise of correctly classifying the samples, when the geometric distance is the largest, the obtained separation hyperplane is optimal. The constraints are as follows:

$$\begin{aligned} \max_{\omega, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \left(\frac{\vec{\omega}}{\|\vec{\omega}\|} \cdot \vec{x}_i + \frac{b}{\|\vec{\omega}\|} \right) \geq \gamma, i = 1, 2, \dots, N \end{aligned}$$

Using the decision boundary function, it can be transformed into the following:

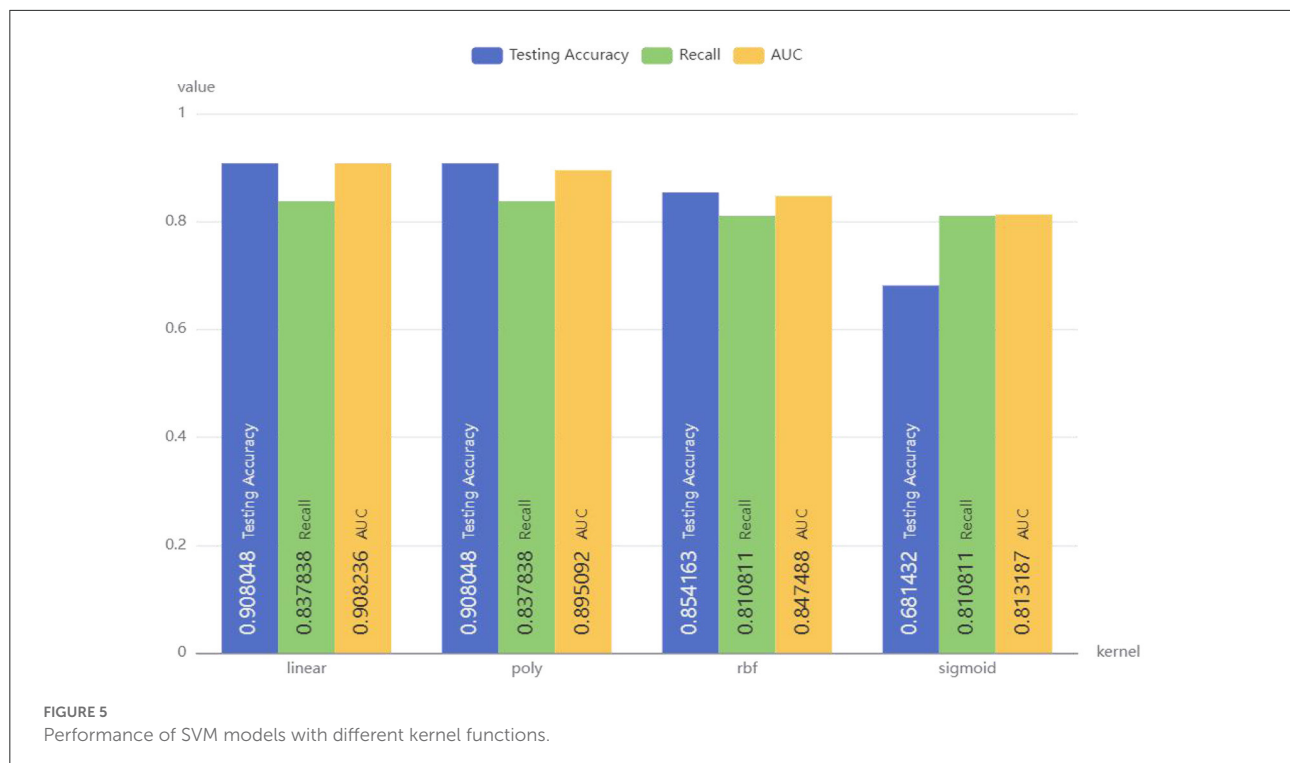
$$\begin{aligned} \min_{w, b} \quad & \frac{\|\vec{\omega}\|^2}{2} \\ \text{s.t.} \quad & y_i (\vec{\omega} \cdot \vec{x}_i + b) \geq 1, i = 1, 2, \dots, N \end{aligned}$$

After introducing the Lagrange operator α_i , it can be transformed into the following:

$$L(\vec{\omega}, b, \alpha) = \frac{1}{2} \|\vec{\omega}\|^2 - \sum_{i=1}^N \alpha_i (y_i (\vec{\omega} \cdot \vec{x}_i + b) - 1) (\alpha_i \geq 0)$$

By taking the partial derivative of a function $L(\vec{\omega}, b, \alpha)$ with respect to $\vec{\omega}$ and b , we can obtain a function about α_i . Let this function be 0 to find the optimal solution, and the optimal hyperplane formula can be obtained as follows.

$$\vec{\omega}^* \cdot \vec{x} + b^* = 0$$



Its corresponding Lagrange operator is optimal, denoted as α_i^* . At this point the classification decision function is listed below.

$$f(\vec{x}) = \text{sign}(\vec{\omega}^* \vec{X} + b^*) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \vec{x}_i \cdot \vec{x} + b^*\right)$$

Using classification decision functions, samples can be classified, which is known as SVM. However, for non-linear problems, a kernel function is required. The function of kernel function is mainly to realize the mapping from a feature space in the support vector machine to another feature space and convert the inner product of high-dimensional vectors into the inner product of low-dimensional vectors.

In the model of LR, the goal of training is to find the best weight and bias for each feature so that the error is minimized. DT is a supervised machine learning algorithm based on if-then-else rules. In the model of RF, it is trained to obtain multiple decision trees by randomly putting back the samples sampled, and finally the results of each decision tree are summed using Bagging algorithm. The above model is built by python to learn the dataset's features.

Model improvement

Linear, poly, rbf, and sigmoid were used as kernel functions for SVM model, and the results are demonstrated in Figure 5. SVM model with poly as the kernel function showed the best

performance, with an accuracy of 0.908048, recall of 0.837838, and AUC of 0.908236. Therefore, the linear kernel function was selected to build the SVM model to predict whether CRC patients have bone metastasis.

To achieve better model performance, a random search method was used for parameter optimization. After parameter optimization, SVM's Accuracy is 0.908, Recall is 0.838, and AUC is 0.926, demonstrating superior performance to previous models.

Results

Demographic and pathological characteristics

A total of 50,566 CRC patients were included in this study. At initial diagnosis, 50,325 patients (99.5%) had no bone metastases and 241 (0.5%) had bone metastases. All patients were randomized into a training set ($n = 40,452$) and a test set ($n = 10,114$) in a ratio of 8:2, and their clinical and pathological characteristics are listed in Table 1.

Model analysis and variable influence on prediction

Pearson correlations between all variables were calculated, and heatmaps were drawn, revealing no significant correlations between variables (Figure 5). For the multivariate

TABLE 1 Clinical and pathological characteristics of training and test sets.

Variables	Training set		Test set		<i>p</i> -value
	NBM(<i>n</i> = 40,248) %	BM(<i>n</i> = 204) %	NBM(<i>n</i> = 10,077) %	BM(<i>n</i> = 37) %	
Age					0.714
<60	13847(34.4)	78(38.2)	3447(34.2)	15(40.5)	
>60	26401(65.6)	126(61.8)	6630(65.8)	222(59.5)	
Sex					0.182
Male	21857(54.3)	116(56.9)	5399(53.6)	20(54.1)	
Female	18391(45.7)	88(43.1)	4678(46.4)	17(45.9)	
Primary tumor site					0.922
Colon	30131(74.9)	135(66.2)	7552(74.9)	20(54.7)	
Rectal	10117(25.1)	69(33.8)	2525(25.1)	17(45.9)	
Size					0.91
<2 cm	4936(11.5)	1(0.5)	1150(11.4)	2(5.4)	
2–5 cm	21397(53.2)	108(52.9)	5385(53.4)	16(43.2)	
>5 cm	14215(35.3)	95(46.6)	3542(35.1)	19(51.4)	
Histology					0.947
Adenocarcinoma	37405(92.9)	189(92.6)	9366(92.9)	34(91.9)	
Mucosal adenocarcinoma	2542(6.3)	8(3.9)	633(6.3)	1(2.7)	
Signet-ring cell carcinoma	301(0.7)	7(3.4)	78(0.8)	2(5.4)	
T stage					0.839
T1/2	10411(25.9)	28(13.7)	2616(26)	4(10.8)	
T3/4	29837(74.1)	176(86.3)	7461(74)	33(89.2)	
N stage					0.108
N0	22046(54.8)	60(29.4)	5607(55.6)	10(27)	
N1/2	18201(45.2)	144(70.6)	4470(44.4)	27(73)	
Grade					0.566
Grade I–II	34486(85.7)	140(68.6)	8655(85.9)	25(67.6)	
Grade III–IV	5762(14.3)	64(31.4)	1422(14.1)	12(32.4)	
CEA level					0.242
Negative	23446(58.3)	32(15.7)	5930(58.8)	5(13.5)	
Positive	16802(41.7)	172(84.3)	4147(41.2)	32(86.5)	
Extraosseous metastases					0.012
No	36227(90)	64(31.4)	9153(90.8)	6(90.6)	
Yes	4021(10)	140(68.6)	924(9.2)	31(9.4)	
Brain metastasis					0.497
No	40219(99.9)	196(96.1)	10071(99.9)	36(97.3)	
Yes	29(0.1)	8(3.9)	6(0.1)	1(2.7)	
Liver metastasis					0.019
No	36655(91.1)	79(38.7)	9249(91.8)	11(29.7)	
Yes	3593(8.9)	125(61.3)	828(8.2)	26(70.3)	
Lung metastasis					0.704
No	39263(97.6)	138(67.6)	9838(97.6)	20(54.1)	
Yes	985(2.4)	66(32.4)	239(2.4)	17(45.9)	

LR model with an enter variable selection method, six characteristics were identified as independent risk factors (Table 2), including primary tumor site ($p < 0.001$), size ($p = 0.042$), histology ($p = 0.018$), grade ($p < 0.001$),

CEA level ($p < 0.001$), EM ($p < 0.001$). According to RF results (Figure 4), the top three most important factors affecting model prediction of BM are EM, CEA, and size. Notably, the influence weight of EM accounts for 42.32%,

TABLE 2 Multivariable logistic regression model with enter variable selection.

Variables	OR (95% CI)	P
Age		
<60 years	Reference	
>60 years	0.155(0.862–1.548)	0.333
Sex		
Male	Reference	0.72
Female	0.949(0.715–1.261)	
Primary tumor site		<0.001
Colon	Reference	
Rectal	1.88(1.39–2.543)	
Size		
<2 cm	Reference	
2–5 cm	11.96(1.661–86.47)	0.014
>5 cm	10.868(1.504–78.531)	0.018
Histology		
Adenocarcinoma	Reference	
Mucosal adenocarcinoma	0.699(0.341–1.433)	0.328
Signet-ring cell carcinoma	3.035(1.316–6.998)	0.009
N stage		
N0	Reference	
N1	1.123(0.815–1.548)	0.479
Grade		<0.001
Grade I–II	Reference	
Grade III–IV	2.118(1.537–2.92)	
CEA level		<0.001
Negative	Reference	
Positive	2.879(1.908–4.344)	
Extraosseous metastases		<0.001
No	Reference	
Yes	12.207(8.805–16.923)	

which may provide some basis for diagnosing clinical auxiliary BM.

Model performance

The training set was used to train the model, and the test set was used to test the accuracy and generalization ability of the model. The performance indicators of the evaluation model were AUC, Accuracy, and Recall. After comparing the performance of different kernel functions in the SVM model (Figure 6), the linear kernel function was selected. The results were compared and analyzed using SVM, LR, DT, RF, and XGB models. The performance comparison of different models is provided in Table 3, showing that SVM model is better than the other models and may be used clinically. Previous models have mostly used LR, and to better compare the improvements

brought about by SVM model, ROC curves were plotted, and Integrated Discrimination Improvement (IDI) was calculated. As displayed in Figure 7, LR AUC is 0.92, and SVM AUC is 0.93, with an IDI of 22.66% (Figure 8), confirming that SVM model outperforms LR in this scenario.

Discussion

The incidence of bone metastasis in CRC patients is only 3–7% (6), but these patients have a poor clinical prognosis and often suffer from SREs, such as pathological fractures, severe bone pain, spinal cord compression, and hypercalcemia (20) which can seriously impair their function and quality of life, and even further affect the outcomes. Therefore, early identification and clinical intervention of bone metastasis are critical to prevent SREs and improve the clinical prognosis.

There remains a lack of accurate and effective methods to predict bone metastasis in CRC patients. Pathological diagnosis is the gold standard, but if the pathological diagnosis is unclear, the identification of bone metastasis in CRC patients relies on SREs and imaging examinations such as X-ray, CT, MRI, emission computed tomography (ECT), and positron emission tomography/computed tomography (PET/CT) (21, 22). However, these imaging modalities are expensive and associated with radiation risks, so they are not recommended as routine screening for CRC patients until SREs occur (12). For this reason, we developed an artificial intelligence model based on SVM algorithm to predict bone metastasis in CRC patients.

The advantage of this model is that it can effectively deal with the imbalance of medical data, as SVM algorithm can effectively solve the problem of inaccurate judgment results caused by small sample data in machine learning, which has stronger practicability (18, 23). In this study, this model displayed better accuracy and generalization than other models (LR, DT, and RF) and can be used to predict the occurrence of bone metastasis in CRC patients, which is helpful for doctors to make timely and effective clinical decisions.

Previous studies have reported risk factors associated with bone metastasis in CRC. Zheng et al. (21) conducted a retrospective study of 106 patients with bone metastasis of CRC, indicating that primary tumor location, lung metastasis, and serum CEA are independent risk factors. Moreover, compared with colon cancer and liver metastasis, colorectal cancer and lung metastasis were more likely to predict disease progression to bone metastasis. Wang et al. (13) determined that the degree of tumor differentiation, N stage, serum alkaline phosphatase (ALP), lactate dehydrogenase (LDH), CEA, liver and lung metastasis were risk factors for bone metastasis of CRC, and further developed a nomogram to evaluate the risk of bone metastasis in CRC patients. In addition, studies have shown that the most common risk factors for BM in CRC patients include cancer site, lymph node invasion, and lung metastasis (6).



FIGURE 6
Results of Pearson correlation analysis between all variables. The heatmap shows the correlation between the variables.

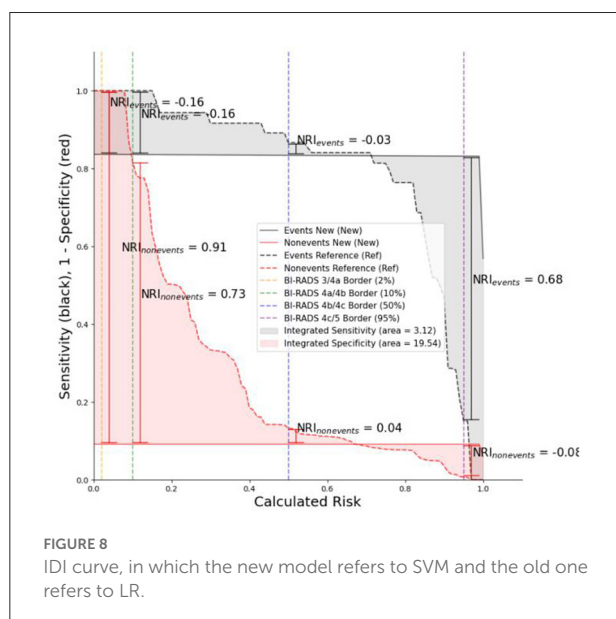
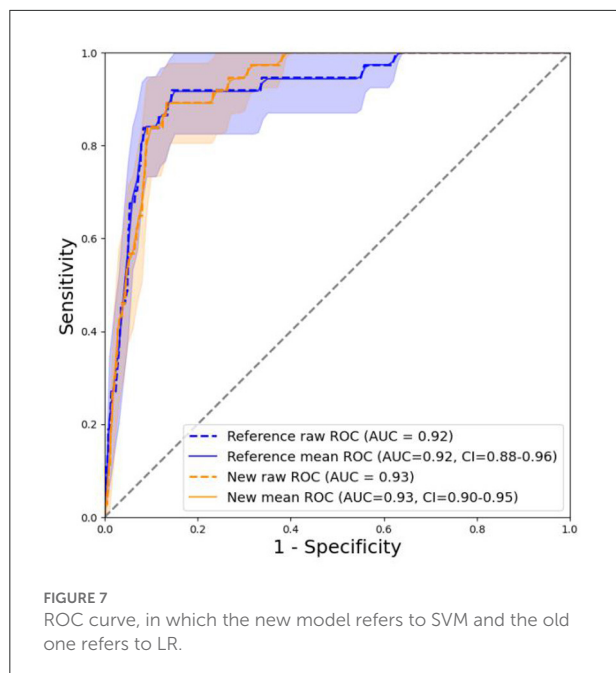
In this study, EM, followed by CEA level and tumor size, were the top three most important factors for developing bone metastasis in CRC patients. Notably, EM has an influence weight of 42.32%, an important predictor of bone metastasis in CRC patients. Studies have depicted that about 25% of CRC patients have distant metastases at the time of diagnosis (24). In our study, EM occurred in 10% of CRC patients and was an independent predictor of bone metastasis in CRC patients, consistent with previous findings (12). Due to the low incidence and insidious symptoms of bone metastasis, it is often identified after the occurrence of SREs, when the disease has already advanced, so the treatment effect and prognosis are poor (25). In addition, due to the environment of a specific organ and its effect on tumor cell adhesion, CRC tends to metastasize first to the liver or lungs before the bones (8, 9, 26). Therefore, for CRC patients with extraosseous metastasis, regular health monitoring and follow-up may be helpful for the early identification of bone metastases.

Serum CEA is considered a specific biomarker for CRC, and its concentration is significantly elevated in patients with

TABLE 3 Comparing the prediction performances of different models for BM.

Models	AUC	Accuracy	Recall
SVM	0.926	0.908	0.838
LR	0.918	0.865	0.865
DT	0.770	0.850	0.703
RF	0.770	0.850	0.676
XGB	0.873	0.882	0.838

metastatic colon cancer (27–29). In this study, CEA level was an independent predictor of bone metastasis in CRC patients, consistent with previous findings (21). Higher CEA levels may be associated with distant metastasis of CRC and nerve infiltration (30). In addition, in the current AJCC TNM staging of CRC, T staging is determined by the depth of the tumor invading the intestinal wall rather than the tumor size, but previous studies have shown that solid tumors, including those



of the gastrointestinal tract, exhibit the potential to spread not only during the vertical invasion but also during horizontal growth; (31). As the tumor size increases, the potential for metastasis is higher (32). A retrospective study by Luo et al. showed that tumor size was positively correlated with distant metastasis of rectal cancer (33). Similarly, our study depicted that size was an independent risk factor for bone metastasis of CRC, with a significantly higher incidence of bone metastases in tumors larger than 2 cm. This may provide some basis for diagnosing CRC patients with bone metastases in the clinic.

Nonetheless, this study has some limitations. First, since the model was not externally validated and was based on retrospective data, prospective cohort studies are needed to validate its accuracy and stability. Second, the model is based on an SVM algorithm, so it may be clinically difficult to interpret key features screened out by the model. In addition, since all study subjects were representative of the US population, the application of this risk model to other countries and ethnicities is limited.

Nowadays, with the rapid development of artificial intelligence technology, deep learning is widely applied in the detection and treatment of various diseases, such as cancer, diabetes, Alzheimer's disease and Parkinson's disease, and better results have been obtained (34, 35). In future research, it is planned to apply deep learning techniques in the prediction of bone metastasis occurring in colorectal cancer.

Conclusion

This study developed and validated an artificial intelligence model based on machine learning algorithms to individually predict the occurrence of bone metastasis in CRC patients by using clinical characteristics and quantifying the major factors leading to the increased risk of bone metastases. Among them, EM, followed by CEA level and size, were the top three most important factors for bone metastasis in CRC patients. Compared with the traditional LR model, the prediction performance of SVM algorithm is better (IDI: 22.66%); consequently, it could be used to timely detect bone metastases providing patients with personalized treatment and allocating health resources more effectively.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of Tianjin Union Medical Center. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

XZ, MZ, and LL conceived the idea for the study. LL and MZ were involved in planning and supervised data collection. HJ and SZ performed data collection. SZ,

YZ, and HJ conducted data analysis. TL and HH drafted the manuscript. MZ, LL, and TS contributed to writing of manuscript. All authors have discussed and decided that this manuscript is the final version and agreed to publish it.

Funding

This study was funded by Foundation of Tianjin Union Medical Center (grant number: 2016RMNK002 and 2019ZDXK01). This work was funded by Tianjin Key Medical Discipline (Specialty) Construction Project.

References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
- Wong MCS, Huang J, Lok V, Wang J, Fung F, Ding H, et al. Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location. *Clin Gastroenterol Hepatol.* (2021) 19:955–66 e61. doi: 10.1016/j.cgh.2020.02.026
- de Krijger I, Mekenkamp LJ, Punt CJ, Nagtegaal ID. MicroRNAs in colorectal cancer metastasis. *J Pathol.* (2011) 224:438–47. doi: 10.1002/path.2922
- Kim HK, Cho JH, Lee HY, Lee J, Kim J. Pulmonary metastasectomy for colorectal cancer: how many nodules, how many times? *World J Gastroenterol.* (2014) 20:6133–45. doi: 10.3748/wjg.v20.i20.6133
- Weiss L, Grundmann E, Torhorst J, Hartveit F, Moberg I, Eder M, et al. Haematogenous metastatic patterns in colonic carcinoma: an analysis of 1541 necropsies. *J Pathol.* (1986) 150:195–203. doi: 10.1002/path.1711500308
- Christensen TD, Jensen SG, Larsen FO, Nielsen DL. Systematic review: Incidence, risk factors, survival and treatment of bone metastases from colorectal cancer. *J Bone Oncol.* (2018) 13:97–105. doi: 10.1016/j.jbo.2018.09.009
- Santini D, Tampellini M, Vincenzi B, Ibrahim T, Ortega C, Virzi V, et al. Natural history of bone metastasis in colorectal cancer: final results of a large Italian bone metastases study. *Ann Oncol.* (2012) 23:2072–7. doi: 10.1093/annonc/mdr572
- Riihimäki M, Hemminki A, Sundquist J, Hemminki K. Patterns of metastasis in colon and rectal cancer. *Sci Rep.* (2016) 6:29765. doi: 10.1038/srep29765
- Sundermeyer ML, Meropol NJ, Rogatko A, Wang H, Cohen SJ. Changing patterns of bone and brain metastases in patients with colorectal cancer. *Clin Colorectal Cancer.* (2005) 5:108–13. doi: 10.3816/CCC.2005.n.022
- Katoh M, Unakami M, Hara M, Fukuchi S. Bone metastasis from colorectal cancer in autopsy cases. *J Gastroenterol.* (1995) 30:615–8. doi: 10.1007/BF02367787
- Farooki A, Leung V, Tala H, Tuttle RM. Skeletal-Related events due to bone metastases from differentiated thyroid cancer. *J Clin Endocrinol Metab.* (2012) 97:2433–9. doi: 10.1210/jc.2012-1169
- Guan X, Ma CX, Quan JC, Li S, Zhao ZX, Chen HP, et al. A clinical model to predict the risk of synchronous bone metastasis in newly diagnosed colorectal cancer: a population-based study. *BMC Cancer.* (2019) 19:704. doi: 10.1186/s12885-019-5912-x
- Wang N, Liu F, Xi W, Jiang J, Xu Y, Guan B, et al. Development and validation of risk and prognostic nomograms for bone metastases in Chinese advanced colorectal cancer patients. *Ann Transl Med.* (2021) 9:875. doi: 10.21037/atm-21-2550
- Xu W, He Y, Wang Y, Li X, Young J, Ioannidis JPA, et al. Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: an umbrella review of systematic reviews and meta-analyses of observational studies. *BMC Med.* (2020) 18:172. doi: 10.1186/s12916-020-01618-6
- Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K, Ranmuthu CKI, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *Lancet Digit Health.* (2022) 4:e466–76. doi: 10.1016/S2589-7500(22)00023-1
- Wang X, Zheng Z, Xie Z, Yu Q, Lu X, Zhao Z, et al. Development and validation of artificial intelligence models for preoperative prediction of inferior mesenteric artery lymph nodes metastasis in left colon and rectal cancer. *Eur J Surg Oncol.* (2022). doi: 10.1016/j.ejso.2022.06.009
- Liu WC, Li MX, Qian WX, Luo ZW, Liao WJ, Liu ZL, et al. Application of machine learning techniques to predict bone metastasis in patients with prostate cancer. *Cancer Manag Res.* (2021) 13:8723–36. doi: 10.2147/CMAR.S330591
- Zhou S. Sparse SVM for sufficient data reduction. *IEEE Trans Pattern Anal Mach Intell.* (2021) 44:5560–71. doi: 10.1109/TPAMI.2021.3075339
- Li X, Hu W, Sun H, Gou H. Survival outcome and prognostic factors for colorectal cancer with synchronous bone metastasis: a population-based study. *Clin Exp Metastasis.* (2021) 38:89–95. doi: 10.1007/s10585-020-10069-5
- Coleman RE. Skeletal complications of malignancy. *Cancer.* (1997) 80:1588–doi: 10.1002/(SICI)1097-0142(19971015)80:8+<1588::AID-CNCR98gt;3.0.CO;2-G
- Zheng H, Zhu ZH, Guo WJ, Zhang N, Cai YY, Ying JS, et al. Retrospective study of predictors of bone metastasis in colorectal cancer patients. *J Bone Oncol.* (2017) 9:25–8. doi: 10.1016/j.jbo.2017.10.003
- Park HS, Chun YJ, Kim HS, Kim JH, Lee CK, Beom SH, et al. Clinical features and KRAS mutation in colorectal cancer with bone metastasis. *Sci Rep.* (2020) 10:21180. doi: 10.1038/s41598-020-78253-x
- Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics.* (2000) 16:906–14. doi: 10.1093/bioinformatics/16.10.906
- Liu C, Wang T, Yang J, Zhang J, Wei S, Guo Y, et al. Distant metastasis pattern and prognostic prediction model of colorectal cancer patients based on big data mining. *Front Oncol.* (2022) 12:878805. doi: 10.3389/fonc.2022.878805
- Baek SJ, Hur H, Min BS, Baik SH, Lee KY, Kim NK. The characteristics of bone metastasis in patients with colorectal cancer: a long-term report from a single institution. *World J Surg.* (2016) 40:982–6. doi: 10.1007/s00268-015-3296-x
- Roth ES, Fetzter DT, Barron BJ, Joseph UA, Gayed IW, Wan DQ. Does colon cancer ever metastasize to bone first? a temporal analysis of colorectal cancer progression. *BMC Cancer.* (2009) 9:274. doi: 10.1186/1471-2407-9-274
- Min YL, Gong YX, Zhu PW, Lin Q, Li B, Shi WQ, et al. CEA as a risk factor in predicting ocular metastasis from colorectal cancer. *J Cancer.* (2020) 11:51–6. doi: 10.7150/jca.31196
- Pakdel A, Malekzadeh M, Naghibalhossaini F. The association between preoperative serum CEA concentrations and synchronous

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

liver metastasis in colorectal cancer patients. *Cancer Biomark.* (2016) 16:245–52. doi: 10.3233/CBM-150561

29. Kanellos I, Zacharakis E, Kanellos D, Pramateftakis MG, Tsahalis T, Altsitsiadis E, et al. Prognostic significance of CEA levels and detection of CEA mRNA in draining venous blood in patients with colorectal cancer. *J Surg Oncol.* (2006) 94:3–8. doi: 10.1002/jso.20549

30. Gao Y, Wang J, Zhou Y, Sheng S, Qian SY, Huo X. Evaluation of Serum CEA, CA19-9, CA72-4, CA125 and ferritin as diagnostic markers and factors of clinical parameters for colorectal cancer. *Sci Rep.* (2018) 8:2732. doi: 10.1038/s41598-018-21048-y

31. Li F, Kishida T, Kobayashi M. Serum iron and ferritin levels in patients with colorectal cancer in relation to the size, site, and disease stage of cancer. *J Gastroenterol.* (1999) 34:195–9. doi: 10.1007/s005350050243

32. Norton L, Massague J. Is cancer a disease of self-seeding? *Nat Med.* (2006) 12:875–8. doi: 10.1038/nm0806-875

33. Luo D, Shan Z, Liu Q, Cai S, Ma Y, Li Q, et al. The correlation between tumor size, lymph node status, distant metastases and mortality in rectal cancer patients without neoadjuvant therapy. *J Cancer.* (2021) 12:1616–22. doi: 10.7150/jca.52165

34. Schneider L, Laiouar-Pedari S, Kuntz S, Krieghoff-Henning E, Hekler A, Kather JN, et al. Integration of deep learning-based image analysis and genomic data in cancer pathology: a systematic review. *Eur J Cancer.* (2022) 160:80–91. doi: 10.1016/j.ejca.2021.10.007

35. Qu C, Zou Y, Ma Y, Chen Q, Luo J, Fan H, et al. Diagnostic performance of generative adversarial network-based deep learning methods for alzheimer's disease: a systematic review and meta-analysis. *Front Aging Neurosci.* (2022) 14:841696. doi: 10.3389/fnagi.2022.841696



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Sadiq Hussain,
Dibrugarh University, India
Mustafa Ghaderzadeh,
Shahid Beheshti University of Medical
Sciences, Iran
Monjoy Saha,
Emory University, United States

*CORRESPONDENCE

Hang Chang
hangchang@znhospital.cn
Shuang Zhou
zshbhtcm@sina.com
Zhiqiang Li
lizhiqiang@znhospital.cn

†These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 26 July 2022

ACCEPTED 15 August 2022

PUBLISHED 21 September 2022

CITATION

Liu X-P, Yang X, Xiong M, Mao X, Jin X,
Li Z, Zhou S and Chang H (2022)
Development and validation of chest
CT-based imaging biomarkers for early
stage COVID-19 screening.
Front. Public Health 10:1004117.
doi: 10.3389/fpubh.2022.1004117

COPYRIGHT

© 2022 Liu, Yang, Xiong, Mao, Jin, Li,
Zhou and Chang. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Development and validation of chest CT-based imaging biomarkers for early stage COVID-19 screening

Xiao-Ping Liu^{1,2†}, Xu Yang^{3†}, Miao Xiong^{4†}, Xuanyu Mao^{5†},
Xiaoqing Jin⁵, Zhiqiang Li^{6*}, Shuang Zhou^{7*} and Hang Chang^{5*}

¹Department of Pathology, Zhongnan Hospital of Wuhan University, Wuhan, China, ²Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, China, ³Key Laboratory of Modern Toxicology of Ministry of Education, School of Public Health, Nanjing Medical University, Nanjing, China, ⁴Department of Radiology, Wuhan Third Hospital, Tongren Hospital of Wuhan University, Wuhan, China, ⁵Department of Emergency, Zhongnan Hospital of Wuhan University, Wuhan, China, ⁶Department of Neurosurgery, Zhongnan Hospital of Wuhan University, Wuhan, China, ⁷Hubei Province Hospital of Traditional Chinese Medicine, Affiliated Hospital of Hubei University of Traditional Chinese Medicine, Hubei Institute of Traditional Chinese Medicine, Wuhan, China

Coronavirus Disease 2019 (COVID-19) is currently a global pandemic, and early screening is one of the key factors for COVID-19 control and treatment. Here, we developed and validated chest CT-based imaging biomarkers for COVID-19 patient screening from two independent hospitals with 419 patients. We identified the vasculature-like signals from CT images and found that, compared to healthy and community acquired pneumonia (CAP) patients, COVID-19 patients display a significantly higher abundance of these signals. Furthermore, unsupervised feature learning led to the discovery of clinical-relevant imaging biomarkers from the vasculature-like signals for accurate and sensitive COVID-19 screening that have been double-blindly validated in an independent hospital (sensitivity: 0.941, specificity: 0.920, AUC: 0.971, accuracy 0.931, F1 score: 0.929). Our findings could open a new avenue to assist screening of COVID-19 patients.

KEYWORDS

Coronavirus Disease 2019 (COVID-19), chest CT image, artificial intelligence, imaging biomarker, biomedical imaging application, multicentric retrospective study

Introduction

Coronavirus Disease 2019 (COVID-19) remains a global pandemic (1, 2). Early detection, early diagnosis, early isolation, and early treatment are essential for the prevention and control of the epidemic. Currently, nucleic acid detection is the most effective tool for COVID-19 diagnosis. However, early COVID-19 detection is still challenging: (1) COVID-19 belongs to a class of highly infectious diseases, with a considerable proportion of patients without obvious clinical symptoms during the onset of disease (2); (2) the critical shortages of resources, including nucleic acid detection kits, also limits the early detection of COVID-19; (3) relatively long time for nucleic

acid extraction and detection, non-standard throat swab sampling; (4) relatively high detection cost; (5) false negative rate and limited sensitivity to a certain extent due to relatively low viral load in the early stage of the disease, non-standard throat swab sampling, heterogeneities in types of samples, degradation samples, presence of PCR inhibitors, evolution of the virus, mutations in the viral genome, etc. (3–5); (6) corresponding medical waste (6–8).

Besides the coronavirus etiology, epidemiological contact history, and clinical symptoms, pulmonary imaging, especially chest computed tomography (CT) imaging, plays a unique role for COVID-19 diagnosis (9). For early-stage COVID-19 patients, unifocal ground-glass opacities (GGOs) may present as the main feature, which are most commonly located in the peripheral and inferior lobe. As the disease progresses, these unifocal GGO can develop into multiple GGOs and infiltrate the lungs, while severe consolidation of these lesion may occur in patients with severe disease (10). Lung CT images can be used not only for the diagnosis of COVID-19, but also for assessing the severity of the disease and tracking the lung changes in patients with COVID-19 who have negative nucleic acid tests (11). Several earlier studies showed high sensitivity of CT for the detection of COVID-19, indicating the potential of CT scan in the screening of COVID-19 (4, 12). Fang et al. confirmed in a cohort study of 51 patients with COVID-19 that the detection rate of chest CT for COVID-19 was 98%, while the detection rate of RT-PCR was only 71% (13). At the same time, their study showed that pulmonary vascular prominence as a key feature of COVID-19 can be found in 45–90% of cases. In another cohort study of 1014 patients, Tao et al. (11) compared the detection rate of CT and RT-PCR for COVID-19. In all 1014 patients, RT-PCR and chest CT scans were positive in 59 and 88%, respectively. Among patients with a positive RT-PCR test, chest CT showed a 97% sensitivity for the detection of COVID-19. Among patients with negative RT-PCR results, 75% had positive chest CT results, and 60–93% of cases had positive chest CT results before (or at the same time as) the initial positive RT-PCR result. Before RT-PCR results turned negative, 42% (24/57) of cases showed improvement on follow-up chest CT scans.

However, the CT image characteristics of COVID-19 patients, especially at early stage, are similar to those found in other common pneumonia patients, including those suffering from H7N9 influenza virus pneumonia, mycoplasma pneumonia, chlamydial pneumonia and bacterial pneumonia (14), which requires immediate investigation of potentially underlying characteristics other than the classical ones. Most recently, several interesting studies used artificial intelligence (AI) for the early diagnosis and GGO detection of COVID-19, including PointNet++ (15) and an AI-driven android application (16), where the former can be used for detection and quantifying GGOs in CT scans of COVID-19 patients as well as assessing the severity of the disease, and the latter

provided a novel Android application that detected COVID-19 infection from chest CT scans using a highly efficient and accurate deep learning algorithm. Furthermore, neural search architecture network (NASNet)-based algorithm has been demonstrated with great potential in a well-designed computer-aided detection (CAD) system for COVID-19 diagnosis (17). And many other deep learning related systems for COVID-19 detection and diagnosis were summarized in (18). In this study, we developed and validated chest CT-based imaging biomarkers (IBs) for early stage COVID-19 patient (i.e., mild and moderate) screening and differential diagnosis combining Artificial Intelligence (AI) and clinical findings on vascular changes in the lung regions of COVID-19 patients within a system biology approach, which could open a new avenue to assist early stage screening of COVID-19 patients. The major advantages of our imaging biomarkers reside in two folds as follows: (1) they provide robust, accurate and cost-effective COVID-19 screening, which can significantly alleviate the shortage of clinical resources, including both nucleic acid detection kits and experienced radiologists; and (2) they provide a non-invasive diagnostic tool that enables world-wide scalable practical applications. We expect that our imaging biomarkers will be of great significance to reduce the workload of clinicians and to assist in differential diagnosis of COVID-19 from other diseases.

Materials and methods

Data collection

The chest CT images in this case-control study were collected from Wuhan Third Hospital (hospital A) and Hubei Provincial Hospital of Traditional Chinese Medicine (hospital B). The inclusion criteria for COVID-19 patients were: (1) patients were diagnosed and confirmed through nucleic acid test from January 2020 to March 2020; (2) patient were with mild or moderate disease status, where the severity was classified according to the Coronavirus Disease 2019 (COVID-19) diagnosis and treatment guideline (trial version 7) (19) issued by the National Health Commission of the People's Republic of China. In addition, both patients with community acquired pneumonia (CAP) and healthy participants (with no obvious abnormalities in chest CT images) were randomly collected from aforementioned two hospitals and used as control groups in training and validation cohorts, independently. The inclusion criteria for control group were: (1) patients who were diagnosed with lung infection on imaging and clinical basis a few months before the onset of the epidemic; (2) patients without severe diseases of respiratory system, cardiovascular or cerebrovascular systems; (3) patients without mental illness or cognitive impairment. This study has been approved by the institutional review board (IRB) of

participating hospitals, and been performed according to the required guidelines.

Imaging protocol for CT chest

Chest CT exams from Hubei Provincial Hospital of Traditional Chinese Medicine were randomly performed with two different scanners: (1) GE Optima 660 CT (GE Healthcare, Milwaukee) and (2) uCT 530 (United imaging, Shanghai), with tube voltage for both scanners at 120 kVp and reconstruction thickness at 0.625 and 1.5 mm, respectively. While, CT exams from Wuhan Third Hospital were performed with GE Discovery CT750 HD (GE Healthcare, Milwaukee) with tube voltage at 120 kVp and reconstruction thickness at 0.625 mm. No intravenous contrast agents were used during scanning in both hospitals.

Vasculature-like structure enhancement

Blood vessels in lung form tubular structures and the corresponding vasculature-like signal is recognized and enhanced using iterative tangential voting (ITV) (20) within pre-segmented lung regions in 3D, where ITV enforces the continuity and strength of local linear structures and the 3D lung segmentation is achieved *via* level-set method (21). Specifically, each 3D chest CT image is resampled into isotropic image space (voxel size = $1.5 \times 1.5 \times 1.5$ mm) with SimpleITK (version 1.2.4), followed by ITV operating on the isotropic chest CT image gradient information with sigma set to be 0.5 and 1.0 on training and validation cohorts, respectively, to accommodate the technical difference across hospitals.

Imaging biomarker detection and visualization

We developed an unsupervised feature learning pipeline based on Stacked Predictive Sparse Decomposition (Stacked PSD) (22) for discovery of underlying 3D characteristics from the “vasculature-like signal” space derived by ITV. Given $\mathbf{V}=[\mathbf{v}_1, \dots, \mathbf{v}_N]$ as a set of 3D “vasculature-like signal” (\mathbf{N}), the formulation of the imaging biomarker mining pipeline is defined as follows.

$$\min_{\mathbf{B}, \mathbf{Z}, \mathbf{W}, \mathbf{G}} \|\mathbf{V} - \mathbf{B}\mathbf{Z}\|_F^2 + \|\mathbf{Z} - \mathbf{G}\sigma(\mathbf{W}\mathbf{V})\|_F^2 + \lambda_1 \|\mathbf{Z}\|_1 \\ \text{s.t. } \|b_i\|_2^2 = 1, \forall i = 1, \dots, h$$

where $\mathbf{B}=[\mathbf{b}_1, \dots, \mathbf{b}_h]$ is a set of imaging biomarkers to be mined (h). $\mathbf{Z}=[\mathbf{z}_1, \dots, \mathbf{z}_N]$ is the sparse biomarker abundance

matrix; \mathbf{W} is the auto-encoder for efficient and effective extraction of sparse biomarker abundance matrix (\mathbf{Z}) from “vasculature-like signal” (\mathbf{V}); $\mathbf{G} = \text{diag}(g_1, \dots, g_h)$ is a scaling matrix with diag being an operator aligning vector, $[g_1, \dots, g_h]$, along the diagonal; $\sigma(\cdot)$ is an element-wise sigmoid function; λ_1 is the regularization constant to ensure the sparsity of \mathbf{Z} , such that only a subset of imaging biomarkers will be utilized during the reconstruction of original “vasculature-like signal.”

The first constraint: $\|\mathbf{V} - \mathbf{B}\mathbf{Z}\|_F^2$, penalizes the reconstruction error of original “vasculature-like signal” (\mathbf{V}) with imaging biomarker (\mathbf{B}) and the corresponding sparse biomarker abundance matrix (\mathbf{Z}); the second constraint: $\|\mathbf{Z} - \mathbf{G}\sigma(\mathbf{W}\mathbf{V})\|_F^2$, penalizes the approximation error of sparse biomarker abundance matrix (\mathbf{Z}) with the auto-encoder; the third constraint: $\|\mathbf{Z}\|_1$, penalizes the sparsity of the biomarker abundance matrix, which helps ensure the utilization/activation of dominant biomarkers during the learning process. The optimization of biomarker pipeline (22) was an iterative process involving ℓ_1 - minimization (23) and stochastic gradient descent. Specifically, in this study, we used single network layer with 256 dictionary elements (i.e., patterns) at a fixed patch size of $20 \times 20 \times 20$ voxels and a fixed random sampling rate of 100 3D patches, where the patch size was optimized against reconstruction error and cross-validation performance on training set (Supplementary Figure 15). After training, Stacked PSD reconstructs vasculature-like structures, at given locations, as a combination of pre-trained patterns, with the reconstruction coefficients as the abundance of the corresponding patterns. In training cohort, 8 of 256 patterns were identified with significant correlation with COVID-19 (FDR < 0.05) through cross-validation (training sample rate: 0.8; bootstrap 100 times). The Out of Bag Error (OOB error) was used to measure the prediction error of model on the training set. At last, these 8 significant patterns (i.e., imaging biomarkers) were utilized to build the random forest classification model for COVID-19 screening. A double-blind study was designed and implemented to validate this pre-built model in an independent hospital with three steps: (1) vasculature-like structure enhancement: apply ITV on the isotopically rescaled 3D CT chest scan; (2) imaging biomarker extraction: apply Stacked PSD with pre-identified imaging biomarkers on “vasculature-like signal” space derived from step (1); and (3) double-blind COVID-19 screening: apply the pre-built random forest model on the abundance of pre-identified imaging biomarkers extracted from validation cohort. Visualization of these imaging biomarkers was created in 3D space using ITK-Snap (version 3.8.0), Python (version 3.7.0), Matplotlib (version 3.1.2), Blender (version 2.82) and Three.js (version r115 on GitHub). Snapshots of the three-dimensional visualization were used to generate two-dimensional visualization that overlays with the original CT scans.

