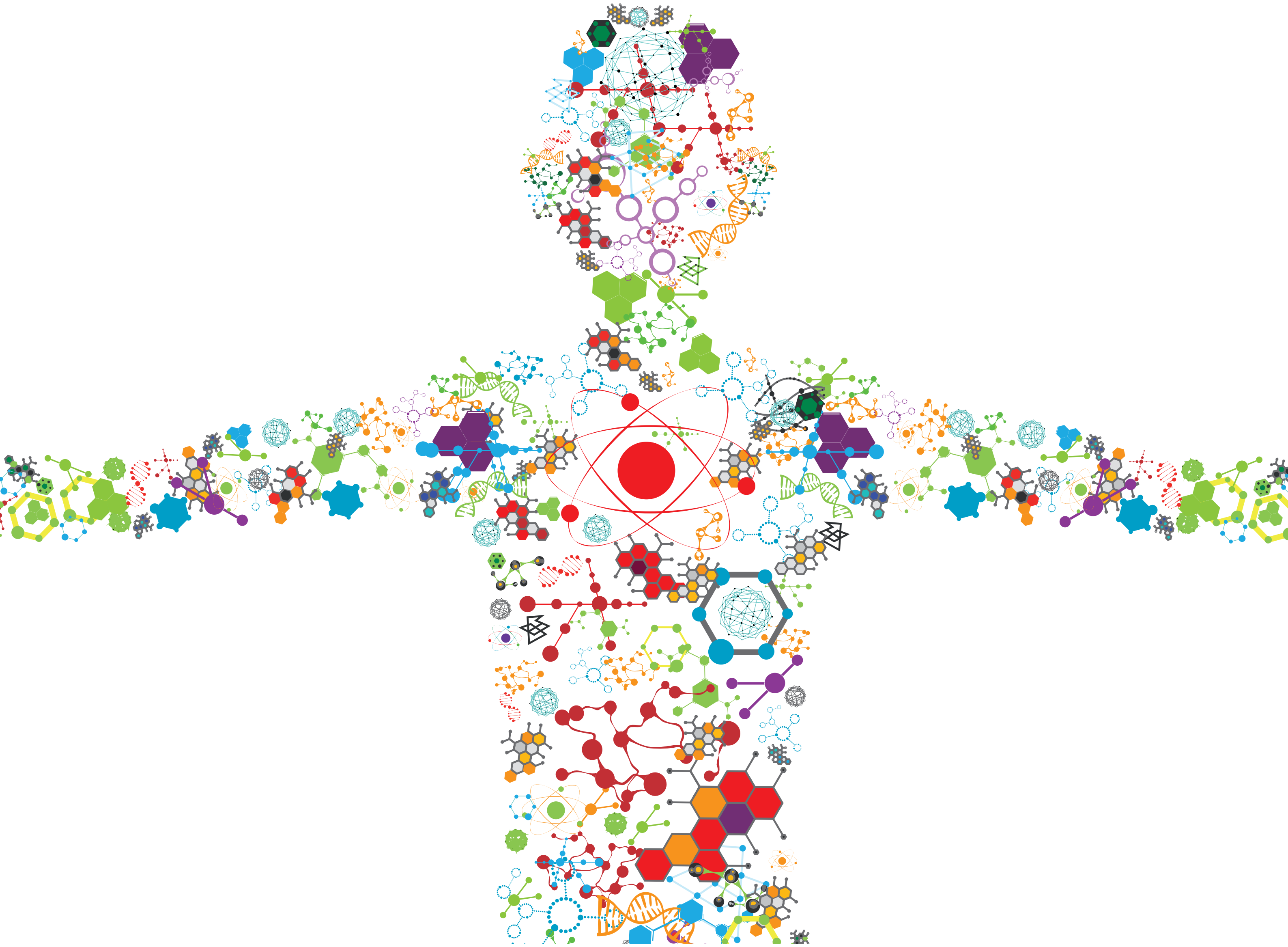
MACHINE LEARNING USED IN BIOMEDICAL COMPUTING AND INTELLIGENCE HEALTHCARE, VOLUME II

EDITED BY: Honghao Gao, Ying Li, Zijian Zhang and Wenbing Zhao

PUBLISHED IN: Frontiers in Bioengineering and Biotechnology,
Frontiers in Genetics and Frontiers in Public Health



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-247-7
DOI 10.3389/978-2-88976-247-7

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 USED IN BIOMEDICAL COMPUTING AND INTELLIGENCE HEALTHCARE, VOLUME II

Topic Editors:
Honghao Gao, Shanghai University, China
Ying Li, Zhejiang University, China
Zijian Zhang, The University of Auckland, New Zealand
Wenbing Zhao, Cleveland State University, United States

Citation: Gao, H., Li, Y., Zhang, Z., Zhao, W., eds. (2022). Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-247-7

Table of Contents

EDITORIAL
published: 09 May 2022
doi: 10.3389/fgene.2022.850667

Editorial: Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume II

Honghao Gao[1]*, *Ying Li*[2], *Zijian Zhang*[3] *and Wenbing Zhao*[4]

[1]*Shanghai University, Shanghai, China,* [2]*Zhejiang University, Hangzhou, China,* [3]*The University of Auckland, Auckland, Netherlands,* [4]*Cleveland State University, Cleveland, OH, United States*

Keywords: machine learning, biomedical computing, intelligence computing, healthcare, data processing and analysis

OPEN ACCESS

Edited and reviewed by:
M. Richard Xu,
Shanghai Jiao Tong University, China

****Correspondence:***
Honghao Gao
gaohonghao@shu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *08 January 2022*
Accepted: *08 April 2022*
Published: *09 May 2022*

Citation:
Gao H, Li Y, Zhang Z and Zhao W (2022) Editorial: Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume II. Front. Genet. 13:850667. doi: 10.3389/fgene.2022.850667

Editorial on the Research Topic

Machine Learning Used in Biomedical Computing and Intelligence Healthcare, Volume II

Biomedical intelligence, especially precision medicine, is considered as one of the most promising directions in healthcare development. However, these technologies have also brought new challenges and issues. In 2021, this research topic was supported by Frontiers again and included three collaborating journals, namely, Frontiers in Genetics, Frontiers in Public Health, and Frontiers in Bioengineering and Biotechnology. Ten papers were accepted for publication from 18 open submissions. A summary of these accepted papers is outlined below.

In the paper entitled "Surveillance Strategy for Barcelona Clinic Liver *Cancer* B Hepatocellular Carcinoma Achieving Complete Response: An Individualized Risk-Based Machine Learning Study" by Qi-Feng Chen et al., the authors proposed a retrospective, real-world study on surveillance strategies for Barcelona Clinical Liver *Cancer* stage B hepatocellular carcinoma (BBHCC) patients with complete response (CR) after curative treatment to support clinical decision making. The data in this paper were collected from the *Cancer* Center of Sun Yat-sen University, part of which were used as training dataset and the rest as validation dataset. The random survival forest method was applied to calculate the disease progression hazard per month, and follow-up schedules were arranged to maximize the capability of progression detection at each visit. The primary endpoint of the study was the delayed-detection months for disease progression. The proposed new surveillance schedule may provide a new perspective concerning follow-up for BBHCC patients with CR.

In the paper entitled "Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models" by Ashwini K et al., the authors presented a short-time Fourier transform (STFT) technique to transform the neonatal cry auditory signals into a spectrogram image. The deep convolutional neural network (DCNN) was used to extract features for neonatal cry classification with spectrogram images as input. Support vector machine (SVM) was used as the classifier based on the features automatically extracted. The study showed that the combination of using DCNN-based feature extraction and using the SVM classifier provided promising results. Specifically, the RBF kernel provided the highest classification accuracy of infant cry.

In the paper entitled "Identification of Key mRNAs as Prediction Models for Early Metastasis of Pancreatic *Cancer* Based on LASSO" by Ke Xue et al., the authors proposed a risk-scoring model to identify potentially robust predictors of metastasis through a minimal number of genes. Enrichment analysis of differential gene expression from multiple datasets was used to gain insight into the mechanism of pancreatic cancer metastasis. The study showed that six Epithelial-Mesenchymal Transition related genes can be used to reliably predict pancreatic cancer metastasis, assess clinical outcomes, and facilitate future personalized treatment for patients with ductal adenocarcinoma of the pancreas.

In the paper entitled "An Intelligent Control Model of Credit Line Computing in Intelligence Health-Care Systems" by Rong Jiang et al., the authors introduced a dynamic permission intelligent access control model that incorporated a credit line calculation to reduce the risk of patient privacy leakage when medical data are accessed. More specifically, the credit limit and credit interval according to the authorization rules were matched to control the access intelligently. The proposed method was validated using real patient data provided by a Grade-III Level-A hospital in Kunming, China.

In the paper entitled "Recurrence Risk of Liver *Cancer* Post-hepatectomy Using Machine Learning and Study of Correlation With Immune Infiltration" by Xiaowen Qian et al., the authors introduced an mRNA-based model to predict the risk of recurrence after hepatectomy for liver cancer and explore the relationship between immune infiltration and the risk of recurrence after hepatectomy. This paper investigated gene expression profiles of liver cancer patients, and selected 18 mRNAs as biomarkers for predicting the risk of recurrence of liver cancer using a machine learning method. The authors evaluated the immune infiltration of the samples and conducted a joint analysis of the recurrence risk of liver cancer. These findings are helpful for early detection, intervention, and the individualized treatment of patients with liver cancer after surgical resection. The scientific novelty in this paper is that it uses Machine Learning to forecast the recurrence risk of liver cancer post-hepatectomy.

In the paper entitled "Two-stage Deep Neural Network via Ensemble Learning for Melanoma Classification" by Jiaqi Ding et al., the authors presented an ensemble method that can integrate different types of classification networks for melanoma classification. U-net was used to segment the lesion area of images to generate a lesion mask, thus resizing images to focus on the lesion. In addition, a squeeze-excitation block was added to models to emphasize the informative features. Then, five classifiers were used to classify dermoscopy images. Finally, the proposed ensemble network was used to integrate the classification results from the five classifiers. The proposed classification framework was validated using the ISCI 2017 challenge dataset with good result. The scientific novelty in this paper is that it uses Deep Neural Network to make full use of the rich and deep feature information of images for melanoma classification.

In the paper entitled "Combining Polygenic Risk Score and Voice Features to Detect Major Depressive Disorders" by Yazheng Di et al., the authors proposed a biomarker of major depressive disorders (MDD) by combining the polygenic risk scores (PRSs) and voice features. Data in this study were collected from 3,580 women with recurrent MDD and 4,016 healthy people. PRS was constructed as a gene biomarker by p value-based clumping and thresholding. Voice features were extracted using the i-vector method. n the paper entitled "Contextualizing Genes by Using Text-Mined Co-Occurrence Features for *Cancer* Gene Panel Discovery" by Hui-O Chen et al., the authors introduced a pipeline that can contextualize genes by using text-mined co-occurrence features. Biomedical Natural Language Processing (BioNLP) techniques were used for literature mining in the cancer gene panel. The produced cancer gene panel was validated with the mutational landscape of different cancer types. The receiver operating characteristic (ROC) curve analysis confirmed that the neural net model has a better prediction performance. The key insight is that the use of text-mined co-occurrence features can contextualize each gene. This study examined several existing gene panels and demonstrated that part of the gene panel set can be used to predict the remaining genes for cancer discovery.

In the paper entitled "Classification of Diabetic Foot Ulcers Using Class Knowledge Banks" by Yi Xu et al., the authors presented a method that used class knowledge banks (CKBs) consisting of trainable units to effectively extract and represent class knowledge, and to better utilize the knowledge in the training data. Each unit in a CKB is used to compute similarity with a representation extracted from an input image. The averaged similarity between units in the CKB and the representation can be regarded as the logit of the considered input. In this way, the prediction depended not only on input images and trained parameters in networks, but also on the class knowledge extracted from the training data and stored in the CKBs. The experimental results showed that the proposed method effectively improved the performance of diabetic foot ulcers infection and ischemia classifications.

\In the paper entitled "Chronological Age Prediction: Developmental Evaluation of DNA Methylation-Based Machine Learning Models" by Haoliang Fan et al., the authors proposed a blood epigenetic clock in Southern Han Chinese (CHS) for chronological age prediction with machine learning algorithms. The correlation coefficient was analyzed for the experimental individuals to select five genes from a candidate set of nine age-associated DNA methylation (DNAm) biomarkers. The DNAm-based profiles of the CHS cohort were generated by the bisulfite targeted amplicon pyrosequencing (BTA-pseq) from 34 cytosine-phosphate-guanine sites of five selected genes. These four chronological age prediction models were evaluated using several machine learning algorithms, including stepwise regression, support vector regression, and random forest regression. The random forest regression appeared to achieve the best performence with a median absolute deviation of 1.15 years for the targeted cohort.

In conclusion, we would like to thank all the authors who submitted their original articles to our Research Topic. We highly appreciate the contributions of the reviewers for their suggestive comments. We would also like to acknowledge the guidance from the Editor-in-Chief and staff members of Frontiers.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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.

frontiers
in Public Health

ORIGINAL RESEARCH
published: 10 June 2021
doi: 10.3389/fpubh.2021.670352

Deep Learning Assisted Neonatal Cry Classification *via* Support Vector Machine Models

Ashwini K [1], P. M. Durai Raj Vincent [1], Kathiravan Srinivasan [1] and Chuan-Yu Chang [2]**

[1] School of Information Technology and Engineering, Vellore Institute of Technology (VIT), Vellore, India, [2] Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan

Neonatal infants communicate with us through cries. The infant cry signals have distinct patterns depending on the purpose of the cries. Preprocessing, feature extraction, and feature selection need expert attention and take much effort in audio signals in recent days. In deep learning techniques, it automatically extracts and selects the most important features. For this, it requires an enormous amount of data for effective classification. This work mainly discriminates the neonatal cries into pain, hunger, and sleepiness. The neonatal cry auditory signals are transformed into a spectrogram image by utilizing the short-time Fourier transform (STFT) technique. The deep convolutional neural network (DCNN) technique takes the spectrogram images for input. The features are obtained from the convolutional neural network and are passed to the support vector machine (SVM) classifier. Machine learning technique classifies neonatal cries. This work combines the advantages of machine learning and deep learning techniques to get the best results even with a moderate number of data samples. The experimental result shows that CNN-based feature extraction and SVM classifier provides promising results. While comparing the SVM-based kernel techniques, namely radial basis function (RBF), linear and polynomial, it is found that SVM-RBF provides the highest accuracy of kernel-based infant cry classification system provides 88.89% accuracy.

Keywords: convolutional neural network, infant cry classification, short time fourier transform, support vector machine, spectrogram

OPEN ACCESS

Edited by:
Di Wu,
Chongqing Institute of Green and Intelligent Technology (CAS), China

Reviewed by:
Ye Yuan,
Chongqing Institute of Green and Intelligent Technology (CAS), China
Olarik Surinta,
Mahasarakham University, Thailand

****Correspondence:***
P. M. Durai Raj Vincent
pmvincent@vit.ac.in
Chuan-Yu Chang
chuanyu@yuntech.edu.tw

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

Received: *05 March 2021*
Accepted: *04 May 2021*
Published: *10 June 2021*

Citation:
K A, Vincent PMDR, Srinivasan K and Chang C-Y (2021) Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. Front. Public Health 9:670352. doi: 10.3389/fpubh.2021.670352

INTRODUCTION

Babies convey their needs through cries. Experienced baby care persons and parents can understand the reason for the baby's cries. Some young working parents struggled to interpret the baby's cries. The baby's cries imply their emotions, physical needs, and pathological problems from internal or external stimulation. Humans can listen to the audio signal in the frequency range from 50 to 15,000 Hz for music, 20 to 20,000 Hz for sounds, and 100 to 4,500 Hz for speech. Within this range, humans can discriminate the audio. Babies do not have control over their vocal tract so that it is more sensitive than adults. Baby cries contain information, and their crying pattern varies based on their physical and emotional state. The researchers found that there is a pattern for each kind of cry. Infant cry classification can be considered pattern recognition or speech recognition. An abnormal cry of the infant can indicate a genetic or pathological problem. Childcare experts can differentiate it. The baby cry-based recognition approach will help us know the infant's feelings from their cries. Techniques such as signal preprocessing, feature extraction, feature selection, and classification are the steps involved in baby cry classification.

Signal preprocessing is crucial to eliminate the unwanted signal present in the audio signals. The audio signal features can be analyzed based on their time, frequency, and time-frequency domain. Neural networks can be able to learn features from the audio itself. The spectral representation of audio plays a crucial role in the classification of audio signals using neural networks. The initial work of infant cry classification was started in the 1960s. Kia et al. (1) designed a system to detect a baby's cries using fast Fourier transform with a fuzzy classifier. Petroni et al. (2) attempted to distinguish the baby's anger, fear, and pain cries. The features were extracted from the Mel cepstrum coefficient, and four kinds of neural networks such as time-delay network, feed-forward network, cascade network, and recurrent network were implemented; the results showed that a fully connected neural network gave a better performance.

Mima and Arakawa (3) examined the frequency analysis of infant cries (hunger, discomfort, sleepiness) and found the difference in Fourier transform tendencies for each state. Jam and Sadjedi (4) carried out work to distinguish the pain and normal infant cries. They observed that, while processing the audio signal, silence elimination, filtering, pre-emphasizing was crucial. Mel frequency and entropy based on multibands were used in the extraction of features. Principal component analysis reduces the dimensions of the feature vector. Multilayer perceptron recognized the infant cries and achieved better results because multiband entropy provides entropy distribution in the spectrum. The work was focused on the linear and the non-linear feature coefficient technique to detect and classify the normal and hearing impaired infant cries. The linear feature coefficients were extracted from linear predictive coefficient (LPC), and those features were optimized by using the hereditary approach. The bilinear nilpotent technique was used to analyze the non-linear signal. Kernel discriminant analysis (KDA) transforms those features into a low-dimensional basis to show the linear and non-linear features' contribution. Support vector machine (SVM) and expectation-maximization (EM) algorithms over an expert system were employed to classify the data. It shows that non-linear feature with an expert system-based classification approach gives better performance (5).

In previous works, the audio signal involves numerous preprocessing techniques; feature extraction and feature selection techniques were used to classify the data. However, deep learning approaches automatically extract the raw data features, even without additional preprocessing methods. Implementing the deep learning approach needs millions of data samples to get the best results. Moreover, this motivates us to enhance the infant cry classification model's performance even with the small dataset by extracting the features using the deep learning technique and classifying the infant cries using a machine learning algorithm with less computational complexity. This work classifies the most common infant cries such as hunger, pain, and sleepiness.

RELATED WORKS

It is a crucial task to discriminate the infant cries, so in this work (6), dealt with K-NN classifier with features such as short-time energy, harmonic to average power ratio (HAPR), Mel frequency coefficient, and harmonicity factor (HF) to recognize the infant cry sounds. In this work (7), convolutional restricted Boltzmann machine was used to analyze the unsupervised auditory filter banks. The network consists of the visible and hidden layers, and the weights were shared between those layers. The non-linear activation of Noisy Leaky Rectifier Linear Unit (NLReLU) was used. The parameters of the network were optimized by using the Adam optimization method. Convolutional restricted Boltzmann machine and discrete cosine transform were applied to reduce the feature dimensions. Those features were compared with MFCC features, and it was found that CNBM-based feature performs well in the discrimination of healthy and pathological auditory cries. In this case (8), they employed a convolutional neural network in infant cry vocalizations. The cry segments were manually extracted from the audio signal and segmented into a 4–8-s duration of segments. Audio signals were represented as spectrogram through short-time Fourier transform, which is based on Fourier transform. The spectrogram is the input for convolutional neural network. The convolution layer can obtain the features from the spectrogram, and the network can successfully discriminate the baby cry vocalizations.

This study (9) investigated the feature extracted from wavelet packet transform based on complex dual-tree form to discriminate the three sets of infant cries such as normal vs. asphyxia, normal vs. deaf, and hunger vs. pain. Various feature selection techniques such as correlation feature selection, principal component analysis, and information gain were applied to select the most relevant and essential features. Extreme machine learning can successfully classify infant cry patterns. This work (10) presented the combined acoustic and prosodic features to distinguish the audio signal's variations. Merge those features and generate a feature matrix for the deep neural network. MFCC features were considered to present the acoustic features. The features such as fundamental frequency, intensity, and formats carry the prosodic feature information. The neural network has an input, two hidden, and an output layer to calculate the weighted prosodic features. Those features were taken as input to the deep learning approach, which is found that the merged features can distinguish the variation present in infant cry signals.

Priscilla Dunstan found that every baby makes certain sounds while crying to convey their needs, such as Owh Heh, Eh, Eair, and Neh, representing tired, discomfort, burp or sleepy, pain, and hunger. Dewi et al. (11) analyze the feature extraction techniques such as linear frequency cepstral coefficient and Mel frequency cepstral coefficient. It extracts the features from the spectrogram. Vector quantization, KNN, and neural network were deployed in the classification of infant cries. It is found that LFCC with KNN classifier gives a better result than other techniques. Felipe et al. (12) discussed the motivation concerning the classification of infant cries. The local visual features such as binary pattern, robust binary pattern, phase quantization, Mel frequency cepstrum coefficient, Mel scale features, and constant Q chromogram were considered to obtain the features from the spectrogram. The best result was obtained from the local binary pattern using SVM with an accuracy of 71.68%. Gujral

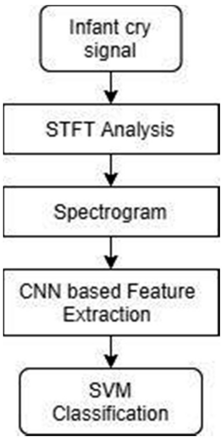


FIGURE 1 | Work flow diagram for baby cry classification.

et al. (13) analyzed and fine tuned the neural network using the transfer learning approach to recognize infant cries. Long short-time memory (LSTM) and convolutional neural network (CNN) were analyzed with and without transfer learning. It is observed that transfer learning based CNN outperforms LSTM with an average recall of 75.7%. The latent factor approach can efficiently obtain the information from high-dimensional and sparse data. A multilayered and randomized latent factor model was adopted to reduce the time complexity and enhance data representation for better understanding. In the case of nonnegative data, β divergence latent factor model is adopted to analyze the performance in recommender systems (14–17).

MATERIALS FOR FEATURE EXTRACTION AND CLASSIFICATION

In our approach, the infant cry signals are taken as input, and short-time Fourier transform (STFT) is deployed to convert the neonatal cry signals into the spectrogram image. The features are extracted from the image using a deep convolutional neural network. Furthermore, the SVM classifier discriminates the neonatal cries as pain, hunger, and sleepiness. **Figure 1** represents the procedure involved in the infant cry classification system.

Audio Signal Analysis

The frequency content of the audio signal varies by time. For that, a standard technique is required to analyze the signal in time-frequency domain. Fourier transform is deployed to characterize the time-varying frequency content present in the signal. It analyzes the signal by converting the domain of time to frequency. As a result of applying fast Fourier transform (FFT), the signal phase and magnitude are obtained. The FFT length of the signal must be equal to two times the power of length needed to get a good frequency resolution in FFT. The STFT simultaneously examines the time and frequency content of the signal. The signals are breakdown into numerous segments called frames; then, each segment multiplied with a window either with or without overlapping. Spectral leakage will happen while converting the signal from time into a frequency domain. The windowing function tries to reduce the spectral leakage and unusual discontinuity in the signal due to segmentation. Several windowing functions are there, such as Hamming, Blackman, uniform, flattop, and exponential. STFT computes the Fourier transform for each windowing segment. The magnitude square of the STFT is spectrogram. It represents the distribution of frequencies present in the signal changes over time (18–21). Short-time Fourier transform and spectrogram are mathematically described as follows.

$$F(m,\omega) = \sum x(n)\, w(n-m)^{e^{-j\omega n}}$$

$$S(m,\omega) = |F(m,\omega)|^2$$

whereas, $F(m, \omega)$ is the short-time Fourier transform, $x(n)$ represents a signal, $w(n–m)$ represents the windowing technique, and $S(m, \omega)$ is the spectrogram. **Figure 2** illustrates the process involved in audio signal analysis.

Convolutional Neural Network

The classical neural network comprises input, hidden, and output layers (22). The data has passed from one layer to another layer. Every layer has several nodes; each node takes a set of values from previous as input and does some mathematical operation that produces a single value as an output to the consecutive layer's nodes. In convolutional network, the image itself is taken as an input and generates an image as an output. It breaks the image into features and detects the particular pattern from that image.

Furthermore, it comprises convolution, pooling, and a fully connected layer. The convolution layer performs convolution with the help of filters. Each node has its own filters, and it extracts the features from the image. It is succeeded by a non-linear function (RELU, sigmoid, tanh) that performs threshold operation. Pooling tends to minimize network complexity. Max pooling and average pooling are types of pooling. It gathers the part of images into small rectangular portions and examines the highest values in max pooling. In average pooling, it calculates the median value of that specific rectangular portion. The rectangular regions are the kernels. The fully connected layer carries information about the number of output classes, and it maps the data to its output. The softmax layer normalizes the data and gives the score a probability for that input to every output data. The classification layer produces a result based on the probability score (23, 24). **Figure 3** shows the simple convolutional neural network.

Support Vector Machine

SVM executes classification by differentiating the data points with a larger margin using hyperplane as a decision boundary. SVM classifier includes hyperplane, margin hyperplane, kernels, and soft margin. The hyperplane is the line that differentiates the discrete data points. Margin is the distance between data samples and the hyperplane. The margin hyperplane divides the dissimilar data with maximum distance from one another. The data samples which are near the hyperplane are named as support

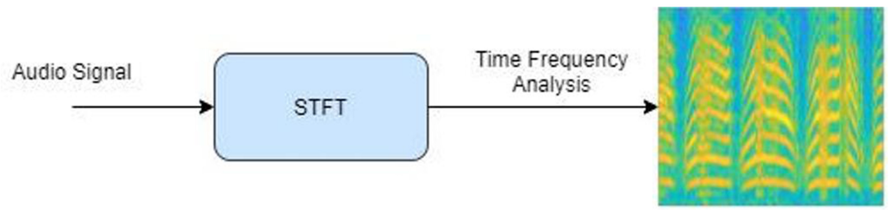


FIGURE 2 | Audio signal transform into spectrogram.

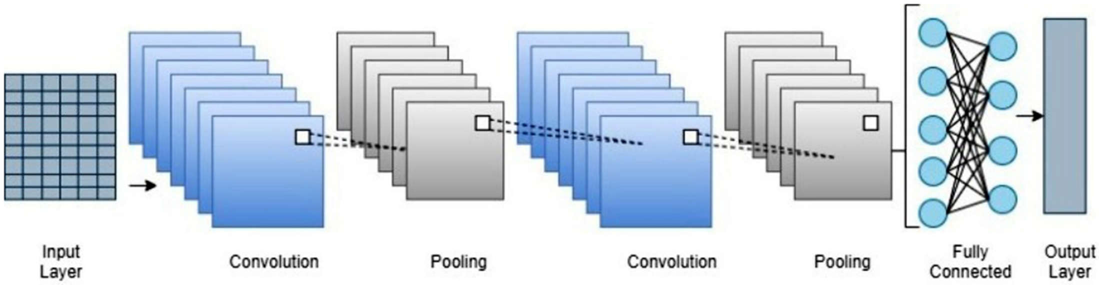


FIGURE 3 | Pictorial representation of convolutional neural network.

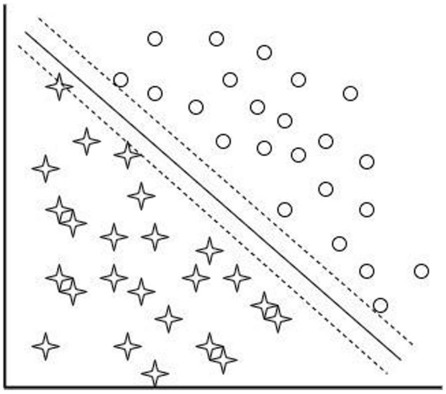


FIGURE 4 | Pictorial representation of linear SVM.

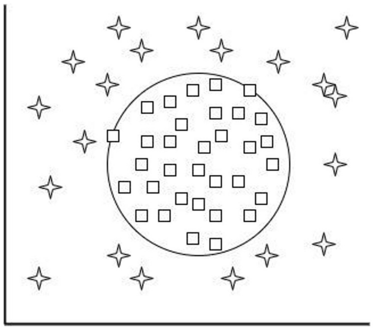


FIGURE 5 | Pictorial representation of non-linear SVM.

vectors. Soft margin in SVM creates accurate models from data that are not able to generalize. The purpose of kernel function converts the data samples into high dimensional feature space. Linear SVM and non-linear SVM are the categories involved in SVM. **Figure 4** shows the simple linear SVM model. In linear SVM, the data points can be distinguishable with a simple straight line. It can be defined as

$$w \cdot a + b = 0$$

where, "w" represents the adjustable weight, "a" defines the input data, and "b" indicates the bias.

When the data samples are not separable using a straight line, non-linear SVM is used to solve the non-linear problems. For that, the kernel functions are used to modify the data into higher feature space. The most common kernel functions are linear, quadratic, polynomial, and radial basis function (RBF) kernel. Linear kernel combines all support vectors linearly to produce the output. It can be described as

$$G\left(x_i, x_j\right) = {x_i}' x_j$$

The quadratic kernel does not require any changes in the parameter to get an efficient result. The polynomial kernel considers each support vector and computes its kernel function. In the polynomial kernel, the polynomial order is usually chosen

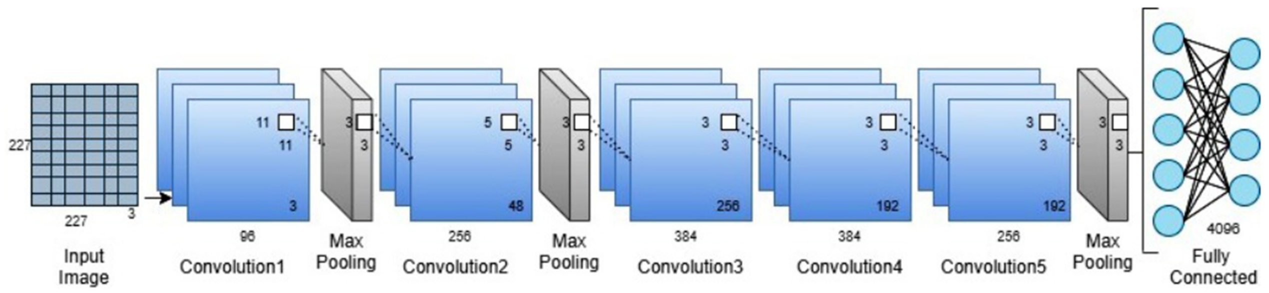


FIGURE 6 | Convolutional neural network for feature extraction.

by more than one. If the polynomial order is one, then it will become a linear kernel. It is mathematically defined as

$$G\left(x_i, x_j\right) = (x_i{'} x_j + 1)^p$$

RBF kernel can effectively generalize the data and perform better to solve the practical problem. In the RBF kernel, the support vectors can automatically determine the number of RBF and its centers. It can be represented as

$$G\left(x_i, x_j\right) = \exp(-\left\|x_i - x_j\right\|^2)$$

where, x_i and x_j represent the observations and p represents the polynomial order. SVM optimization can be performed by increasing the margin space between data samples and selecting the precise kernel function for our system demands (25, 26). **Figure 5** shows the non-linear SVM model.

RESULTS AND DISCUSSION

The dataset of pain, hunger, and sleepiness cries was collected from the infants born in National Taiwan University Hospital Yunlin Branch, Taiwan (27, 28). The infants' cries were recorded from the healthy infants' age range from 1 to 10 days. There were no pathological problems or any complications found in the babies, even during birth and after birth. In this study, we consider 300 audio records, in that every 100 audio samples for hunger, pain, and sleepiness cries. Each audio signal is in the length of 4 s data with a sampling frequency of 8 kHz. Eighty percent of the data is utilized for training, and the remaining data is used for testing. The whole experiment is implemented in MATLAB using the Deep Learning Toolbox and Statistics and Machine Learning Toolbox. The convolutional neural network-based feature extraction process is shown in **Figure 6**.

At first, the neonatal cry auditory signals are transformed into spectrogram images by applying the short-time Fourier transform. The audio signals are broken down into numerous segments called frames; then, the windowing function multiplies with each frame. In this study, the auditory signals are segmented into 128 sections with 64 windows overlapping. The Hamming window function is deployed here. Fourier transform is computed; for that purpose, 256 discrete Fourier transform points are considered. The magnitude square of the STFT gives the spectrogram image. **Figure 7** illustrates the overall process involved in this study.

The data augmented is done to resize the image into 227*227*3 to meet the pretrained deep convolutional network's requirement. The convolutional neural network comprises eight layers, five convolution layers succeeded by RELU activation layer and pooling, and three fully connected layers. The deep network breaks the images into features using multiple sets of layers. The foremost layer of the network is taking the input data and normalizes the input image. The first layer of convolution has 96 filters; each filter size is 11*11 with four strides and zero padding, succeeded by RELU activation and max pooling of size 3*3 with a stride of two. The second layer of convolution comprises 256 filters, 5*5 filter size with one stride, succeeded by RELU and max pooling, in which 3*3 pooling size with two strides and zero paddings. The third layer of convolution has 384 filters, 3*3 filter size with one stride and one padding succeeded by RELU. The fourth layer of convolution consists of 384 filters, 3*3 size of filters with one stride succeeded by the RELU layer. The last layer of convolution has 256 filters, 3*3 size of filters with one stride succeeded by RELU and max pooling, in which 3*3 pooling size with two stride and zero paddings. The convolution layer description is shown in **Table 1**. We get the features from a fully connected layer instead of convolution, making it easier to execute the model with crucial features. The obtained features from the convolutional network are fed into the machine learning classifier. SVM with several kernel techniques such as polynomial, linear, and radial basis function is implemented to discriminate the baby cries. We have used error correcting output code (ECOC) approach with a one vs. one coding design to train the multiclass SVM model. In this case, we have considered three kinds of baby cries, for that the approach yields three binary learners which use all combinations of infant cries and return a multiclass model. To avoid overfitting or underfitting, we had cross validated the model using 10-fold cross validation, efficiently estimating the model with conventional variance.

The confusion matrix analyzes the performance of the approach. It has attributes such as true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The 3*3 confusion matrix is defined in **Table 2** where, $\mathrm{TP_A}$ defines the

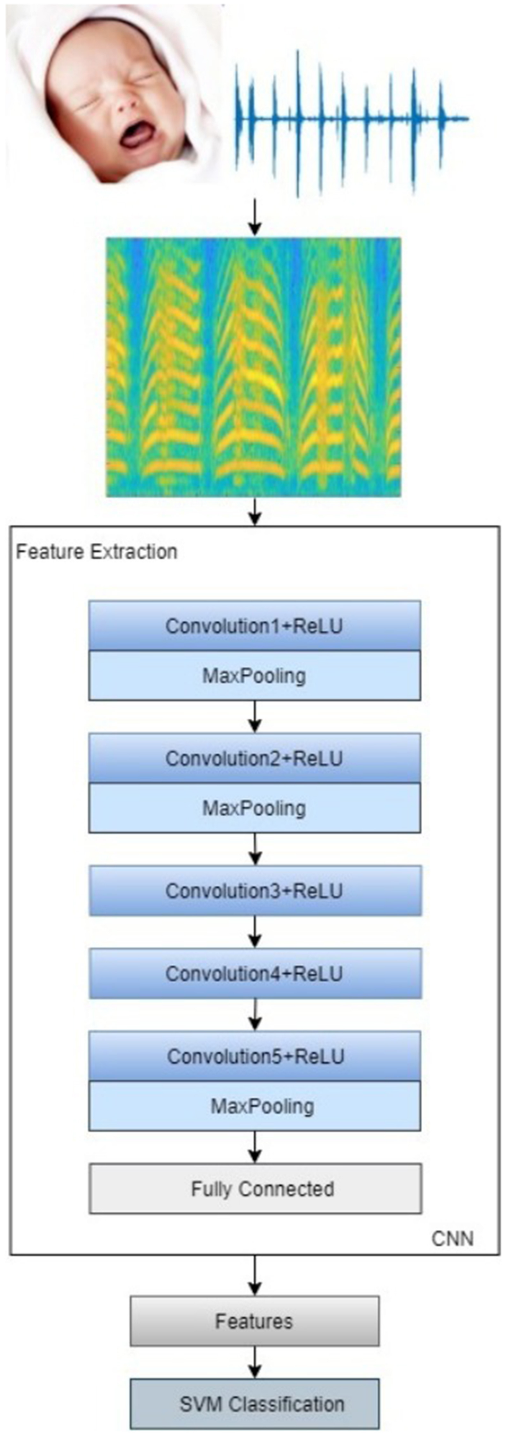


FIGURE 7 | Proposed infant cry classification system.

TABLE 1 | Convolution layer description of the network.

Layer number	Layers	Number of filters	Filter size	Number of channels
1	Conv1	96	11*11	3
2	Conv2	256	5*5	48
3	Conv3	384	3*3	256
4	Conv4	384	3*3	192
5	Conv5	256	3*3	192

TABLE 2 | 3*3 confusion matrix.

	A	B	C
A	TP_{A}	E_{AB}	E_{AC}
B	E_{BA}	TP_{B}	E_{BC}
C	E_{CA}	E_{CB}	TP_{C}

number of samples that are classified as class A. TP_{B} represents the number of data samples that are correctly recognized in class B. TP_{C} describes the number of data samples that are precisely classified in class C. E_{AB} represents the number of data samples from class A which are mislabeled as class B. E_{AC} describes the number of samples from class A that are misinterpreted as class C. E_{BA} represents the number of data samples from class B which are mislabeled as class A. E_{BC} describes the number of data samples from class B that are mislabeled as class C. E_{CA} represents the number of data samples from class C which are misinterpreted as class A. E_{CB} defines the number of data samples from class C that are misinterpreted as class B. For class A, false negative (FN_{A}), false positive (FP_{A}), and true negative (TN_{A}) can be calculated as

$$\begin{aligned}
\mathrm{FP}_{\mathrm{A}} &= E_{\mathrm{AB}} + E_{\mathrm{AC}} \\
\mathrm{FN}_{\mathrm{A}} &= E_{\mathrm{BA}} + E_{\mathrm{CA}} \\
\mathrm{TN}_{\mathrm{A}} &= E_{\mathrm{BC}} + E_{\mathrm{CB}} + \mathrm{TP}_{\mathrm{B}} + \mathrm{TP}_{\mathrm{C}}
\end{aligned}$$

For class B, false negative (FN_{B}), false positive (FP_{B}), and true negative (TN_{B}) can be calculated as

$$\begin{aligned}
\mathrm{FP}_{\mathrm{B}} &= E_{\mathrm{BA}} + E_{\mathrm{BC}} \\
\mathrm{FN}_{\mathrm{B}} &= E_{\mathrm{AB}} + E_{\mathrm{CB}} \\
\mathrm{TN}_{B} &= E_{\mathrm{CA}} + E_{\mathrm{AC}} + \mathrm{TP}_{\mathrm{A}} + \mathrm{TP}_{\mathrm{C}}
\end{aligned}$$

For class C, false negative (FN_{C}), false positive (FP_{C}), and true negative (TN_{C}) can be calculated as

$$\begin{aligned}
\mathrm{FP}_{\mathrm{C}} &= E_{\mathrm{CA}} + E_{\mathrm{CB}} \\
\mathrm{FN}_{\mathrm{C}} &= E_{\mathrm{AC}} + E_{\mathrm{BC}} \\
\mathrm{TN}_{\mathrm{C}} &= E_{\mathrm{AB}} + E_{\mathrm{BA}} + \mathrm{TP}_{\mathrm{A}} + \mathrm{TP}_{\mathrm{B}}
\end{aligned}$$

Those are used to compute the performance metrics such as precision, accuracy, recall, F1 score, and specificity. Accuracy compares the actual and desired output. Specificity shows the proportions of all negative cases, and recall represents the proportions of all positive cases. Precision shows the proportions

TABLE 3 | Performance evaluation of SVM-RBF.

Performance metrics	Hunger	Pain	Sleepy	Average measures
Specificity	0.8571	0.8235	1.0000	0.8935
Sensitivity	0.9032	0.9643	0.9677	0.9450
Precision	0.8000	0.9333	0.9333	0.8888
Accuracy	0.8889	0.9111	0.9778	0.9259
F1 score	0.8276	0.8750	0.9655	0.8893

TABLE 4 | Performance evaluation of SVM-polynomial.

Performance metrics	Hunger	Pain	Sleepy	Average measures
Specificity	0.8125	0.8750	0.9231	0.8702
Sensitivity	0.9310	0.9655	0.9063	0.9342
Precision	0.8667	0.9333	0.8000	0.8666
Accuracy	0.8889	0.9333	0.9111	0.9111
F1 score	0.8387	0.9032	0.8571	0.8663

TABLE 5 | Performance evaluation of SVM-linear.

Performance metrics	Hunger	Pain	Sleepy	Average measures
Specificity	0.8125	0.8571	0.8667	0.8454
Sensitivity	0.9310	0.9032	0.9333	0.9225
Precision	0.8667	0.8000	0.8667	0.8444
Accuracy	0.8889	0.8889	0.9111	0.8963
F1 score	0.8387	0.8276	0.8667	0.8443

of positive which are actually positive. F measure/F1 score computes the mean of recall and precision.

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

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

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

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

$$\text{F1 Score} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})}$$

Tables 3–5 represent the performance metrics such as specificity, sensitivity, precision, accuracy, and F1 score of infant cries based on SVM-RBF, polynomial, and linear kernels.

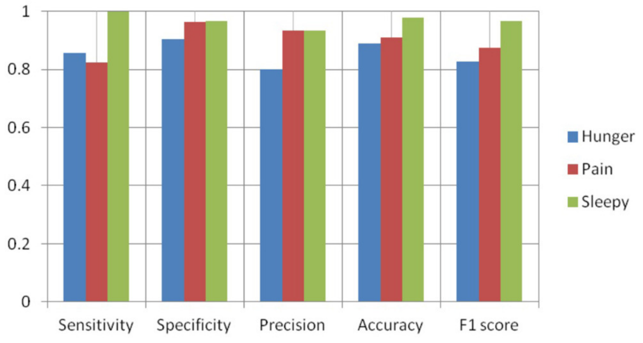


FIGURE 8 | Performance measures of SVM-RBF.

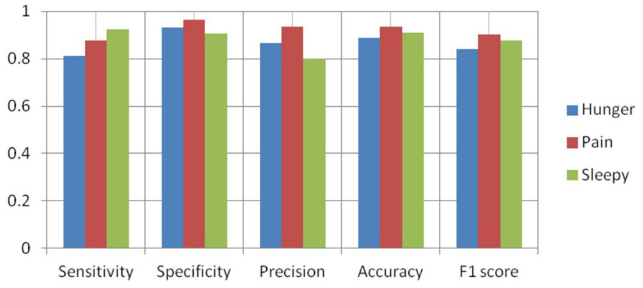


FIGURE 9 | Performance measures of SVM-polynomial.

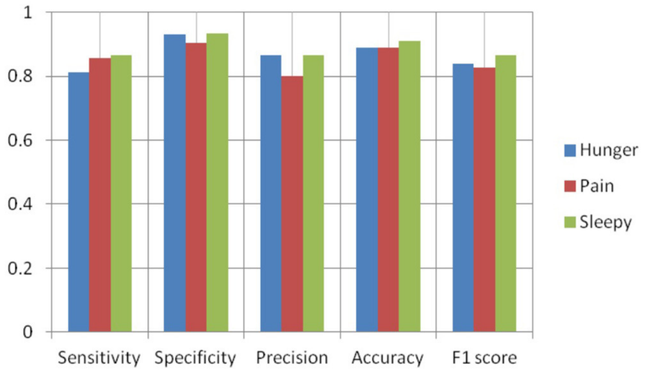


FIGURE 10 | Performance measures of SVM-linear.

Figures 8–10 show the CNN-SVM-based infant cry classification model's performance measures with various kernel functions. It is clearly shown that SVM with RBF performs better than other kernel functions. Overall, the deep convolutional network-based feature extraction and SVM with the RBF classification-based model by the parameters of $c = 1$ and gamma = 1 provides the highest accuracy of 88.89% with a generalized classification error of 5.56% and standard deviation of 0.0835.

Receiver operating characteristics (ROC) curve illustrates the correlation between the true-positive rate (TPR) and false-positive rate (FPR). It is a crucial tool to estimate the performance of the approach. **Figures 11–13** represent the ROC curve for SVM-based polynomial, linear, and RBF kernel. The area under the curve for the polynomial kernel is 90.3%, the linear kernel is 87.9%, and the RBF kernel is 91.9%.

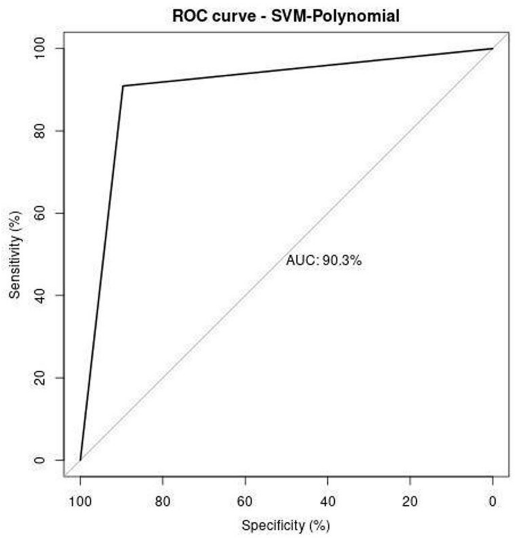


FIGURE 11 | ROC analysis of SVM-polynomial.

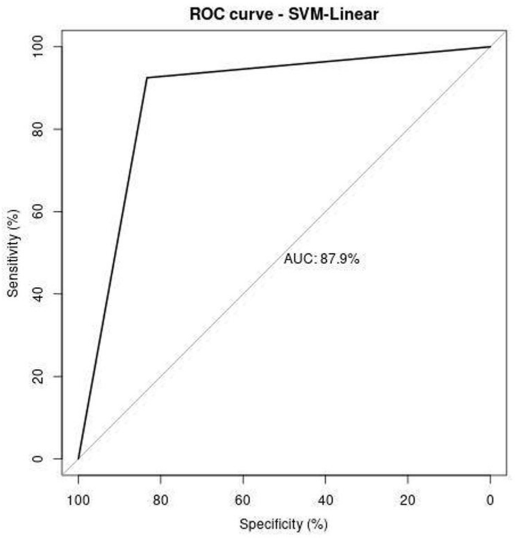


FIGURE 12 | ROC analysis of SVM-linear.

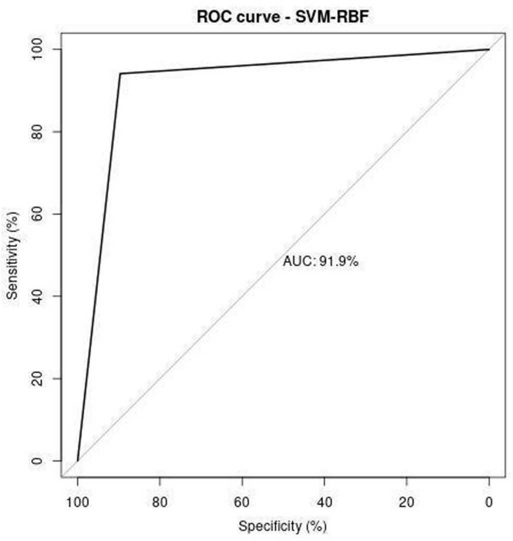


FIGURE 13 | ROC analysis of SVM-RBF.

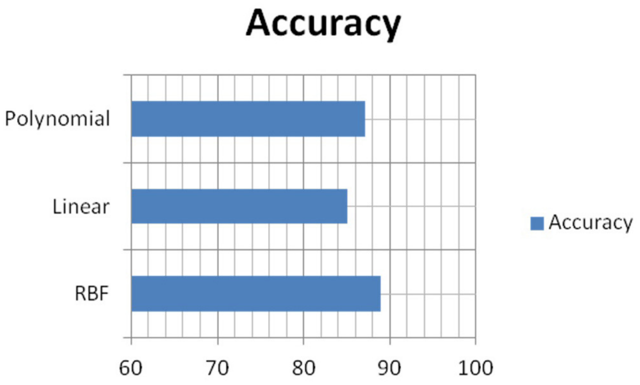


FIGURE 14 | Comparison of various kernels in SVM.

Figure 14 shows the comparison of overall accuracy obtained from various kernel functions in SVM. SVM-polynomial, linear, and RBF kernel's accuracy is 86.67, 84.44, and 88.89%. It clearly shows that the performance of the SVM-RBF kernel gives more accurate results than other kernel functions. We observe that by varying the kernel functions in SVM, the classification system's performance changes drastically.