Performance comparison between 3d imaging biomarkers and experienced chest radiologists

We invited two experienced chest radiologists to independently and blindly assess the CT images in our validation cohort, who have 8 and 10 years of clinical imaging diagnosis experience, respectively. And both radiologists have more than 2 months of intense and continuous diagnosis experience of COVID-19 in Wuhan, China. Specifically, de-identified and randomized chest CT images were given to the chest radiologists and their diagnosis were achieved according to their chest CT based clinical practice during COVID-19 diagnosis. Sensitivity and specificity were utilized for performance comparison, with nucleic acid test results as the ground-truth.

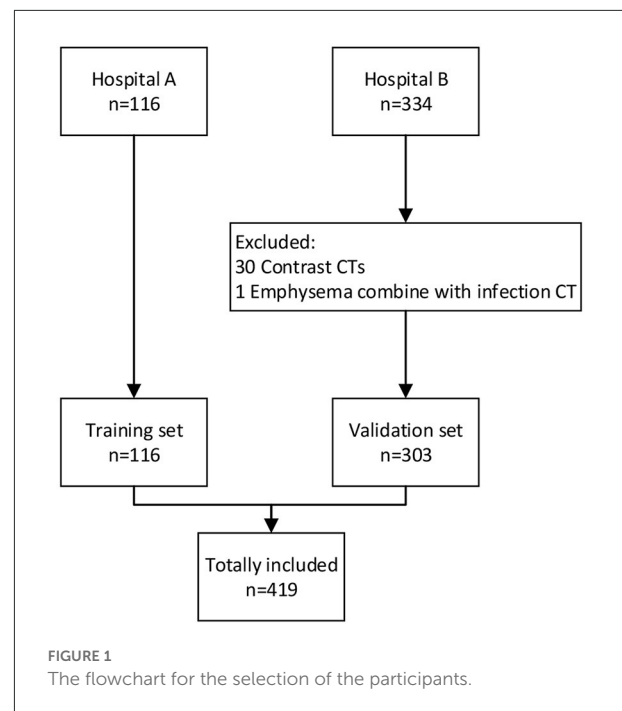
Statistical analysis

The difference in vasculature-like signals and abundance of individual imaging biomarker among different groups were assessed by Mann-Whitney non-parametric test, and association between signatures and COVID-19 were evaluated by logistic regression. The importance of individual imaging biomarker during COVID-19 screening was assessed by random forest package (version 4.6-14) in R (version 3.6.1). Principle component analysis (PCA) and heatmap were performed in R (version 3.6.1) and MATLAB (version 2012b), respectively. The screening performance was characterized with sensitivity, specificity and area under the ROC curve (AUC). Calibration of the screening model was characterized with Hosmer-Lemeshow test in R (version 3.6.1).

Results

Study population characteristics

The flowchart of participant selection in our case-control study was illustrated in Figure 1. The characteristics of cohorts are summarized in Table 1. A total of 419 participants were included in this study. The cohort ($n = 116$) from Hospital A served as training set, the cohort ($n = 303$) from the Hospital B as a double-blind validation set (Figure 2). The median ages of participants in training and validation cohorts were 42 (range: 14–76) and 51 (range: 15–89), respectively. There were 53 (45.7%) females and 63 (54.3%) males in training cohort, and 161 (53.1%) females and 142 (46.9%) males in validation cohort. Training cohort contained 47 (40.5%) COVID-19 patients, 20 (17.2%) healthy and 49 (42.2%) CAP patients, while validation cohort had 153 (50.5%) COVID-19 patients, 60 (19.8%) healthy, and 90 CAP (29.7%) patients.



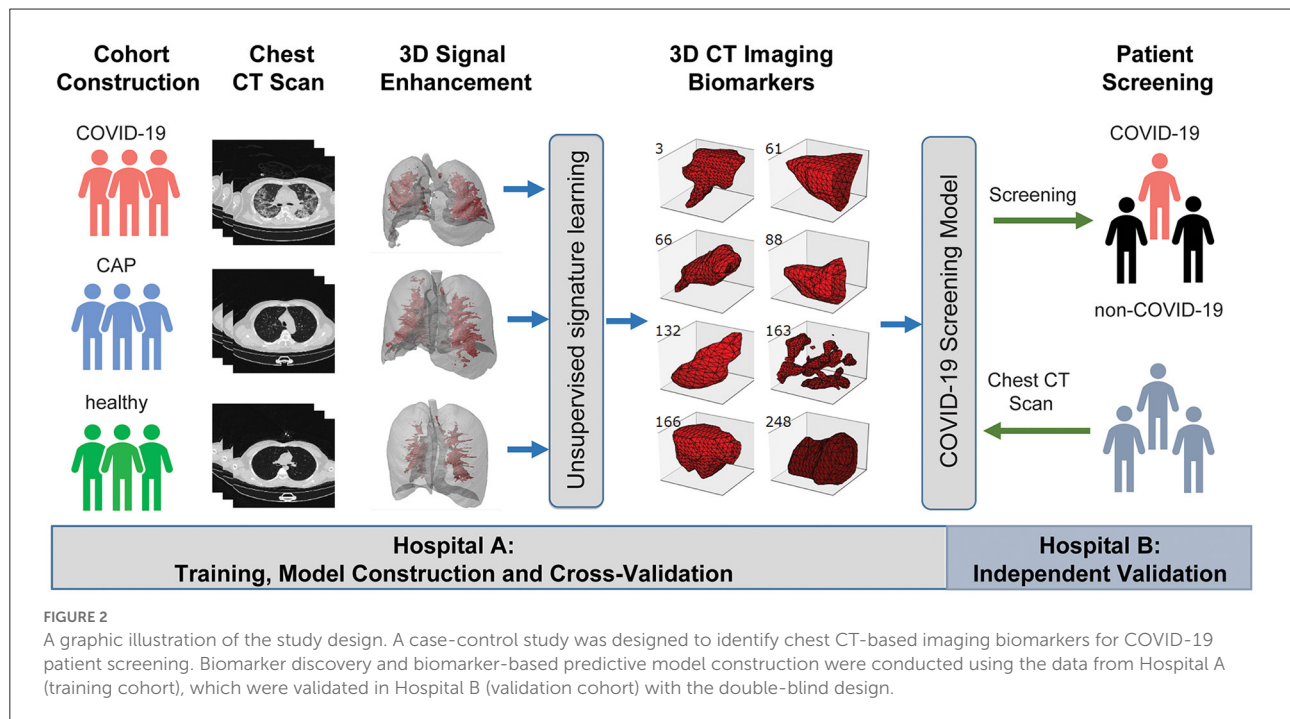
Vasculature-like structure enhancement

Inspired by recent findings on vascular changes in lung tissue of COVID-19 patients, including vascular congestion/enlargement, small vessels hyperplasia and vessel wall thickening (24–26), we hypothesize that, compared with healthy and CAP patients, COVID-19 patients have significantly more vascular changes in the lung. Therefore, we built a machine learning pipeline on enhanced vasculature-like structures formed by blood vessels to discover underlying characteristics from chest CT of early stage COVID-19 patients. Specifically, the vasculature-like structure was recognized and enhanced with ITV (20) in both training and validation cohorts as a pre-processing step. Interestingly in training cohort, the mean vasculature-like signal (i.e., the average intensity of vasculature-like structures recognized and enhanced by ITV in lung region) reveals significant differences ($p < 0.05$) between healthy, CAP and COVID-19 patients (Figure 3B). Examples of vasculature-like structure enhancement are illustrated in Figures 4A–D and Supplementary Videos 1–3 for COVID-19, CAP, and healthy cases, respectively. These findings are not only consistent with the clinical observations (24–26), but also leads to remarkable differentiation between COVID-19 and non-COVID-19 groups in training cohort [AUC = 0.721 (95% CI (0.536, 0.861)), Supplementary Figure 1, blue curve] with logistic regression. Altogether, it encourages us to identify imaging biomarkers from the “vasculature-like signal” space to assist accurate early stage COVID-19 screening.

TABLE 1 Characteristics of participants included in this study.

Variables	Training				Validation			
	COVID-19 (n = 47)	Healthy (n = 20)	CAP (n = 49)	P-value	COVID-19 (n = 153)	Healthy (n = 60)	CAP (n = 90)	P-value
Age ~ Median [Min, Max]	53.0 [31.0, 74.0]	29.0 [14.0, 50.0]	37.0 [16.0, 76.0]	<0.001	64.0 [20.0, 89.0]	41.0 [19.0, 67.0]	38.0 [15.0, 85.0]	<0.001
Gender								
Female	24 (51.1%)	7 (35.0%)	22 (44.9%)	0.477	81 (52.9%)	37 (61.7%)	43 (47.8%)	0.247
Male	23 (48.9%)	13 (65.0%)	27 (55.1%)		72 (47.1%)	23 (38.3%)	47 (52.2%)	
GGO								
No	6 (12.8%)	20 (100%)	33 (67.3%)	<0.001	12 (7.8%)	60 (100%)	55 (61.1%)	<0.001
Yes	41 (87.2%)	0 (0%)	16 (32.7%)		141 (92.2%)	0 (0%)	35 (38.9%)	
Consolidation								
No	43 (91.5%)	20 (100%)	26 (53.1%)	<0.001	123 (80.4%)	60 (100%)	46 (51.1%)	<0.001
Yes	4 (8.5%)	0 (0%)	23 (46.9%)		30 (19.6%)	0 (0%)	44 (48.9%)	

COVID-19, Coronavirus Disease 2019; CAP, community acquired pneumonia; GGO, ground-glass opacities.



Imaging biomarker detection and COVID-19 screening

Next, we applied Stacked PSD (22) on the “vasculature-like signal” space from training cohort. Two hundred fifty-six dictionary elements were learned and optimized, where 8 of them have significant positive correlations with COVID-19 (FDR < 0.05, [Supplementary Tables 1, 2](#), [Supplementary Figures 2, 3](#)). These eight COVID-19-relevant signatures (i.e., imaging biomarkers, [Figure 2](#) 3D CT

Imaging Biomarkers panel, and [Supplementary Figures 4, 5](#)) allow the construction of multispectral staining in the entire lung region ([Figure 3A](#)), which is further demonstrated in 3D ([Supplementary Videos 4–6](#)) and 2D ([Supplementary Videos 7–9](#)) animations. The 8 imaging biomarkers clearly separate COVID-19 patients from others in training cohort by PCA ([Figure 3C](#)) and clustering ([Figure 5A](#)) analysis, where each individual biomarker has significantly different abundance between COVID-19 patients and others ([Figure 5B](#)). Finally, we built a random forest classification

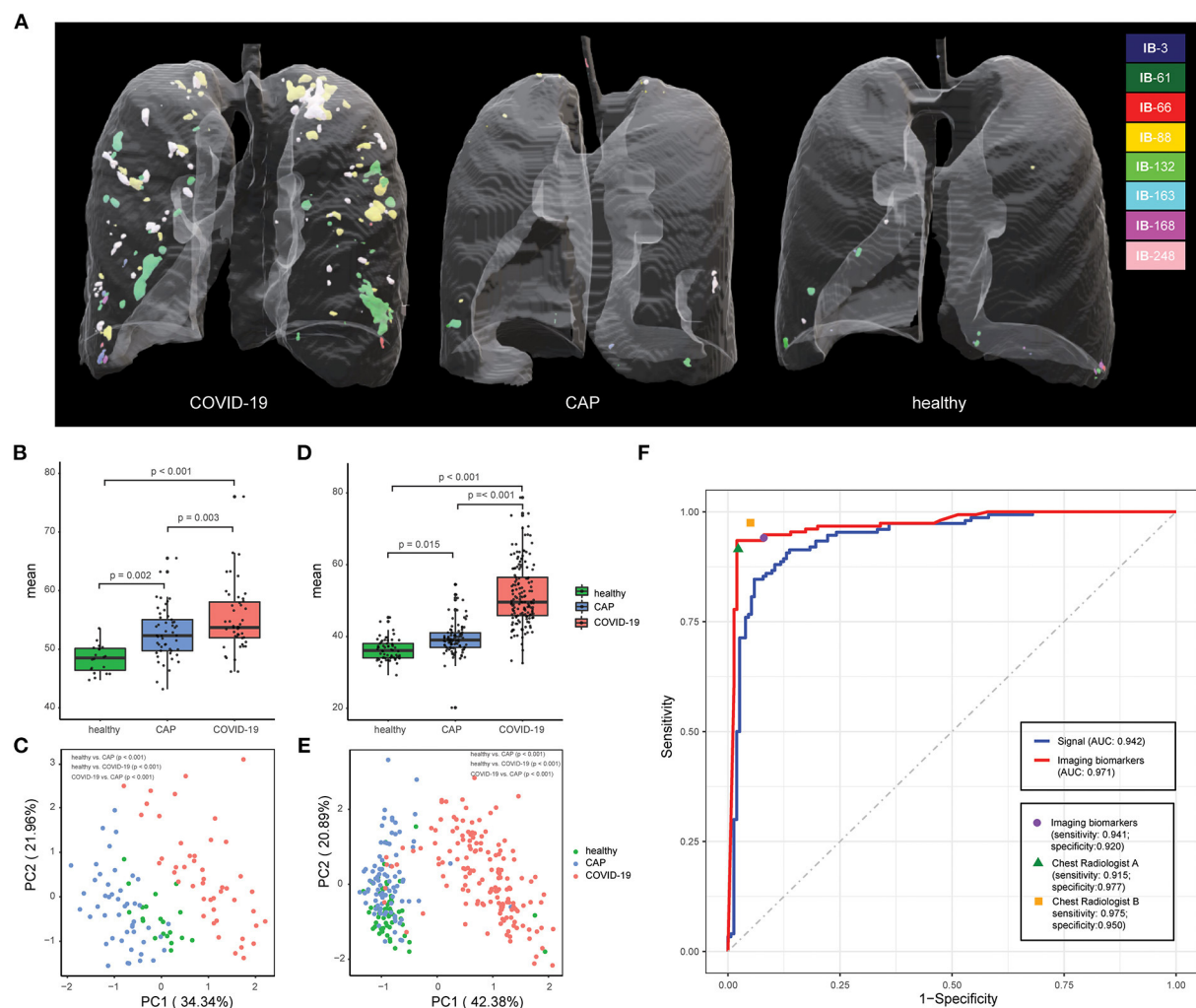


FIGURE 3

Chest CT-based imaging biomarkers accurately predicts COVID-19. **(A)** Representative examples for 3D multispectral imaging biomarker visualization in COVID-19, CAP and healthy samples. **(B)** The boxplot shows differences in the vasculature-like signals among healthy, community acquired pneumonia (CAP), and COVID-19 patients in the training cohort. The p -values were obtained by the non-parametric Mann–Whitney test. **(C)** PCA of 8 imaging biomarkers in the training cohort. Twenty healthy participants (green dots), 49 CAP patients (blue dots), and 47 COVID-19 patients (red dots). The p -values were obtained from permutational multivariate analysis of variance (PERMANOVA). **(D)** The boxplot shows differences in the vasculature-like signals among healthy, community acquired pneumonia (CAP), and COVID-19 patients in the validation cohort. The p -values were obtained by the non-parametric Mann–Whitney test. **(E)** PCA of 8 imaging biomarkers in the validation cohort. Sixty healthy participants (green dots), 90 CAP patients (blue dots), and 153 COVID-19 patients (red dots). The p -values were obtained from permutational multivariate analysis of variance (PERMANOVA). **(F)** Screening performance of signal-based model, imaging biomarker-based model, and two COVID-19 experienced radiologist on validation cohort.

model for COVID-19 screening based on these imaging biomarkers within training cohort [the OOB error = 3.26%, 95% CI (1.09–6.52%); AUC = 1.000, 95% CI (0.982, 1.000); Sensitivity = 1.000, 95% CI (0.800, 1.000); Specificity = 1.000, 95% CI (0.930, 1.000); F1 score = 0.966, 95% CI (0.923, 1.000); accuracy = 0.964, 95% CI (0.900, 1.000); precision = 1.000, 95% CI (0.875, 1.000); [Supplementary Figure 1](#), red curve]. Additionally, we show that each individual imaging biomarker contribute differently during screening, where IB-163 played the most important role ([Supplementary Figure 1b](#)),

with the best single biomarker performance [AUC = 0.893, 95% CI (0.842, 0.953), [Supplementary Figures 1c, 6](#), [Supplementary Table 3](#)].

Double-blind test of imaging biomarkers in validation cohort

The vasculature-like structure enhancement process was applied onto validation cohort, followed by biomarker

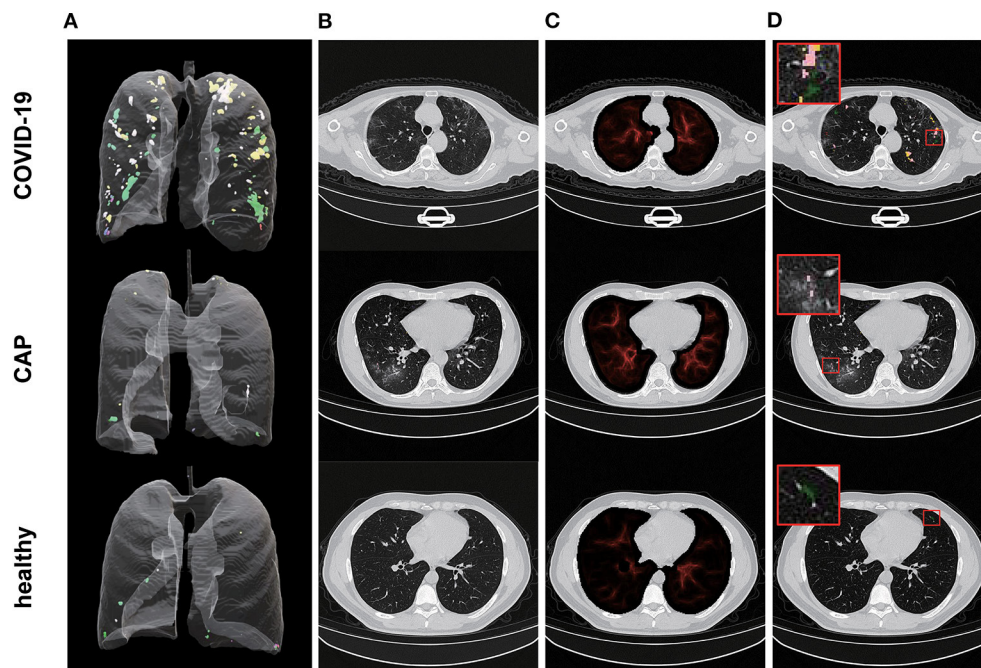


FIGURE 4

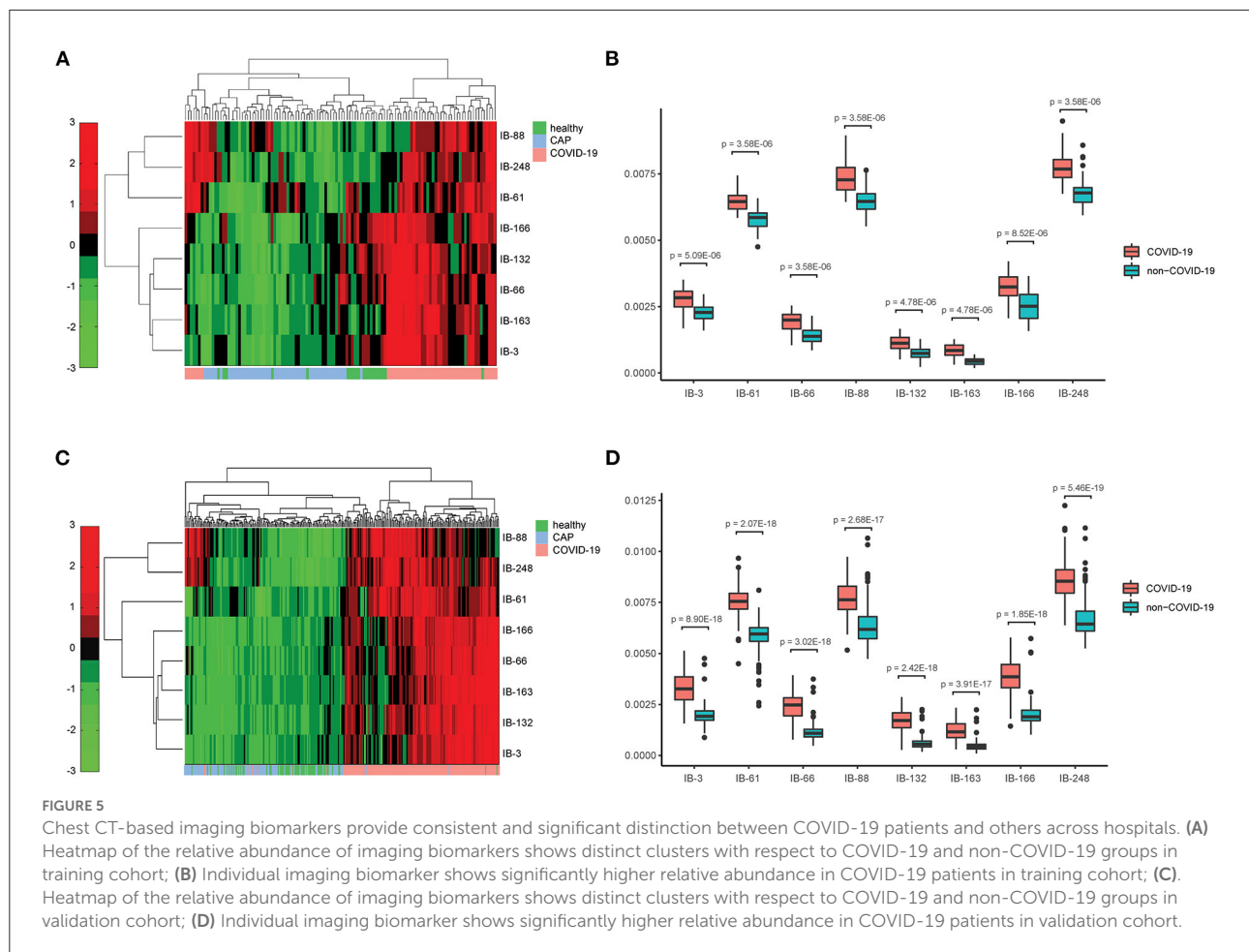
Illustration of representative CT image and the corresponding vasculature-structure enhancement and multi-spectral staining in COVID-19, CAP and healthy samples. (A) Representative examples for 3D multispectral imaging biomarker visualization (3D animations are provided by [Supplementary Videos 4–6](#)); (B) Representative 2D CT images; (C) Corresponding 2D vasculature-structure enhancement (enhancement for entire chest CTs are provided by [Supplementary Videos 1–3](#)); (D) Corresponding 2D multi-spectral staining (2D multi-spectral staining for entire chest CTs are provided by [Supplementary Videos 7–9](#)).

extraction. As seen in training cohort, we observed the distinction of mean vasculature-like signal between different groups ([Figure 3D](#)). The logistic regression model pre-built on training cohort with mean vasculature-like signal led to accurate prediction in validation cohort ($AUC = 0.942$, [Figure 3F](#), blue curve). The combination of 8 pre-identified imaging biomarkers also clearly separates the COVID-19 patients from others in validation cohort ([Figures 3E, 5C](#)), where each individual biomarker consistently revealed significantly different abundance ([Figure 5D](#)). Excitingly, we found the pre-built random forest model based on pre-obtained imaging biomarkers predict COVID-19 with excellent sensitivity (0.941), specificity (0.920), accuracy (0.931), precision (0.939), F1 score (0.929), and AUC (0.971), which is competitive with two COVID-19 experienced chest radiologists ([Figure 3F](#)): radiologist A (sensitivity = 0.915; specificity = 0.977, accuracy = 0.944, precision = 0.898, F1 score = 0.946, radiologist B (sensitivity = 0.975; specificity = 0.950, accuracy = 0.974, precision = 0.973, F1 score = 0.973). In addition, the competitiveness is further demonstrated using bootstrapping strategy (100 iterations, 80% sampling rate) on various performance metrics between imaging biomarkers and two radiologists ([Supplementary Figure 7](#)). Furthermore, the Hosmer-Lemeshow test suggested no departures from perfect fit

on both training ($p = 0.867$) and validation ($p = 1.000$) cohorts ([Supplementary Figure 8](#)).

Case study

We further examined the capability of our imaging biomarkers with misdiagnosed cases by our participating radiologists, where a COVID-19 patient (female, 65 years old, [Figure 6A](#)), and a CAP patient (male, 21 years old, [Figure 6E](#)) were included. Due to the lack of typical abnormality ([Figure 6C](#), both experts misdiagnosed the COVID-19 patient. Meanwhile, the CAP patient showed subtle misleading characteristics (i.e., GGO) in the upper lobe of both lungs ([Figure 6G](#), red arrows), and led to false positive decision by one of the experts. Obviously, in real-world clinical practice, chest CT based early screening of COVID-19 can be challenging for both clinical experts, and typical-abnormality-driven end-to-end AI systems, due to either lack of typical abnormality in COVID-19 cases or presence of misleading characteristics in non-COVID-19 cases. In contrast, our imaging biomarkers provided both perceptual ([Figure 6B](#) vs. [Figure 6F](#), [Supplementary Video 10](#) vs. [Supplementary Video 11](#); [Figure 6D](#) vs. [Figure 6H](#), [Supplementary Video 12](#) vs. [Supplementary Video 13](#)) and



quantitative (Figure 6I) distinctions (except for IB-88) for these ambiguous cases, and therefore enables accurate screening with high confidence (Figures 6A,E; over 96% confidence for both cases).

Further comprehensive justification of the robustness of imaging biomarkers

We (1) switched the role of two hospitals with Hospital B as training cohort and A as validation cohort [sensitivity: 0.957, specificity: 0.841, accuracy: 0.888, precision: 0.951, F1 score: 0.892 and AUC: 0.961 (95% CI (0.932, 0.994))]; and (2) combined two cohorts for cross-validation with random training sample rate at 80% and 100 bootstrap iterations [Supplementary Table 4, Supplementary Figure 9; sensitivity: 0.950 (95% CI (0.875, 1.000)), specificity: 0.977 (95% CI (0.909, 1.000)), accuracy: 0.953 (95% CI (0.909, 0.995)), precision: 0.973 (95% CI (0.902, 1.000)), F1 score: 0.951 (95% CI (0.903, 0.994)) and AUC: 0.980 (95% CI (0.937, 0.999))], which further demonstrated the robustness of our imaging biomarkers. Also, we performed age-group-wised (<60 and

≥60 years old) study (27) on combined cohorts to evaluate the age impact on our imaging biomarkers. As shown in Supplementary Table 5, age was comparable between the two groups both in training and validation set in ≥60 years old groups. It is clear that (Supplementary Figure 10), for all signatures (except IB-88), (1) within all age groups, the imaging biomarker has significantly higher abundance in COVID-19 patients; (2) across age groups, the imaging biomarker has significant higher abundance in category (COVID-19, <60 years old) than in category (non-COVID-19, ≥60 years old). Additionally, correlation analysis (Supplementary Table 6, Supplementary Figure 11) revealed (1) statistically non-significant ($FDR > 0.05$) “poor correlation” (28) between age and single/imaging biomarkers within COVID-19 group; and (2) three statistically significant ($FDR < 0.05$) “poor/fair correlation” (28) between age and (IB-3, IB-61, and IB-166) within Non-COVID-19 group. Also, we investigated the abundance of imaging biomarkers between age groups on both training and validation sets (Supplementary Figure 12), and confirmed that most biomarkers were significantly different between COVID-19 and non-COVID-19 age groups on both training and validation sets, except for IB-61, IB-88 and

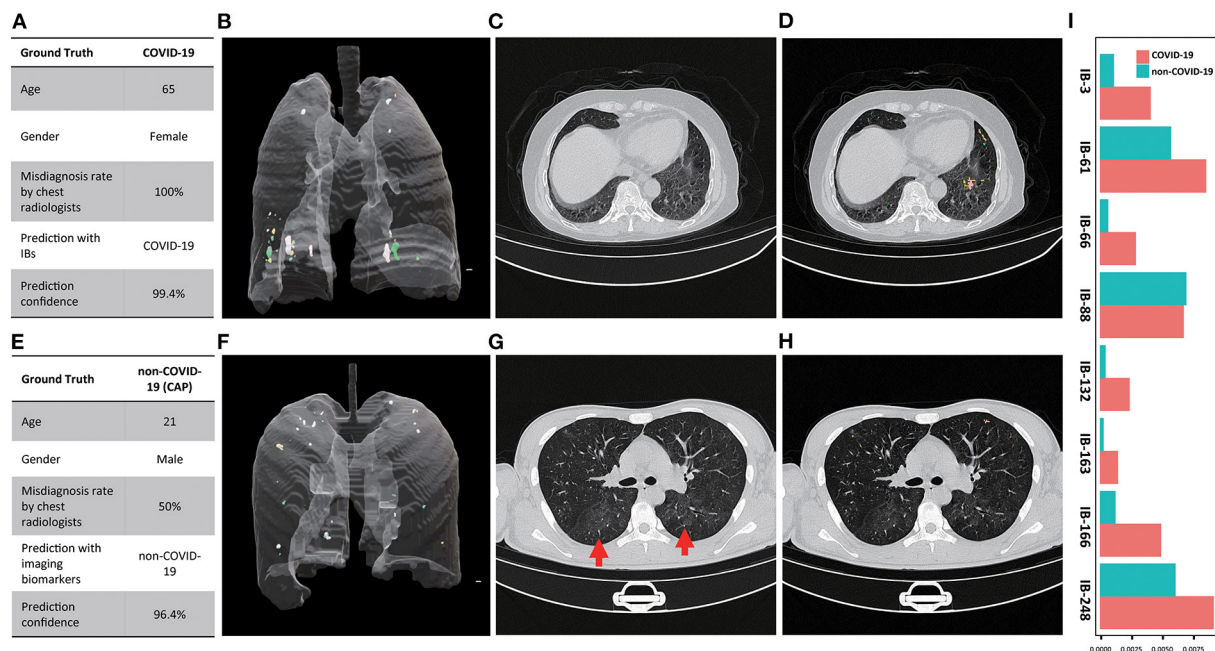


FIGURE 6

Examples of misdiagnosed cases by participating chest radiologist(s). (A) Characteristics of the COVID-19 patient and the corresponding diagnosis (chest radiologists) and screening (imaging biomarkers) results; (B) 3D multi-spectral staining of the COVID-19 patient (3D animation can be found in [Supplementary Video 10](#)); (C) Representative CT image slice of the COVID-19 shows no typical abnormality related to COVID-19, which led to the false negative decision of both chest radiologist; (D) the corresponding 2D multi-spectral staining of the selected CT image slice (2D animation of the entire CT scan can be found in [Supplementary Video 12](#)); (E) Characteristics of the CAP patient and the corresponding diagnosis and screening results; (F) 3D multi-spectral staining of the CAP patient (3D animation can be found in [Supplementary Video 11](#)); (G) Representative CT image slice of the CAP patient shows the typical white subtle image characteristics (GGO, marked by red arrows) of the COVID-19 in the upper lobe of both lungs, which led to the false positive decision by one of the chest radiologists; (H) The corresponding 2D multi-spectral staining of the selected CT image slice (2D animation of the entire CT scan can be found in [Supplementary Video 13](#)); (I) Relative abundance of imaging biomarkers differentiate the COVID-19 from CAP patient.

IB-248, potentially due to the limited sample numbers in each age group. In addition, we showed that the prediction model built upon our 8 biomarkers and patient age yielded statistically identical performance compared to the original prediction model with our 8 biomarkers only on training cohort ([Supplementary Table 7](#), [Supplementary Figure 13](#); $p > 0.05$; 100 bootstrap iterations with random training sample rate at 80%), which was further confirmed by the quantitative evaluation of these two pre-built models on validation cohort ([Supplementary Table 8](#), [Supplementary Figure 14](#)). These evidences indicate that age does not impact our imaging biomarker nor the corresponding screening model.

Potential underlying molecular and biological mechanisms

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection triggers a reverse host immunity response, followed by propagation of the virus especially to the ACE2 rich organs, among which lungs remain to be the mostly

affected organ resulting in severe respiratory disease in many individuals. Also, the unrestrained immune response triggers lung inflammation with unfavorable outcomes, where reactive oxygen species (ROS) are key signaling molecules with an important role in the progression of inflammatory disorders (29). Recent studies on SARS-CoV-2 revealed the potential molecular and biological mechanisms strikingly similar to what have been seen in pulmonary vascular disease development, including inflammation, hypoxia, oxidative stress, and DNA damage, that contribute to the promotion of endothelia dysfunction, vascular leak, and pulmonary microthrombi (30–36). Furthermore, SARS-CoV-2 leads to cytokine outburst, including IL-6, IL-1b, IL-2, IL-10, and monocyte chemoattractant protein-1 (MCP-1), which are also associated with vascular dysfunction and vascular disease such as atherosclerosis, abdominal aortic aneurysm, varicose veins and hypertension (37). Consequently, the SARS-CoV-2-related disease (COVID-19) revealed significant effects on the lungs and the pulmonary vasculature. In addition to parenchymal abnormalities, pulmonary microthrombi, ventilation-perfusion mismatch, and hypoxemia are also observed which are

due to disseminated intravascular coagulation, endothelial dysfunction, and impaired hypoxic pulmonary vasoconstriction. Importantly, our findings are consistent with these molecular- and biological-driven effects on pulmonary vasculature, which provides the underlying molecular and biological mechanism for our imaging biomarkers. Furthermore, our study indicates that these molecular and biological effects on pulmonary vasculature exist and can be quantitatively captured even at the early stage of COVID-19. With above molecular and biological potentials, we believe our imaging biomarkers could help assess the severity as well as the treatment outcome of COVID-19 patients.

Discussion

In this study, we developed and validated chest CT based 3D imaging biomarkers for early stage COVID-19 screening. We suggest, compared to healthy and CAP patients, COVID-19 patients may have significantly more vascular changes in lung tissue (24–26), which leads to the discovery of robust imaging biomarkers for early stage COVID-19 screening. Our double-blind validation across hospitals and CT scanners confirms (1) the hypothesis on the quantitative difference of vascular changes among COVID-19 and non-COVID-19 groups; (2) the robustness and effectiveness of our imaging biomarkers in real-world clinical settings with considerable technical variations; and (3) the competitiveness with COVID-19 experienced chest radiologists. Detailed case study further demonstrates the capability of our imaging biomarkers especially for ambiguous cases, which is common during early-stage COVID-19 screening. Further comprehensive evaluation suggests our imaging biomarkers are independent from hospital (batch effect free) and age (independent value). In addition, the robustness and effectiveness of our vasculature-related imaging biomarkers attribute to the effects of COVID-19 on the lungs and the pulmonary vasculature, including pulmonary microthrombi, ventilation-perfusion mismatch and hypoxemia, which are resulted from the potential mechanisms of SARS-CoV-2, including inflammation, hypoxia, oxidative stress, and DNA damage, that contribute to the promotion of endothelial dysfunction, vascular leak, and pulmonary microthrombi. For example, the structure of our best performing single imaging biomarker: IB-163 (Figure 2), potentially resembles the phenomenon related to vascular leak.

Specifically, our demonstrated screening capability was built upon biomedical evidence, robustness, interpretability, scalability, and accuracy to maximize its clinical impact. Different from many existing end-to-end solutions (38), our work was realized by seamless integration of the blood-vessel-related clinical insights within an highly compact and scalable unsupervised learning framework with feed-forward

biomarker extraction strategy involving only element-wise non-linearity and matrix multiplication (22), which helped alleviating challenges due to the (1) absence or subtle typical abnormal characteristics in chest CT especially for early stage COVID-19 patients; (2) presence of misleading characteristics in chest CT from non-COVID-19 cases; and (3) requirement of large training cohort and excessive computational resources by many end-to-end AI models. Subsequently, it enables the discovery of robust biomedical-relevant imaging biomarkers effectively from a small training cohort ($n = 116$), and thereafter scalable [~ 50 s *via* Matlab with Intel(R) Xeon(R) CPU E5-2630 v3], superior and stable screening performance.

The major limitation of our study is the exclusion of non-image information, including clinical symptoms and laboratory findings, which are valuable for COVID-19 diagnosis (39, 40). However, given (1) our current focus on imaging biomarker development and validation, and (2) the nature of biomarker detection and utilization (different from end-to-end AI systems), it is straightforward to combine non-image information with our imaging biomarkers to realized multi-modality screening capability *via* scalable techniques (e.g., random forest). Additionally, the CAP patients included in this study were from patients with pneumonia before the outbreak, which were clinically diagnosed (based on imaging findings) and treated with empirical drugs. Therefore, like many retrospective studies (38, 39, 41), the CAP patients cannot be classified according to specific pathogens, which requires a future prospective study. Chest CT scan also has certain shortcomings: first, similar to RT-PCR, chest CT scan also has certain false negative rates when the viral load is relatively low. Second, lung CT imaging is relatively expensive compared to RT-PCR testing, which may limit its use in less developed areas. Third, if the lung CT scan environment is not sufficiently disinfected, it may cause cross-infection among the tested persons. In the early stage of this epidemic, due to the high false negative rate of RT-PCR and the long return time of the test results, the chest CT scan has made up for the shortcomings of RT-PCR, and a large number of patients have been timely diagnosed, isolated and treated (42, 43). Even with the improvement of RT-PCR detection technology, chest CT still remains useful for auxiliary diagnosis and assessment of disease severity and prognosis (44–47), as well as for its potential screening capability in consideration of the possible variation of the virus during RT-PCR test. We also realized that the accessibility of CT scanner may potentially impact the utilization of our findings. However, given the (1) the demonstrated clinical implications; and (2) the prognostic potential of our imaging biomarkers combining with clinical information, we strongly believe the potential of our study in providing a valuable alternative besides nucleic acid toolkit for early-stage COVID-19 screening with world-wide impact.

To summarize, COVID-19 epidemic is a world-wide threat (48), consuming the medical resources in some countries

(49). Facing the short supply of nucleic acid detection kits in many countries, most chest CT based computational studies were built upon typical abnormality in an end-to-end fashion, which can suffer due to the lack/subtle amount of such typical characteristics in early stage COVID-19 patients, or even misleading characteristics in others. To overcome these challenges, we identified robust imaging biomarkers from vasculature-like signal in chest CT scans for accurate early stage COVID-19 screening with major advantages as follows: (1) they provide robust, accurate and cost-effective COVID-19 screening, which can significantly alleviate the shortage of clinical resources, including both nucleic acid detection kits and experienced chest radiologists; and (2) they provide a non-invasive diagnostic tool that enables world-wide scalable practical applications. Our merits originate from the system biology approach, and thus provide important clinical insights/knowledge that is beyond existing clinical practice as well as the capability/scope of many existing end-to-end AI systems. As future work, our imaging biomarkers may (1) be combined with non-image information to improve screening performance; and (2) facilitate the prediction of COVID-19 patients' prognosis and clinical outcome at early stage.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Wuhan Third Hospital; and Hubei Provincial Hospital of Traditional Chinese Medicine. Written

informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

HC, ZL, MX, and SZ designed the study. X-PL, XM, XY, XJ, and HC performed the analysis. HC, ZL, and X-PL wrote the manuscripts. All authors revised the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.1004117/full#supplementary-material>

References

1. Velavan TP, Meyer CG. The COVID-19 epidemic. *Trop Med Int Health*. (2020) 25:278–80. doi: 10.1111/tmi.13383
2. Sun P, Lu X, Xu C, Sun W, Pan B. Understanding of COVID-19 based on current evidence. *J Med Virol*. (2020) 92:548–51. doi: 10.1002/jmv.25722
3. Tahamtan A, Ardebili A. Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev Mol Diagn*. (2020) 20:453–4. doi: 10.1080/14737159.2020.1757437
4. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing. *Radiology*. (2020) 296:E41–5. doi: 10.1148/radiol.20200343
5. Pecoraro V, Negro A, Pirotti T, Trenti T. Estimate false-negative RT-PCR rates for SARS-CoV-2. A systematic review and meta-analysis. *Eur J Clin Invest*. (2022) 52:e13706. doi: 10.1111/eci.13706
6. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, Zambrano-Achig P, Del Campo R, Ciapponi A, et al. False-negative results of initial RT-PCR assays for COVID-19: a systematic review. *PLoS ONE*. (2020) 15:e0242958. doi: 10.1371/journal.pone.0242958
7. Alsharif W, Qurashi A. Effectiveness of COVID-19 diagnosis and management tools: a review. *Radiography*. (2021) 27:682–7. doi: 10.1016/j.radi.2020.09.010
8. Wang X, Tan L, Wang X, Liu W, Lu Y, Cheng L, et al. Comparison of nasopharyngeal and oropharyngeal swabs for SARS-CoV-2 detection in 353 patients received tests with both specimens simultaneously. *Int J Infect Dis*. (2020) 94:107–9. doi: 10.1016/j.ijid.2020.04.023
9. Nakajima K, Kato H, Yamashiro T, Izumi T, Takeuchi I, Nakajima H, et al. COVID-19 pneumonia: infection control protocol inside computed tomography suites. *Jpn J Radiol*. (2020) 38:391–3. doi: 10.1007/s11604-020-00948-y
10. Han R, Huang L, Jiang H, Dong J, Peng H, Zhang D. Early clinical and ct manifestations of coronavirus disease 2019 (COVID-19) pneumonia. *Am J Roentgenol*. (2020) 215:338–43. doi: 10.2214/AJR.20.22961

11. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*. (2020) 296:E32–40. doi: 10.1148/radiol.2020200642
12. Huang P, Liu T, Huang L, Liu H, Lei M, Xu W, et al. Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion. *Radiology*. (2020) 295:22–3. doi: 10.1148/radiol.2020.00330
13. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology*. (2020) 296:E115–7. doi: 10.1148/radiol.2020200432
14. Dai WC, Zhang HW, Yu J, Xu HJ, Chen H, Luo SP, et al. CT imaging and differential diagnosis of COVID-19. *Can Assoc Radiol J*. (2020) 71:195–200. doi: 10.1177/0846537120913033
15. Saha M, Amin SB, Sharma A, Kumar TKS, Kalia RK. AI-driven quantification of ground glass opacities in lungs of COVID-19 patients using 3D computed tomography imaging. *PLoS ONE*. (2022) 17:e0263916. doi: 10.1371/journal.pone.0263916
16. Verma A, Amin SB, Naeem M, Saha M. Detecting COVID-19 from chest computed tomography scans using AI-driven android application. *Comput Biol Med*. (2022) 143:105298. doi: 10.1016/j.compbiomed.2022.105298
17. Ghaderzadeh M, Asadi F, Jafari R, Bashash D, Abolghasemi H, Aria M. Deep convolutional neural network-based computer-aided detection system for COVID-19 using multiple lung scans: design and implementation study. *J Med Internet Res*. (2021) 23:e27468. doi: 10.2196/27468
18. Ghaderzadeh M, Asadi F. Deep learning in the detection and diagnosis of COVID-19 using radiology modalities: a systematic review. *J Healthc Eng*. (2021) 2021:6677314. doi: 10.1155/2021/9868517
19. Harmon SA, Sanford TH, Brown GT, Yang C, Mehralivand S, Jacob JM, et al. Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. *JCO Clin Cancer Inform*. (2020) 4:367–82. doi: 10.1200/CCI.19.00155
20. Chang H, Wen Q, Parvin B. Coupled segmentation of nuclear and membrane-bound macromolecules through voting and multiphase level set. *Pattern Recognit*. (2015) 48:882–93. doi: 10.1016/j.patcog.2014.10.005
21. Chan TF, Vese LA. Active contours without edges. *IEEE Trans Image Process*. (2001) 10:266–77. doi: 10.1109/83.902291
22. Chang H, Zhou Y, Borowsky A, Barner K, Spellman P, Parvin B. Stacked predictive sparse decomposition for classification of histology sections. *Int J Comp Vis*. (2014) 13:3–18. doi: 10.1007/s11263-014-0790-9
23. Lee H, Ekanadham C, Ng AY. Sparse deep belief net model for visual area V2. In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Vancouver, BC: Curran Associates Inc (2007). p. 873–80.
24. Li M, Lei P, Zeng B, Li Z, Yu P, Fan B, et al. Coronavirus disease (COVID-19): spectrum of CT findings and temporal progression of the disease. *Acad Radiol*. (2020) doi: 10.1016/j.acra.2020.03.003
25. Luo W, Yu H, Gou J, Li X, Sun Y, Li J, et al. Clinical pathology of critical patient with novel coronavirus pneumonia (COVID-19). *Preprints*. (2020). doi: 10.1097/TP.00000000000003412
26. Ackermann M, Verleden SE, Kuehnel M, Haverich A, Welte T, Laenger F, et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19. *N Engl J Med*. (2020) 383:120–8. doi: 10.1056/NEJMoa2015432
27. Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med*. (2020) 26:506–10. doi: 10.1038/s41591-020-0822-7
28. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med*. (2018) 18:91–3. doi: 10.1016/j.tjem.2018.08.001
29. Mittal M, Siddiqui MR, Tran K, Reddy SP, Malik AB. Reactive oxygen species in inflammation and tissue injury. *Antioxid Redox Signal*. (2014) 20:1126–67. doi: 10.1089/ars.2012.5149
30. Chen L, Li X, Chen M, Feng Y, Xiong C. The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. *Cardiovasc Res*. (2020) 116:1097–100. doi: 10.1093/cvr/cvaa078
31. Giannis D, Ziogas IA, Gianni P. Coagulation disorders in coronavirus infected patients: COVID-19, SARS-CoV-1, MERS-CoV and lessons from the past. *J Clin Virol*. (2020) 127:104362. doi: 10.1016/j.jcv.2020.104362
32. Luftig MA. Viruses and the DNA damage response: activation and antagonism. *Annu Rev Virol*. (2014) 1:605–25. doi: 10.1146/annurev-virology-031413-085548
33. Schwarz KB. Oxidative stress during viral infection: a review. *Free Radic Biol Med*. (1996) 21:641–9. doi: 10.1016/0891-5849(96)00131-1
34. Tang D, Comish P, Kang R. The hallmarks of COVID-19 disease. *PLOS Pathogens*. (2020) 16:e1008536. doi: 10.1371/journal.ppat.1008536
35. Tay MZ, Poh CM, Réna L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol*. (2020) 20:363–74. doi: 10.1038/s41577-020-0311-8
36. Varga Z, Flammer AJ, Steiger P, Haberecker M, Andermatt R, Zinkernagel AS, et al. Endothelial cell infection and endotheliitis in COVID-19. *Lancet*. (2020) 395:1417–8. doi: 10.1016/S0140-6736(20)30937-5
37. Sprague AH, Khalil RA. Inflammatory cytokines in vascular dysfunction and vascular disease. *Biochem Pharmacol*. (2009) 78:539–52. doi: 10.1016/j.bcp.2009.04.029
38. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology*. (2020) 296:E65–71. doi: 10.1148/radiol.2020200905
39. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*. (2020) 181:1423–33.e11. doi: 10.1016/j.cell.2020.04.045
40. Mei X, Lee H-C, Diao K-y, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med*. (2020) 26:1224–8. doi: 10.1038/s41591-020-0931-3
41. Han Z, Wei B, Hong Y, Li T, Cong J, Zhu X, et al. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. *IEEE Trans Med Imaging*. (2020) 39:2584–94. doi: 10.1109/TMI.2020.2996256
42. Liu WH, Wang XW, Cai ZQ, Wang X, Huang XL, Jin ZG. Chest CT as a screening tool for COVID-19 in unrelated patients and asymptomatic subjects without contact history is unjustified. *Quant Imaging Med Surg*. (2020) 10:876–7. doi: 10.21037/qims.2020.04.02
43. Huang Y, Cheng W, Zhao N, Qu H, Tian J. CT screening for early diagnosis of SARS-CoV-2 infection. *Lancet Infect Dis*. (2020) 20:1010–1. doi: 10.1016/S1473-3099(20)30241-3
44. Sardanelli F, Di Leo G. Assessing the value of diagnostic tests in the coronavirus disease 2019 pandemic. *Radiology*. (2020) 296:E193–4. doi: 10.1148/radiol.2020201845
45. American College of Radiology. *ACR Recommendations for the Use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection*. American College of Radiology (2020). Available online at: <https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection> (accessed June 1, 2021).
46. Society of Thoracic Radiology. *STR/ASER COVID-19 Position Statement (March 11, 2020)*. Society of Thoracic Radiology (2020). Available online at: https://thoracicrad.org/?page_id=2879 (accessed June 1, 2021).
47. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the fleischner society. *Radiology*. (2020) 296:172–80. doi: 10.1148/radiol.2020201365
48. Paterlini M. Lockdown in Italy: personal stories of doing science during the COVID-19 quarantine. *Nature*. (2020). doi: 10.1038/d41586-020-01001-8
49. Kamerow D. Covid-19: the crisis of personal protective equipment in the US. *BMJ*. (2020) 369:m1367. doi: 10.1136/bmj.m1367



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Shibiao Wan,
St. Jude Children's Research Hospital,
United States
Jin Cho,
University of Tennessee at
Chattanooga, United States
Igor Saveljic,
University of Kragujevac, Serbia

*CORRESPONDENCE

Michael T. Mapundu
michael.mapundu@wits.ac.za

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 10 July 2022

ACCEPTED 31 August 2022

PUBLISHED 27 September 2022

CITATION

Mapundu MT, Kabudula CW,
Musenge E, Olago V and Celik T (2022)
Performance evaluation of machine
learning and Computer Coded Verbal
Autopsy (CCVA) algorithms for cause
of death determination: A comparative
analysis of data from rural South Africa.
Front. Public Health 10:990838.
doi: 10.3389/fpubh.2022.990838

COPYRIGHT

© 2022 Mapundu, Kabudula, Musenge,
Olago and Celik. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa

Michael T. Mapundu^{1*}, Chodziwadziwa W. Kabudula^{1,2},
Eustasius Musenge¹, Victor Olago³ and Turgay Celik^{4,5}

¹Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa, ²MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of the Witwatersrand, Johannesburg, South Africa, ³National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa, ⁴Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa, ⁵School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

Computer Coded Verbal Autopsy (CCVA) algorithms are commonly used to determine the cause of death (CoD) from questionnaire responses extracted from verbal autopsies (VAs). However, they can only operate on structured data and cannot effectively harness information from unstructured VA narratives. Machine Learning (ML) algorithms have also been applied successfully in determining the CoD from VA narratives, allowing the use of auxiliary information that CCVA algorithms cannot directly utilize. However, most ML-based studies only use responses from the structured questionnaire, and the results lack generalisability and comparability across studies. We present a comparative performance evaluation of ML methods and CCVA algorithms on South African VA narratives data, using data from Agincourt Health and Demographic Surveillance Site (HDSS) with physicians' classifications as the gold standard. The data were collected from 1993 to 2015 and have 16,338 cases. The random forest and extreme gradient boosting classifiers outperformed the other classifiers on the combined dataset, attaining accuracy of 96% respectively, with significant statistical differences in algorithmic performance ($p < 0.0001$). All our models attained Area Under Receiver Operating Characteristics (AUROC) of greater than 0.884. The InterVA CCVA attained 83% Cause Specific Mortality Fraction accuracy and an Overall Chance-Corrected Concordance of 0.36. We demonstrate that ML models could accurately determine the cause of death from VA narratives. Additionally, through mortality trends and pattern analysis, we discovered that in the first decade of the civil registration system in South Africa, the average life expectancy was approximately 50 years. However, in the second decade, life expectancy significantly dropped, and the population was dying at a

much younger average age of 40 years, mostly from the leading HIV related causes. Interestingly, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to effective health intervention programmes. Through a structure and semantic analysis of narratives where experts disagree, we also demonstrate the most frequent terms of traditional healer consultations and visits. The comparative approach also makes this study a baseline that can be used for future research enforcing generalization and comparability. Future study will entail exploring deep learning models for CoD classification.

KEYWORDS

cause of death, machine learning, Verbal Autopsy, CCVA, algorithms

1. Introduction

More than 65% of the population in the world lacks high quality information on the cause of death (CoD) since every year about sixty million deaths worldwide are not assigned a medically certified cause (1). As such, most of the countries in the world fail to meet the United Nations 90% death registration coverage requirement, as deaths in many Low to Medium Income Countries (LMICs) are not captured in civil registration systems (2, 3). On the contrary, the CoD information is vital for public health monitoring, informing critical health policies and priorities. Therefore, in the absence of clinically oriented sources, CoD information should be derived from alternative sources. Verbal Autopsy (VA) is the most used tool worldwide as an alternative source of CoD information. VA is common in LMICs and is a process that is used to determine CoD where deaths occur outside health facilities and is not certified by a medical practitioner (4). These sentiments are supported by Mapoma et al. (5), who also reports on the importance of the VA process in determining CoD in countries where there are no active civil registration systems. The VA process is conducted by non-medical personnel who seek to elicit valuable information using both structured questions and an open narrative section with the next of kin of the deceased about circumstances and events that led to death (1). Two doctors are given the full set of responses, both from structured questions and open narratives for assessment and to reach a consensus on the CoD and if not a third physician is consulted, a process known as Physician Coded Verbal Autopsy (PCVA). PCVA is the most used process for determining CoD. However, it is widely criticized because of its lack of robustness, cost, time, inconsistencies, and inaccuracies as it is subjective and prone to errors among many drawbacks (6). This results in PCVAs mostly employed for the training and validation of computational approaches. The surge of technological advances has availed a plethora of automated methods for determining CoD which are faster, efficient, and cost effective (1). Most of the research that reports on ML applications in the VA domain mainly uses the responses

from the questionnaire as the classical dataset. As such, this affects comparability and generalisability. In this study, we validate the performance of various ML techniques using various VA data types for determining CoD using a comparative analysis approach. We apply enhanced data standardization and normalization strategies to achieve optimum transparency and accuracy through addressing most model limitations and applying recommendations that are reported in Reeves and Quigley (7) and Mujtaba et al. (8). We assess the robustness of several classifiers including; random forest (RF), k-nearest neighbor (KNN), decision tree (DT), support vector machine (SVM), logistic regression (LR), artificial neural network (ANN), Bayes Classifier (BC), bagging and eXtreme Gradient Boosting (XGBoost) as ensemble classifiers. We also validate our dataset using the common conventional Computer Coded Verbal Autopsy (CCVA) algorithm; InterVA.

1.1. Computer Coded Verbal Autopsy algorithms

Previous studies report on the most commonly used VA algorithms also known as CCVA algorithms. These CCVA approaches use expert-driven rules to determine CoD from VAs (9–13). The VA algorithms make use of the responses from the standardized structured World Health Organization questionnaire that denote signs or symptoms based on the deceased health history prior to death. Most of these VA algorithms take input from VA data derived from real deaths, and symptom-cause information (SCI) which is a repository of information about symptoms that are related to each probable CoD. Additionally, they make use of logic that entails a logical algorithm that combines the SCI and VA data to identify cause-specific mortality fractions (CSMF), so as to assign a specific CoD.

The InterVA uses the Bayes rule to compute the probability of cause of death, given the availability of indicators such as SCI from the VAs. This approach is reported in the study of Clark

et al. (14), Leitao et al. (15), Miasnikof et al. (13), and Murray et al. (16).

These VA approaches have been widely criticized in terms of their credibility and reliability. The study of Kalter et al. (17) reports on the evaluation of VA expert algorithms and deduces that population level accuracy is similar to that of ML approaches with CSMF in the range of 57–96%. Similar findings are also presented in the study of Quigley et al. (18) who did a study where they validated data derived algorithms against the gold standard of physician review using various disease categories based on the CSMF. Leitao et al. (15) argues that, there is little evidence to justify the CCVA as a possible replacement of the gold standard which is the PCVA. Therefore, there is a need for further investigations and research with large datasets to train and test models on CoD classification.

Little research exists in the VA domain on the application of ML to determine CoD from VA narratives. These ML algorithms make use of automated computer programs that can take input of data to learn new trends and patterns from complex data by applying optimization techniques for VA classification (19).

1.2. Machine learning in VA

Most ML model predictors commonly use only responses from the standardized questionnaire, attaining Sensitivity scores of around 60% for individual CoD classification, using various numbers of CoD categories. On the contrary, the study of Jebblee et al. (1) demonstrates that the VA narratives have valuable rich information that can be used for CoD determination. ML can avail real-time results that are similar to that of physicians/experts (20). Alternative complex ML approaches exist in the literature and can be used as substitutes for the PCVA and CCVA algorithms as approaches to determining CoD.

Moran et al. (21) applied the Bayesian hierarchical factor regression models to infer CoD using VA narratives and report an improvement in model performance on inferring CoD and CSMF. However, they used thirty-four disease categories. Idicula-Thomas et al. (22) applied six different ML algorithms (SVM, ANN, KNN, DT, C5.0, and gradient boosting). Their results report the SVM as the best classifier with an Accuracy of more than 80%. However, they used six disease categories. Similar results are reported in the study of Mujtaba et al. (23), with SVM attaining a Precision of 78.1%, Recall of 78.3%, F-score of 78.2%, and overall Accuracy of 78.25% for 16 disease categories. Their study used text classification techniques to predict CoD from forensic autopsy reports. Other studies by Danso et al. (24), Mujtaba et al. (25), and Koopman et al. (26) also found similar results and deduce that feature extraction approaches are grossly affected by variations in words and word combinations.

The study of Mwanyangala et al. (27) used the LR model to determine the completion rate of VA and factors associated with undetermined CoD. They report a completion rate of 83–89%. They ascertain that 94% of deaths submitted to physicians were assigned a specific cause, and on the other hand, 31% were labeled as undetermined. Quigley et al. (28) reports various common diseases that lead to death using CSMF and LR classifier and they achieved 80% Specificity. Boule et al. (29) applied ANN to classify CoD from VAs and achieved a Sensitivity of 45.3%. They concluded that more explorations are needed with large datasets and large training samples to improve the results of the ANN. The study of Flaxman et al. (30) used the RF classifier to assign CoD categories and affirmed that the RF algorithm performed better if not as the PCVA approach. Additionally, they point out that the RF classifier was better than PCVA on overall chance concordance and CSMF accuracy for both adults and children.

Related work that has also used VA data for cause of death determination is also reported elsewhere in Danso et al. (31). They conclude that using word occurrences produced better results as compared to word occurrence features and suggest using large datasets in order to improve model performance. Their sentiments are echoed in the study of Pestian et al. (32) and Murtaza et al. (33). Additionally, Mujtaba et al. (8, 23, 25, 34) have done vast work in the VA domain and argue that uni-grams are better feature extraction techniques, Term Frequency (TF) and TF-IDF are better feature representation schemes, and Chi-squared is a better dimensionality reduction approach. They recommend employing effective data cleaning strategies and feature engineering techniques to get improved performance.

Despite the reported results in the literature, both CCVA algorithms and ML models applied to VA data to determine CoD, suffer from challenges and limitations as they lack concrete evidence where there is a limited expert diagnosis and cannot be fully utilized to inform health priorities (2). Most of the CCVA approaches use statistical concepts and scores to determine CoD (9, 35). Moreover, these approaches are affected in terms of optimal performance because of their dependency on sample size, age group, causes of death, and characteristics of the sample (4, 13, 17, 35, 36). Other issues that affect VA data quality, emanate from having interviewers being untrained, incompetent, and unqualified to appropriately elicit relevant and appropriate symptoms on causes of death. Additionally, language barriers call for the need for the interviewer and interviewee to speak the same language so as to derive the best results. Soleman et al. (4) recommended incorporating fully trained multiple translators. The other downside is the length of the recall period which can create a bias in the collected VA data. The heterogeneity of various autopsies in terms of the non-intersecting dialects of the English language (terms being in the native language) compromises data quality as most of these approaches tend to omit such autopsies in their model prediction, yet they might entail valuable information.

All the discussed challenges and limitations affect the VA data quality that is taken as input to the CCVA and ML approaches. Therefore, we can deduce that there is a great need to address these challenges in order to remove room for any bias and misinterpretations of the models, thereby enforcing generalisability and comparability. This study demonstrates the robust assessment of ML approaches and CCVA algorithms in determining CoD, thus availing a baseline ML framework that can be used for comparability and generalization across all VA dataset types.

2. Methods

2.1. Study design

This is a retrospective cross-sectional study that uses secondary data analysis. All the cleaned VA datasets, model performance, and classification results of various tasks are pushed from a Python Jupyter Notebook environment and housed within a PostgreSQL Version 4.2 object-relational database management system.

2.2. Population

This study uses VA narrative data from the study area of the Agincourt Health and Demographic Surveillance System (HDSS). The HDSS came into existence in 1992 and is located in the rural Sub-district of Bushbuckridge under Ehlanzeni District, in Mpumalanga Province, in north-eastern South Africa. The study area covers approximately 420 km². According to the Agincourt fact sheet of 2019, the population was at 1,16,247 individuals residing in 28 villages with 22,716 households, with men being 55 961, women being 60,280, children under 5 years being 11,724, and school going children with ages from 5 – 19 being 35,928 (37).

2.3. Data source

The source of our data for this study is the Agincourt HDSS. It is a surveillance site that specifically provides evidence based health monitoring that seeks to strengthen health priorities, practice and inform policy. The VA narratives data is from 1993 to 2015. However, physician diagnosis was done from 1993 to 2010, and this target variable of the doctors' diagnosis is enough for model training and prediction.

In this study, we used three types of datasets such as the responses from the standard questionnaire, narratives, and a combination of the responses and the narratives. The whole dataset had 287 columns/features and 16,338 records/observations. For the responses only, we took all

TABLE 1 Twelve disease classes and the number of data samples before and after data balancing.

Class labels and corresponding number of samples			
Disease category	Label	Samples before data balancing	Samples after data balancing
HIV/TB	0	3,388	3,388
Other infectious	1	964	3,388
Metabolic	2	242	3,388
Cardiovascular	3	140	3,388
Indeterminate	4	1,468	3,388
Maternal and Neonatal	5	121	3,388
Abdominal	6	117	3,388
Neoplasms	7	93	3,388
External causes	8	89	3,388
Neurological	9	57	3,388
Respiratory	10	46	3,388
Other NCD	11	21	3,388

features that had responses from the standard questionnaire as our predictors and the CoD assigned by physicians using the International Classification of Diseases-10 (ICD-10) code for each record in the dataset as our target variable. Ultimately, we had 231 predictors (all symptoms, age at death, and gender) and 1 target variable, and all our features were in English. The predictions using the narratives were done using age at death, gender, the narrative feature, and 1 target variable.

For the combined VA dataset, we used 232 predictors and 1 target variable. We only added the VA narrative feature to the responses dataset in order to have our combined dataset. We further created 12 CoD categories with corresponding labels, and class distribution with the number of samples for each class before and after data balancing for our training dataset, as shown in Table 1. The CoD categories were derived based on the InterVA user guide and literature studies of Byass et al. (11), Danso et al. (24), King and Lu (38), and Jeblee et al. (1).

Figure 1 illustrates the logical steps that we follow for this study's experiments. We first do data acquisition of our VA narratives as a comma separated value text file (csv), followed by data exploration and cleaning. Additionally, we do feature engineering and data balancing and feed our data to our models for training, validation, and testing. Finally, we do CoD classification.

2.4. Data pre-processing and encoding

For the questionnaire responses dataset, we cleaned and replaced all nulls with zeros, implying that there was no symptom assigned for a missing value in a record. All symptoms

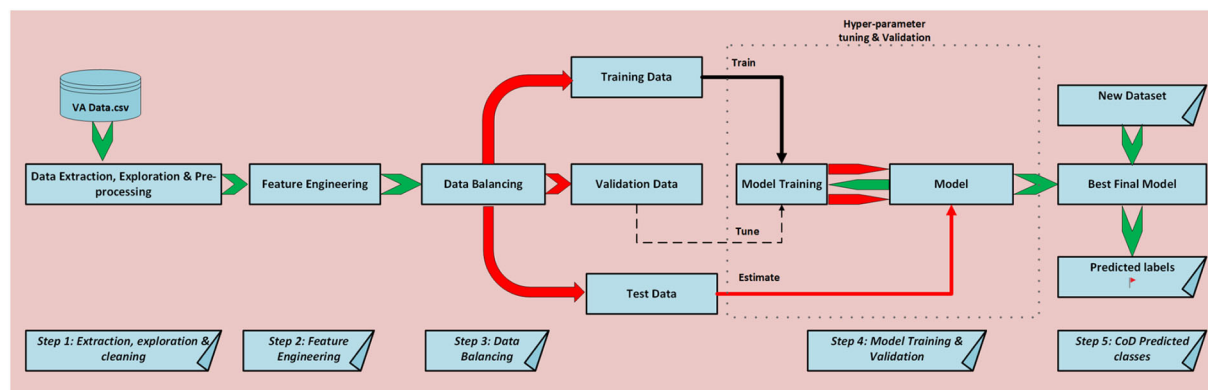


FIGURE 1
Schematic diagram of ML process followed.

that had a 'Y' were encoded as a 1 meaning that the record had a present symptom value. On the other hand, all symptoms that had an 'N', were encoded as a 0 meaning that those records had no symptoms present. In order to normalize and standardize the narrative feature used with the combined dataset, we pre-processed in order to retain with only relevant data. Data were first imported in comma separated value format, followed by pre-processing. The pre-processing stage entailed converting all text to lowercase and removing all punctuation, spaces, numbers, and special characters. Tokenisation was done by splitting a document (seen as a string) into tokens. Stopword removal was then applied to do away with insignificant words using the NLTK library of English stopwords. We applied normalization using the Python spacy package, a process known as lemmatization. Lemmatization uses a dictionary of known word forms and considers the role of a word in a sentence with the aim of extracting some normal form of a word. Finally, we applied feature engineering to determine the most representative features, as we then aimed at retaining only relevant words in the vector space by applying a weighting scheme (39). All categorical data was encoded using the one-hot encoding technique to create numeric vectors. This was followed by concatenating the narratives and the questionnaire response datasets using horizontal stacking which was pushed to our models for training, validation, and testing.

2.5. Feature engineering

We did feature engineering in order to derive new input features from existing ones. This process was done in three phases namely; feature extraction, feature selection, and feature value representation. Feature extraction was applied in order to get only relevant and useful features from textual data using n-gram models. The n-grams are a set of words that are sequential as they make use of the continuous number of items such as

characters or words from a given sequence of narratives. n-gram models can be of the form; a) $n = 1$ (unigram), b) when $n = 2$ (bigram), c) $n = 3$ (trigram), and d) hybrid-grams (mixture of unigram, bigram, and trigram) (8, 23). This was followed by the feature value representation stage employing the TF-IDF approach. In this phase, we sought to create a numeric vector of features, where each feature will have a corresponding numeric value that can be used for model learning. TF-IDF considers a feature important if it occurs frequently in the VA narratives belonging to one class and less frequently available in narratives belonging to another class. Finally, we applied feature selection in order to attain the most useful subset of features from the narratives. This was achieved using Singular Value Decomposition (SVD) as a selection approach to reduce the dimensionality of our feature space, thus removing noise in our dataset. This dimensionality reduction technique creates a matrix that only has relevant information producing an exact representation of data in a low dimensional space without any loss of data (40, 41).

2.6. Data balancing and feature scaling

We applied data balancing to the training set to address data imbalance where one or more classes are less represented than the other classes, meaning that the majority classes have more samples as compared to other minority classes. As such, this creates a bias in the minority classes as they will have fewer data points that can cause large misclassification errors. The ratio of the majority against the minority class was 1 : 160. In order to address the issue of data imbalance, we explored various techniques (under sampling, over sampling, threshold, and class weight). We attained optimal results when using the Random Over Sampling Examples (ROSE) and Synthetic Minority Oversampling Technique (SMOTE). After our experiments, we chose SMOTE as the best choice for our dataset. This

possibly suggests that our dataset was well suited for SMOTE as a data balancing technique. Moreover, our balanced datasets behaved better than imbalanced datasets. SMOTE was applied by generating artificial samples for the minority class, through interpolation between the positive instances that lie together. This approach addresses the issue of over-fitting caused by the general oversampling approach that replicates existing positive cases (8). We ended up having 3,388 samples per class. We did feature scaling using the Python Standard scaler library in order to get all our features within the same range as the target variable. After data balancing and feature scaling, we fed the data into our 12 models for training and validation.

2.7. Machine learning models for CoD prediction

We specifically applied supervised ML techniques to predict our target variable given input data. We aimed at predicting the related CoD by taking input of; questionnaire responses only, narratives only, and combined questionnaire responses and narratives. The input was then fed into nine classifiers (SVM, DT, XGBoost, KNN, RF, Bagging, LR, BC, and ANN). These ML approaches are reported elsewhere (1, 22, 23, 27, 33, 34, 39–47). Using the questionnaire responses only, we created a feature space made up of binary responses as predictors and our target variable was a categorical ICD-10 code for CoD. Similarly, we did the same for the narratives only dataset. For the combined dataset only, we added the narrative column to the list of our predictors.

2.8. Model training, validation, and testing

In this study, we perform multi-class classification, where we generated individual prediction models for each of the 12 disease categories. Data were split into 70% training, 20% validation, and 10% testing on unseen or new data, for all our nine models. We evaluated model performance by assessing the robustness of the nine classifiers by applying 10-fold cross-validation supplemented by the GridSearch algorithm. The k-fold cross-validation (k=10 in our study) is advantageous in that, it uses all observations for both training and validation, with each observation used for validation exactly once. On the contrary, this approach has the disadvantage of having to define the number of folds manually. In order to address the limitations of the k-fold cross-validation technique, we also used the automated GridSearch approach that eliminates the random setting of parameters and chooses optimum parameters automatically for a specific model.

In order to attain a better estimate of the generalization performance, we used 10-fold cross-validation to evaluate the performance of each parameter combination, instead of using

TABLE 2 Model optimal hyperparameters.

Selected hyperparameters	
Model name	Hyperparameters
XGBoost	L1, max_depth=10, objective=multi:softmax, learning_rate =0.1, alpha=0
RF	gini, max_depth =10, n_estimators=100, min_samples_leaf=1
ANN	relu, alpha=0.0001, solver=adam
KNN	minkowski, n_neighbors=5, p=2
SVM	gamma=scale, kernel=rbf, C=1.0
Bagging	KNN, max_samples, max_features
DT	gini, min_samples_split=2, min_samples_leaf=1,
LR	L_2 , $C = 1.0$
BC	alpha=1.0, fit_prior=True, class_prior=None

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.

a single split into a training and validation set. First, we specified the parameters for searching stored in a dictionary. GridSearch cross validation function then performed all the necessary model fits. All dictionary keys were the names of the parameters that we wanted to tune, and the values were the parameter settings that we wanted to test out. Applying cross-validation, we managed to choose the optimal parameters that gave us the best model performance based on the accuracy of the test set or unseen data.

We used optimisation parameters such as; cost complexity pruning and tuning parameter alpha through k-fold validation (tree based models). Moreover, we also used the Mean Squared Error (MSE) and Cross Entropy Error (CEE), Minkowski and Gini as cost functions to compute the minimal cost error between our predictor and the response using the k-fold cross-validation approach to optimize model performance. These cost functions are described in Zaki and Meira (40). Additionally, we also employ L_1 and L_2 regularization approaches to further optimize some of our models. L_1 regularization involves eliminating features that are not useful for model prediction by setting some weights close to zero. On the contrary, L_2 regularization tends to penalize large weights more and small weights less (41). Table 2 depicts some of the model hyperparameters used in our models.

3. CCVA algorithms

We followed the same preprocessing steps of our dataset and fed it into our commonly applied CCVA algorithm InterVA. The data preprocessing steps entailed de-duplication based on the identifier field (ID), dropping observations with peculiar IDs, filtering out observations with recorded age at death above 110 years, and any observation with the year of death before 1992

and after 2016. All records with unspecified sex were dropped from the raw dataset. All modeling for the InterVA was done in R. Libraries such as knitr were used for dynamic report generation, lubridate was used for date and time functions, tidyr for organizing and tidying of data, tidyverse for loading core packages, ggplot for plotting graphs, readxl for reading our excel raw data, and InterVA for our CCVA algorithm. In order to determine the most probable CoD, we used the InterVA libraries for analysis in our R-statistical analysis software guided by the study of Li et al. (48) and McCormic et al. (12). Since InterVA and InSilico are correlated, we decided to only validate the InterVA algorithm for comparability with ML approaches.

4. Identification of contradicting cases and best model predictors

In order to identify contradicting cases, where physicians were not agreeing on the diagnosis, we extracted a separate dataset. We used simple text mining techniques known as n-gram models for identifying the contradicting cases and best features for our models (refer to Section 2.5).

4.1. ML techniques model evaluation

Performance evaluation of classifiers is evaluated using various metrics and we report the metrics based on studies by Mujtaba et al. (23, 34). We validated our results using one vs. all with Accuracy, Precision, Recall, F-score, and AUROC as our metrics for evaluation.

Accuracy denotes all classes with classified results that have been predicted correctly in fraction terms. Precision also known as the Positive Predictive Value (PPV) defines the proportion of VA narratives correctly predicted as positive to the total of positively predicted VA narratives. Recall also known as Sensitivity or True Positive Rate (TPR) defines the proportion of VA narratives correctly predicted as positive to all VA narratives in the actual positive category. F-measure computes the average or harmonic mean of Precision and Recall.

True Positives (TP) and True Negatives (TN) represent the number of outcomes in which our prediction model correctly classifies positive and negative cases, respectively. In our case, TP denotes predicted positive VA narratives with a particular disease category from the 12 classes and are actually positive and TN denotes predicted negative VA narratives with a particular disease category from the 12 classes and are actually negative. Conversely, False Positives (FP), and False Negatives (FN) denote the number of outcomes where our models incorrectly predict the positive and negative classes, respectively. As in our case, the FP implies predicted positive VA narratives with a particular disease category from the 12 classes but are actually negative and FN depicts the predicted negative VA narratives

with a particular disease category from the 12 classes but are actually positive.

The AUROC visualizes the TPR against the false positive rate (FPR). The area under the ROC curve applies the principle of plotting a curve specific to a machine learning algorithm where the classifier is evaluated relative to a weighting on the area under the curve. Good performance of the algorithm is given a weight of close to 1, thus graph is AUROC closer to the upper left corner and the poor performance of an algorithm is given a weight of 0.5 and below. Specificity computes the ratio of negative VA narratives that are correctly predicted as negative.

4.2. CCVA techniques model evaluation

We explicitly validated the InterVA algorithm using CSMF accuracy and Overall Chance-Corrected Concordance (CCC). CCC computes the accuracy of individual cause assignment and ranges from 0 to 1 and the lower the CCC, the larger the error type on the accuracy of the underlying cause (14, 49). On the other hand, CSMF accuracy defines accuracy as having a value between 0 and 1. This metric assumes the worst possible case for predicting CSMF and assigns a weight on the least possible CSMF value that matches the total absolute error (50).

5. Statistical analysis

We applied statistical tests for comparing the performance of our nine algorithms. We computed the variance of our models using descriptive statistics such as mean and standard deviation based on the results of our AUROC. Moreover, we computed some tests using 10-fold cross-validation using the mean and standard deviation. Furthermore, we conducted some non-parametric tests since our data distribution was non-normal using the Kruskal-Wallis test, to test if the model's mean is different or the same. For the Kruskal-Wallis test, we considered $p < 0.005$ statistically significant. We applied the pairwise model comparisons using McNemar statistical tests, in order to be able to state objectively whether one model performs better than the other (51). Since we did eight different tests, we used the Bonferroni corrected p-value of 0.0065, derived from $0.05/8$, where the denominator is the total number of tests. We used Python version 5.2.2 and STATA version 17 SE edition for all these statistical tests.

6. Results

In this section, we present the results attained from CCVA algorithms and various classification techniques employed to determine CoD from various VA datasets (using only narratives as predictors, using questionnaire responses only, and results of the combined features).

TABLE 3 Comparison of nine ML models using narratives only.

Model evaluation						
Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROCMA	AUROCMAA
XGBoost	96	96	96	96	0.927	0.906
RF	96	96	96	96	0.998	0.996
ANN	94	94	94	94	0.982	0.964
KNN	93	93	93	92	0.989	0.987
SVM	92	92	92	92	0.917	0.917
Bagging	91	91	91	91	0.997	0.995
DT	85	84	85	84	0.910	0.910
LR	82	82	82	82	0.977	0.959
BC	71	75	71	72	0.921	0.920

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.

6.1. Performance evaluation of ML classifiers

We validated our results using one vs. all with Precision, Recall, Accuracy, F1-score, and AUROC. We report on Precision, Recall, Accuracy, F-score measure, and AUROC for our nine classifiers in the CoD categorization of the 12 disease classes for narratives only in [Table 3](#), questionnaire responses only in [Table 4](#), combined questionnaire responses and narratives in [Table 5](#).

6.1.1. Results from only the VA narrative predictors

The XGBoost and RF classifier outperformed all the other classifiers with a Precision of 96%, Recall of 96%, F1-score of 96%, and Accuracy of 96%, respectively. The least performing classifier was the statistical BC classifier with an Accuracy of 71%. Overall, our nine models had an AUROCMA (Area Under Receiver Operating Characteristics Micro Average) and AUROCMAA (Area Under Receiver Operating Characteristics Macro Average) between 0.910 – 0.998 and 0.910 – 0.996, respectively. [Table 3](#) shows the detailed performance evaluation results of our nine models using VA narratives only.

6.1.2. Results from using questionnaire responses only as predictors

The ANN and XGBoost outperformed all the other classifiers when using questionnaire responses from the standardized questionnaire attaining a Precision, Recall, F1-score, and Accuracy of 100%, respectively. It was followed by Bagging our ensemble classifier and KNN both recorded a Precision, Recall, F1-score, and Accuracy of 98%, respectively. Our statistical classifiers LR and BC were on the lower ranking

of our evaluation recording an Accuracy in the range of 74–83%, respectively. All of our models attained the highest AUROCMA within the range of 0.869 and 1, respectively. Our nine models XGBoost, RF, ANN, Bagging, SVM, LR, DT, and KNN record high scores and the BC a bit lower AUROCMA of 0.869. Additionally, the same nine models attained the highest AUROCMAA within the range of 0.976 and 1, respectively. On the other hand, the BC achieved an AUROCMAA score of 0.884. [Table 4](#) shows the detailed performance evaluation results of our nine models using questionnaire responses only.

6.1.3. Results from using combined narratives and questionnaire responses

The XGBoost, ANN, and the RF classifier outperformed all the other classifiers with a Precision of 96%, Recall of 96%, F1-score of 96%, and Accuracy of 96%, respectively. On the contrary, BC and SVM were the least performing classifiers with Accuracy in the range of 68 – 72%. All of our models attained the highest AUROCMA within the range of 0.910 and 0.998, respectively. The RF, XGBoost, ANN, Bagging, and KNN recorded high scores and the rest a bit lower scores. Additionally, our models attained the highest AUROCMAA within the range of 0.907 and 0.996, respectively. Similarly, the RF, XGBoost, ANN, Bagging, and KNN recorded high scores and the rest a bit lower comparable scores. However, the BC attained the lowest AUROCMA of 0.869 and 0.884, respectively. [Table 5](#) shows the detailed performance evaluation results of our nine models using combined questionnaire responses and narratives. Additionally, [Figure 2](#) shows the model validation done using AUROC.

We report on the performance validation of our nine algorithms using descriptive statistics such as the mean and SD based on the Micro and Macro averages of our AUROC

TABLE 4 Comparison of nine ML models using questionnaire responses only.

Model evaluation						
Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROCMA	AUROCMAA
XGBoost	100	100	100	100	1	1
ANN	99	99	99	99	1	1
Bagging	98	98	98	98	0.998	0.998
KNN	98	98	98	98	0.997	0.997
RF	97	97	97	97	0.999	0.998
DT	97	97	97	97	0.976	0.976
SVM	94	94	94	94	0.990	0.988
LR	83	83	83	83	0.990	0.980
BC	74	77	74	75	0.869	0.884

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.

TABLE 5 Comparison of nine ML models using combined narratives and questionnaire responses.

Model evaluation						
Model name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROCMA	AUROCMAA
XGBoost	96	96	96	96	0.994	0.990
RF	96	96	96	96	0.998	0.996
ANN	96	95	96	95	0.995	0.991
Bagging	93	92	93	92	0.994	0.994
KNN	91	91	91	90	0.982	0.981
DT	87	87	87	87	0.928	0.928
LR	76	76	76	76	0.985	0.973
BC	72	73	72	73	0.910	0.907
SVM	68	68	68	66	0.969	0.958

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier; AUROCMA, Area Under Receiver Operating Characteristics Micro Average; AUROCMAA, Area Under Receiver Operating Characteristics Macro Average.

reported in Table 5. We report on 0.010282 and 0.010105 variance for AUROCMAA and AUROCMA for our dataset, respectively. Table 6 shows the mean and standard deviation scores for each model throughout the 10-fold cross-validation training of the algorithms. The rank sums for each model column depict the Kruskal-Wallis test conducted. The test revealed that the mean observation was not the same ($Chi = 85.383, p = 0.0001$) across the nine models. This, therefore, implies that there was a statistically significant difference in mean observation between the nine models. We also report a p -value greater than the significance level of 0.05, hence, we fail to reject the null hypothesis and conclude that the nine model observations are not normally distributed. All model variances are very low or insignificant, implying that our dataset had a low degree of spread. Therefore, we can confidently state that our models were consistent in making predictions, thus even if different training

data were used, they could still make a good estimate of the target variable. Additionally, we can infer that our sampled data points were very close to where our nine models predicted they would be.