In the SVM polynomial kernel, the 3rd and 4th order of polynomial gives an accuracy of 86.67%, the 5th order acquires 84.45%, and the 6th polynomial order provides 82.22%. It is observed that the variations in the polynomial order affect the performance of the model. By increasing the polynomial order, the system's performance (accuracy) gradually decreases. In the classification of infant cries' physiological needs, the pretrained network, which uses stochastic gradient descent, acquires 76% accuracy, and stochastic gradient descent with momentum obtained 82% accuracy (29). It is also compared with convolutional neural network feature extraction based on other machine learning techniques such as KNN, Naïve Bayes, and Decision Tree, which acquire 84.69, 83.56, and 84.45% accuracy. While comparing this CNN architecture with another pretrained CNN architecture which comprises 13 convolution layers followed by three fully connected layers, the features were extracted based on those layers. They passed the features to the SVM classifier, which gives 87.22% accuracy, respectively. MFCC was used to analyze the infant cry audio signal, which acquires an accuracy of 85.76%. It is found that STFT outperformed baseline MFCC. Compared with these, the proposed approach gives better performance than the existing approach concerning infant cry classification and the SVM classifier performs better than KNN and Naïve Bayes. It is observed that the time taken to train the pretrained network takes more time than the convolutional feature extraction-based machine learning classification. The neonatal cry classification model helps the new parents to know their infants once they discover the need for baby cries. They can respond to their baby's needs more quickly and effectively.

CONCLUSION

Infant cries carry information about the infant's feelings. This study combines the deep learning and machine learning model to enhance the infant cry classification model's efficiency even with small datasets. The audio cry signals are converted into a spectrogram image using STFT. The spectrogram images are fed into the deep convolutional network. The convolutional network is good at extracting features from images. The extracted features are taken as input for the SVM technique. The experimental result exhibits that the proposed method acquires the highest classification accuracy of 88.89% compared with all other approaches considered in the literature. It is found that CNN can extract the features from the time-frequency representation of audio signals. To the best of the authors' knowledge, the demonstration of CNN feature extraction and machine learning classifier is reported for the first time in infant cry classification. Convolutional feature extraction-based machine learning classifier provides good results even with the moderate dataset, but tuning the SVM technique's hyperparameters is computationally expensive. In the future, we would like to experiment with this deep neural network feature extraction with hybrid or embedded machine learning based classifiers. Also, much more focus will be given to implementing the machine learning model's optimization techniques, which may enhance these approaches' efficiency.

DATA AVAILABILITY STATEMENT

The original contributions generated for this 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 Human Research Ethics Committee at National Cheng Kung University, Taiwan. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

C-YC: conceptualization, resources, project administration, and funding acquisition. PV and AK: methodology and software. KS and C-YC: validation and writing—review and editing. PV: writing—original draft preparation. All authors contributed to the article and approved the submitted version.

FUNDING

This research was partially funded by the Intelligent Recognition Industry Service Research Center from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. Grant number: N/A and the APC were funded by the aforementioned Project.

REFERENCES

1. Kia M, Kia S, Davoudi N, Biniazan R. A detection system of infant cry using fuzzy classification including dialing alarm calls function. In:*Second International Conference on the Innovative Computing Technology (INTECH 2012)*. Casablanca: IEEE (2012) p. 224–9. doi: 10.1109/INTECH.2012.6457776
2. Petroni M, Malowany AS, Johnston CC, Stevens BJ. Classification of infant cry vocalizations using artificial neural networks (ANNs). In: *International Conference on Acoustics, Speech, Signal Processing*. Detroit: IEEE (1995) p. 3475–8. doi: 10.1109/ICASSP.1995.479734
3. Mima Y, Arakawa K. Cause estimation of younger babies' cries from the frequency analyses of the voice-Classification of hunger, sleepiness, and discomfort. In*: International Symposium on Intelligent Signal Processing and Communications.* Yonago: IEEE (2006) p. 29–32. doi: 10.1109/ISPACS.2006.364828
4. Jam MM, Sadjedi H. A system for detecting of infants with pain from normal infants based on multi-band spectral entropy by infant's cry analysis. In: *Second International Conference on Computer and Electrical Engineering*. Dubai: IEEE (2009) p. 72–6. doi: 10.1109/ICCEE.2009.164
5. Peralta-Malváez L, López-Rincón O, Rojas-Velazquez D, Valencia-Rosado LO, Rosas-Romero R, Etcheverry G. Newborn cry nonlinear features extraction and classification. *J Intell Fuzzy Syst.* (2018) 34:3281–9. doi: 10.3233/JIFS-169510
6. Bano S, RaviKumar KM. Decoding baby talk: a novel approach for normal infant cry signal classification. In: *International Conference on Soft-Computing and Networks Security (ICSNS).* Taipei: IEEE (2015) p. 1–4. doi: 10.1109/ICSNS.2015.7292392
7. Sailor HB, Patil HA. Auditory filterbank learning using ConvRBM for infant cry classification. In*: INTERSPEECH*. (2018) p. 706–10. doi: 10.21437/Interspeech.2018-1536
8. Anders F, Hlawitschka M, Fuchs M. Automatic classification of infant vocalization sequences with convolutional neural networks. *Speech Commun.* (2020) 119:36–45. doi: 10.1016/j.specom.2020.03.003
9. Lim WJ, Muthusamy H, Vijean V, Yazid H, Nadarajaw T, Yaacob S. Dual-tree complex wavelet packet transform and feature selection techniques for infant cry classification. *J Telecommun Electron Comput Eng.* (2018) 10:75–9. Available online at: https://jtec.utem.edu.my/jtec/article/view/4098
10. Ji C, Xiao X, Basodi S, Pan Y. Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features. In*: International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData).* IEEE (2019) p. 1233–40. doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00206
11. Dewi SP, Prasasti AL, Irawan B. The study of baby crying analysis using MFCC and LFCC in different classification methods. In*: IEEE International Conference on Signals and Systems (ICSigSys).* Bandung: IEEE. (2019) p. 18–23. doi: 10.1109/ICSIGSYS.2019.8811070

12. Felipe GZ, Aguiar RL, Costa YM, Silla CN, Brahnam S, Nanni L, et al. Identification of infants' cry motivation using spectrograms. In*: International Conference on Systems, Signals and Image Processing (IWSSIP).* Osijek: IEEE. (2019) 181–6. doi: 10.1109/IWSSIP.2019.8787318
13. Gujral A, Feng K, Mandhyan G, Snehil N, Chaspari T. Leveraging transfer learning techniques for classifying infant vocalizations. In: *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI),* IEEE. (2019) p. 1–4. doi: 10.1109/BHI.2019.8834666
14. Yuan Y, He Q, Luo X, Shang M. A Multilayered-and-Randomized Latent Factor Model for High-Dimensional and Sparse Matrices. *IEEE TransactBig Data*. (2020). doi: 10.1109/TBDATA.2020.2988778. [Epub ahead of print].
15. Yuan Y, Luo X, Shang MS. Effects of preprocessing and training biases in latent factor models for recommender systems. *Neurocomputing.* (2018) 275:2019–30. doi: 10.1016/j.neucom.2017.10.040
16. Yuan Y, Luo X, Shang M, Wu D. A generalized and fast-converging non-negative latent factor model for predicting user preferences in recommender systems. In*: Proceedings of The Web Conference.* (2020) p. 498–507. doi: 10.1145/3366423.3380133
17. Luo X, Yuan Y, Zhou M, Liu Z, Shang M. Non-negative latent factor model based on β-divergence for recommender systems. *IEEE Transact Syst.* (2019). doi: 10.1109/TSMC.2019.2931468. [Epub ahead of print].
18. Yuan Y, Xun G, Jia K, Zhang A. A multi-view deep learning method for epileptic seizure detection using short-time fourier transform. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, Health Informatics*. Boston (2017) p. 213–22. doi: 10.1145/3107411.3107419
19. Ouelha S, Touati S, Boashash B. An efficient inverse short-time Fourier transform algorithm for improved signal reconstruction by time-frequency synthesis: optimality and computational issues. *Digit Sign Proc.*(2017) 65:81–93. doi: 10.1016/j.dsp.2017.03.002
20. Decorsière R, Søndergaard PL, MacDonald EN, Dau T. Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations. *IEEE/ACM Transact Audio Speech Lang Proc.* (2014) 23:46–56. doi: 10.1109/TASLP.2014.2367821
21. Flandrin P. Time–frequency filtering based on spectrogram zeros. *IEEE Sign Proc Lett.* (2015) 22:2137–41. doi: 10.1109/LSP.2015.2463093
22. Sanchez-Riera J, Srinivasan K, Hua KL, Cheng WH, Hossain MA, Alhamid MF. Robust RGB-D hand tracking using deep learning priors. *IEEE Transact Circ Syst Video Technol.* (2017) 28:2289–301. doi: 10.1109/TCSVT.2017.2718622
23. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Into Imaging.* (2018) 9:611–29. doi: 10.1007/s13244-018-0639-9
24. Tu F, Yin S, Ouyang P, Tang S, Liu L, Wei S. Deep convolutional neural network architecture with reconfigurable computation patterns. *IEEE Transact Very Large Scale Integr Syst.* (2017) 25:2220–33. doi: 10.1109/TVLSI.2017.2688340
25. Scholkopf B, Smola AJ. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Adaptive Computation and Machine Learning Series*. Vienna: IEEE (2018). doi: 10.7551/mitpress/4175.001.0001
26. Alam S, Kang M, Pyun JY, Kwon GR. Performance of classification based on PCA, linear SVM, and Multi-Kernel SVM. In: *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN).* IEEE (2016) p. 987–9. doi: 10.1109/ICUFN.2016.7536945
27. Chang CY, Chang CW, Kathiravan S, Lin C, Chen ST. DAG-SVM based infant cry classification system using sequential forward floating feature selection. *Multidimension Syst Sign Proc.* (2017) 28:961–76. doi: 10.1007/s11045-016-0404-5
28. Chen ST, Srinivasan K, Lin C, Chang CY. Neonatal cry analysis and categorization system via directed acyclic graph support vector machine. In: *Big Data Analytics for Sensor-Network Collected Intelligence*. (2017) p. 205–22. doi: 10.1016/B978-0-12-809393-1.00010-6
29. Ashwini K, Durai Raj Vincent PM. A deep convolutional neural network based approach for effective neonatal cry classification. *Recent Adv Comput Sci Commun.* (2020). doi: 10.2174/2666255813999200710135408. [Epub ahead of print].

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.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 17 August 2021
doi: 10.3389/fbioe.2021.701039

Identification of Key mRNAs as Prediction Models for Early Metastasis of Pancreatic Cancer Based on LASSO

Ke Xue[1], Huilin Zheng[2], Xiaowen Qian[1], Zheng Chen[3], Yangjun Gu[4], Zhenhua Hu[5,3,6], Lei Zhang[2]* and Jian Wan[1]**

[1]Department of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China, [2]Department of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou, China, [3]Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, Fourth Affiliated Hospital, School of Medicine, Zhejiang University, Yiwu, China, [4]Shulan Hospital Affiliated to Zhejiang Shuren University Shulan International Medical College, Hangzhou, China, [5]Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, First Affiliated Hospital, School of Medicine, Key Laboratory of Combined Multi-Organ Transplantation, Ministry of Public Health Key Laboratory of Organ Transplantation, Zhejiang University, Hangzhou, China, [6]Division of Hepatobiliary and Pancreatic Surgery, Yiwu Central Hospital, Yiwu, China

Pancreatic cancer is a highly malignant and metastatic tumor of the digestive system. Even after surgical removal of the tumor, most patients are still at risk of metastasis. Therefore, screening for metastatic biomarkers can identify precise therapeutic intervention targets. In this study, we analyzed 96 pancreatic cancer samples from The Cancer Genome Atlas (TCGA) without metastasis or with metastasis after R0 resection. We also retrieved data from metastatic pancreatic cancer cell lines from Gene Expression Omnibus (GEO), as well as collected sequencing data from our own cell lines, BxPC-3 and BxPC-3-M8. Finally, we analyzed the expression of metastasis-related genes in different datasets by the Limma and edgeR packages in R software, and enrichment analysis of differential gene expression was used to gain insight into the mechanism of pancreatic cancer metastasis. Our analysis identified six genes as risk factors for predicting metastatic status by LASSO regression, including *zinc finger BED-Type Containing 2 (ZBED2), S100 calcium-binding protein A2 (S100A2), Jagged canonical Notch ligand 1 (JAG1), laminin subunit gamma 2 (LAMC2), transglutaminase 2 (TGM2), and the transcription factor hepatic leukemia factor (HLF)*. We used these six EMT-related genes to construct a risk-scoring model. The receiver operating characteristic (ROC) curve showed that the risk score could better predict the risk of metastasis. Univariate and multivariate Cox regression analyses revealed that the risk score was also an important predictor of pancreatic cancer. In conclusion, 6-mRNA expression is a potentially valuable method for predicting pancreatic cancer metastasis, assessing clinical outcomes, and facilitating future personalized treatment for patients with ductal adenocarcinoma of the pancreas (PDAC).

Keywords: pancreatic cancer, metastasis, EMT, bioinformatics, precision medicine

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Jessica Jiang,
Columbia University, United States
Xiaoxian Yang,
Shanghai Second Polytechnic University, China
Ying Zhong,
Harvard Medical School, United States

****Correspondence:***
Huilin Zheng
zhenghuilin@zust.edu.cn
Lei Zhang
papaver_rhoeas@126.com
Jian Wan
wanjian@zust.edu.cn

Specialty section:
This article was submitted to Biosafety and Biosecurity, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *27 April 2021*
Accepted: *09 August 2021*
Published: *17 August 2021*

Citation:
Xue K, Zheng H, Qian X, Chen Z, Gu Y, Hu Z, Zhang L and Wan J (2021) Identification of Key mRNAs as Prediction Models for Early Metastasis of Pancreatic Cancer Based on LASSO. Front. Bioeng. Biotechnol. 9:701039. doi: 10.3389/fbioe.2021.701039

Abbreviations: AUC, Area under Curve; DEGs, Differential expressed genes; GEO, Gene Expression Omnibus; GO, Gene Ontology; GSEA, Gene Set Enrichment Analysis; HR, Hazard ratio; LASSO, Least Absolute Shrinkage and Selection Operator; OS, Overall Survival; PC, Pancreatic Cancer; PDAC, Ductal Adenocarcinoma of Pancreas; PPI, Protein–protein interaction; RF, Random Forest; ROC, Receiver operating characteristic; SVM, Support Vector Machine; TCGA, The Cancer Genome Atlas.

INTRODUCTION

Pancreatic cancer (PC) typically progresses rapidly and tends to metastasize early in the course of the disease. Metastasis is the primary cause of its high mortality and low cure rate. Current cohort studies of metastasis are classified as follows: normal tissue and metastatic foci in metastatic tissue (Barry et al., 2013), primary tumors and metastases (McDonald et al., 2017), primary and para-cancerous tissues (Barry et al., 2013), and different metastatic foci to define metastases (Chaika et al., 2012). Only a few studies on postoperative recurrence and metastasis have been performed. We screened patients both without and with metastasis after clean postoperative tumor (R0) resection. We screened only patients with R0 to reduce the impact of clinical disturbances. The samples were collected during surgery and their RNA was sequenced.

The use of biomarkers to predict pancreatic cancer metastasis and prognosis has gained increasing attention from researchers and clinicians worldwide. Carbohydrate antigen199 (CA199) and carcinoembryonic antigen (CEA) are commonly used pancreatic cancer biomarkers (Chen et al., 2015b), although serum interleukin 6 family cytokine (*LIF*) is more effective than CA199 and CEA in predicting lymph node and distant metastasis (Jiang et al., 2021). An increasing number of studies have used machine-learning strategies to identify metastatic risk factors and predict risk. Using the adaptive least absolute shrinkage and selection operator (LASSO), Cox Boost and Elastic net, Zemmour compared the three algorithms and successfully identified patients with higher risk of breast cancer metastasis (Zemmour et al., 2015). Li et al. reported a random forest (RF) algorithm to identify biomarkers of pancreatic cancer (Li et al., 2018), although our own analyses revealed that the degree of model generalization is not sufficient. Other studies have used support vector machine (SVM) to predict colorectal cancer metastasis (Zhi et al., 2018). Although this classification algorithm can solve the problem of nonlinear classification, SVM is sensitive to the selection of parameters and kernel function and has a low efficiency. Moreover, the RF and SVM algorithms produce a binary 0–1 classification. Unlike RF and SVM, LASSO is a regression analysis that performs variable selection and combines the risk-scoring formula with the prognosis of pancreatic cancer, making it a suitable algorithm for this study.

The aim of this study was to identify potentially robust predictors of metastasis through a minimal number of genes. Based on a previous study of PDAC biomarkers, we integrated data from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases and our cell lines, BXPC-3 and BXPC-3-M8 (Owusu-Ansah et al., 2019). We then applied the LASSO regression model classification model to identify potential predictors associated with pancreatic cancer metastasis.

MATERIALS AND METHODS

Epithelial-Mesenchymal Transition Related Genes

To further demonstrate the role of EMT related genes in pancreatic cancer metastasis, we collected 994 EMT-related genes from published literature and database. (**Supplementary Table S1**).

Differential EMT Gene Expression in Pancreatic Cancer Metastasis

To identify EMT differential genes, we screened 994 EMT-related genes from the GSEA gene set (**Supplementary Table S1**). We then explored the relationship between EMT-related genes and pancreatic cancer metastasis. We included the public database TCGA human pancreatic cancer transcriptome sequencing data, our own cell lines BXPC-3 and BXPC-3-M8 transcriptome sequencing data, and two sets of microarray data from the public database GEO (GSE23952, GSE21654). For TCGA pancreatic adenocarcinoma data, only pathological samples obtained through R0 resection from patients for which survival time was available were selected. Survival time was defined as the period from surgery to death or to the end date of follow-up. Patients were grouped into two categories depending on tumor recurrence after the end of the treatment (patients with no tumor, or patients with metastatic tumors), which was determined by clinical follow-up. Samples from new or metastatic tumors were not included in this study. A total of 96 patients were screened, including 51 metastasis-free and 45 metastatic patients. All samples were sequenced before metastasis (**Figure 1**).

BxPC-3 is a human pancreatic cancer cell line from which the metastatic line BxPC-3-M8 has been derived. BxPC-3 and BxPC-3-M8 were kindly provided by Donghai Jiang (Owusu-Ansah et al., 2019). Total RNA was extracted using TRIZOL Reagent (Life Technologies) and purified using the RNAClean XP Kit (Beckman Coulter) and RNase-Free DNase Set (QIAGEN). For the BxPC-3 and BxPC-3-M8 cell lines, cDNA was generated using SuperScript II Reverse Transcriptase (Invitrogen). RNA-seq libraries were created using the VAHTS Stranded mRNA-seq Library Prep Kit (Illumina) on an Agilent 2,100 sequencer. Sample read quality was determined using FASTQC. The data volume was approximately 6G/sample, and the proportion of base quality greater than 20 (Q20) was not less than 90%. Finally, all libraries were sequenced using an Illumina Novaseq 6,000 sequencer (Illumina). Each sample was mapped to hg38 using Hisat2 (version 2.0.4). We applied Seqtk to filter unqualified reads. We removed reads containing linker sequences, with 3′ end quality Q less than 20 bases, length less than 25 reads, and ribosomal RNA reads of the species.

We also obtained two datasets from GEO: GSE23952 and GSE21654. In GSE23952, the TGFβ-induced group was regarded as the metastatic group and the control group as the free group. GSE21654 is a model of 22 epithelial and mesenchymal cell lines. Mesenchymal-like cell lines were used as the metastatic group, and epithelioid cell lines were used as the free group. We downloaded the processed probe matrix data of these two GEO data sets. We matched the expression data for 994 EMT-related genes in human samples and cell lines. Next, we analyzed the difference between the free and metastatic groups for each set of data. The Limma package of R Software 3.6.2 (https://www.r-project.org/) was used to process chip data, and the edgeR package of R software was used to screen the mRNAs differentially expressed between groups. The criteria |log fold

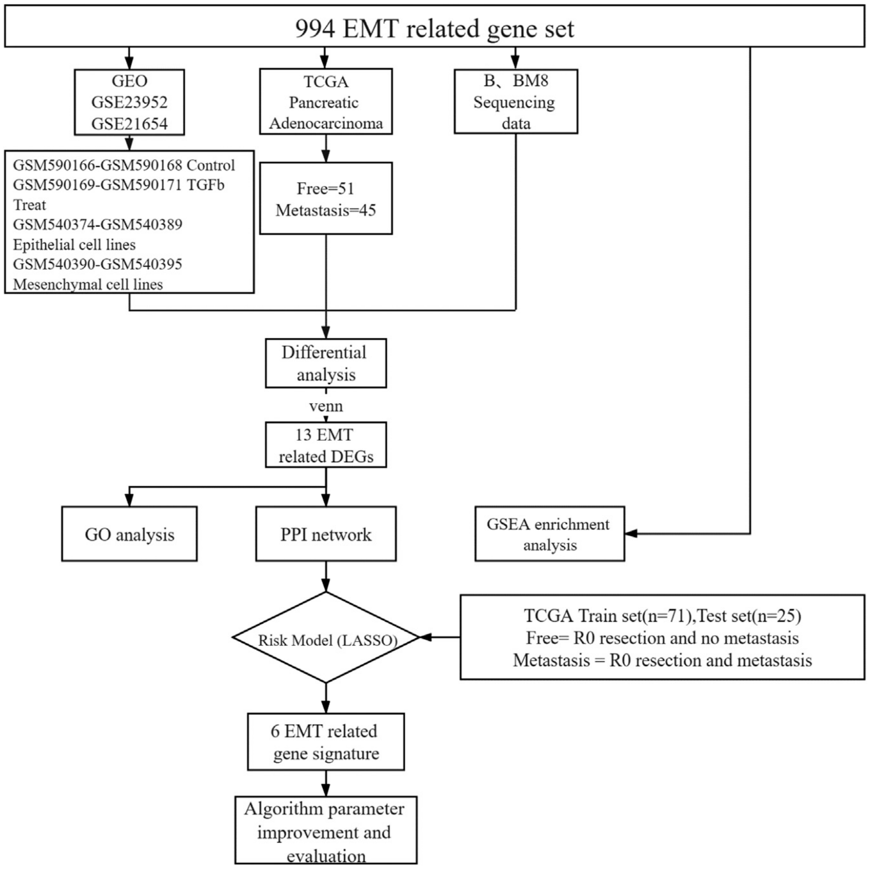


FIGURE 1 | Schematic representation of the process followed in this article.

change| > 1 and adjusted p value < 0.05 were set as threshold criteria. We extracted the overlapping differentially expressed genes (DEGs) from the GEO, BXPC-3, BXPC-3-M8, and TCGA datasets for subsequent analyses.

Protein-Protein Interaction Network Construction and Enrichment Analysis

Functional interactions between the 13 proteins were analyzed using Search Tool for the Retrieval of Interacting proteins database (STRING, http://string-db.org). PPI networks for the 13 genes retrieved were depicted using Cytoscape software 3.7.1 (http://www.cytoscape.org/). GO analysis was performed for 13 selected genes using DAVID (https://david.ncifcrf.gov/). Gene Set Enrichment Analysis (GSEA) software (http://software.broadinstitute.org/gsea/index.jsp) was used to perform enrichment analysis of all genes. The data sets c2.cp.kegg.v7.1.symbols.gmt, c2.cp.v7.1.symbols. gmt, and c5. all.v7.1.symbols. gmt were used as the reference gene sets. The selected threshold criteria were FDR <0.25, or p < 0.05.

LASSO Regression Model Construction

Thirteen genes were differentially expressed in all datasets analyzed. To construct a risk-score model for pancreatic cancer metastasis prediction, we developed risk scores using the LASSO regression algorithm, we chose the perfect penalty parameter λ associated with the minimum 10-fold cross-validation within the training set. Finally, six genes and their coefficients were defined by the minimum binominal deviance (Tibshirani, 1997; Goeman, 2010). The formula for the risk score was:

$$\boldsymbol{Risk\ Score} = \sum_{i=1}^{n} \left(\boldsymbol{Coef}_{\boldsymbol{i}} * \boldsymbol{x}_{\boldsymbol{i}}\right)$$

where $\boldsymbol{Coef}_{\boldsymbol{i}}$ is coefficient, and $\boldsymbol{x}_{\boldsymbol{i}}$ is the expression level of the corresponding gene in the sample.

Statistical Analysis

The “pheatmap” package in R software was used to generate heat maps. The Kaplan-Meier survival curve was analyzed using the log-rank test. The GEPIA website (http://gepia.cancer-pku.cn/) was used to analyze the survival of pancreatic cancer patients from TCGA. Univariate Cox regression was used to estimate the hazard ratio (HR) under different factors, and multivariate Cox regression analysis was used to analyze independent factors. The area under the ROC curve (AUC) is an accurate indicator in diagnostic tests (Obuchowski and Bullen, 2018), which is used to

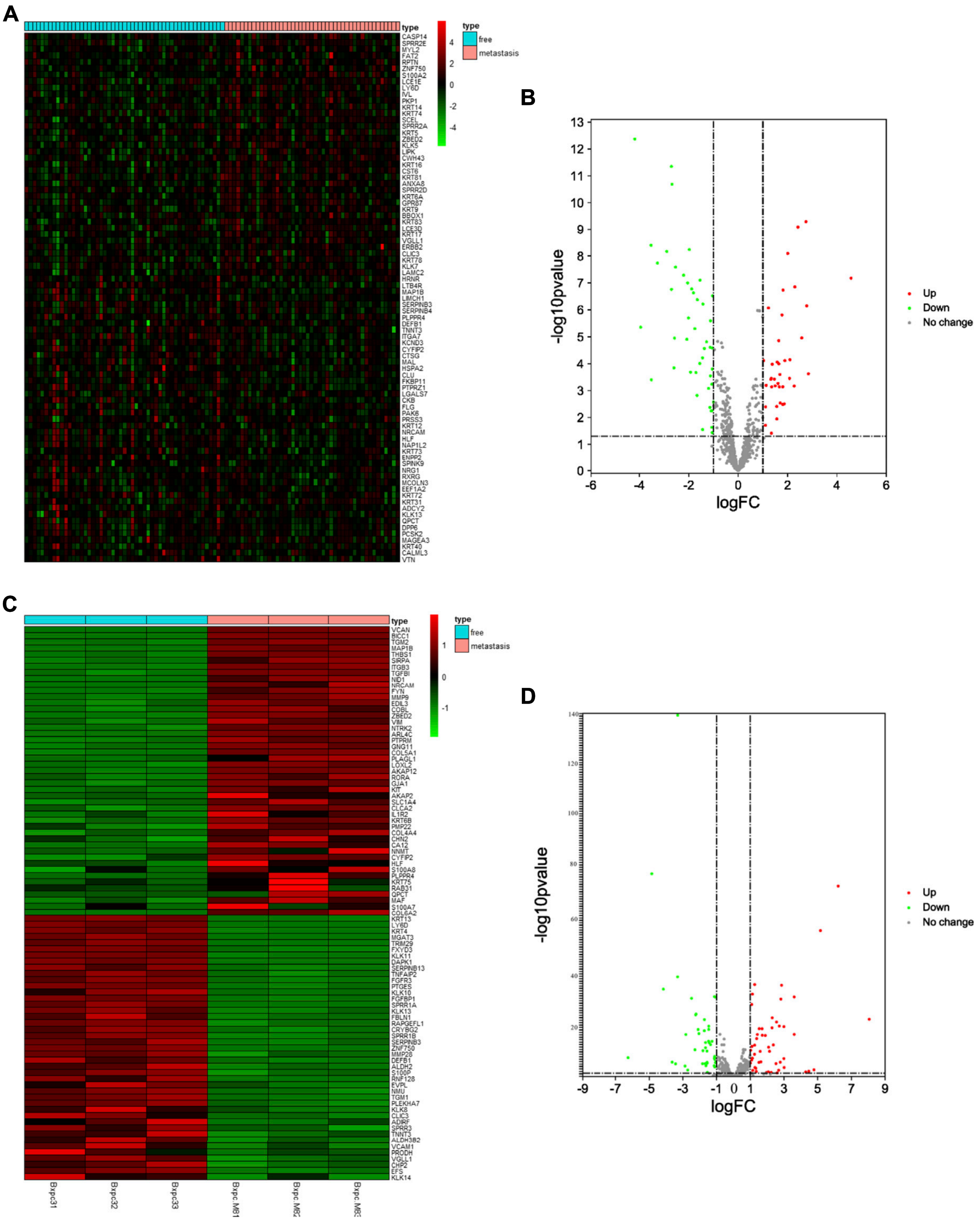


FIGURE 2 | Identification of Differentially Expressed EMT-related Genes **(A, C)** The heat map of DEGs in pancreatic cancer from TCGA dataset and BxPC-3-M8 and BxPC-3 cells (p-value < 0.05 and |log FC| > 1). Red color indicates up-regulated genes, and green indicates down-regulated genes **(B, D)** The volcano plot of DEGs in pancreatic cancer from TCGA dataset **(B)** and BxPC-3-M8 and BxPC-3 cells **(D)**. The red dots and green dots represent upregulated DEGs and downregulated DEGs with significance (p-value < 0.05 and |log FC| > 1), respectively. The gray dots are those DEGs without significance.

evaluate the quality of The model. R software and GraphPad Prism 7.0 software (https://www.graphpad.com/) were used for data analysis. Statistical significance was set at $p < 0.05$. We have added the code with proper instructions to the attached materials in the form of GitHub (https://github.com/xkeke77/Paper_Code).

RESULTS

Identification of Differentially Expressed EMT-Related Genes

Current research often makes it difficult to validate metastasis-associated candidate genes obtained through screening of clinical samples at the cellular level or for further mechanistic studies, or to validate metastasis-associated candidate genes obtained through screening at the cellular level in patients because of the presence of individual and cellular heterogeneity. To make the results of bioinformatics analysis more reliable, three data models were incorporated in this paper, including patient's sample data, metastasis-related cell line data, and highly metastatic cell lines from the same parental cells. We finally regressed to clinical significance with the patient's sample data as the most dominant baseline.

First, we selected the data of interest from TCGA database according to the criteria described in the Methods section, resulting in a total of 96 patients, and divided them into groups according to the occurrence of metastasis after R0 resection. The total sample size was divided in the training and test set with a ratio of 7:3, resulting in 71 samples in the training set and 25 samples in the test set. **Figure 1** shows a flowchart of the entire study. We analyzed transcriptomic data from TCGA pancreatic cancer patients to further investigate the differences between the free and metastatic pancreatic cancer groups. The results of the analyses are depicted as a heat map (**Figure 2A**) and as a volcanic map (**Figure 2B**). A total of 38 significantly upregulated genes and 45 significantly downregulated genes were identified in the data retrieve from TCGA. We then investigated the transcriptomic profile of BxPC-3-M8 and BxPC-3 cells, which were considered as models of metastatic-cancer and metastasis-free cancer conditions, respectively. The heat map of the DEGs (**Figure 2C**) and the volcano plot distribution map of the differentially expressed genes (**Figure 2D**) revealed that there were a total of 47 upregulated and 43 downregulated genes between the two conditions, all of which were statistically significant. Furthermore, we analyzed the differences between the free and metastatic pancreatic cancer groups using the transcriptomic profile of GSE23952 (model 1), which included data from the normal group and TGFβ-treated group, and GSE21654 (model 2), which included data from both epithelial and mesenchymal cell lines. Results from model 1 and 2 analyses are shown in **Supplementary Figures S1, S2**, respectively.

Gene Screening

To identify reliable predictors of metastasis, we selected genes whose expression was significantly different in at least two datasets, as well as consistent with the trend in the TCGA dataset. Therefore, the intersections of the differentially up- and downregulated genes in the TCGA pancreatic cancer dataset, BxPC-3 and BxPC-3-M8, model 1, and model 2 were obtained in pairs (**Figures 3A,B**). The Venn results showed that 25 differential genes were screened out, 17 differentially upregulated genes and eight differentially downregulated genes, respectively (**Supplementary Table S2**). Finally, 13 genes were extracted, consistent with the expression trends in TCGA pancreatic cancer data (**Table 1**).

Establishment of the Six mRNAs Metastatic Signature, and the Risk Score is a Good Predictor of Metastasis Performance

Next, we put 13 genes into the lasso model. The LASSO algorithm was applied to select the penalty coefficient according to the least squares deviation standard, and *ZBED2*, *S100A2*, *JAG1*, *LAMC2*, *TGM2*, and *HLF* were selected to construct the risk model. The risk score was calculated for each TCGA pancreatic cancer sample using a formula for gene coefficient and gene expression screened by LASSO (**Figure 4A**). The AUC of the training set was 0.711, and that of the test set was 0.729. These results indicated that the risk score model could be used as a classifier to predict the metastatic status of patients (**Figures 4B,C**). We also constructed classification models based on the RF and SVM algorithms for the selected pancreatic cancer samples. In the RF model, the AUC of the training and test sets was 0.682 and 0.629, respectively (**Supplementary Figures S3A,B**). In the SVM model, the AUC of the training and test sets was 0.708 and 0.706, respectively (**Supplementary Figures S3C,D**). These results indicate that the RF and SVM classification models were less efficient than the LASSO model, confirming its applicability in this study.

Potential Mechanisms Associated With Tumor Metastasis

To further investigate the function of the identified DEGs, we performed functional enrichment analysis on GO terms. The DEGs were enriched in EMT, cell adhesion, extracellular matrix organization, endothelial cell migration, blood vessel remodeling, chondrocyte differentiation, collagen fibril organization, multicellular organism development, basement membrane, and extracellular space (**Figures 5A,B**).

Then, we performed a correlation analysis to elucidate the interactions between the 13 genes that were involved in EMT. We found that there was a significant correlation between inducers and transcription factors in the same signal transduction pathway (**Supplementary Figure S4A**). Functional interactions between these 13 genes and others were determined using STRING (**Supplementary Figure S4B**). Next, GSEA was used to identify the processes associated with metastasis of pancreatic cancer, which included NOTCH signaling pathway, TGF-beta signaling pathway, positive regulation of calcium development exocytosis, regulation of epidermis development, cell adhesion

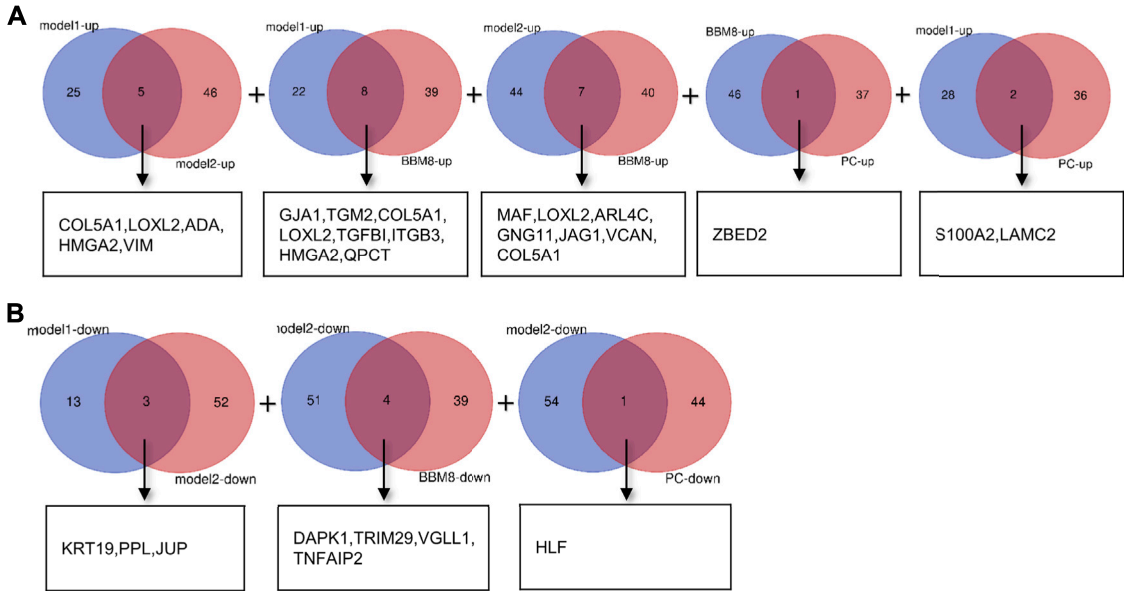


FIGURE 3 | Gene screening **(A)** Intersection of differentially upregulated genes of Model 1, Model 2, BxPC-3 and BxPC-3-M8, and TCGA pancreatic cancer samples, respectively. No differentially up-regulated genes were shared between Model 2 and TCGA pancreatic cancer samples **(B)** Intersection of differentially down-regulated genes of Model 1, Model 2, BxPC-3 and BxPC-3-M8, and TCGA pancreatic cancer samples, respectively. No differentially down-regulated genes were shared between Model 1 and BxPC-3, and BxPC-3-M8 and TCGA pancreatic cancer samples, and between BxPC-3 and BxPC-3-M8 and TCGA pancreatic cancer samples.

TABLE 1 | Selected 13 EMT genes in TCGA.

Gene symbol	Log FC	Regulation
S100A2	2.2929	up
ZBED2	1.7747	up
LAMC2	1.0317	up
TGM2	0.6931	up
HMGA2	0.6648	up
LOXL2	0.4358	up
TGFBI	0.4011	up
SNAI2	0.3758	up
ARL4C	0.2516	up
JAG1	0.1947	up
COL5A1	0.1133	up
ADA	0.0322	up
HLF	−1.8114	down

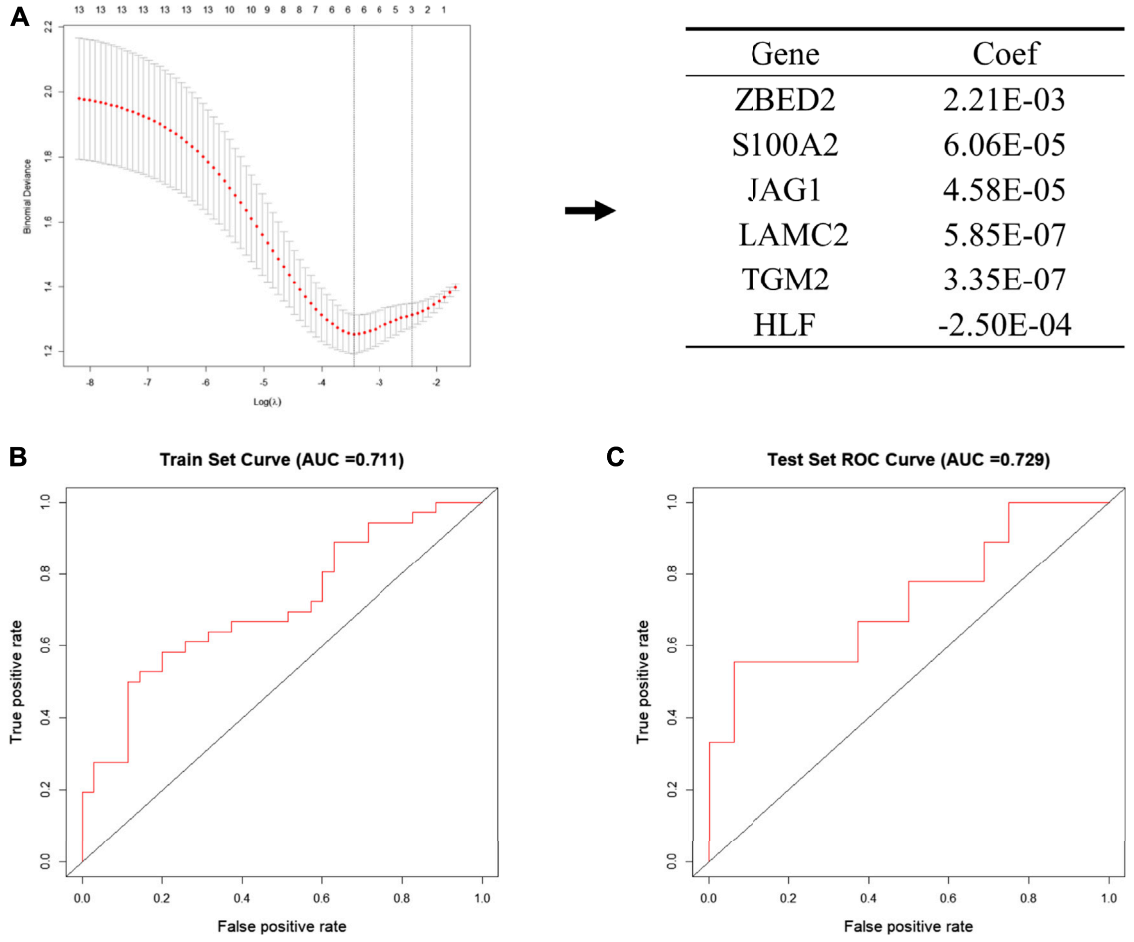
molecule binding and regulation of epidermal cell differentiation (**Supplementary Figure S4C**).

The Association Between Pancreatic Cancer Metastasis and Patient Characteristics, and Heat Map of Six Genes in Pancreatic Cancer Metastasis

Multi-group heat-map analysis showed that *ZBED2*, *S100A2*, *JAG1*, *LAMC2*, *TGM2* were highly expressed, and *HLF* was lowly expressed in both the training and test sets (**Figures 6A,B**), confirming the validity of their selection and the metastasis risk model constructed with them. We compared the metastasis status with other patient characteristics, which were analyzed by a chi-square test. There was a significant relationship between grade (p = 0.037), survival state (p = 0.002), with metastasis of pancreatic cancer in the training set (**Figure 6A**). In the test set, age (p = 0.022) and survival status (p = 0.041) was significantly associated with metastasis (**Figure 6B**). In the training and test set, metastasis was only significantly associated with survival, indicating that the metastatic state of pancreatic cancer is potentially related to prognosis. Survival curves for high- and low-risk patients, as well as univariate and multivariate Cox analyses, were performed to evaluate the prognostic value of risk scores of pancreatic cancer metastasis combined with patient characteristics.

Risk Score is an Important Predictor of Prognosis in Patients With Pancreatic Cancer

Many factors affect the prognosis of pancreatic cancer patients, such as age, tumor size, the degree of invasion, the tissue affected, and the presence of metastasis. To explore the relationship between metastasis risk score and prognosis in pancreatic cancer patients, we used the software X-tile to obtain the optimal cutoff value, and then divided the patients into low and high-risk groups. In the training set, the p value of overall survival (OS) between the low-risk and high-risk groups was 0.034, which was statistically significant between the two groups (**Supplementary Figure S5A**). These results were further confirmed in the test set. Though p-values were not significant in the test set, there was a trend in survival curves and patients in the high risk group had a poor prognosis (p = 0.252) (**Supplementary Figure S5B**).



Gene	Coef
ZBED2	2.21E-03
S100A2	6.06E-05
JAG1	4.58E-05
LAMC2	5.85E-07
TGM2	3.35E-07
HLF	-2.50E-04

FIGURE 4 | Establishment of the metastatic signature based on the six selected genes **(A)** Schematic representation of the model building process. In the left pane, the 13 genes that were used to construct the models are shown, among six genes were selected according to the minimum binomial deviance. The right panel shows the coefficients of six genes screened by Lasso **(B, C)** The ROC curve of risk signature of train set **(B)** and test set **(C)** of LASSO risk scoring model.

Representation of the expression levels of the six genes selected to construct the model in the TCGA data sets using histograms showed that the expression of the *ZBED2*, *S100A2*, *JAG1*, *LAMC2*, and *TGM2* was significantly upregulated in the metastasis group compared with the metastasis-free group. However, *HLF* was significantly down-regulated in the metastasis group compared with the metastasis-free group. Meanwhile, OS analysis of genes in TCGA pancreatic cancer using the GEPIA website showed that high expression of five of these six genes was associated with poor prognosis of pancreatic cancer, and low expression of HLF was associated with poor prognosis of pancreatic cancer (**Supplementary Figure S6**).

To further verify the effect of risk score and patient characteristic features on the prognosis of pancreatic cancer, univariate and multivariate Cox analyses were performed for risk score, age, sex, histological grade, and TNM stage. In the training set, univariate Cox analysis showed that the risk score, histological grade, TNM stage, gender, and age were risk factors (HR > 1), and risk score and histological grade were significant ($p < 0.05$). Multivariate Cox analysis showed that the risk score, TNM stage, gender, and age were risk factors (HR > 1), and the *p*-value of risk score and TNM stage was less than 0.05 (**Supplementary Figures S7A,B**). In the test set, although the *p*-value was greater than 0.05, this may be due to the small sample size (**Supplementary Figures S7C,D**). Univariate and multivariate Cox analyses showed that risk score may be associated with OS, even after accounting for other patient characteristics. These results confirm that the risk score constructed by LASSO is an important predictor of prognosis in pancreatic cancer patients.

DISCUSSION

Pancreatic cancer presents a high rate of metastasis. The expression of EMT-related genes are highly associated with metastasis and poor prognosis of pancreatic cancer (Zheng et al., 2015; Tao et al., 2020). We found that DEGs related to EMT predicted potential biological mechanisms, key signaling pathways, and malignancy markers of pancreatic cancer metastasis. In addition, a model constructed with the expression levels of six EMT-related genes could predict pancreatic cancer metastasis.

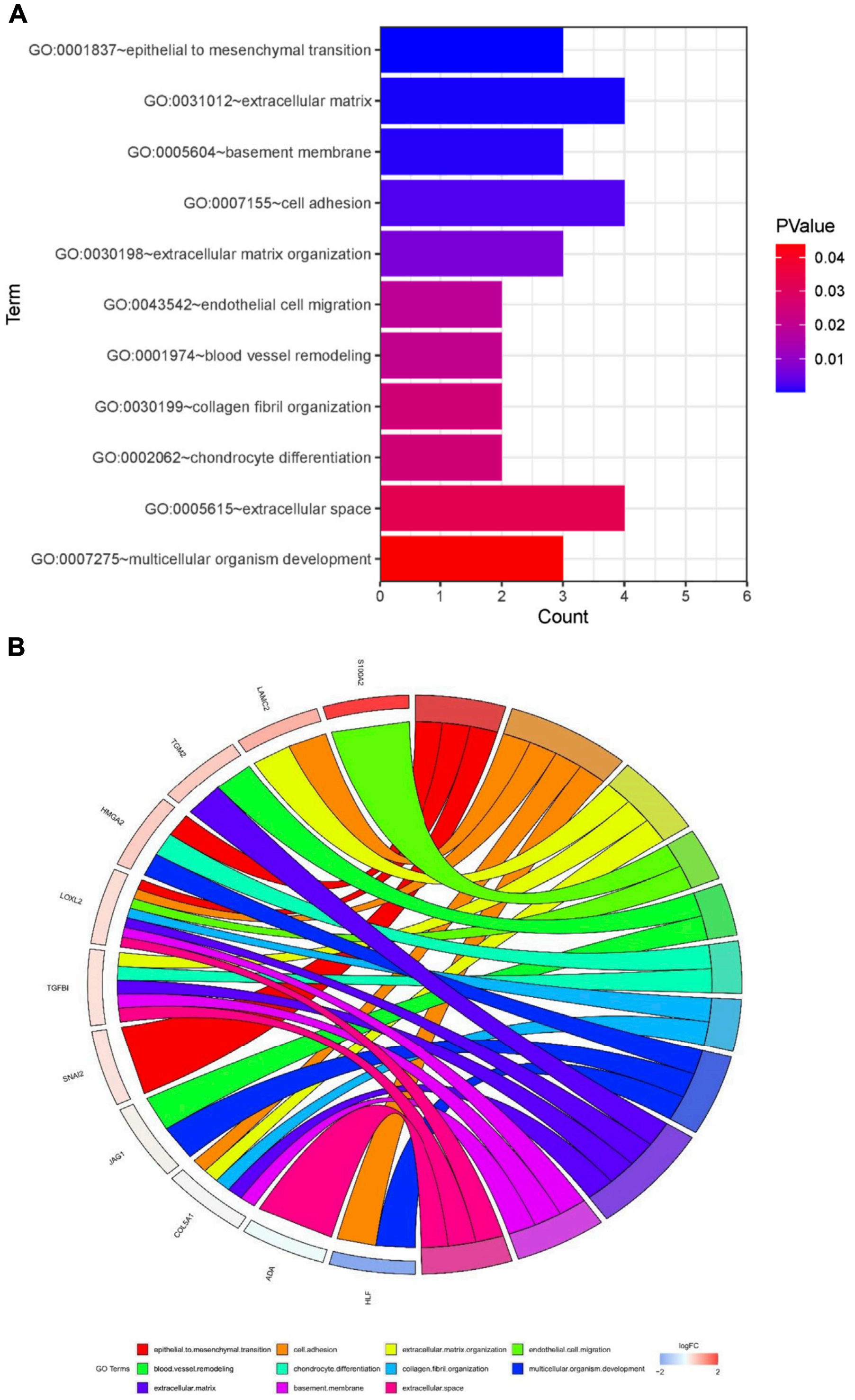


FIGURE 5 | Potential mechanisms associated with tumor metastasis **(A)** GO enrichment analysis of overlapping differentially expressed genes (DEGs) **(B)** Gene Ontology (GO) analysis of 13 epithelial-mesenchymal transition related genes. Analyses are shown in a chord plot, in which the left side represents the 13 selected genes, and the right side shows the GO enrichment pathway of the selected genes. Genes were ranked from large to small according to the change of log FC expression in the sample. The upper side represent up-regulated genes, and the lower side depicts down-regulated genes. Different colors represent different clusters in GO terms.