The results of the McNemar tests on validating the performance of our nine models suggest good performance on the XGBoost and RF classifiers. The pairwise tests on XGBoost and RF suggest that there is a significant difference between the classifiers ($p < 0.0001$), which is smaller than our significance threshold ($\alpha = 0.0065$). Therefore, we reject our null hypothesis. We discovered that the XGBoost got 868 predictions right that RF got wrong. On the contrary, RF got 555 predictions correct that XGBoost got wrong. As such, based on this 1.5 : 1 ratio, we may conclude that XGBoost performed substantially better than RF. Additionally, we performed comparative pairwise tests on all our models (LR, KNN, DT, SVM, ANN, BC, and Bagging)

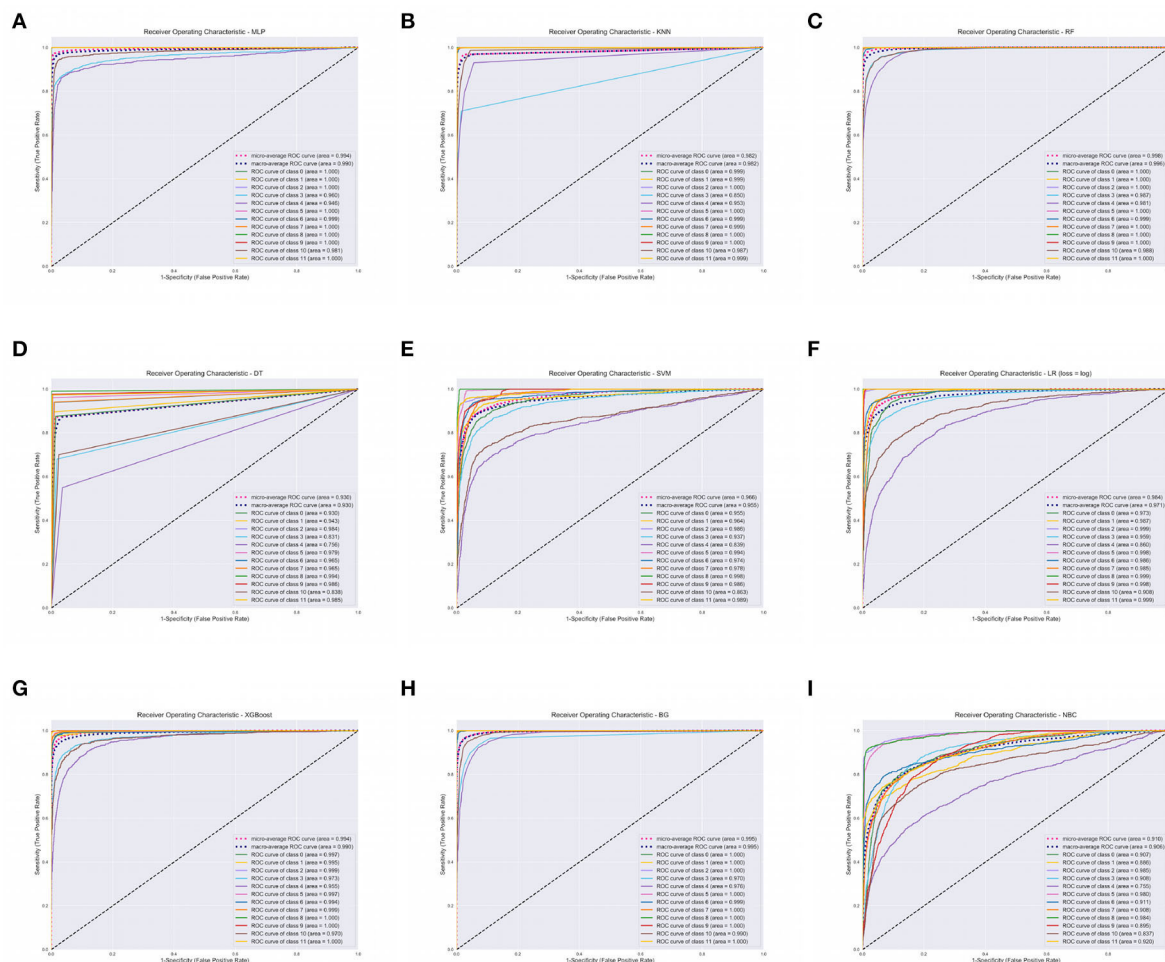


FIGURE 2

Area Under Receiver Operating Characteristics (AUROC) of our nine classifiers using combined questionnaire responses and narratives. (A) AUROC for ANN. (B) AUROC for KNN. (C) AUROC for RF. (D) AUROC for DT. (E) AUROC for SVM. (F) AUROC for LR. (G) AUROC for XGBOOST. (H) AUROC for BG. (I) AUROC for BC. XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.

and our best classifiers XGBoost and RF. Based on the tests, we can objectively reject our null hypothesis and state that there is a significant difference between our two best classifiers and the other seven classifiers in terms of model performance ($p < 0.0001$) smaller than our significance threshold ($\alpha = 0.0065$).

7. CCVA algorithm evaluation using CSMFs

This extract highlights how the InterVA and InSilico algorithms were evaluated using CSMFs. We also present CSMF and CCC as evaluation metrics for the InterVA algorithm.

Figure 3 presents the 12 leading causes of death over time as determined by the InterVA algorithm using only one CoD. We observe that between the years 1993 and 2015, HIV/AIDS was the leading CoD across the population (CSMF=0.2739). This was

closely followed by Pulmonary Tuberculosis (CSMF=0.1987) and Other Infectious/parasitic diseases (CSMF=0.1385). These three causes alone accounted for up to 61% of all deaths in the population during this period. The InterVA algorithm performance using CSMF accuracy and CCC attained values of 83% and 0.36, respectively.

8. Trend and pattern analysis using ML and CCVA approaches

8.1. CCVA algorithms

This section presents mortality trend and pattern analysis using conventional CCVA algorithms based on gender (Figure 4A), age (Figure 4B), and population over time

(Figure 4C), using data from structured questions. The visualizations are given in Figure 4.

We investigated the average age at death for the 12 leading causes of death. We discovered that both men and women

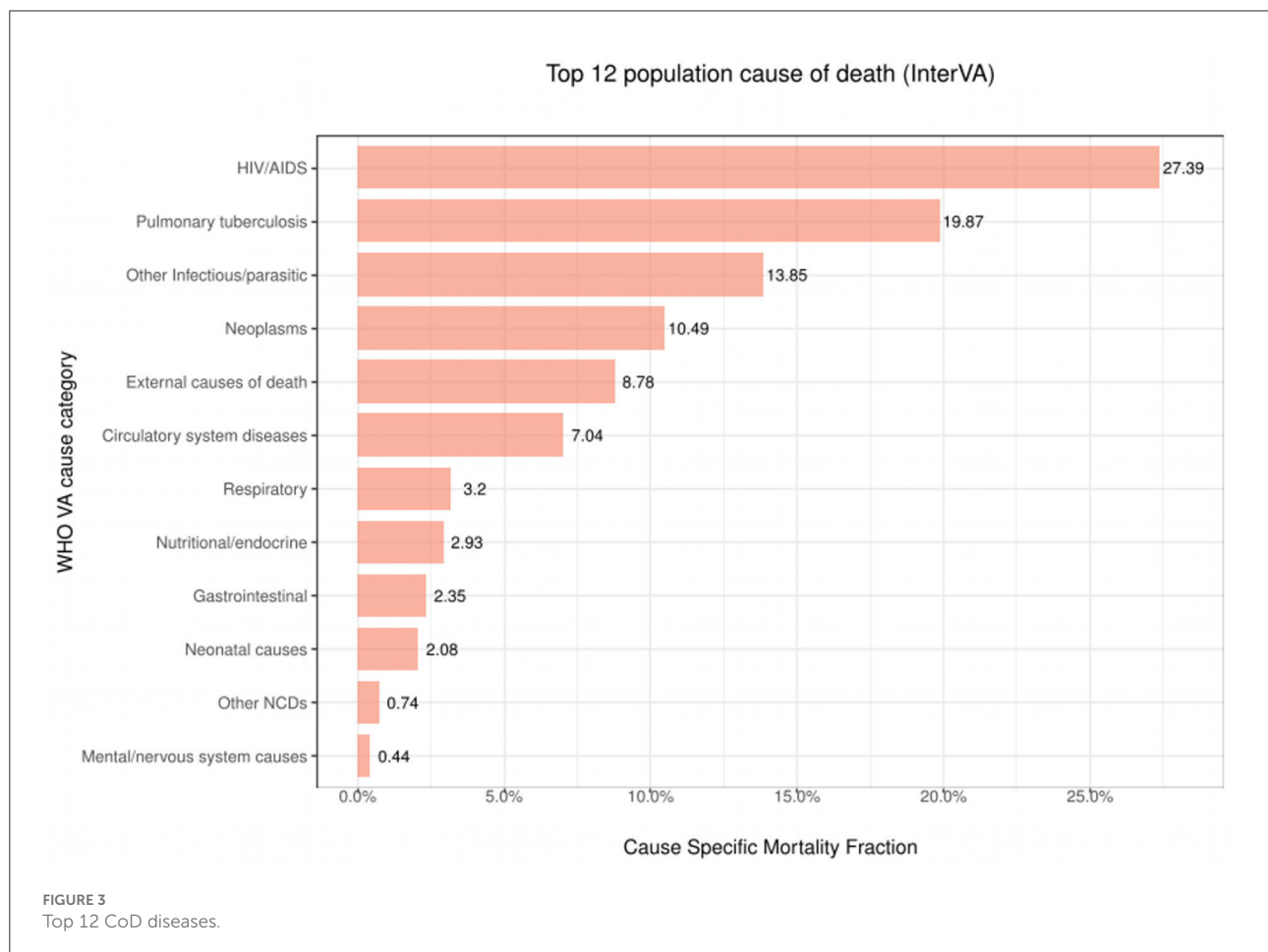
TABLE 6 Statistical tests of our nine models.

Model name	Model scores		
	Mean	Standard deviation	Rank sum
XGBoost	0.9622614	0.003209	836.00
RF	0.9566394	0.0030548	735.50
ANN	0.9530553	0.0025771	663.50
Bagging	0.9216445	2.91e+07	585.00
KNN	0.9015075	0.0033769	447.00
DT	0.8671503	0.003984	255.00
LR	0.7509405	0.0124037	155.00
BC	0.698092	0.0081906	55.00
SVM	0.6783361	0.0054433	50.00

XGBoost, eXtreme Gradient Boosting; RF, Random Forest; ANN, Artificial Neural Network; KNN, K-Nearest Neighbor; SVM, Support Vector Machine; BG, Bagging; DT, Decision Tree; LR, Logistic Regression; BC, Bayes Classifier.

were more likely to die from any disease at an average age of 40 years (mean=40, median=39, IQR=36, SD=26), despite the sex. We notice more women's deaths from HIV and circulatory diseases. On the contrary, we notice more male deaths from other infectious diseases, tuberculosis, and external causes (refer to Figure 4A). However, these differences were not statistically significant.

Figure 4B depicts percentages of mortality trends across all age groups. To determine mortality across age groups, the data were grouped into five bins "0–4," "5–14," "15–49," "50–64," and "65+." We significantly notice a declining trend in the number of deaths among persons aged between 0 and 4 years over time. In the earlier years of the Agincourt HDSS, there appears a declining trend in the number of deaths among individuals 65 years and above. However, this pattern is reversed and the mortality in the same age category is gradually increasing since the mid-2000s. Similarly, we also notice an almost comparable trend in the 50–64 age group to that of the 65+ group but the trend is gentle and stable. Among the 5–14 and 15–49 age groups, the number of deaths is appearing to be constant over time.



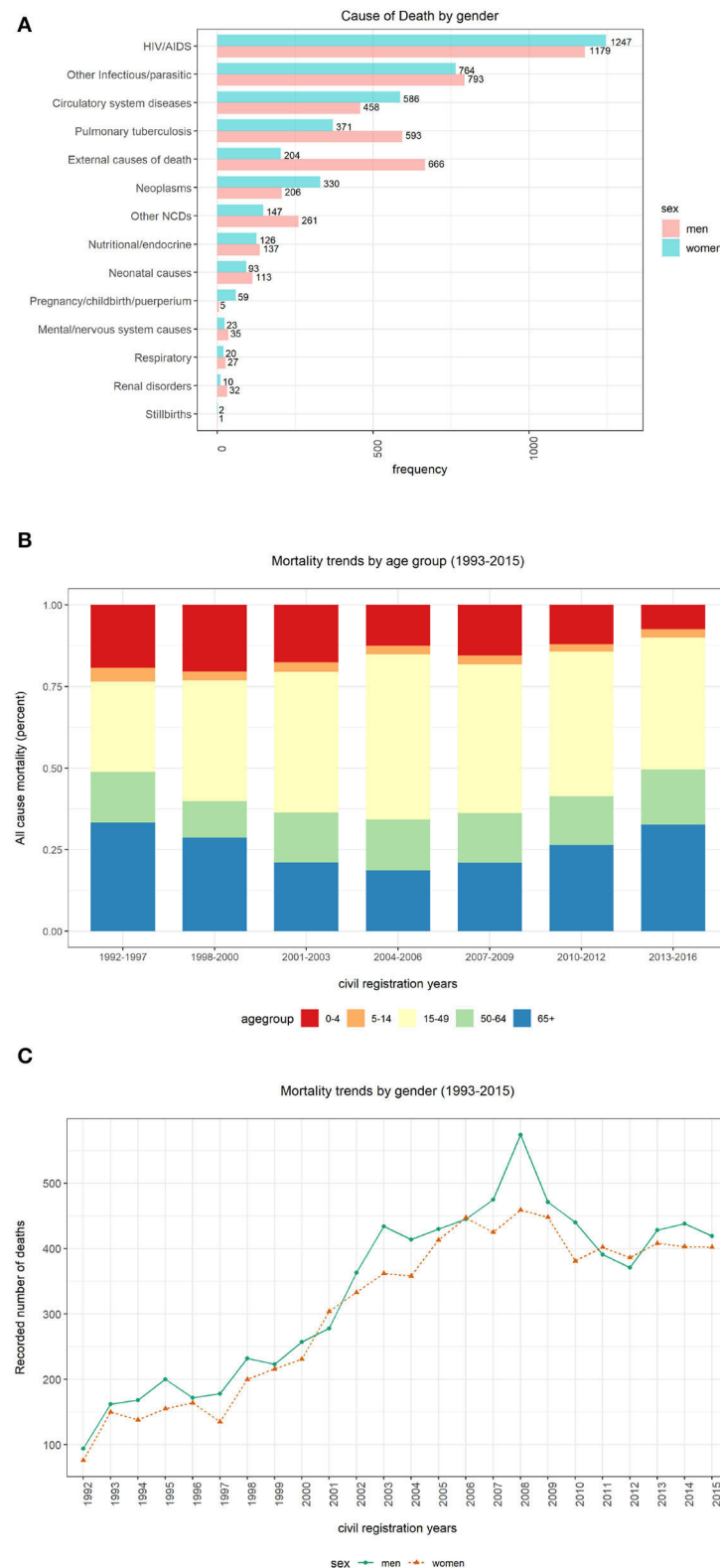


FIGURE 4

Computer Coded Verbal Autopsy (CCVA) mortality trends based on age, population, and gender. (A) Cause of death by sex. (B) Percentage of deaths by age group. (C) Yearly mortality trends by gender.

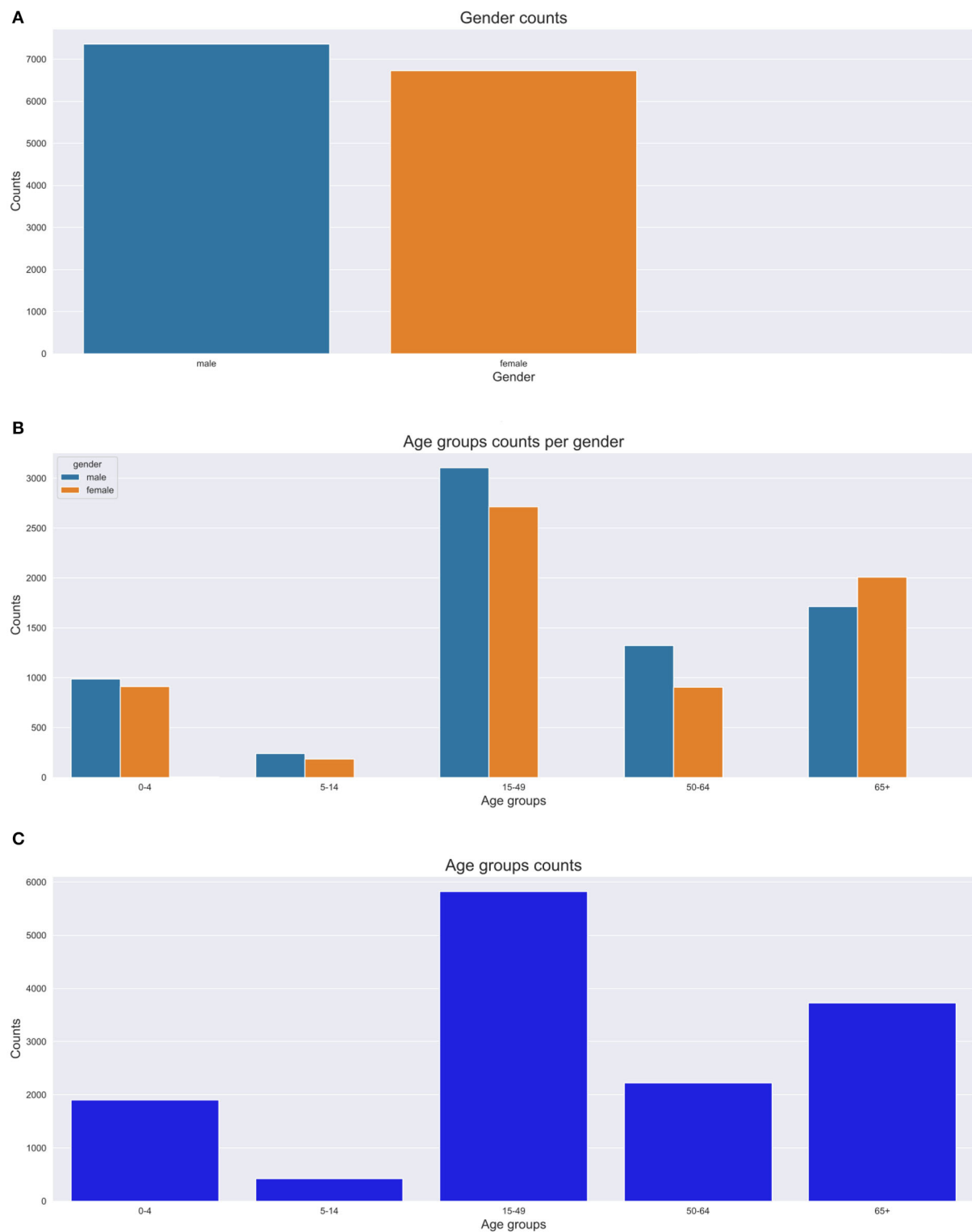


FIGURE 5
Gender and age group counts graphs. (A) Gender count. (B) Age group count per gender. (C) Age group count.

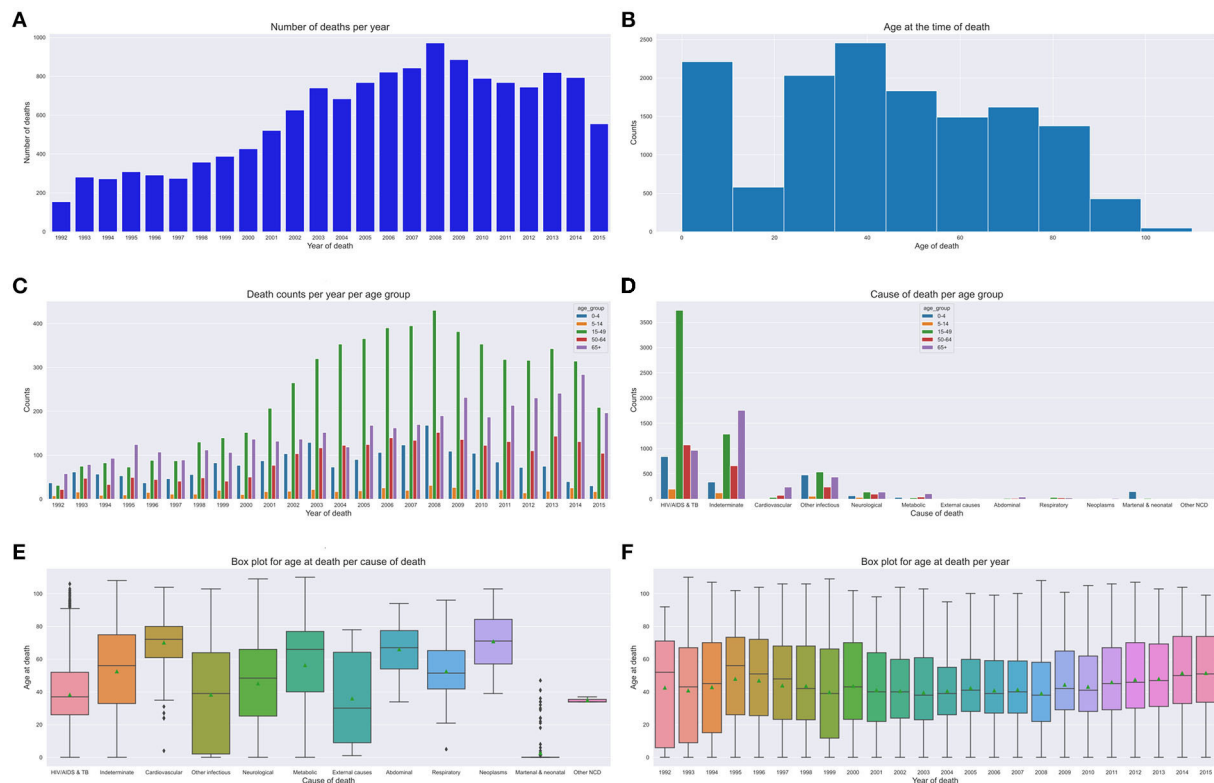


FIGURE 6
Mortality trends across age groups. (A) Number of deaths over time. (B) Age at death count. (C) Yearly death count across age groups. (D) Age group CoD count. (E) CoD and age death. (F) Age at death per year.

Figure 4C shows mortality trends based on gender over time. A total of 16,063 observations was collected, and composed of 52% men ($n = 8,354$) and 48% women ($n = 7,709$). We observe a gentle but steady increase in mortality between the years 1993 and 2000. This pattern rapidly accelerates among men between 2001 and 2008 before gradually declining.

8.2. ML techniques

Figures 6A–C depict the number of deaths over time, age at death count and yearly death count across age groups respectively. In this section, we present the results of our trend and pattern analysis using ML approaches to mortality based on gender, age, and population over time using narrative data combined with structured questions. We start by looking at the general distribution of our population based on gender, as depicted in Figure 5A, age groups (Figures 5B,C). All these graphs are depicted in Figure 5. We observe that there were more male deaths than female deaths. Most of the deaths were within the 15–49 and 65+ age groups.

We analyzed our mortality trend and pattern based on age groups as in Figure 5. We observe that most deaths are within

the 15–49 and 65+ age groups. The 65+ age group had more deaths recorded in the 1990s with a gradual increase till 2014. We also discovered that the 15–49 age group trend sharply increases till 2008 and then steadily goes down till 2015. We notice a constant trend for the 5–14 year age group over time. There is a high number of deaths from HIV causes affecting mostly the 15–49 age group. We also notice that most deaths appear to be common in the 0–10 year age group and 30–80 years age groups. Conversely, we notice fewer deaths for 80+ years.

Figures 6E,F depict our boxplots on CoD and age at death over time and age at death per year. On average, the population died of HIV/AIDS or tuberculosis which was the leading CoD at a median age of 38 years. The plots depict an average death age of 66 years succumbed to cardiovascular, neoplasm, metabolic, or abdominal diseases. Worth taking note of is the death from other infectious disease causes that show a dissimilar trend across all age groups. Additionally, on average, most of the cases died of metabolic causes at an elderly age of 65 years. Other Non Communicable Diseases (NCDs) causes of death were more prevalent in the 30–35-year-old age group and neonatal and maternal causes in their first year (shown by the narrow IQRs). CoD from neurological and respiratory causes show a mortality

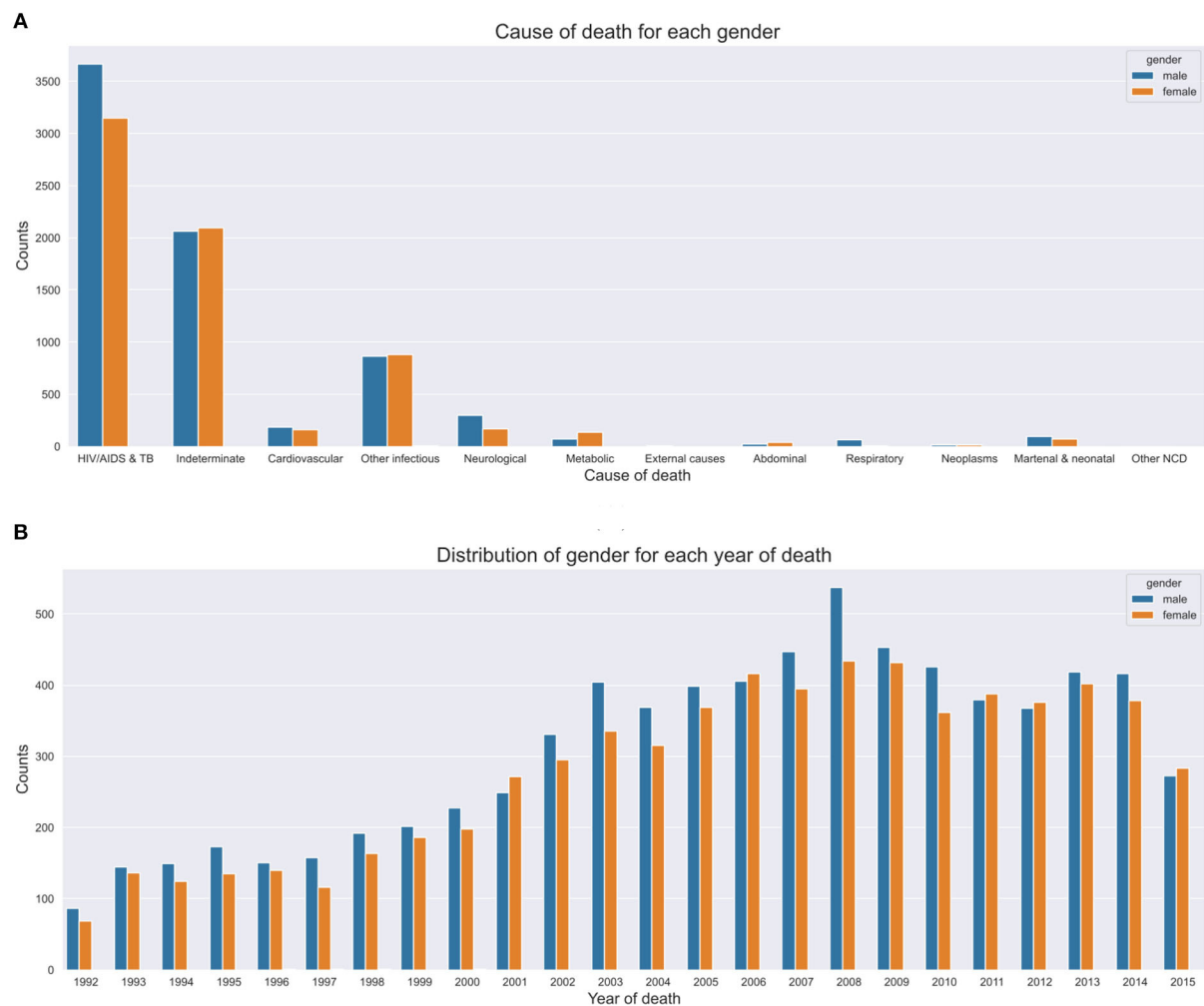


FIGURE 7
Yearly CoD based on gender. (A) CoD based on gender. (B) Yearly CoD based on gender.

trend and pattern that illustrates an average age at death of 50 years. We observe that there were more deaths in men than women, despite the cause. There is a gradual up-trend from 1992 (less than 100 deaths) to 2008 (almost 500 deaths) and a steady decline in the number of deaths from 2009 (refer to Figures 7A,B). Figure 6D illustrates that between the years 1993 and 1997, the average life expectancy was approximately 50 years. However, from 1998 to 2010, life expectancy significantly dropped and the population was dying at a much younger age of 40 years on average. From the year 2011, we see a gentle improvement in life expectancy.

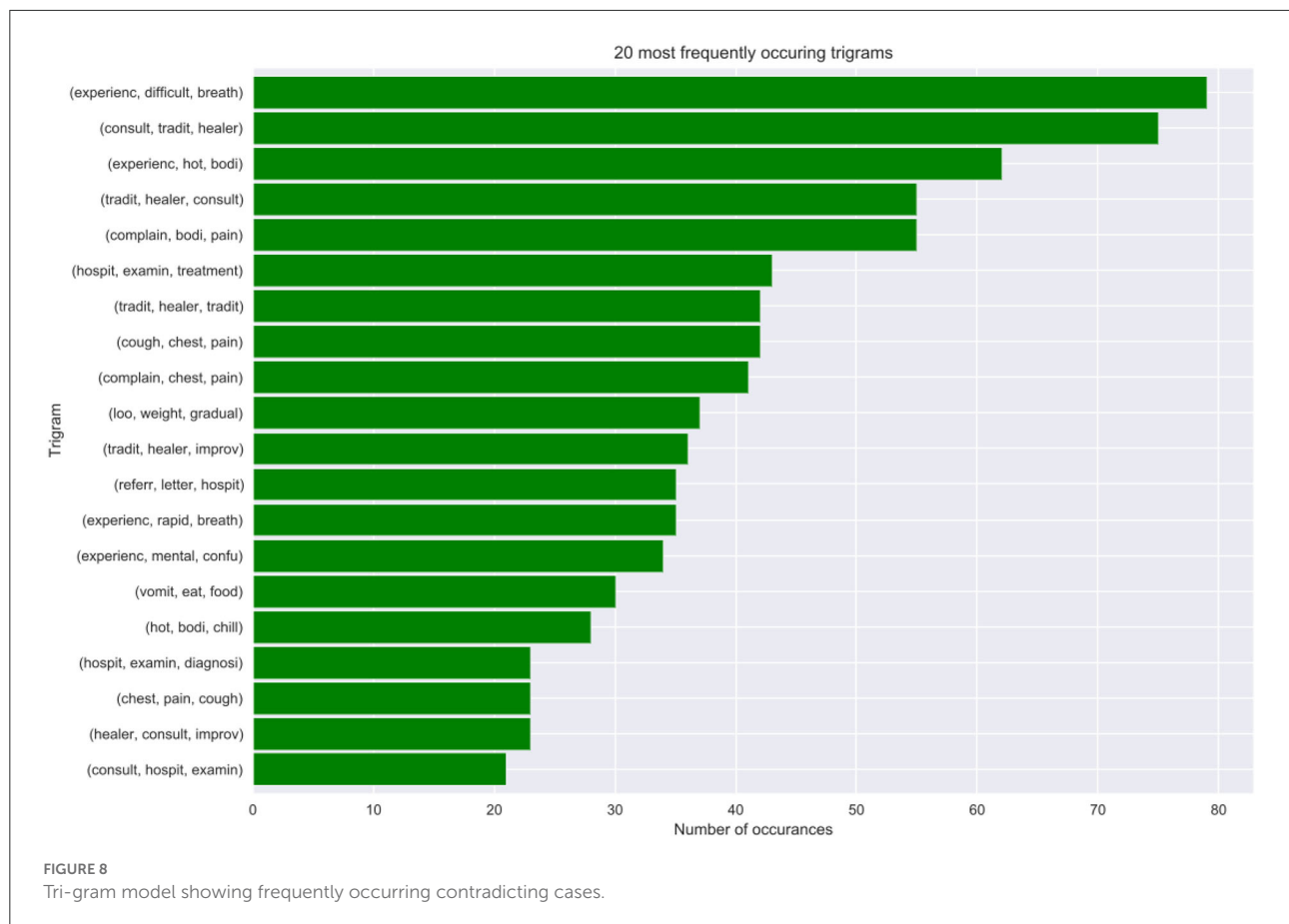
8.3. Analysis of contradicting cases

The extract details an analysis of the structure and semantics of features in cases where the doctors are in disagreement. We

discovered that approximately 16% of the observations in the VA dataset denote cases where the experts are in disagreement. Further analysis of the structure and semantics availed insights that in most cases the narratives entail information related to traditional healers' visits and consultations. Additionally, we deduced that also this is a result of cases where the captured information entails the imminent loss of weight, vomiting, and having a fever leading to an unexplainable sudden death. Figure 8 below shows our n-gram model of what was mined from the contradicting narratives.

8.4. Model best predictors

This part discusses how the most important narrative features were identified as the best predictors of our models. We chose the bi-grams as they show an evenly distributed



frequency analysis of features (refer to Figure 9). As *n* increases the features start having more or less the same frequency.

9. Discussion

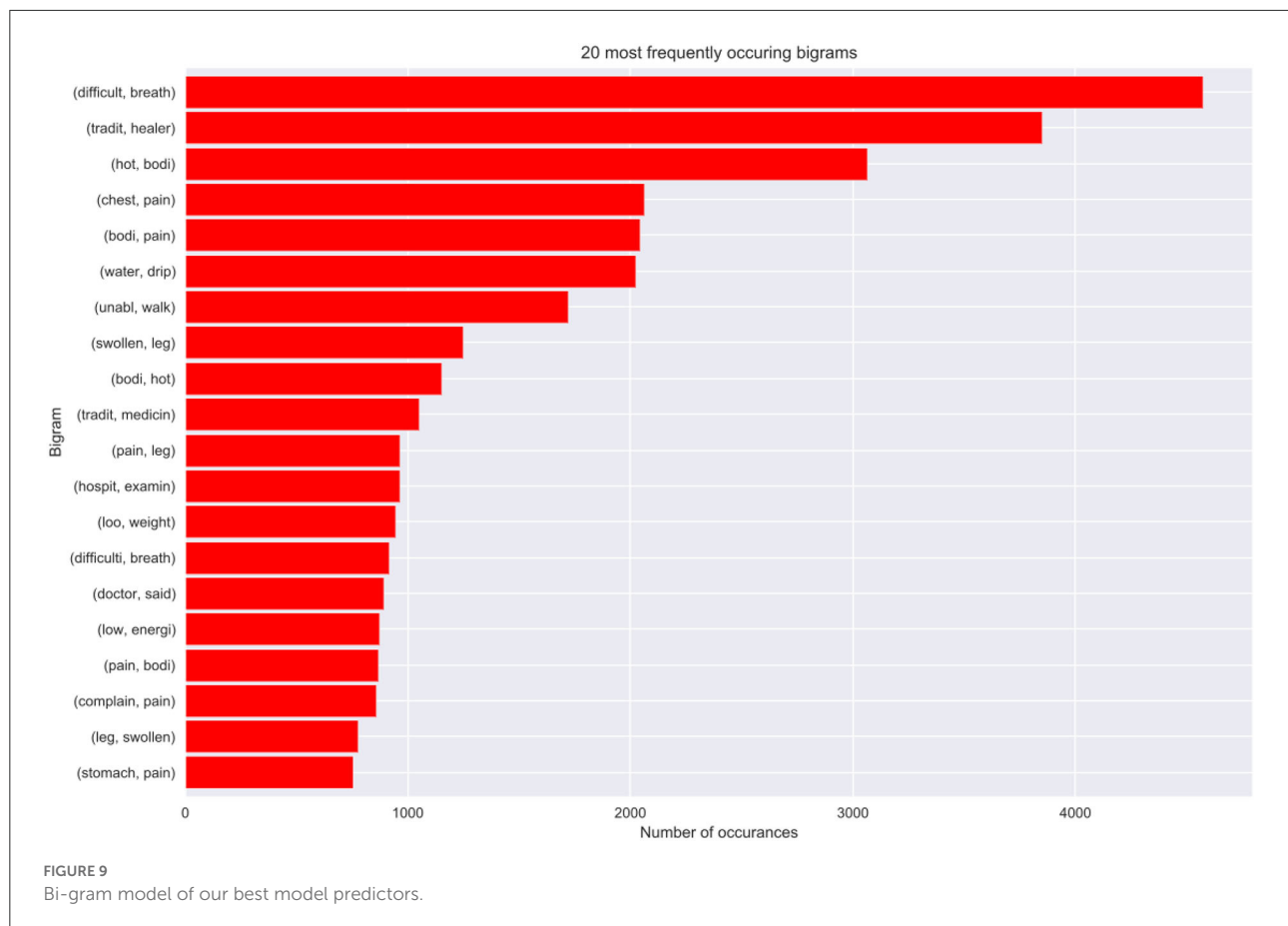
The process of determining causes of death using VAs still remains a manual task and suffers from many drawbacks (refer to Section 1). This negatively affects the VA reporting process, despite it being vital for strengthening health priorities and informing civil registration systems. Therefore, under such circumstances, there is a great need for innovative novel automated approaches to address these problems thereof.

In this study, we explore various VA data types, despite most studies in literature reporting results based on the classical dataset for CoD determination using ML approaches. Our aim is to investigate if the narratives can improve or enhance model prediction if they are added to the responses from the structured questionnaire. Our deductions suggest that the VA narratives have vital valuable information that should be used in model prediction. Consequently, we identify the best model predictors from the narratives. We further do a mortality pattern and trend analysis based on age, population, and gender over time. We also

do a structure and semantic analysis of narratives in cases where the experts agree and also disagree. To add to our findings, we also investigate the best features that contribute to our models from the narratives.

Generally, the results of all our ML models used in this study, demonstrate that our models exhibited consistent superior performance on all datasets. This further reinforces the notion that ML approaches can be used as alternatives to conventional approaches for CoD determination using VAs. Ensemble classifiers (XGBoost, bagging), tree based models (DT, RF), ANN and KNN performed exceptionally well on all datasets. Our results of the combined dataset do not exhibit a consistent model performance, as most models slightly drop in model performance. This can be attributed to the fact that the combined dataset creates high dimensionality of the feature space and this triggers model complexity with too many noisy data points. The CCVA approach, InterVA, attained a CSFM accuracy of 83% and CCC of 36%.

Our CCVA approaches and ML techniques produced similar mortality trends and patterns based on age, population, and gender. Interestingly, we observed that in the first decade of the civil registration system, the average life expectancy was approximately 50 years. However, in the second decade,



life expectancy significantly dropped and the population was succumbing to death at a slightly lower average age of 40 years. This suggests CoD mostly from the leading HIV and tuberculosis related causes. Interestingly, in the third decade, we see a gradual improvement in life expectancy, possibly attributed to the implementation of effective health intervention programmes. We notice that cardiovascular, neoplasm, neurological, respiratory, and metabolic CoD mainly affected the elderly. We observe that other infectious diseases and external causes affected the population disproportionately across all age groups, with the latter having an average age at death of 30 years. Despite the expected CoD from neonatal and maternal causes, we can also infer that those with HIV had a lower life expectancy as compared to the other CoDs. Of interest, is that most undetermined causes of death are found within the 65+ age group. This suggests that as the elderly population grows older, their health state deteriorates and they succumb to many symptoms that can lead to untimely hard to explain deaths. Other NCDs, causes of death were more prevalent in the younger age groups. We also discovered that sudden deaths are common in the elderly, suggesting symptoms, such as imminent loss of weight, vomiting, and having a fever leading to an unexplainable

premature death. Generally, we notice more deaths in men than women.

We, therefore, propose that optimal model performance should be set at 80% accuracy. In cases where the ML model fails to reach a threshold value of 80% accuracy in terms of performance, we propose an expert's intervention for further exploration and assessment. Conversely, in cases where the experts are failing or do not reach a consensus, we recommend the help of the machine to make predictions. Most of these cases where the machine can assist, entail narratives where the interviewee details most content about the deceased circumstances and events that led to death based on traditional healer visits and consultation. Interestingly, we still found out that traditional healer consultations are a common practice in the population as they occurred frequently in our model best predictors. This cements the notion that most people in the HDSS seek traditional ways for their terminal illnesses, rather than western means. This finding opened exciting avenues for future study, which will focus on sequential text modeling with the aim of fully understanding treatment sequences for terminal illnesses. Nevertheless, in cases where the physicians were in agreement, these narrations about traditional healer's consultations were supplemented by enough

symptoms that made it possible for the experts to give a proper diagnosis. We also discovered that our model's best predictors entail matching symptoms with those in the responses to the structured questionnaire.

The results of this study, consistent with several studies that used VA data to determine CoD, suggest that ML approaches can accurately classify CoD from VA narratives. However, in most cases, statistical approaches and CCVA approaches are always outperformed by ML approaches (1, 8, 9, 13, 23, 25, 34). Therefore, it is imperative for future research studies to incorporate effective data handling strategies (8). This study adds to the existing body of literature, suggesting that automated approaches can be used as alternatives to PCVA in a cost effective way, producing real-time results that are consistent, accurate, and error free, thus strengthening health priorities. As such, VA processes are still key in capturing civil registration data where death occurs outside health facilities, up until a point when deaths start to take place in areas where it can be documented. Given these complexities, there is a great need for novel automated approaches that can be used as alternatives (22).

The strength of this study lies in the application of various ML and CCVA algorithms to various VA data types. Moreover, our sample size was large and representative of deaths that occurred at Agincourt HDSS that were captured in a standard way. Moreover, our mortality trend and pattern analysis gave us valuable insights into our HDSS and this can be used to inform policy and practice. This enforces generalization and comparability across studies. On the contrary, this study had limitations of data quality described in Section 1.2.

10. Conclusion

In general, this study demonstrates that ML techniques can be used as alternatives in determining CoD from VA narratives producing results comparable to physician diagnosis. Our findings should be used to inform policy and practice and enforce effective health intervention programmes and resource prioritization to reduce the mortality rate and prolong life expectancy. As such, they can help close the gap in civil registration systems. Our comparative analysis of the ML models on various VA datasets enforces comparability and generalization, thus availing a baseline study for future research. Future work will entail exploring deep learning methods and employing novel techniques such as transfer learning to determine CoD.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by University of the Witwatersrand, Faculty of Health Sciences Ethics Committee (Certificate No. M1911132). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s), nor the minor(s)' legal guardian/next of kin, for the publication of any potentially no identifiable images or data are presented in the manuscript or data included in this article.

Author contributions

MM and VO did all algorithm experiments with the help of TC. MM and all the other authors drafted and critically revised the study. The first draft of the manuscript was written by MM and all the authors commented on previous versions of the manuscript. All the authors contributed to the study's conception and design, read, and approved the final manuscript.

Funding

This study was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) (Grant No. DEL-15-005). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the United Kingdom government.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. *BMC Med Inform Decis Mak.* (2019) 19:127. doi: 10.1186/s12911-019-0841-9
- Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, et al. The WHO 2016 verbal autopsy instrument: an international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS Med.* (2018) 15:e1002486. doi: 10.1371/journal.pmed.1002486
- Thomas LM, D'mbruso L, Balabanova D. Verbal autopsy in health policy and systems: a literature review. *BMJ Global Health.* (2018) 3:e000639. doi: 10.1136/bmjgh-2017-000639
- Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: current practices and challenges. *Bull World Health Organ.* (2006) 84:239–245. doi: 10.2471/BLT.05.027003
- Mapoma CC, Munkombwe B, Mwango C, Bwalya BB, Kalindi A, Gona NP. Application of verbal autopsy in routine civil registration in Lusaka District of Zambia. *BMC Health Serv Res.* (2021) 21:408. doi: 10.1186/s12913-021-06427-y
- Lozano R, Lopez AD, Atkinson C, Naghavi M, Flaxman AD, Murray CJ. Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr.* (2011) 9:1–13. doi: 10.1186/1478-7954-9-32
- Reeves BC, Quigley M. A review of data-derived methods for assigning causes of death from verbal autopsy data. *Int J Epidemiol.* (1997) 26:1080–9. doi: 10.1093/ije/26.5.1080
- Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl.* (2019) 116:494–520. doi: 10.1016/j.eswa.2018.09.034
- Desai N, Aleksandrowicz L, Miasnikof P, Lu Y, Leitao J, Byass P, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low- and middle-income countries. *BMC Med.* (2014) 12:20. doi: 10.1186/1741-7015-12-20
- James SL, Flaxman AD, Murray CJ. Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. *Popul Health Metr.* (2011) 9:31. doi: 10.1186/1478-7954-9-31
- Byass P, Herbst K, Fottrell E, Ali MM, Odhiambo F, Amek N, et al. Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia. *J Glob Health.* (2015) 5: 010402. doi: 10.7189/jogh.05.010402
- McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. *J Am Stat Assoc.* (2016) 111:1036–49. doi: 10.1080/01621459.2016.1152191
- Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D, et al. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med.* (2015) 13:286. doi: 10.1186/s12916-015-0521-2
- Clark SJ, Li Z, McCormick TH. Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. *arXiv preprint arXiv:180307141.* (2018). doi: 10.48550/arXiv.1803.07141
- Leitao J, Desai N, Aleksandrowicz L, Byass P, Miasnikof P, Tollman S, et al. Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low- and middle-income countries: systematic review. *BMC Med.* (2014) 12:22. doi: 10.1186/1741-7015-12-22
- Murray CJ, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC Med.* (2014) 12:1–19. doi: 10.1186/1741-7015-12-5
- Kalter HD, Perin J, Black RE. Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *J Glob Health.* (2016) 6:010601. doi: 10.7189/jogh.06.010601
- Quigley MA, Chandramohan D, Setel P, Binka F, Rodrigues LC. Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Trop Med Int Health.* (2000) 5:33–9. doi: 10.1046/j.1365-3156.2000.00517.x
- Nithya B, Ilango V. Predictive analytics in health care using machine learning tools and techniques. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. Madurai: IEEE (2017). p. 492–9.
- Flaxman AD, Vos T. Machine learning in population health: opportunities and threats. *PLoS Med.* (2018) 15:e1002702. doi: 10.1371/journal.pmed.1002702
- Moran KR, Turner EL, Dunson D, Herring AH. Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *J R Stat Soc Ser C Appl Stat.* (2019) 70:532–57. doi: 10.1111/rssc.12468
- Idicula-Thomas S, Gawde U, Jha P. Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: national survey of 18,000 verbal autopsies in the Million Death Study in India. *BMC Public Health.* (2021) 21:1–11. doi: 10.1186/s12889-021-11829-y
- Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: a comparative study. *J Forensic Leg Med.* (2018) 57:41–50. doi: 10.1016/j.jflm.2017.07.001
- Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:14024380.* (2014). doi: 10.48550/arXiv.1402.4380
- Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Classification of forensic autopsy reports through conceptual graph-based document representation model. *J Biomed Inform.* (2018) 82:88–105. doi: 10.1016/j.jbi.2018.04.013
- Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Inform Decis Mak.* (2015) 15:53. doi: 10.1186/s12911-015-0174-2
- Mwanyangala MA, Urassa HM, Rutashobya JC, Mahutanga CC, Lutambi AM, Maliti DV, et al. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of Tanzania. *Popul Health Metr.* (2011) 9:41. doi: 10.1186/1478-7954-9-41
- Quigley MA, Chandramohan D, Rodrigues LC. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *Int J Epidemiol.* (1999) 28:1081–7. doi: 10.1093/ije/28.6.1081
- Boulle A, Chandramohan D, Weller P. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *Int J Epidemiol.* (2001) 30:515–20. doi: 10.1093/ije/30.3.515
- Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr.* (2011) 9:29. doi: 10.1186/1478-7954-9-29
- Danso S, Johnson O, Ten Asbroek A, Soromekun S, Edmond K, Hurt C, et al. A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death. *ICAME J.* (2013) 37:37–69.
- Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: A content analysis. *Biomed Inform Insights.* (2010) 3:BII-S4706. doi: 10.4137/BII.S4706
- Murtaza SS, Kolpak P, Bener A, Jha P. Automated verbal autopsy classification: using one-against-all ensemble method and Naive Bayes classifier. *Gates Open Res.* (2018) 2:63. doi: 10.12688/gatesopenres.12891.1
- Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS ONE.* (2017) 12:e0170242. doi: 10.1371/journal.pone.0170242
- Clark SJ. A guide to comparing the performance of VA algorithms. *arXiv preprint arXiv:180207807.* (2018). doi: 10.48550/arXiv.1802.07807
- Chandramohan D, Setel P, Quigley M. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *Int J Epidemiol.* (2001) 30:509–14. doi: 10.1093/ije/30.3.509
- Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. *Glob Health Action.* (2014) 7:25596. doi: 10.3402/gha.v7.25596
- King G, Lu Y, et al. Verbal autopsy methods with multiple causes of death. *Stat Sci.* (2008) 23:78–91. doi: 10.1214/07-STS247
- Korde V, Mahender CN. Text classification and classifiers: a survey. *Int J Artif Intell Appl.* (2012) 3:85. doi: 10.5121/ijaia.2012.3208
- Zaki MJ, Meira Jr W. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press (2019).
- Leskovec J, Rajaraman A, Ullman JD. *Mining of Massive Data Sets*. Cambridge: Cambridge University Press (2020).
- Pičulin M, Smole T, Žunković B, Kokalj E, Robnik-Šikonja M, Kukar M, et al. Disease progression of hypertrophic cardiomyopathy: modeling using machine learning. *JMIR Med Inform.* (2022) 10:e30483. doi: 10.2196/30483

43. Yang Y, Zheng J, Du Z, Li Y, Cai Y, et al. Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: retrospective study. *JMIR Med Inform.* (2021) 9:e30277. doi: 10.2196/30277
44. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. *Syst Sci Control Eng.* (2014) 2:602–9. doi: 10.1080/21642583.2014.956265
45. Poole DL, Mackworth AK. *Artificial Intelligence: Foundations of Computational Agents.* Cambridge: Cambridge University Press (2010).
46. Byrne MD. Machine learning in health care. *J PeriAnesthesia Nurs.* (2017) 32:494–6. doi: 10.1016/j.jopan.2017.07.004
47. Iqbal Z, Ilyas R, Shahzad W, Inayat I. A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings. In: *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE).* Penang: IEEE (2018). p. 406–11.
48. Li ZR, McCormick TH, Clark SJ. InterVA4: an R package to analyze verbal autopsy data. In: *Center for Statistics and the Social Sciences, University of Washington.* Vienna, Austria: R Foundation for Statistical Computing (2014).
49. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popul Health Metr.* (2011) 9:1–11. doi: 10.1186/1478-7954-9-28
50. Murray CJ, Lopez AD, Black R, Ahuja R, Ali SM, Baqui A, et al. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Popul Health Metr.* (2011) 9:27. doi: 10.1186/1478-7954-9-27
51. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* (1998) 10:1895–923. doi: 10.1162/089976698300017197



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Slavko Žitnik,
University of Ljubljana, Slovenia
Yafei Wu,
Xiamen University, China

*CORRESPONDENCE

Sheng-Feng Sung
richard.sfsung@gmail.com;
sfusng@cych.org.tw

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 01 August 2022

ACCEPTED 13 September 2022

PUBLISHED 29 September 2022

CITATION

Tsai H-C, Hsieh C-Y and Sung S-F
(2022) Application of machine learning
and natural language processing for
predicting stroke-associated
pneumonia.
Front. Public Health 10:1009164.
doi: 10.3389/fpubh.2022.1009164

COPYRIGHT

© 2022 Tsai, Hsieh and Sung. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Application of machine learning and natural language processing for predicting stroke-associated pneumonia

Hui-Chu Tsai¹, Cheng-Yang Hsieh^{2,3} and Sheng-Feng Sung^{4,5*}

¹Department of Radiology, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi, Taiwan, ²Department of Neurology, Tainan Sin Lau Hospital, Tainan, Taiwan, ³School of Pharmacy, Institute of Clinical Pharmacy and Pharmaceutical Sciences, College of Medicine, National Cheng Kung University, Tainan, Taiwan, ⁴Division of Neurology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi, Taiwan, ⁵Department of Nursing, Min-Hwei Junior College of Health Care Management, Tainan, Taiwan

Background: Identifying patients at high risk of stroke-associated pneumonia (SAP) may permit targeting potential interventions to reduce its incidence. We aimed to explore the functionality of machine learning (ML) and natural language processing techniques on structured data and unstructured clinical text to predict SAP by comparing it to conventional risk scores.

Methods: Linked data between a hospital stroke registry and a deidentified research-based database including electronic health records and administrative claims data was used. Natural language processing was applied to extract textual features from clinical notes. The random forest algorithm was used to build ML models. The predictive performance of ML models was compared with the A²DS², ISAN, PNA, and ACDD⁴ scores using the area under the receiver operating characteristic curve (AUC).

Results: Among 5,913 acute stroke patients hospitalized between Oct 2010 and Sep 2021, 450 (7.6%) developed SAP within the first 7 days after stroke onset. The ML model based on both textual features and structured variables had the highest AUC [0.840, 95% confidence interval (CI) 0.806–0.875], significantly higher than those of the ML model based on structured variables alone (0.828, 95% CI 0.793–0.863, $P = 0.040$), ACDD⁴ (0.807, 95% CI 0.766–0.849, $P = 0.041$), A²DS² (0.803, 95% CI 0.762–0.845, $P = 0.013$), ISAN (0.795, 95% CI 0.752–0.837, $P = 0.009$), and PNA (0.778, 95% CI 0.735–0.822, $P < 0.001$). All models demonstrated adequate calibration except for the A²DS² score.

Conclusions: The ML model based on both textual features and structured variables performed better than conventional risk scores in predicting SAP. The workflow used to generate ML prediction models can be disseminated for local adaptation by individual healthcare organizations.

KEYWORDS

machine learning, natural language processing, pneumonia, prediction, risk score, stroke

Introduction

The global burden of stroke is huge and rising (1). According to the most updated statistics from the World Stroke Organization, the global incidence of strokes exceeds 12 million annually and the number of prevalent strokes is more than 100 million worldwide (2). Apart from direct neurological damage, stroke patients are prone to medical complications such as infection (3). Approximately 21–30% of stroke patients develop post-stroke infections, with pneumonia accounting for a third to half of them (4, 5). Stroke-associated pneumonia (SAP) is not only associated with substantial morbidity and mortality (6–8) but also increases direct healthcare costs (9). Despite the advances in acute stroke treatment over the past decades, the frequency of SAP remains unchanged (4). Effective strategies and interventions are therefore urgently needed to reduce the burden of pneumonia, a potentially preventable complication of stroke.

To prevent SAP, a fundamental first step is the early recognition of high-risk patients, for whom appropriate preventive measures can be taken. Besides, the high-risk patient group is also the main target population for which clinical trials can be designed to test novel interventions for the prevention of pneumonia. Analysis of patient data stored in the Virtual International Stroke Trials Archive showed that most post-stroke pneumonias occurred in the first week and its incidence peaked on the third day after stroke onset (10). Consequently, the risk of developing pneumonia should be assessed as early as possible following stroke. To date, several integer-based risk scores have been developed for predicting SAP (11). Most of the risk models make predictions based on similar predictor variables, such as age, stroke severity, and the presence of dysphagia (11). Hence it is no surprise that these risk models perform comparably regarding discrimination and calibration (11–13). On the other hand, almost all existing SAP prediction models were developed using logistic regression analysis, thus ignoring the potential complex interactions between variables.

With the advances in data science and artificial intelligence, data-driven machine learning (ML) approaches have been increasingly used to develop prediction models in the medical domain (14). These approaches have also been introduced to develop SAP prediction models (15, 16). Compared to conventional parametric techniques like logistic regression, ML approaches have several advantages such as the capability of dealing with high-dimensional data and modeling complex and non-linear relations between data. Furthermore, the ubiquitous adoption of electronic health record (EHR) systems provides an opportunity to use various types of structured and unstructured data for data-driven prediction of clinical outcomes (17–19). Using natural language processing techniques, information extracted from unstructured clinical text has the potential to improve the performance of clinical prediction models (20, 21).

Inspired by these ideas, we aimed to explore the value of combining both structured and unstructured textual data in developing ML models to predict SAP.

Materials and methods

Data sources

The data sources for this study were the hospital stroke registry and the Ditmanson Research Database (DRD), a deidentified database comprising both administrative claims data and EHRs for research purposes. [Supplementary Table 1](#) lists the general specifics of the data sources. The DRD currently holds clinical information of over 1.4 million patients, including 0.6 million inpatient and 21.5 million outpatient records. It includes both structured data (demographics, vital signs, diagnoses, prescriptions, procedures, and laboratory results) and unstructured textual data (physician notes, nursing notes, laboratory reports, radiology reports, and pathology reports). The hospital stroke registry has prospectively registered all consecutive hospitalized stroke patients since 2007 conforming to the design of Taiwan Stroke Registry (22). Currently, it has enrolled over 12,000 patients. The stroke registry consists of structured data only. Stroke severity was assessed using the National Institutes of Health Stroke Scale (NIHSS) while functional status was evaluated using the modified Rankin Scale (mRS). Information regarding patients' demographics, risk factor profiles, treatments and interventions, complications, and outcomes were collected by trained stroke case managers. To create the dataset for this study, the stroke registry was linked to the DRD using a unique encrypted patient identifier. The study protocol was approved by the Ditmanson Medical Foundation Chia-Yi Christian Hospital Institutional Review Board (approval number: 2022060). Study data were maintained with confidentiality to ensure the privacy of all participants.

Study population

The derivation of the study population is shown in [Supplementary Figure 1](#). The stroke registry was queried for all stroke hospitalizations, including both acute ischemic stroke (AIS) and intracerebral hemorrhage (ICH), between Oct 2010 and Sep 2021. Only the first hospitalization was considered for each patient. Patients who suffered an in-hospital stroke or already had pneumonia on admission and those whose records could not be linked were excluded. Patients with missing data that made the calculation of pneumonia risk scores impossible were excluded. The study population was randomly split into a training set that consisted of 75% of the patients and a holdout test set comprising the remaining 25% of the patients.

Predictor and outcome variables

The outcome variable was SAP occurring within the first 7 days after stroke onset (23). As per the protocol of the Taiwan Stroke Registry (22), the diagnosis of SAP was made according to the modified Centers for Disease Control and Prevention criteria (23). Because risk stratification at an early stage after stroke is preferred so that appropriate interventions can be applied, only information available within 24 h of admission was considered. Candidate predictors comprised demographics, pre-stroke dependency (defined as an mRS score of ≥ 3), risk factors and comorbidities, prior use of medications, physiological measurements, neurological assessment (NIHSS, Glasgow coma scale, and bedside dysphagia screening), as well as routine blood tests (Supplementary Table 2). For predictors that had multiple measurements after admission, such as physiological measurements, neurological assessment, and routine blood tests, only the first measurement was used. Missing values for continuous variables were imputed using the mean of non-missing values. Then each continuous variable was rescaled to a mean of zero and a standard deviation of one.

In the study hospital, admission notes are written in English. To extract predictor features from clinical text, we experimented with three approaches of text representation: a simple “bag-of-words” (BOW) approach, a fastText embedding approach (24), and a deep learning approach using the bidirectional encoder representations from transformers (BERT) (25).

The free text from the History of Present Illness (HPI) section of the admission note was preprocessed through the following steps: spell checking, abbreviation expansion, removal of non-word symbols, lowercase conversion, lemmatization, marking of negated words with the suffix “_NEG” using the Natural Language Toolkit mark_negation function with default parameters (https://www.nltk.org/_modules/nltk/sentiment/util.html), and stop-word removal. Lemmatization, negation marking, and stop-word removal were not needed for the BERT approach.

Supplementary Figure 2 shows an example of feature extraction and preprocessing using the BOW approach. Having no prior knowledge of what information the text can provide, we used an “open-vocabulary” approach (26) to detect features predictive of SAP. We built a document-term matrix where each column represents each unique feature (word or phrase) from the text corpus while the rows represent each patient’s clinical document. The preprocessed text was vectorized using the BOW approach with three different types of feature representation (27). In other words, the cells of the document-term matrix represent the counts of each word within each document (term frequency), the absence or presence of each word within each document (binary representation), or the term frequency with inverse document frequency weighting, respectively. Because medical terms are commonly comprised of two words or even more, we also experimented with adding word bigram features

(two-word phrases) to the basic BOW model. To reduce noise such as redundant and less informative features as well as to improve training efficiency (28), feature selection was performed by selecting the top 20 words or phrases that appeared in the documents of patients with SAP and those without based on chi-square statistics (29). Supplementary Figures 3–6 show the top 20 selected words or phrases for each feature representation method.

The fastText subword embedding model is an extension of Word2Vec, which uses skip-gram model to represent each word in the form of character n-grams (24). It allows handling out-of-vocabulary words in the training samples. We resumed training of the model from a pre-trained model called BioWordVec using the training set. Then the clinical text was vectorized using the trained model. BioWordVec was originally created from unlabeled biomedical text from PubMed and Medical Subject Headings using the fastText subword embedding model (30). Later, the original BioWordVec was extended by adding the Medical Information Mart for Intensive Care III clinical notes to the training text corpus (31).

The BERT model is a contextualized word representation model, which allows modeling long-distance dependencies in text. The BERT model is pre-trained based on masked language modeling and next sentence prediction using bidirectional transformers on the general Toronto BookCorpus and English Wikipedia corpus (25). For this study, we used a domain-specific BERT model, i.e., ClinicalBERT (32), which was pre-trained on the Medical Information Mart for Intensive Care III clinical notes. We fine-tuned the BERT model using the training set to predict SAP. The text from the training set was preprocessed and split into BERT tokens. Since the BERT model can only accommodate 512 tokens, the input text was truncated to 512 tokens. For BERT fine-tuning, the batch size was set at 16. The learning rate of the Adam optimizer was set at 2×10^{-5} and the number of epochs was 3. Then text from the training and test sets was vectorized by averaging all contextualized word embeddings output by the fine-tuned BERT model.

SAP risk scores

To compare the predictive performance of ML models, four conventional SAP risk scores (Table 1) were used as comparison models based on variables available in the dataset. The total score of each SAP risk score is calculated by summing up the scores of all its items. A higher total score indicates a greater risk of developing SAP. The A^2DS^2 score was derived from clinical data of patients with AIS from the Berlin Stroke Register (33). It comprised age (1 point for ≥ 75), atrial fibrillation (1 point), dysphagia (2 points), male sex (1 point), and NIHSS (3 points for 5–15 and 5 points for ≥ 16). The 22-point ISAN score was developed using data of patients with AIS or ICH from

TABLE 1 Risk scores for predicting stroke-associated pneumonia.

	A ² DS ²	ISAN	PNA	ACDD ⁴
Age				
≥70			+1	
≥75	+1			+1
60–69		+3		
70–79		+4		
80–89		+6		
≥90		+8		
Male	+1	+1		
Diabetes			+1	
AF	+1			
CHF				+1
Pre-stroke dependency		+2		
NIHSS				
5–15	+3	+5	+3	
≥16	+5		+5	
16–20		+8		
≥21		+10		
Dysphagia	+2			+4
Dysarthria				+1

AF, atrial fibrillation; CHF, congestive heart failure; NIHSS, National Institutes of Health Stroke Scale.

a national United Kingdom registry (34). It consisted of pre-stroke dependency (2 points), male sex (1 point), age (3 points for 60–69, 4 points for 70–79, 6 points for 80–89, and 8 points for ≥90), and NIHSS (5 points for 5–15, 8 points for 16–20, and 10 points for ≥21). The PNA score, created using data of AIS patients from a single academic institution, included age (1 point for ≥70), history of diabetes (1 point), and NIHSS (3 points for 5–15 and 5 points for >15) (35). The ACDD⁴ score, developed based on a single-site cohort of patients with AIS or ICH, was composed by age (1 point for ≥75), congestive heart failure (1 point), dysarthria (1 point), and dysphagia (4 point) (36).

Machine learning models

ML models were constructed based on structured variables, features extracted from the text, or a combination of both (Supplementary Figure 7). For comparison of classifier performance, simple logistic regression was used as the baseline. Because the performance of ML classifiers can be affected by class imbalance, we experimented with both oversampling and under-sampling methods to maintain the ratio of majority and minority classes as 1:1, 2:1, or 3:1 (37). The random forest (RF) algorithm was used to build classifiers. RF is a classifier ensemble method that consists of a set of decision tree classifiers. During the learning process, RF iteratively adopts the bootstrap

aggregating method to select samples and randomly selects a subset of predictors. In each iteration, each set of bootstrap samples with a subset of predictors is used to generate a decision tree. In the end, the algorithm outputs a whole forest of decision trees, which can be used for prediction by a majority vote of the trees.

During the training process (Supplementary Figure 7), we first experimented with different combinations of text vectorization techniques and resampling methods without hyperparameter tuning. We repeated 10-fold cross-validation 10 times to estimate the performance of classifiers. The best combination of text vectorization and resampling methods was determined based on the area under the receiver operating characteristic curve (AUC). Next, for each text vectorization technique with its corresponding best resampling method, we trained classifiers with hyperparameter tuning using 10 times of 10-fold cross-validation to determine the best number of decision trees in the random forest. Then we trained the final ML models from the whole training set using the best hyperparameter. The generated ML models were tested on the holdout test set. Shapley additive explanations (38) was used to interpret the model output. The experiments were carried out by using scikit-learn, imbalanced-learn, gensim, transformers, sentence-transformers, and SHAP libraries within Python 3.7 environment.

Statistical analysis

Categorical variables were presented with counts and percentages. Continuous variables were reported as medians and interquartile ranges. Differences between groups were tested by Chi-square tests for categorical variables and Mann-Whitney *U* tests for continuous variables.

Because accuracy may not be appropriate for model evaluation under imbalanced scenarios (39), the AUC was chosen as the primary evaluation metric for comparing the performance of prediction models on the holdout test set. The AUC for SAP risk scores was calculated using the receiver operating characteristic (ROC) analysis to determine the ability of each risk score to predict SAP. The method for ROC analysis was detailed in the Supplementary Methods in the Supplementary material. AUCs were calculated and compared using DeLong's method (40). The AUC ranges from 0 to 1, with 0.5 indicating random guess and 1 indicating perfect model discrimination. A model with an AUC value above 0.7 is considered acceptable for clinical use (41). The point closest to the upper left corner of the ROC curve (42), which represents the optimal trade-off between sensitivity and specificity, was considered the cut-off value for each SAP score. Then each SAP score was transformed into a binary variable for calculating accuracy, precision (positive predictive value), recall (sensitivity), and F1 score. Model calibration was evaluated by

the Hosmer-Lemeshow test and visualized by the calibration plot (43), which depicts the observed risk vs. the predicted risk.

All statistical analyses were performed using Stata 15.1 (StataCorp, College Station, Texas) and R version 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria). Two-tailed *P* values of 0.05 were considered significant.

Results

Characteristics of the study population

The study population consisted of 5,913 patients including 4,947 (83.7%) with AIS and 966 (16.3%) with ICH. A total of 450 (7.6%) patients developed SAP. Table 2 lists their baseline characteristics. Patients with SAP were older, more likely to be male, and more likely to have atrial fibrillation, congestive heart failure, pre-stroke dependency, dysarthria, and dysphagia, but less likely to have hyperlipidemia. They had a higher pre-stroke mRS, NIHSS, and white blood cell (WBC) count as well as a lower consciousness level than those without SAP. The training set consisted of 4,434 patients and the remaining 1,479 patients comprised the holdout test set (Supplementary Table 3).

Construction of ML models

Supplementary Figure 8 shows the estimates of AUC obtained from 10 times of 10-fold cross-validation in the training set. In general, the RF algorithm outperformed logistic regression when structured variables or both structured and textual features were used to build classifiers. By contrast, logistic regression models had higher AUCs than RF classifiers when only textual features were used. Resampling methods generally improved the performance of ML classifiers. Overall, RF classifiers based on both structured variables and textual features attained higher AUCs than the other classifiers. Text representation using the BOW approach performed better than that using the fastText embedding or BERT approach. The highest AUC was achieved by the ML model using the combination of text vectorization with BOW (binary representation) and 1:2 under-sampling of data.

Supplementary Table 4 shows the performance of ML models on the holdout test set and the number of decision trees used to build the RF classifiers. Supplementary Table 5 lists *P* values for pairwise comparisons of AUCs between these models. In general, ML models based on both structured and textual features achieved higher AUCs than those based on textual features alone. The ML model using the combination of text vectorization with BOW (binary representation) also had the highest AUC among all ML models. Therefore, it was chosen as the final model (ML Model A). For comparison with

conventional risk scores, the ML model based on structured variables alone (ML Model B) was also evaluated.

Comparison with conventional risk scores

By determining the point closest to the upper left corner of the ROC curve (42) the cut-off value for predicting SAP was 4.5 points for A²DS², 9.5 points for ISAN, 4.5 points for PNA, and 1.5 points for ACDD⁴, respectively. The cut-off value for ML models was set at the probability of 0.5. Accuracy, precision, recall, and F1 score were calculated based on these cut-off values. Table 3 lists the performance of ML models and conventional SAP risk scores on the holdout test set. Among all prediction models, ML Model A attained the highest AUC, accuracy, and F1 score. Figure 1 plots the ROC curves of the four SAP risk scores and two ML models. All the prediction models achieved an AUC value >0.7. ML Model A had the highest AUC [0.840, 95% confidence interval (CI) 0.806–0.875], which was significantly higher than those of ML Model B (0.828, 95% CI 0.793–0.863, *P* = 0.040), ACDD⁴ (0.807, 95% CI 0.766–0.849, *P* = 0.041), A²DS² (0.803, 95% CI 0.762–0.845, *P* = 0.013), ISAN (0.795, 95% CI 0.752–0.837, *P* = 0.009), and PNA (0.778, 95% CI 0.735–0.822, *P* < 0.001). Figure 2 shows the calibration plots and *P* values for the Hosmer-Lemeshow test for the prediction models. ML Model A was well-calibrated over the entire risk range with all points lying close to the 45-degree line (*P* = 0.579). All the other prediction models also demonstrated adequate calibration except for the A²DS² score (*P* = 0.023).

Influential features selected by ML models

Figure 3A shows the top 20 most influential features selected by ML Model A ordered by the mean absolute Shapley value, which indicates the global importance of each feature on the model output. Figure 3B presents the beeswarm plot depicting the Shapley value for every patient across these features, demonstrating each feature's contribution to the model output. According to the magnitude and direction of the Shapley value, higher values of NIHSS, WBC count, heart rate, blood glucose, international normalization ratio, and aspartate aminotransferase were associated with a higher risk of SAP, while lower values of Glasgow coma scale total score and its component (verbal, motor, and eye) scores, body mass index, platelet count, and triglyceride were associated with a higher risk of SAP. Male patients and those with dysphagia, dysarthria, or current smoking were more likely to have SAP. Among the textual features, the presence of “numbness”, “deny”, or “acute” in the HPI of the admission note was associated with a decreased

TABLE 2 Baseline characteristics of the study population.

Characteristic	Total (N = 5,913)	SAP (N = 450)	No SAP (N = 5,463)	P†
Age	70 (59–78)	72 (61–80)	69 (59–78)	<0.001
Male	3,643 (61.6)	308 (68.4)	3,335 (61.0)	0.002
Hypertension	4,739 (80.2)	361 (80.2)	4,378 (80.1)	0.966
Diabetes	2,422 (41.0)	188 (41.8)	2,234 (40.9)	0.714
Hyperlipidemia	3,167 (53.6)	187 (41.6)	2,980 (54.6)	<0.001
AF	822 (13.9)	106 (23.6)	716 (13.1)	<0.001
CHF	226 (3.8)	30 (6.7)	196 (3.6)	0.001
COPD	397 (6.7)	34 (7.6)	363 (6.6)	0.458
Smoking	2,431 (41.1)	202 (44.9)	2,229 (40.8)	0.090
Pre-stroke dependency	562 (9.5)	80 (17.8)	482 (8.8)	<0.001
Pre-stroke mRS	0 (0–0)	0 (0–1)	0 (0–0)	<0.001
NIHSS	5 (3–11)	17 (9–27)	5 (3–10)	<0.001
GCS	15 (14–15)	13 (8–15)	15 (15–15)	<0.001
Dysphagia	1,195 (20.2)	282 (62.7)	913 (16.7)	<0.001
Dysarthria	3,039 (51.4)	338 (75.1)	2,701 (49.4)	<0.001
Glucose (mmol/L)	7.38 (6.11–9.99)	7.77 (6.27–10.43)	7.33 (6.11–9.96)	0.030
WBC (10 ⁹ /L)	7.68 (6.19–9.61)	8.49 (6.63–10.96)	7.63 (6.16–9.47)	<0.001
A ² DS ²	4 (1–5)	6 (4–6)	3 (1–5)	<0.001
ISAN	7 (4–10)	11 (8–14)	7 (4–9)	<0.001
PNA	4 (1–5)	5 (4–6)	4 (1–5)	<0.001
ACDD ⁴	1 (0–2)	5 (2–5)	1 (0–2)	<0.001

†P values are comparisons between patients with SAP and those without SAP for each variable.

Data are given as n (%) and median (interquartile range).

AF, atrial fibrillation; CHF, congestive heart failure; COPD, chronic obstructive pulmonary disease; GCS, Glasgow coma scale; mRS, modified Rankin Scale; NIHSS, National Institutes of Health Stroke Scale; SAP, stroke-associated pneumonia; WBC, white blood cells.

TABLE 3 Performance of prediction models for predicting SAP.

Model	AUC (95% CI)	Accuracy	Precision	Recall	F1 score
ML model A	0.840 (0.806–0.875)	83.2%	0.254	0.634	0.363
ML model B	0.828 (0.793–0.863)	76.3%	0.212	0.786	0.334
A ² DS ²	0.803 (0.762–0.845)	75.1%	0.197	0.741	0.311
ISAN	0.795 (0.752–0.837)	76.9%	0.202	0.696	0.313
PNA	0.778 (0.735–0.822)	75.9%	0.189	0.661	0.294
ACDD ⁴	0.807 (0.766–0.849)	73.5%	0.193	0.786	0.310

AUC, area under the receiver operating characteristic curve; CI, confidence interval; ML, machine learning; SAP, stroke-associated pneumonia.

risk of SAP. The top 20 most influential features selected by ML Model B are shown in [Supplementary Figure 9](#) for reference.

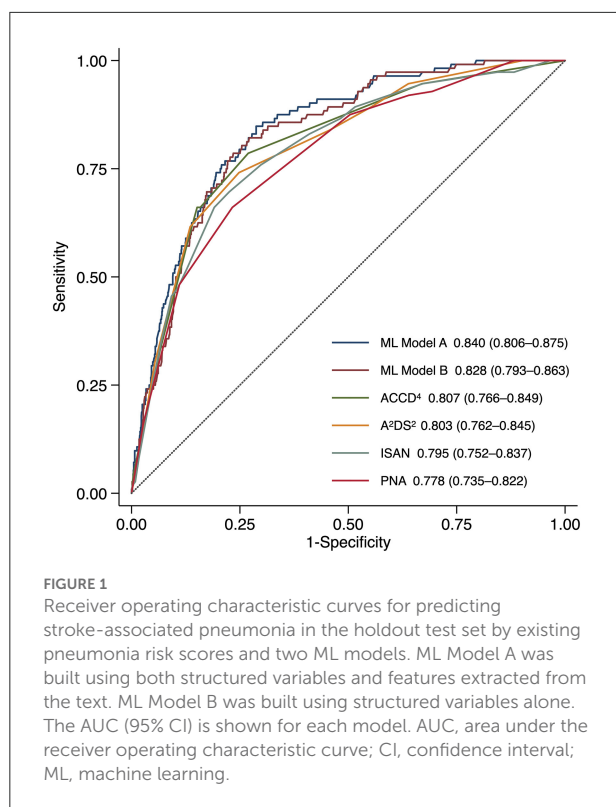
Discussion

In this exploratory study, the predictive performance of ML models was nominally higher than those using conventional SAP risk scores in terms of discrimination. Notably, the ML model built on both structured and unstructured textual data performed significantly better than the ML model built on structured data alone as well as all the conventional risk scores.

Besides, we discovered several influential features or predictors of SAP using Shapley values. These predictors might help early stratification of stroke patients who are more likely to develop SAP.

Predictors of SAP

Among the top 20 influential predictors selected by the ML model, NIHSS score, Glasgow coma scale score, dysphagia, dysarthria, current smoking, male sex, WBC count, and blood



glucose were known predictors of SAP, which have been included in conventional SAP risk scores (11, 33–36). A higher value of international normalized ratio in the context of stroke generally denotes the use of vitamin K antagonist and preexisting atrial fibrillation, which is also a known risk factor for SAP (11, 33). Interestingly, the ML model identified additional predictors, such as lower values of body mass index, platelet count, and triglyceride as well as higher values of heart rate and aspartate aminotransferase. Previous studies have found significantly lower body mass index, platelet count, and triglyceride as well as higher aspartate aminotransferase in stroke patients with SAP than those without (16, 44, 45). All these factors indicate poorer nutritional status, which may have a role in the development of SAP (45). Higher heart rate at rest was associated with poorer functional status in the elderly and predicted subsequent functional decline independently of cardiovascular risk factors (46). Higher initial in-hospital heart rate also predicted poorer stroke outcomes (47). The potential influence of these additional predictors on the development of SAP may warrant further research. We speculate that these factors are missing in conventional SAP risk scores either because logistic regression models cannot handle complex interactions and non-linear relationships among variables, or simply because they were not expected to be predictors of SAP and thus not investigated in previous studies.

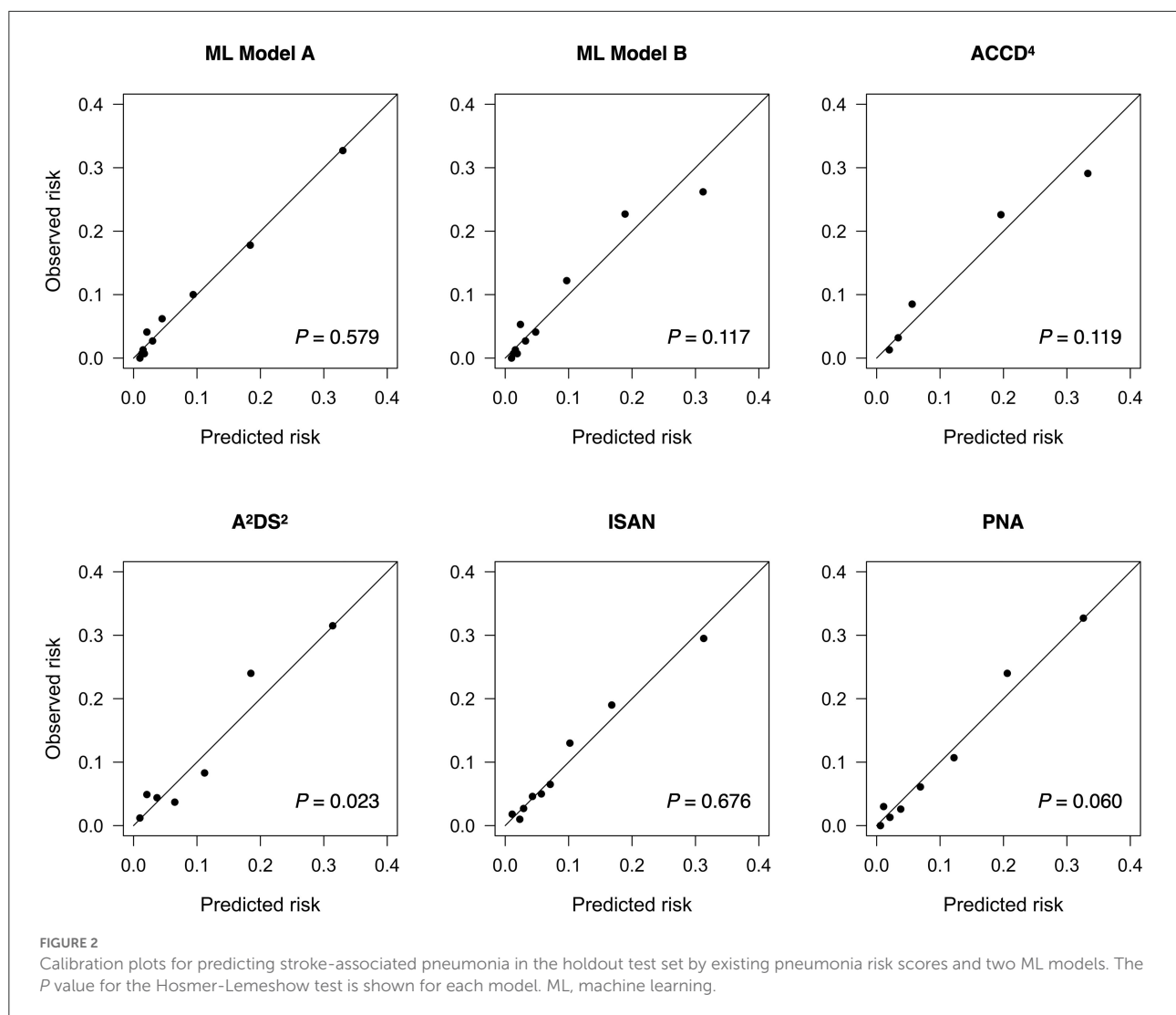
Hidden information from clinical text

The key finding of the present study was that the information extracted from unstructured clinical text could improve the prediction of SAP. However, the reason why the identified textual features (words) were associated with the risk of SAP may not be readily discernible unless these words and their context are examined simultaneously. For example, stroke patients who complain of “numbness” are generally fully conscious and may suffer a pure sensory stroke or sensorimotor stroke due to a small ischemic lesion (48, 49), which carries a low risk of pneumonia. Likewise, patients who can provide a history of their illness and “deny” the presence of certain symptoms are likely to have clear consciousness and may have mild neurological impairment. Furthermore, the mode of symptom onset can influence the pre-hospital delay of stroke patients (50). Patients experiencing “acute” symptoms are generally admitted to the stroke unit earlier while stroke unit care is associated with a lower frequency of SAP (4). These findings demonstrate that useful and informative predictors could be uncovered from unstructured clinical text through natural language processing and ML without human curation.

Clinical significance and implications

SAP has traditionally been attributed to aspiration secondary to dysphagia, impaired cough reflex, or reduced level of consciousness (3). Nonetheless, up to 40% of SAP may be unrelated to aspiration (8). Other causes such as bacteremia due to dysfunction of the gut immune barrier (51) and stroke-induced immune suppression (3, 52) may also contribute to the development of SAP. So far there is no sufficient evidence from clinical trials to demonstrate the effect of dysphagia screening protocols on the prevention of SAP (53). Meta-analyses of randomized trials have also failed to support the use of preventive antibiotic therapy to decrease the risk of SAP in acute stroke patients (54, 55). Furthermore, only weak evidence exists about whether intensified oral hygiene care reduces the risk of SAP (56, 57). Therefore, it is still a major challenge to find new therapeutic approaches to prevent SAP.

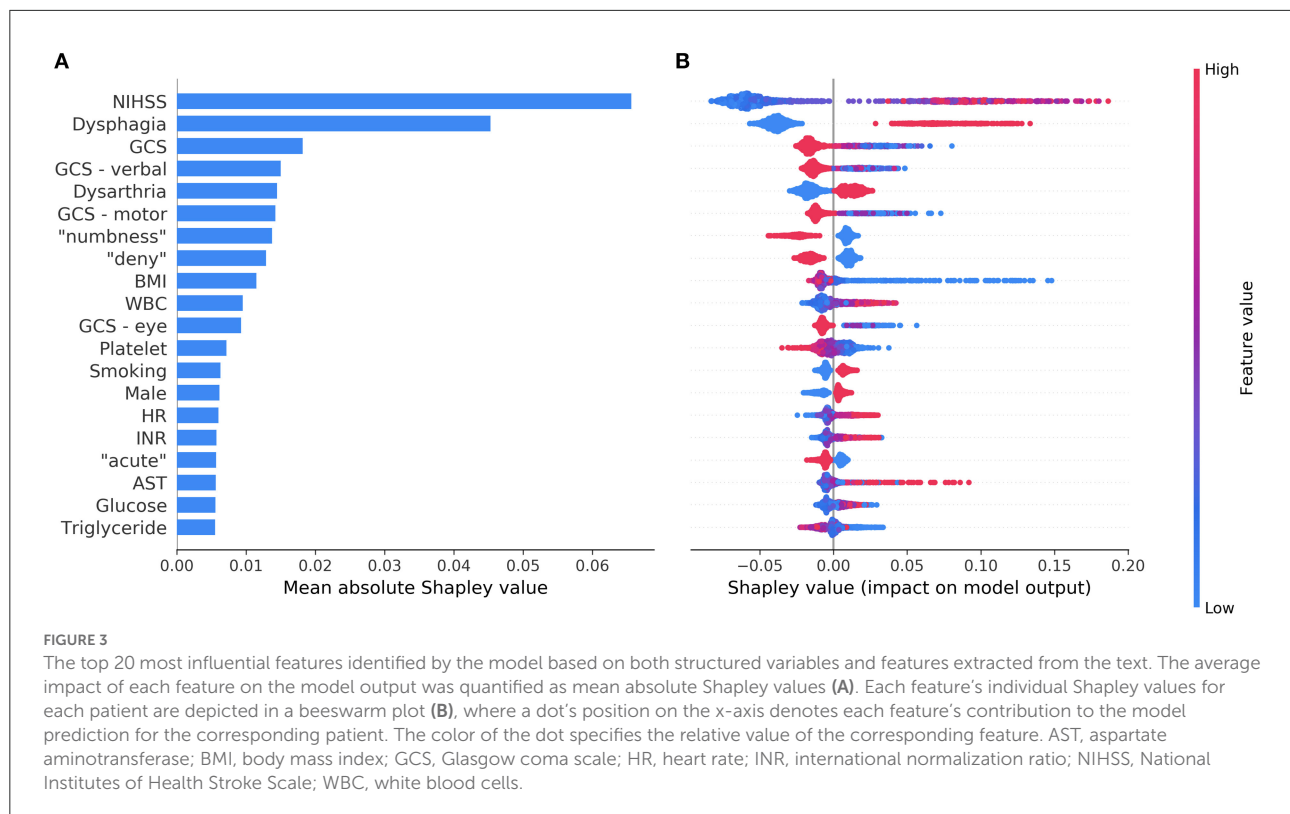
Despite this, adequate stratification of SAP risk is not without value. First, a good understanding of the risk of this serious complication of stroke will improve communication between physicians, patients, and caregivers. Second, the identification of at-risk patient groups allows recruiting suitable patients into clinical trials to test preventive interventions for SAP. Up to two-thirds of SAP occurs in the first week, with a peak incidence on the third day after stroke onset (10). Therefore, early stratification of SAP risk is beneficial in both clinical practice and research settings. The ML model developed in this study, which was based on information available within 24 h of admission, is well-suited for use in this context.