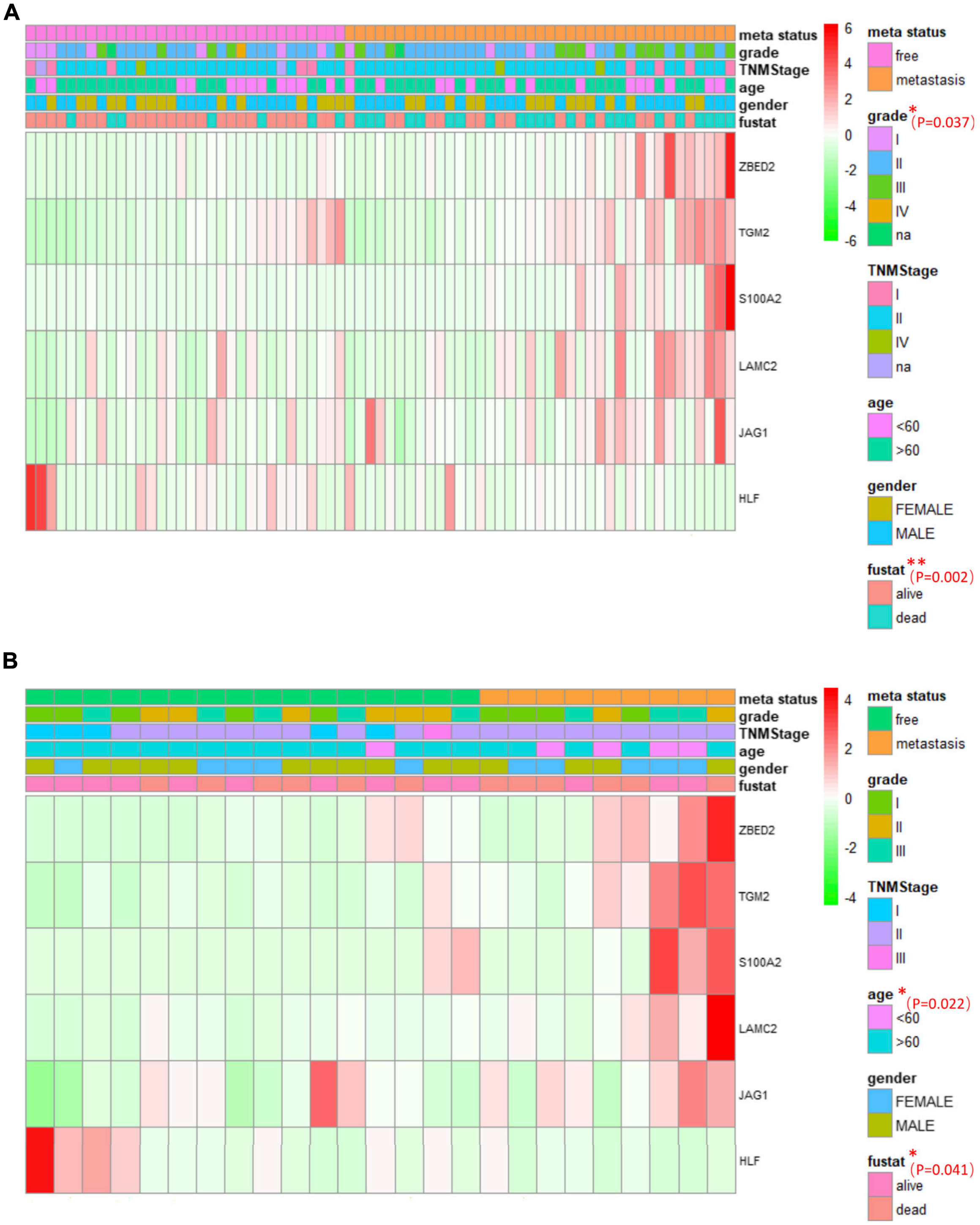


FIGURE 6 | The association between pancreatic cancer metastasis and patient characteristics, and heat maps of six genes in pancreatic cancer metastasis **(A, B)** The heat maps of six EMT related genes in the free and metastatic groups in the train set **(A)** and test set **(B)**. The distribution of patient characteristics was compared between the free and metastatic groups. *$p < 0.05$ and **$p < 0.01$.

In our study, we not only demonstrated the effectiveness of the 6-mRNA based risk score model for predicting pancreatic cancer metastasis, but also demonstrated that it is an effective predictor of prognosis compared to other known features (Stratford et al., 2010; Chen et al., 2015a; Birnbaum et al., 2017). In comparison with other previously reported models, our model is derived from a more comprehensive database and is more generalizable. We also used a machine learning algorithm to construct RF and SVM classification models using 96 pancreatic cancer samples to predict pancreatic cancer metastasis. The AUC of these models indicated that a risk model constructed using LASSO presented a better performance that RF- and SVM-based models. The RF model was constructed using 500 trees, which results in a stable model with good performance. Since the RF algorithm is suitable for processing high-dimensional data, but the dataset used in this study had a small sample size, the RF-based model could not produce good classification, and presented an AUC smaller that of the other models. The SVM algorithm requires the selection of a kernel. However, there is no proper method to determine the kernel of the mapping function of each high-dimensional space in this situation. Even if the kernel function is determined, quadratic optimization of the solution function is required when solving the problem of classification, issues that remain to be further explored further in the future. Finally, the LASSO model can construct a risk scoring model by reducing feature dimension and calculating the corresponding feature coefficients, which can also be used to explore the association between risk score and pancreatic cancer prognosis in combination with clinical indicators. Altogether, these characteristics support choosing the LASSO algorithm for this study.

Previously, we found several EMT-related mRNAs that could be used as biomarkers to predict metastasis of pancreatic cancer, further verifying the relationship between EMT and tumor metastasis (Krebs et al., 2017). In this study, we selected six of these genes to construct our risk model. The expression levels *ZBED2*, *S100A2*, *JAG1*, *LAMC2*, and *TGM2* were upregulated in the metastatic group, while the expression level of *HLF* was downregulated in the metastatic group. *ZBED2* had the highest coefficient value, suggesting that this gene may be related to tumor metastasis. *ZBED2* has previously been shown to promote pancreatic cancer cell invasion by inhibiting the IFN response (Somerville et al., 2020). Our results indicate that high expression of *ZBED2* is associated with metastasis of cancer and can be used as a diagnostic molecular marker. *S100A2* presented the second-largest coefficient. This gene has been considered as a biomarker for pancreatic cancer therapy (Bachet et al., 2013; Feng et al., 2021). GO enrichment analysis indicated that *S100A2* is involved in endothelial cell migration. Previously, it has been demonstrated that *S100A2* upregulation in pancreatic cancer is associated with tumor invasion and poor prognosis (Ohuchida et al., 2007). *JAG1* is mainly involved in Notch signal transduction, which is consistent with our GO and GSEA enrichment results. Moreover, *JAG1* has been reported to be associated with the EMT process of pancreatic cancer and resistance to anticancer drugs (Lee et al., 2020; Zhao et al., 2021). Furthermore, *LAMC2* and *TGM2* have been found to promote the migration of pancreatic cancer cells. *TGM2* knockdown has been reported to inhibit the proliferation and invasion of pancreatic cancer cells (Sagini et al., 2018; Wang et al., 2020). Finally, *HLF* was the only down-regulated gene in the TCGA pancreatic cancer metastasis group. *HLF* has been reported to inhibit proliferation and metastasis of glioma cells (Chen et al., 2016). Our study also suggests that *HLF* may inhibit the invasive process of pancreatic cancer. Notably, this is the first study to show that *HLF* is closely associated with pancreatic cancer metastasis.

Multiple studies have shown that tumor grade is an important prognostic factor in pancreatic cancer (Lüttges et al., 2000; Macías et al., 2018). Although TNM showed a statistically significant hazard ratio (HR > 1) in univariate and multivariate Cox regression analyses, it did not present a significant association in the multivariate Cox analysis of the training set, which may be due to the uneven distribution of subgroups. Approximately 70% of patients had stage II tumors, and only a small proportion of patients presented stage I and IV tumors. In the test set, neither univariate nor multivariate Cox analysis were significant for any of the indicators, which may be due to the small sample size, but the HRs of some indicators were significant (HR > 1).

EMT is thought to increase the resistance of malignant cells and reduce the effectiveness of treatment (Creighton et al., 2009). However, we know that EMT-related genes are not targeted by traditional drugs and antibodies (Palena et al., 2014). Currently, there is no feasible method to inhibit the metastasis process by inhibiting EMT. However, Otsuki et al. reported that injection of snail-specific small interfering RNA (siRNA) into melanoma had a negative effect on tumor migration, accelerated tumor-specific lymphocyte growth, and enhanced the immune response in mice (Otsuki et al., 2018). Furthermore, it has been suggested that the immune microenvironment is crucial for managing tumor development (Elinav et al., 2013). Therefore, targeting EMT-related genes may be a promising therapeutic strategy.

CONCLUSION

In conclusion, this study examined the underlying mechanism of pancreatic cancer metastasis and constructed a model to predict the function and prognostic potential of pancreatic cancer metastasis. Our findings are important for future exploration of the role of EMT in pancreatic cancer metastasis.

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.

AUTHOR CONTRIBUTIONS

HZ designed the study. KX carried out literature search, made charts, and participated in manuscript writing and proofreading. HZ and LZ revised the manuscript. LZ and JW oversee research and critically read drafts. All authors participated in revising, reading, and approving the submitted version.

FUNDING

This study was supported by the Basic Public Welfare Research Program of Zhejiang Province (LGF20H180004), Zhejiang Province Key Research and Development Projects (2020C03057, 2021C03145).

ACKNOWLEDGMENTS

The authors gratefully acknowledge contributions from the GDC and GSE23952 (Sartor MA) and GSE21654 (Euger E). Thanks to Donghai J for help in the construction of BXPC-3 and BXPC-3-M8.

REFERENCES

Bachet, J.-B., Maréchal, R., Demetter, P., Bonnetain, F., Cros, J., Svrcek, M., et al. (2013). S100A2 Is a Predictive Biomarker of Adjuvant Therapy Benefit in Pancreatic Adenocarcinoma. *Eur. J. Cancer* 49 (12), 2643–2653. doi:10.1016/j.ejca.2013.04.017

Barry, S., Chelala, C., Lines, K., Sunamura, M., Wang, A., Marelli-Berg, F. M., et al. (2013). S100P Is a Metastasis-Associated Gene that Facilitates Transendothelial Migration of Pancreatic Cancer Cells. *Clin. Exp. Metastasis* 30 (3), 251–264. doi:10.1007/s10585-012-9532-y

Birnbaum, D. J., Finetti, P., Lopresti, A., Gilabert, M., Poizat, F., Raoul, J.-L., et al. (2017). A 25-gene Classifier Predicts Overall Survival in Resectable Pancreatic Cancer. *BMC Med.* 15 (1), 170. doi:10.1186/s12916-017-0936-z

Chaika, N. V., Yu, F., Purohit, V., Mehla, K., Lazenby, A. J., DiMaio, D., et al. (2012). Differential Expression of Metabolic Genes in Tumor and Stromal Components of Primary and Metastatic Loci in Pancreatic Adenocarcinoma. *PLoS One* 7 (3), e32996. doi:10.1371/journal.pone.0032996

Chen, D.-T., Davis-Yadley, A. H., Huang, P.-Y., Husain, K., Centeno, B. A., Permuth-Wey, J., et al. (2015a). Prognostic Fifteen-Gene Signature for Early Stage Pancreatic Ductal Adenocarcinoma. *PLoS One* 10 (8), e0133562. doi:10.1371/journal.pone.0133562

Chen, S., Wang, Y., Ni, C., Meng, G., and Sheng, X. (2016). HLF/miR-132/TTK axis Regulates Cell Proliferation, Metastasis and Radiosensitivity of Glioma Cells. *Biomed. Pharmacother.* 83, 898–904. doi:10.1016/j.biopha.2016.08.004

Chen, Y., Gao, S.-G., Chen, J.-M., Wang, G.-P., Wang, Z.-F., Zhou, B., et al. (2015b). Serum CA242, CA199, CA125, CEA, and TSGF Are Biomarkers for the Efficacy and Prognosis of Cryoablation in Pancreatic Cancer Patients. *Cell Biochem. Biophys.* 71 (3), 1287–1291. doi:10.1007/s12013-014-0345-2

Creighton, C. J., Li, X., Landis, M., Dixon, J. M., Neumeister, V. M., Sjolund, A., et al. (2009). Residual Breast Cancers after Conventional Therapy Display Mesenchymal as Well as Tumor-Initiating Features. *Proc. Natl. Acad. Sci.* 106 (33), 13820–13825. doi:10.1073/pnas.0905718106

Elinav, E., Nowarski, R., Thaiss, C. A., Hu, B., Jin, C., and Flavell, R. A. (2013). Inflammation-induced Cancer: Crosstalk between Tumours, Immune Cells and Microorganisms. *Nat. Rev. Cancer* 13 (11), 759–771. doi:10.1038/nrc3611

Feng, Z., Li, K., Lou, J., Wu, Y., and Peng, C. (2021). An EMT-Related Gene Signature for Predicting Response to Adjuvant Chemotherapy in Pancreatic Ductal Adenocarcinoma. *Front. Cell Dev. Biol.* 9, 665161. doi:10.3389/fcell.2021.665161

Goeman, J. J. (2009). L1Penalized Estimation in the Cox Proportional Hazards Model. *Biom. J.* 52 (1), NA. doi:10.1002/bimj.200900028

Jiang, W., Bai, W., Li, J., Liu, J., Zhao, K., and Ren, L. (2021). Leukemia Inhibitory Factor Is a Novel Biomarker to Predict Lymph Node and Distant Metastasis in Pancreatic Cancer. *Int. J. Cancer* 148 (4), 1006–1013. doi:10.1002/ijc.33291

Krebs, A. M., Mitschke, J., Lasierra Losada, M., Schmalhofer, O., Boerries, M., Busch, H., et al. (2017). The EMT-Activator Zeb1 Is a Key Factor for Cell Plasticity and Promotes Metastasis in Pancreatic Cancer. *Nat. Cell Biol.* 19 (5), 518–529. doi:10.1038/ncb3513

Lee, J., Lee, J., and Kim, J. H. (2020). Association of Jagged1 Expression with Malignancy and Prognosis in Human Pancreatic Cancer. *Cell Oncol.* 43 (5), 821–834. doi:10.1007/s13402-020-00527-3

Li, C., Zeng, X., Yu, H., Gu, Y., and Zhang, W. (2018). Identification of Hub Genes with Diagnostic Values in Pancreatic Cancer by Bioinformatics Analyses and Supervised Learning Methods. *World J. Surg. Onc* 16 (1), 223. doi:10.1186/s12957-018-1519-y

Lüttges, J., Schemm, S., Vogel, I., Hedderich, J., Kremer, B., and Klöppel, G. (2000). The Grade of Pancreatic Ductal Carcinoma Is an Independent Prognostic Factor and Is superior to the Immunohistochemical Assessment of Proliferation. *J. Pathol.* 191 (2), 154–161. doi:10.1002/(SICI)1096-9896(200006)191:2<154::AID-PATH603>3.0.CO;2-C

Macías, N., Sayagués, J. M., Esteban, C., Iglesias, M., González, L. M., Quiñones-Sampedro, J., et al. (2018). Histologic Tumor Grade and Preoperative Bilary Drainage Are the Unique Independent Prognostic Factors of Survival in Pancreatic Ductal Adenocarcinoma Patients after Pancreaticoduodenectomy. *J. Clin. Gastroenterol.* 52 (2), e11–e17. doi:10.1097/mcg.0000000000000793

McDonald, O. G., Li, X., Saunders, T., Tryggvadottir, R., Mentch, S. J., Warmoes, M. O., et al. (2017). Epigenomic Reprogramming during Pancreatic Cancer Progression Links Anabolic Glucose Metabolism to Distant Metastasis. *Nat. Genet.* 49 (3), 367–376. doi:10.1038/ng.3753

Obuchowski, N. A., and Bullen, J. A. (2018). Receiver Operating Characteristic (ROC) Curves: Review of Methods with Applications in Diagnostic Medicine. *Phys. Med. Biol.* 63 (7), 07tr01. doi:10.1088/1361-6560/aab4b1

Ohuchida, K., Mizumoto, K., Miyasaka, Y., Yu, J., Cui, L., Yamaguchi, H., et al. (2007). Over-expression ofS100A2 in Pancreatic Cancer Correlates with Progression and Poor Prognosis. *J. Pathol.* 213 (3), 275–282. doi:10.1002/path.2250

Otsuki, Y., Saya, H., and Arima, Y. (2018). Prospects for New Lung Cancer Treatments that Target EMT Signaling. *Dev. Dyn.* 247 (3), 462–472. doi:10.1002/dvdy.24596

Owusu-Ansah, K., Song, G., Chen, R., Edoo, M., Li, J., Chen, B., et al. (2019). COL6A1 Promotes Metastasis and Predicts Poor Prognosis in Patients with Pancreatic Cancer. *Int. J. Oncol.* 55 (2), 391–404. doi:10.3892/ijo.2019.4825

Palena, C., Fernando, R. I., and Hamilton, D. H. (2014). An Immunotherapeutic Intervention against Tumor Progression. *Oncoimmunology* 3 (1), e27220. doi:10.4161/onci.27220

Sagini, M. N., Zepp, M., Bergmann, F., Bozza, M., Harbottle, R., and Berger, M. R. (2018). The Expression of Genes Contributing to Pancreatic Adenocarcinoma Progression Is Influenced by the Respective Environment. *Genes Cancer* 9 (3-4), 114–129. doi:10.18632/genesandcancer.173

Somerville, T. D. D., Xu, Y., Wu, X. S., Maia-Silva, D., Hur, S. K., de Almeida, L. M. N., et al. (2020). ZBED2 Is an Antagonist of Interferon Regulatory Factor 1 and Modifies Cell Identity in Pancreatic Cancer. *Proc. Natl. Acad. Sci. USA* 117 (21), 11471–11482. doi:10.1073/pnas.1921484117

Stratford, J. K., Bentrem, D. J., Anderson, J. M., Fan, C., Volmar, K. A., Marron, J. S., et al. (2010). A Six-Gene Signature Predicts Survival of Patients with Localized Pancreatic Ductal Adenocarcinoma. *Plos Med.* 7 (7), e1000307. doi:10.1371/journal.pmed.1000307

Tao, C., Huang, K., Shi, J., Hu, Q., Li, K., and Zhu, X. (2020). Genomics and Prognosis Analysis of Epithelial-Mesenchymal Transition in Glioma. *Front. Oncol.* 10, 183. doi:10.3389/fonc.2020.00183

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statist. Med.* 16 (4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Wang, H., Cai, J., Du, S., Wei, W., and Shen, X. (2020). LAMC2 Modulates the Acidity of Microenvironments to Promote Invasion and Migration of Pancreatic Cancer Cells via Regulating AKT-dependent NHE1 Activity. *Exp. Cell Res.* 391 (1), 111984. doi:10.1016/j.yexcr.2020.111984

Zemmour, C., Bertucci, F., Finetti, P., Chetrit, B., Birnbaum, D., Filleron, T., et al. (2015). Prediction of Early Breast Cancer Metastasis from DNA Microarray Data Using High-Dimensional Cox Regression Models. *Cancer Inform.* 14s2 (Suppl. 2), CIN.S17284–138. doi:10.4137/cin.S17284

Zhao, Z., Shen, X., Zhang, D., Xiao, H., Kong, H., Yang, B., et al. (2021). miR-153 Enhances the Therapeutic Effect of Radiotherapy by Targeting JAG1 in Pancreatic Cancer Cells. *Oncol. Lett.* 21 (4), 300. doi:10.3892/ol.2021.12561

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2021.701039/full#supplementary-material

Zheng, X., Carstens, J. L., Kim, J., Scheible, M., Kaye, J., Sugimoto, H., et al. (2015). Epithelial-to-mesenchymal Transition Is Dispensable for Metastasis but Induces Chemoresistance in Pancreatic Cancer. *Nature* 527 (7579), 525–530. doi:10.1038/nature16064

Zhi, J., Sun, J., Wang, Z., and Ding, W. (2018). Support Vector Machine Classifier for Prediction of the Metastasis of Colorectal Cancer. *Int. J. Mol. Med.* 41 (3), 1419–1426. doi:10.3892/ijmm.2018.3359

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.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 31 August 2021
doi: 10.3389/fbioe.2021.667641

Surveillance Strategy for Barcelona Clinic Liver Cancer B Hepatocellular Carcinoma Achieving Complete Response: An Individualized Risk-Based Machine Learning Study

Qi-Feng Chen[1,2,3†], *Lin Dai*[4†], *Ying Wu*[1,2,3†], *Zilin Huang*[1†], *Minshan Chen*[5] *and Ming Zhao*[1] *

[1]Department of Medical Imaging and Interventional Radiology, Sun Yat-sen University Cancer Center, Guangzhou, China, [2]State Key Laboratory of Oncology in South China, Guangzhou, China, [3]Collaborative Innovation Center for Cancer Medicine, Guangzhou, China, [4]Cancer Prevention Center, Sun Yat-sen University Cancer Center, Guangzhou, China, [5]Department of Liver Surgery, Sun Yat-sen University Cancer Center, Guangzhou, China

OPEN ACCESS

Edited by:
Wenbing Zhao,
Cleveland State University,
United States

Reviewed by:
Fernanda Aurora Seixas Travassos,
University of Trás-os-Montes and Alto Douro, Portugal
Chao An,
First Affiliated Hospital of Chinese PLA General Hospital, China
Bilal Ahmad Hijazi,
King Saud Medical City, Saudi Arabia

****Correspondence:***
Ming Zhao
drmingzhao@outlook.com

[†]These authors share first authorship

Specialty section:
This article was submitted to Biosafety and Biosecurity, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *14 February 2021*
Accepted: *30 June 2021*
Published: *31 August 2021*

Citation:
Chen Q-F, Dai L, Wu Y, Huang Z, Chen M and Zhao M (2021) Surveillance Strategy for Barcelona Clinic Liver Cancer B Hepatocellular Carcinoma Achieving Complete Response: An Individualized Risk-Based Machine Learning Study. Front. Bioeng. Biotechnol. 9:667641. doi: 10.3389/fbioe.2021.667641

Background: For patients with complete response (CR) of Barcelona Clinical Liver Cancer (BCLC) stage B hepatocellular carcinoma (HCC), there is no consensus regarding the monitoring strategy. Optimal surveillance strategies that can detect early progression of HCC within a limited visit after treatment have not yet been investigated. A retrospective, real-world study was conducted to investigate surveillance strategies for BCLC stage B HCC (BBHCC) patients with CR after curative treatment to support clinical decision making.

Methods: From January 2007 to December 2019, 546 BBHCC patients with CR after radical treatment were collected at Sun Yat-sen University Cancer Center. Seventy percent of patients were subjected to the train cohort randomly; the remaining patients comprised the validation cohort to verify the proposed arrangements. The random survival forest method was applied to calculate the disease progression hazard per month, and follow-up schedules were arranged to maximize the capability of progression detection at each visit. The primary endpoint of the study was the delayed-detection months for disease progression.

Results: The cumulative 1, 2, and 3-years risk-adjusted probabilities for the train/validation cohorts were 32.8%/33.7%, 54.0%/56.3%, and 64.0%/67.4%, respectively, with peaks around approximately the 9th month. The surveillance regime was primarily concentrated in the first year posttreatment. The delayed-detection months gradually decreased when the total follow-up times increased from 6 to 11. Compared with controls, our schedule reduced delayed detection. Typically, the benefits of our surveillance regimes

Abbreviations: BCLC, barcelona clinical liver cancer: HCC, hepatocellular carcinoma: BBHCC, BCLC B stage HCC: CR, complete response: TACE, transarterial chemoembolization: RSF, random survival forest: CT, computed tomography: MRI, magnetic resonance imaging: ECIO, European conference on interventional oncology: ESOI, European society of oncologic imaging: PFS, progression-free survival

were obvious when the patients were followed seven times according to our schedule. The optional schedules were 5, 7, 9, 11, 17, 23, and 30 months.

Conclusion: The proposed new surveillance schedule may provide a new perspective concerning follow-up for BBHCC patients with CR.

Keywords: Barcelona clinic liver cancer B, hepatocellular carcinoma, complete response, surveillance strategy, machine learning

INTRODUCTION

Hepatocellular carcinoma (HCC) was the 6th most diagnosed cancer type and the 4th leading cause of cancer death worldwide in 2018 (Bray et al., 2018). The Barcelona Clinical Liver Cancer (BCLC) algorithm is a useful HCC staging classifier that is utilized worldwide (Forner et al., 2018). The BCLC staging system has been extensively validated clinically, and it is the most commonly used system for HCC. Following BCLC guidelines, only early-stage patients (BCLC 0/A) should be treated with radical therapies (surgery/ablation). For BCLC B stage HCC (BBHCC) cases with large multifocal tumors and without vascular invasion or spread outside of the liver, transarterial chemoembolization (TACE) is recommended if liver function is maintained. Nevertheless, recent advances in technology and appropriate patient selection have gradually reduced the morbidity and mortality of radical treatments, which have been considered for BBHCC patients with promising results in terms of postoperative outcomes (Day et al., 2016; Chen et al., 2017). For example, Labgaa I et al. systematically analyzed 1,730 BBHCC patients and found that compared with TACE, surgery improved long-term survival; postoperative mortality was equivalent (Labgaa et al., 2020). In our previous study, ablation-TACE combination therapy had a better clinical efficacy than TACE monotherapy for BBHCC (Chen et al., 2017; Zhang et al., 2018). Therefore, selected BBHCC patients might benefit from radical therapies.

Regardless, follow-up is a confusing issue in the BCLC staging system, which is important for assessing treatment success and detecting disease progression. The practical monitoring strategies in guidelines are mainly based on expert opinions. It is recommended that cancer survivors should be monitored regularly after treatment (radiological examination three to 6 months on average) to expedite detection of disease progression (Kanwal and Singal, 2019; Chen et al., 2020a). Tumors may relapse after radical therapies, leading to an early diagnosis of tumor relapse being more likely to be treated curatively, which can better manage the disease and prolong survival (Trinchet et al., 2011; Vogel et al., 2018). At present, the question of what is the best monitoring strategy that can detect tumor progression in a timely manner after treatment remains. Although recent guidelines suggest follow-up strategies for monitoring after curative treatment (Chen et al., 2020a), there is a lack of a specific surveillance algorithm for curatively treated HCCs, especially for BBHCC patients who show a complete response (CR) after radical treatment. The guidelines do not recommend specific monitoring intervals for BBHCC patients with CR, cases that are more complicated and likely to relapse earlier than BCLC stage 0/A cases. In addition, it remains unclear whether the current surveillance strategies are adequate.

In this study, we applied a random survival forest (RSF) analysis, a machine learning method, to calculate the probability of disease progression for each month. Thereafter, a risk-associated surveillance program was established on the basis of the abovementioned disease progression probabilities. The surveillance regime was evaluated by calculating the total number of delayed-detection days, followed by comparison to other surveillance proposals. Our surveillance strategy for BBHCC patients with CR after radical therapy will support clinical follow-up decision making.

MATERIALS AND METHODS

Patient Datasets and Processing

We retrospectively collected BBHCC patients who underwent radical treatment (surgery/ablation) from an institutional database at Sun Yat-sen University Cancer Center from January 2007 to December 2019. All cases were diagnosed with HCC according to pathology or clinical criteria (Xie et al., 2020). A total of 2,193 consecutive BBHCC patients were initially eligible. This study included BBHCC patients who received radical treatment and achieved CR. Patients underwent multidetector computed tomography (CT) and/or magnetic resonance imaging (MRI) routinely to evaluate the local or distant extension of the primary tumors. After radical treatment, the patients were instructed to undergo multiphasic cross-sectional chest, abdomen, and pelvis high-quality imaging checks within first month, and every 2–6 months thereafter. CR is defined as no disease progression (death or local/distant tumor progression) during first follow-up after radical treatment. We excluded patients who had any of the following criteria: <18 or >80 years, mixed liver cancer, or death due to postoperative complications. Clinical and blood tests were performed at diagnosis and surveillance. After excluding 1,639 patients according to the exclusion criteria, 546 patients were included in the study. All patients received radical treatment, with some being treated with TACE (considered noncurative treatment) before radical treatment. The train cohort consisted of 382 patients (70%); 164 patients (30%) were used as the validation cohort. Considering the retrospective nature of the study, our cancer center institutional ethics committee approved the study protocol and waived the requirement for informed consent.

Disease Progression Probability Calculation

To determine the optimal surveillance strategy, we first assessed the cumulative disease progression probabilities over 3 years in the two cohorts (train/validation cohorts) through the RSF method and calculated the probability of disease progression every month. As a machine learning tool, RSF can conduct right-censored survival statistical analysis (Taylor, 2011). The RSF method has a number of appealing features, with a major feature concerning our study being that none of the variables is deleted or selected, such that all of the variables influence the predicted result. In addition, RSF can incorporate situations in which the complex relationship between predictor and response variables occurs and predictors have nonlinear patterns and interactions. The RSF method plotted survival curves for two cohorts, involving all hazard-modified variables. Processing of the RSF method was conducted using the R package of random forest SRC.

Development of a Risk-Based Follow-Up Schedule

After calculating the probability of disease progression for each month, a total number of follow-up times from the minimum of 6 (follow-up every 6 months) to the maximum of 11 (follow-up every 3 months from 4th month) was set. The follow-up times were assigned based on the progression probability of each month; the best strategy is to strike a balance between timely progression detection and minimal follow-up times.

We assessed the surveillance strategy based on the total delayed-detection days and compared it to a typical surveillance strategy (set as control), as follows: 7 times (Maas et al., 2020) (1, 3, 6, 9, 12, 18, 24, 30, and 36 months, which was put forward in a cooperative meeting of ECIO (European Conference on Interventional Oncology) and ESOI (European Society of Oncologic Imaging).

Therefore, within a 2-year period, the full supervision times for the hazard-based surveillance regime should be adjusted from a maximum of 11 to a minimum of 6. To explore an ideal follow-up schedule for rapidly revealing disease progression under minimal follow-up, we subsequently created a surveillance program covering a 3-years supervision ranging from 6 to 11 times. We assigned follow-ups to those months in which disease would more likely progress when one supervision was arranged for each month at most (Zhou et al., 2020).

Delayed-Detection Calculation

Then, we compared our surveillance schedule with the control strategies. The capability of the supervision strategy was quantified by counting the total delayed-detection months in the train cohort. Delayed-detection months were defined as the time from disease progression to the next closest follow-up. As an example, if a patient progressed on the 200th day and the following most recent scheduled day was 240, then the delayed-detection days for that patient was 40. We calculated the total number of delayed-detection months for our plans and compared it with the control strategy. Strategy that reduced the sum of delayed-detection months with less follow-up time were considered preferable. The arrangements of the proposed schedule were also applied to patients in the validation cohort.

Statistical Analysis

Disease progression was deemed death or local/distant tumor progression. Progression-free survival (PFS) was measured from the date of CR to disease progression or the last follow-up evaluation (August 2020). The $\chi 2$ or Fisher's exact probability test was used for categorical variables. In the train group, the risk-based surveillance schedule was conducted by RSF. The time differences in delayed detection between our model and the recommended model were compared using the paired *t*-test or Kruskal-Wallis test. R language (version 3.6.0; R Foundation) was utilized for all analyses. A two-sided $p < 0.05$ indicated that the difference was of statistical significance.

RESULTS

Patient Characteristics

A total of 382 patients in the train cohort and 164 in the validation cohort were enrolled. The flowchart is illustrated in **Figure 1**. **Table 1** shows the basic patient characteristics. Demographics were similar with regard to sex, age, hepatitis virus infection, α-fetoprotein (AFP) level, tumor size, tumors numbers, cirrhosis, combined TACE therapy, surgery or ablation treatment, tumor differentiation, satellite nodules, venous invasion, perineural invasion, capsule invasion, disease progression status, and PFS time between the train and validation cohorts ($p > 0.05$). In the whole cohort, progression occurred in 309 patients, with a median PFS of 13.7 months (interquartile range (IQR], 7.80–25.70 months).

Calculation of Disease Progression Probability

To measure the monthly disease progression hazard of the train/validation cohorts, the monthly probability of disease progression was calculated by the RSF method, which was adjusted with clinical factors. **Figure 2A** shows the progression probabilities of patients in the train/validation cohorts. The cumulative 1, 2, and 3-years adjusted risk probabilities for the train cohort were 32.8, 54.0, and 64.0%, respectively, and those for the validation cohort were 33.7, 56.3, and 67.4%, respectively (**Supplementary Table S1**).

Then, the calculation of progression probability at a specific time was performed. The probability patterns in both the train/validation cohorts were quite similar. According to the data shown in **Figure 2B**, the disease progression incidence rose rapidly, reached a peak around approximately the 9th month, and decreased smoothly to a plateau less than 2% (**Supplementary Table S2**).

Development of a Risk-Based Follow-Up Schedule

Next, a risk-based surveillance regime was established depending on the disease progression probability of each month by the

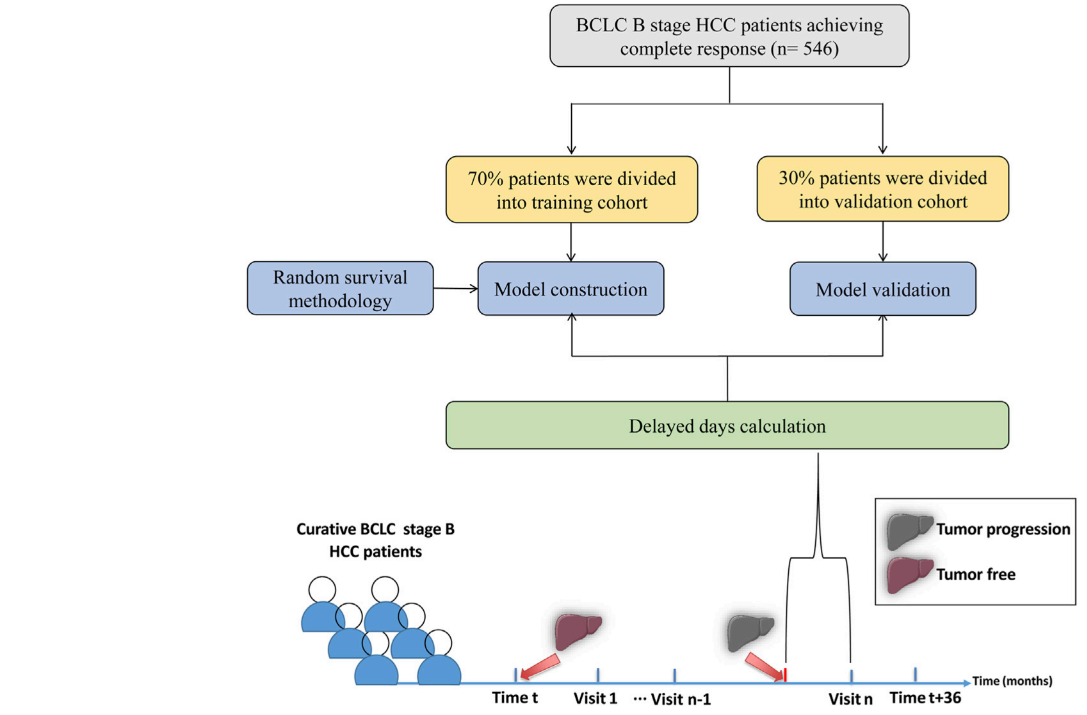

BCLC B stage HCC patients achieving complete response (n= 546)
70% patients were divided into training cohort
30% patients were divided into validation cohort
Random survival methodology
Model construction
Model validation
Delayed days calculation
Curative BCLC stage B HCC patients
Tumor progression
Tumor free
Time t
Visit 1
··· Visit n-1
Visit n
Time t+36
Time (months)

FIGURE 1 | General design of the present study.

prescribed method. The follow-up schedule with total follow-up times ranging from 6 to 11 for the first 3 years is depicted in **Figure 3**. The surveillance regime was concentrated primarily in the first year posttreatment with rather less supervision during the following years. The third year had relatively fewer follow-up times and more follow-up times were allocated in the second half year.

Delayed Month Comparison

We compared our model performance (the ability to detect disease progression in a timely manner) with that of controls (**Figure 4**). As shown in **Figures 4A,B**, the delayed-detection time gradually decreased when the total follow-up time increased from 6 to 11. The delayed-detection months of our surveillance regime (blue dots with gray curves connected) with that of the control (the red points indicated a 7 times follow-up strategy) were also compared. As presented in **Figures 4A,B**, under the same number of follow-up times, our monitoring arrangement significantly reduced the delayed-detection months, which was more efficient than in the controls. Typically, when patients were followed seven times according to our schedule, the advantage of our surveillance schedule was of significance.

Our recommended supervision schedules are as follows. Our surveillance schedule involves seven times within 3 years (5, 7, 9, 11, 17, 23, and 30 months, respectively). The detailed schedule for each follow-up is shown in **Figure 5A**. In general, monitoring should be concentrated in the first year posttreatment. The proposed supervision schedules were further verified in individual disease progressed cases of train cohort and the validation cohort that had clinical characteristics that were almost consistent with those of the train cohort. For disease progressed patients, the surveillance strategy recommended by us significantly decreased the delayed-detection time compared with the control (**Figures 5B**,C, both *p*-values< 0.01).

DISCUSSION

Currently, there are limited validated data showing an optional surveillance schedule for BBHCC patients with CR. Given the potential for disease progression posttreatment, continued monitoring is necessary for these patients. In this population-based real-world study, we applied an RSF method to determine the risk of disease progression for each month. Thereafter, we propose a surveillance plan that is able to detect disease progression effectively at each follow-up. Typically, our model was more efficient than control schedules. Despite the fact that our project was generated using the BBHCC population, our risk-related monitoring method can be applied to develop monitoring strategies for other BCLC-stage HCC patients and can generally help develop personalized surveillance schedules after treatment.

Many professional associations have put forward guidelines for posttreatment management and have provided universal surveillance recommendations for HCC patients. However, most of these recommendations were derived from early HCC cases. Thus, due to the substantial differences in biology, therapy, and the way disease progresses, early HCC monitoring experience may not be applicable to BBHCC (Chen et al., 2019; Liu et al., 2020). Boas FE et al. enrolled 910 patients receiving 1,766 successive

TABLE 1 | Patient characteristics in the train and validation cohorts.

	Overall (546)	Train (382)	Validation (164)	*p*-value
Gender = Male/Female (%)	480/66 (87.9/12.1)	338/44 (88.5/11.5)	142/22 (86.6/13.4)	0.631
Age (year) = <45/≥45 (%)	137/409 (25.1/74.9)	99/283 (25.9/74.1)	38/126 (23.2/76.8)	0.568
Hepatatis virus = No/Yes (%)	74/472 (13.6/86.4)	51/331 (13.4/86.6)	23/141 (14.0/86.0)	0.941
AFP(ng/ml) = <400/≥400 (%)	296/250 (54.2/45.8)	206/176 (53.9/46.1)	90/74 (54.9/45.1)	0.912
Tumor size = <50 mm/≥50 mm (%)	256/290 (46.9/53.1)	182/200 (47.6/52.4)	74/90 (45.1/54.9)	0.654
Tumor number = <4/≥4 (%)	458/88 (83.9/16.1)	318/64 (83.2/16.8)	140/24 (85.4/14.6)	0.624
Cirrhosis = No/Yes (%)	189/357 (34.6/65.4)	129/253 (33.8/66.2)	60/104 (36.6/63.4)	0.592
Combined TACE = No/Yes (%)	312/234 (57.1/42.9)	218/164 (57.1/42.9)	94/70 (57.3/42.7)	0.999
Surgery/Ablation (%)	436/110 (79.9/20.1)	307/75 (80.4/19.6)	129/35 (78.7/21.3)	0.734
Differentiation (%)	—	—	—	0.713
Well	175 (32.1)	118 (30.9)	57 (34.8)	—
Moderated	243 (44.5)	175 (45.8)	68 (41.5)	—
Poor	27 (4.9)	20 (5.2)	7 (4.3)	—
Unknown	101 (18.5)	69 (18.1)	32 (19.5)	—
Satellite nodules (%)	—	—	—	0.736
No	401 (73.4)	284 (74.3)	117 (71.3)	—
Yes	41 (7.5)	27 (7.1)	14 (8.5)	—
Unknown	104 (19.0)	71 (18.6)	33 (20.1)	—
Venous invasion (%)	—	—	—	0.753
No	303 (55.5)	216 (56.5)	87 (53.0)	—
Yes	139 (25.5)	95 (24.9)	44 (26.8)	—
Unknown	104 (19.0)	71 (18.6)	33 (20.1)	—
Perineural invasion (%)	—	—		0.487
No	439 (80.4)	308 (80.6)	131 (79.9)	—
Yes	3 (0.5)	3 (0.8)	0 (0.0)	—
Unknown	104 (19.0)	71 (18.6)	33 (20.1)	—
Capsule invasion (%)	—	—	—	0.879
No	188 (34.4)	133 (34.8)	55 (33.5)	—
Yes	255 (46.7)	179 (46.9)	76 (46.3)	—
Unknown	103 (18.9)	70 (18.3)	33 (20.1)	—
PFS = No/Yes (%)	237/309 (43.4/56.6)	167/215 (43.7/56.3)	70/94 (42.7/57.3)	0.897
PFS (median (IQR))	13.70 (7.80, 25.70)	14.15 (8.03, 25.35)	13.00 (7.47, 27.65)	0.572

TACE, transcatheter arterial chemoembolization; AFP, α-fetoprotein; PFS, progression-free survival; IQR, interquartile range.

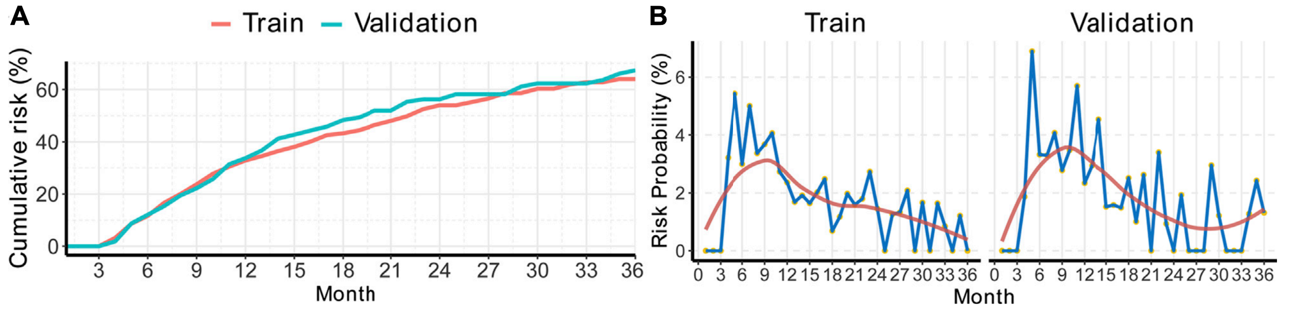
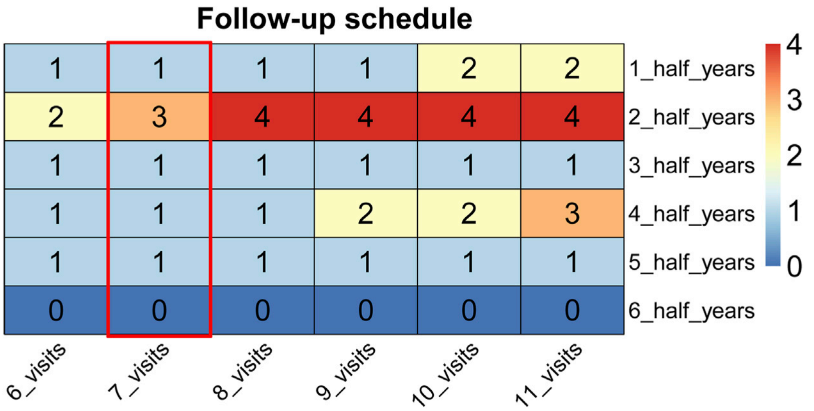


FIGURE 2 | Cumulative risk curves and time-specific progression probabilities of BBHCC CR patients. **(A)** Cumulative risk curves. **(B)** Each month's progression probability. BBHCC: Barcelona clinical liver cancer stage B hepatocellular carcinoma; CR: complete response.

operations, including TACE, radioembolization, and ablation, at a single institution regardless of patient stage between 2006 and 2011 (Boas et al., 2015). Consistent with our results, they demonstrated that more recurrence occurred in the first year after treatment, leading to much more frequent screening in the first year. In april 2018, a joint session from ECIO and ESOI produced a recommendation based on the literature and expert opinion that the total number of follow-up times is 7 for liver-directed cases (first year: 1, 3, 6, 9, and 12 months; every 6 months thereafter) (Maas et al., 2020). However, the quality and quantity of evidence were limited. The recommendations conferred were based in part on expert opinion and consensus that applied to all liver cancer patients; moreover, it remains unknown whether the follow-up guidelines are most effective. As a result, we focused on BBHCC patients with CR and developed the present new superior surveillance strategy.

The purpose of tumor monitoring is to detect disease progression as early as possible (Wu et al., 2020a; Chen et al., 2020b). Thus, patients would ideally be checked every day, which is not practical in



Follow-up schedule

	6_visits	7_visits	8_visits	9_visits	10_visits	11_visits
1_half_years	1	1	1	1	2	2
2_half_years	2	3	4	4	4	4
3_half_years	1	1	1	1	1	1
4_half_years	1	1	1	2	2	3
5_half_years	1	1	1	1	1	1
6_half_years	0	0	0	0	0	0

FIGURE 3 | The supervision arrangements ranging from 6 to 11 follow-up times. The follow-up schedules of seven times are highlighted in the red box.

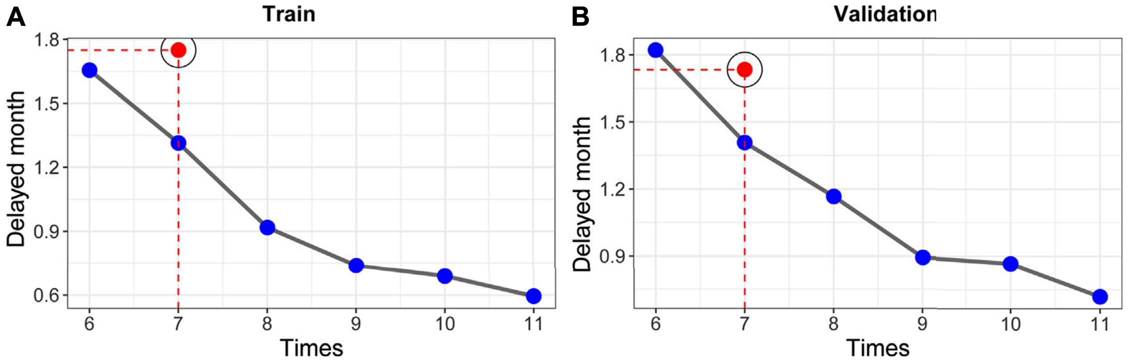


FIGURE 4 | Establishment of risk-based surveillance arrangements. **(A,B)** In contrast with the control strategy (7 times, which was put forward in a cooperative meeting of ECIO and ESOI), our surveillance arrangements (blue points with gray curve connected) had fewer delayed-detection days.

clinical practice. Regardless, the surveillance strategy can be improved based on disease progression probabilities per month by scheduling as close as possible surveillance time to the expected time. We assumed that the damage caused by delayed detection was proportionally associated with delayed detection. In fact, there might be a threshold value that could be useful for clinical decision making. When patients were followed within the threshold value, delayed-detection days could be ignored clinically. However, no published research solves this problem in a quantitative manner.

Currently, it is generally believed that the detection of disease progression as soon as possible is of utmost importance, as a number of cases of early progression can be treated effectively (Wu et al., 2020b; Li et al., 2020). Tsilimigras DI reported that 154 BCLC B/C patients underwent resection with an annual recurrence rate of 38.3% during the first postoperative year (Tsilimigras et al., 2020). In the study by Di Sandro S et al. (2019), relapse-free survival of 131 BBHCC patients receiving surgery was 34.4, 21.4, 15.3, 6.1, and 2.3% for 2, 4, 6, 8 and 10 years, respectively. In our previous research, outcomes of BBHCC patients receiving TACE improved when treatment was combined with ablation therapy, regardless of whether the patients achieved CR (Zhang et al., 2018). We found that the median time of tumor progression was 10.14 months in the TACE-ablation group, with disease progression rates of 26.0, 52.2, 65.0, and 68.2% at 6, 12, 18, and 24 months, respectively. In the present analysis, the occurrence peak of disease progression occurred approximately around the 9th months; and almost 45% of the patients did not experience tumor progression following the first 3 years after treatment. The superiority in tumor control may be attributed to the fact that the BBHCC patients in our cohort all had CR after treatment. Accordingly, the schedule should be concentrated in the first year posttreatment.

We acknowledge the following limitations of our study. First, our research was retrospectively conducted in a single center for more than 10 years. The next step is to complete multicenter data collection to expand the sample size. Second, the endpoint needs to be defined more specifically, since PFS includes local or regional relapse and metastatic organs beyond the liver. Therefore, the follow-up schedules in these different disease progression types need to be further explored. Third, cost-effectiveness should be analyzed. Last, there are other reported prognostic indicators that were not entered into the RSF tool.