Limitations

This study has several limitations to be addressed. First, even though data-driven ML modeling has the potential to identify novel predictors, the predictor-outcome relationships discovered from data do not translate into a causal relationship (58). Second, we only extracted textual information from the HPI section of the admission note and did not investigate other clinical notes such as nursing notes and image reports. Further studies may examine the usefulness of information extracted from different kinds of clinical notes. Third, this study used oversampling and under-sampling techniques to solve the problem of data imbalance. Other data preprocessing approaches, such as synthetic minority oversampling technique

or its variants (37), can be explored in future studies. Fourth, several criteria exist to determine the most appropriate cut-off value for tests with continuous outcomes (42). The use of different criteria can result in different cut-off values for SAP risk scores, hence different results of accuracy, precision, recall, and F1 score. Fifth, high percentages of missingness for certain potential predictors, such as glycosylated hemoglobin, might prevent the ML algorithm from identifying their significance. Finally, this is a single-site study, and the generalizability of the study findings is limited. For example, the vocabulary and terms used for clinical documentation may differ across healthcare settings. Nevertheless, the procedure of model development can be replicated in individual hospitals to generate customized versions of SAP prediction models.



Conclusions

We demonstrated that it is feasible to build ML models to predict SAP based on both structured and unstructured textual data. Using natural language processing, pertinent information extracted from clinical text can be applied to improve the performance of SAP prediction models. In addition, ML algorithms identified several novel predictors of SAP. The workflow used to generate these models can be disseminated for local adaptation by individual healthcare organizations.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions. The data used in this study cannot be made available because of restrictions regarding the use of EHR data. Requests to access these datasets should be directed to S-FS, sfusng@cych.org.tw.

Ethics statement

The studies involving human participants were reviewed and approved by the Ditmanson Medical Foundation Chia-Yi

Christian Hospital Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

Study concept and design: H-CT and S-FS. Acquisition of data and study supervision: S-FS. Drafting of the manuscript: H-CT and C-YH. All authors analysis and interpretation of data, critical revision of the manuscript for important intellectual content, and had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This research was supported in part by the Ditmanson Medical Foundation Chia-Yi Christian Hospital-National Chung Cheng University Joint Research Program [grant number CYCH-CCU-2022-14]. The funder of the research had no role in the design and conduct of the

study, interpretation of the data, or decision to submit for publication.

Acknowledgments

The authors thank the help from the Clinical Data Center, Ditmanson Medical Foundation Chia-Yi Christian Hospital for providing administrative and technical support. This study is based in part on data from the Ditmanson Research Database (DRD) provided by Ditmanson Medical Foundation Chia-Yi Christian Hospital. The interpretation and conclusions contained herein do not represent the position of Ditmanson Medical Foundation Chia-Yi Christian Hospital. The authors also thank Ms. Li-Ying Sung for English language editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships.

References

1. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol.* (2021) 20:795–820. doi: 10.1016/S1474-4422(21)00252-0
2. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World stroke organization (WSO): global stroke fact sheet 2022. *Int J Stroke.* (2021) 17:18–29. doi: 10.1177/17474930211065917
3. Elkind MSV, Boehme AK, Smith CJ, Meisel A, Buckwalter MS. Infection as a stroke risk factor and determinant of outcome after stroke. *Stroke.* (2020) 51:3156–68. doi: 10.1161/STROKEAHA.120.030429
4. Badve MS, Zhou Z, van de Beek D, Anderson CS, Hackett ML. Frequency of post-stroke pneumonia: systematic review and meta-analysis of observational studies. *Int J Stroke.* (2018) 14:125–36. doi: 10.1177/1747493018806196
5. Westendorp WF, Nederkoorn PJ, Vermeij J-D, Dijkgraaf MG, van de Beek D. Post-stroke infection: a systematic review and meta-analysis. *BMC Neurol.* (2011) 11:110. doi: 10.1186/1471-2377-11-110
6. Hong KS, Kang DW, Koo JS, Yu KH, Han MK, Cho YJ, et al. Impact of neurological and medical complications on 3-month outcomes in acute ischaemic stroke. *Eur J Neurol.* (2008) 15:1324–31. doi: 10.1111/j.1468-1331.2008.02310.x
7. Vermeij FH, Scholte op Reimer WJ, de Man P, van Oostenbrugge RJ, Franke CL, de Jong G, et al. Stroke-associated infection is an independent risk factor for poor outcome after acute ischemic stroke: data from the Netherlands stroke survey. *Cerebrovasc Dis.* (2009) 27:465–71. doi: 10.1159/000210093
8. Teh WH, Smith CJ, Barlas RS, Wood AD, Bettencourt-Silva JH, Clark AB, et al. Impact of stroke-associated pneumonia on mortality, length of hospitalization, and functional outcome. *Acta Neurol Scand.* (2018) 138:293–300. doi: 10.1111/ane.12956
9. Katzan IL, Dawson NV, Thomas CL, Votruba ME, Cebul RD. The cost of pneumonia after acute stroke. *Neurology.* (2007) 68:1938–43. doi: 10.1212/01.wnl.0000263187.08969.45
10. de Jonge JC, van de Beek D, Lyden P, Brady MC, Bath PM, van der Worp HB, et al. Temporal profile of pneumonia after stroke. *Stroke.* (2022) 53:53–60. doi: 10.1161/STROKEAHA.120.032787
11. Kishore AK, Vail A, Bray BD, Chamorro A, Napoli MD, Kalra L, et al. Clinical risk scores for predicting stroke-associated pneumonia: a systematic review. *Eur Stroke J.* (2016) 1:76–84. doi: 10.1177/2396987316651759

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.1009164/full#supplementary-material>

12. Ni J, Shou W, Wu X, Sun J. Prediction of stroke-associated pneumonia by the A2DS2, AIS-APS, and ISAN scores: a systematic review and meta-analysis. *Expert Rev Res Med.* (2021) 15:1–12. doi: 10.1080/17476348.2021.1923482
13. Zapata-Arriaza E, Moniche F, Blanca P-G, Bustamante A, Escudero-Martínez I, Uclés O, et al. External validation of the ISAN, A2DS2, and AIS-APS scores for predicting stroke-associated pneumonia. *J Stroke Cerebrovasc Dis.* (2018) 27:673–6. doi: 10.1016/j.jstrokecerebrovasdis.2017.09.059
14. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA.* (2018) 319:1317. doi: 10.1001/jama.2017.18391
15. Ge Y, Wang Q, Wang L, Wu H, Peng C, Wang J, et al. Predicting post-stroke pneumonia using deep neural network approaches. *Int J Med Inform.* (2019) 132:103986. doi: 10.1016/j.ijmedinf.2019.103986
16. Li X, Wu M, Sun C, Zhao Z, Wang F, Zheng X, et al. Using machine learning to predict stroke-associated pneumonia in Chinese acute ischaemic stroke patients. *Eur J Neurol.* (2020) 27:1656–63. doi: 10.1111/ene.14295
17. Ruiz VM, Goldsmith MP, Shi L, Simpao AF, Gálvez JA, Naim MY, et al. Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records. *J Thorac Cardiovasc Surg.* (2022) 164:211–22.e3. doi: 10.1016/j.jtcvs.2021.10.060
18. Sung S-F, Hsieh C-Y, Hu Y-H. Early prediction of functional outcomes after acute ischemic stroke using unstructured clinical text: retrospective cohort study. *JMIR Med Inform.* (2022) 10:e29806. doi: 10.2196/29806
19. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open.* (2021) 4:00ab011. doi: 10.1093/jamiaopen/00ab011
20. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Crit Care Med.* (2018) 46:1125–32. doi: 10.1097/CCM.00000000000003148
21. Sung S, Chen C, Pan R, Hu Y, Jeng J. Natural language processing enhances prediction of functional outcome after acute ischemic stroke. *J Am Heart Assoc.* (2021) 10:e023486. doi: 10.1161/JAHA.121.023486
22. Hsieh F-I, Lien L-M, Chen S-T, Bai C-H, Sun M-C, Tseng H-P, et al. Get with the guidelines-stroke performance indicators: surveillance of stroke care in the taiwan stroke registry. *Circulation.* (2010) 122:1116–23. doi: 10.1161/CIRCULATIONAHA.110.936526

23. Smith CJ, Kishore AK, Vail A, Chamorro A, Garau J, Hopkins SJ, et al. Diagnosis of stroke-associated pneumonia. *Stroke*. (2015) 46:2335–40. doi: 10.1161/STROKEAHA.115.009617
24. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Transact Assoc Comput Linguis*. (2017) 5:135–46. doi: 10.1162/tacl_a_00051
25. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Minneapolis USA: Curran Associates, Inc. (2019). p. 4171–86
26. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE*. (2013) 8:e73791. doi: 10.1371/journal.pone.0073791
27. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khawaja K, et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl*. (2019) 116:494–520. doi: 10.1016/j.eswa.2018.09.034
28. Deng X, Li Y, Weng J, Zhang J. Feature selection for text classification: a review. *Multimed Tools Appl*. (2018) 78:3797–816. doi: 10.1007/s11042-018-6083-5
29. Culpeper J. Keyness: words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *Int J Corpus Linguis*. (2009) 14:29–59. doi: 10.1075/ijcl.14.1.03cul
30. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data*. (2019) 6:52. doi: 10.1038/s41597-019-0055-0
31. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. 2019 *IEEE Int Conf Healthc Informatics ICHI*. (2019) 00:1–5. doi: 10.1109/ICHI.2019.8904728
32. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. *Proc 2nd Clin Nat Lang Process Work*. (2019) pp. 72–78. doi: 10.18653/v1/W19-1909
33. Hoffmann S, Malzahn U, Harms H, Koennecke H-C, Berger K, Kalic M, et al. Development of a clinical score (A2DS2) to predict pneumonia in acute ischemic stroke. *Stroke*. (2012) 43:2617–23. doi: 10.1161/STROKEAHA.112.653055
34. Smith CJ, Bray BD, Hoffman A, Meisel A, Heuschmann PU, Wolfe CDA, et al. Can a novel clinical risk score improve pneumonia prediction in acute stroke care? A UK multicenter cohort study. *J Am Heart Assoc*. (2015) 4:e001307. doi: 10.1161/JAHA.114.001307
35. Friedant AJ, Gouse BM, Boehme AK, Siegler JE, Albright KC, Monlezun DJ, et al. A simple prediction score for developing a hospital-acquired infection after acute ischemic stroke. *J Stroke Cerebrovasc Dis*. (2015) 24:680–6. doi: 10.1016/j.jstrokecerebrovasdis.2014.11.014
36. Kumar S, Marchina S, Massaro J, Feng W, Lahoti S, Selim M, et al. ACDD4 score: a simple tool for assessing risk of pneumonia after stroke. *J Neurol Sci*. (2017) 372:399–402. doi: 10.1016/j.jns.2016.10.050
37. Branco P, Torgo L, Ribeiro RP, A. survey of predictive modeling on imbalanced domains. *ACM Comput Surv (CSUR)*. (2016) 49:1–50. doi: 10.1145/2907070
38. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9
39. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: review of methods and applications. *Expert Syst Appl*. (2017) 73:220–39. doi: 10.1016/j.eswa.2016.12.035
40. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. (1988) 44:837–45. doi: 10.2307/2531595
41. LaValley MP. Logistic regression. *Circulation*. (2008) 117:2395–9. doi: 10.1161/CIRCULATIONAHA.106.682658
42. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Medica*. (2016) 26:297–307. doi: 10.11613/BM.2016.034
43. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. (2010) 21:128–38. doi: 10.1097/EDE.0b013e3181c30fb2
44. Li W, He C. Association of platelet-to-lymphocyte ratio with stroke-associated pneumonia in acute ischemic stroke. *J Healthc Eng*. (2022) 2022:1033332. doi: 10.1155/2022/1033332
45. Quesada AS, Aliaga AA, Julia, Saumell JB, Galano MEH. Relationship between indicators of nutritional status and the development of pneumonia associated with ischemic stroke. *Finlay*. (2020) 10:231–9.
46. Ogliari G, Mahinrad S, Stott DJ, Jukema JW, Mooijart SP, Macfarlane PW, et al. Resting heart rate, heart rate variability and functional decline in old age. *CMAJ*. (2015) 187:E442–9. doi: 10.1503/cmaj.150462
47. Kuo Y-W, Lee M, Huang Y-C, Lee J-D. Initial in-hospital heart rate is associated with three-month functional outcomes after acute ischemic stroke. *BMC Neurol*. (2021) 21:222. doi: 10.1186/s12883-021-02252-2
48. Staaf G, Samuelsson M, Lindgren A, Norrving B. Sensorimotor stroke: clinical features, MRI findings, and cardiac and vascular concomitants in 32 patients. *Acta Neurol Scand*. (1998) 97:93–8. doi: 10.1111/j.1600-0404.1998.tb00616.x
49. Arboix A, García-Plata C, García-Eroles L, Massons J, Comes E, Oliveres M, et al. Clinical study of 99 patients with pure sensory stroke. *J Neurol*. (2005) 252:156–62. doi: 10.1007/s00415-005-0622-5
50. Derex L, Adeleine P, Nighoghossian N, Honnorat J, Trouillas P. Factors influencing early admission in a french stroke unit. *Stroke*. (2002) 33:153–9. doi: 10.1161/hs0102.100533
51. Stanley D, Mason LJ, Mackin KE, Srikhanta YN, Lyras D, Prakash MD, et al. Translocation and dissemination of commensal bacteria in post-stroke infection. *Nat Med*. (2016) 22:1277–84. doi: 10.1038/nm.4194
52. Shi K, Wood K, Shi F-D, Wang X, Liu Q. Stroke-induced immunosuppression and poststroke infection. *Stroke Vasc Neurol*. (2018) 3:34–41. doi: 10.1136/svn-2017-000123
53. Smith EE, Kent DM, Bulsara KR, Leung LY, Lichtman JH, Reeves MJ, et al. Effect of dysphagia screening strategies on clinical outcomes after stroke. *Stroke*. (2018) 49:e123–8. doi: 10.1161/STR.0000000000000159
54. Vermeij J, Westendorp WF, Dippel DW, van de Beek D, Nederkoorn PJ. Antibiotic therapy for preventing infections in people with acute stroke. *Cochrane Database Syst Rev*. (2018) 2018:CD008530. doi: 10.1002/14651858.CD008530.pub3
55. Westendorp WF, Vermeij J-D, Smith CJ, Kishore AK, Hodsoll J, Kalra L, et al. Preventive antibiotic therapy in acute stroke patients: a systematic review and meta-analysis of individual patient data of randomized controlled trials. *Eur Stroke J*. (2021) 6:385–94. doi: 10.1177/23969873211056445
56. Lyons M, Smith C, Boaden E, Brady MC, Brocklehurst P, Dickinson H, et al. Oral care after stroke: where are we now? *Eur Stroke J*. (2018) 3:347–54. doi: 10.1177/2396987318775206
57. Yuan D, Zhang J, Wang X, Chen S, Wang Y. Intensified oral hygiene care in stroke-associated pneumonia: a pilot single-blind randomized controlled trial. *Inquiry*. (2020) 57:0046958020968777. doi: 10.1177/0046958020968777
58. Li J, Liu L, Le TD, Liu J. Accurate data-driven prediction does not mean high reproducibility. *Nat Mach Intell*. (2020) 2:13–5. doi: 10.1038/s42256-019-0140-2



OPEN ACCESS

EDITED BY
Riccardo Nucera,
University of Messina, Italy

REVIEWED BY
Ayman Raouf Khalifa,
October 6 University, Egypt
Antonino Lo Giudice,
University of Catania, Italy

*CORRESPONDENCE
Monica Macri
m.macri@unich.it

SPECIALTY SECTION
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 15 April 2022
ACCEPTED 07 October 2022
PUBLISHED 01 November 2022

CITATION
Macri M and Festa F (2022)
Three-dimensional evaluation using
CBCT of the mandibular asymmetry
and the compensation mechanism in a
growing patient: A case report.
Front. Public Health 10:921413.
doi: 10.3389/fpubh.2022.921413

COPYRIGHT
© 2022 Macri and Festa. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Three-dimensional evaluation using CBCT of the mandibular asymmetry and the compensation mechanism in a growing patient: A case report

Monica Macri* and Felice Festa

Department of Innovative Technologies in Medicine and Dentistry, University "G. D'Annunzio" of Chieti-Pescara, Chieti, Italy

Background: This case report aims to evaluate the development and the compensation mechanisms of the mandibular asymmetry in a growing male patient using cone beam computed tomography (CBCT). In this case, the menton deviated on the right, a sporadic condition, which may be the consequence of a disorder in the mandibular growth.

Case presentation: The young male patient was treated with rapid palatal expander (RPE) and Fränkel functional regulator III (FR-3). The initial CBCT was acquired at the beginning of therapy when the patient was 8 years old, and the final CBCT was developed at the end of the treatment when the patient was 12 years old. The patient's CBCT was performed with the head oriented according to the Natural Head Position (NHP); the NHP is a physiological and reproducible posture defined for morphological analysis. The 3D image of the cranium was oriented in the Dolphin software according to NHP posture, and cephalometric measurements were taken in the software's frontal, laterolateral right and left, posteroanterior, and submentovertex views. The therapy lasted 3.8 years and ended with significant regression of the mandibular asymmetry from moderate grade (4.2 mm) to slight grade (1.3 mm).

Conclusion: The literature shows that the left hemi-mandible has grown more than the right side, which affirms that in case of deviation of the menton >4 mm, the bone volume increases on the non-deviated side.

KEYWORDS

facial asymmetry, dental midline deviation, CBCT, rapid palatal expander, Fränkel-III, orthopedic therapy, orthodontic therapy

Background

Facial asymmetry is the difference in shape, size, position, or function between the two sides of the face (1). In most cases, the asymmetry is not clinically detectable; it is also known as subclinical, minor, or normal facial asymmetry (2).

A dominant half-face is recognized in all subjects: in 80% of cases, it corresponds to the right side, with no differences in distribution according to sex and age (3). The dominance of the right side is explained by the migration of the cells of the cranial neural crest (NCC): migration begins earlier on the right side than on the left side, but it ends simultaneously on both sides; for this reason, there is an evident dominance on the right side of the face; consequently, the menton left shift (the most inferior point on mandibular symphysis) is more frequent than the right shift (4).

In addition, mandibular asymmetry is more frequent than maxillary asymmetry. The growth of the maxilla is more stable due to the connection with the cranial base synchondroses, and it is less vulnerable to environmental factors influence; differently, the mandible is the only mobile bone in the skull, and for this reason, it is highly prone to environmental impacts (5).

The right shift of the menton is a rare condition that may result from a disorder in the mandibular growth (i.e., facial trauma, TMJ ankylosis, bad habits, prone sleep position, premature tooth loss, and iatrogenic causes) (6).

The craniofacial growth can be compromised if a pathogenic noxa affects an evolutionary age, producing deformities and asymmetries in the head–neck district.

It is essential to detect dentofacial asymmetries in orthodontic practice: the dental midline is a reference landmark that must coincide with the center of the mouth (the imaginary line that joins the center of the philtrum with the center of the palatine raphe). The mandibular midline corresponds to the inferior interincisal line (7).

When a clinician observes a mandibular asymmetry in children, he has to think of a functional asymmetry, which must be corrected to prevent its transformation into a skeletal and joint asymmetry. Using the Frankel function regulator, it is possible to re-center the two arches and restore muscle function, breathing and vocalization. If a mandibular asymmetry is detected within 6 years of age, it can be fully recovered, preventing TMD (8) and joint problems in future adult patients (9).

Treating mandibular asymmetry as soon as it is detected is important, and it has practical results if treated during primary dentition. Frankel's function regulator type 3 is very effective, especially in treating third-class malocclusions, even if treated in early mixed dentition (10).

With its particular shape and design, the device promotes maxillary growth by retracting soft tissues that block it and

stimulating the periosteum, directing mandibular growth (11). The device consists of four resin shields: two on the anterior part and the other on the sides. The upper anterior shields eliminate the pressure of the upper lip on the underdeveloped jaw. The two vestibular shields act superiorly by stimulating the periosteum and relieving the pressure of the buccinator (12).

Controlled retrospective studies show that the craniofacial changes following the treatment with Frankel-III are stable. There is no significant inhibition of mandibular growth but the closure of the gonial angle. Intermaxillary and interdental changes are maintained and stable over time (13).

Some authors recommended that to be effective, long-term appliance wear (more than 5 years) is necessary to achieve clinically valuable results in FR-3 appliances (14).

The present case report describes the successful orthopedic and orthodontic treatments of an 8-year-old Caucasian patient with an anterior crossbite and severe mandibular deviation to the right side.

The orthopedic–orthodontic treatment lasted 3.8 years and was divided into two phases: the first phase with the RPE and the second phase with the FR-3. The patient was 8 years at the beginning of therapy and 12 years at the end. The CBCT scans were acquired at the treatment's beginning (T0) and the end (T1).

Case presentation

Diagnosis and etiology

An 8-year-old male patient visited the Orthodontic Department at G. D'Annunzio University in Chieti, Italy, with a chief complaint of anterior crossbite and mandibular asymmetry. No systemic pathologies or maxillofacial disorders were found in the medical history.

The facial evaluation showed a straight profile and a soft-tissue asymmetry of the lower face with a mandible shift to the right side. Intraorally, the dentition was mildly crowded in the upper arch, and a class III molar relationship was observed on the left and right sides. The mandibular dental midline was deviated 4 mm to the right, whereas the upper dental midline coincided with the facial midline.

The patient exhibited a normal overbite and an anterior crossbite with a -2.0 mm overjet.

The dental cast analysis at T0 revealed a maxillary transverse deficiency: the upper arch width was 2.5 mm narrower than the lower arch in the first molar region.

The cephalometric analysis at T0 reveals a class I skeletal profile (15) (ANB: $+0.9^\circ$), mesocephalic (16) (SN—GoGn: 30.1°), hypodivergent growth pattern (17) (FH—GoGn: 13.6°), and moderate right shift of the menton (4.2 mm) (18).

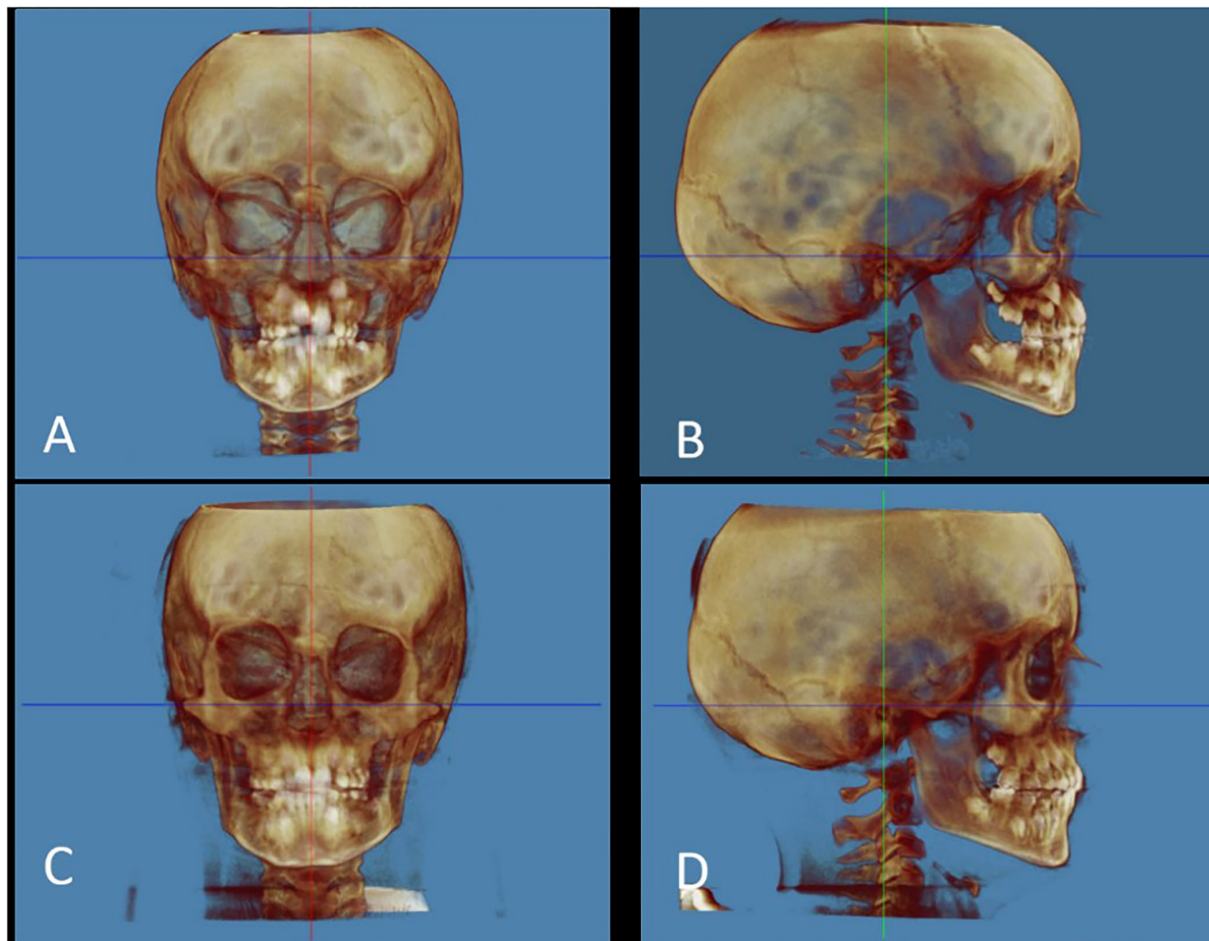


FIGURE 1

Natural head position. (A) Pre-treatment frontal view; (B) Pre-treatment lateral view (right); (C) Post-treatment frontal view; (D) Post-treatment lateral view (right). The red line corresponds to the sagittal plane. The green line corresponds to the coronal plane. The blue line corresponds to the transverse plane. The reference landmarks used for cephalometric measurements are shown in [Table 1](#).

Cone beam CT analysis

All CBCT examinations were taken at T0 and T1 and were performed by the Planmeca ProMax[®] 3D MID unit (Planmeca Oy, Helsinki, Finland) according to the low-dose protocol (19) with these parameters: large FOV, standard resolution quality images, 80 kVp, 5 Ma, and acquisition time of 15 s resulted in an effective dose of 35 microsieverts (μ Sv) (20).

The three-dimensional graphic rendering software used for the cephalometric measurements was Dolphin Imaging 11.95 Premium (Patterson Technology, Chatsworth, CA). The software processes the 3D-CT scan images in 2D-Digital Imaging and Communications in Medicine (DICOM) files.

The patient's CBCT was performed with the head oriented according to the NHP; the patient was in a sitting position with the back perpendicular to the floor as much as possible.

The head was stabilized with ear rods in the external auditory meatus. The patient was instructed to look into their eyes in a mirror 1.5 m away to obtain NHP. The NHP is a physiological and reproducible posture defined for the morphological analysis described in the orthodontic and anthropological literature (21).

The 3D image of the cranium was oriented in the Dolphin software according to NHP posture before taking cephalometric measurements.

The NHP orientation was carried out by the widgets present in Dolphin; hard and soft tissue views were checked for orientation in the software by visualizing the head from the front, right, and left sides. In the NHP, there are three reference planes ([Figure 1](#)), perpendicular to each other, which are identified on the software for the patient's cephalometric measurements.

TABLE 1 Reference cephalometric landmarks.

Landmark	Abbreviation	Description
Crista Galli	Cg	The most superior point of the crista Galli of the ethmoid bone
Basion	Ba	The median point on the anterior margin of the foramen magnum
Porion	Po	The highest point on the roof of the external auditory meatus
Orbitale	Or	The deepest point on the infraorbital margin
Condylion superius	Cdsup	The most superior point of the condyle head
Condylion medialis	Cdmed	The most medial point of the condyle head
Condylion lateralis	Cdlat	The most lateral point of the condyle head
Condylion posterius	Cdpost	The most posterior point of the condyle head
Sigmoid notch	S	The most inferior point of the sigmoid notch
Gonion lateralis	Golat	The most lateral point of the gonion area
Gonion posterius	Gopost	The most posterior point of the gonion area
Gonion inferius	Goinf	The most inferior point of the gonion area
Menton	Me	The most inferior point on the mandibular symphysis
First maxillary molar	6	Occlusal fossa of the maxillary first molar
Mandibular canine	3	Cuspal tip of the mandibular canine

1. The transverse plane coincides with the Frankfurt plane (FH), a plane passing through two points: Orbital (Or) and Porion (Po);
2. The sagittal plane coincides with the mid-sagittal plane (MSP), a plane perpendicular to the plane FH and passing through two points: crista galli (Cg) and basion (Ba);
3. The coronal plane coincides with the anteroposterior (PO) plane, perpendicular to the FH and MSP, passing through the right and left Porion.

The CBCT measurements (Table 2) were performed in frontal, laterolateral (LL) right, LL left, posteroanterior (PA), and submentovertex (SMV) views. Each measurement was performed on the initial and final CBCT. Also, the size of the right and left masseter muscles was evaluated with a widget

present in Dolphin. In the frontal view, the size of each muscle was measured by adjusting the translucency instrument to discriminate soft from hard tissues.

Treatment objectives

Based on the clinical and radiographic findings, the primary objectives of treatment were planned as follows: (1) correction of the dental and skeletal mandibular midlines, (2) correction of the dental class III malocclusion, (3) correction of the anterior crossbite, (4) making space on the maxillary dentition for guiding eruption and correction of the mild crowding, and (5) correction of the negative overjet.

Treatment alternatives

Option 1. The orthopedic–orthodontic treatment with RPE and FR-3 was proposed as the first-choice treatment based on the treatment objectives.

Option 2. The orthopedic treatment with a class III protrusion facemask was proposed as an alternative treatment.

Option 3. If orthopedic–orthodontic treatment (options 1 and 2) could not be performed, orthognathic surgery could be a choice after completing skeletal growth. However, option 3 was poorly recommended because of the surgical risks and costs of surgical intervention, whereas option one was highly recommended and chosen with the consent of the patient's parents.

Treatment progress

The orthopedic therapy was performed in two phases: the first phase with a rapid palatal expander (RPE) and the second phase with the Fränkel function regulator III (FR-3).

The first phase of the treatment uses the RPE, which provides a transverse expansion of the maxilla; the RPE was initially activated on the chair by performing a complete turn of the screw, which corresponds to four activations (1 mm). The patient was instructed to activate the RPE at home two times daily (0.5 mm expansion a day) for 10 days. The same RPE was used as a passive retainer to prevent transverse maxillary relapse for 6 months, and the screw was locked with a light-cure flow composite. The appliance was removed after 6 months after its last activation. The second phase with the FR-3 corrected skeletal deformities and prognathism. The therapeutic principle is based on eliminating all factors that could arrest maxillary development and, at the same time, prevent excessive mandibular growth (19).

TABLE 2 Cephalometric measurements.

	Landmarks	Pre-treatment	Post-treatment	Results
Frontal view (F)				
Menton deviation	Distance from Me to MSP	4.2 mm (moderate deviation)	1.3 mm (slight deviation)	Δ : −2.9 mm
Right masseter muscle	Maximum length and width	Length: 55.4 mm	Length: 61.5 mm	Δ : +6.1 mm
		Width: 15.7 mm	Width: 19.4 mm	Δ : +3.7 mm
Left masseter muscle	Maximum length and width	Length: 51.0 mm	Length: 54.3 mm	Δ : +3.3 mm
		Width: 11.3 mm	Width: 14.9 mm	Δ : +3.6 mm
Laterolateral view (LL)				
Vertical facial growth pattern	Angle from SN to GoGn	30.1° (mesofacial)	32.5° (mesofacial)	Δ : +2.4°
Frankfort-mandibular plane angle (FMA)	The angle from FH to GoGn	13.6° (hypodivergent)	16.8° (hypodivergent)	Δ : +3.2°
Sagittal facial growth pattern (ANB)	The angle from A to N to B	0.9° (class I)	2.5° (class I)	Δ : +1.6°
Right–left difference in lateral Ramal inclination	The angle from Cd post—Go post to FH	Right: 74.7°	Right: 77.1°	Δ : +2.4°
		Left: 73.4 mm	Left: 71.8°	Δ : −1.6°
Right–left difference in ramus length (without condyle and gonial angle)	Distance from Copost gopost	Right: 37.7 mm	Right: 38.9 mm	Δ : +1.2 mm
		Left: 33.8 mm	Left: 41.6 mm	Δ : +7.8 mm
Right–left difference in ramus length (with condyle and gonial angle)	Distance from Cdsup to Go inf	Right: 50.9 mm	Right: 55.8 mm	Δ : +4.9 mm
		Left: 48.9 mm	Left: 54.6 mm	Δ : +5.7 mm
Right–left difference in condylar height	Distance from Cdsup to S	Right: 18.3 mm	Right: 17.3 mm	Δ : −1.0 mm
		Left: 18.1 mm	Left: 20.2 mm	Δ : +2.1 mm
Postero-anterior view (PA)				
Right–left difference in maxillary height	6 to FH	Right: 29.0 mm	Right: 35.8 mm	Δ : +6.8 mm
		Left: 27.2 mm	Left: 37.0 mm	Δ : +9.8 mm
Right–left difference in frontal Ramal inclination	The angle from Cdlat-Golat to MSP	Right: 20.4°	Right: 14.9°	Δ : −5.5°
		Left: 16.5°	Left: 16.9°	Δ : +0.5°
Right–left difference in mandibular body height	Distance from 3 to GoGn	Permanent canines not erupted	Right: 53.1 mm	Not evaluabe
			Left: 33.3 mm	
Intercondilar distance	Distance from right Cdmed to left Cdmed	74.0 mm	83.3 mm	Δ : +9.3 mm
Extracondilar distance	Distance from right Cdlat to left Cdlat	102.7 mm	107.9 mm	Δ : +5.2 mm
Maximum width of the left condyle	Distance from Cdlat to Cdmed	15.0 mm	16.1 mm	Δ : +1.1 mm
Maximum width of the right condyle	Distance from Cdlat to Cdmed	15.1 mm	16.5 mm	Δ : +1.4 mm
Right–left difference in condyle—MSP distance	Distance from Cdlat to MSP	Right: 50.3 mm	Right: 51.7 mm	Δ : +1.4 mm
		Left: 52.3 mm	Left: 53.8 mm	Δ : +1.5 mm
Sub-mentovertex view (SMV)				
Right–left difference in mandibular body length	Me-Gopost,	Right: 76.4 mm	Right: 77.4 mm	Δ : +1 mm
		Left: 74.9 mm	Left: 82.4 mm	Δ : +7.5 mm

Δ Difference (post-treatment data – pre-treatment data), FH, Frankfort Horizontal plane; PO, anteroposterior reference plane; MSP, mid-sagittal reference plane; GoGn, mandibular plane.

Treatment results

The facial evaluation showed an improved soft-tissue symmetry in the lower face. Intraorally, ideal occlusion, proper overjet, and I molar relationship were achieved.

The dental cast analysis revealed the achieving of proper maxillary and mandibular intermolar widths and revealed a partial re-centring of the mandibular midline was achieved (2.9 mm to the left), as confirmed by CBCT; however, at the end of the therapy, the menton still deviated

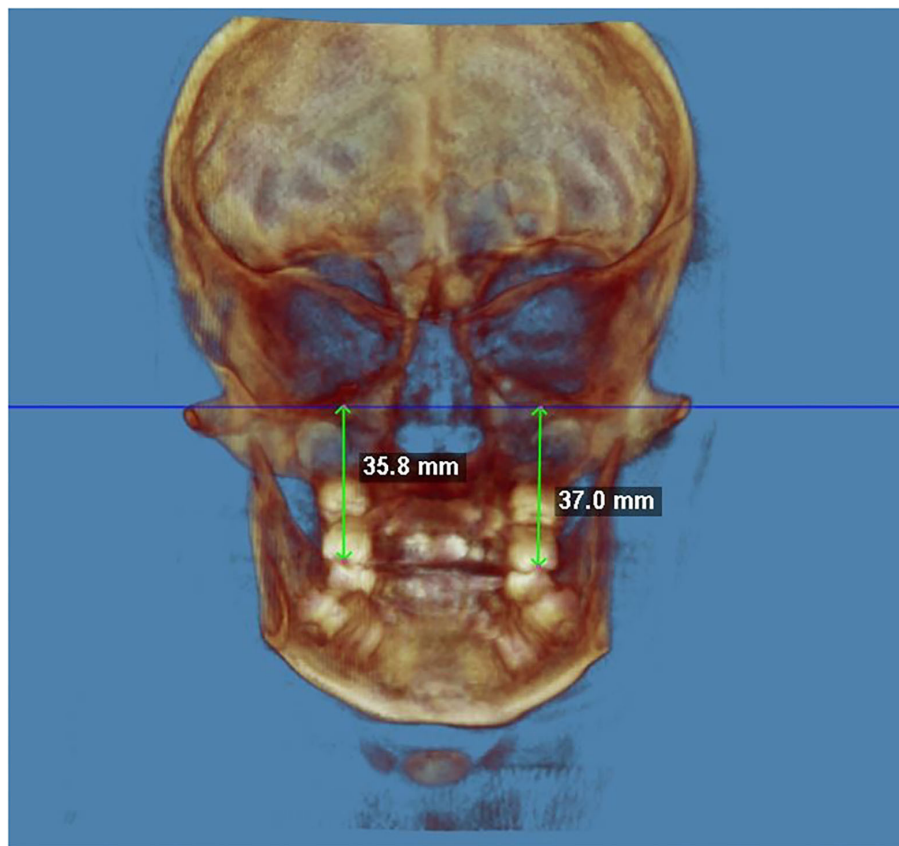


FIGURE 2

The right–left difference in maxillary height at the end of the treatment (PA view). The maxillary height was calculated from FH to the occlusal fossa of the maxillary first molar.

1.2 mm to the right (slight deviation) (22). The CBCT cephalometric analysis before and after the treatment is shown in Table 2.

As described in the literature (23), the menton point is the most inferior point on mandibular symphysis in the median plane. In this case report, the mandibular deviation was evaluated, calculating the deviation of the menton from the MSP. At T0, the menton deviation was 4.2 mm (moderate deviation) and after the treatment was 1.3 mm (slight deviation).

After the treatment, the menton point moved 2.9 mm toward the reference midline.

The cephalometric analysis of the masseter muscles (Figure 2) showed that both muscles developed similarly in thickness but not in length. The maximum length of the right masseter muscle was 55.4 mm at t0, and 61.5 mm at t1, with a difference of +6.1 mm. The maximum length of the left masseter muscle was 51.0 mm at t0, and 54.3 mm at t1, with a difference of +3.3 mm. The length of the right muscle has increased more than the left muscle, and this result positively affected the re-centring of the menton points toward the MSP. This

finding is significant because it was shown that if mandibular asymmetry is not corrected, the mandible may grow and develop asymmetrically due to lateral displacement and asymmetric muscle function.

On the laterolateral view (LL), the cephalometric analysis evaluated the vertical facial growth pattern, the Frankfort–mandibular plane angle (FMA), the Sagittal facial growth pattern (ANB), the right–left difference in lateral Ramal inclination; the right–left difference in ramus length (without condyle and gonial angle), the right–left difference in ramus length (with condyle and gonial angle), and the right–left difference in condylar height.

SN.GoGn and FMA were the most reliable indicators in assessing facial vertical growth patterns. An FMA of $25 \pm 4^\circ$ is within a normal range (hypodivergent $< 21^\circ$, hyperdivergent $> 29^\circ$). An SN.GoGn of $32 \pm 4^\circ$ is within a normal range (hypodivergent $< 28^\circ$, hyperdivergent $> 36^\circ$) (24).

The facial divergence was evaluated with the Sella–Nasion and Gonion–Gnathion angle (SN^{GoGn}); the SN^{GoGn} angle is an angular measurement that quantifies the inclination of the mandibular base concerning the cranial base. A SN^{GoGn} of 32

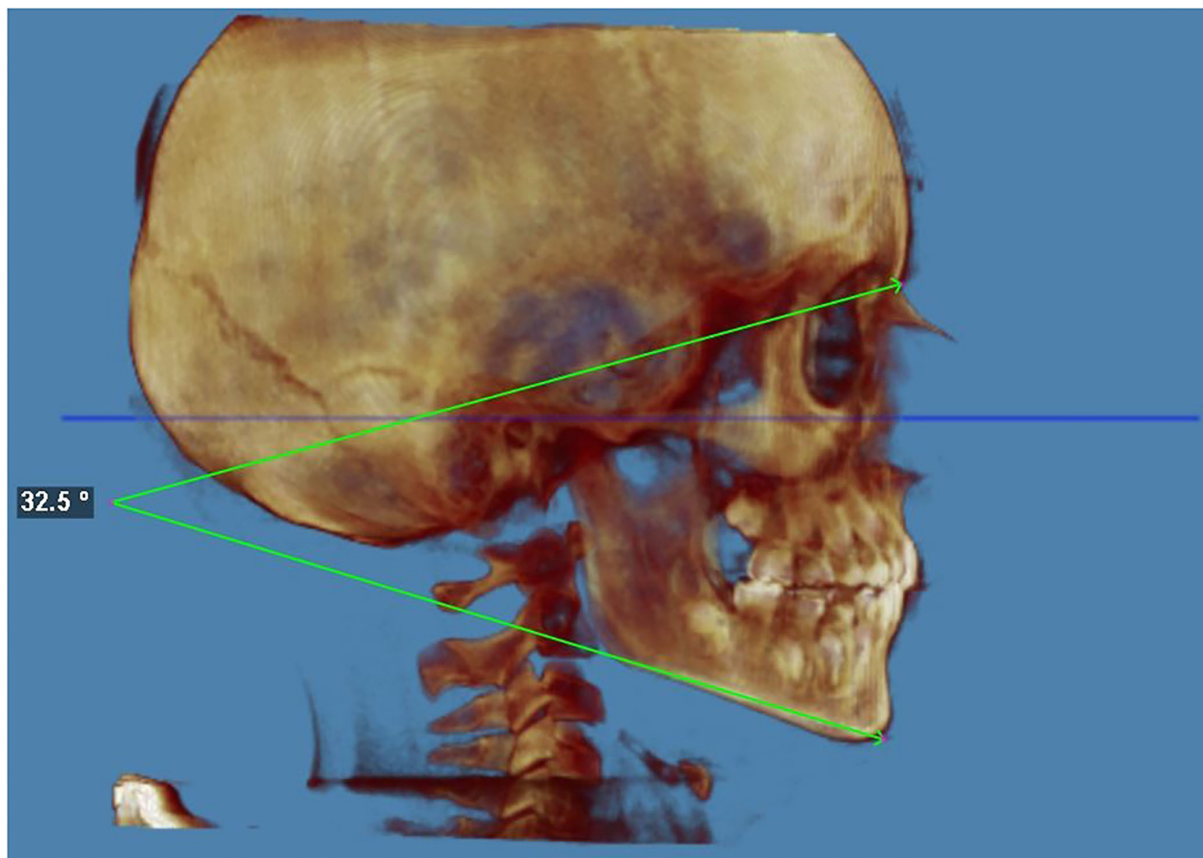


FIGURE 3

Facial vertical growth pattern at the end of the treatment. The facial divergence was evaluated with the Sella–Nasion and Gonion–Gnathion angle ($SN^G oGn$); the $SN^G oGn$ angle is an angular measurement that quantifies the inclination of the mandibular base about the cranial base. The angle from SN to GoGn was 32.5 (mesofacial) at the end of the treatment. An SN.GoGn of 32 ± 4 degrees is within normal range (hypodivergent $< 28^\circ$ and hyperdivergent $> 36^\circ$).

$\pm 4^\circ$ is within a normal range (brachyfacial $< 28^\circ$, dolichofacial $> 36^\circ$) (25, 26) found a decrease from 36° to 31° between 6 and 16 years of age.

The angle from SN to GoGn was 30.1° (mesofacial) at t0 and 32.5° (mesofacial) at t1, with a difference of $+2.4^\circ$ (Figure 3).

The Frankfort horizontal plane–gonion–gnathion angle ($FH^G oGn$) is formed by the intersection of the Frankfort horizontal plane (FH) and the mandibular plane (GoGn). A FMA of $25 \pm 5^\circ$ is within a normal range (hyperdivergent $> 30^\circ$, hypodivergent $< 20^\circ$).

The FMA was 13.6° (hypodivergent) at t0 and 16.8° (hypodivergent) at t1, with a difference of $+3.2^\circ$ (Figure 4). This result does not differ much from the $SN^G oGn$.

The subspinale–nasion–supramental angle (ANB) indicates the skeletal relationship between the maxilla (at the level of point A) and mandible (at the level of point B). The ANB angle (Figure 5) is commonly used to determine the sagittal facial

growth pattern in cephalometric analysis, and an ANB of $2 \pm 2^\circ$ is within a normal range (class II $> 4^\circ$, class III $< 0^\circ$).

The sagittal facial growth pattern (ANB) was 0.9° (class I) at T0 and 2.5° (class I) at T1, with a difference of $+1.6^\circ$.

The inclination of the mandibular ramus was calculated with the angle between Cd post–Go post and FH). The inclination of the right ramus has increased ($+2.4^\circ$); instead, the inclination of the left ramus has decreased (-1.6°). Also, this result positively affected the re-centring of the menton points toward the MSP.

The height of the mandibular ramus was calculated in different ways: the ramus length without condyle and gonial angle (distance from Copost to gopost), the ramus length with condyle and gonial angle (distance from Cdsup to Go inf), and the condylar height (distance from Cdsup to S). In each case, the right side was significantly higher than the left side at t0. At the end of the treatment, the right side was slightly higher than the left side, a sign of more growth on the left side.

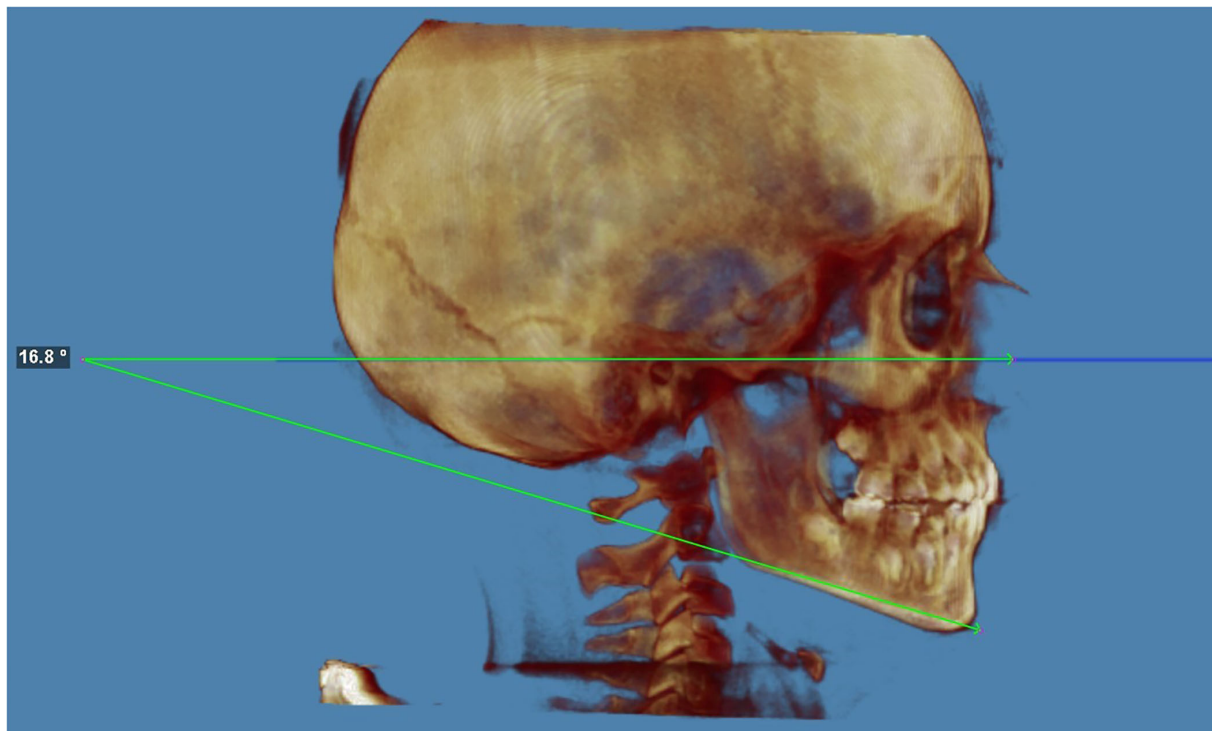


FIGURE 4

Frankfort-mandibular plane angle (FMA) at the end of the treatment. The FMA is the angle from FH to GoGn. The FMA was 16,8° (hypodivergent) at t1 with a difference of +3.2°. An FMA of $25 \pm 4^\circ$ is within normal range (hypodivergent $< 21^\circ$, hyperdivergent $> 29^\circ$).

The maxillary height was calculated from FH to the occlusal fossa of the maxillary first molar. The right hemimaxilla was slightly higher than the left hemimaxilla at t0. At the end of the treatment, the left hemimaxilla was marginally higher than the right hemimaxilla, a sign of more growth on the left side (+9.8 mm).

The frontal Ramal inclination was calculated with the angle between Cdlat-Golat to MSP. The inclination of the right ramus has decreased (-5.5°); instead, the inclination of the left ramus has increased ($+0.5^\circ$).

After the treatment, the inclination of the right mandibular ramus has changed more than the left one, as shown in LL and PA view, instead of the inclination of the left mandibular ramus, which has remained relatively unchanged. However, the height evaluation showed that the left ramus had grown more than the right ramus.

The height of the hemi-mandible was evaluated as the distance from the cuspal tip of the mandibular canine to GoGn. The height of the left hemi-mandible was shorter

than the right hemi-mandible after the treatment. The pre-treatment height was not evaluated as the canines did not erupt yet.

The intercondylar distance (from the right Cdmed to the left Cdmed) was 74.0 mm at t0 when the patient was 8 years old. After the treatment, when the patient was 12 years old, the intercondylar distance was 83.3 mm, increasing +by 9.3 mm.

The extracondylar distance was 102.7 mm at t0 and 107.9 mm at t1, increasing +by 5.2 mm.

On SMV view, the cephalometric analysis evaluated the length of the hemimandibular body.

The length of the hemimandibular body was calculated with the distance between the menton point and the gopost point. The right side was slightly longer than the left side at T0.

After the treatment, the length of the right side has slightly increased (+1.0 mm); instead, the length of the left side has significantly increased (+7.5 mm). The left side resulted longer than the right side at the end of the treatment. Also, this result positively affected the re-centring of the menton points toward the MSP.

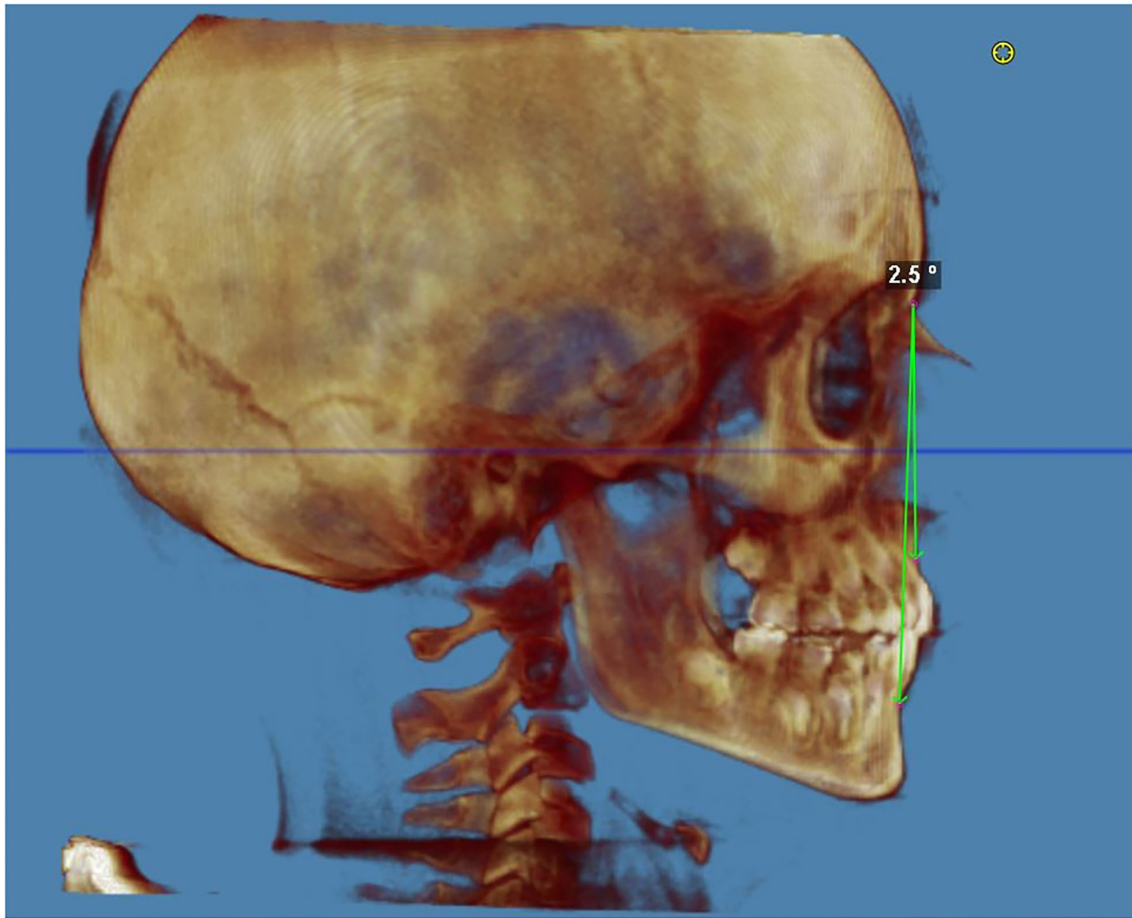


FIGURE 5

Sagittal facial growth pattern (ANB) at the end of the treatment (LL view). The subspinale–nasion–supramental angle (ANB) indicates the skeletal relationship between the maxilla (at the level of point A) and mandible (at the level of point B). The ANB angle is commonly used to determine the sagittal facial growth pattern in cephalometric analysis, and an ANB of $2 \pm 2^\circ$ is within the normal range (class II $> 4^\circ$, class III $< 0^\circ$). The sagittal facial growth pattern (ANB) was 2.5° (class I) at the end of the treatment.

Discussion

The purpose of this case report was to evaluate the development and the compensation mechanisms of the mandibular asymmetry in a growing male patient using cone beam computed tomography (CBCT) after treatment with RPE and FR-3 (21).

A low-dose CBCT protocol was used to identify landmarks better and reduce the patient's radiation exposure. The first phase of the treatment consists of using the RPE, which provides a transverse expansion of the maxilla. Maxillary transverse deficiency (MTD or maxillary hypoplasia) is a common problem that affects the normal development of the maxillofacial complex. Therefore, early diagnosis and correction of MTD are essential to achieve a normal transverse skeletal relationship between the maxilla and mandible (21). There are three types of MPS disjunction: RPE (with dental support), miniscrew-assisted

rapid palatal expansion (MARPE) with skeletal support, and surgically assisted rapid palatal expansion (SARPE). MARPE and SARPE are used in fused MPS or compromised dental support. The introduction of CBCT in orthodontics allows an accurate analysis of sagittal and vertical growth patterns, which helps decide whether to use conventional (RPE) or unconventional maxillary expansion (MARPE or SARPE). A recent study addressed the potential spontaneous adaptive dentoalveolar compensation of the lower arch during RME (27).

The second phase of the treatment consists of using the FR-3 appliance that promotes mandibular growth in a vertical direction and the growth of the maxilla. Compatible with the present case report, many authors (13, 28) reported that the FR-3 appliance promotes an increase in overjet. The increased ANB angle shows that point A advanced sagittally more than point B; therefore, the maxilla has grown more than the mandible. The left hemi-mandible has grown more than the right one and the

height of the left half-maxilla compared to the right one. The increase in bone volume on the non-deviated side is due to the compensation mechanisms that occur when the deviation of the menton is >4 mm (29). A recent study found that RME (with both TB and BB anchorage) could determine a slight opening of the sfero-occipital synchondrosis, with questionable clinical relevance in terms of promoting maxillary protraction helpful during the functional and orthopedic treatment of class III (30).

In bone specific, the most important vertical bone growth occurs at the left mandibular ramus; therefore, the condyle and the goniatic angle on the left side have grown more than on the right side.

The growth of the left hemi-mandible was also confirmed by measuring the inclination of the left ramus external border: the angle with MSP decreased in opposition to the right side, which was slightly increased, proving a strong growth of bone in the transverse direction on the left hemi-mandible, also confirmed by the SMV view. In conclusion, the growing patient with moderate right menton deviation was successfully treated using RPE and FR-3. There was a significant regression of the mandibular asymmetry from moderate grade (4.2 mm) to slight grade (1.3 mm), in addition to the correction of dental characteristics (dental class III and anterior crossbite). These therapeutic goals result from a compensation mechanism: the left hemi-mandible has grown more than the right side, by the literature, which affirms that in case of deviation of the menton >4 mm, the bone volume increases on the non-deviated side.

This treatment protocol is recommended for mandibular asymmetry cases and to use on large samples to better know the effects.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

1. Ercan I, Ozdemir ST, Etoz A, Sigirli D, Tubbs RS, Loukas M, et al. Facial asymmetry in young healthy subjects evaluated by statistical shape analysis. *J Anat.* (2008) 213:663–9. doi: 10.1111/j.1469-7580.2008.01002.x
2. Thiesen G, Gribel BF, Freitas MP. Facial asymmetry: a current review. *Dental Press J Orthod.* (2015) 20:110–25. doi: 10.1590/2177-6709.20.6.110-125.sar
3. Hafezi F, Javdani A, Naghibzadeh B, Ashtiani AK. Laterality and left-sidedness in the nose, face, and body: a new finding. *Plast Reconstr Surg Global Open.* (2017) 5:12. doi: 10.1097/GOX.0000000000001590
4. Geschwind N, Galaburda AM. Cerebral lateralization: Biological mechanisms, associations, and pathology: II. A hypothesis and a program for research. *Arch Neurol.* (1985) 42:521–52.
5. Haraguchi S, Iguchi Y, Takada K. Asymmetry of the face in orthodontic patients. *Angle Orthod.* (2008) 78:421–6. doi: 10.2319/022107-85.1
6. Liu MT, Iglesias RA, Sekhon SS, Li Y, Larson K, Totonchi A, et al. Factors contributing to facial asymmetry in identical twins. *Plast Reconstr Surg.* (2014) 134:638–46. doi: 10.1097/PRS.0000000000000554
7. Nold SL, Horvath SD, Stampf S, Blatz MB. Analysis of select facial and dental esthetic parameters. *Int J Periodontics Restor Dent.* (2014) 34:623–9. doi: 10.11607/prd.1969
8. Festa F, Rotelli C, Scarano A, Navarra R, Caulo M, Macri M. Functional magnetic resonance connectivity in patients with temporomandibular joint disorders. *Front Neurol.* (2021) 12:629211. doi: 10.3389/fneur.2021.629211
9. Deshayes M-J. Traiter orthopédiquement les asymétries avant six ans ou comment symétriser la croissance crano-faciale et optimiser le fonctionnement temporo-mandibulaire. *L'Orthodontie Française.* (2010) 81:189–207. doi: 10.1051/orthodfr/2010021
10. Aytan S, Yukay F, Cigler S, Enacar A, Aksoy A, Telli AE. Frankel III appliance. *Türk Ortodonti Dergisi Ortodonti Derneği'nin Resmi Yayın Organidir Turk J Orthod.* (1989) 2:338–45.
11. Roser C, Hodecker LD, Koebel C, Lux CJ, Ruckes D, Rues S, et al. Mechanical properties of CAD/CAM-fabricated in comparison to

Ethics statement

The studies involving human participants were reviewed and approved by University of Chieti G. D'Annunzio. Ethics approval (number 23) was obtained by the hospital's Independent Ethics Committee of Chieti. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the patient for publication of this report and any accompanying images.

Author contributions

FF and MM performed and documented the orthodontic case and exams, MM performed the analysis of the CBCT images. MM conducted a review of the literature and drafted the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

conventionally fabricated functional regulator 3 appliances. *Sci Rep.* (2021) 11:1–10. doi: 10.1038/s41598-021-94237-x

12. Yang X, Li C, Bai D, Su N, Chen T, Xu Y, et al. Treatment effectiveness of Fränkel function regulator on the Class III malocclusion: a systematic review and meta-analysis. *Am J Orthod Dentofac Orthop.* (2014) 146:143–54. doi: 10.1016/j.ajodo.2014.04.017

13. Levin AS, McNamara JA, Franchi L, Baccetti T, Fränkel C. Short-term and long-term treatment outcomes with the FR-3 appliance of Fränkel. *Am J Orthod Dentofac Orthop.* (2008) 134:513–24. doi: 10.1016/j.ajodo.2006.10.036

14. McGuinness N. Short term and long-term treatment outcomes with the FR-3 appliance of Frankel. *Orthod Update.* (2009) 2:29–29. doi: 10.12968/ortu.2009.2.1.29

15. Gasgoos SS, Al-Saleem NA, Awni K. Cephalometric features of skeletal Class I, II and III (A comparative study). *Al-Rafidain Dent J.* (2007) 7:122–30. doi: 10.33899/rden.2007.8956

16. Franco FC, Araujo TM, Vogel CJ, Quintão CC. Brachycephalic, dolichocephalic and mesocephalic: Is it appropriate to describe the face using skull patterns? *Dent Press J Orthod.* (2013) 18:159–63. doi: 10.1590/S2176-94512013000300025

17. DiPietro GJ, Moergeli JR. Significance of the Frankfort-mandibular plane angle to prosthodontics. *J Prost Dent.* (1976) 36:624–35.

18. Haraguchi S, Takada K, Yasuda Y. Facial asymmetry in subjects with skeletal Class III deformity. *Angle Orthod.* (2002) 72:28–35. doi: 10.1043/0003-3219(2002)072<0028:FAISWS>2.0.CO;2

19. Fränkel R. A functional approach to orofacial orthopaedics. *Br J Orthod.* (1980) 7:41–51.

20. Feragalli B, Rampado O, Abate C, Macri M, Festa F, Stromei F, et al. Cone beam computed tomography for dental and maxillofacial imaging: technique improvement and low-dose protocols. *La Radiol Med.* (2017) 122:581–8. doi: 10.1007/s11547-017-0758-2

21. Macri M, Toniato E, Murmura G, Varvara G, Festa F. Midpalatal suture density as a function of sex and growth-pattern-related variability via CBCT evaluations of 392 adolescents treated with a rapid maxillary expander appliance. *Appl Sci.* (2022) 12:2221. doi: 10.3390/app12042221

22. Gribel BF, Thiesen G, Borges TS, Freitas MP. Prevalence of mandibular asymmetry in skeletal Class I adult patients. *J Res Dent.* (2014) 2:189–97. doi: 10.19177/jrd.v2e22014189-97

23. Boel T, Sofyanti E, Sufarnap E. Analysing menton deviation in posteroanterior cephalogram in early detection of temporomandibular disorder. *Int J Dent.* (2017) 2017:5604068. doi: 10.1155/2017/5604068

24. Ahmed M, Shaikh A, Fida M. Diagnostic performance of various cephalometric parameters for the assessment of vertical growth pattern. *Dental Press J Orthod.* (2016) 21:41–9. doi: 10.1590/2177-6709.21.4.041-049.oar

25. Subramaniam P, Naidu P. Mandibular dimensional changes and skeletal maturity. *Contemp Clin Dent.* (2010) 1:218.

26. Valletta R, Rongo R, Pango Madariaga AC, Baiano R, Spagnuolo G, D'Antò V. Relationship between the Condylion–Gonion–Menton angle and dentoalveolar heights. *Int J Environ Res Public Health.* (2020) 17:3309. doi: 10.3390/ijerph17093309

27. Lo Giudice A, Ronsivalle V, Lagravere M, Leonardi R, Martina S, Isola G. Transverse dentoalveolar response of mandibular arch after rapid maxillary expansion (RME) with tooth-borne and bone-borne appliances: a CBCT retrospective study. *Angle Orthod.* (2020) 90:680–7. doi: 10.2319/0425-20-353.1

28. Kerr WJ, TenHave TR, McNamara JA. A comparison of skeletal and dental changes produced by function regulators (FR-2 and FR-3). *Eur J Orthod.* (1989) 11:235–42.

29. Kim SJ, Lee KJ, Lee SH, Baik HS. Morphologic relationship between the cranial base and the mandible in patients with facial asymmetry and mandibular prognathism. *Am J Orthod Dentofac Orthop.* (2013) 144:330–40. doi: 10.1016/j.ajodo.2013.03.024

30. Leonardi R, Ronsivalle V, Lagravere MO, Barbato E, Isola G, Lo Giudice A. Three-dimensional assessment of the spheno-occipital synchondrosis and clivus after tooth-borne and bone-borne rapid maxillary expansion: a retrospective CBCT study using voxel-based superimposition. *Angle Orthod.* (2021) 91:822–9. doi: 10.2319/013021-86.1



OPEN ACCESS

EDITED BY

Yi-Ju Tseng,
National Yang Ming Chiao Tung
University, Taiwan

REVIEWED BY

Jose A. Vega,
University of Oviedo, Spain
Felice Festa,
University of Studies G. d'Annunzio
Chieti and Pescara, Italy

*CORRESPONDENCE

Xin Xiong
drxiongxin@scu.edu.cn

SPECIALTY SECTION

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

RECEIVED 16 September 2022

ACCEPTED 04 November 2022

PUBLISHED 17 November 2022

CITATION

Zhu R, Zheng Y-H, Zhang Z-H,
Fan P-D, Wang J and Xiong X (2022)
Development of a new category
system for the profile morphology of
temporomandibular disorders patients
based on cephalograms using cluster
analysis.
Front. Public Health 10:1045815.
doi: 10.3389/fpubh.2022.1045815

COPYRIGHT

© 2022 Zhu, Zheng, Zhang, Fan, Wang
and Xiong. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Development of a new category system for the profile morphology of temporomandibular disorders patients based on cephalograms using cluster analysis

Rui Zhu¹, Yun-Hao Zheng², Zi-Han Zhang², Pei-Di Fan²,
Jun Wang² and Xin Xiong^{2,3*}

¹The State Key Laboratory of Oral Diseases and National Clinical Research Center for Oral Diseases, Department of Prosthodontics, West China Hospital of Stomatology, Sichuan University, Sichuan, China, ²The State Key Laboratory of Oral Diseases and National Clinical Research Center for Oral Diseases, Department of Orthodontics, West China Hospital of Stomatology, Sichuan University, Sichuan, China, ³Department of Temporomandibular Joint, West China Hospital of Stomatology, Sichuan University, Sichuan, China

Objective: This study aims to develop a new category scheme for the profile morphology of temporomandibular disorders (TMDs) based on lateral cephalometric morphology.

Methods: Five hundred and one adult patients (91 males and 410 females) with TMD were enrolled in this study. Cluster tendency analysis, principal component analysis and cluster analysis were performed using 36 lateral cephalometric measurements. Classification and regression tree (CART) algorithm was used to construct a binary decision tree based on the clustering results.

Results: Twelve principal components were discovered in the TMD patients and were responsible for 91.2% of the variability. Cluster tendency of cephalometric data from TMD patients were confirmed and three subgroups were revealed by cluster analysis: (a) cluster 1: skeletal class I malocclusion; (b) cluster 2: skeletal class I malocclusion with increased facial height; (c) cluster 3: skeletal class II malocclusion with clockwise rotation of the mandible. Besides, CART model was built and the eight key morphological indicators from the decision tree model were convenient for clinical application, with the prediction accuracy up to 85.4%.

Conclusion: Our study proposed a novel category system for the profile morphology of TMDs with three subgroups according to the cephalometric morphology, which may supplement the morphological understanding of TMD and benefit the management of the categorical treatment of TMD.

KEYWORDS

temporomandibular disorders, cluster analysis, cephalometric analysis, classification and regression tree (CART), morphological category

Introduction

Temporomandibular disorders (TMDs) are a set of clinical conditions associated with the temporomandibular joint (TMJ), masticatory muscles, and orofacial structures (1–4). Generally, approximately 5% of the population suffered from these disorders with a prevalence between 5 and 15% in adults (5, 6). However, the situation of TMDs is not encouraging recently. Evidence shows that the prevalence of TMDs is increasing recently, with an overall prevalence of 31% in adults and 11% in children and adolescence (7). Besides, the most frequent TMD related symptoms including restricted mouth opening, TMJ sounds, and TMJ pain have been up to 50% in adults (8), which greatly affects the patients' quality of life.

Nowadays, in spite of various methods with well diagnostic reliability and validity developed for diagnosing TMDs (9–11), the Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) is still the most widely utilized, thorough and accurate diagnostic criteria worldwide for assessment and classification of TMD (12), which comprehensively takes both characterization of the disease in the joint and muscle (Axis I) and psychosocial disability (Axis II) into consideration (13). Although DC/TMD is an excellent tool to diagnose and classify the TMDs, there also exists several vacancies about lateral cephalograms and further efforts are still needed for relevant research.

Lateral cephalometric radiograph, an easily accessible and non-invasive examination, can supply abundant data concerning the cranial, facial bony and soft tissue structures. For its economy and convenience, lateral cephalometric radiograph has been not only widely used as facial analysis before and after orthodontic treatment, but also utilized to explore the association between TMD including its symptoms and the characteristics of craniofacial morphology (14–17). Already in 1995, lateral cephalometry was applied to investigate the association between morphologic features and internal derangements of the temporomandibular joint (15). Recently, the craniofacial morphology of TMD and has been well-investigated (16) and it is reported that patients with TMD exhibit specific craniofacial features compared to patients without TMD (16, 17). Our previous study (14) also validated the results and further observed a significant difference in Frankfort-mandibular plane angle (FMA) between patients with and without TMDs. Besides, we found there existed specific craniofacial features between TMD patients with and without TMJ pain as well (14). At present, although these studies revealed the significant relationship between TMD and morphologic features, the indicators from lateral cephalometric radiograph were still mainly applied to judge the skeletal pattern of the patients by orthodontic diagnosis and only partially reflected the features of TMD, which might help little for the treatment of TMD patients. Consequently, it is necessary to develop a new

category system specific to TMD to integrate those significant features for clinical application.

Clustering analysis is an unsupervised learning model widely used in data mining (18) and has been utilized to determine the subtypes of many diseases according to their numerous indicators such as idiopathic inflammatory myopathies (19), class III malocclusion (20) and others (21). However, there was no clustering analysis based on the cephalograms in the research of TMD.

In this study, in order to make the most of these indicators from lateral cephalometric radiograph, we develop a new category system for the profile morphology of TMD patients using cluster analysis according to thirty-six cephalometric parameters.

Materials and methods

Subjects and study design

The research was conducted at the Department of Orthodontics, West China Hospital of Stomatology, Sichuan University, from June 2021 to October 2021. All patients were investigated and diagnosed by one TMD specialist who had received extensive training and calibration in the use of the DC/TMD (12).

The inclusion criteria were as follows: (a) patients diagnosed with TMD for the first time; (b) patients aged 18 years or above; and (c) patients with available chart, lateral cephalograms, and photographs. The exclusion criteria were: (a) presence of tumor, trauma and/or surgery history in the maxilla and facial area; (b) presence of clefts and other craniofacial anomalies.

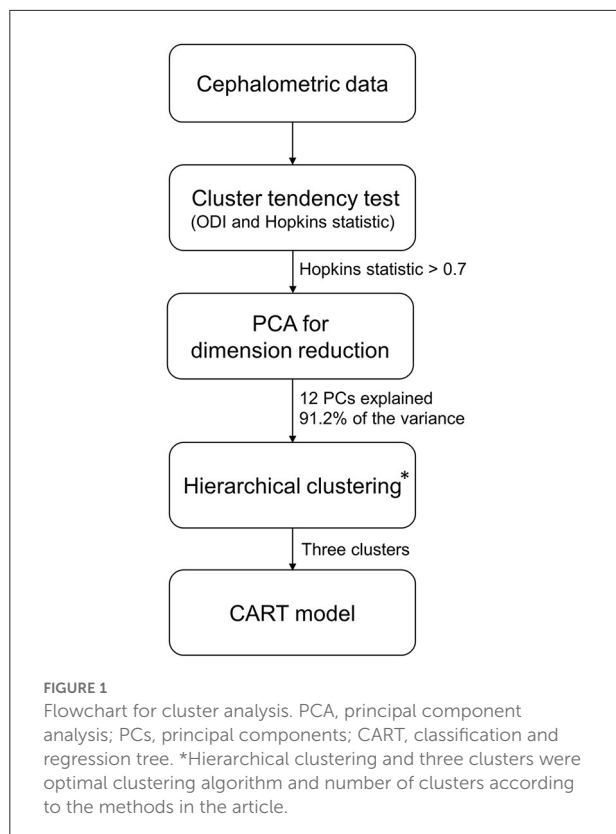
The study was approved by the Ethics Committee of West China School of Stomatology of Sichuan University (Ethics number: 2021-396) and was conducted in accordance with the Declaration of Helsinki. Informed consents were provided with all the patients.

This study was carried out based on multiple clustering approaches and general procedures were given in the flowchart (Figure 1).

Cephalometric analysis

All the patients' lateral cephalograms were collected before they started to receive orthodontic treatment by the same radiologist. Patients had to maintain the natural head position with the mandible in the maximum intercuspal position by request (22). The Uceph software (Chengdu Yaxun, Chengdu, China) was applied for cephalometric analysis after collecting the lateral cephalograms.

Table 1 showed the thirty-six cephalometric parameters measured in the study. The measurements were conducted by



two researchers blinded to the patients' details. According to the approach described by Xiong et al. (23), inter-observer and intra-observer reliability were examined to ensure the accuracy of the measurements. For inter-observer reliability, 20 lateral cephalograms were selected randomly and measured by the examiners for the first time. After a washout period of 4 weeks, the observer repeated the measurement. The intra-class correlation coefficient (ICC) was calculated to test the repeatability of the results. The examiners were eligible when ICC was over 0.75.

Cluster tendency analysis

The dissimilarity matrix based on Euclidean distance metrics between the normalized samples was calculated and reordered to form an ordered dissimilarity image (ODI). The visual assessment of cluster tendency algorithm (VAT) was used to visualize the ODI (24). Considering that clustering algorithms will locate and specify clusters in data even if none are present, Hopkins statistic H was used to validate cluster tendency. The significance level was set to $H > 0.7$, which meant that data had a cluster tendency and the clustering results were meaningful (25).

TABLE 1 Cephalometric variables.

Cranial base	S-Go (mm)	Interincisal angle (U1-L1) (°)
Saddle/Sella angle (°)	Mandibular body length (Go-Me) (mm)	U1-SN(°)
Anterior cranial base (S-N) (mm)	Intermaxillary	UPDH (U6-PP) (mm)
Posterior cranial base (S-Ar) (mm)	Midface length (Co-A) (mm)	LPDH (L6-MP) (mm)
Maxilla	ANB (°)	U1-ANS (mm)
SNA (°)	Y-axis (°)	L1-Me (mm)
PP-FH (°)	Y-axis length (mm)	MP-OP (°)
Mandible	Wits appraisal (mm)	PP-OP (°)
SNB (°)	Anterior face height (N-Me) (mm)	OP-FH (°)
Gonial/Jaw angle (Ar-Go-Me)(°)	FMA (FH-MP) (°)	Overbite (mm)
Ramus height (Ar-Go) (mm)	ANS-Xi-Pm (°)	Overjet (mm)
Articular angle (S-Ar-Go) (°)	Dental	Soft Tissue
Dc-Xi-Pm (°)	IMPA (L1-MP) (°)	Upper lip to E-plane (UL-EP) (mm)
SN-MP (°)	FMIA (L1-FH) (°)	Lower lip to E-plane (LL-EP) (mm)

Boldface indicates six categories of the thirty-six cephalometric parameters.

Principal components analysis

Principal components (PCs) are a series of mutually orthogonal variables formed by linear combinations of the original data variables and are arranged in descending order according to their ability to describe the variance of the original data.

To calculate the principal components, the data matrix needed to be normalized first, and the variables of the normalized data matrix Z are then linearly combined as principal components in the form of equation (1) through algorithms (e.g., maximum projection variance, singular value decomposition, etc.) making the data have the largest variance in the first principal component, followed by the second principal component, and so on.

$$PC_k = \sum_{i=1}^N a_{ik} Z_i \quad (1)$$

where PC_k is the k -th principal component, a_{ik} is the coefficient of the linear combination obtained according to a specific

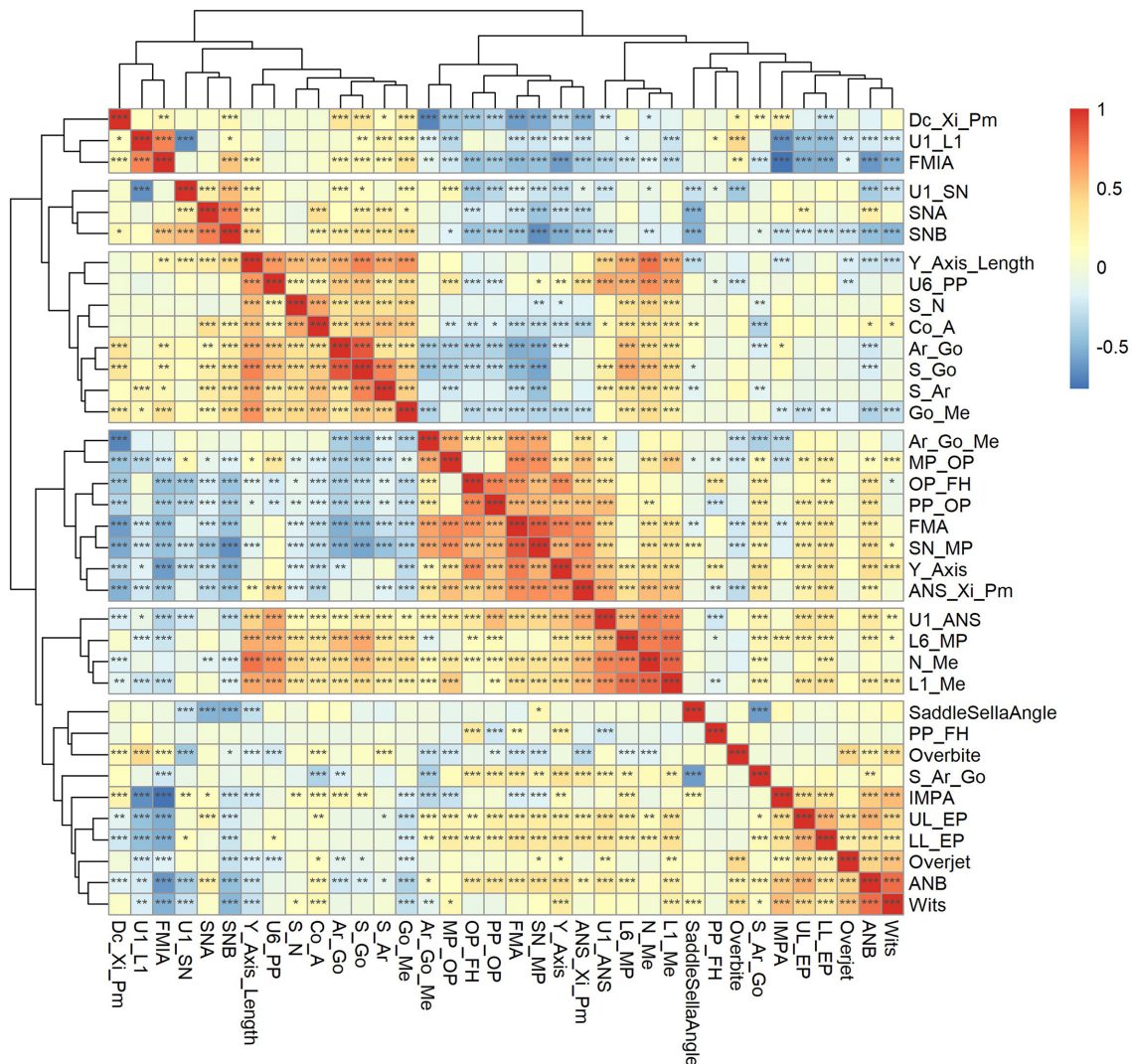


FIGURE 2
Pearson's correlation coefficient heat map and hierarchical clustering dendrogram for cephalometric variables. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

algorithm, and Z_i is the i -th column of the centralized matrix Z , i.e., the i -th variable.

The first n principal components are selected to satisfy (i) the cumulative percentage of variance exceeds 90%; (ii) the $(n + 1)$ -th to N -th principal components have sufficiently small contribution to the variance to be used as pre-processed data for modeling.

Optimization of number of clusters and clustering algorithm

The number of clusters was evaluated by using 26 indices such as CH index and Dula index, and the optimal

number of clusters was selected according to the “majority voting” principle (26). The optimal clustering algorithm was selected by calculating the connectivity, Dunn and Silhouette indices of three common clustering methods, namely hierarchical clustering, K-means clustering and partitioning around medoids (PAM), for the selected number of clusters.

Hierarchical clustering on principal components

Hierarchical clustering was performed based on Ward's minimum variance method on the basis of principal component

TABLE 2 Baseline characteristics of the cephalometric variables.