We developed an RSF machine learning method to calculate the disease progression risk per month for BBHCC patients with CR. Afterwards, we established a surveillance strategy that was

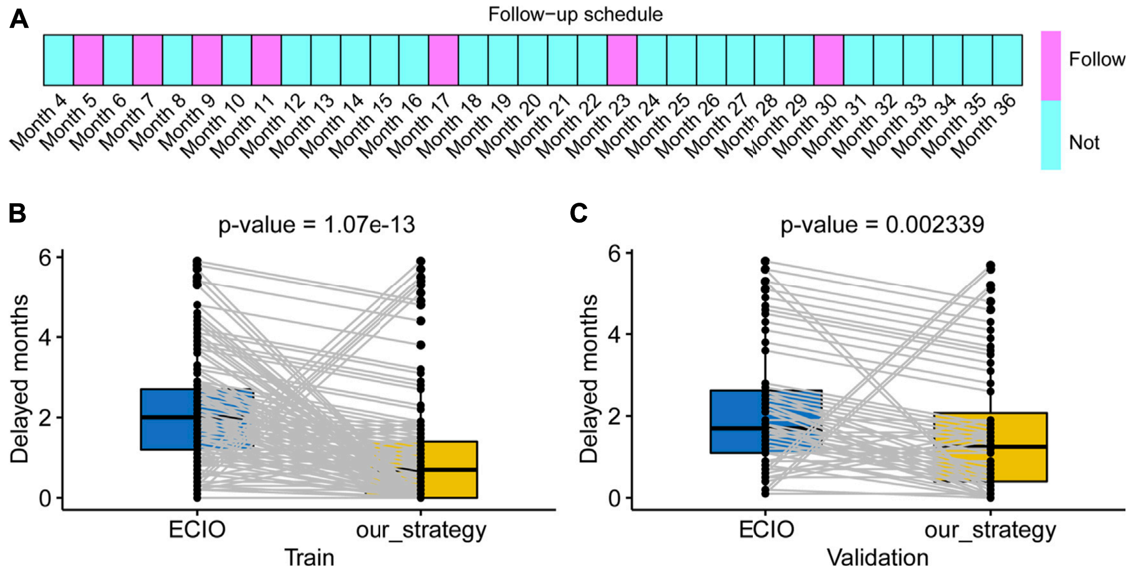


FIGURE 5 | Our schedule and the clinical application. **(A)** The boxes panel show the months that should be followed. **(B,C)** Our schedule produces less delayed detection in both the train and validation cohorts. ECIO: European Conference on Interventional Oncology.

more effective than the existing surveillance strategies. Our follow-up schedule might shed light on individualized surveillance for BBHCC CR patients.

DATA AVAILABILITY STATEMENT

The original contribution 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 the Sun Yat-sen University Cancer Center institutional ethics committee. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Q-FC and MZ designed this study. All authors collected data. Q-FC performed the statistical analysis and wrote the draft. All authors critically analysed the draft, reviewed literature, and approved the final version of the manuscript.

FUNDING

This work was primarily supported by the National Natural Science Foundation of China (No. 8207070122) and the Natural Science Foundation of Guangdong (2019A1515010751) and National Key Research and Development Program (2019YFC1316204).

ACKNOWLEDGMENTS

We thank Sipeng Shen, an assistant professor from department of biostatistics of the Nanjing Medical University, for his review concerning the statistical methods of this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2021.667641/full#supplementary-material

REFERENCES

Boas, F. E., Do, B., Louie, J. D., Kothary, N., Hwang, G. L., Kuo, W. T., et al. (2015). Optimal Imaging Surveillance Schedules after Liver-Directed Therapy for Hepatocellular Carcinoma. *J. Vasc. Interv. Radiol.* 26 (1), 69–73. doi:10.1016/j.jvir.2014.09.013

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer J. clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Chen, L.-T., Martinelli, E., Cheng, A.-L., Pentheroudakis, G., Qin, S., Bhattacharyya, G. S., et al. (2020). Pan-Asian Adapted ESMO Clinical Practice Guidelines for the Management of Patients with Intermediate and

Advanced/relapsed Hepatocellular Carcinoma: a TOS-ESMO Initiative Endorsed by CSCO, ISMPO, JSMO, KSMO, MOS and SSO. *Ann. Oncol.* 31 (3), 334–351. doi:10.1016/j.annonc.2019.12.001

Chen, Q.-F., Jia, Z.-Y., Yang, Z.-Q., Fan, W.-L., and Shi, H.-B. (2017). Transarterial Chemoembolization Monotherapy versus Combined Transarterial Chemoembolization-Microwave Ablation Therapy for Hepatocellular Carcinoma Tumors ≤5 Cm: A Propensity Analysis at a Single Center. *Cardiovasc. Intervent Radiol.* 40 (11), 1748–1755. doi:10.1007/s00270-017-1736-8

Chen, Q.-F., Li, W., Wu, P., Shen, L., and Huang, Z.-L. (2019). Alternative Splicing Events Are Prognostic in Hepatocellular Carcinoma. *Aging* 11 (13), 4720–4735. doi:10.18632/aging.102085

Chen, Q. F., Huang, T., Si-Tu, Q. J., Wu, P., Shen, L., Li, W., et al. (2020). Analysis of Competing Endogenous RNA Network Identifies a Poorly Differentiated Cancer-specific RNA Signature for Hepatocellular Carcinoma. *J. Cel Biochem* 121 (3), 2303–2317. doi:10.1002/jcb.29454

Day, R. W., Brudvik, K. W., Vauthey, J.-N., Conrad, C., Gottumukkala, V., Chun, Y.-S., et al. (2016). Advances in Hepatectomy Technique: Toward Zero Transfusions in the Modern Era of Liver Surgery. *Surgery* 159 (3), 793–801. doi:10.1016/j.surg.2015.10.006

Di Sandro, S., Centonze, L., Centonze, L., Pinotti, E., Lauterio, A., De Carlis, R., et al. (2019). Surgical and Oncological Outcomes of Hepatic Resection for BCLC-B Hepatocellular Carcinoma: a Retrospective Multicenter Analysis Among 474 Consecutive Cases. *Updates Surg.* 71 (2), 285–293. doi:10.1007/s13304-019-00649-w

Forner, A., Reig, M., and Bruix, J. (2018). Hepatocellular Carcinoma. *The Lancet* 391 (10127), 1301–1314. doi:10.1016/S0140-6736(18)30010-2

Kanwal, F., and Singal, A. G. (2019). Surveillance for Hepatocellular Carcinoma: Current Best Practice and Future Direction. *Gastroenterology* 157 (1), 54–64. doi:10.1053/j.gastro.2019.02.049

Labgaa, I., Taffé, P., Martin, D., Clerc, D., Schwartz, M., Kokudo, N., et al. (2020). Comparison of Partial Hepatectomy and Transarterial Chemoembolization in Intermediate-Stage Hepatocellular Carcinoma: A Systematic Review and Meta-Analysis. *Liver cancer* 9 (2), 138–147. doi:10.1159/000505093

Li, W., Chen, Q.-F., Huang, T., Wu, P., Shen, L., and Huang, Z.-L. (2020). Identification and Validation of a Prognostic lncRNA Signature for Hepatocellular Carcinoma. *Front. Oncol.* 10, 780. doi:10.3389/fonc.2020.00780

Liu, J.-N., Kong, X.-S., Huang, T., Wang, R., Li, W., and Chen, Q.-F. (2020). Clinical Implications of Aberrant PD-1 and CTLA4 Expression for Cancer Immunity and Prognosis: A Pan-Cancer Study. *Front. Immunol.* 11, 2048. doi:10.3389/fimmu.2020.02048

Maas, M., Beets-Tan, R., Gaubert, J.-Y., Gomez Munoz, F., Habert, P., Klompenhouwer, L. G., et al. (2020). Follow-up after Radiological Intervention in Oncology: ECIO-ESOI Evidence and Consensus-Based Recommendations for Clinical Practice. *Insights Imaging* 11 (1), 83. doi:10.1186/s13244-020-00884-5

Taylor, J. M. G. (2011). Random Survival Forests. *J. Thorac. Oncol.* 6 (12), 1974–1975. doi:10.1097/JTO.0b013e318233d835

Trinchet, J.-C., Chaffaut, C., Bourcier, V., Degos, F., Henrion, J., Fontaine, H., et al. (2011). Ultrasonographic Surveillance of Hepatocellular Carcinoma in Cirrhosis: a Randomized Trial Comparing 3- and 6-month Periodicities. *Hepatology* 54 (6), 1987–1997. doi:10.1002/hep.24545

Tsilimigras, D. I., Bagante, F., Moris, D., Hyer, J. M., Sahara, K., Paredes, A. Z., et al. (2020). Recurrence Patterns and Outcomes after Resection of Hepatocellular Carcinoma within and beyond the Barcelona Clinic Liver Cancer Criteria. *Ann. Surg. Oncol.* 27 (7), 2321–2331. doi:10.1245/s10434-020-08452-3

Vogel, A., Cervantes, A., Chau, I., Daniele, B., Llovet, J. M., Meyer, T., et al. (2018). Hepatocellular Carcinoma: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* 29 (Suppl. 4), iv238–iv255. doi:10.1093/annonc/mdy308

Wu, F., Chen, Q., Liu, C., Duan, X., Hu, J., Liu, J., et al. (2020). Profiles of Prognostic Alternative Splicing Signature in Hepatocellular Carcinoma. *Cancer Med.* 9 (6), 2171–2180. doi:10.1002/cam4.2875

Wu, Y., Shen, L., Qi, H., Cao, F., Chen, S., Xie, L., et al. (2020). Surveillance Strategy for Patients with BCLC Stage B Hepatocellular Carcinoma after Achieving Complete Remission: Data from the Real World. *Front. Oncol.* 10, 574804. doi:10.3389/fonc.2020.574804

Xie, D.-Y., Ren, Z.-G., Zhou, J., Fan, J., and Gao, Q. (2020). 2019 Chinese Clinical Guidelines for the Management of Hepatocellular Carcinoma: Updates and Insights. *Hepatobiliary Surg. Nutr.* 9 (4), 452–463. doi:10.21037/hbsn-20-480

Zhang, R., Shen, L., Zhao, L., Guan, Z., Chen, Q., and Li, W. (2018). Combined Transarterial Chemoembolization and Microwave Ablation versus Transarterial Chemoembolization in BCLC Stage B Hepatocellular Carcinoma. *Diagn. Interv. Radiol.* 24 (4), 219–224. doi:10.5152/dir.2018.17528

Zhou, G.-Q., Wu, C.-F., Deng, B., Gao, T.-S., Lv, J.-W., Lin, L., et al. (2020). An Optimal Posttreatment Surveillance Strategy for Cancer Survivors Based on an Individualized Risk-Based Approach. *Nat. Commun.* 11 (1), 3872. doi:10.1038/s41467-020-17672-w

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.

frontiers
in Public Health

ORIGINAL RESEARCH
published: 10 September 2021
doi: 10.3389/fpubh.2021.718594

An Intelligent Control Model of Credit Line Computing in Intelligence Health-Care Systems

Rong Jiang [1,2,3], *Wenxuan Wu* [1,2,3,4], *Yimin Yu* [1,2,3,4]* *and Feng Ma* [4]

[1] Institute of Intelligence Applications, Yunnan University of Finance and Economics, Kunming, China, [2] Key Laboratory of Service Computing and Safety Management of Yunnan Provincial Universities, Kunming, China, [3] Kunming Key Laboratory of Information Economy & Information Management, Kunming, China, [4] School of Information, Yunnan University of Finance and Economics, Kunming, China

Technologies such as machine learning and artificial intelligence have brought about a tremendous change to biomedical computing and intelligence health care. As a principal component of the intelligence healthcare system, the hospital information system (HIS) has provided great convenience to hospitals and patients, but incidents of leaking private information of patients through HIS occasionally occur at times. Therefore, it is necessary to properly control excessive access behavior. To reduce the risk of patient privacy leakage when medical data are accessed, this article proposes a dynamic permission intelligent access control model that introduces credit line calculation. According to the target given by the doctor in HIS and the actual access record, the International Classification of Diseases (ICD)-10 code is used to describe the degree of correlation, and the rationality of the access is formally described by a mathematical formula. The concept of intelligence healthcare credit lines is redefined with relevance and time Windows. The access control policy matches the corresponding credit limit and credit interval according to the authorization rules to achieve the purpose of intelligent control. Finally, with the actual data provided by a Grade-III Level-A hospital in Kunming, the program code is written through machine learning and biomedical computing-related technologies to complete the experimental test. The experiment proves that the intelligent access control model based on credit computing proposed in this study can play a role in protecting the privacy of patients to a certain extent.

Keywords: machine learning, biomedical computing, intelligence healthcare, privacy security, intelligent access control, credit line

OPEN ACCESS

Edited by:
Ying Li,
Zhejiang University, China

Reviewed by:
Qiang Hu,
Qingdao University of Science and Technology, China
Xianwei Li,
Waseda University, Japan
Donghui Yang,
Southeast University, China

****Correspondence:***
Yimin Yu
yuyimin@ynufe.edu.cn

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

Received: *01 June 2021*
Accepted: *31 July 2021*
Published: *10 September 2021*

Citation:
Jiang R, Wu W, Yu Y and Ma F (2021) An Intelligent Control Model of Credit Line Computing in Intelligence Health-Care Systems. Front. Public Health 9:718594. doi: 10.3389/fpubh.2021.718594

INTRODUCTION

Medical big data (1) is a branch of big data in the field of biomedicine. It refers to the data related to life, health, and medical care generated in activities related to human health, mainly from intelligent medical systems such as clinical data, hospital, operation, biomedical research, disease prevention and control, health protection and food safety, public health and health management data, health care and other aspects (2). In these massive amounts of data, there are opportunities. If the data generated by smart medical care can be flexibly called after biomedical calculations, data pressure can be converted into data advantage (3, 4).

In terms of biomedicine, individual users have become an important source of data. The private information generated by smart medical treatment often means unspeakable pain for individuals. The leakage of such negative information has become a huge hidden danger in the era of big data (5). In the past, most patients maintained their personality and dignity through self-forgetting and the privacy of medical institutions (6). Nowadays, the ubiquitous smart medical equipment and cloud storage and cloud computing functions, such as placing users in a transparent glass room. Our every move may be recorded, and the electronic health records generated by the widely used smart medical system and smart medical equipment make it difficult for patients to hide their privacy. According to a security report released by Trust wave, more than 90% of the investigators believe that there are more and more cyberattacks against the medical field, but the budget for protecting sensitive patient information is <10% (7). Once the criminals steal medical data, they can easily learn the name, home address, contact information, test report, diagnosis results, and even medical insurance and other important information of the patient, and use this to falsify the data to defraud or purchase medical equipment. Therefore, the consequences of data theft in the medical field are very serious. More than two million people in the United States will become victims each year. The loss caused by this is as high as $13,500, and it will take hundreds of hours to solve this problem. In 2015, the social security system became the hardest hit area for personal information leakage, etc. (8). These incidents seriously violated the privacy and legal rights of users. At present, both the public and the government have begun to pay attention to personal privacy issues in medicine (9). In the United States, electronic health data are also being prepared for an online transformation. Dosia (a non-profit coalition of major employers), Google Health, Microsoft Health Vault, and other network services are driving this transformation. These services are seeking expanded role in the United States health-care system that values 21,000 dollars (10).

With China's accession to the WTO and the acceleration of social information, whether it has a fully functional intelligent medical system, it has become an important indicator to measure the comprehensive strength of a hospital (11, 12). A perfect hospital information system (HIS) includes outpatient management, hospitalization management, drug management, multiple subsystems, such as electronic medical records and financial management. The high integration between the various subsystems improves the overall operating efficiency of the hospital, improves the medical environment of the patient, and at the same time provides data-driven support for the management of hospital, clinical, etc. The electronic medical record system (13) includes the electronic medical record of the doctor and nurse. The main function is to save the medical records of the patients electronically. It not only includes the medication information of the patient, but also includes the treatment record, laboratory and examination records, and other information of the patient. The doctor is giving the past medical records and medical history of the patient during treatment. It can more accurately analyze the condition of the patient and treat the patient (7). However, due to the use of HIS, doctors can access a large amount of medical information, and the resulting medical problem of privacy leakage is also very tricky. When the system security and data security are not guaranteed, the intelligent medical system is fragile, which will not only cause great troubles for medical work, but also greatly reduce the prestige of the intelligent medical system (14).

Given the medical privacy leakage risk arising from the widespread use of intelligent medical systems today, this study proposes an access control model based on credit line calculations for intelligent medical systems. In this model, when doctors use the intelligent medical system to diagnose patients, they use historical records to calculate credit lines, and dynamically restrict doctors' access rights based on their credit capabilities. Don't give unnecessary permissions, and will not affect the normal work of doctors, try to comply with the A principle (15). The main steps of model realization are as follows:

(1) Through similarity function calculation, the results obtained by the mathematical method can be used to describe whether the inquiring behavior of the doctor is reasonable.
(2) The appropriate weight calculation method is used to obtain the weight value so that the unreasonable behavior of the doctor is easy to lead to the decline of the credit limit, but the reasonable behavior of the doctor will not affect his trust limit.
(3) According to the credit limit of the user, match the corresponding trust interval to achieve the ability to limit the access authority of the user. Doctors with a high credit limit will become larger and larger. On the contrary, doctors with a low credit limit will become smaller and smaller, until it is lower than the credit line threshold, and the visit is forbidden. In the existing model, doctors select medical records based on randomly assigned work goals, or medical records are selected based on the work goals selected by the doctors themselves, and this study defines that the doctors may not necessarily choose an honest work goal based on the preliminary examination information of the patient. In addition, the doctor will give a more accurate final diagnosis only after checking the medical records. The model can properly describe the real diagnosis process of the doctor, and it is more in line with the actual situation.

The contributions of this study are as follows:

1. Some contents have been added to the doctor behavior model in the study (16), which improves the performance of the model in screening curious doctors.
2. In a relatively mature intelligent medical system, the concept of the credit line is introduced as the carrier of medical trust computing.
3. After comparison, more appropriate trust calculation and weight calculation methods are selected to achieve the effect of using historical records to restrain the behavior of doctors and reduce the risk of privacy disclosure in the medical field.

RELATED WORKS

If divided by authorization strategy, the access control model can be divided into the following: traditional access control model (DAC/MAC), role-based access control (RBAC) model (17), task and workflow-based access control (TBAC) model (18), task-based and role-based access control (TRBAC) model, etc. (19). RBAC model permissions are associated with roles, and users become members of corresponding roles, which greatly simplifies the management of permissions. However, the RBAC model cannot be directly used for more complex forms of access control (20). Goyal et al. (21) proposed an attribute-based mechanism to protect data and avoid setting data owner rules. However, the main disadvantage of using attribute-based methods is that it will bring a high workload to the user side. For most ordinary users, with limited knowledge of rules or strategy design, creating a complex data access mechanism in a medical environment is an arduous task (22). The workload brought by traditional access

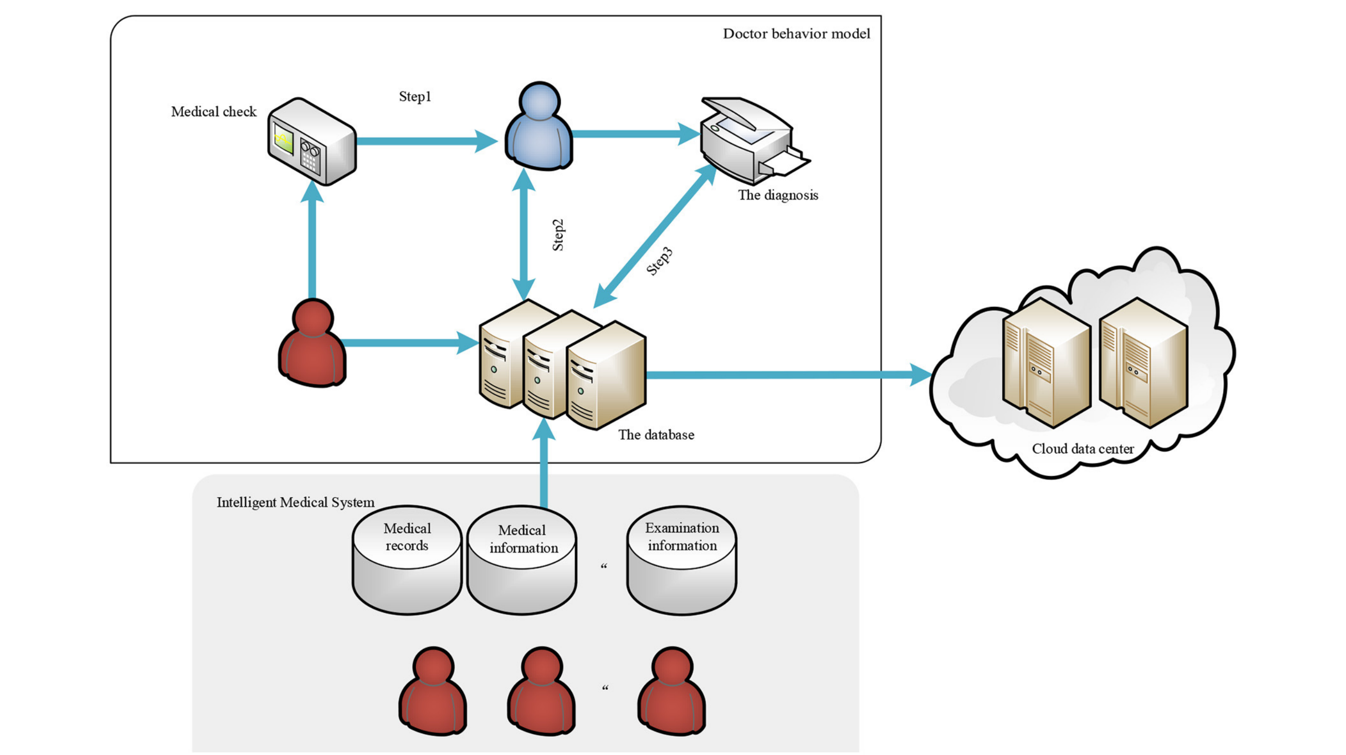


FIGURE 1 | Doctor behavior model.

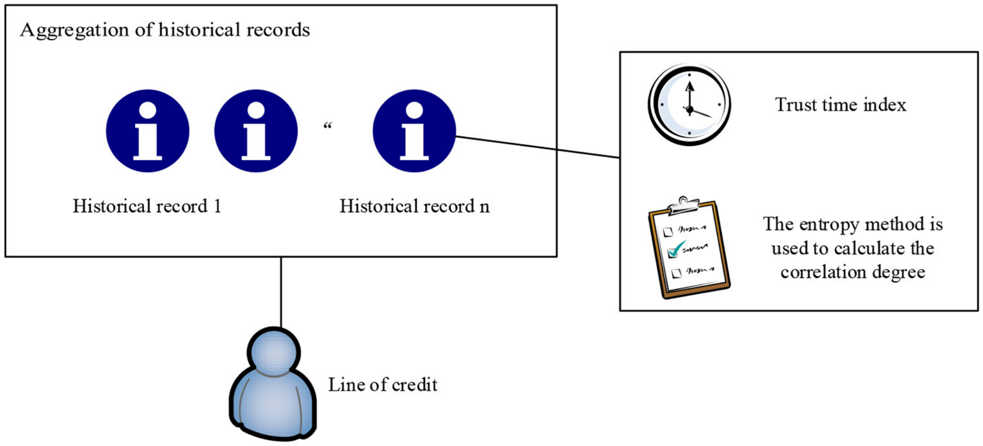

Aggregation of historical records
"
Historical record 1
Historical record n
Trust time index
The entropy method is used to calculate the correlation degree
Line of credit

FIGURE 2 | Trust calculation process.

control obviously cannot adapt to the situation of massive data in the context of big data.

Health and medical big data are important basic strategic resources of the country. Traditional database achieves security and privacy protection through data granularity-based security control, but the operation of big data still lacks effective security protection measures (23). The realization technology of medical big data information security includes access control and password technology. Data privacy implementation technologies include obfuscation, anonymity, differential privacy, and encryption (16). At present, the prominent problems in the use of HIS medical data mainly include the following:

(a) Security issues: dynamic permissions are granted. The existing medical information system does not consider the wishes of patients, and the scope of medical data that doctors can access is not detailed enough. According to the actual needs of doctors, there is little research on medical information systems that dynamically grant data access rights to achieve fine-grained data access.
(b) Data sharing issues: Doctors and researchers have strict restrictions when accessing and sharing medical data (24).

The important issue studied in this study is access control, with the focus on protecting information from unauthorized access (25). Wang et al. (26) designed a secure authentication algorithm to limit the access rights of access objects in the electronic medical record (EMR) system. Zhu et al. proposed a user-friendly, easy-to-manage, attribute-based access control (ABAC) for cloud storage services in 2015. This mechanism defines the priority of attributes and refines the granularity of data access control in the cloud environment (27). Liu et al. (28) is based on the trust-based access control model, which combines dynamic hierarchical fuzzy systems with trust evaluation, layered the attributes related to trust in the cloud manufacturing environment, and proposed a multi-attribute fuzzy trust evaluation access control scheme. Gao (29) built a flexible dynamic access control model to make up for the lack of static policies, making the original role, the static authorization access of permissions is transformed into a model that can dynamically authorize users. Zhang and Zhou (30) to solve the problems of access resources insufficient flexibility and preset allocation of permissions in the traditional role-based access control system, improved compatibility of access control, refine the granularity of access control, and propose a dynamic multilevel access control model based on trust. The static role and dynamic trust degree of the user obtain the corresponding authority authorization (31). Based on the traditional free access control (DAC) and RBAC model, a context-sensitive access control method is proposed, which strictly follows regulatory and technical standards in the health-care field to ensure authorized access (32). The literature found that existing symmetric and asymmetric encryption technologies have complex key management and certificate management problems. In response

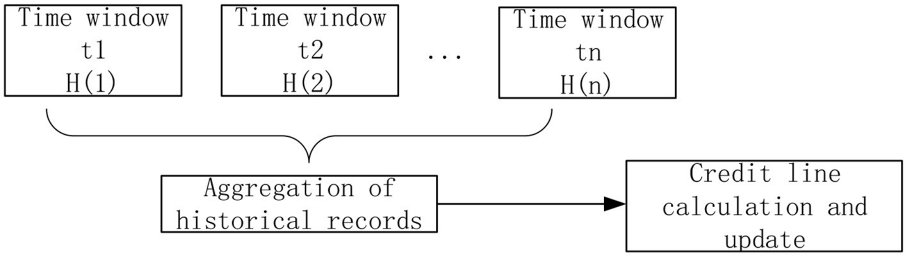

Time window
t1
H(1)
Time window
t2
H(2)
...
Time window
tn
H(n)
Aggregation of historical records
Credit line calculation and update

FIGURE 3 | History aggregation.

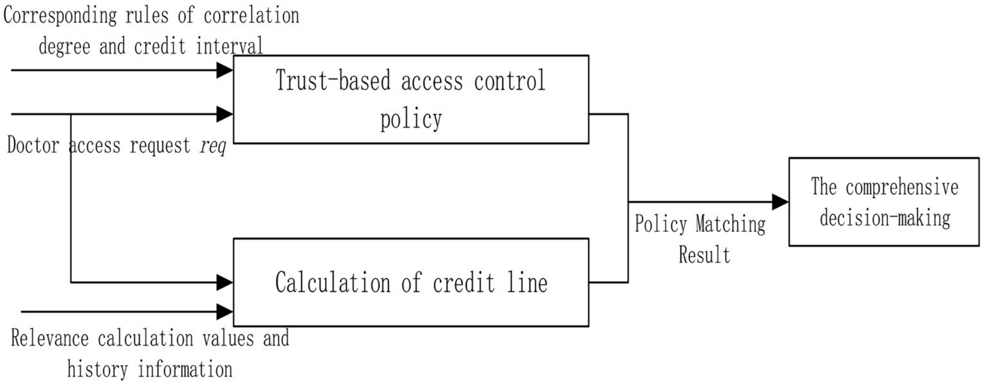

Corresponding rules of correlation degree and credit interval
Doctor access request *req*
Trust-based access control policy
Calculation of credit line
Relevance calculation values and history information
Policy Matching Result
The comprehensive decision-making

FIGURE 4 | Access control scheme.

to these problems, the policy-based access control (PBAC) model and the purpose of joining conditions are proposed. IBE encryption technology medical data can access control scheme. Yang et al. (33) proposed a privacy protection medical big data system with adaptive access control. Through a new dual access control mechanism, the mechanism has adaptive capabilities for both normal and emergency medical data access. In summary, although various access control models have been expanded by previous researchers, yet studies on trust computing and access control models in the field of intelligent medical research are inadequate (34). The authorization method for access control in the process of diagnosis by doctors and treatment is relatively simple and restrictive. The problem of insufficient binding still exists.

Therefore, it is more necessary to explore a dynamic trust computing method and access control scheme, which are more suitable for specific occasions of medical access, and dynamically adjust the permissions of doctors to access medical resources through the results of trust computing, to improve the privacy protection performance of the model.

MODEL DESIGN

Based on the research of a large number of existing intelligent medical systems (35), this section presents the following behavior model which is more suitable for the actual situation. Taking the process of diagnosis by doctors and treatment of patients as the research object, the diagnosis and treatment behavior of doctors is abstracted into a model from the three aspects of examination information of doctors, access of doctors to medical data, and the diagnosis results are given. Then, the definition of credit limit, correlation calculation, and weight determination method are given.

Doctor Behavior Model

This is shown in **Figure 1**, for the diagnosis and treatment process of each patient, we call it a task. In the task, the diagnosis by doctor and treatment steps are generally the following: first browse the basic information of the patient, such as name, age, and past medical history. If the patient has a medical history in the hospital, the doctor can check the past examination items and results of the patient through the HIS database. Then, the patient will receive new test results according to the arrangement of the doctor, whose arrangement is also stored in the HIS database. The doctor can view the test results of the patient and certain related medical records (such as medical imaging data of other similarly diagnosed patients in the database, etc.) to obtain the final diagnosis for the patient. The medical data in the HIS database that doctors can view involve some sensitive information about the patient, but considering moral factors and the cost of leaks, hypothetical model doctors in China will not disclose any information about their patients. Based on the

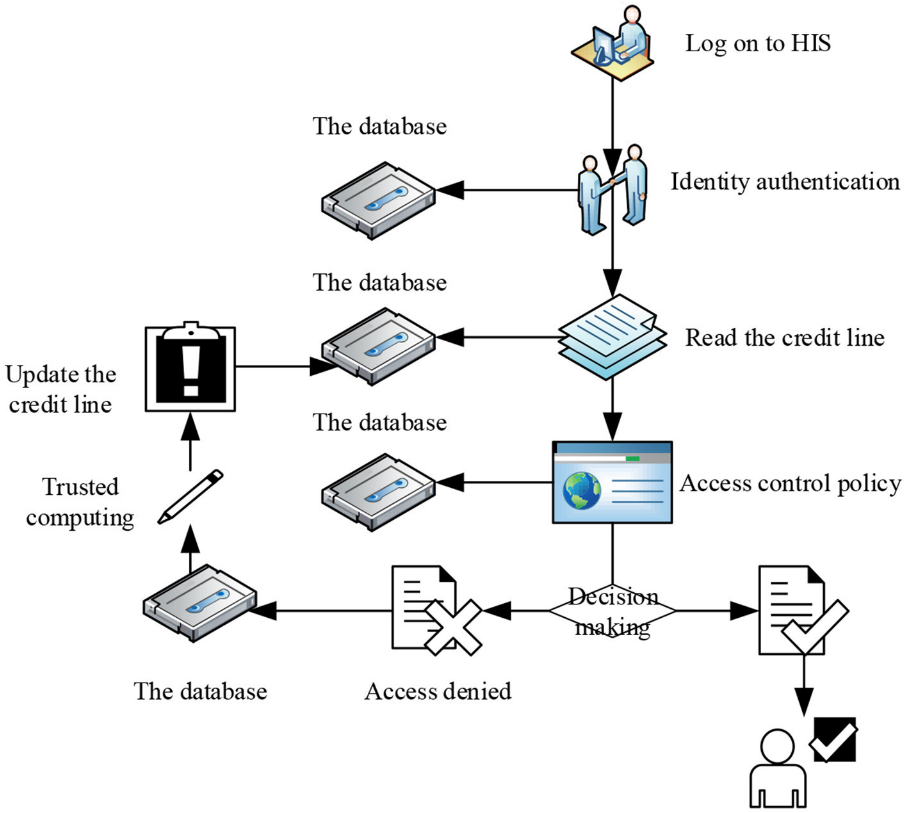


FIGURE 5 | Access control flow chart.

above behavior model, the behavior of privacy leakage of curious doctors will occur in the following three steps:

Step 1. The correlation between the examination information of the patient and the initial diagnosis given by the doctor is low. For example, the results of the examination of a patient can directly indicate that the patient is unlikely to have an infectious disease. However, the doctor still gave a preliminary diagnosis that the patient may have an infectious disease, and then consulted the relevant medical records of the patient with infectious disease based on the false preliminary diagnosis.
Step 2. Suppose that the doctor gave a correct preliminary diagnosis consistent with the examination information in Step 1, but inquired about unnecessary medical records when accessing the medical records based on the preliminary diagnosis.
Step 3. Assume that the doctor is operating normally in Step 1 and Step 2, but the final diagnosis has a low correlation with the medical records queried. It is suspected that the doctor had accessed unnecessary medical records.

Give the formal description of the symbol as follows, and abstract the process:

E: A collection of examination information;
P: A collection of primary diagnoses;
F: A collection of final diagnoses;
R: A collection of medical records.

$S_1 : E, P \rightarrow [0, 1]$: Define the correlation function between inspection information and preliminary diagnosis, $e \in E, p \in P$, where the return value of the function reflects the degree of correlation between the two in a certain diagnosis and treatment process.
$S_2 : P, R \rightarrow [0, 1]$: Define the correlation function between the initial diagnosis and medical records, $p \in P, r \in R$, where the return value of the function reflects the degree of correlation between the two in a certain diagnosis and treatment process.
$S_3 : R, F \rightarrow [0, 1]$: Define the correlation function between medical records and the final diagnosis, $r \in R, f \in F$, where the return value of the function reflects the degree of correlation between the two in a certain diagnosis and treatment process.

Trust Attribute System

The existing trust system for access control is relatively single, usually divided into direct trust and indirect trust. In the context of diagnosis and treatment by doctors, this trust model cannot accurately assess the credibility of behavior of doctors. Doctors have extensive access to patients and medical records in intelligent health-care systems, but there is a lack of effective direct trust between each doctor or between patients and other doctors (who do not diagnose themselves). According to the behavioral characteristics of diagnosis and treatment by doctors and the particularity of the structure of medical resource system, indirect trust is not considered in the trust attribute, but only the historical visit records of doctors will directly affect their credit line.

The concept of credit originates from the financial field and refers to the funds provided by banks to non-financial users, including but not limited to various businesses, such as loans. The credit line means the highest credit value given to users by the bank after calculation and evaluation during the credit period.

This study introduces the concept of credit line in the intelligent medical system, and redefines it as a comprehensive evaluation of the history records, access behavior, and other factors of medical information system, and calculates and grants credit line of the doctor user for overdraft use. The credit limit is calculated by reading the history of the doctor through HIS. The continuous integrity behavior record of the doctor can help increase the credit limit, and high-risk behaviors will lead to a reduction in the credit limit, thereby realizing dynamic access control to medical data. The history record includes two sub attributes of the trust time window and operational relevance. This is shown in **Figure 2**.

Correlation Calculation

In the intelligent medical system, doctors use electronic medical records to record the medical treatment of patients during the medical treatment process. This study introduces the International Classification of Diseases (ICD) as the code used by doctors in the diagnosis of electronic medical records. Suppose a certain disease in the electronic medical record is represented by an ICD code. Afterward, you can use the element group to write $a_1, a_2, a_3, ..., a_n$, the elements representing the disease at each location are divided into different subcategories according to the ICD code and expressed as $a_{i1}, a_{i2}, a_{i3}, ..., a_{in}$, where n represents the number of subcategories of the disease. To calculate the similarity, prepare to construct the initial judgment matrix EQ from the diseases in the medical records as follows:

$$EQ = \begin{Bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{Bmatrix} \tag{1}$$

The three steps involved in diagnosis process by a doctor have different risks of privacy leakage. This section focuses on the calculation method of the correlation function. There are many methods to measure similarity (36), and the distance measurement is Minkowski distance, Euclidean distance, Manhattan distance, Hamming distance, etc. Commonly used similarity coefficients include cosine similarity, Pearson correlation coefficient, Jaccard correlation coefficient, etc. The traditional methods for measuring the similarity of two individuals are commonly used. Cosine similarity is defined as follows: regarding user information as an n-dimensional vector,

TABLE 1 | Correlation degree—credit interval rules.

	Correlation	The line of credit is in the range
1	0.8	$t_3 < t \le t_4$
2	0.6	$t_2 < t \le t_3$
3	0.4	$t_1 < t \le t_2$
4	0.2	$t_0 < t \le t_1$

the similarity is calculated as the cosine of the angle between the vectors. The similarity between individual i and j is recorded as shown:

$$\mathrm{sim}\left(i,j\right)=\cos\left(\overrightarrow{i},\overrightarrow{j}\right)=\frac{\overrightarrow{i}\cdot\overrightarrow{j}}{\left\|\overrightarrow{i}\right\|*\left\|\overrightarrow{j}\right\|}\tag{2}$$

$\overrightarrow{i}\cdot\overrightarrow{j}$ is the inner product and $\left\|\overrightarrow{i}\right\|*\left\|\overrightarrow{j}\right\|$ is the vector product.

In addition to cosine similarity, there is also correlation similarity. The similarity is measured by calculating the correlation coefficient between i and j of the item. Determine the common user set U of i and j, and the correlation similarity is defined as follows:

$$\mathrm{sim}\left(i,j\right)=\frac{\sum_{u\in U}\left(R_{ui}-R_i\right)\left(R_{uj}-R_j\right)}{\sqrt{\sum_{u\in U}\left(R_{ui}-R_i\right)^2}\sqrt{\sum_{u\in U}\left(R_{uj}-R_j\right)^2}}\tag{3}$$

The similarity is obtained by using cosine or correlation similarity, because the medical data category base is relatively large, and the data are dense, and the wrong conclusion with high similarity is obtained. When calculating the similarity, the Jaccard similarity coefficient is introduced to calculate the privacy leakage risk of the doctor in each step of the diagnosis process. The Jaccard similarity coefficient is also called the Jaccard index, which is used to compare the similarity and difference statistics of a limited sample set (37). Assume that sets I and F are the

TABLE 2 | Expert opinion weights.

	α	β	γ
Expert 1	0.6	0.2	0.2
Expert 2	0.3	0.6	0.1
Expert 3	0.5	0.3	0.2
…	…	…	…
Expert 9	0.4	0.3	0.3
Expert 10	0.2	0.3	0.5
MSD calculate	0.49	0.16	0.35

TABLE 3 | The results of doctors were judged by maximum score deviation (MSD) weights.

		Expert 3			**Expert 10**			**MSD**		
		Step 1	**Step 2**	**Step 3**	**Step 1**	**Step 2**	**Step 3**	**Step 1**	**Step 2**	**Step 3**
Special access	1		○			×			○	
	2		•			×			○	
Malicious access	1	○			•			○		
	2		○			○			○	
	3			•			○			○

•, Unable to determine; ○, right; ×, false.

TABLE 4 | Changes of historical aggregate values of doctors.

Doctor	**CT(1)**	**CT(2)**	**CT(3)**	**CT(4)**	**CT(5)**	**CT(6)**	**CT(7)**	**CT(8)**
Honest doctor	0.66	0.57	0.62	0.59	0.67	0.63	0.60	0.59
Special doctor	0.63	0.60	0.58	0.56	0.64	0.51	0.61	0.56
Curious doctor	0.51	0.43	0.57	0.64	0.66	0.67	0.58	0.47

TABLE 5 | The variation of the mean value of credit line with α.

Doctor	**Δ**	**α = 0.95**	**Δ**	**α = 0.5**	**Δ**	**α = 0.2**	**<Δ**
Honest doctor	0.16	0.596	0.11	0.604	0.12	0.583	0.13
Special doctor	0.31	0.557	0.12	0.585	0.14	0.602	0.19
Curious doctor	0.34	0.531	0.26	0.538	0.17	0.551	0.25

initial diagnosis and the final diagnosis is described using ICD codes. Each code contains n public attributes, which indicate the category and subcategory of the disease. Each attribute in the code consists of a number or letter. To facilitate the calculation, the number will be represented by the set of 0 and 1. The Jaccard index can be written as J (I, F). The definition of the Jaccard index is as follows (38):

$$J\left(I,F\right)=\frac{I\bigcap F}{I\bigcup F} \tag{4}$$

Define that when I=F=Ø, J(I, F) = 1, the value range is [0,1], the larger the J value, the greater the similarity between the two samples. From this, the Jaccard distance can be obtained, and dJ (I, F) is used to represent the difference between the two samples:

$$\mathrm{d}_J\left(I,F\right)=1-J\left(I,F\right)=\frac{\left|I\bigcup F\right|-\left|I\bigcap F\right|}{\left|I\bigcup F\right|} \tag{5}$$

Taking the stomatology department as an example, suppose that the initial diagnosis given by a doctor is periodontitis. Define M11 as the number of ones in both I and F; M01 is the number of attributes of F that are 1 when the attribute of the set I is 0; M10 is the number of attributes of F in the set I that is 0 when the attribute is 1; M11 is the number of attributes of the set I and F that are 1. According to the above assumptions, the calculation method of Jaccard index and Jaccard coefficient can be obtained as follows:

$$M_{11}+M_{01}+M_{10}+M_{00}=\mathrm{n} \tag{6}$$

$$J=\frac{M_{11}}{M_{11}+M_{01}+M_{10}} \tag{7}$$

$$\mathrm{d}_J=\frac{M_{01}+M_{10}}{M_{11}+M_{01}+M_{10}} \tag{8}$$

According to the analysis of the above doctor behavior part, the doctor can directly contact the medical data of the non-attending patient during the diagnosis process, which is a high-risk reason for privacy leakage. To avoid a single error in the similarity calculation, cross-entropy is then used (39) to introduce the calculation of the similarity between two random variables.

Entropy is the expected value of the amount of information (40). Assuming that there is a random variable x with a value range of set X, its probability distribution function can be expressed as $\mathrm{p}\left(x\right) = \Pr\left(X=x\right), x \in X$, and defines the amount of information as $I\left(x_1\right) = -\lg\left(p\left(x_1\right)\right)$, the greater the probability of an event, the more $p\left(x_1\right)$ larger, the smaller the amount of information it carries (41). In the extreme case $p\left(x_1\right) = 1$, the amount of information is equal to zero, which means that when the probability of an event happening is 100%, then the occurrence of this event will not introduce too much information. When we know the amount of information to measure the uncertainty of the occurrence of an event, we can calculate the expectation $\left(E\left[I\left(x\right)\right]\right)$ for the additional information brought by all possible results, and the entropy can be defined as follows:

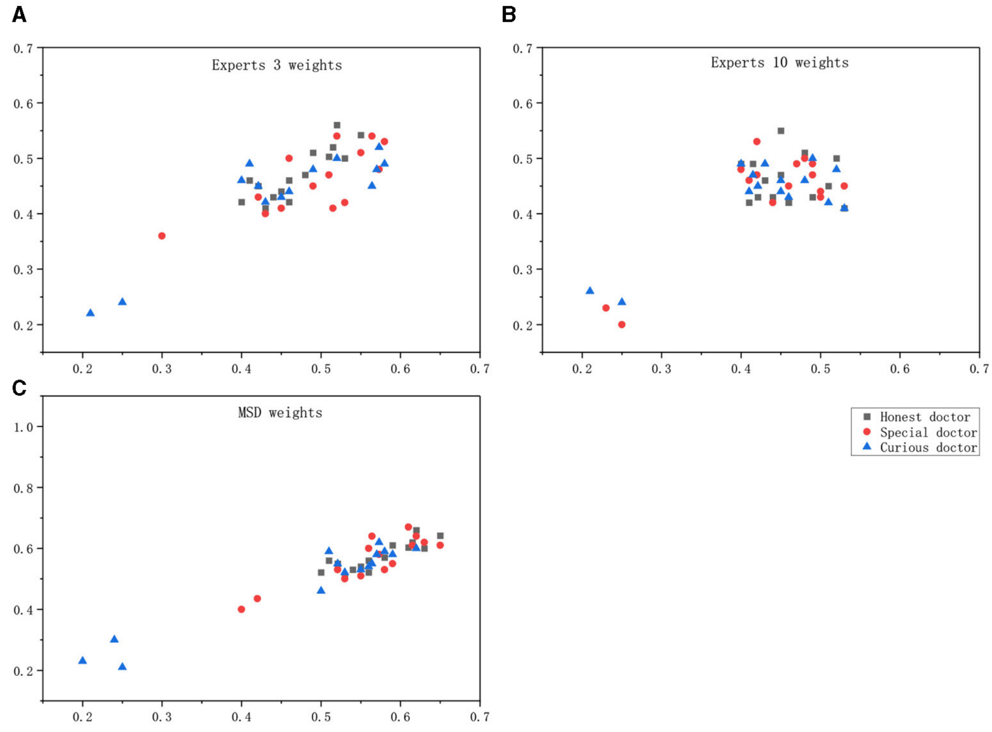


FIGURE 6 | Maximum score deviation (MSD) weight differentiating doctor effect. **(A)** Experts 3 weights, **(B)** Experts 10 weights, **(C)** MSD weights.

$$H(X) = Ep\left[\lg p(x)\right] = -\sum x \in Xp(x) \lg p(x) \tag{9}$$

According to the diseases and symptoms covered by ICD coding statistics, the average information required for each preliminary diagnosis (disease) was calculated as the threshold value. Suppose that the preliminary diagnosis obeys a random distribution p, and the interview records of a doctor obey a random distribution q. Then, cross-entropy is introduced to calculate the similarity degree of p and q. The expectation obtained according to the distribution p is *H (P)*. For the diagnosis process of doctors, the access records are discrete variables, and the p distribution is represented by the q distribution, which is called the cross-entropy.

$$H(\mathrm{p}) = \sum_i p(i) * \lg \frac{1}{p(i)} \tag{10}$$

$$H(p,q) = \sum_i p(i) * \lg \frac{1}{q(i)} \tag{11}$$

Assuming that p of a disease can be expressed as $[1, 0, 0]$ $[1, 0, 0]$, and q obtained by a doctor A's visit to the historical record is $[0.5, 0.4, 0.1]$ $[0.5, 0.4, 0.1]$, then according to the calculation method of cross-entropy of formula (11), the cross-entropy between the visit behavior of doctor in the process of diagnosis and the initial diagnosis given by doctor A can be obtained as follows:

$$\begin{aligned}
&H(p = [1, 0, 0], q = [0.5, 0.4, 0.1]) \\
&= -(1 * \lg 0.5 + 0 * \lg 0.4 + 0 * \lg 0.1) \\
&\approx 0.3H(p = [1, 0, 0], q = [0.5, 0.4, 0.1]) \\
&= -(1 * \lg 0.5 + 0 * \lg 0.4 + 0 * \lg 0.1) \\
&\approx 0.3
\end{aligned}$$

If the interview record Q of Doctor B with the same preliminary diagnosis is$[0.8, 0.1, 0.1]$ $[0.8, 0.1, 0.1]$, then, the cross-entropy between the visit and the preliminary judgment of Doctor B in the process of this diagnosis is follows:

$$\begin{aligned}
&H(p = [1, 0, 0], q = [0.8, 0.1, 0.1]) \\
&= -(1 * \lg 0.8 + 0 * \lg 0.1 + 0 * \lg 0.1) \\
&\approx 0.1H(p = [1, 0, 0], q = [0.8, 0.1, 0.1]) \\
&= -(1 * \lg 0.8 + 0 * \lg 0.1 + 0 * \lg 0.1) \\
&\approx 0.1
\end{aligned}$$

It can be seen from the calculation results that cross entropy value of Doctor B is small, that is, the operational correlation is higher. If the threshold value of this disease is known to be 0.2, then, it

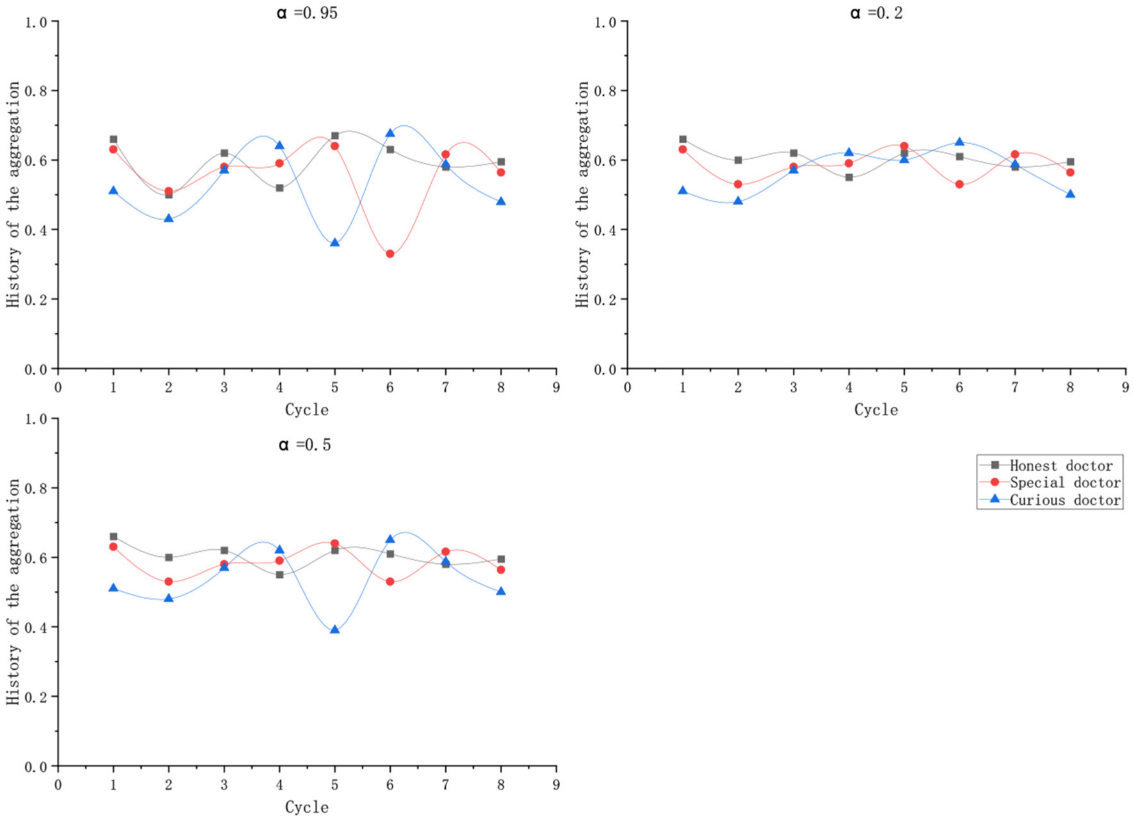


FIGURE 7 | Fluctuation of credit line.

can be concluded that Doctor B is accessing medical data safely, and Doctor A is suspected of a large privacy breach.

To calculate the accuracy of similarity, two calculation methods, namely, Jaccard coefficient and cross-entropy, were used to calculate the correlation degree of the diagnosis and treatment process. The final formula for calculating the correlation degree of the diagnosis and treatment process was as follows:

$$S=\frac{(1+\mathrm{d}_1)\,\alpha}{2H_1d_1}+\frac{(1+d_2)\,\beta}{2H_2d_2}+\frac{(1+d_3)\,\gamma}{2H_3d_3}\;(\alpha+\beta+\gamma=1) \qquad (12)$$

Because the weight cannot simply be given a definite value, it is determined by the vague advice given by experienced experts.

A review of the relevant literature and consultation with medical professionals has been discussed as follows:

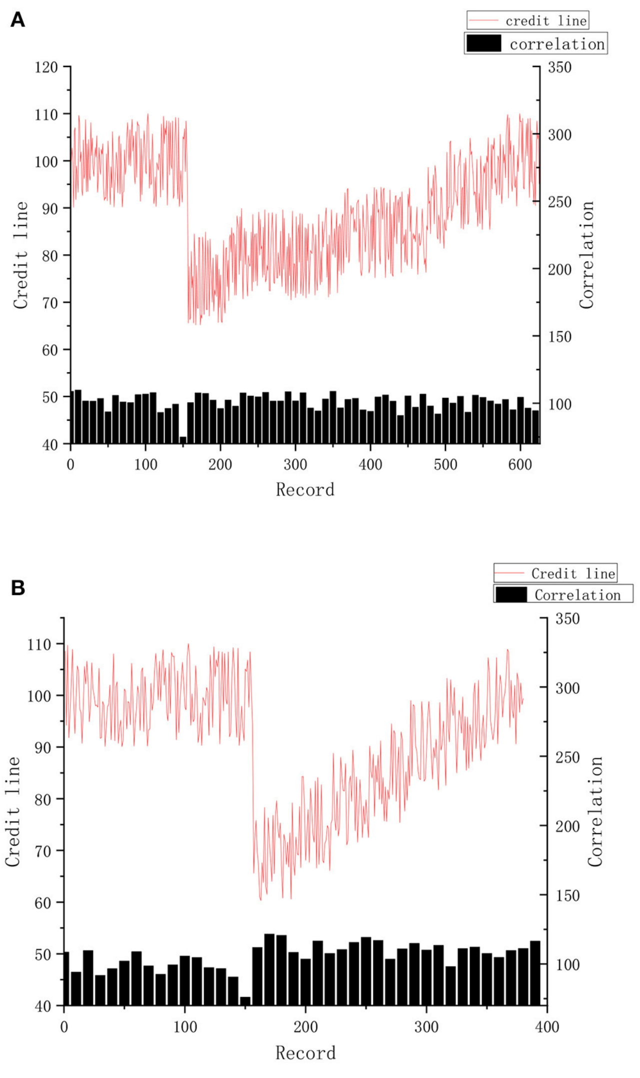


FIGURE 8 | (A,B) Change of credit limit and correlation degree.

Hypothesis A: To obtain certain medical records, a curious doctor falsifies the information of the primary diagnosis that does not match the inspection information, thereby, rationalizing the second step. However, even if qualified doctors encounter patients with special circumstances, they will not make a preliminary diagnosis with a correlation below the threshold based on the examination information of the patient. Therefore, the rules at this stage are very strong. Once the initial diagnosis is wrong, the correlation between Step 2 and Step 3 is normal, and the risk of leakage of medical record privacy is relatively high. Therefore, the weight corresponding to Step 1 needs to be relatively large. In this case, even if a curious doctor performs normal operations in Step 2 and Step 3, the credibility of the calculation will be greatly reduced.

Hypothesis B: If a qualified doctor diagnoses a patient with rare symptoms, the doctor needs to refer to more medical records to determine what disease the patient has. At this time, Step 1 is normal and the correlation of Step 2 is decreased, but according to medical records, the final judgment Step 3 should also be normal. Therefore, during the diagnosis, the weight of Step 2 can be appropriately relaxed, so that doctors can have a larger space for resource selection, and the diagnosis process of doctors in complicated cases will not be restricted.

Hypothesis C: If a curious doctor tries to imitate the behavior of an ordinary doctor in Hypothesis B, the curious doctor will naturally give a final diagnosis with low relevance to the medical record. Assuming that in the context of the medical environment, all doctors will perform their duties. A patient will not be diagnosed by only one doctor, so curious doctors will not insist on making a wrong final diagnosis to steal medical data. Therefore, in this case, the conclusion of the doctor based on a large number of irrelevant medical records will be less relevant to the initial diagnosis given by malicious intent.

Therefore, to distinguish between hypothesis B and hypothesis C, the weight of S3, namely γ, should also be large.

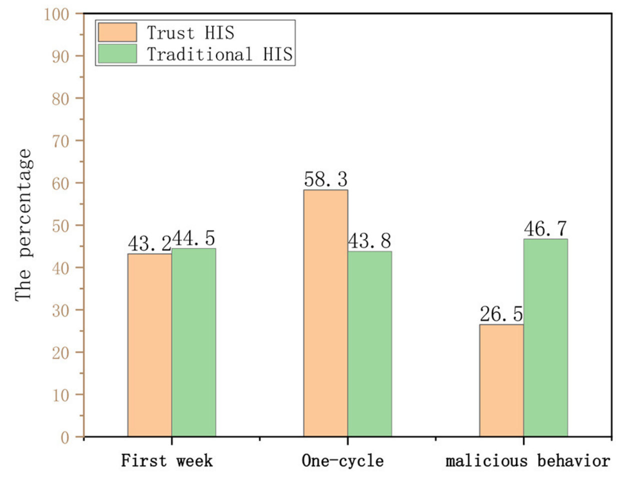


FIGURE 9 | Comparison of health information system (HIS) correlation degree and behavioral constraint ability.

According to the above hypothesis analysis, in each step, appropriate weights can provide doctors with a certain space for fault-tolerant visits or special situations requiring additional resources and can also effectively screen behaviors of curious doctors. Therefore, it is necessary to introduce appropriate weight determination technology. Xu and Zhou (42) further developed the maximum score deviation (MSD) method to obtain the weight of each index. The principle of the MSD method is that when multiple experts evaluate the evaluation factors, the higher the similarity with the evaluation of other experts, the less weight should be given. In theory, if two experts give exactly the same assessment because it does not help to draw consensus from the disagreement, the weight can be set to zero. For each expert $PF \cdot p_i$, we introduce a function $D_{ki}(x)$ to represent the scoring deviation between the evaluated step and the remaining steps:

$$D_{\mathrm{ki}}(x) = \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) w_k \right| \tag{13}$$

where h_{ki} and h_{kt} are hesitation probability fuzzy numbers, $S(x)$ is a scoring function, and w_k is the weight of expert $PF \cdot p_i$ $i, t = 1, 2, ..., N$ and $k = 1, 2, ..., K$.

Thus, the total score deviation for all the steps evaluated by expert $PF \cdot p_i$ can be expressed as $D_{ki}(x)$ follows:

$$D_k(x) = \sum_{i=1}^{N} D_{ki}(x) = \sum_{i=1}^{N} \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) w_k \right| \tag{14}$$

To obtain the optimal weight vector, since the general weight vector meets the normalization in cognition of people, Zhou Wei introduced constraint condition Equation (15) based on Wang (43), transformed w_k into $\overline{w_k}$ through Equation (16), and obtained the weight vector $\overline{\mathrm{w}} = (\overline{w_1}, \overline{w_2}, ..., \overline{w_k})$. In this study, a developed MSD method was adopted.

$$\sum_{k=1}^{k} (w_k)^2 = 1 \tag{15}$$

$$\overline{\mathrm{w}_k} = \frac{w_k}{\sum_{k=1}^{k} w_k} \tag{16}$$

Based on the above analysis and setting, the following objective function is constructed to obtain an optimal weight vector that can maximize the deviation value of overall scores of all doctors for each expert evaluation.

$$D(x) = \sum_{K=1}^{K} D_k(x) = \sum_{K=1}^{K} \sum_{i=1}^{N} \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) w_k \right| \tag{17}$$

To solve the weight vector, the following model and Lagrange function are constructed:

$$\mathrm{maxD}(w) = \max \left\{ \sum_{K=1}^{K} \sum_{i=1}^{N} \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) w_k \right| \right\} \tag{18}$$

$$s.t \begin{cases} \sum_{k=1}^{k} (w_k)^2 = 1 \\ w_k \geq 0, \quad k = 1, 2, ..., k \end{cases} \tag{19}$$

$$L(w, \eta) = \sum_{K=1}^{K} \sum_{i=1}^{N} \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) \right| w_k + \frac{\eta}{2} \left(\sum_{k=1}^{k} (w_k)^2 = 1 \right) \tag{20}$$

Combined with the above formula, we can get the following:

$$\overline{w_k} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) w_k \right|}{\sum_{K=1}^{K} \sum_{i=1}^{N} \sum_{t=1}^{N} \left| s\left(h_{ki}\right) w_k - s\left(h_{kt}\right) w_k \right|} \tag{21}$$

According to the expert advice and MSD method, the optimal weights of each step in the similarity calculation can be obtained.

Calculation and Update of Credit Line

Aggregation of Historical Records

In the previous section, we calculated a value describing the behavior of doctors, correlation.

The historical record of each doctor is composed of calculated correlations. In a period, the doctor will generate a large number of historical records. When calculating the credit limit, the historical records are summarized according to the timeline. In the process of calculating and updating the credit limit, the time recorded in history is the time when each doctor diagnosed a certain patient. The influence of early historical records on credit lines will diminish over time. On the contrary, if the behavior of curious doctors occurs recently, the impact on credit will be even greater. As a penalty, the credit limit will remain low for a period.

Since the historical visit record is composed of similarity and time window, the value of similarity is a percentage, in the range of [0,1], so there is no need for standardization. However, the time of each historical access record needs to be mapped in the range of [0,1], that is, the data are standardized. Suppose the time window before processing is $A = (\mathrm{a}_1, \mathrm{a}_2, ..., \mathrm{a}_\mathrm{n})$, and the time window after standardization is $B = (\mathrm{b}_1, \mathrm{b}_2, ..., \mathrm{b}_\mathrm{n})$.

The mapping method is as follows:

$$B_\mathrm{i} = \begin{cases} \frac{a_i - (a_i)_{\min}}{(a_i)_{\max} - (a_i)_{\min}}, & a_i > 0 \\ \frac{(a_i)_{\max} - a_i}{(a_i)_{\max} - (a_i)_{\min}}, & a_i < 0 \end{cases} \tag{22}$$

To make the calculation of credit limit more objective and authentic, the earlier historical record in real life will have less impact on the current credit, that is, the longer the historical record is, its value will decay over time. Suppose the set $HT = \left\{T_{hk}\left(1 \le k \le q\right)\right\}\left(q = |HT|\right)$ of medical history records and the corresponding time window $B = \left\{b_k \middle| 1 \le k \le q\right\}$, then the time attenuation function for a task $h_k\left(1 \le k \le q\right)$ is as follows:

$$\phi\left(t\right) = \frac{1 - b_k / \sum_{k=1}^{q} b_k}{\sum\limits_{k=1}^{q}\left(1 - b_k / \sum_{k=1}^{q} b_k\right)} \tag{23}$$

When calculating the credit limit, the model proposed by Caverlee et al. (44) is modified. Each history record is distinguished by a time window. The structure diagram of the aggregate value calculated according to the historical record of the user in the past N cycles is shown in the **Figure 3**:

The aggregate calculation formula for history of a user $H\left(1\right) \ldots H\left(n\right)$ in the past N cycles is as follows:

$$H\left(old\right) = \frac{1}{\gamma} \times \sum_{k=1}^{N} H_K \times \alpha^{N-K} \tag{24}$$

wherein $\gamma = \sum\limits_{k=1}^{N} \alpha^{N-K}$, γ is used to limit the credit value obtained after aggregation to remain within the original credit value range; α is the adjusting parameter of the influence of historical records to the current trust evaluation. The value range of α is $0 < \alpha < 1$, the smaller the α is, the less important the historical record is.

The updated formula of the credit limit can be obtained based on the aggregate results of the above historical records:

$$H_{\text{new}} = \begin{cases} H_{\text{old}} \cdot \left[1 + \varphi\left(\Delta H\right)\right], & t > t_0 \\ H_{old} \cdot \phi\left(t\right), & t < t_0 \end{cases} \tag{25}$$

H_{old} represents the initial line of credit that is aggregated according to the historical records for the first time, H_{new} represents the value of the line of credit after constant updates, $t_0$represents the effective time of the set time window, and the time decay function. When the time interval t is less than t_0, it means that the current operation occurs within the same time window as the last one. At this time, the credit line is not updated, and the time decay function is used for processing. When t is greater than t_0, it means that within the next time window, the new aggregate value and the increment $\Delta H = H_{\text{new}} - H_{old}$ of the historical aggregate value are used to recalculate and update the value of the credit line.

DYNAMIC ACCESS CONTROL BASED ON TRUST

Overview

In this study, the concept of the credit line is introduced to improve the access control of the consultation process of doctors in the existing intelligent medical system. The doctor logs in HIS according to the identity information (the doctor logs in the device, time, place, etc.), and each user calculates the corresponding credit limit according to the system, which is used to match the reasonable permissions according to the access control strategy.

After the doctor finishes each diagnosis, the data such as the visit record from HIS will be saved in the historical record. Through trust calculation, the credit limit of the user within a period can be obtained.

The trust interval corresponds to the degree of openness of the permission. For example, the line of credit of a doctor is t (t2 < t ≤ t3). According to the access control strategy, the doctor is only allowed to visit contents with relevance of 0.6 during the diagnosis and treatment process. If the doctor visits too many irrelevant contents, the decline of the operational relevance will lead to the reduction of the credit limit of the doctor. The access request of the user is denied when the amount is insufficient. The

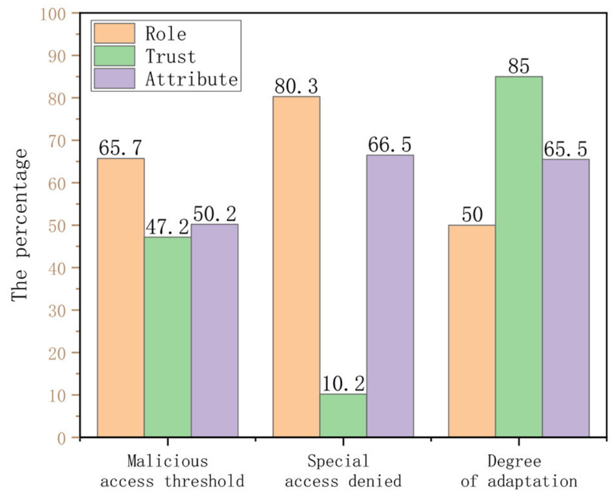


FIGURE 10 | Traditional HIS vs. trust-based access control HIS.

TABLE 6 | User evaluation table of doctors.

	Difficulty of malicious access			Risk of rejection of special access requests			Degree of system automation		
Level	I	II	III	I	II	III	I	II	III
Attribute	73	128	55	42	189	25	16	127	13
Role	156	56	44	206	34	16	116	25	15
Trust	80	84	92	10	34	212	3	64	189

credit interval of the proposed scheme and overall flow chart is shown in the **Figure 4**.

Access Control Policies

An access control policy is the key point of the access control model, which is the access rule set and condition constraint set of subject to object. In the background of the HIS system in this study, the subject is set as the doctor, and the object is the medical record. The access control flow for this article is shown in **Figure 5**.

Step 1. Read the trust value of the doctor from the library and compare the threshold $t_0(DT_i \epsilon DT, t_0 < t \leq t_4)$. The value range of T is shown in **Table 1**.

(a) if$(DT_i.t<t_0)$return false;

If credit limit of the doctor is below the threshold value t_0, the decision to deny access request is returned and recorded.

(b) if$(DT_{\,i}.t \geq t_0)$

When credit line of the doctor is higher than the threshold value t_0, proceed to Step 2.

Step 2. Match the trust interval according to credit limit of the doctor:

```
(c) if ($t_0 < DT_i.t \leq t_1$) trust = 1;
    else if ($t_1 < DT_i.t \leq t_2$) trust = 2;
    else if ($t_2 < DT_i.t \leq t_3$) trust = 3;
    else trust = 4;
```

If the trust value of the doctor belongs to (t_0, t_1), then 1 is returned, indicating that access rights belong to level 1.

If the trust value of the doctor is (t_1, t_2), then return 2, indicating that the access is level 2.

If the trust value of the doctor is (t_2, t_3), then return 3, which means that the access is level 3.

If the trust value of the doctor belongs to (t_3, t_4), then 4 is returned, representing that the access authority belongs to level 4.

Step 3. Match the corresponding relevance requirements according to the credit interval.

```
Switch(trust)
{
case 1 : pre(S)=0.9;
case 2 : pre(S)=0.6;
case 3 : pre(S)=0.4;
case 4 : pre(S)=0.2;
}
```

pre (S) specifies the minimum value of the access relevancy. If it is lower than this value, it will be reflected in the historical record, which will greatly affect the next round of credit evaluation.

EXPERIMENTAL ANALYSIS

Data Sources

Relying on the project of the National Natural Science Foundation of China, this study completed relevant research experiments according to the medical data set provided by a third-class hospital of Kunming, the cooperative unit of the project. The data set contains rich text data and image data, with a total of five databases, the size of which is 1,200 G, including 1,360 data tables and a total of 21,39,373 records. In this experiment, part of medical data was extracted to simulate visits of doctors in the process of diagnosis and treatment.

Experimental Settings

The purpose of the experiment is to verify whether the access control model based on HIS proposed in this study can calculate the line of credit through the historical behavior records of doctors, and well control the access rights of doctors through the value of the line of credit. The data of HIS account access records of three doctors in a department provided by the cooperative hospital were selected for calculation, including one doctor who simulated the behavior of a curious doctor and one honest doctor who simulated a special visit situation as the experimental group.

Weights

Ten medical experts were asked to directly give the weights of the relevant calculations. The weights calculated by the MSD method according to Equation (21) are shown in the **Table 2**. To verify whether the weight calculated according to the weight calculation method, MSD, is better than the weight directly given by the expert, randomly select the weights of two groups of experts and the weights calculated by the MSD method for comparison experiments.

The three doctors are honest doctors, non-malicious doctors with special circumstances (hereinafter referred to as special doctors), and curious doctors. In HIS, each doctor completed 15 diagnoses, among which the curious doctor completed three malicious behaviors, and the special doctor completed two special case diagnoses.

In addition, weights were set according to the weights directly given by Expert 3 and Expert 10 as well as the MSD calculation results, and the scatterplot drawn could intuitively see the calculation results of the correlation degree as shown in the **Figure 6**.

Maximum score deviation weights can distinguish between malicious behavior, special behavior, and normal behavior.

According to the obtained images, the analysis in the following **Table 3** can be obtained. The correlation calculated by the weight given by an expert alone cannot make an accurate judgment on the behavior of doctors, especially in the discrimination between curious doctor and special doctor.

Aggregation

According to Equation (25), with a period of 1 month, the aggregate value of historical records is used to calculate the changes in the credit lines of the three types of doctors in eight periods, as shown in the following **Table 4**:

According to the calculation, the average value of the credit line in eight periods is obtained as **Table 5**.

Three historical record influencing parameters α were given: 0.95, 0.5, and 0.2, and the credit limit of three kinds of doctors was calculated based on the aggregation of historical records in eight cycles. In the table, Δ represents the maximum fluctuation range of the line of credit under the corresponding value of α. Column $\bar{\Delta}$ records the mean fluctuation range of the line of credit.

As shown in the **Figure 7**, an appropriate α can keep the credit limit of doctors who maintain normal behavior during the diagnosis process in a relatively stable state, but they are sensitive to the malicious behavior of curious doctors.

Access Control Experiment

The period N of the historical record for the calculation of the credit limit was 1 month. Assuming that each doctor arranges 3 days a week to diagnose patients, the average daily medical record is about 50. According to the results of the experiment, when malicious visit of the doctor occurred, the credit line completely returned and stabilized at the original level, which required about 650 records, which took nearly a month. This situation is shown in **Figure 8A**.

If doctors intend to increase their average interview relevancy after malicious visits, as shown in the experimental results in the **Figure 8B**, it requires about 250 records, nearly half a month, to completely stabilize the original level of the credit line.

The experiment proves that when malicious access occurs, the value of the credit limit will be immediately affected. As punishment for privacy risk, the credit limit will be kept at a low value for a long period to warn users of their bad behavior and achieve the effect of access control at the same time.

Contrast Experiment

The hospitals that our project cooperates with are currently using traditional HIS without access control. Hundred doctors from the hospital were randomly selected for a black-box test. The doctors were divided into two groups, and the traditional HIS and the HIS of the trusted access control model proposed in this study were used for a 1 month comparison test. In the case that the doctor does not know the contents of the experiment, the historical records of the two groups of doctors are analyzed.

It can be seen from the **Figure 9** that there is no significant difference in the historical visit records (correlation) of the doctors using the two HISs within 1 week of the experiment. Throughout the experimental cycle, the relevance of doctors using traditional HIS has not changed significantly, while the trust HIS model has been significantly improved, indicating that the proposed credit line can regulate user behavior to a certain extent.

Then, we conducted a questionnaire survey of some doctors in the hospital. The purpose is to compare the credit line model proposed in this study with the role-based access control model (hereinafter referred to as role) and ABAC model (hereinafter referred to as attribute). The feedback from all 256 users is shown in the **Table 6**.

According to the table data and **Figure 10**, the following chart shows that the trust-based HIS access control model proposed in this study has a good performance in terms of flexibility of access control, preventing malicious access behavior from occurring, and the degree of system automation.

CONCLUSION

Aiming at HIS in the context of medical big data, this study proposes a dynamic access control model for doctors in the process of diagnosis and treatment. First, according to the diagnosis and treatment process of the doctor, the behavior model of the doctor is designed, and three hypotheses of privacy leakage are proposed. Then, according to the operation correlation of the doctor, time index, and other factors, the behavior of the doctor in the diagnosis and treatment process is described, and the purpose is to calculate the rationality of the diagnosis process of the doctor through mathematical methods. Finally, by calculating the credit limit, the access control strategy using the credit limit interval dynamically restricts the access ability of the doctor in the diagnosis and treatment process. Experiments prove that the model designed in this study can accurately identify bad doctors and inhibit their visits by trust value, and the ability to prevent patient privacy leakage is better than traditional HIS.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

RJ proposed the idea for this paper. RJ, WW, and YY designed the study. WW wrote the paper. WW and FM performed the experimental analysis. All authors reviewed and edited the manuscript and read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 71972165, 61763048), Science and Technology Foundation of Yunnan Province (No. 202001AS070031).

REFERENCES

1. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang G-Z. Big data for health. *IEEE J Biomed Health Inform.* (2015) 19:1193–208. doi: 10.1109/JBHI.2015.2450362
2. Shi T, Ma J, Cao H, Meng L, Zhang C. Research progress of medical big data privacy protection technology. *China Med Equip*. (2019) 34:163–6. doi: 10.3969/j.issn.1674-1633.2019.05.042
3. Priyanka K, Kulennavar N. A survey on big data analytics in health care. *Int J Comp Sci Inform Technol.* (2014) 5:5865–8. doi: 10.1109/ICSSIT46314.2019.8987882
4. Dolley S. Big data's role in precision public health. *Front Public Health.* (2018) 6:68. doi: 10.3389/fpubh.2018.00068
5. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med.* (2019) 25:37–43. doi: 10.1038/s41591-018-0272-7

6. Wang Q. Ethics predicament and protection path of medical privacy in the era of big data. *Chin Med Ethics*. (2016) 29:685–9. doi: 10.12026/j.issn.1001-8565.2016.04.43
7. Liu J. Hospital Information System (HIS) applications in the hospital[J]. *Med Inf (Surg Sect)*. (2010) 5:966–7. doi: 10.3969/j.issn.1006-1959.2010.04.214
8. Guo Z, Luo Y, Cai Z, Zheng T. Overview of privacy protection technology of big data in healthcare. *Comp Sci Explor*. (2021) 15:389–402. doi: 10.3778/j.issn.1673-9418.2009071
9. Lu X, Gu C. Analysis on causes and protective strategy of user privacy disclosure in the big data environment. *Modern Intellig*. (2016) 36:66–70. doi: 10.3969/j.issn.1008-0821.2016.11.012
10. Steinbrook R. Personally controlled online health data—the next big thing in medical care? *N Engl J Med*. (2008) 358:1653–6. doi: 10.1056/NEJMp0801736
11. Ma Z, Zhang L. Role of HIS in the modernization efforts of hospitals[J]. *Chin J Hosp Manage*. (2006) 22:350–1. doi: 10.3760/j.issn:1000-6672.2006.05.027
12. Dong J. Current status and cause analysis of hospital information system in my country[J]. *Chin J Hosp Manage*. (2003) 19:228–30. doi: 10.3760/j.issn:1000-6672.2003.04.014
13. Xue W. The research development of electronic medical records in China. *Chin Hosp Manage*. (2005) 25:17–9. doi: 10.3969/j.issn.1001-5329.2005.02.006
14. Ren L, Wang J. Lessons from the establishment and application of hospital information systems. *Chin J Hosp Manage*. (2002) 21–3. doi: 10.3760/j.issn:1000-6672.2002.05.007
15. Sandhu RS, Samarati P. Access-control-principles and practice. *IEEE Commun Mag*. (1994) 32:40–8. doi: 10.1109/35.312842
16. Hao L, Min Z, Deng G, Zhen H. Research on big data access control. Chinese *J Comp*. (2017) 40:72–91. doi: 10.11897/SP.J.1016.2017.00072
17. Mohammed I, Dilts DM. Design for dynamic user-role-based security. *Comp Secur*. (1994) 13:661–71. doi: 10.1016/0167-4048(94)90048-5
18. Thomas RK, Sandhu RS. Conceptual foundations for a model of task-based authorizations. In: *Proceedings The Computer Security Foundations Workshop VII*. Franconia, NH: IEEE (1994). p. 66–79.
19. Shen H, Hong F. Overview of access control model research. *Comp Appl Res*. (2005) 9–11. doi: 10.3969/j.issn.1001-3695.2005.06.003
20. Sandhu RS. Role-based access control. *Adv Comput*. (1998) 466:237–86. doi: 10.1016/S0065-2458(08)60206-5
21. Goyal V, Pandey O, Sahai A, Waters B. Attribute-based encryption for fine grained access control of encrypted data. In: *Proceedings of the 13th ACM Conference on Computer and Communications Security*. Alexandria, VA (2006). p. 89–98.
22. Wang X, Wang L, Li Y, Gai K. Privacy-aware efficient fine-grained data access control in Internet of medical things based fog computing. *IEEE Access*. (2018) 6:47657–65. doi: 10.1109/ACCESS.2018.2856896
23. Xu P, Huang K. The Status, Problems and Countermeasures of the Big Data of Health Care in China. *China Dig Med*. (2017) 12:24–6. doi: 10.3969/j.issn.1673-7571.2017.05.008
24. Xue T, Fu Q, Wang C, Wang X. Research on blockchain-based medical data sharing model. *Acta Automat Sin*. (2017) 43:1555–62. doi: 10.16383/j.aas.2017.c160661
25. Narayanan HAJ, Güneş MH. Ensuring access control in cloud provisioned healthcare systems. In: *Consumer Communications and Networking Conference*. Las Vegas, NV (2011). p. 247–51.
26. Wang M, Wang J, Guo L, Harn L. Inverted XML access control model based on ontology semantic dependency. *Comp Mater Continua*. (2018) 55:465–82. doi: 10.3970/cmc.2018.02568
27. Zhu Y, Huang D, Hu C-J, Wang X. From RBAC to ABAC: constructing flexible data access control for cloud storage services. *IEEE Transac Serv Comput*. (2014) 8:601–16. doi: 10.1109/TSC.2014.2363474
28. Liu Y, Zhang W, Wang X. Access control scheme based on multi-attribute fuzzy trust evaluation in cloud manufacturing environment. *Comp Integr Manuf Syst*. (2018) 24:321–30. doi: 10.13196/j.cims.2018.02.005
29. Gao P. *Research and Design of Dynamic Access Control Model Based on Trust and Role*. Tianjin: Tianjin University (2014).
30. Zhang P, Zhou L. Trust-based dynamic multi-level access control model. *Comp Modern*. (2019) 116–247. doi: 10.3969/j.issn.1006-2475.2019.07.020
31. Khan MFF, Sakamura K. Fine-grained access control to medical records in digital healthcare enterprises. *In: 2015 International Symposium on Networks*. Computers and Communications (ISNCC). Yasmine Hammamet: IEEE (2015). p. 1–6.
32. Zhang Y, Fu Y, Yang M, Luo J. Access control scheme for medical data based on PBAC and IBE. *J Commun*. (2015) 36:200–11. doi: 10.11959/j.issn.1000-436x.2015329
33. Yang Y, Zheng X, Guo W, Liu X, Chang V. Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. *Inform Sci*. (2019) 479:567–92. doi: 10.1016/j.ins.2018.02.005
34. Shi M, Jiang R, Hu X, Shang J. A privacy protection method for health care big data management based on risk access control. *Health Care Manage Sci*. (2020) 23:427–42. doi: 10.1007/s10729-019-09490-4
35. Aggelidis VP, Chatzoglou PD. Hospital information systems. *J Biomed Inform*. (2012) 45:566–79. doi: 10.1016/j.jbi.2012.02.009
36. Zhang X, Fu Y, Chu P. Application of Jackard similarity coefficient in recommender system. *Comp Technol Dev*. (2015) 25:158–226. doi: 10.3969/j.issn.1673-629X.2015.04.036
37. Hamers L. Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Inform Proc Manage*. (1989) 25:315–8. doi: 10.1016/0306-4573(89)90048-4
38. Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S. Using of Jaccard coefficient for keywords similarity. In: *Proceedings of the International Multiconference of Engineers and Computer Scientists*. Hong Kong (2013). p. 380–4.
39. Jamin A, Humeau-Heurtier A. (Multiscale) cross-entropy methods: a review. *Entropy*. (2020) 22:15. doi: 10.3390/e22060644
40. Dong X, Qian M, Jiang R. Packet classification based on the decision tree with information entropy. *J Supercomp*. (2020) 76:4117–31. doi: 10.1007/s11227-017-2227-z
41. De Boer PT, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. *Ann Operat Res*. (2005) 134:19–67. doi: 10.1007/s10479-005-5724-z
42. Xu Z, Zhou W. Consensus building with a group of decision makers under the hesitant probabilistic fuzzy environment. *Fuzzy Optimiz Decis Mak*. (2017) 16:481–503. doi: 10.1007/s10700-016-9257-5
43. Wang YM. Using the method of maximizing deviations to make decision for multi-indicies. *Syst Eng Electron*. (1998) 20:24–6.
44. Caverlee J, Liu L, Webb S. The SocialTrust framework for trusted social information management: architecture and algorithms. *Inform Sci*. (2010) 180:95–112. doi: 10.1016/j.ins.2009.06.027

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fgene.2021.771435

Contextualizing Genes by Using Text-Mined Co-Occurrence Features for Cancer Gene Panel Discovery

Hui-O Chen[1,2], *Peng-Chan Lin*[1,2,3,4]*, *Chen-Ruei Liu*[1,2], *Chi-Shiang Wang*[1,2] and *Jung-Hsien Chiang*[1,2]*

[1]Department of Computer Science and Information Engineering, College of Electrical Engineering and Computer Science, National Cheng Kung University, Tainan, Taiwan, [2]Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan, [3]Department of Oncology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan, [4]Department of Genomic Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan

OPEN ACCESS

Edited by:
Ying Li,
Zhejiang University, China

Reviewed by:
Rafael Rosell,
Catalan Institute of Oncology, Spain
Ehsan Nazemalhosseini-Mojarad,
Shahid Beheshti University of Medical Sciences, Iran
Ruan Maomei,
Shanghai Chest Hospital, Shanghai Jiaotong University, China

***Correspondence:**
Peng-Chan Lin
pengchanlin@gmail.com
Jung-Hsien Chiang
jchiang@mail.ncku.edu.tw

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 06 September 2021
Accepted: 11 October 2021
Published: 25 October 2021

Citation:
Chen H-O, Lin P-C, Liu C-R, Wang C-S and Chiang J-H (2021) Contextualizing Genes by Using Text-Mined Co-Occurrence Features for Cancer Gene Panel Discovery. Front. Genet. 12:771435. doi: 10.3389/fgene.2021.771435

Developing a biomedical-explainable and validatable text mining pipeline can help in cancer gene panel discovery. We create a pipeline that can contextualize genes by using text-mined co-occurrence features. We apply Biomedical Natural Language Processing (BioNLP) techniques for literature mining in the cancer gene panel. A literature-derived 4,679 × 4,630 gene term-feature matrix was built. The *EGFR* L858R and T790M, and *BRAF* V600E genetic variants are important mutation term features in text mining and are frequently mutated in cancer. We validate the cancer gene panel by the mutational landscape of different cancer types. The cosine similarity of gene frequency between text mining and a statistical result from clinical sequencing data is 80.8%. In different machine learning models, the best accuracy for the prediction of two different gene panels, including MSK-IMPACT (Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets), and Oncomine cancer gene panel, is 0.959, and 0.989, respectively. The receiver operating characteristic (ROC) curve analysis confirmed that the neural net model has a better prediction performance (Area under the ROC curve (AUC) = 0.992). The use of text-mined co-occurrence features can contextualize each gene. We believe the approach is to evaluate several existing gene panels, and show that we can use part of the gene panel set to predict the remaining genes for cancer discovery.

Keywords: biomedical natural language processing, machine learning, topic modeling, cancer gene panel, text mining

INTRODUCTION

Scientific articles provide text mining (TM) applications in cancer biology (Zhu et al., 2013; Azam et al., 2019; Wang et al., 2020). Several solutions are currently available to meet the growing need for different cancer gene panels. Several commercial gene panels constitute a "one-size-fits-all" solution. In a clinical investigation, we need to design gene panels specifically tailored for particular questions or individual cancers (Hyman et al., 2015). The precision of the designed panel for different tumors plays an important role. They rely on literature reviews and cancer genomics databases. The reason for selecting somatic and germline mutation profiling is also complicated. Emerging TM techniques such as Gene2Vec offer some answers to information interpreting problems. Gene2Vec is a study that explored the idea of gene embedding, distributed representation of genes, in the spirit of word embedding (Demeester et al., 2016; Du et al., 2019). However, we cannot explain the biomedical meaning of the vector in the neural embedding model. The goal of explainability is very important and would be very useful. The ability to provide additional gene suggestions for a gene panel with an explanation would be hugely valuable but also really challenging. Therefore, we developed a biomedical-explainable and validatable text mining pipeline for cancer gene panel discovery.

Firstly, we find a system for predicting genes and interesting applications for a gene panel discovery. The use of text-mined co-occurrences features for each gene can contextualize each gene, and as input for a machine learning system. We extract NER names mentioned in the literature, such as gene NER (Leaman et al., 2013) and disease NER (Wei et al., 2013). The use of PubTator (Wang et al., 2016) along with MeSH (Ikonomakis et al., 2005) is a good way of getting as good enrichment for biomedical relevant terms. The frequency-inverse document frequency (TF-IDF) was used to construct the document-term matrix (Ikonomakis et al., 2005). Machine learning-based and biomedical-explainable approaches have recently become the most popular approaches in the study of the document-term matrix. For example, M. Ikonomakis et al. introduced several machine learning (ML) algorithms applied to text classification such as naïve-Bayes, decision trees, neural networks, nearest neighbors, and support vector machines (Devarajan et al., 2015). Wei Xu et al. proposed a novel document-clustering method based on non-negative matrix factorization (Choo et al., 2013). Choo et al. presented a user-driven topic modeling based on interactive non-negative matrix factorization capable of tuning the topic model result by integrating user interactions (Pedregosa et al., 2011). Summarizing the abovementioned studies, we established a fully integrated text mining pipeline to find the gene term-feature, mutational landscape heatmap, and cancer information topic.

With next-generation sequencing (NGS) technologies (Shabani Azim et al., 2018), many targeted panels have been developed to detect hereditary cancer and monitor somatic mutation changes in progressive cancer (McCabe et al., 2019). The Memorial Sloan Kettering Cancer Center has developed MSK-IMPACT (Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets), a hybridization capture-based next-generation sequencing assay for deep target sequencing of all exons and selected introns of 410 essential cancer genes in tumors (Hyman et al., 2015; Cheng et al., 2015). The MSK-IMPACT panel performed well not only in the above study but also in a large-scale clinical sequencing project with more than 10,000 patients (Zehir et al., 2017). They provided a comprehensive gene panel database including actionable drug targets, cancer susceptibility genes in hematological malignancies, and solid tumors. For solid tumors, the Oncomine Cancer Panel (OCP) is only used for the clinical screening of actionable genetic mutations in solid tumors (Luthra et al., 2017). They significantly provide druggable target databases. We validate the biomedical literature mining through the MSK-IMPACT or OCP cancer gene panel NGS database.

We create a pipeline that can suggest additional genes for a gene panel given an existing set of genes. And we believe the approach is to evaluate several existing gene panels, and show that we can use part of the gene panel set to predict the remaining genes.

MATERIALS AND METHODS

PUBMED

PubMed, a free database of more than 30 million literature citations for biomedical literature, includes the fields of biomedicine and health. We extracted the abstracts that mentioned genes related to human cancer from PubMed and took the gene's context by gene window.

Machine Learning Model and Analysis

K nearest neighbors, linear support vector machine (SVM), Gaussian process, decision tree, random forest, neural net, and naive Bayes were used to conduct supervised machine learning. All the models were built by python with the scikit-learn package and used five-fold cross-validation (Wei et al., 2015). The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were used to evaluate the model's performance.

Biomedical Term Tagging

PubTator

PubTator (Wei et al., 2013) is a web-based PubMed abstract biomedical named entity recognition (NER) system. PubTator can tag the gene, disease, chemical, species, and mutation in PubMed abstracts, and the tagging result could be accessed *via* the RESTful API. We used PubTator as a part of the biomedical term tagger.

Medical Subject Heading

MeSH is a hierarchically organized medical vocabulary thesaurus used for indexing articles for PubMed. PubMed Articles curated by NLM are indexed with several related MeSH terms; every MeSH term has unique id and hierarchical categories. With these characteristics of MeSH term and our tagging algorithm, we could tag biomedical terms that are not tagged by PubTator. Our algorithm started from the MeSH terms of each PubMed article. For each MeSH term in an article, we first created a MeSH term-mapping set that mapped a MeSH term to another set that contained itself and its lower hierarchy MeSH term. Second, for each MeSH term in the MeSH term-mapping set, we tried matching all of the entry terms, synonyms of a specific MeSH term, to every word in the article. If a word in the article matched any entry names of a MeSH term, we tagged that word as a biomedical term. This way, those terms having the same concepts could be merged and analyzed.

Gene Term-Feature Term Frequency–Inverse Document Frequency Matrix Construction

For a particular gene, considering all of its gene windows in the whole corpus, we calculated the frequency of the co-occurrence of the gene and features (terms) tagged by our algorithm in the window as the term frequency of the feature. The higher the term frequency is, the stronger the association of the gene and feature. In our study, term frequency (TF) was calculated using the following formula:

$$TF_{gene,\ feature} = \log\left(1 + tf_{gene,feature}\right)$$

To calculate the inverse document frequency of each term feature, we simply count the occurrences of the term feature in all genes as document frequency. The inverse document frequency (IDF) was calculated using the following formula:

$$IDF_{feature} = \log\left(1 + n_{gene}/df_{feature}\right)$$

The higher the IDF, the more specific the term feature is to a particular gene. Finally, by multiplying TF and IDF, the gene term -feature matrix was constructed.

Term Feature Selection by the Hypergeometric Test

We filtered out genes that had less than ten term features. We identified the critical term feature according to the gene panel using the *p*-values of hypergeometric tests as follows. We input the MSK-IMPACT (Hyman et al., 2015) panel. Ns is the size of the MSK-IMPACT panel set S, N is the size of the set $S^{'}$, which contains 500 non-MSK genes (randomly selected from the gene term-feature matrix) and all of the MSK genes, N_t is the number of genes in the set S' that contain term feature t, and N_{st} is the number of genes in the set S containing t. The random variable y representing several genes containing the term feature in the set S is a hypergeometric random variable with parameters $N_s, N_t,$ and N (Westlake and Larson, 1970). The probability distribution of y is shown as follows:

$$\mathrm{P}(\mathrm{y}) = \frac{\binom{N_t}{y}\binom{N - N_t}{N_s - y}}{\binom{N}{N_s}}$$

From N_{st}, we compute the *p*-value, the probability of the observed (N_{st}), as follows:

$$\text{Pvalue} = \sum_{y=N_{st}}^{\min(N_s,N_t)} P(y)$$

The *p*-value reflects significant phrases in S compared with all of the genes in the gene term -feature matrix. A low *p*-value indicates that we observe a rare event and that the observed term feature represents a statistical discovery, suggesting that it is essential in the MSK-IMPACT panel.

Topic Modeling

Our topic modeling was based on the algorithms of non-negative matrix factorization (NMF) (Yeganova et al., 2014). Given a nonnegative matrix $\mathrm{X} \in \mathbb{R}^{m\times n}$, when the desired lower dimension is k, the goal of NMF is to find the two matrixes, $\mathrm{W} \in \mathbb{R}^{m\times k}$ and $\mathrm{H} \in \mathbb{R}^{k\times n}$, having only non-negative entries such that $\mathrm{X} \approx \mathrm{WH}$.

The objective function is shown as the following formula:

$$\min_{\mathrm{W}\geq 0,\mathrm{H}\geq 0} f(W,H) = ||X - WH||_F^2$$

The function is the most commonly used formulation based on the Frobenius norm. K represents the number of topics we expected, X represents the gene term-feature matrix, W represents the gene-topic matrix, and H represents the topic text-feature matrix. Since the weights in W and H have been calculated, we used the top 20 genes and the top 20 text features with the highest importance for each topic to interpret the biomedical meaning.

Gene Window

We take the gene's context as its gene window. Each gene window contains three sentences. The sentence contains the gene, the previous sentence, and the next sentence. We want to eliminate the redundant part. Using the gene window algorithm, we could iterate through the full abstracts containing specific genes in the text and grip the most critical section for further analysis. We pick three sentences based on the concept that the sentence that is closer to the gene is more relevant to it. Since the closest ones are previous and the next one, so we picked three.

RESULTS

Study Design and Workflow

This study develops a gene panel analysis framework that can discover the characteristics of a gene panel based on biomedical literature mining. The method overview is shown in **Figure 1**. First, we extracted the PubMed abstracts, which mentioned genes related to humans. The method is shown as **Figure 2**. In this step, approximately 430,000 PubMed abstracts regarding genes were filtered out from all of the current PubMed corpus (approximately 30 million articles). Second, we performed biomedical named entity recognition (NER) on the extracted PubMed abstracts using PubTator (Wang et al., 2016) and MeSH (Medical Subject Headings). Third, we used the biomedical term to construct the gene term-feature matrix, which has a concept similar to that of the document-term matrix. Fourth, we performed term feature selection according to individual gene panels to make the term feature generated by the previous step stronger and correspond to the target gene panel.

Here, we explored the idea of the hypergeometric distribution. For each term feature, by comparing the distribution of occurrences in the target gene set and the whole gene set, the term features that correlated more with the target gene panel would be enriched. This approach is flexible in regard to different target gene sets, such as the Oncomine Cancer Panel or cardiovascular gene panels.

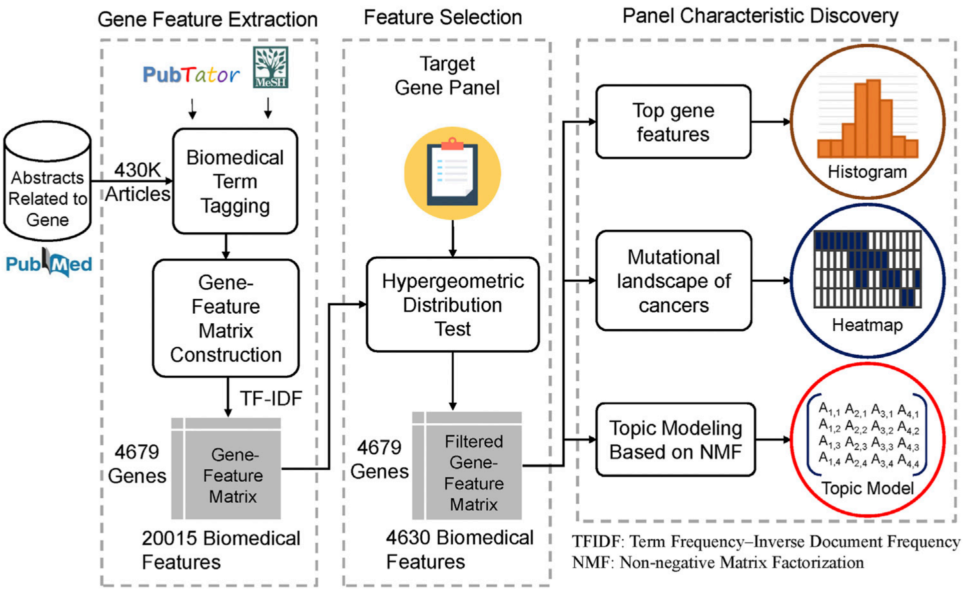
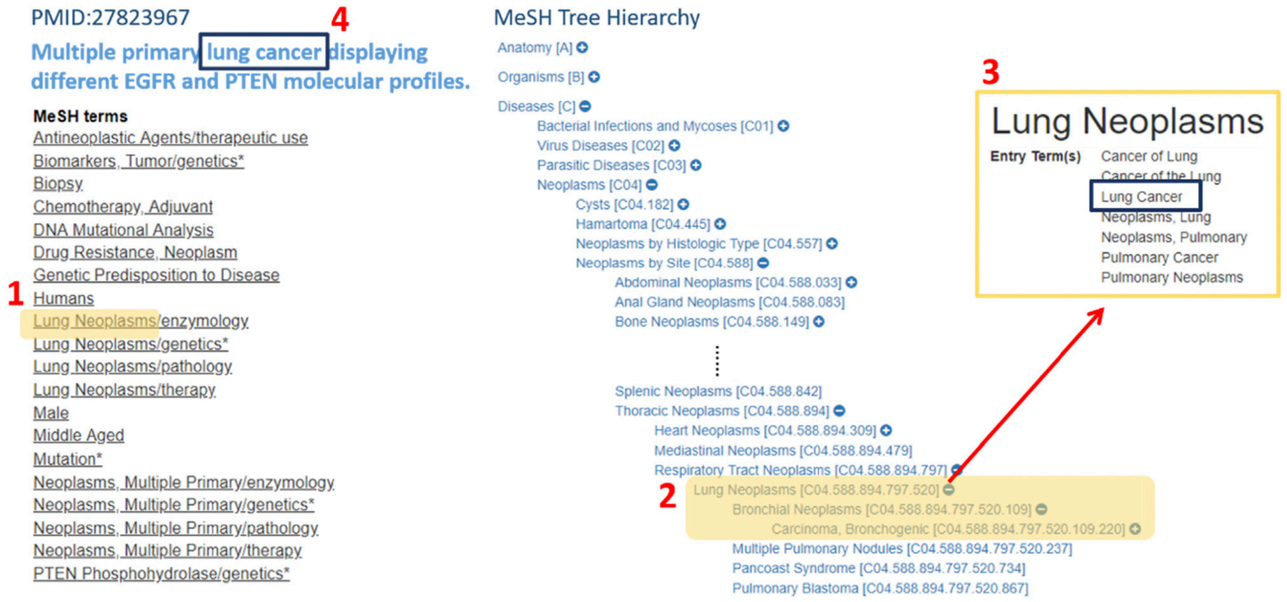

Gene Feature Extraction
Feature Selection
Panel Characteristic Discovery
PubTator
MeSH
Abstracts Related to Gene
PubMed
430K Articles
Biomedical Term Tagging
Gene-Feature Matrix Construction
TF-IDF
4679 Genes
Gene-Feature Matrix
20015 Biomedical Features
Target Gene Panel
Hypergeometric Distribution Test
4679 Genes
Filtered Gene-Feature Matrix
4630 Biomedical Features
Top gene features
Histogram
Mutational landscape of cancers
Heatmap
Topic Modeling Based on NMF
A1,1 A2,1 A3,1 A4,1
A1,2 A2,2 A3,2 A4,2
A1,3 A2,3 A3,3 A4,3
A1,4 A2,4 A3,4 A4,4
Topic Model
TFIDF: Term Frequency–Inverse Document Frequency
NMF: Non-negative Matrix Factorization

FIGURE 1 | Study design and workflow The flowchart shows the overall analysis framework of this study. We first extracted 430,000 abstracts that mentioned genes related to humans in the PubMed corpus. Second, biomedical named entity recognition (NER) was performed to obtain biomedical terms, such as gene name, disease name, and drug name, using PubTator and MeSH. Third, we used the biomedical term tagged by the previous step to construct the gene term-feature matrix whose concept was similar to the document-term matrix. Fourth, we performed term feature selection according to a particular gene panel. We took the MSK-IMPACT panel as an example and made the term features generated by the previous step correspond more to the target gene panel using the hypergeometric distribution. Finally, several analyses, including identifying the top gene term features, creating the mutational landscape of cancers, and topic modeling based on nonnegative matrix factorization, were conducted to determine and interpret the biomedical characteristics of the target gene panel.