Variables	Male (<i>n</i> = 91)	Female (<i>n</i> = 410)	Total (<i>n</i> = 501)
Age	29.25 (9.95)	32.16 (10.63)	31.63 (10.56)
Cranial base			
Saddle/Sella Angle	126.05 (5.10)	125.62 (5.29)	125.70 (5.25)
S-N	65.39 (2.91)	61.90 (2.85)	62.53 (3.16)
S-Ar	35.08 (3.50)	32.13 (2.90)	32.66 (3.22)
Maxilla			
SNA	82.52 (3.63)	81.87 (3.49)	81.98 (3.52)
PP-FH	0.14 (2.76)	0.19 (2.76)	0.18 (2.76)
Mandible			
SNB	78.28 (4.14)	77.57 (3.76)	77.70 (3.84)
Ar-Go-Me	116.91 (7.21)	117.49 (5.97)	117.39 (6.21)
Ar-Go	50.73 (5.10)	46.01 (4.15)	46.87 (4.70)
S-Ar-Go	148.69 (7.12)	151.24 (6.45)	150.78 (6.64)
Dc-Xi-Pm	37.11 (5.64)	37.01 (5.64)	37.03 (5.63)
SN-MP	31.78 (6.34)	34.53 (5.96)	34.03 (6.12)
S-Go	82.57 (6.64)	75.66 (5.38)	76.91 (6.22)
Go-Me	70.98 (6.11)	68.07 (4.39)	68.60 (4.87)
Intermaxillary			
Co-A	84.96 (6.63)	79.40 (4.18)	80.41 (5.18)
ANB	4.23 (2.71)	4.29 (2.69)	4.28 (2.69)
Y-Axis	60.87 (3.64)	61.29 (3.45)	61.21 (3.48)
Y-Axis length	121.40 (7.28)	114.76 (5.96)	115.96 (6.72)
Wits	1.31 (3.68)	0.53 (3.43)	0.67 (3.49)
N-Me	119.48 (6.69)	113.93 (6.13)	114.94 (6.58)
FMA	22.37 (5.83)	24.68 (5.39)	24.26 (5.54)
ANS-Xi-Pm	46.39 (4.44)	47.31 (4.73)	47.14 (4.69)
Dental			
IMPA	98.58 (7.70)	97.26 (7.55)	97.50 (7.59)
FMIA	59.04 (8.31)	58.04 (8.56)	58.22 (8.52)
U1-L1	126.27 (11.01)	125.72 (12.19)	125.82 (11.97)
U1-SN	103.34 (7.77)	102.46 (8.62)	102.62 (8.47)
U6-PP	23.60 (2.44)	22.38 (2.10)	22.60 (2.21)
L6-MP	33.67 (2.90)	31.48 (2.61)	31.88 (2.80)
U1-ANS	29.00 (2.94)	28.39 (2.51)	28.50 (2.60)
L1-Me	41.71 (3.11)	39.54 (3.06)	39.93 (3.18)
MP-OP	15.20 (3.86)	16.20 (4.10)	16.02 (4.08)
PP-OP	6.74 (3.69)	8.16 (3.43)	7.91 (3.52)
OP-FH	7.24 (4.14)	8.48 (3.78)	8.25 (3.87)
Overbite	2.76 (2.05)	2.52 (1.79)	2.56 (1.84)
Overjet	4.00 (1.89)	4.04 (1.66)	4.03 (1.70)
Soft Tissue			
UL-EP	0.81 (2.81)	0.37 (2.60)	0.45 (2.64)
LL-EP	0.98 (2.54)	0.76 (2.64)	0.80 (2.62)

Mean (SD), SD, standard deviation.

analysis (PCA), and the initial partitions obtained from the hierarchical clustering were improved by K-means clustering (27). The PCA step can be considered as a denoising step which can lead to a more stable clustering.

Classification and regression tree

Classification and regression tree (CART) algorithm was used to construct a binary decision tree to help dentists classify TMD according to patients' cephalometric characteristics easily. We performed cross-validation to select the optimal tree and performed multiple runs to avoid overfitting. Cephalometric dataset was split into 70% as training set and 30% as validation set and the classification tree model was evaluated by the accuracy of prediction. Confusion matrix was made to visualize and summarize the performance of the CART model (Supplementary Figure 1).

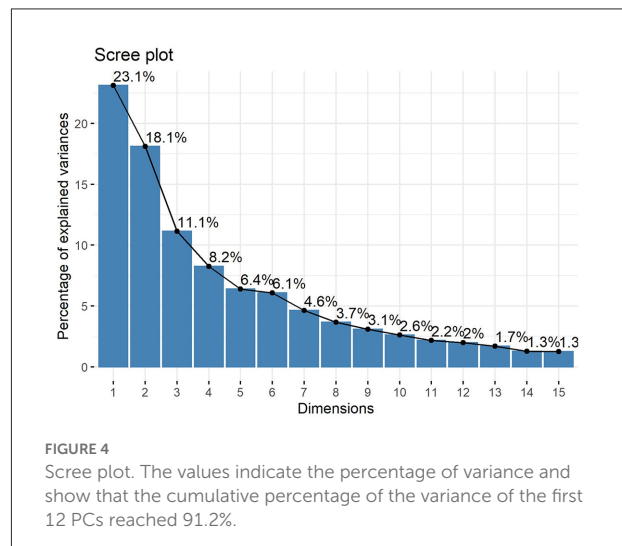
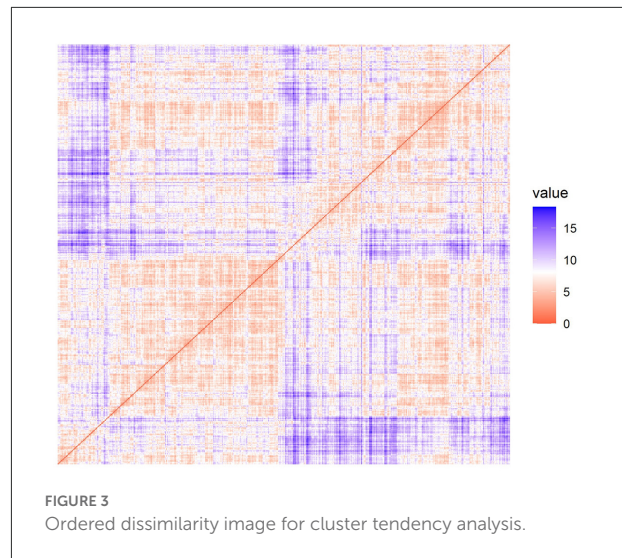
Statistical analysis

One-way ANOVA, Tukey HSD *post hoc* test, Kruskal-Wallis test, Dunn *post hoc* test and Bonferroni correction were used for hypothesis testing. Pearson's correlation coefficient was used to explore the correlation of the normalized variables in cephalometric data and was visualized by a heat map (Figure 2). The Pearson's correlation coefficient of 0.40, 0.60, and 0.80 were considered weak, moderate and strong associations respectively. At the same time, hierarchical clustering was performed on the normalized variables. Feature selection and feature transformation was conducted to improve the final clustering effect. All statistical analyses were based on Language R, version 4.1.3 (R Foundation for Statistical Computing, Vienna, Austria). $P < 0.05$ was considered statistically significant.

Results

Baseline characteristics of the cephalometric variables

Five hundred and one adult orthodontic patients diagnosed with TMD were included in this study. The mean age of the patients was 31.63 ± 10.56 years. Of the 501 patients, 91 were males and 410 were females (81.8%). Thirty-six cephalometric parameters shown in Table 1 were measured to reflect the TMD patients' maxillofacial features in six categories, including cranial base, maxilla, mandible, intermaxillary relation, teeth and soft tissue (Table 2).



Cluster tendency of cephalometric data

According to the ODI, it was observed that the dissimilarity matrix presented a block phenomenon along the inverse diagonal direction (Figure 3), indicating that cephalometric data had a cluster tendency. The Hopkins statistic ($H = 0.736 > 0.7$) also showed a significant cluster tendency of cephalometric data, which ensured the statistical significance of clustering analysis.

Principal component analysis for cephalometric data

A strong linear correlation was found by correlation analysis and cluster tendency analysis implied that feature selection and feature transformation should be conducted to improve the

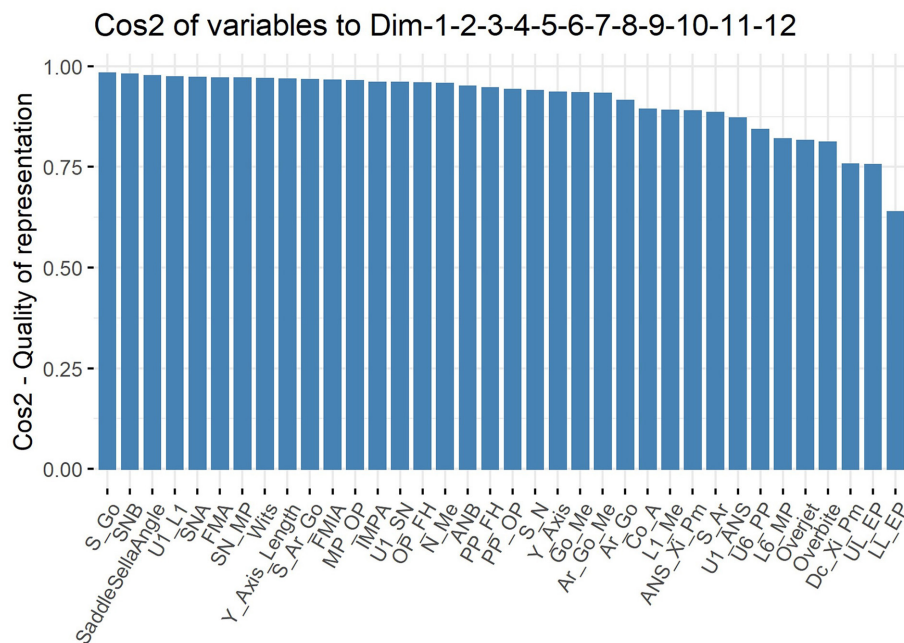


FIGURE 5

Cos2 plot. 24 of the 36 cephalometric variables had Cos2 values >0.9 (66.7%) and 33 were >0.8 (91.7%).

final clustering effect (Figure 2). Therefore, it was necessary to perform PCA to combine variables.

The cumulative percentage of the variance of the first 12 PCs was calculated to be 91.2%, and the percentage of the variance of each PC after the 13th PC < 2% (Figure 4). Consequently, the first 12 PCs were chosen to represent the entire data. The Cos2 of the first 12 PCs on each variable was calculated, and the results showed that the first 12 PCs were able to represent each variable to a good extent (Figure 5).

Clusters of cephalometric data of TMD patients

Twenty-six indices were used to evaluate different numbers of clusters for the data after principal component analysis from two to nine, and fifteen indices recommended that the data should be divided into three clusters, accounting for 57.7% (Figure 6). Connectivity, Dunn and Silhouette indices of three common clustering algorithms, including hierarchical clustering, K-means clustering and PAM, were calculated. The results (Table 3) showed that the optimal number of clusters was three and the optimal algorithm was the hierarchical clustering algorithm.

Hierarchical clustering on PCs divided cephalometric data of TMD patients into three clusters which 34 of the 36 cephalometric parameters (94.4%), as well as age and sex,

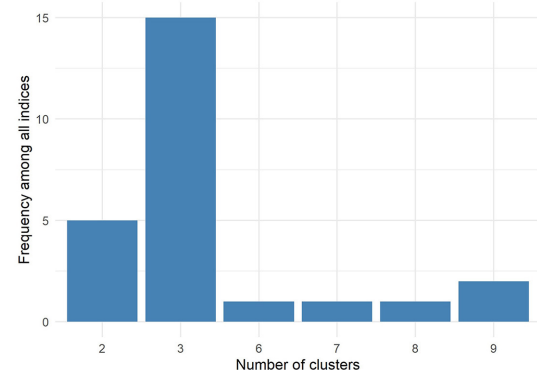


FIGURE 6

Fifteen of the twenty-six indices (57.7%) showed that the optimal number of clusters was three.

showed significant differences (Table 4). The projection of scatter plot for three clusters on the first two PCs (Figure 7) visualized the clustering result and the clear clustering boundaries indicated the reliability of our clustering result. The cluster dendrogram (Figure 8) visualized the clustering result from another perspective, which could show that there were no outliers in the clusters, supporting the reasonability and reliability of the clustering result.

TABLE 3 Connectivity, dunn, and silhouette indices of three commonly used clustering methods in three clusters.

Indices*	Hierarchical clustering	K-means clustering	PAM
Connectivity	68.49	272.6	334.5
Dunn	0.182	0.158	0.142
Silhouette	0.152	0.152	0.100

*Dunn and Silhouette indices are positively correlated with the clustering effect while the Connectivity index is negatively correlated with the clustering effect.

Patients with TMD were divided into three groups and each group could be given clinical meanings according to the cephalometrics in orthodontics as visualized in Figure 9: (a) cluster 1: skeletal class I malocclusion; (b) cluster 2: skeletal class I malocclusion with increased facial height; (c) cluster 3: skeletal class II malocclusion with clockwise rotation of the mandible and anterior open bite. Patients in cluster 1 only showed skeletal class I malocclusion ($ANB = 3.27^\circ$) and normo-divergent ($SN-MP = 30.74^\circ$, $FMA = 21.17^\circ$). Patients in cluster 2 presented skeletal class I malocclusion ($ANB = 3.67^\circ$), normo-divergent ($SN-MP = 30.57^\circ$, $FMA = 21.73^\circ$), increased posterior facial height ($S-Go = 84.98$ mm), increased anterior facial height ($N-Me = 121.42$ mm) and a slight protrusion of upper lip ($UL-EP = 0.57$ mm). Patients in cluster 3 exhibited skeletal class II malocclusion ($ANB = 5.68^\circ$), hyperdivergent ($SN-MP = 39.38^\circ$, $FMA = 28.89^\circ$), tendency of protrusive incisors ($U1-L1 = 120.74^\circ$), anterior overjet (4.35 mm) and anterior open bite.

CART model for prediction of cephalometric data category

A CART model was built based on the clustering results (Figure 10) to easily classify TMDs into the three clusters. The data were split into training and validation set by 70:30 and the prediction accuracy was 85.4%, which indicated the CART model had effective predictive power for our previously proposed clusters of TMD patients. Confusion matrix also showed good performance of the CART model (Supplementary Figure 1).

Discussion

The diagnosis and classification of TMDs has been discussed since last century. However, the evaluated systems did not meet the diagnostic criteria until the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) was proposed in 1992 (13). After years of expansion and refinement, the DC/TMD was released on the basis of RDC/TMD in 2014, which improved axis II procedures and delineated 12 disorders in detail. With years of practical application, the current dominant two-axis approach was not enough in clinical application and

Enriqueta C. Bond noted that a broader exploration to the painful TMD beyond the two axes was necessary in future research (13). Lateral cephalometric radiograph, recognized as the most commonly used examination during orthodontal treatment (28), has been already widely applied to explore the associations between TMD and craniofacial morphology (15, 29, 30). Although the specific craniofacial features of the TMD patients were observed in many studies through cephalometric analysis, the features could not reflect the whole morphology and were difficult for clinical application. Therefore, in this study, through analyzing the features of TMD obtained from cephalometric radiograph, we developed a new category system and proposed a CART model of TMD for clinical application based on cephalometric morphology aiming to make progress for the morphological understanding of TMD. For this study was to identify the subgroups only among TMD patients, healthy populations without TMD were not included. This is the first study to classify TMD using unsupervised analysis according to lateral cephalometric radiographs in a large population ($n = 501$). The gender distribution in our study was consist with the clinical situation that females account for the majority of TMD patients (31) and the cluster analysis was conducted according to 36 morphological features, which assured a reliable and comprehensive evaluation.

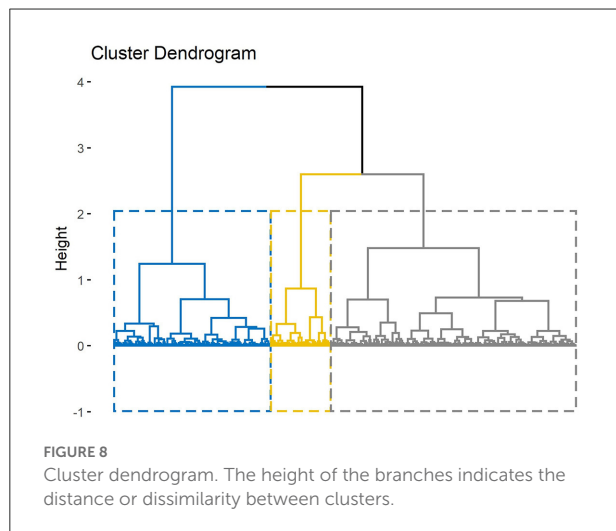
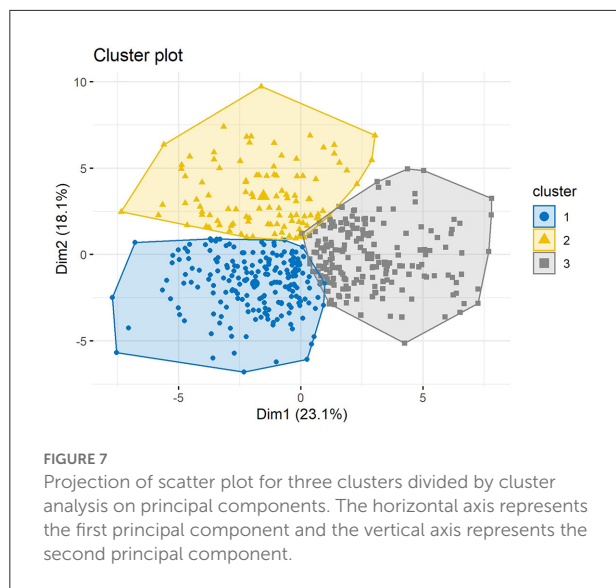
In the cluster analysis, three subgroups were identified from the 36 variables among the 501 participants. In this procedure, the clustering algorithm was performed for a range of 2–9 clusters separately. According to our results, fifteen of the twenty-six indices (57.7%) showed that the optimal number of clusters was three. Intriguingly, three subgroups were also identified in another cluster analysis with a large sample including 1,031 chronic TMD cases and 3,247 TMD-free controls, which was consistent in the cluster numbers calculated in our study (32). Thus, we determined three subtypes of patients with TMD based on the cluster analysis. To our delight, each group corresponded to the entity with distinct features.

For patients in cluster 1, the values of the morphological features were mostly in the normal range (33), indicating that this group of patients did not exhibit much difference in their appearances compared with normal population, which may explain why some researchers did not find distinct relationship between morphologic features of the face and TMD when the sample size was not large enough (15). Since there was not

TABLE 4 Comparison of cephalometric variables among three clusters.

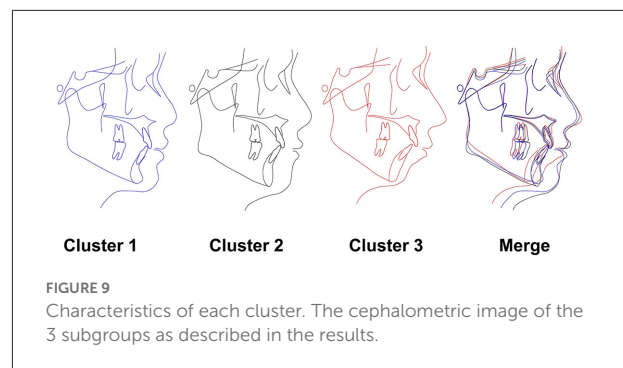
Variables	Cluster1 (n = 202)	Cluster2 (n = 106)	Cluster3 (n = 193)	P-value	Multiple comparisons
Age ^k	29.83 (8.90)	31.43 (10.87)	33.63 (11.65)	0.007**	3 > 1
Sex(M/F) ^c	18/184	54/52	19/174	<0.001***	2 > (1, 3)
Cranial Base					
Saddle/Sella	125.72 (5.12)	125.20 (4.63)	125.94 (5.71)	0.511	–
Angle^a					
S-N ^k	61.88 (2.65)	65.46 (3.23)	61.62 (2.64)	<0.001***	2 > (1, 3)
S-Ar ^a	32.26 (2.80)	35.83 (2.71)	31.35 (2.72)	<0.001***	2 > 1 > 3
Maxilla					
SNA ^a	82.37 (3.27)	83.33 (3.61)	80.84 (3.38)	<0.001***	2 > 1 > 3
PP-FH ^k	0.10 (2.62)	−0.44 (2.90)	−0.34 (2.79)	0.093	–
Mandible					
SNB ^k	79.09 (3.18)	79.65 (3.69)	75.17 (3.13)	<0.001***	(1, 2) > 3
Ar-Go-Me ^k	115.90 (5.91)	115.74 (6.87)	119.85 (5.30)	<0.001***	3 > (1, 2)
Ar-Go ^a	46.16 (3.64)	52.26 (4.22)	44.65 (3.53)	<0.001***	2 > 1 > 3
S-Ar-Go ^k	149.03 (6.09)	149.57 (6.19)	153.27 (6.69)	<0.001***	3 > (1, 2)
Dc-Xi-Pm ^k	38.60 (5.34)	38.23 (5.10)	34.73 (5.47)	<0.001***	(1, 2) > 3
SN-MP ^k	30.74 (4.35)	30.57 (4.70)	39.38 (4.36)	<0.001***	3 > (1, 2)
S-Go ^k	75.58 (4.52)	84.98 (4.77)	73.88 (4.46)	<0.001***	2 > 1 > 3
Go-Me ^a	68.42 (4.10)	73.33 (4.32)	66.18 (3.98)	<0.001***	2 > 1 > 3
Intermaxillary					
Co-A ^k	79.72 (4.13)	85.55 (5.42)	78.31 (4.05)	<0.001***	2 > 1 > 3
ANB ^k	3.27 (2.49)	3.67 (2.62)	5.68 (2.32)	<0.001***	3 > (1, 2)
Y-Axis ^a	59.07 (2.72)	60.70 (2.98)	63.74 (2.75)	<0.001***	3 > 2 > 1
Y-Axis	113.20 (4.75)	124.65 (5.58)	114.09 (4.87)	<0.001***	2 > (1, 3)
Length^k					
Wits ^k	−0.22 (3.33)	0.52 (3.63)	1.70 (3.31)	<0.001***	3 > (1, 2)
N-Me ^k	109.96 (4.56)	121.42 (5.50)	116.58 (4.75)	<0.001***	2 > 3 > 1
FMA ^k	21.17 (4.14)	21.73 (4.54)	28.89 (3.97)	<0.001***	3 > (1, 2)
ANS-Xi-Pm ^k	43.88 (3.84)	46.92 (3.49)	50.67 (3.39)	<0.001***	3 > 2 > 1
Dental					
IMPA ^k	95.85 (8.24)	98.49 (7.38)	98.70 (6.66)	0.001**	(2, 3) > 1
FMIA ^a	62.97 (7.39)	59.77 (6.87)	52.41 (6.85)	<0.001***	1 > 2 > 3
U1-L1 ^a	130.48 (11.81)	126.20 (10.99)	120.74 (10.60)	<0.001***	1 > 2 > 3
U1-SN ^k	102.91 (8.49)	104.72 (8.94)	101.17 (7.93)	<0.001***	2 > 3
U6-PP ^a	21.48 (1.78)	24.58 (1.78)	22.68 (2.05)	<0.001***	2 > 3 > 1
I6-MP ^k	30.02 (2.02)	34.60 (2.37)	32.34 (2.25)	<0.001***	2 > 3 > 1
U1-ANS ^k	26.57 (2.07)	29.84 (2.26)	29.78 (1.95)	<0.001***	(2, 3) > 1
L1-Me ^k	37.35 (2.07)	42.65 (2.54)	41.14 (2.40)	<0.001***	2 > 3 > 1
MP-OP ^k	14.13 (3.54)	15.22 (3.66)	18.44 (3.58)	<0.001***	3 > (1, 2)
PP-OP ^a	6.51 (3.02)	6.45 (3.03)	10.17 (3.05)	<0.001***	3 > (1, 2)
OP-FH ^a	7.04 (3.24)	6.56 (3.43)	10.45 (3.69)	<0.001***	3 > (1, 2)
Overbite ^k	2.84 (1.68)	2.76 (1.93)	2.16 (1.89)	<0.001***	1 > 3
Overjet ^k	3.85 (1.53)	3.80 (1.79)	4.35 (1.78)	0.026*	–
Soft tissue					
UL-EP ^k	−0.82 (2.08)	0.57 (2.66)	1.73 (2.53)	<0.001***	3 > 2 > 1
LL-EP ^a	−0.58 (2.30)	0.82 (2.18)	2.24 (2.38)	<0.001***	3 > 2 > 1

Mean (SD), SD: standard deviation. *P < 0.05, **P < 0.01, ***P < 0.001. ^aOne-way ANOVA and Tukey HSD *post hoc* test. ^cChi-square test. Bonferroni's method was used for multiple comparisons. The result showed the sex composition of Cluster 2 was significantly different from Cluster 1 and Cluster 3 with more male patients, while there were no significant differences between the sex composition of Cluster 1 and Cluster 2. ^kKruskal–Wallis test and Dunn *post hoc* test. Bonferroni's method was used for multiple comparisons.



much difference in the appearance of the patients compared with normal population in cluster 1, the TMD could be more likely developed by psychological distress than intra-articular lesion, which the latter more or less affected the morphological features of TMD patients (14, 16, 17, 34–37). Consequently, conservative therapy and psychological intervention may be the first choice for treating TMD patients in cluster 1.

Most of the cephalometric C of angles in cluster 2 were quite similar with those of cluster 1. However, the cephalometric measurements of linear distances in cluster 2 were larger than those of cluster 1, indicating cluster 2 exhibited a larger craniofacial size than cluster 1 with significant increases in posterior facial height, anterior facial height and S-Go/N-Me (70.0%). The differences may be mainly attributed to the gender factor with the percentage of males in cluster 1 and cluster



2 being 8.9 and 51% respectively. A previous study on TMD classification reported a cluster with equal gender distribution exhibited “normal” psychological conditions but were more sensitive to muscle pain (32). It can be extrapolated that patients in cluster 2 with even gender balance may also presented the same symptoms. Therefore, conservative therapy especially pain management may be optimal for treating TMD of cluster 2 for the first time. However, the validation of the abovementioned suggestion is still reserved for future work.

Specific craniofacial features observed in patients with TMD in many studies may mainly refer to the cluster 3 patients in our study (14, 16, 17). Previous studies compared the craniofacial morphology of patients with and without TMD and found that patients with TMDs exhibited specific craniofacial features such as skeletal class II malocclusion, hyperdivergent growth pattern, increased FMA, clockwise rotation of the mandible, anterior open bite and others (14, 16, 17, 34–37), reflecting the craniofacial morphology of TMD patients in cluster 3. Considering the great differences in craniofacial morphology, patients in cluster 3 may suffer from more severe TMD symptoms than cluster 1 and cluster 2. Studies revealed that the clockwise rotation of the mandible was associated with disk displacement (DD) and can be aggravated with the development of DD (38, 39). A recent study published in June 2022 suggested that the abnormality of craniofacial structures resulted from TMJ pain could be reversed by pain control therapy. Therefore, in spite of conservative therapy including pain management, it could be more important for the TMD patients in cluster 3 to improve the risky facial type. Orthodontic therapies such as passive aligners (4) or even surgical method may be considered during the treatment of TMD.

The assessment and classification of TMDs remains a challenge for dentists these days, despite multiple relevant researches in this field. This is because TMDs are a group of disease and patients can be diagnosed as multiple TMDs simultaneously due to the complex etiologies and various symptoms of TMD (7). For simple and convenient application in clinic, a CART model was designed to help dentists classify TMD according to patients’ cephalometric characteristics and

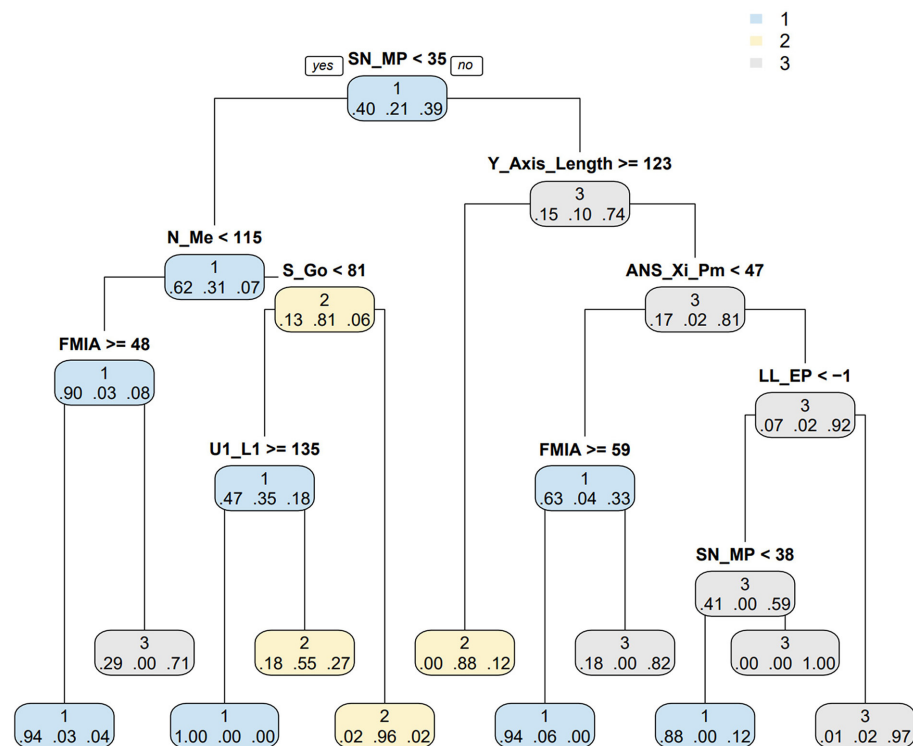


FIGURE 10

CART model for predicting cephalometric data category. The left branch of the binomial tree indicates the cases that meet the conditions while the right branch do not. The decimal below the branch indicates the probability that the sample belongs to cluster 1, 2, or 3 at this time. CART: classification and regression tree.

make a preliminary judgment of the TMD to which cluster they belonged, with the accuracy rate mostly above 80%. It will be even more easily and quickly when our category system is applied in cephalometric software with artificial intelligent analysis. In this CART model, the critical values of 8 key morphological indicators identified to distinguish among these three clusters were observed great similarity with the critical points of the cephalometrics in orthodontics. For example, the critical value of SN-MP was 35° in the CART model, which was also the critical point for distinguishing whether the mandibular plane is steep or not. The LL-EP = -1 mm in the CART model was the critical point for discriminating the retraction lower lip as well. The association reflected the accuracy and reliability of our study.

Several limitations still remained in our study. Firstly, the category system was only based on the morphological analysis, and the clinical symptoms were not involved in this system. This is because the study was a retrospective study under orthodontic background and the detailed clinical symptoms such as TMJ pain and others of the patients were not recorded. Thus, we will cooperate with the clinicians in the department of TMJ in the next step to supplement this system with clinical symptoms of TMD. Secondly, the study primarily proposed a new category

system for the profile morphology of TMD, which lacked clinical verification. Further studies will be needed to verify the reliability and validity of this category system. Despite these limitations, our research creatively classified TMD according to the lateral cephalometric radiographs, which made a step toward morphological understanding of TMD.

Conclusion

Our study primarily proposed a novel category system for the profile morphology of TMDs with 3 subgroups according to the cephalometric morphology, which dentists can easily recognize TMDs according to our CART model. This study may make a step toward the morphological understanding of TMD and benefit the management of the categorical treatment of TMD.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of West China School of Stomatology of Sichuan University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

RZ: conceptualization, methodology, software, investigation, formal analysis, writing, and original draft. Y-HZ: data curation, writing, and original draft. Z-HZ: visualization and investigation. P-DF: resources and supervision. JW: software and validation. XX: conceptualization, funding acquisition, resources, supervision, writing—review, and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants to XX from the Technology Innovation Project of Science and Technology Bureau of Chengdu (2022-YF05-01691-SN).

References

- Schiffman EL, Look JO, Hodges JS, Swift JQ, Decker KL, Hathaway KM, et al. Randomized effectiveness study of four therapeutic strategies for TMJ closed lock. *J Dent Res.* (2007) 86:58–63. doi: 10.1177/154405910708600109
- Rauch A, Körner A, Kiess W, Hirsch C, Schierz O. Relationship between age-dependent body constitution and temporomandibular joint sounds in adolescents. *J Clin Med.* (2020) 9:E3236–927. doi: 10.3390/jcm9123927
- Macri M, Murmura G, Scarano A, Festa F. Prevalence of temporomandibular disorders and its association with malocclusion in children: a transversal study. *Front Public Health.* (2022) 10:860833. doi: 10.3389/fpubh.2022.860833
- Festa F, Rotelli C, Scarano A, Navarra R, Caulo M, Macri M. Functional magnetic resonance connectivity in patients with temporomandibular joint disorders. *Front Neurol.* (2021) 12:629211. doi: 10.3389/fneur.2021.629211
- De Rossi SS, Greenberg MS, Liu F, Steinkeler A. Temporomandibular disorders: evaluation and management. *Med Clin North Am.* (2014) 98:1353–84. doi: 10.1016/j.mcna.2014.08.009
- Durham J, Newton-John TRO, Zakrzewska JM. Temporomandibular disorders. *Br Med J.* (2015) 350:h1154–62. doi: 10.1136/bmj.h1154
- Valesan LE, Da-Cas CD, Réus JC, Denardin ACS, Garanhan RR, Bonotto D, et al. Prevalence of temporomandibular joint disorders: a systematic review and meta-analysis. *Clin Oral Invest.* (2021) 25:441–53. doi: 10.1007/s00784-020-03710-w
- Li DTS, Leung YY. Temporomandibular disorders: current concepts and controversies in diagnosis and management. *Diagnostics.* (2021) 11:459. doi: 10.3390/diagnostics11030459
- Monaco A, Cattaneo R, Marci MC, Pietropaoli D, Ortu E. Central sensitization-based classification for temporomandibular disorders: a pathogenetic hypothesis. *Pain Res Manag.* (2017) 2017:5957076–88. doi: 10.1155/2017/5957076
- Pimenta de Silva Machado L, de Macedo Nery MB, de Góis Nery C, Leles CR. Profiling the clinical presentation of diagnostic characteristics of

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.1045815/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Confusion matrix for CART model. Color represents the proportion of number in reference cluster. CART, classification and regression tree.

a sample of symptomatic TMD patients. *BMC Oral Health.* (2012) 12:26–33. doi: 10.1186/1472-6831-12-26

11. Yap AU, Zhang MJ, Lei J, Fu K-Y. Accuracy of the fonseca anamnestic index for identifying pain-related and/or intra-articular temporomandibular disorders. *Cranio J Craniomandib Pract.* (2021) 16:1–8. doi: 10.1080/08869634.2021.1954375

12. Schiffman E, Ohrbach R, Truelove E, Look J, Anderson G, Goulet J-P, et al. Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: recommendations of the international RDC/TMD consortium network* and orofacial pain special interest group†. *J Oral Facial Pain Headache.* (2014) 28:6–27. doi: 10.11607/jop.1151

13. Bond EC, Mackey S, English R, Liverman CT, Yost O. *Temporomandibular Disorders: Priorities for Research and Care.* Washington, DC: National Academies Press (2020).

14. Yan Z-B, Wan Y-D, Xiao C-Q, Li Y-Q, Zhang Y-Y, An Y, et al. Craniofacial morphology of orthodontic patients with and without temporomandibular disorders: a cross-sectional study. *Pain Res Manag.* (2022) 2022:1–11. doi: 10.1155/2022/9344028

15. Brand JW, Nielson KJ, Nanda RH, Currier GF, Owen WL. Lateral cephalometric analysis of skeletal patterns in patients with and without internal derangement of the temporomandibular joint - PubMed. *Am J Orthod Dentofacial Orthop.* (1995) 107:121–8. doi: 10.1016/S0889-5406(95)70126-5

16. Hwang C-J, Sung S-J, Kim S-J. Lateral cephalometric characteristics of malocclusion patients with temporomandibular joint disorder symptoms. *Am J Orthod Dentofacial Orthop.* (2006) 129:497–503. doi: 10.1016/j.ajodo.2004.12.019

17. Manfredini D, Segù M, Arveda N, Lombardo L, Siciliani G, Alessandro Rossi null, Guarda-Nardini L. Temporomandibular joint disorders in patients with different facial morphology: a systematic review of the literature. *J Oral Maxillofac Surg Off.* (2016) 74:29–46. doi: 10.1016/j.joms.2015.07.006

18. Yu J, Li H, Liu D. Modified immune evolutionary algorithm for medical data clustering and feature extraction under cloud computing environment. *J Healthc Eng.* (2020) 2020:1051394–404. doi: 10.1155/2020/1051394

19. Mariampillai K, Granger B, Amelin D, Guiguet M, Hachulla E, Maurier F, et al. Development of a new classification system for idiopathic inflammatory myopathies based on clinical manifestations and myositis-specific autoantibodies. *JAMA Neurol.* (2018) 75:1528–37. doi: 10.1001/jamaneurol.2018.2598
20. Li C, Cai Y, Chen S, Chen F. Classification and characterization of class iii malocclusion in Chinese individuals. *Head Face Med.* (2016) 12:31. doi: 10.1186/s13005-016-0127-8
21. Chen C-Z, Wang L-Y, Ou C-Y, Lee C-H, Lin C-C, Hsiue T-R. Using cluster analysis to identify phenotypes and validation of mortality in men with COPD. *Lung.* (2014) 192:889–96. doi: 10.1007/s00408-014-9646-x
22. Solow B, Tallgren A. Natural head position in standing subjects. *Acta Odontol Scand.* (1971) 29:591–607. doi: 10.3109/00016357109026337
23. Xiong X, Huang Y, Liu W, Wu Y, Yi Y, Wang J. Distribution of various maxilla-mandibular positions and cephalometric comparison in Chinese skeletal class II malocclusions. *J Contemp Dent Pract.* (2020) 21:822–8. doi: 10.5005/jp-journals-10024-2897
24. Kumar D, Bezdek JC. Visual approaches for exploratory data analysis: a survey of the visual assessment of clustering tendency (VAT) family of algorithms. *IEEE Syst Man Cybern Mag.* (2020) 6:10–48. doi: 10.1109/MSMC.2019.2961163
25. Brian H, Skellam JG. A new method for determining the type of distribution of plant individuals. *Annal Bot.* (1954) 18:213–27. doi: 10.1093/oxfordjournals.aob.a083391
26. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. *J Stat Softw.* (2014) 61:1–36. doi: 10.18637/jss.v061.i06
27. Lê S, Josse J, Husson F. FactoMineR : an R package for multivariate analysis. *J Stat Softw.* (2008) 1:25. doi: 10.18637/jss.v025.i01
28. Devereux L, Moles D, Cunningham SJ, McKnight M. How important are lateral cephalometric radiographs in orthodontic treatment planning? *Am J Orthod Dentofacial Orthop.* (2011) 139:e175–81. doi: 10.1016/j.ajodo.2010.09.021
29. Paço M, Duarte JA, Pinho T. Orthodontic treatment and craniocervical posture in patients with temporomandibular disorders: an observational study. *Int J Environ Res Public Health.* (2021) 18:3295–306. doi: 10.3390/ijerph18063295
30. Derwich M, Mitus-Kenig M, Pawlowska E. Is the temporomandibular joints' reciprocal clicking related to the morphology and position of the mandible, as well as to the sagittal position of lower incisors? A case-control study. *Int J Environ Res Public Health.* (2021) 18:4994–5007. doi: 10.3390/ijerph18094994
31. Bueno CH, Pereira DD, Pattussi MP, Grossi PK, Grossi ML. Gender differences in temporomandibular disorders in adult populational studies: a systematic review and meta-analysis. *J Oral Rehabil.* (2018) 45:720–9. doi: 10.1111/joor.12661
32. Bair E, Gaynor S, Slade GD, Ohrbach R, Fillingim RB, Greenspan JD, et al. Identification of clusters of individuals relevant to temporomandibular disorders and other chronic pain conditions: the OPPERA study. *Pain.* (2016) 157:1266–78. doi: 10.1097/j.pain.0000000000000518
33. Xiong X, Zhang Q, Liu Y. Correlations between mandibular ramus height and occlusal planes in Han Chinese individuals with normal occlusion: a cross-sectional study. *APOS Trends Orthod.* (2022) 11:295–300. doi: 10.25259/APOS_78_2021
34. Mollabashi V, Heidari A, Ebrahimi Zadeh H, Seyed Tabib M. The study of facial morphology in patients with vertical growth pattern (hyperdivergent) lacking or showing temporomandibular disorders symptoms. *J Stomatol Oral Maxillofac Surg.* (2020) 121:233–7. doi: 10.1016/j.jormas.2019.10.001
35. Ahn S-J, Kim T-W, Nahm D-S. Cephalometric keys to internal derangement of temporomandibular joint in women with Class II malocclusions. *Am J Orthod Dentofacial Orthop.* (2004) 126:486–94. doi: 10.1016/j.ajodo.2003.08.029
36. Ahn S-J, Baek S-H, Kim T-W, Nahm D-S. Discrimination of internal derangement of temporomandibular joint by lateral cephalometric analysis. *Am J Orthod Dentofacial Orthop.* (2006) 130:331–9. doi: 10.1016/j.ajodo.2005.02.019
37. Byun E-S, Ahn S-J, Kim T-W. Relationship between internal derangement of the temporomandibular joint and dentofacial morphology in women with anterior open bite. *Am J Orthod Dentofacial Orthop.* (2005) 128:87–95. doi: 10.1016/j.ajodo.2004.01.028
38. Sakar O, Çalisir F, Öztas E, Marsan G. Evaluation of the effects of temporomandibular joint disk displacement and its progression on dentocraniofacial morphology in symptomatic patients using lateral cephalometric analysis. *J Investig Clin Dent.* (2011) 29:211–8. doi: 10.1179/crn.2011.030
39. Gidarakou IK, Tallents RH, Kyrkanides S, Stein S, Moss ME. Comparison of skeletal and dental morphology in asymptomatic volunteers and symptomatic patients with unilateral disk displacement without reduction. *Angle Orthod.* (2003) 73:121–7. doi: 10.1043/0003-3219(2003)73%3C121:CO\$ADM%3E2.0.CO;2

Frontiers in Public Health

Explores and addresses today's fast-moving healthcare challenges

One of the most cited journals in its field, which promotes discussion around inter-sectoral public health challenges spanning health promotion to climate change, transportation, environmental change and even species diversity.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



Frontiers in Public Health