PMID:27823967

Multiple primary lung cancer displaying different EGFR and PTEN molecular profiles.

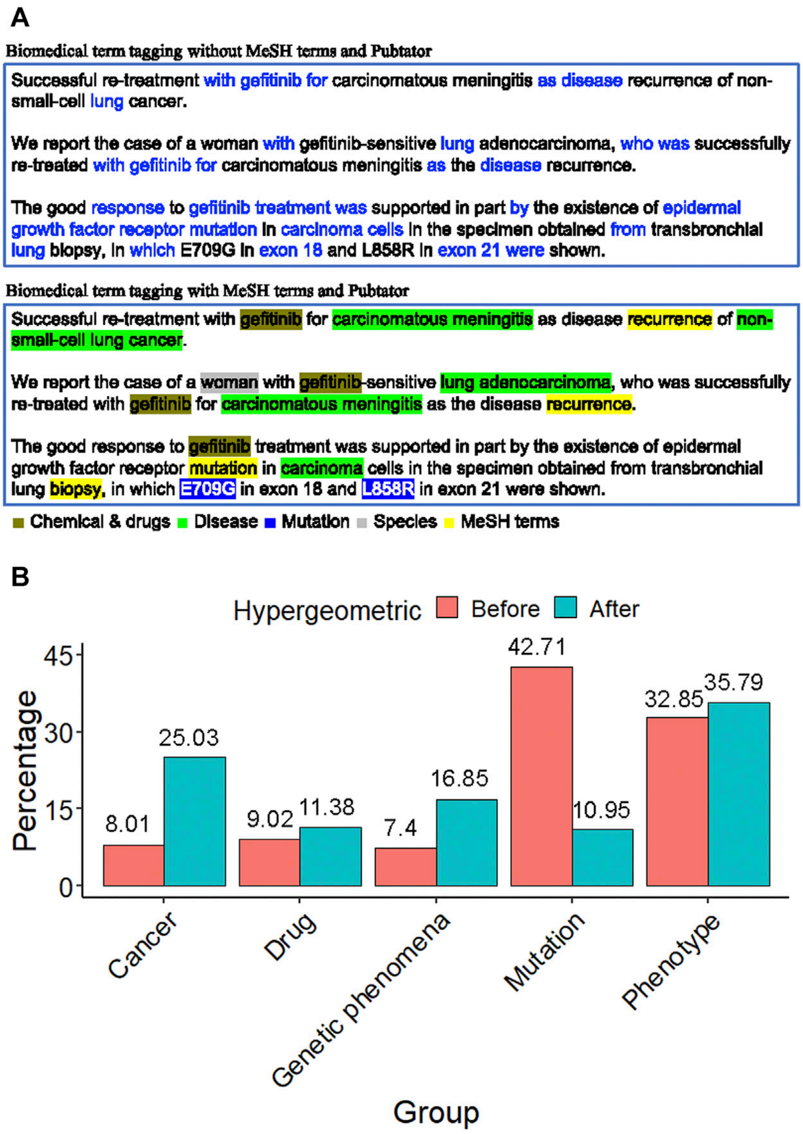
MeSH terms
Antineoplastic Agents/therapeutic use
Biomarkers, Tumor/genetics*
Biopsy
Chemotherapy, Adjuvant
DNA Mutational Analysis
Drug Resistance, Neoplasm
Genetic Predisposition to Disease
Humans
Lung Neoplasms/enzymology
Lung Neoplasms/genetics*
Lung Neoplasms/pathology
Lung Neoplasms/therapy
Male
Middle Aged
Mutation*
Neoplasms, Multiple Primary/enzymology
Neoplasms, Multiple Primary/genetics*
Neoplasms, Multiple Primary/pathology
Neoplasms, Multiple Primary/therapy
PTEN Phosphohydrolase/genetics*

MeSH Tree Hierarchy
Anatomy [A]
Organisms [B]
Diseases [C]
Bacterial Infections and Mycoses [C01]
Virus Diseases [C02]
Parasitic Diseases [C03]
Neoplasms [C04]
Cysts [C04.182]
Hamartoma [C04.445]
Neoplasms by Histologic Type [C04.557]
Neoplasms by Site [C04.588]
Abdominal Neoplasms [C04.588.033]
Anal Gland Neoplasms [C04.588.083]
Bone Neoplasms [C04.588.149]
⋮
Splenic Neoplasms [C04.588.842]
Thoracic Neoplasms [C04.588.894]
Heart Neoplasms [C04.588.894.309]
Mediastinal Neoplasms [C04.588.894.479]
Respiratory Tract Neoplasms [C04.588.894.797]
Lung Neoplasms [C04.588.894.797.520]
Bronchial Neoplasms [C04.588.894.797.520.109]
Carcinoma, Bronchogenic [C04.588.894.797.520.109.220]
Multiple Pulmonary Nodules [C04.588.894.797.520.237]
Pancoast Syndrome [C04.588.894.797.520.734]
Pulmonary Blastoma [C04.588.894.797.520.867]

Lung Neoplasms
Entry Term(s) Cancer of Lung
Cancer of the Lung
Lung Cancer
Neoplasms, Lung
Neoplasms, Pulmonary
Pulmonary Cancer
Pulmonary Neoplasms

FIGURE 2 | An example displays how the term "lung cancer", being tagged in MeSH hierarchical structure. The way "lung cancer" being tagged is as follows. First, we iterate through the MeSH terms of the index of PMID: 27823967 and found "Lung Neoplasms" was one of the MeSH terms, which its synonyms contain "Lung Cancer." Second, if the term "Lung Cancer" also appeared in the article, the MeSH tagging algorithm would tag this word and take its MeSH ID for further analysis.

Finally, we filtered out 4,630 term features from 20,015 term features. The filtered gene term-feature matrix, whose size is 4,679 (genes) x 4,630 (term features), will be used in the following analysis. Thus, we can discover the top 20 gene term features, the mutational landscape of the cancer genome, and topic modeling of cancer information. In



A

Biomedical term tagging without MeSH terms and Pubtator

Successful re-treatment with gefitinib for carcinomatous meningitis as disease recurrence of non-small-cell lung cancer.

We report the case of a woman with gefitinib-sensitive lung adenocarcinoma, who was successfully re-treated with gefitinib for carcinomatous meningitis as the disease recurrence.

The good response to gefitinib treatment was supported in part by the existence of epidermal growth factor receptor mutation in carcinoma cells in the specimen obtained from transbronchial lung biopsy, in which E709G in exon 18 and L858R in exon 21 were shown.

Biomedical term tagging with MeSH terms and Pubtator

Successful re-treatment with gefitinib for carcinomatous meningitis as disease recurrence of non-small-cell lung cancer.

We report the case of a woman with gefitinib-sensitive lung adenocarcinoma, who was successfully re-treated with gefitinib for carcinomatous meningitis as the disease recurrence.

The good response to gefitinib treatment was supported in part by the existence of epidermal growth factor receptor mutation in carcinoma cells in the specimen obtained from transbronchial lung biopsy, in which E709G in exon 18 and L858R in exon 21 were shown.

FIGURE 3 | Biomedical term extraction **(A)** The term feature of an EGFR-related abstract. The former was filled with many redundant words, such as with, for, and after. The latter contains lots of biologically meaningful terms, such as gefitinib (chemical), non-small cell lung cancer (disease), L858R (mutation), the woman (species), and recurrence (MeSH). This phenomenon shows that the tagging approach with MeSH and PubTator terms is essential to gene term-feature extraction. **(B)** The proportion distribution bar chart of the MSK-IMPACT panel in each term feature group before and after the hypergeometric distribution test. It shows that after term feature selection, the proportion of the term feature groups of interest increases, such as cancer, drug, genetic phenomena, and phenotype.

this way, we can find the potential characteristics of the gene panel.

Biomedical Term Extraction by Hypergeometric Test

In the field of biomedical literature mining, tagging the biomedical term is an important issue. For an abstract of the biomedical literature, only biomedical words are what we are interested in, such as drug name, disease name, or gene name. PubTator was capable of tagging the gene, disease, chemical, species, and mutation in PubMed abstracts. **Figure 3A** shows the term feature extraction result of an *EGFR*-related abstract compared to the term features extracted by raw text TF-IDF scoring without biomedical term tagging. The biomedical term features were filled with redundant words, such as “with”, “for”, and “after”.

On the other hand, the term feature extraction approach with MeSH terms and PubTator resulted in term features that contained lots of biologically meaningful terms, such as gefitinib (chemical), non-small cell lung cancer (disease), L858R (mutation), the woman (species), and recurrence (MeSH). This phenomenon shows that the tagging approach is essential for gene term feature extraction.

To discover the characteristics of a gene panel, we used the hypergeometric distribution test. According to MeSH terms and PubTator categories, all the term features can be divided into five groups: cancer, drug, genetic phenomena, mutation, and phenotype

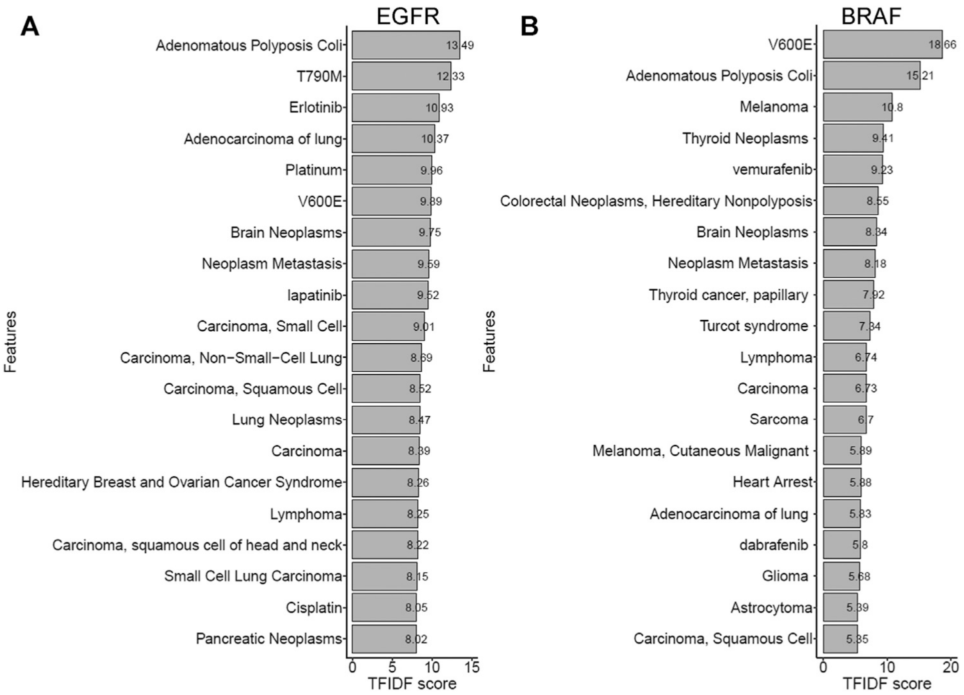


FIGURE 4 | Top term features in *EGFR* and *BRAF* genes **(A)** The bar chart shows the TF-IDF scores of term features related to *EGFR*. Most of the identified term features for *EGFR* were associated with syndromes (e.g., lung adenocarcinoma and non-small cell lung carcinoma), mutations (e.g., T790M), and therapies (e.g., erlotinib and lapatinib) for lung cancer. **(B)** The bar chart shows the TF-IDF scores of term features related to *BRAF*. Biomedical term features, including cancer types (e.g., melanoma and thyroid cancer), mutations (e.g., V600E), and inhibitors (e.g., vemurafenib and dabrafenib) for *BRAF*, were consistent with known findings.

(**Supplementary Table S1**). Take the MSK-IMPACT panel as a target gene panel, for example. The distribution of the MSK-IMPACT panel shows that the percentage increases in some term feature groups after using the hypergeometric distribution test (**Figure 3B**). We filtered out the unimportant genes and found the critical term features according to the gene panel using a hypergeometric distribution test. The proportion of term feature groups in our interest increases, such as cancer, drug, genetic phenomena, and phenotype. The percentage after using the hypergeometric distribution test showed a noticeable improvement from 8.01 to 25.03% in the cancer group. The proportion increased from 9.02 to 11.38% in the drug group and grew from 7.4 to 16.85% in the genetic phenomenon group. There was a slight increase from 32.85 to 35.79% in the phenotype group. After the term feature selection, the proportion decreased from 42.71 to 10.95% in the mutation group. The MSK-IMPACT panel stands for integrated mutation profiling of actionable cancer targets, so the percentage in these groups increases after the hypergeometric distribution test.

3.3 Literature-derived Gene Term Features

The biomedical term features extracted from the literature were directly or indirectly related to each gene. Here, we took some cancer-related genes as examples for further demonstration. **Figures 4A,B** show the top twenty biomedical term features with the highest TF-IDF scores for *EGFR* (the score range from 8.02 to 13.49) and *BRAF* (the score range from 5.35 to 18.66). For *EGFR*, which has been recognized for its importance in lung cancer (Paez et al., 2004; Shepherd et al., 2005), most of the term features directly represent lung cancer or its subtypes, such as "Adenocarcinoma of the lung," "Carcinoma, small cell," and "Carcinoma, Non-Small Cell Lung." "T790M" is a drug resistance mutation frequently observed in patients with lung cancer (Zhou et al., 2009). "Erlotinib" is an effective tyrosine kinase inhibitor (TKI) targeting *EGFR* for non-small cell lung carcinoma (NSCLC). "Lapatinib" is a dual *EGFR/ERBB2* TKI for metastatic breast cancer (Burris, 2004). Some term features were indirectly relevant to *EGFR*, such as "Platinum" and "cisplatin," which are both standard chemotherapy in NSCLC (Arriagada et al., 2004). *EGFR* TKIs are commonly compared with conventional platinum-based therapies. Another example is *BRAF*, whose mutations are widely detected in melanoma, thyroid cancer, and colorectal cancer (Chapman et al., 2011). "V600E" is a crucial mutation that causes the constitutive activation of the cellular signaling pathway (Chapman et al., 2011). "Vemurafenib" and "dabrafenib" are competitive inhibitors designed for *BRAF* with the V600E mutation (Hauschild et al., 2012). The other examples, such as *BRCA1*, *BRCA2*, *MLH1*, and *ERBB2*, are shown in **Supplementary Figure S1**. Nearly all of the biomedical term features relevant to these genes were consistent with current knowledge.

Mutational Landscape of the Actionable Cancer Genome From Biomedical Literature Mining Validated by NGS Database

We constructed the gene-cancer association matrix from the filtered gene term -feature matrix to understand the

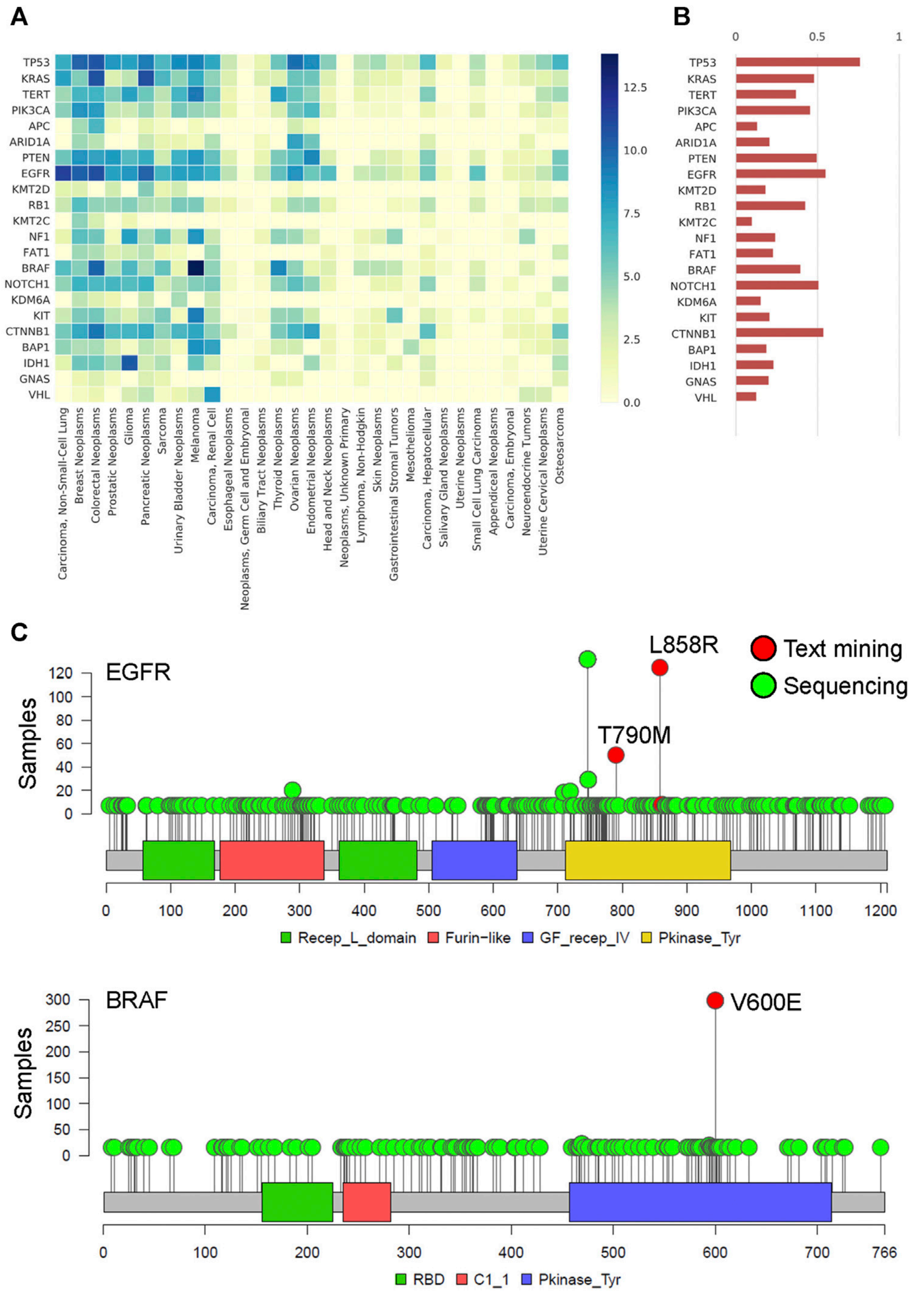
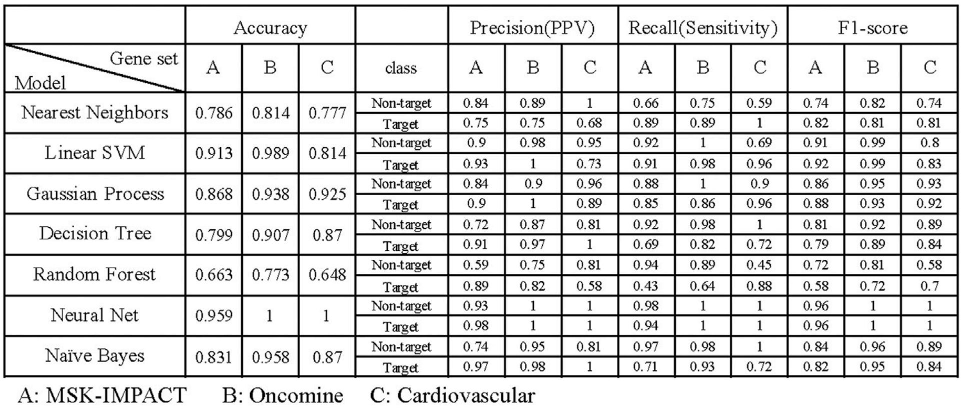


FIGURE 5 | The spectrum and frequency of actionable genetic mutation by literature mining **(A)** Heatmap of cancer genomics by the TF-IDF matrix. The X-axis represents the 31 common cancer types, and the y-axis represents the recurrent somatic genes. The darker color indicates a higher association between genes and cancer. **(B)** The bar plot shows the gene frequency within all of the cancer types. The data is validated by the MSK-IMPACT Clinical Sequencing Cohort, which is targeted sequencing of 10,000 clinical cases using the MSK-IMPACT assay. The cosine similarity of gene frequency between text mining and a statistical result from clinical sequencing data is 80.8%. **(C)** Lollipop plot of *EGFR* and *BRAF* in the MSK-IMPACT pan-cancer cohort. The critical gene mutation term features found by text mining are shown and labeled in red. Other gene mutations are labeled in green.

associations between cancer types and gene mutations. The recurrent common cancer-associated genes are shown in **Figure 5A**. The most common cancer-associated genes were *TP53*, *EGFR*, *CTNNB1*, *NOTCH1*, and *PTEN*, as shown in **Figure 5B**. Using two genes, *EGFR* and *BRAF*, as examples, we found that *EGFR* L858R and T790M and *BRAF* V600E were important mutation term features in text mining and were frequently mutated in MSK samples (**Figure 5C**). The cosine

A



Model \ Gene set	Accuracy				Precision(PPV)			Recall(Sensitivity)			F1-score		
	A	B	C	class	A	B	C	A	B	C	A	B	C
Nearest Neighbors	0.786	0.814	0.777	Non-target	0.84	0.89	1	0.66	0.75	0.59	0.74	0.82	0.74
				Target	0.75	0.75	0.68	0.89	0.89	1	0.82	0.81	0.81
Linear SVM	0.913	0.989	0.814	Non-target	0.9	0.98	0.95	0.92	1	0.69	0.91	0.99	0.8
				Target	0.93	1	0.73	0.91	0.98	0.96	0.92	0.99	0.83
Gaussian Process	0.868	0.938	0.925	Non-target	0.84	0.9	0.96	0.88	1	0.9	0.86	0.95	0.93
				Target	0.9	1	0.89	0.85	0.86	0.96	0.88	0.93	0.92
Decision Tree	0.799	0.907	0.87	Non-target	0.72	0.87	0.81	0.92	0.98	1	0.81	0.92	0.89
				Target	0.91	0.97	1	0.69	0.82	0.72	0.79	0.89	0.84
Random Forest	0.663	0.773	0.648	Non-target	0.59	0.75	0.81	0.94	0.89	0.45	0.72	0.81	0.58
				Target	0.89	0.82	0.58	0.43	0.64	0.88	0.58	0.72	0.7
Neural Net	0.959	1	1	Non-target	0.93	1	1	0.98	1	1	0.96	1	1
				Target	0.98	1	1	0.94	1	1	0.96	1	1
Naïve Bayes	0.831	0.958	0.87	Non-target	0.74	0.95	0.81	0.97	0.98	1	0.84	0.96	0.89
				Target	0.97	0.98	1	0.71	0.93	0.72	0.82	0.95	0.84

A: MSK-IMPACT B: Oncomine C: Cardiovascular

B

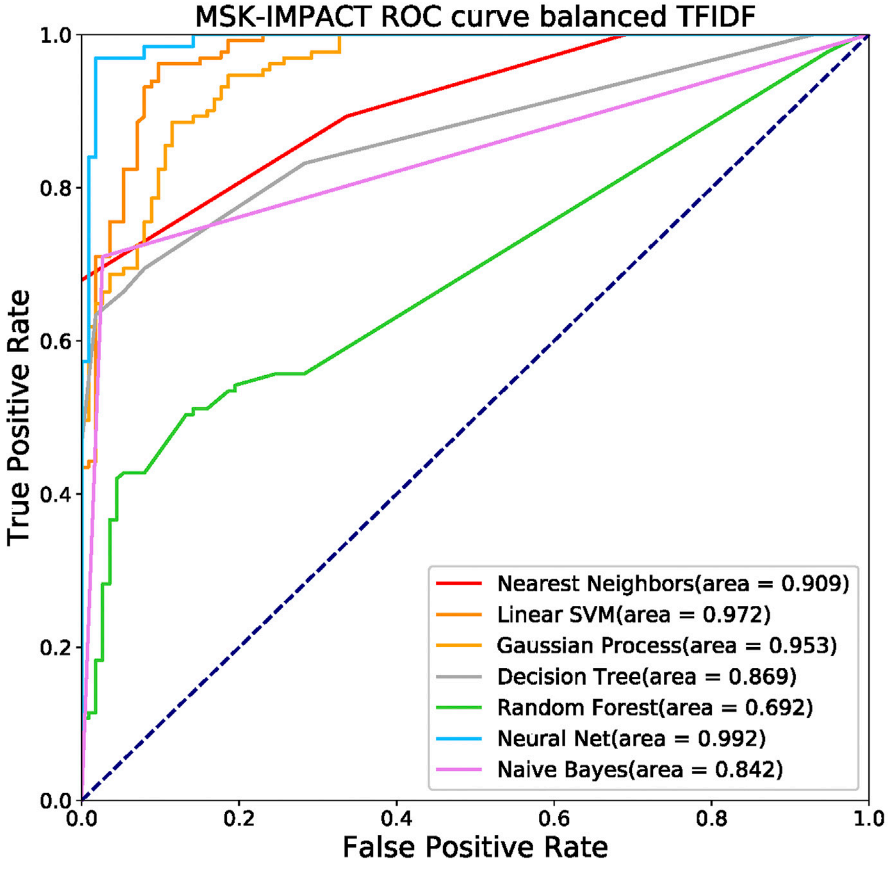


FIGURE 6 | Performance of the machine learning models with the gene panel **(A)** Evaluation of the overall accuracy, precision (positive predictive value, PPV), recall (sensitivity), and F1-score of every prediction model. Each gene could be labeled a target or non-target, indicating whether the gene is in the given target panel. The following seven prediction models were used: nearest neighbors, linear support vector machine (SVM), Gaussian process, decision tree, random forest, neural net, and Naive Bayes. The target gene panels were MSK-IMPACT, Oncomine Comprehensive Assay, and cardiovascular gene panels. **(B)** Receiver operating characteristic (ROC) curves of the models with the MSK-IMPACT 410-cancer gene panel. The neural net model had the highest area under the ROC curve (AUC), which was 0.992.

similarity of gene frequency between text mining and a statistical result from clinical sequencing data (Demeester et al., 2016) is 80.8% (**Figure 5B**). To understand the time series of the association between gene mutations and cancer types in the last decade, we constructed the gene-cancer TF-IDF matrixes of the years from 2011 to 2015 and the years from 2016 to 2019. As shown in **Supplementary Figure S2A and S2B**, we found that cancer immunotherapy was a major issue in the past 5 years. The rank of CD274 was increased, and CTLA4 first appeared (Seidel et al., 2018). In addition, the TF-IDF value of *BRAF* mutation in colorectal cancers increased because of the better outcomes of the *BRAF*-mutant CRC tumors with microsatellite instability (MSI) in immunotherapy (Rosenbaum et al., 2016). The results indicate that we can design a series of cancer gene panels by updating the literature mining time frame.

Gene Panel Prediction by Machine Learning Models

Seven machine learning prediction models, including nearest neighbors, linear support vector machine (SVM), Gaussian process, decision tree, random forest, neural net, and Naive Bayes (Wei et al., 2015), were used to verify the specific gene panel (**Figure 6A**). The MSK-IMPACT, Oncomine Comprehensive Assay (Rhodes et al., 2007), and cardiovascular gene panels (Paige et al., 2018) represent different gene characteristics. There are 410 essential cancer genes in the MSK-IMPACT panel. The Oncomine Comprehensive Assay includes 161 cancer-related genes. We used the congenital heart defect focus panel of 115 genes associated with congenital heart defects (CHDs) as the cardiovascular gene panels.

Each gene can be labeled as a target or non-target, which indicates whether the gene is in the given target panel. We performed five-fold cross-validation on our dataset to evaluate the models' efficiency and evaluate the overall accuracy of each prediction model. We measured the target and non-target genes in each prediction model separately with precision (positive predictive value, PPV), recall (sensitivity), and F1-score. The accuracies for nearest neighbors, linear SVM, Gaussian process, decision tree, random forest, neural net, and naive Bayes in the MSK-IMPACT panel were 0.786, 0.913, 0.868, 0.799, 0.663, 0.959, and 0.831, respectively; the accuracies for all models in the OCP gene panel were 0.814, 0.989, 0.938, 0.907, 0.773, 1 and 0.958; and the accuracies for all the models in the cardiovascular gene panel were 0.777, 0.814, 0.925, 0.87, 0.648, 1, and 0.87. The receiver operating characteristic (ROC) curve analysis confirmed that the neural net model had a better prediction performance; the area under the ROC curve (AUC) was 0.992 (**Figure 6B**). The AUCs of nearest neighbors, linear SVM, Gaussian process, decision tree, random forest, and naive Bayes were 0.909, 0.972, 0.953, 0.869, 0.692, and 0.842, respectively. The results of the biomedical term feature set prediction models are good, and the performance can reach up to 0.9. This means that the term feature sets can contain most of the information in the gene panel.

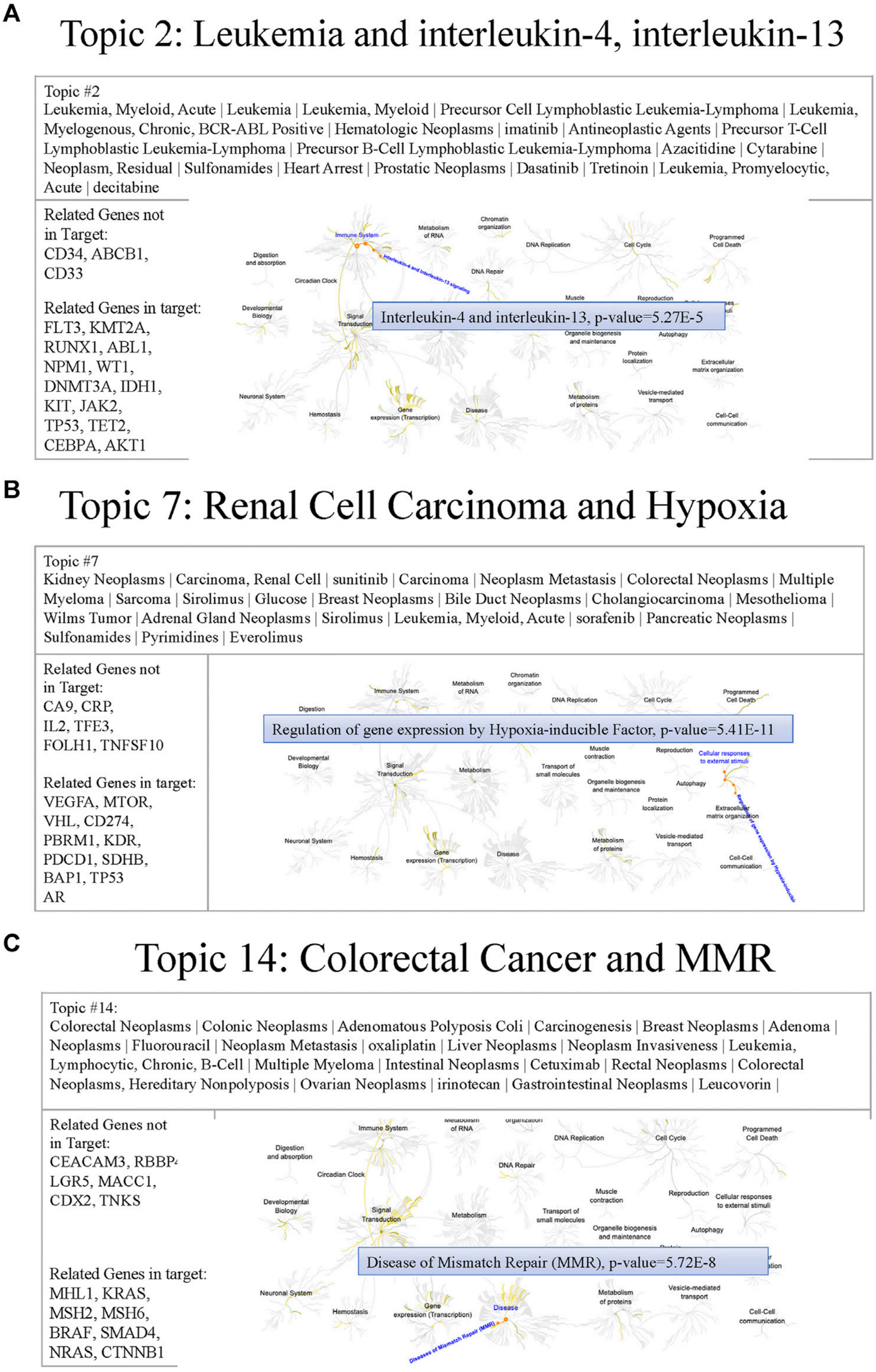
Design of Cancer-Related Gene Panels Based on Topic Modeling

To understand the MSK-IMPACT panel characteristics, we generated thirty topics that potentially represented different biomedical meanings. The following are some examples of issues relevant to genes in the MSK-IMPACT panel. **Figure 7** shows the text features, genes, and related pathways derived from the Reactome of topics 2, 7, and 14, including hematologic, and malignancies. In topic two, leukemia subtypes and targeted inhibitors (e.g., imatinib, dasatinib, and decitabine) were mined. Heart arrest, a common side effect of inhibitors for leukemia, was also been reported (Hochhaus et al., 2009). The related MSK-IMPACT panel in topic two was involved in the signaling of interleukin-4 and interleukin-13 (p = 5.27e-5), which was associated with the apoptosis of leukemia cells (Chaouchi et al., 1996; Peña-Martínez et al., 2018) (**Figure 7A**). These results indicated that topic two was associated with leukemia, a hematological malignancy. In topic seven, key text features such as kidney neoplasms, carcinoma, renal cell, and Wilms tumor implied the relationship between topic seven and kidney cancer. Inhibitors for kidney cancer, such as sorafenib and everolimus, were also identified (Martín-Aguilar et al., 2021; Ren et al., 2021). The hypoxia pathway enriched by *VHL*, *VEGFA*, and *PBRM1* (p = 5.41e-11) played a crucial role in the governance of cancer stem cells of renal cancer (Myszczyszyn et al., 2015) (**Figure 7B**). In topic 14, colorectal neoplasms, hereditary nonpolyposis, adenomatous polyposis coli, oxaliplatin, and cetuximab were associated with colon cancer. Related genes (e.g., *MLH1*, *MSH2*, and *MSH6*) in topic 14 were involved in mismatch repair (p = 5.72e-8), which has clinical importance in Lynch syndrome (Truninger et al., 2005) (**Figure 7C**). Other examples of different cancers, including brain cancer, gynecologic cancer, and breast cancer, are shown in **Supplementary Figure S3**. These results indicated that most of the genes in the MSK-IMPACT panel were collected for either therapeutic usage or biological relevance to various cancer types. In the future, we could design a small subset of multiple-gene groups by cancer topic.

DISSCUSSION

It is helpful to gain insight into the field that bridges the knowledge gap between valuable biomedical information and free text by text mining (Sachin Kumar Deshmukh, 2020). With biomedical text mining advances and its applications in cancer research, we can design cancer gene panels by the semantic interpretation of comprehensive cancer narratives. Here, we used a biomedical literature mining model to discover the characteristics of a gene panel. Importantly, we demonstrated and validated the performance of the machine learning approach in text mining of cancer information. Our results highlight the following important points. 1) We developed a gene panel analysis framework based on a biomedical text mining pipeline. 2) Our pipeline can enrich the term features of cancer gene panels. 3) We demonstrated and validated the patterns of the cancer mutational landscape by NGS database. 4) The non-negative matrix factorization (NMF) method and topic modeling are useful for generating cancer information. Biomedical literature mining is valuable for discovering the inherent characteristics of gene panels. These results could be applied to the classification of cancer-related information and strategies for novel cancer gene panel designs.

The hypergeometric distribution test is one of the practical machine learning tools in TM. It can be used to select and extract term features from various genomic characterizations (Pal, 2017).



A

Topic 2: Leukemia and interleukin-4, interleukin-13

Topic #2
Leukemia, Myeloid, Acute | Leukemia | Leukemia, Myeloid | Precursor Cell Lymphoblastic Leukemia-Lymphoma | Leukemia, Myelogenous, Chronic, BCR-ABL Positive | Hematologic Neoplasms | imatinib | Antineoplastic Agents | Precursor T-Cell Lymphoblastic Leukemia-Lymphoma | Precursor B-Cell Lymphoblastic Leukemia-Lymphoma | Azacitidine | Cytarabine | Neoplasm, Residual | Sulfonamides | Heart Arrest | Prostatic Neoplasms | Dasatinib | Tretinoin | Leukemia, Promyelocytic, Acute | decitabine

Related Genes not in Target:
CD34, ABCB1, CD33

Related Genes in target:
FLT3, KMT2A, RUNX1, ABL1, NPM1, WT1, DNMT3A, IDH1, KIT, JAK2, TP53, TET2, CEBPA, AKT1

Interleukin-4 and interleukin-13, p-value=5.27E-5

B

Topic 7: Renal Cell Carcinoma and Hypoxia

Topic #7
Kidney Neoplasms | Carcinoma, Renal Cell | sunitinib | Carcinoma | Neoplasm Metastasis | Colorectal Neoplasms | Multiple Myeloma | Sarcoma | Sirolimus | Glucose | Breast Neoplasms | Bile Duct Neoplasms | Cholangiocarcinoma | Mesothelioma | Wilms Tumor | Adrenal Gland Neoplasms | Sirolimus | Leukemia, Myeloid, Acute | sorafenib | Pancreatic Neoplasms | Sulfonamides | Pyrimidines | Everolimus

Related Genes not in Target:
CA9, CRP, IL2, TFE3, FOLH1, TNFSF10

Related Genes in target:
VEGFA, MTOR, VHL, CD274, PBRM1, KDR, PDCD1, SDHB, BAP1, TP53 AR

Regulation of gene expression by Hypoxia-inducible Factor, p-value=5.41E-11

C

Topic 14: Colorectal Cancer and MMR

Topic #14:
Colorectal Neoplasms | Colonic Neoplasms | Adenomatous Polyposis Coli | Carcinogenesis | Breast Neoplasms | Adenoma | Neoplasms | Fluorouracil | Neoplasm Metastasis | oxaliplatin | Liver Neoplasms | Neoplasm Invasiveness | Leukemia, Lymphocytic, Chronic, B-Cell | Multiple Myeloma | Intestinal Neoplasms | Cetuximab | Rectal Neoplasms | Colorectal Neoplasms, Hereditary Nonpolyposis | Ovarian Neoplasms | irinotecan | Gastrointestinal Neoplasms | Leucovorin |

Related Genes not in Target:
CEACAM3, RBBP4 LGR5, MACC1, CDX2, TNKS

Related Genes in target:
MHL1, KRAS, MSH2, MSH6, BRAF, SMAD4, NRAS, CTNNB1

Disease of Mismatch Repair (MMR), p-value=5.72E-8

FIGURE 7 | Examples of cancer topics containing relevant text features, genes, and pathways **(A)** Figure showing the text features, genes, and pathways of topic 2. Cancer types (e.g., leukemia) and inhibitors (e.g., imatinib) were reported in this topic. Reactome pathway analysis revealed that the related genes of the MSK-IMPACT panel in topic 2 (e.g., FLT3) were involved in interleukin-4 and interleukin-13 signaling (p = 5.27e-5). **(B)** Figure showing the text features, genes, and pathways of topic 7. Text features including cancer types (e.g., kidney neoplasms) and inhibitors (e.g., sorafenib) implied the relationship between topic seven and kidney cancer. The hypoxia pathway enriched by related genes (e.g., VHL) of the MSK-IMPACT panel in topic 7 (p = 5.41e-11) played a crucial role in the governance of cancer stem cells of renal cancer. **(C)** Figure showing the text features, genes, and pathways of topic 14. Many text features containing cancer types (e.g., colorectal neoplasms) and inhibitors (e.g., oxaliplatin) indicated the association between topic 14 and colon cancer. Related genes of the MSK-IMPACT panel in topic 14 (e.g., MLH1) were involved in the mismatch repair pathway (p = 5.72e-8).

We identified the critical term features according to the gene panel using *p*-values based on a hypergeometric test. Our term feature selection methods can distinguish in different gene panels. This implicates a high-performance prediction model for different datasets, including the MSK-IMPACT panel, Oncomine Cancer Panel, and cardiovascular gene panels. Although many gene recommendation algorithms have been developed, little is known about gene panel design.

Our biomedical term tagging algorithm provides a compressive cancer gene panel and related information. With our tagging algorithm, most of the essential biomedical terms in the text have been tagged. The construction of a gene term-feature matrix in different categories provides useful profiling for the characteristics of the genes. In this study, we constructed a biologically meaningful platform to analyze gene panels in terms of the diseases, chemicals, mutations, and MeSH terms related to genes. We can implement more biomedical term feature matrixes, such as a drug-feature matrix and disease-feature matrix. These different types of forms can provide strategies to analyze biology. With NMF topic modeling, we can capture cancer gene-drug information compatible with our knowledge. It will be useful to design a small subset of cancer gene panels by interpreting the topic model.

For the discovery of cancer gene panels, **Figure 5A** and **Figure 7C** illustrate an example of a cancer gene panel design for colorectal cancer. The most frequent genes are *KRAS*, *EGFR*, *BRAF*, *PTEN*, *TP53*, *MLH1*, *PIK3CA*, *CTNNB1* in colorectal cancer by the heatmap. Hereditary nonpolyposis colon cancer (HNPCC) is caused by inherited mismatch repair genetic mutations, including *MLH1, MSH2,* and *MSH6.* The lifetime ovarian cancer risk increased in HNPCC. We can find ovarian cancer and a gene panel including *MLH1*, *MSH2*, *MSH6*, *BRAF*, *KRAS*, *SMAD4*, *NRAS*, *CTNNB1* by topic model. In our study, we can design the two different cancer panels by phenotype. These results indicated the platform could provide an opportunity to construct a cancer gene panel recommendation by different cancer subtypes. There are some text mining limitations in our study. The entity-term based features are based only on co-occurrence in three sentences. However, related entities may have distinct relationships, which are not necessarily co-occurred. The features were obtained from only one resource, PubMed abstracts. Many curated databases have many useful biological features of genes or diseases or drugs; for example, Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium., 2017) contains GO terms that describe genes by the functions of genes or cellular components. It may provide a benefit to the cancer researcher. Unfortunately, the TF-IDF table is going to weight toward common diseases and omit those that are critical in identifying rare diseases. The gene panels are not useful for the identification of unknown or rare gene mutations that are important for treatment. Simultaneously, the manuscripts and supplementary materials may also provide more critical results, but the lack of standardization in accessing this information is a significant problem. The text mining method often focuses on a few sentences due to the challenges of creating a complicated relationship between several critical keywords.

As we know, the random forest algorithm performed well than the decision tree in most of pattern classification cases. However, we found that the random forest approach presented a worse ability for cancer gene panel prediction in the experiments. Several reasons may cause this situation in the model training and evaluation, such as whether or not we specify the maximum number of features to be included at each node split. One of the reasons is that the random forest builds subtrees by randomly choosing features from amounts of features in our study. Unlike the other methods, they calculated the weights for each feature by determining the importance of all features. Thus, the performance might be increased when we increase the number of trees in the random forest. Because the subtrees increased, the model will be seen more features to build more diverse trees. Therefore, the model will become robust and make an excellent performance. Nevertheless, in this paper, we are focusing on a pipeline that can contextualize genes. We used the default parameter in most of the methods in our study. Although we are not emphasizing the methods and parameters optimization, it is also an important issue that we will study in our future works.

Several text mining systems have been developed for mutation-disease association (Erdogmus and Sezermen., 2007; Yeniterzi and Sezerman., 2009; Singhal et al., 2016). An automated pipeline using the full-length biomedical literature was recently established and validated by evidence-based gene panels (Saberian et al., 2020). All these methods focus on mutation-disease associations. In contrast, we contextualized the genes for clinical precision medicine. We provide information about druggable targets, mutations in hereditary cancer syndrome, and disease subtypes.

Although many text mining-based gene panel algorithms were developed, there is still little known to validate the gene panel characteristics. This study provides a biomedical literature mining pipeline in gene panel discovery and interpretation. The platform validated by NGS database could provide an opportunity to construct a gene recommendation and annotation system for precision medicine.

CONCLUSIONS

In conclusion, this study highlights the importance of biomedical literature mining in gene panel discovery and interpretation. The platform could provide an opportunity to construct a gene recommendation and annotation system for precision medicine.

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.

AUTHOR CONTRIBUTIONS

Conception and study design: H-OC, P-CL, C-RL, C-SW, and J-HC; Development of methodology: H-OC, P-CL, J-HC; Acquisition of data: H-OC; C-RL, and C-SW; Statistical and computational analysis: P-CL, H-OC, C-RL, and J-HC; Writing, review, and revision of the manuscript: H-OC, P-CL, C-RL, and J-HC; Study supervision: J-HC; All authors have read and approved the manuscript. All authors agree for publication.

FUNDING

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan under Research Grant of MOST 110-2634-F-006-014 and MOST 110-2634-F-006-020, Ministry of Health and Welfare (MOHW110-TDU-B-211-144018). All authors have read and approved the manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the significant contribution of Kimforest LTD. Taiwan, KD Yang, CC Pan, and CJ Lee for the bioinformatics support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.771435/full#supplementary-material

REFERENCES

Arriagada, R., Bergman, B., Dunant, A., Le Chevalier, T., Pignon, J. P., and Vansteenkiste, J. (2004). & International Adjuvant Lung Cancer Trial Collaborative GroupCisplatin-Based Adjuvant Chemotherapy in Patients with Completely Resected Non-small-cell Lung Cancer. *N. Engl. J. Med.* 350 (4), 351–360. doi:10.1056/NEJMoa031644

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Azam, F., Musa, A., Dehmer, M., Yli-Harja, O. P., and Emmert-Streib, F. (2019). Global Genetics Research in Prostate Cancer: A Text Minning and Computational Network Theory Approach. *Front. Genet.* 10, 70. doi:10.3389/fgene.2019.00070

Burris, H. A., 3rd (2004). Dual Kinase Inhibition in the Treatment of Breast Cancer: Initial Experience with the EGFR/ErbB-2 Inhibitor Lapatinib. *Oncologist* 9 (Suppl. 3), 10–15. doi:10.1634/theoncologist.9-suppl_3-10

Chaouchi, N., Wallon, C., Goujard, C., Tertian, G., Rudent, A., Caput, D., et al. (1996). Interleukin-13 Inhibits Interleukin-2-Induced Proliferation and Protects Chronic Lymphocytic Leukemia B Cells from *In Vitro* Apoptosis. *Blood* 87 (3), 1022–1029.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., et al.BRIM-3 Study Group (2011). Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *N. Engl. J. Med.* 364 (26), 2507–2516. doi:10.1056/NEJMoa1103782

Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., et al. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* 17 (3), 251–264. doi:10.1016/j.jmoldx.2014.12.006

Choo, J., Lee, C., Reddy, C. K., and Park, H. (2013). UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Trans. Vis. Comput. Graph.* 19 (12), 1992–2001. doi:10.1109/TVCG.2013.212

Demeester, T., Sutskever, I., Chen, K., Dean, J., and Corado, G. (2016). Distributed Representations of Words and Phrases and Their Compositionality. *EMNLP 2016 – Conf. Empir. Methods Nat. Lang. Process. Proc.*, 1389–1399. arXiv: 1606.08359.

Devarajan, K., Wang, G., and Ebrahimi, N. (2015). A Unified Statistical Approach to Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. *Mach. Learn.* 99 (1), 137–163. doi:10.1007/s10994-014-5470-z

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. (2019). Gene2vec: Distributed Representation of Genes Based on Co-expression. *BMC Genomics* 20 (Suppl. 1), 82. doi:10.1186/s12864-018-5370-x

Erdogmus, M., and Sezerman, O. U. (2007). Application of Automatic Mutation-Gene Pair Extraction to Diseases. *J. Bioinform. Comput. Biol.* 5 (6), 1261–1275. doi:10.1142/s021972000700317x

Hauschild, A., Grob, J. J., Demidov, L. V., Jouary, T., Gutzmer, R., Millward, M., et al. (2012). Dabrafenib in BRAF-Mutated Metastatic Melanoma: a Multicentre, Open-Label, Phase 3 Randomised Controlled Trial. *Lancet* 380 (9839), 358–365. doi:10.1016/S0140-6736(12)60868-X

Hochhaus, A., O'Brien, S. G., Guilhot, F., Druker, B. J., Branford, S., Foroni, L., et al.IRIS Investigators (2009). Six-year Follow-Up of Patients Receiving Imatinib for the First-Line Treatment of Chronic Myeloid Leukemia. *Leukemia* 23 (6), 1054–1061. doi:10.1038/leu.2009.38

Hyman, D. M., Solit, D. B., Arcila, M. E., Cheng, D. T., Sabbatini, P., Baselga, J., et al. (2015). Precision Medicine at Memorial Sloan Kettering Cancer Center: Clinical Next-Generation Sequencing Enabling Next-Generation Targeted Therapy Trials. *Drug DiscovToday* 20 (12), 1422–1428. doi:10.1016/j.drudis.2015.08.005

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. *WSEAS Trans. Comput.* 4.

Kumar Deshmukh, S. (2020). Machine Learning for Precision Medicine in Cancer-Transforming Drug Discovery and Treatment. *J. Cancer Biol.* 1, 20–22. doi:10.46439/cancerbiology.1.005

Leaman, R., Islamaj Dogan, R., and Lu, Z. (2013). DNorm: Disease Name Normalization with Pairwise Learning to Rank. *Bioinformatics* 29 (22), 2909–2917. doi:10.1093/bioinformatics/btt474

Luthra, R., Patel, K. P., Routbort, M. J., Broaddus, R. R., Yau, J., Simien, C., et al. (2017). A Targeted High-Throughput Next-Generation Sequencing Panel for Clinical Screening of Mutations, Gene Amplifications, and Fusions in Solid Tumors. *J. Mol. Diagn.* 19 (2), 255–264. doi:10.1016/j.jmoldx.2016.09.011

Martín-Aguilar, A. E., Núñez-López, H., and Ramirez-Sandoval, J. C. (2021). Sorafenib as a Second-Line Treatment in Metastatic Renal Cell Carcinoma in Mexico: a Prospective Cohort Study. *BMC Cancer* 21, 1–9. doi:10.1186/s12885-020-07720-5

McCabe, M. J., Gauthier, M. A., Chan, C. L., Thompson, T. J., De Sousa, S., Puttick, C., et al. (2019). Development and Validation of a Targeted Gene Sequencing Panel for Application to Disparate Cancers. *Sci. Rep.* 9 (1), 17052. doi:10.1038/s41598-019-52000-3

Myszczyszyn, A., Czarnecka, A. M., Matak, D., Szymanski, L., Lian, F., Kornakiewicz, A., et al. (2015). The Role of Hypoxia and Cancer Stem Cells in Renal Cell Carcinoma Pathogenesis. *Stem Cel Rev. Rep.* 11 (6), 919–943. doi:10.1007/s12015-015-9611-y

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., et al. (2004). EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science* 304 (5676), 1497–1500. doi:10.1126/science.1099314

Paige, S. L., Saha, P., and Priest, J. R. (2018). Beyond Gene Panels: Whole Exome Sequencing for Diagnosis of Congenital Heart Disease. *Circ. Genom. Precis. Med.* 11 (3), e002097. doi:10.1161/CIRCGEN.118.002097

Pal, R. (2017). *Feature Selection and Extraction from Heterogeneous Genomic Characterizations*. Predictive Modeling of Drug Sensitivity, 45–81. doi:10.1016/b978-0-12-805274-7.00003-8

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., and Grisel, O. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. arXiv: 201.0490.

Peña-Martínez, P., Eriksson, M., Ramakrishnan, R., Chapellier, M., Högberg, C., Orsmark-Pietras, C., et al. (2018). Interleukin 4 Induces Apoptosis of Acute Myeloid Leukemia Cells in a Stat6-dependent Manner. *Leukemia* 32 (3), 588–596. doi:10.1038/leu.2017.261

Ren, Z., Niu, Y., Fan, B., Wei, S., Ma, Y., Zhang, X., et al. (2021). Clinical Analysis of Everolimus in the Treatment of Metastatic Renal Cell Carcinoma. *Ann. Palliat. Med.* 10. doi:10.21037/apm-20-2465

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia* 9 (2), 166–180. doi:10.1593/neo.07112

Rosenbaum, M. W., Bledsoe, J. R., Morales-Oyarvide, V., Huynh, T. G., and Mino-Kenudson, M. (2016). PD-L1 Expression in Colorectal Cancer Is Associated with Microsatellite Instability, BRAF Mutation, Medullary Morphology and Cytotoxic Tumor-Infiltrating Lymphocytes. *Mod. Pathol.* 29 (9), 1104–1112. doi:10.1038/modpathol.2016.95

Saberian, N., Shafi, A., Peyvandipour, A., and Draghici, S. (2020). MAGPEL: an autoMated Pipeline for Inferring vAriant-Driven Gene PanEls from the Full-Length Biomedical Literature. *Sci. Rep.* 10 (1), 12365. doi:10.1038/s41598-020-68649-0

Seidel, J. A., Otsuka, A., and Kabashima, K. (2018). Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations. *Front. Oncol.* 8, 86. doi:10.3389/fonc.2018.00086

Shabani Azim, F., Houri, H., Ghalavand, Z., and Nikmanesh, B. (2018). Next Generation Sequencing in Clinical Oncology: Applications, Challenges and Promises: A Review Article. *Iran. J. Public Health* 47, 1453–1457. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6277731/.

Shepherd, F. A., Rodrigues Pereira, J., Ciuleanu, T., Tan, E. H., Hirsh, V., Thongprasert, S., et al. (2005). Erlotinib in Previously Treated Non-small-cell Lung Cancer. *N. Engl. J. Med.* 353 (2), 123–132. doi:10.1056/NEJMoa050753

Singhal, A., Simmons, M., and Lu, Z. (2016). Text Mining for Precision Medicine: Automating Disease-Mutation Relationship Extraction from Biomedical Literature. *J. Am. Med. Inform. Assoc.* 23 (4), 766–772. doi:10.1093/jamia/ocw041

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology Knowledgebase and Resources. *Nucleic Acids Res.* 45 (D1), D331–D338. doi:10.1093/nar/gkw1108

Truninger, K., Menigatti, M., Luz, J., Russell, A., Haider, R., Gebbers, J. O., et al. (2005). Immunohistochemical Analysis Reveals High Frequency of PMS2 Defects in Colorectal Cancer. *Gastroenterology* 128 (5), 1160–1171. doi:10.1053/j.gastro.2005.01.056

Wang, C. C. N., Jin, J., Chang, J. G., Hayakawa, M., Kitazawa, A., Tsai, J. J. P., et al. (2020). Identification of Most Influential Co-occurring Gene Suites for Gastrointestinal Cancer Using Biomedical Literature Mining and Graph-Based Influence Maximization. *BMC Med. Inform. Decis. Mak.* 20, 1–12. doi:10.1186/s12911-020-01227-6

Wang, Y., Wu, S., Li, D., Mehrabi, S., and Liu, H. (2016). A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval. *J. Biomed. Inform.* 63, 379–389. doi:10.1016/j.jbi.2016.08.026

Wei, C. H., Kao, H. Y., and Lu, Z. (2015). GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *Biomed. Res. Int.* 918710. doi:10.1155/2015/918710

Wei, C. H., Kao, H. Y., and Lu, Z. (2013). PubTator: a Web-Based Text Mining Tool for Assisting Biocuration. *Nucleic Acids Res.* 41, W518. doi:10.1093/nar/gkt441

Westlake, A. J., and Larson, H. J. (1970). Introduction to Probability Theory and Statistical Inference. *Stat* 19, 352.

Yeganova, L., Kim, W., Kim, S., and Wilbur, W. J. (2014). Retro: Concept-Based Clustering of Biomedical Topical Sets. *Bioinformatics* 30 (22), 3240–3248. doi:10.1093/bioinformatics/btu514

Yeniterzi, S., and Sezerman, U. (2009). EnzyMiner: Automatic Identification of Protein Level Mutations and Their Impact on Target Enzymes from PubMed Abstracts. *BMC bioinformatics* 10 (Suppl. 8Suppl 8), S2. doi:10.1186/1471-2105-10-S8-S2

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., et al. (2017). Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nat. Med.* 23 (6), 703–713. doi:10.1038/nm.4333

Zhou, W., Ercan, D., Chen, L., Yun, C. H., Li, D., Capelletti, M., et al. (2009). Novel Mutant-Selective EGFR Kinase Inhibitors against EGFR T790M. *Nature* 462 (7276), 1070–1074. doi:10.1038/nature08622

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., et al. (2013). Biomedical Text Mining and its Applications in Cancer Research. *J. Biomed. Inform.* 46 (2), 200–211. doi:10.1016/j.jbi.2012.10.007

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fgene.2021.733654

Recurrence Risk of Liver Cancer Post-hepatectomy Using Machine Learning and Study of Correlation With Immune Infiltration

Xiaowen Qian[1], Huilin Zheng[2], Ke Xue[1], Zheng Chen[3], Zhenhua Hu[3,4,5], Lei Zhang[1,2]* and Jian Wan[1]**

[1]Department of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China, [2]Department of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou, China, [3]Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, Fourth Affiliated Hospital, School of Medicine, Zhejiang University, Yiwu, China, [4]Key Laboratory of Combined Multi-Organ Transplantation, Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, First Affiliated Hospital, School of Medicine, Zhejiang University, Ministry of Public Health Key Laboratory of Organ Transplantation, Hangzhou, China, [5]Division of Hepatobiliary and Pancreatic Surgery, Yiwu Central Hospital, Yiwu, China

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Takayuki Shiomi,
International University of Health and Welfare, Japan

****Correspondence:***
Huilin Zheng
zhenghuilin@zust.edu.cn
Lei Zhang
papaver_rhoeas@126.com
Jian Wan
wanjian@zust.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 June 2021*
Accepted: *24 November 2021*
Published: *08 December 2021*

Citation:
Qian X, Zheng H, Xue K, Chen Z, Hu Z, Zhang L and Wan J (2021) Recurrence Risk of Liver Cancer Post-hepatectomy Using Machine Learning and Study of Correlation With Immune Infiltration. Front. Genet. 12:733654. doi: 10.3389/fgene.2021.733654

Postoperative recurrence of liver cancer is the main obstacle to improving the survival rate of patients with liver cancer. We established an mRNA-based model to predict the risk of recurrence after hepatectomy for liver cancer and explored the relationship between immune infiltration and the risk of recurrence after hepatectomy for liver cancer. We performed a series of bioinformatics analyses on the gene expression profiles of patients with liver cancer, and selected 18 mRNAs as biomarkers for predicting the risk of recurrence of liver cancer using a machine learning method. At the same time, we evaluated the immune infiltration of the samples and conducted a joint analysis of the recurrence risk of liver cancer and found that B cell, B cell naive, T cell $CD4^+$ memory resting, and T cell $CD4^+$ were significantly correlated with the risk of postoperative recurrence of liver cancer. These results are helpful for early detection, intervention, and the individualized treatment of patients with liver cancer after surgical resection, and help to reveal the potential mechanism of liver cancer recurrence.

Keywords: liver cancer, recurrence risk, machine learning, immune infiltration, TCGA

INTRODUCTION

Liver cancer is a common malignancy with morbidity and mortality ranking sixth and fourth, respectively. (Bray et al., 2018). Surgical resection is the most common treatment method for liver cancer. However, the high recurrence and metastasis rate of liver cancer patients after resection poses a great challenge to liver cancer treatment. According to statistics, the recurrence rate of liver cancer patients 3 years after surgery is approximately 40–50%, and the recurrence rate 5 years after surgery is as high as 60–70%. (Nakayama and Takayama, 2014). Therefore, it is of great clinical significance to identify high-risk patients with recurrence of liver cancer after radical surgical resection.

With the development of high-throughput sequencing technology, some molecular biomarkers have been reported to predict liver cancer recurrence after hepatectomy. In previous studies, Erb-B2 receptor tyrosine kinase 2 (*ERBB2)* and NUF2 component of the NDC80 Kinetochore Complex (*NUF2)* were reported to be biomarkers of hepatocellular carcinoma (HCC) recurrence after surgery. (Li et al., 2017a), (Feng et al., 2021) D. Wang et al. (Chen et al., 2020) demonstrated that the level of interleukin 11 (*IL11*) increased after hepatectomy, which led to the growth of HCC. Zhou et al. (Chen et al., 2021) confirmed that WNK lysine deficient protein kinase 2 (*WNK2)* is a driving factor of HCC and a risk factor for early recurrence through genome sequencing. However, due to the complex etiology of liver cancer (such as hepatitis virus infection, alcohol-related liver disease, nonalcoholic fatty liver) and the difficulty of collecting liver cancer samples with a history of recurrence after hepatectomy, the predictive effect of individual genes screened from a limited number of samples in previous studies may be limited and may not be generally applicable to liver cancer caused by different etiologies. Our study included 306 liver cancer samples caused by different etiologies from The Cancer Genome Atlas, which is the largest sample size that can be incorporated with both recurrent disease history and sequencing data currently, improving the reliability of prediction.

Machine learning is a mathematical method for finding patterns in data to achieve artificial intelligence, and is now widely used in medical image detection and medical aid diagnosis. (Li et al., 2017a; Chen et al., 2020; Chen et al., 2021; Feng et al., 2021). Some studies have used machine learning methods to predict the recurrence of liver cancer. Ho et al.(Ho et al., 2006) constructed a KNN model based on 18 patients to predict HCC recurrence after resection. Liang et al. (Liang et al., 2014) analyzed 83 patients with HCC after radio frequency ablation and constructed an SVM model to predict HCC recurrence. However, it is worth noting that KNN and SVM have good predictive performance, but they are commonly used to solve classification problems and are not very explanatory for variables. Random forest in machine learning is an ensemble classifier, and the tree-based ensemble makes it suitable for handling with redundant features. (Reif et al., 2006). Unlike KNN, SVM and random forest, logistic regression is a regression analysis that gives formulaic results to quantify the probability of event occurrence, with good interpretation of variables. (Nick and Campbell, 2007).

In our study, we aimed to identify potential biomarkers of liver cancer recurrence after resection, and construct a model to quantify the risk of recurrence based on the biomarkers. The above advantages of logistic regression make it a suitable algorithm for our study. In order to solve the problem that logistic regression is difficult to handle with high-dimensional variables, we performed random forest to screen variables. The model constructed using this combined method allows for the quantification of the risk of liver cancer recurrence and good predictive performance.

MATERIALS AND METHODS

Patient Selection

The RNA-seq data and clinical data of samples with liver cancer were downloaded from The Cancer Genome Atlas database (https://portal.gdc.cancer.gov/). The detailed clinical data included age, sex, TNM stage, histologic grade, and recurrence (**Supplementary Table S1**). Each file downloaded was the mRNA expression data of one sample, and we collated all sample data into one file to obtain an mRNA expression matrix of all samples, while matching the clinical information to the corresponding samples. The inclusion criteria for the cohort were as follows: 1) Normal tissue samples were removed. 2) Samples with R0 excision were selected for further analyses. Next, according to whether new tumor events occurred after the initial treatment, the patients were divided into recurrence and non-recurrence groups. Patients who had a new tumor event after the initial treatment and the type of new tumor event were intrahepatic recurrence, locoregional recurrence, or extrahepatic recurrence were included in the recurrence group. Samples that did not develop new tumor events after the initial treatment were included in the non-recurrence group. Finally, 306 usable samples were obtained, including 158 non-recurrent and 148 recurrent samples. The clinical characteristics of the recurrence and non-recurrence groups are shown in **Supplementary Table S2**.

In the subsequent analysis, the samples were further divided into two subgroups based on etiology: the alcohol-associated liver disease subgroup (ALD, n = 57) and the hepatitis virus infection-related subgroup (HVI, n = 98). The HVI subgroup included those associated with hepatitis B virus (HBV, n = 71) and hepatitis C virus (HCV, n = 27) infections. Considering the small sample size of HCV, HCV and HBV were not split into two subgroups and were uniformly classified as the HVI subgroup. The remaining patients (n = 151) were not included in the subgroup analysis due to missing etiologies.

Screening of DE-mRNAs

DESeq and EdgeR packages in R software (https://www.r-project.org/) were used to screen DE-mRNAs between the recurrence and non-recurrence groups. The threshold was set at $p < 0.05$, and |log2 (fold change)|>1. Overlapping DE-mRNAs screened using the two packages were used for further analyses.

Functional Enrichment and Co-expression Network

Gene Ontology (GO) functional enrichment of these DE-mRNAs was conducted using the database for annotation, visualization, and integrated discovery (https://david.ncifcrf.gov/), and $p < 0.05$, was set as the cutoff value. The co-expression network between DE-mRNAs was predicted using the STRING database (https://string-db.org/) and visualized using Cytoscape (https://cytoscape.org/).

Estimation of Immune Cells and Identification of Differential Immune Cells

Based on the mRNA expression levels of the recurrence and non-recurrence groups, we used the TIMER database (http://timer.cistrome.org/) to estimate the expression of immune cells in the two groups.

Development of the Risk Assessment Model

The overall flow chart of this study was shown in **Supplementary Figure S1**. 306 patients were assigned to the training cohort (n = 215) or validation cohort (n = 91) with a 7:3 split ratio by applying simple random sampling. In the training cohort, 60 DE-mRNAs were entered into random forest (RF) as variables. RF is an ensemble method based on multiple decision trees, and the decision trees in RF are built by randomly selected samples made from bootstrap and a randomly selected subset of variables. Some original samples were not selected to construct trees, which are called out-of-bag (OOB) dataset. (Breiman, 2001). After inputting 60 variables into RF, mean decrease accuracy (MDA) and mean decrease Gini (MDG) provided in RF were calculated to assess the importance scores of each variable for liver cancer recurrence. MDA quantifies the importance of a variable by calculating the mean decrease accuracy in the OOB before and after the permutation of the variable, and MDG quantifies the importance of a variable by measuring the mean decrease in Gini impurity caused by the variable when it is used to form a split in the random forest. (Díaz-Uriarte and de Andrés, 2006; Wang et al., 2016). These were achieved by the "randomForest" package in R software. Next, a variable ranking was obtained by ranking the importance scores. Larger values of the importance scores represented greater influence of DE-mRNAs on recurrence. The variable ranking can be described as:

$$X_{Rank} = \left(x^{r(1)}, x^{r(2)}, \cdots, x^{r(d)}\right)$$

where $r(i), i = 1, 2, \cdots, d$ is the index of variable x_i in the descending ranking. We took MDA as the importance scores to obtain the top 30 importance ranking of variables $\tilde{X}_{Rank1}$, and took MDG as the importance scores to obtain the top 30 importance ranking of variables $\tilde{X}_{Rank2}$. In order to obtain a small scale of variables and increase reliability, we took the intersection of the variables in $\tilde{X}_{Rank1}$ and $\tilde{X}_{Rank2}$ as the variables to be put into the risk assessment model, which can be presented as:

$$X = \tilde{X}_{Rank1} \cap \tilde{X}_{Rank2}$$

After the above steps, 18 DE-mRNAs with important effects on liver cancer recurrence were obtained as the input variables for the risk assessment model.

Multivariate logistic regression was used to construct the risk assessment model for liver cancer recurrence. The calculation formula of risk score can be presented as:

$$Risk\ score = \frac{1}{1 + e^{-\left(\omega^T X + b\right)}}$$

In our study, X is the expression level of each DE-mRNA, and the parameters ω and b can be learned from the training data. The parameter ω can be interpreted as the relative impact of each DE-mRNA in the recurrence of liver cancer. In the training cohort, 18 DE-mRNAs screened by random forest were entered into logistic regression, and the model was validated using the validation cohort.

In addition, a logistic regression model without variable screening (model 2), a logistic model with stepwise regression (model 3), and a logistic model with L1 regularization (model 4) were constructed for comparison. In model 2, a logistic regression model was constructed directly with 60 DE-mRNAs. Model 3 was constructed by stepwise logistic regression with Akaike information criterion (AIC) for feature selection, which is a classic variable selection method. (Agostini et al., 2015; Sanchez-pinto et al., 2018; Hu et al., 2019).Model 4 added the L1 regularization term to the logistic regression, which can have a dimensionality reduction effect. (Tibshirani, 1996; Wang et al., 2014). The "glmnet" package in R software was used for the analyses. StromalScore, ImmuneScore, and ESTIMATEScore are three scores used to assess the level of infiltrating stromal and immune cells. (Yoshihara et al., 2013). They were calculated to compare with the risk score presented in our study, using the "estimate" package in R software.

Performance Evaluation of the Prediction Models

The receiver operating characteristics (ROC) curve and the area under the ROC curve (AUC) were used to evaluate the performance of the models. As measures of model performance, they exhibit some desirable properties and are good ways to visualize model performance. (Bradley, 1997). In the field of big biological data and cancer-related research, ROC and AUC are widely used to evaluate the performance of machine learning models. (Le et al., 2020; Yu et al., 2020; Kudo et al., 2021; Le et al., 2021). ROC and AUC analyses were performed by R software.

Statistical Analysis

All statistical analyses were performed using GraphPad Prism version 8.0 software (GraphPad Software Inc.). A two-tailed *t*-test was used to identify immune cells that were significantly different between the recurrence and non-recurrence groups. In all analyses, a two-tailed *p*-value less than 0.05, was considered statistically significant. Random forest and logistic regression models were performed using R software (version 4.0.2). The codes for this study have been uploaded on GitHub (https://github.com/polarbbbear/code).

RESULTS

Identification of DE-mRNAs and the Link Between DE-mRNAs

The clinical information of the entire data, training data, and validation data are presented in **Supplementary Table S1**. Kaplan-Meier curves indicated a significant difference in prognosis between the recurrence and non-recurrence groups (**Figure 1A**). Using the edgeR and DESeq packages of R, 199 and 204 mRNAs that were significantly different between the recurrence and non-recurrence groups were identified, respectively (**Figures 1B, C**). Next, 60 overlapping DE-mRNAs were obtained through the cross between the differentially expressed mRNAs identified by the edgeR package and DESeq package (**Figure 1D**). To unveil the relationships among all DE-mRNAs, we constructed a network diagram of protein interactions by the string database (**Supplementary Figure S2A**). In order of the ability to interact with the others, the top 10 genes were CEA Cell Adhesion Molecule 5 (*CEACAM5*), Mucin 1, Cell Surface Associated (*MUC1*), Cathepsin G (*CTSG*),

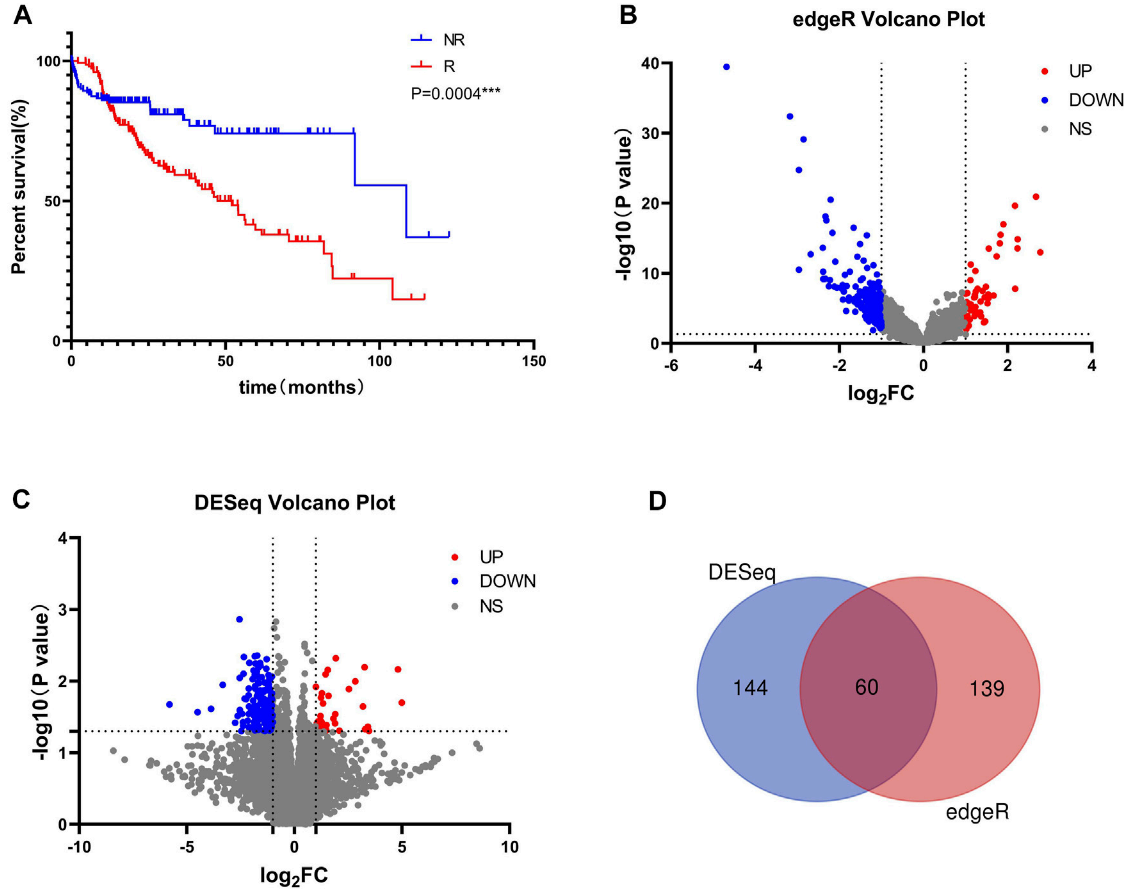


FIGURE 1 | Identification of DE-mRNAs **(A)** Kaplan-Meier curves of OS between the recurrence and non-recurrence groups across the entire dataset **(B)** DE-mRNAs that identified using the edgeR package **(C)** DE-mRNAs that identified using the DESeq package **(D)** Overlapping DE-mRNAs between the selection methods.

Ret Proto-Oncogene (*RET*), CD79a Molecule (*CD79A*), Collagen Type XI Alpha 2 Chain (*COL11A2*), GLI Family Zinc Finger 2 (*GLI2*), Collagen Type X Alpha 1 Chain (*COL10A1*), Myosin Light Chain 3 (*MYL3*), Tectorin Beta (*TECTB*). Interestingly, we found that most of these key node genes play important roles in tumorigenesis (*RET*, *CEACAM5*), immune regulation (*CTSG*, *CD79A*) and the EMT pathway (*COL10A1*, *COL11A2*), suggesting their significant role in the recurrence of liver cancer. To quantify the interaction relationships between DE-mRNAs, we calculated the correlation of DE-mRNAs and found that many genes of the immunoglobulin superfamily were highly positively correlated (**Supplementary Figure S2B**), such as genes of the Immunoglobulin Heavy Variable (IGHV) and Immunoglobulin Kappa Variable (IGKV) regions. The combination of these genes may co-regulate immune function in patients and affect the recurrence of liver cancer after surgery.

Functional Enrichment for mRNAs Co-expressed

To comprehensively study the potential mechanism of liver cancer recurrence, functional enrichment was performed in both groups. The results of the GO enrichment analysis are shown in **Figures 2A, B**. It was noted that the DE-mRNAs between the two groups were significantly enriched in immune-related pathways such as antigen binding, Fc-gamma receptor signaling pathway involved in phagocytosis, regulation of immune response, immune response, immunoglobulin receptor binding, positive regulation of B cell activation, B cell receptor signaling pathway, immunoglobulin complex and circulating, innate immune response, and B cell activation. Moreover, GSEA enrichment results showed that there were significant differences in the B cell receptor signaling pathway, T cell receptor signaling pathway, cell cycle, RNA polymerase, and DNA replication between the recurrence and non-recurrence groups. The B cell receptor signaling pathway and the T cell receptor signaling pathway-related genes were significantly enriched in the non-recurrence group, while the cell cycle, RNA polymerase, and DNA replication pathway-related genes were significantly enriched in the recurrence group (**Figures 2C–G**). These results suggested that the changes of immune-related functions may be the mechanism influencing liver cancer recurrence.

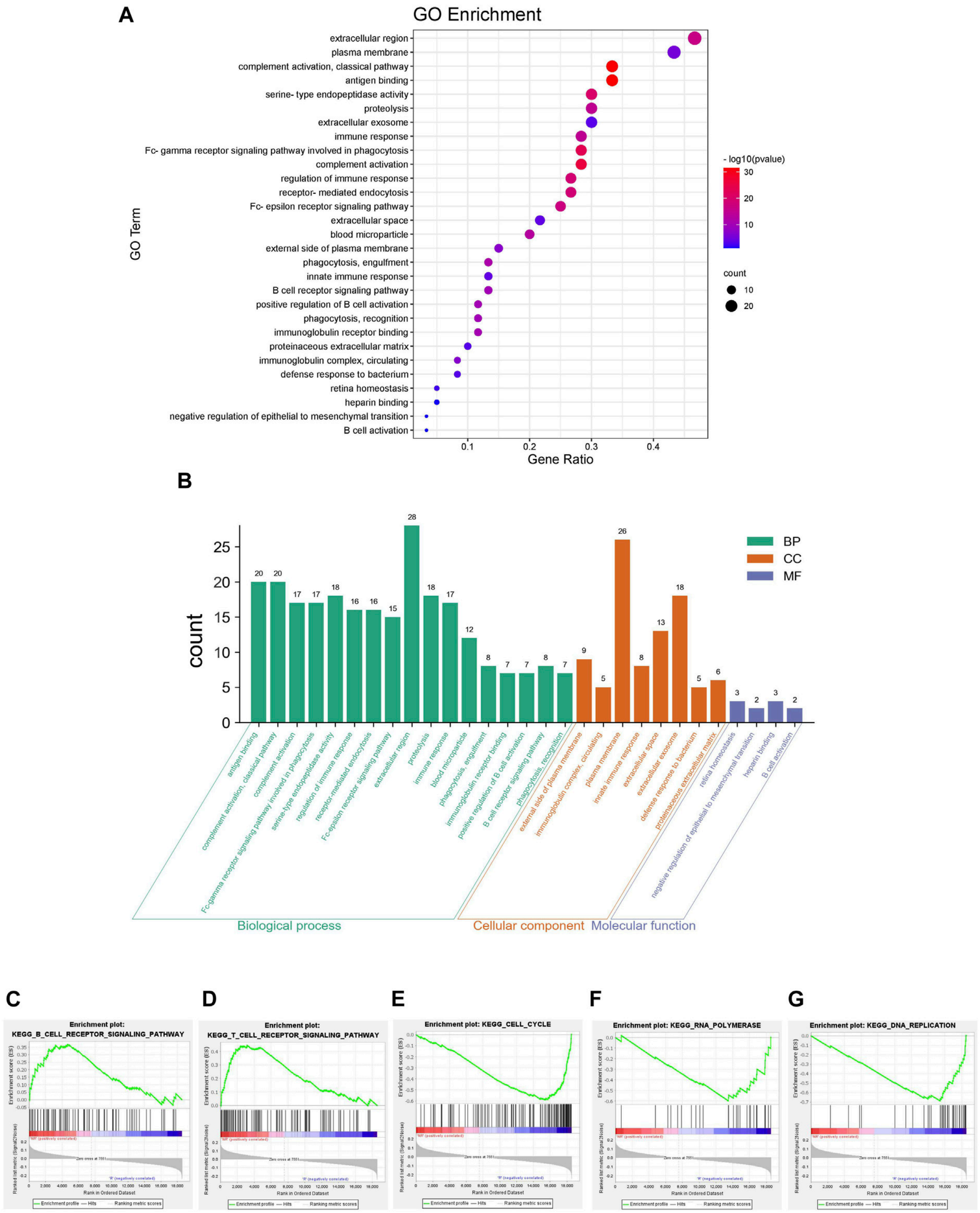


FIGURE 2 | Functional enrichment for mRNAs co-expressed **(A)** The bubble pattern shows the enrichment pathways with Gene Ratio, gene count and *p*-value **(B)** The histogram shows the enrichment of molecular function, biological process and cellular component. Results of GSEA enrichment on **(C)** B cell receptor signaling pathway **(D)** T cell receptor signaling pathway **(E)** Cell cycle **(F)** RNA polymerase, and **(G)** DNA replication.

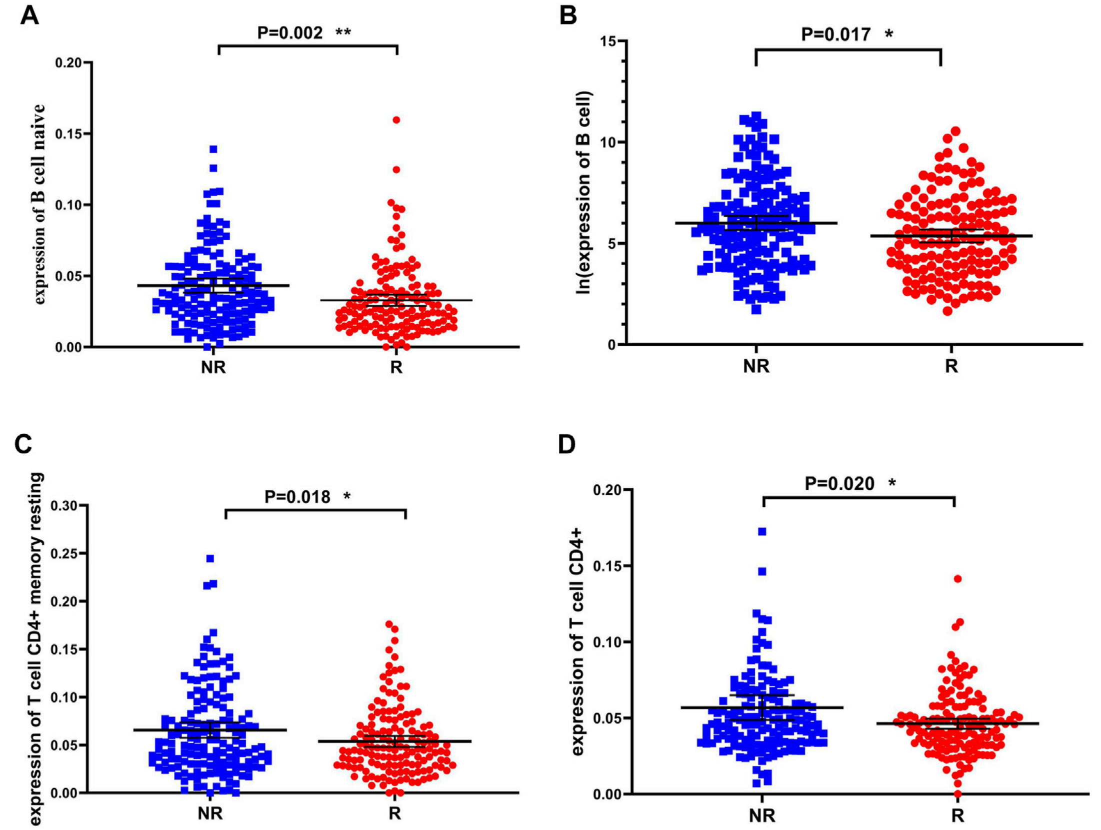


FIGURE 3 | Immune cells that are significantly different between the recurrence and non-recurrence groups **(A)** B cell naive expression between the two groups **(B)** B cell expression between the two groups **(C)** T cell $CD4^+$ memory resting expression between the two groups **(D)** T cell $CD4^+$ expression between the two groups.*$p <$ 0.05, **$p < 0.01$ and ***$p < 0.001$.

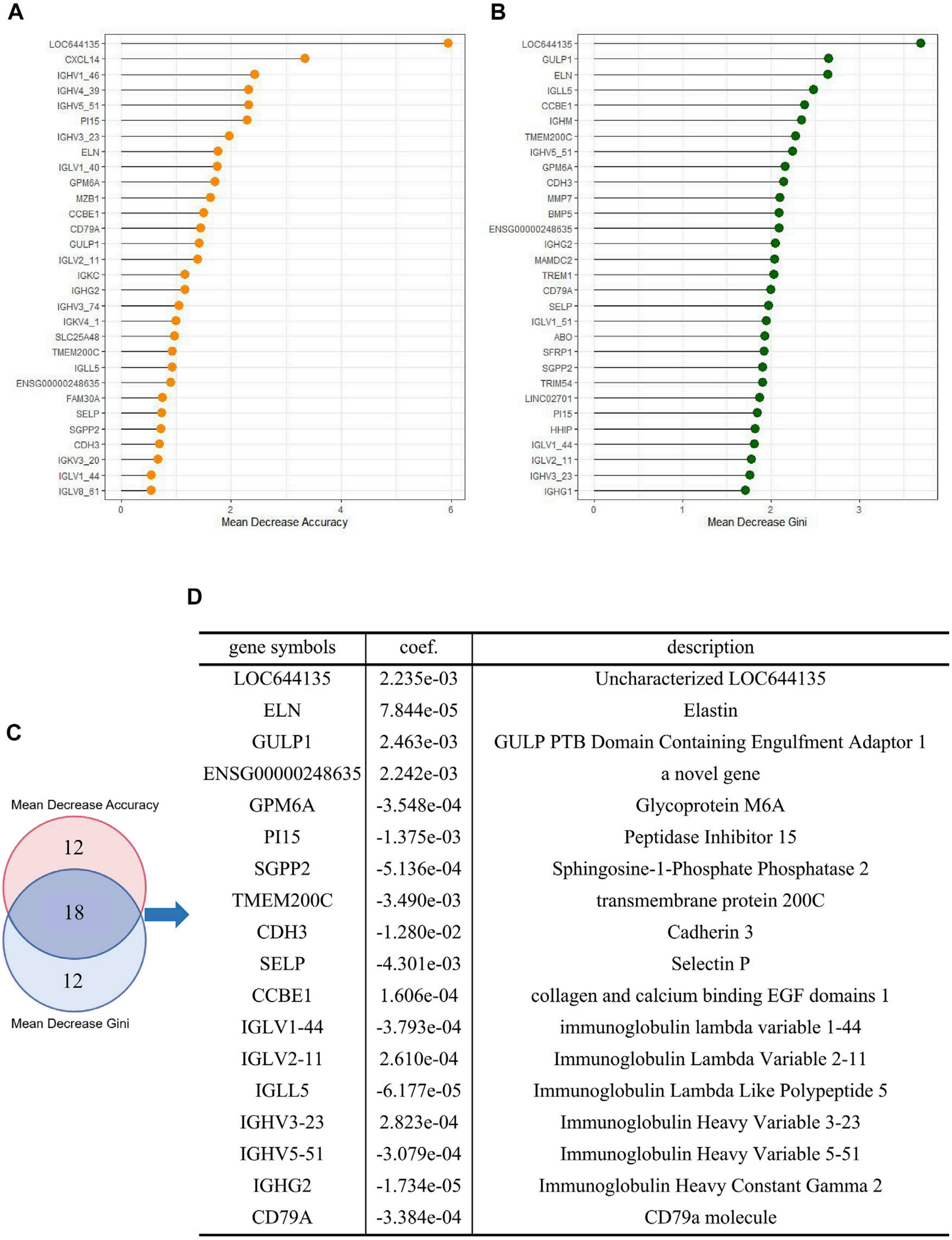
Identification of Immune Cells With Significant Difference Between Two Groups

The results of functional enrichment suggested that there were significant differences in immune-related pathways between the recurrence and non-recurrence groups. In order to investigate which immune cells are involved in the mechanisms influencing the recurrence of liver cancer, we used the TIMER database (http://timer.cistrome.org) to estimate the expression level of immune cells in the two groups, and the *t*-test was used to determine whether there were significant differences in the expression of immune cells between the two groups. Immune cells showed significant differences between the recurrence and non-recurrence groups (**Figures 3A–D**), and the expression of naive B cells, B cells, T cell $CD4^+$ memory resting, and T cell $CD4^+$ were downregulated in the recurrence group. The results demonstrated the significant influence of these four immune cells on liver cancer recurrence.

Construction and Performance Evaluation of the Risk Assessment Model

To identify the mRNAs that play an important role in liver cancer recurrence, we constructed 500 decision trees using random forest with 60 DE-mRNAs as features and measured the importance of each DE-mRNA on recurrence by calculating the mean decrease Gini and mean decrease accuracy for each DE-mRNA. The top 30 DE-mRNAs ranked by mean decrease Gini and mean decrease accuracy were selected (**Figures 4A, B**). The top 30 mRNAs were intersected, and 18 overlapping mRNAs were finally selected to construct the risk assessment model (**Figures 4C, D**).

Subsequently, we trained an 18-mRNA risk assessment model in the training cohort using logistic regression analysis. These 18 mRNAs were uncharacterized LOC644135 (*LOC644135),* elastin (*ELN),* GULP PTB domain-containing engulfment adaptor 1 (*GULP1*), *ENSG00*000248635 (a novel gene), Glycoprotein M6A (GPM6A), peptidase inhibitor 15 (PI15), Sphingosine-1-Phosphate phosphatase 2 (SGPP2), transmembrane protein 200C (TMEM200C), cadherin 3 (CDH3), selectin P (SELP), collagen and calcium binding EGF domains 1 (CCBE1), immunoglobulin lambda variable 1–44 (IGLV1-44), Immunoglobulin Lambda Variable 2–11 (IGLV2-11), Immunoglobulin Lambda Like Polypeptide 5 (IGLL5), Immunoglobulin Heavy Variable 3–23 (IGHV3-23), Immunoglobulin Heavy Variable 5–51 (IGHV5-51), Immunoglobulin Heavy Constant Gamma 2 (IGHG2), and CD79a molecule (CD79A). The risk assessment model was developed based on the coefficients of mRNAs and the constant derived from this analysis. The value of the constant



gene symbols	coef.	description
LOC644135	2.235e-03	Uncharacterized LOC644135
ELN	7.844e-05	Elastin
GULP1	2.463e-03	GULP PTB Domain Containing Engulfment Adaptor 1
ENSG00000248635	2.242e-03	a novel gene
GPM6A	-3.548e-04	Glycoprotein M6A
PI15	-1.375e-03	Peptidase Inhibitor 15
SGPP2	-5.136e-04	Sphingosine-1-Phosphate Phosphatase 2
TMEM200C	-3.490e-03	transmembrane protein 200C
CDH3	-1.280e-02	Cadherin 3
SELP	-4.301e-03	Selectin P
CCBE1	1.606e-04	collagen and calcium binding EGF domains 1
IGLV1-44	-3.793e-04	immunoglobulin lambda variable 1-44
IGLV2-11	2.610e-04	Immunoglobulin Lambda Variable 2-11
IGLL5	-6.177e-05	Immunoglobulin Lambda Like Polypeptide 5
IGHV3-23	2.823e-04	Immunoglobulin Heavy Variable 3-23
IGHV5-51	-3.079e-04	Immunoglobulin Heavy Variable 5-51
IGHG2	-1.734e-05	Immunoglobulin Heavy Constant Gamma 2
CD79A	-3.384e-04	CD79a molecule

FIGURE 4 | Construction of a risk assessment model **(A)** Top 30 DE-mRNAs with mean decrease Gini **(B)** Top 30 DE-mRNAs with mean decrease accuracy **(C)** Overlapping DE-mRNAs between the selection methods **(D)** Logistic risk model constructed by overlapping DE-mRNAs.

b in the logistic regression formula was 0.3883, and the coefficients of mRNAs were shown in **Figure 4D**.

In the training cohort, the performance of the risk assessment model was good, with an AUC value of 0.7356 (**Figures 5A, B**). Subsequently, we assessed the robustness and accuracy of this 18-mRNA signature by applying the same statistical model to the validation cohort. In the validation cohort, the 18-mRNA biomarkers also showed significant diagnostic accuracy in

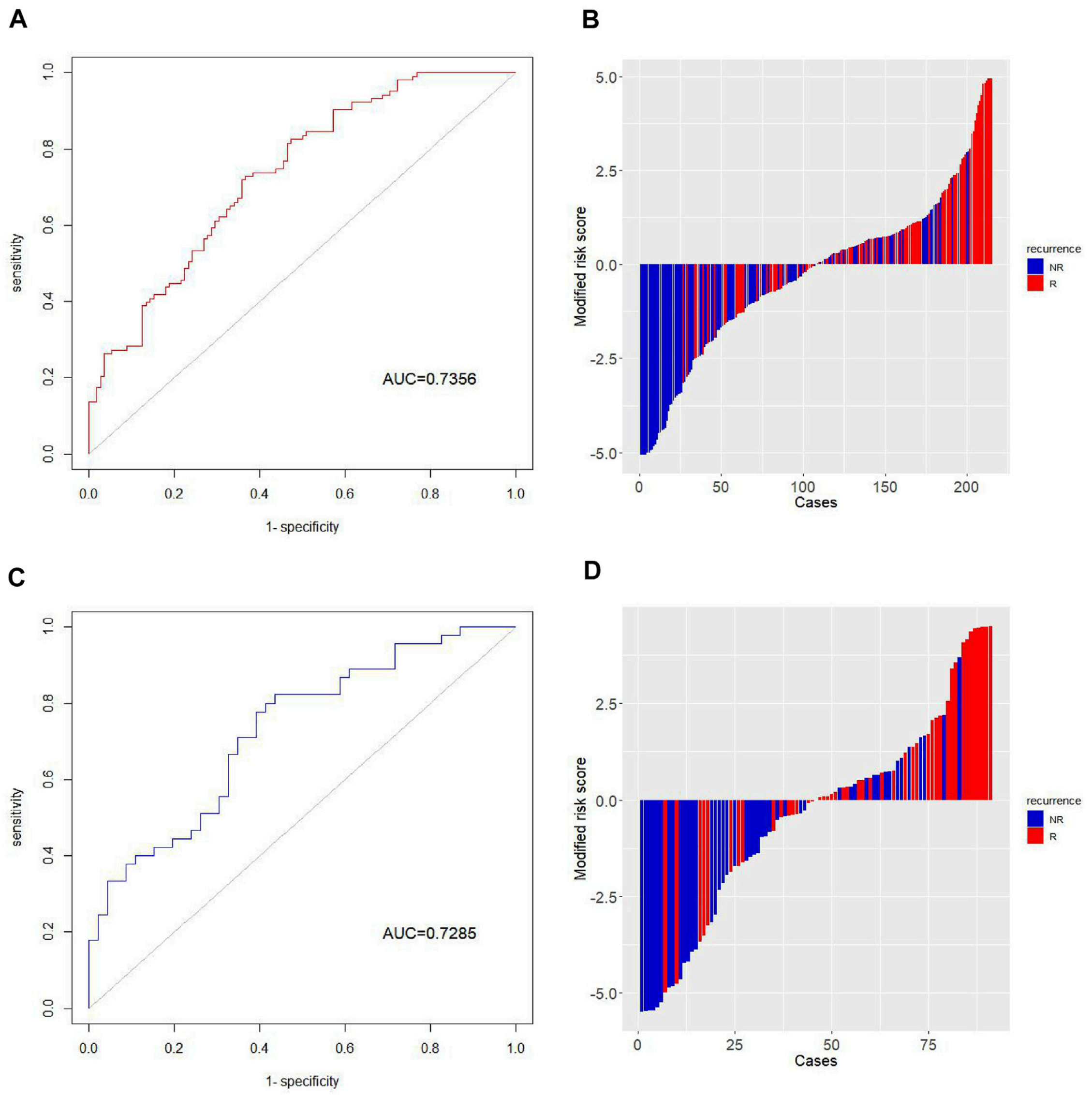


FIGURE 5 | Performance evaluation of the risk assessment model.The ROC curves demonstrate the diagnostic performance of the model in distinguishing recurrent patients in the **(A)** training cohort and **(C)** validation cohort. The histograms show the risk score distribution in the **(B)** training and **(D)** validation cohorts. For convenience of display, the risk score is subtracted from the median and magnified 10 times to obtain the modified risk score.

identifying postoperative recurrence in patients with liver cancer, with an AUC value of 0.7285 (**Figures 5C, D**).

Liver cancer is multi-centric and is significantly affected by background diseases. Different disease backgrounds may have influenced the results of the model. Therefore, we tried to verify whether the 18-mRNA signature can show good predictive performance in different etiological sources of liver cancer subgroups. Restricted by limited etiological data on patients, we could only divide the samples into ALD (n = 57) and HVI (n = 98) subgroups. Within the ALD subgroup, we randomly re-divided the training and validation cohorts in a 7:3 ratio and reconstructed a logistic model with the 18 mRNAs based on the training cohort to predict recurrence in patients in the ALD group and validated them in the validation cohort. The same procedure was performed for the HVI subgroup. These were executed to reduce the effect of background disease on the prediction results of the 18 mRNAs. In both subgroups, the 18-mRNA logistic regression models exhibited good and similar predictive performances. In the ALD subgroup, the AUC values of the 18-mRNA model were 0.8107 and 0.7273 on the training and validation sets, respectively. In the HVI subgroup, the AUC values were 0.8795 and 0.7764 for the training and validation sets, respectively (**Supplementary Figure S3A–D**). In summary, the 18-mRNA signature performed well in predicting liver cancer of different etiological sources (viral infection and alcohol-related), suggesting that our predictive markers may be universally applicable for predicting the recurrence of liver cancer from different etiological sources.

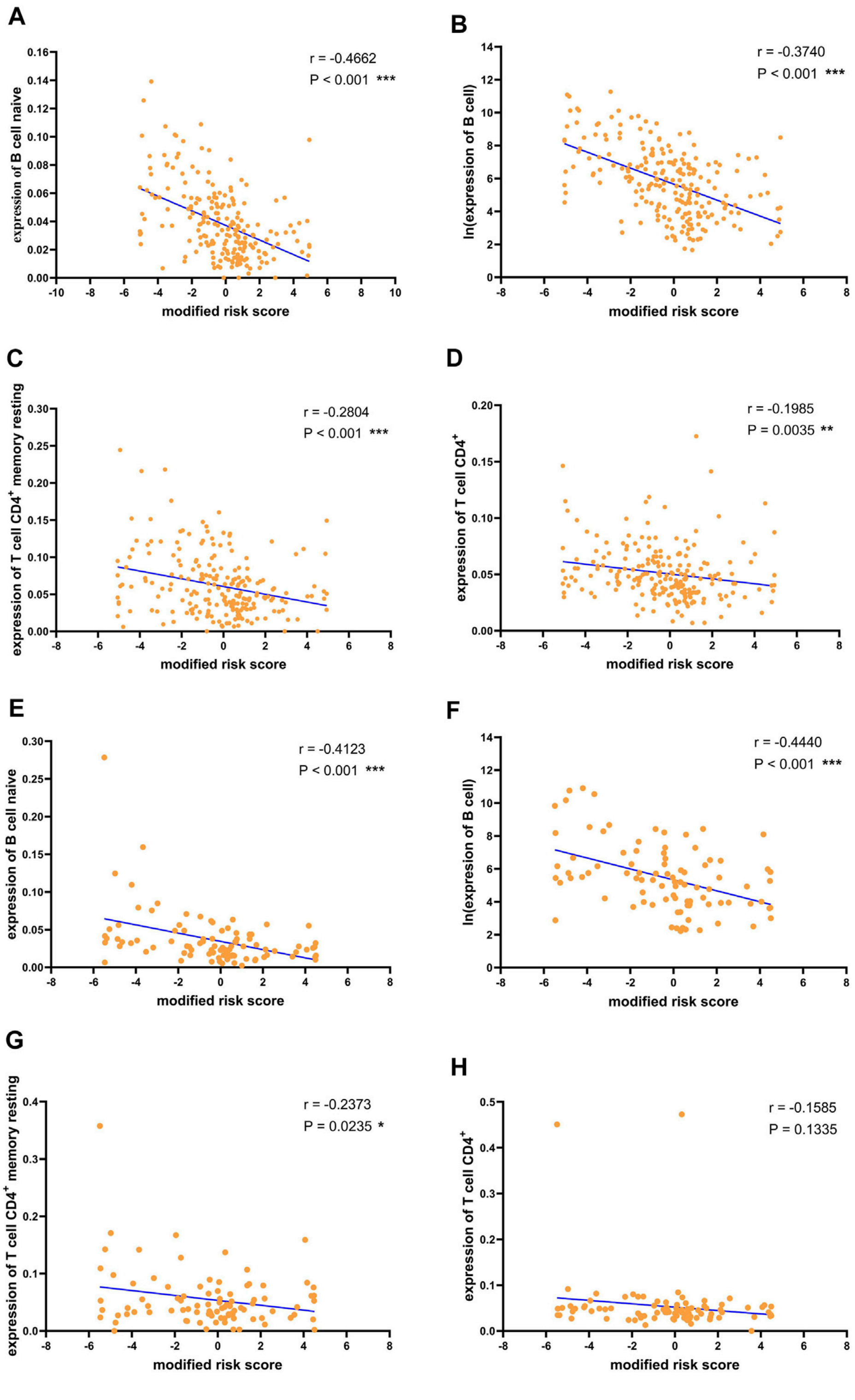


FIGURE 6 | Correlation between risk scores and immune cells.Correlation between risk score and **(A)** B cell naive **(B)** B cell **(C)** T cell CD4$^+$ memory resting **(D)** T cell CD4$^+$ in the training data. Correlation between risk score and **(E)** B cell naive **(F)** B cell **(G)** T cell CD4$^+$ memory resting **(H)** T cell CD4$^+$ in the validation data. For convenience of display, the risk score is subtracted from the median and magnified 10 times to obtain the modified risk score. $^*p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$.

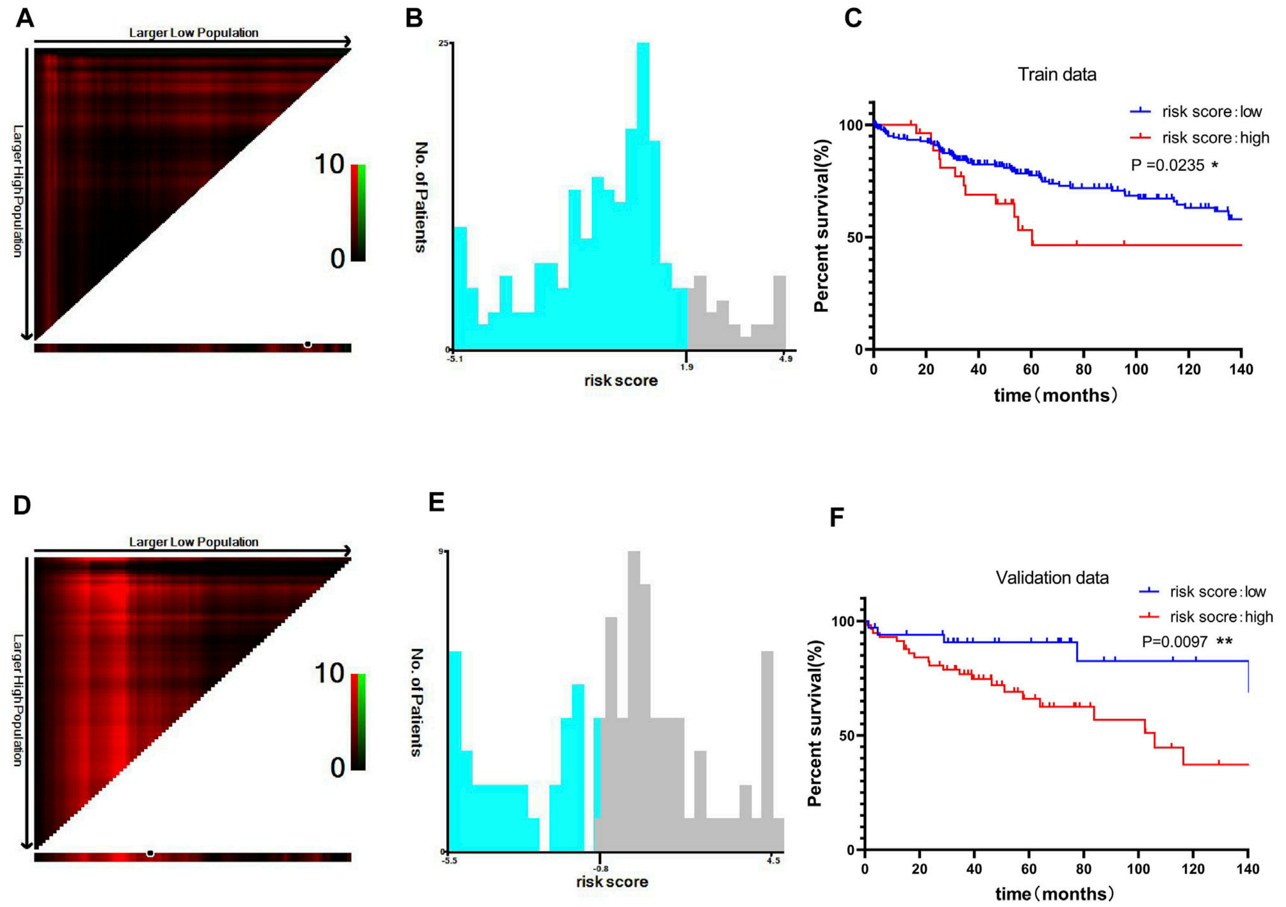


FIGURE 7 | Prognostic analysis using the risk scores in the training and validation sets. Optimal cutoff value of the risk score divided by X-tile software in the training data **(A, B) (C)** Kaplan-Meier curve of the two groups divided by the cutoff value in the training data. The optimal cutoff value of the risk score was divided by the X-tile software in the validation data **(D, E) (F)** Kaplan-Meier curve of the two groups divided by the cutoff value in the validation data. *$p < 0.05$, **$p < 0.01$ and ***$p < 0.001$.

In order to form a comparison with the logistic regression model constructed by the above method, we used two other variable screening methods and constructed three different logistic models: a logistic regression model without variable filtering (model 2), a logistic regression model with variable filtering by stepwise regression (model 3), and a logistic regression model with variable filtering by L1 regularization (model 4). The prediction results of these three models are good in the training set, but the results in the validation set are much worse than those in the training set (**Supplementary Figure S4A–F**), that is, the models constructed by the three methods show serious overfitting. In contrast, the logistic model constructed after filtering with the random forest algorithm showed similar results and good predictions for both the training and validation cohorts.

We also compared the risk score with the StromalScore, ImmuneScore, and ESTIMATEScore. The StromalScore, ImmuneScore, and ESTIMATEScore were calculated based on the expressions of mRNAs. In the training set, the AUC values of the StromalScore, ImmuneScore, and ESTIMATEScore for predicting recurrence of liver cancer were 0.5581, 0.5599, and 0.5643, respectively, and 0.5923, 0.5942 and 0.5948 in the validation set, respectively (**Supplementary Figure S5A–F**). This indicated that the risk score proposed in our study had better performance in predicting the recurrence of liver cancer compared to StromalScore, ImmuneScore, and ESTIMATEScore.

Correlation Between Risk Score and Immune Cells

The previous analysis results showed that the recurrence of liver cancer after surgery was significantly associated with the expression of immune cells. Furthermore, to confirm the relationship between the risk of recurrence and immune infiltration, we performed correlation analysis between the recurrence risk score estimated using the previously established risk assessment model and immune cells. In the training cohort, B cell naive, B cell, T cell CD4^{+} memory resting, and T cell CD4^{+} expression levels were significantly negatively correlated with the risk score (**Figures 6A–D**). The negative association between these immune cells and the risk score was also verified in the validation cohort (**Figures 6E–H**). The results demonstrated the validity of the risk score predicted by the 18-mRNA model and further confirmed the negative association between these four immune cells and postoperative recurrence of liver cancer.

Prognostic Analysis Using the Risk Scores in the Training and Validation Sets

To improve the clinical prognostic significance of the risk score, we then grouped patients by risk score and performed Kaplan-Meier survival analysis. X-tile software was used to obtain the best truncation value, and patients in the training cohort were divided into low-risk and high-risk groups (**Figures 7A, B**). We observed a significant difference in overall survival (OS) between the low-risk and high-risk groups in the training cohort (p = 0.0235) (**Figure 7C**). Similarly, in the validation cohort, there was a significant difference in the prognosis between the high- and low-risk groups classified by the X-tile software (p = 0.0097) (**Figures 7D–F**). The results indicated that risk score had good prognostic value for liver cancer patients.

Univariate and Multivariate Analyses of the mRNA Signature Prognostic Abilities

To verify the prognostic value of the 18-mRNA signature independently from the clinicopathological characteristics, we performed Cox univariate and multivariate analyses that included 18-mRNA risk score, age, sex, histologic grade, and TNM stage as co-variables in the training and validation cohorts. In univariate and multivariate analyses of the training cohort, TMN stage was found to be related to OS (**Supplementary Figure S6A–B**). However, in univariate and multivariate analyses of the validation cohort, the TMN stage was no longer significantly correlated with OS, and the risk score was a significant variable related to OS (**Supplementary Figure S6C–D**). HR values of the risk score were significant but not high, which was not perfect in clinical prognosis prediction. In addition, histologic grade and TMN stage appeared to have high but not significant HR values, which may be due to excessive standard errors of the variables. (Katz and Hauck, 1993).

DISCUSSION

In this study, we constructed a risk scoring model with 18 mRNAs to predict post-hepatectomy recurrence of liver cancer. The 18 mRNAs were LOC644135, ELN, GULP1, ENSG00000248635, GPM6A, PI15, SGPP2, TMEM200C, CDH3, SELP, CCBE1, IGLV1-44, IGLV2-11, IGLL5, IGHV3-23, IGHV5-51, IGHG2, and CD79A.The risk assessment model could accurately distinguish between low- and high-risk samples for the recurrence of liver cancer after resection, with good prognostic performance. The identified 18 mRNAs also had good predictive performance in liver cancer samples caused by different etiologies. B cell naive, B cell, T cell CD4$^+$ memory resting, and T cell CD4$^+$ were significantly different in recurrence versus non-recurrence liver cancer samples and were found to be negatively correlated with the risk scores predicted by the constructed model by correlation analysis.

Compared to the classification algorithms in machine learning, regression analysis has better explanatory power for variables. The advantage of our method is that it combines random forest with regression analysis to screen biomarkers of liver cancer recurrence and construct a formulaic risk assessment model of liver cancer recurrence, which ensures the interpretability of variables with good predictive performance. We also constructed a logistic regression model without feature screening, a stepwise regression logistic model, and an L1 regularized logistic model, which revealed that the random forest-based logistic regression model had better generalization performance on the validation set. Meanwhile, compared to StromalScore, ImmuneScore, and ESTIMATEScore, our risk score showed better performance to predict recurrence risk of liver cancer.

In previous studies, machine learning methods have also been used to predict liver cancer recurrence. Wang et al.(Wang et al., 2020) used lasso and Cox regressions to screen five mRNAs to predict HCC recurrence, and the predicted AUC values for 1-year, 2-year, and 3-year RFS rates from the independent validation data were 0.752, 0.651, and 0.677, respectively. Iizuka et al.(Iizuka et al., 2003) used Fisher's linear classifier algorithm to predict intrahepatic recurrence in hepatocellular carcinoma patients within 1 year after resection, using 18 mRNAs. In the validation sample, the predictive accuracy was 92.6% (25/27). Based on this, Somura et al.(Somura et al., 2008) selected three of the mRNAs and constructed a prediction model with the same algorithm, with a correct prediction accuracy of 81.4% (35/43) in the validation set. However, selection bias or publication bias in the small sample sizes may produce inflated over-promising results. (Ntzani and Ioannidis, 2003). In addition, a few previous studies have used a large number of genetic biomarkers to predict the prognosis of liver cancer. (Lee et al., 2006; Hoshida et al., 2008; Woo et al., 2008). However, due to the different sample selection criteria and definition of the outcomes, there was little gene overlap between these studies. To our knowledge, the present study includes the largest sample size through the machine learning method compared to other studies in HCC recurrence prediction, which may enhance the validity of the gene signature predicted by our study. Our screened 18 mRNAs showed good performance in predicting liver cancer recurrence, with AUCs of 0.7356 and 0.7285 in the training and validation cohorts, respectively.

We report for the first time that LOC644135, ELN, GULP1, ENSG00000248635, GPM6A, and PI15 are associated with the risk of liver cancer recurrence. In addition, some genes that we identified have been reported to play a role in the development, recurrence, and metastasis of cancer in previous studies. It has been reported that Nudix hydrolase 21 promotes tumor growth and metastasis through modulating SGPP2 in gastric cancer. (Zhu et al., 2021). TMEM200C is reported to be hypomethylated, and candidate oncogenes linked to early metastasis in uveal melanoma. (Ness et al., 2021). CDH3, a classical cell adhesion molecule, has been reported to be related to a variety of human cancers. L. Li et al. (Li et al., 2019) found that Kruppel-like factor 4 (KLF4)-mediated upregulation of CDH3 inhibits the growth and migration of human hepatoma cells through GSK-3 β signaling. Overexpression of CDH3 has been reported to promote the movement of pancreatic cancer cells by interacting with P120ctn and activating Rho family GTPases. (Taniuchi et al., 2005). CDH3 has also been reported to be associated with prognosis in patients with colorectal adenocarcinoma and lung cancer. (Xu et al., 2019; Hsiao et al., 2020). SELP is a member of the selectin family of cell adhesion molecules, and it mediates heterotypic aggregation of activated platelets to cancer cells and adhesion of cancer cells to stimulated endothelial cells. (Chen and Geng, 2006). It has been demonstrated that SELP plays an important role in the growth and metastasis of human colon carcinoma *in vivo*. (Kim et al., 1998). CCBE1 is essential in lymphatic vascular development. Song et al.(Song et al., 2020) demonstrated the procarcinogenic role of CCBE1 in promoting lymphangiogenesis and metastasis in colorectal cancer, and it has been reported that targeting of CCBE1 by miR-330-3p promotes tumor metastasis in breast cancer. (Mesci et al., 2017). Our results contribute to further understanding of the impact of these genes on recurrence of liver cancer. It is interesting to note that many of the genes we report are immunoglobulin genes (IGLV1-44, IGLV2-11, IGLL5, IGHV3-23, IGHV5-51, IGHG2). Previous studies have shown that some immunoglobulins regulate the tumor microenvironment and influence the prognosis of patients with breast cancer, diffuse large B-cell lymphoma, and chronic lymphocytic leukemia (Baliakas et al., 2015; Zhang et al., 2015; Stamatopoulos et al., 2018; Xu-Monette et al., 2019), and our results suggest that the above immunoglobulin genes may play an important role in recurrence of liver cancer. CD79A is also an immune function-related gene. CD79A and CD79b molecule (CD79B) heterodimers are important signaling components of the B cell receptor (BCR) complex, which plays a crucial role in B cell development and antibody production. (Li et al., 2017b). These immune-related genes identified in our study may provide insights into the underlying mechanism of liver cancer recurrence.

Previous studies have reported that immune cells play a role in cancer recurrence. (Bindea et al., 2013; Chevalier et al., 2017; Zhou et al., 2019). The enrichment results showed that DE-mRNAs were enriched in many immune-related pathways. Therefore, on the basis of predicting liver cancer recurrence, we analyzed the differences in immune cells between patients with and without recurrence, and explored the relationship between the predicted risk of recurrence and differential immune cells in patients with liver cancer. We demonstrated that compared with the non-recurrence group, B cell naive, B cell, T cell CD4$^+$ memory resting, and T cell CD4$^+$ were significantly downregulated in the recurrence group and were inversely associated with the recurrence risk evaluated by the model we built. Previous studies have shown that tumor metastasis is associated with the presence of CD4$^+$ T cells and B cells. Olkhanud et al. (Olkhanud et al., 2011) found that tumor-evoked regulatory B cells promote breast cancer metastasis by converting resting CD4$^+$ T cells to T-regulatory cells. Ou et al.(Ou et al., 2015) found that tumor microenvironment B cells increase bladder cancer metastasis via modulation of the IL-8/androgen receptor (AR)/MMPs signals. In addition, Guy et al. (Guy et al., 2016) found that tumor-specific CD4$^+$ T cells and B cells play important functions and major roles in anticancer immunity. Many studies have shown that a variety of immunosuppressive signals regulate the tumor microenvironment, which plays an important regulatory role in the process of tumorigenesis, and its heterogeneity can lead to multiple aspects, including patient prognosis and treatment response. (Fridman et al., 2012; Galon et al., 2014; Hui and Chen, 2015). Sun et al. (Sun et al., 2021) reported that recurrent HCC has a unique immune ecosystem compared with the primary HCC tumor microenvironment. Therefore, our findings may provide a new approach for the immunotherapy of liver cancer recurrence.

In conclusion, the risk assessment model developed in this study may serve as a complementary tool to provide useful information for predicting disease outcomes after hepatectomy in patients with liver cancer and guide adjuvant therapy. Meanwhile, the immune cells reported in this study could be the targets of immunotherapy for patients with liver cancer. While these predictions are valuable, the current study has some limitations. The 18-mRNA biomarkers were screened by bioinformatics and machine learning methods, and their expression and specific functions need to be validated by biological experiments. In addition, the predictive performance of the 18-mRNA risk assessment model constructed in our study needs to be validated using large independent samples.

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.

AUTHOR CONTRIBUTIONS

XQ and HZ designed the project. ZH, LZ and JW provided administrative, technical, and material support. XQ, KX, and ZC performed statistical analyses. XQ wrote the manuscript. HZ revised the paper. All authors reviewed the manuscript and agreed to its submission.

FUNDING

This study was supported by the Zhejiang Province Key Research and Development Projects (2020C03057, 2021C03145, 2020C03071) and the National Natural Science Foundation of China (Grant no. 61972358).

ACKNOWLEDGMENTS

We sincerely thank the researchers and study participants for their contributions toward this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.733654/full#supplementary-material

REFERENCES

Agostini, M., Zangrando, A., Pastrello, C., D'Angelo, E., Romano, G., Giovannoni, R., et al. (2015). A Functional Biological Network Centered on XRCC3: A New Possible Marker of Chemoradiotherapy Resistance in Rectal Cancer Patients. *Cancer Biol. Ther.* 16, 1160–1171. doi:10.1080/15384047.2015.1046652

Baliakas, P., Agathangelidis, A., Hadzidimitriou, A., Sutton, L.-A., Minga, E., Tsanousa, A., et al. (2015). Not all IGHV3-21 Chronic Lymphocytic Leukemias Are Equal: Prognostic Considerations. *Blood* 125, 856–859. doi:10.1182/blood-2014-09-600874

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39, 782–795. doi:10.1016/j.immuni.2013.10.003

Bradley, A. P. (1997). The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition* 30, 1145–1159. doi:10.1016/s0031-3203(96)00142-2

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer J. Clinicians* 68, 394–424. doi:10.3322/caac.21492

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chen, J., Ying, H., Liu, X., Gu, J., Feng, R., Chen, T., et al. (2020). A Transfer Learning Based Super-resolution Microscopy for Biopsy Slice Images: The Joint Methods Perspective. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18, 1. doi:10.1109/TCBB.2020.2991173

Chen, M., and Geng, J.-G. (2006). P-selectin Mediates Adhesion of Leukocytes, Platelets, and Cancer Cells in Inflammation, Thrombosis, and Cancer Growth and Metastasis. *Arch. Immunol. Ther. Exp.* 54, 75–84. doi:10.1007/s00005-006-0010-6

Chen, T., Liu, X., Feng, R., Wang, W., Yuan, C., Lu, W., et al. (2021). Discriminative Cervical Lesion Detection in Colposcopic Images with Global Class Activation and Local Bin Excitation. *IEEE J. Biomed. Health Inform.* [in press]. doi:10.1109/JBHI.2021.3100367

Chevalier, M. F., Trabanelli, S., Racle, J., Salomé, B., Cesson, V., Gharbi, D., et al. (2017). ILC2-modulated T Cell-To-MDSC Balance Is Associated with Bladder Cancer Recurrence. *J. Clin. Invest.* 127, 2916–2929. doi:10.1172/JCI89717

Díaz-Uriarte, R., and de Andrés, S. A. (2006). Gene Selection and Classification of Microarray Data Using Random forest. *BMC Bioinformatics* 7, 3. doi:10.1186/1471-2105-7-3

Feng, R., Liu, X., Chen, J., Chen, D. Z., Gao, H., and Wu, J. (2021). A Deep Learning Approach for Colonoscopy Pathology WSI Analysis: Accurate Segmentation and Classification. *IEEE J. Biomed. Health Inform.* 25, 3700–3708. doi:10.1109/JBHI.2020.3040269

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The Immune Contexture in Human Tumours: Impact on Clinical Outcome. *Nat. Rev. Cancer* 12, 298–306. doi:10.1038/nrc3245

Galon, J., Mlecnik, B., Bindea, G., Angell, H. K., Berger, A., Lagorce, C., et al. (2014). Towards the Introduction of the 'Immunoscore' in the Classification of Malignant Tumours. *J. Pathol.* 232, 199–209. doi:10.1002/path.4287

Guy, T. V., Terry, A. M., Bolton, H. A., Hancock, D. G., Zhu, E., Brink, R., et al. (2016). Collaboration between Tumor-specific CD4+ T Cells and B Cells in Anti-cancer Immunity. *Oncotarget* 7, 30211–30229. doi:10.18632/oncotarget.8797

Ho, M.-C., Lin, J.-J., Chen, C.-N., Chen, C.-C., Lee, H., Yang, C.-Y., et al. (2006). A Gene Expression Profile for Vascular Invasion Can Predict the Recurrence after Resection of Hepatocellular Carcinoma: A Microarray Approach. *Ann. Surg. Oncol.* 13, 1474–1484. doi:10.1245/s10434-006-9057-1

Hoshida, Y., Villanueva, A., Kobayashi, M., Peix, J., Chiang, D. Y., Camargo, A., et al. (2008). Gene Expression in Fixed Tissues and Outcome in Hepatocellular Carcinoma. *N. Engl. J. Med.* 359, 1995–2004. doi:10.1056/NEJMoa0804525

Hsiao, T.-F., Wang, C.-L., Wu, Y.-C., Feng, H.-P., Chiu, Y.-C., Lin, H.-Y., et al. (2020). Integrative Omics Analysis Reveals Soluble Cadherin-3 as a Survival Predictor and an Early Monitoring Marker of EGFR Tyrosine Kinase Inhibitor Therapy in Lung Cancer. *Clin. Cancer Res.* 26–3229. doi:10.1158/1078-0432.CCR-19-3972

Hu, T., Wang, S., Huang, L., Wang, J., Shi, D., Li, Y., et al. (2019). A Clinical-Radiomics Nomogram for the Preoperative Prediction of Lung Metastasis in Colorectal Cancer Patients with Indeterminate Pulmonary Nodules. *Eur. Radiol.* 29, 439–449. doi:10.1007/s00330-018-5539-3

Hui, L., and Chen, Y. (2015). Tumor Microenvironment: Sanctuary of the Devil. *Cancer Lett.* 368, 7–13. doi:10.1016/j.canlet.2015.07.039

Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., et al. (2003). Oligonucleotide Microarray for Prediction of Early Intrahepatic Recurrence of Hepatocellular Carcinoma after Curative Resection. *The Lancet* 361, 923–929. doi:10.1016/S0140-6736(03)12775-4

Katz, M. H., and Hauck, W. W. (1993). Proportional Hazards (Cox) Regression. *J. Gen. Intern. Med.* 8, 702–711. doi:10.1007/bf02598295

Kim, Y. J., Borsig, L., Varki, N. M., and Varki, A. (1998). P-selectin Deficiency Attenuates Tumor Growth and Metastasis. *Pnas* 95, 9325–9330. doi:10.1073/pnas.95.16.9325

Kudo, S.-e., Ichimasa, K., Villard, B., Mori, Y., Misawa, M., Saito, S., et al. (2021). Artificial Intelligence System to Determine Risk of T1 Colorectal Cancer Metastasis to Lymph Node. *Gastroenterology* 160, 1075–1084. doi:10.1053/j.gastro.2020.09.027

Le, N. Q. K., Do, D. T., Hung, T. N. K., Lam, L. H. T., Huynh, T.-T., and Nguyen, N. T. K. (2020). A Computational Framework Based on Ensemble Deep Neural Networks for Essential Genes Identification. *Ijms* 21, 9070. doi:10.3390/ijms21239070

Le, N. Q. K., Hung, T. N. K., Do, D. T., Lam, L. H. T., Dang, L. H., and Huynh, T.-T. (2021). Radiomics-based Machine Learning Model for Efficiently Classifying Transcriptome Subtypes in Glioblastoma Patients from MRI. *Comput. Biol. Med.* 132, 104320. doi:10.1016/j.compbiomed.2021.104320

Lee, J.-S., Heo, J., Libbrecht, L., Chu, I.-S., Kaposi-Novak, P., Calvisi, D. F., et al. (2006). A Novel Prognostic Subtype of Human Hepatocellular Carcinoma Derived from Hepatic Progenitor Cells. *Nat. Med.* 12, 410–416. doi:10.1038/nm1377

Li, H., Li, Y., Zhang, X., Wang, Y., Zhang, W., Wu, X., et al. (2017). Molecular Characterization of the CD79a and CD79b and its Role against Aeromonas Hydrophila Infection in Chinese Sucker (*Myxocyprinus asiaticus*). *Fish. Physiol. Biochem.* 43, 1571–1585. doi:10.1007/s10695-017-0394-8

Li, L., Yu, S., Wu, Q., Dou, N., Li, Y., and Gao, Y. (2019). KLF4-Mediated CDH3 Upregulation Suppresses Human Hepatoma Cell Growth and Migration via GSK-3β Signaling. *Int. J. Biol. Sci.* 15, 953–961. doi:10.7150/ijbs.30857

Li, S., Jiang, H., and Pang, W. (2017). Joint Multiple Fully Connected Convolutional Neural Network with Extreme Learning Machine for Hepatocellular Carcinoma Nuclei Grading. *Comput. Biol. Med.* 84, 156–167. doi:10.1016/j.compbiomed.2017.03.017

Liang, J.-D., Ping, X.-O., Tseng, Y.-J., Huang, G.-T., Lai, F., and Yang, P.-M. (2014). Recurrence Predictive Models for Patients with Hepatocellular Carcinoma after Radiofrequency Ablation Using Support Vector Machines with Feature Selection Methods. *Comp. Methods Programs Biomed.* 117, 425–434. doi:10.1016/j.cmpb.2014.09.001

Mesci, A., Huang, X., Taeb, S., Jahangiri, S., Kim, Y., Fokas, E., et al. (2017). Targeting of CCBE1 by miR-330-3p in Human Breast Cancer Promotes Metastasis. *Br. J. Cancer* 116, 1350–1357. doi:10.1038/bjc.2017.105

Nakayama, H., and Takayama, T. (2014). Role of Surgical Resection for Hepatocellular Carcinoma Based on Japanese Clinical Guidelines for Hepatocellular Carcinoma. *Wjh* 7, 261–269. doi:10.4254/wjh.v7.i2.261

Ness, C., Katta, K., Garred, Ø., Kumar, T., Olstad, O. K., Petrovski, G., et al. (2021). Integrated Differential DNA Methylation and Gene Expression of Formalin-Fixed Paraffin-Embedded Uveal Melanoma Specimens Identifies Genes Associated with Early Metastasis and Poor Prognosis. *Exp. Eye Res.* 203, 108426. doi:10.1016/j.exer.2020.108426

Nick, T. G., and Campbell, K. M. (2007). Logistic Regression. *Methods Mol. Biol.* 404, 273–301. doi:10.1007/978-1-59745-530-5_14

Ntzani, E. E., and Ioannidis, J. P. (2003). Predictive Ability of DNA Microarrays for Cancer Outcomes and Correlates: An Empirical Assessment. *The Lancet* 362, 1439–1444. doi:10.1016/S0140-6736(03)14686-7

Olkhanud, P. B., Damdinsuren, B., Bodogai, M., Gress, R. E., Sen, R., Wejksza, K., et al. (2011). Tumor-Evoked Regulatory B Cells Promote Breast Cancer Metastasis by Converting Resting CD4+ T Cells to T-Regulatory Cells. *Cancer Res.* 71, 3505–3515. doi:10.1158/0008-5472.CAN-10-4316

Ou, Z., Wang, Y., Liu, L., Li, L., Yeh, S., Qi, L., et al. (2015). Tumor Microenvironment B Cells Increase Bladder Cancer Metastasisviamodulation of the IL-8/androgen Receptor (AR)/MMPs Signals. *Oncotarget* 6, 26065–26078. doi:10.18632/oncotarget.4569

Reif, D. M., Motsinger, A. A., McKinney, B. A., Crowe, J. E., and Moore, J. H. (2006). "Feature Selection Using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types," in IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, Toronto, ON, Canada, September 28-29, 2006, 1–8. doi:10.1109/CIBCB.2006.330987

Sanchez-pinto, L. N., Venable, L. R., Fahrenbach, J., and Churpek, M. M. (2018). Comparison of Variable Selection Methods for Clinical Predictive Modeling. *Int. J. Med. Inform.* 116, 10–17. doi:10.1016/j.ijmedinf.2018.05.006

Somura, H., Iizuka, N., Tamesa, T., Sakamoto, K., Hamaguchi, T., Tsunedomi, R., et al. (2008). A Three-Gene Predictor for Early Intrahepatic Recurrence of Hepatocellular Carcinoma after Curative Hepatectomy. *Oncol. Rep.* 19, 489–495. doi:10.3892/or.19.2.489

Song, J., Chen, W., Cui, X., Huang, Z., Wen, D., Yang, Y., et al. (2020). CCBE1 Promotes Tumor Lymphangiogenesis and Is Negatively Regulated by TGFβ Signaling in Colorectal Cancer. *Theranostics* 10, 2327–2341. doi:10.7150/thno.39740

Stamatopoulos, B., Smith, T., Crompot, E., Pieters, K., Clifford, R., Mraz, M., et al. (2018). The Light Chain IgLV3-21 Defines a New Poor Prognostic Subgroup in Chronic Lymphocytic Leukemia: Results of a Multicenter Study. *Clin. Cancer Res.* 24, 5048–5057. doi:10.1158/1078-0432.CCR-18-0133

Sun, Y., Wu, L., Zhong, Y., Zhou, K., Hou, Y., Wang, Z., et al. (2021). Single-cell Landscape of the Ecosystem in Early-Relapse Hepatocellular Carcinoma. *Cell* 184, 404–421. doi:10.1016/j.cell.2020.11.041

Taniuchi, K., Nakagawa, H., Hosokawa, M., Nakamura, T., Eguchi, H., Ohigashi, H., et al. (2005). Overexpressed P-cadherin/CDH3 Promotes Motility of Pancreatic Cancer Cells by Interacting with P120ctn and Activating Rho-Family GTPases. *Cancer Res.* 65, 3092–3099. doi:10.1158/0008.5472.CAN-04-3646

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Wang, H., Yang, F., and Luo, Z. (2016). An Experimental Study of the Intrinsic Stability of Random forest Variable Importance Measures. *BMC Bioinformatics* 17, 60. doi:10.1186/s12859-016-0900-5

Wang, J., Zhou, J., Liu, J., Wonka, P., and Ye, J. (2014). "A Safe Screening Rule for Sparse Logistic Regression," in Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, December 8 - 13, 2014, 2, 1053–1061.

Wang, Z., Zhang, N., Lv, J., Ma, C., Gu, J., Du, Y., et al. (2020). A Five-Gene Signature for Recurrence Prediction of Hepatocellular Carcinoma Patients. *Biomed. Res. Int.* 2020, 1–13. doi:10.1155/2020/4037639

Woo, H. G., Park, E. S., Cheon, J. H., Kim, J. H., Lee, J.-S., Park, B. J., et al. (2008). Gene Expression-Based Recurrence Prediction of Hepatitis B Virus-Related Human Hepatocellular Carcinoma. *Clin. Cancer Res.* 14, 2056–2064. doi:10.1158/1078-0432.CCR-07-1473

Xu, Y., Zhao, J., Dai, X., Xie, Y., and Dong, M. (2019). High Expression of CDH3 Predicts a Good Prognosis for colon Adenocarcinoma Patients. *Exp. Ther. Med.* 18, 841–847. doi:10.3892/etm.2019.7638

Xu-Monette, Z. Y., Li, J., Xia, Y., Crossley, B., Bremel, R. D., Miao, Y., et al. (2019). Immunoglobulin Somatic Hypermutation Has Clinical Impact in DLBCL and Potential Implications for Immune Checkpoint Blockade and Neoantigen-Based Immunotherapies. *J. Immunotherapy Cancer* 7, 272. doi:10.1186/s40425-019-0730-x

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Yu, J., Deng, Y., Liu, T., Zhou, J., Jia, X., Xiao, T., et al. (2020). Lymph Node Metastasis Prediction of Papillary Thyroid Carcinoma Based on Transfer Learning Radiomics. *Nat. Commun.* 11, 965–975. doi:10.1038/s41467-020-18497-3

Zhang, N., Deng, H., Fan, X., Gonzalez, A., Zhang, S., Brezski, R. J., et al. (2015). Dysfunctional Antibodies in the Tumor Microenvironment Associate with Impaired Anticancer Immunity. *Clin. Cancer Res.* 21, 5380–5390. doi:10.1158/1078-0432.CCR-15-1057

Zhou, G., Sprengers, D., Mancham, S., Erkens, R., Boor, P. P. C., van Beek, A. A., et al. (2019). Reduction of Immunosuppressive Tumor Microenvironment in Cholangiocarcinoma by *Ex Vivo* Targeting Immune Checkpoint Molecules. *J. Hepatol.* 71, 753–762. doi:10.1016/j.jhep.2019.05.026

Zhu, Y., Zhang, R., Zhang, Y., Cheng, X., Li, L., Wu, Z., et al. (2021). NUDT21 Promotes Tumor Growth and Metastasis through Modulating SGPP2 in Human Gastric Cancer. *Front. Oncol.* 11, 670353. doi:10.3389/fonc.2021.670353

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.

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fgene.2021.761141

Combining Polygenic Risk Score and Voice Features to Detect Major Depressive Disorders

Yazheng Di[1,2], Jingying Wang[3], Xiaoqian Liu[1,2] and Tingshao Zhu[1,2]*

[1]Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China, [3]School of Optometry, Faculty of Health and Social Sciences, Hong Kong Polytechnic University, Hong Kong, China

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Bingxin Zhao,
Purdue University, United States
Hongde Liu,
Southeast University, China
Stelios Fuentes,
University of Leicester,
United Kingdom

***Correspondence:**
Tingshao Zhu
tszhu@psych.ac.cn

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 19 August 2021
Accepted: 12 November 2021
Published: 20 December 2021

Citation:
Di Y, Wang J, Liu X and Zhu T (2021)
Combining Polygenic Risk Score and
Voice Features to Detect Major
Depressive Disorders.
Front. Genet. 12:761141.
doi: 10.3389/fgene.2021.761141

Background: The application of polygenic risk scores (PRSs) in major depressive disorder (MDD) detection is constrained by its simplicity and uncertainty. One promising way to further extend its usability is fusion with other biomarkers. This study constructed an MDD biomarker by combining the PRS and voice features and evaluated their ability based on large clinical samples.

Methods: We collected genome-wide sequences and utterances edited from clinical interview speech records from 3,580 women with recurrent MDD and 4,016 healthy people. Then, we constructed PRS as a gene biomarker by *p* value-based clumping and thresholding and extracted voice features using the i-vector method. Using logistic regression, we compared the ability of gene or voice biomarkers with the ability of both in combination for MDD detection. We also tested more machine learning models to further improve the detection capability.

Results: With a *p*-value threshold of 0.005, the combined biomarker improved the area under the receiver operating characteristic curve (AUC) by 9.09% compared to that of genes only and 6.73% compared to that of voice only. Multilayer perceptron can further heighten the AUC by 3.6% compared to logistic regression, while support vector machine and random forests showed no better performance.

Conclusion: The addition of voice biomarkers to genes can effectively improve the ability to detect MDD. The combination of PRS and voice biomarkers in MDD detection is feasible. This study provides a foundation for exploring the clinical application of genetic and voice biomarkers in the diagnosis of MDD.

Keywords: biomarkers, polygenic risk score (PRS), computer technology, major depressive disorder (MDD), voice biomarkers, depression

1 INTRODUCTION

The deployment of bioinformatic evaluations in psychiatry would revolutionize the ability to diagnose, treat, and prevent major depressive disorder (MDD). MDD affects nearly 1 in 10 people (Kessler et al., 2003; Demyttenaere et al., 2004) and has lately been recognized as the world's leading cause of disability (World Health Organization, 2017). However, only approximately half of the population suffering from MDD is currently identified and treated (Wells et al., 1989;

Goldberg 1995). The difficulty in identifying MDD is one of the key barriers to the effective utilization of current medications. Diagnosis remains based on clinical interviews and mental status examination (Regier et al., 2013); screening instruments are hindered by poor specificity and sensitivity, and there are no reliable biomarkers. Furthermore, because MDD is a syndromic diagnosis, it possibly comprises several different diseases, each with its own set of symptoms and treatment response (Alexopoulos et al., 1997; Kendler et al., 2001, 2006; Kendler et al., 2013; Gustafsson et al., 2015; Masters et al., 2015; Peterson et al., 2018).

The study of constructing MDD biomarkers has shown two different orientations. On the one hand, researchers have been devoted to finding the biological basis of depression (Schneider and Prvulovic 2013), for example, genetic factors (23andMe Research Team et al., 2019; CONVERGE Consortium, 2015; eQTLGen et al., 2018), on which to build valid biomarkers. On the other hand, studies have started from behavioral indices that are easily accessible and nonintrusive, such as patient speech voice (Low et al., 2020). The studies focus on improving diagnostic accuracy by developing machine learning (ML) algorithms.

Researchers have spent decades looking for the genetic foundation for developing more accurate MDD diagnosis models (Reus et al., 2017; Mullins et al., 2019; Rantalainen et al., 2020). The results from genome-wide association studies (GWAS) (23andMe Research Team et al., 2019; CONVERGE Consortium, 2015; eQTLGen et al., 2018) suggested that MDD is polygenic, which means that hundreds of DNA variants impact its hereditary influences with very small effects. Polygenic risk scores (PRSs) provide an estimated risk for individuals suffering from MDD. PRS is calculated as a weighted sum of an individual's risk alleles, where their weights are specified by loci and their assessed effects found by GWAS (Chatterjee et al., 2016). Advances in biotechnology have made sequencing technologies less expensive and the genetic screening of individuals easier. However, the utility of PRS in MDD prediction is currently constrained by its simplicity and uncertainty, which, to date, captures only part of the genetic contribution to MDD risk (Murray et al., 2021). Moreover, other non-genetic risk factors, such as lifestyles, also play important roles in MDD. As a result, extending the PRS models with other MDD biomarkers may be a more practical solution to addressing this problem (Torkamani et al., 2018).

Benefitting from the development of speech recognition technology, voice-based diagnostic models for depression have been validated and have achieved a high level of accuracy. Speech biomarkers can be used not only to identify depression (Low et al., 2020) but also to recognize the severity of depression (Shin et al., 2021) and predict depression-related symptoms (Zhang et al., 2020). One of the main obstacles hindering the application of voice biomarkers is its poor generalization ability, as traditional voice feature distribution can easily change due to different speech content and speakers (Wang et al., 2019). To address this issue, researchers developed the i-vector method, extracting the factors from voice features that are independent of speaker and channel variabilities (Dehak et al., 2011; Cummins et al., 2014). A study recognizing MDD in 1,808 clinical samples proved that voice i-vectors are effective and robust (Di et al., 2021). Therefore, combining technologies in speech recognition and integrating them into existing genetic models are likely to enable clinical diagnosis in general populations.

To construct biomarkers for clinical disease detection, researchers have combined PRS with known risk factors (Hoang et al., 2021; Kapoor et al., 2021), neuroimaging, metabolites (Badhwar et al., 2020), or body indicators (Moldovan et al., 2021). However, to the best of our knowledge, there are no studies combining genes and voice in detecting MDD, which may be due to the difficulty in obtaining multiple types of samples of the same subject simultaneously. Evidence from clinical samples is needed to prove their ecological validity (Zhang et al., 2020; Murray et al., 2021). The combination and cross-validation of biological and behavioral biomarkers hold great promise to take us one step closer to the objective clinical diagnosis of MDD.

Here, based on a large sample of women with recurrent MDD diagnosed clinically, we used the PRS together with voice i-vectors to detect MDD. We examined whether their combination could surpass a single biomarker. We constructed models on different single nucleotide polymorphisms (SNPs) to examine their robustness. We also tested various ML models to find the better model.

2 MATERIALS AND METHODS

We used a fivefold cross-validation design in this study. As shown in **Figure 1**, we split 80% of the samples into a training group and the rest into a test group. Firstly, we used voice data from the training samples to train the universal background model (UBM), and we used this UBM to extract i-vectors for each individual. Then, we used clumped SNP data from the training samples to train the PRS model, through which we calculated the PRS for each individual. Finally, we used the PRS and voice i-vectors from the training samples to train the ML models and used the same features from the test samples to validate the model performance. The details of each step are explained below.

2.1 Data Collection

The database used in this study was developed from the China, Oxford, and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE). The CONVERGE study, designed for a genome-wide association of major depression disorders, recruited 11,670 Han Chinese women. There were 5,303 women with recurrent MDD aged between 30 and 60 years whose first episodes of MDD met the DSM-IV criteria (Association 1994). A total of 5,337 controls were recruited from patients undergoing minor surgical procedures at general hospitals or from local community centers. Only women were included in this study to minimize genetic heterogeneity because approximately 45% of the genetic liability to MDD is not shared between sexes (Kendler et al., 2007; Sullivan et al., 2000). The subject inclusion criteria and interview process were strictly controlled, as detailed in CONVERGE Consortium (2015).

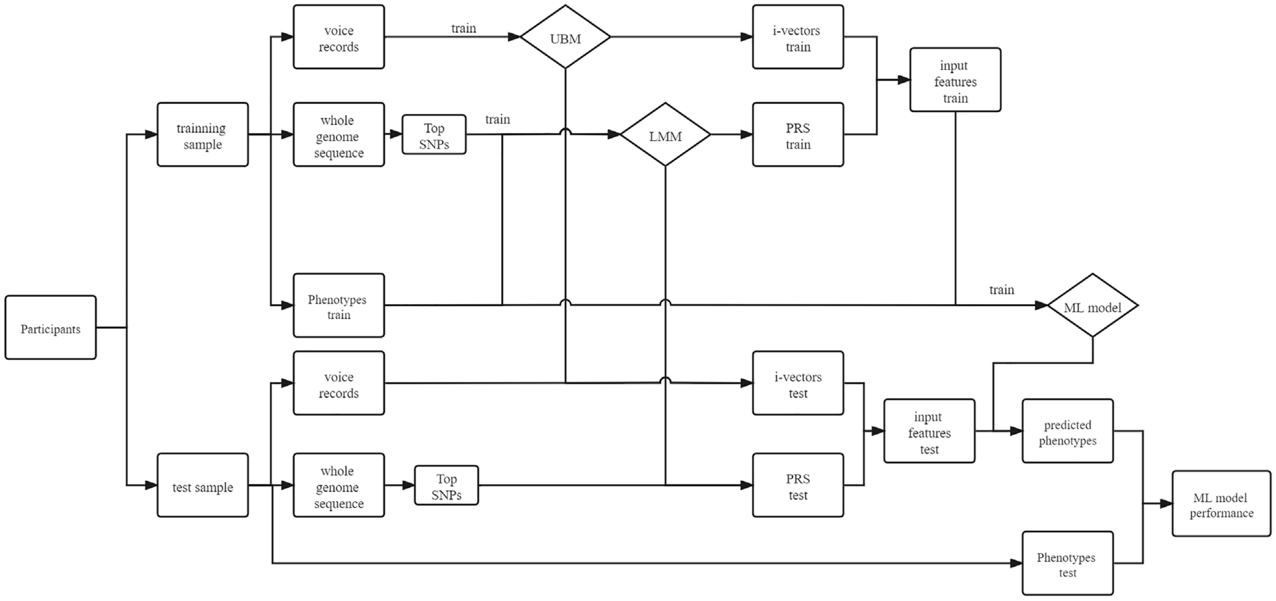


FIGURE 1 | Fivefold cross-validation of voice–gene data. In each fold, the samples were split into a training group and a test group. Voice and genetic sequence data of the training group were used to train the universal background model (UBM) and linear mixed model (LMM) separately. Then, i-vectors for the training and test groups were extracted through the UBM, and the polygenic risk score (PRS) can be calculated through the LMM. The i-vectors and PRS will be concatenated as input features for a machine learning (ML) model.

The voice data of the patients were from the records during the semi-structured interview, which included assessments of psychopathology, demographic and personal characteristics, and psychosocial functioning. These voice data are characterized by a high degree of phonetic and content variety. A detailed description of the interview protocol is in Di et al. (2021).

2.2 Data Preprocessing

2.2.1 Genetic Data

DNA sequencing, variant calling, and the genotype likelihood calculation and imputation processes are described in CONVERGE Consortium, (2015). We used PLINK (Chang et al., 2015) to select SNPs with minor allele frequency (MAF) >0.5% and imputation quality INFO score >0.9 and clumped the SNP set using $r^2 = 0.5$ with 50-kb windows. A total of 359,515 SNPs passed the filter.

2.2.2 Voice Data

The utterances of participants were edited from recordings of the conversations between doctors and patients through the following steps. Firstly, voice segments from the participants were selected and labeled. Then, all the segments of one participant were combined into one utterance. Due to the variety of interviews, not all voice samples of participants had segments >2 s for the latter analysis. Thus, samples with both genetic data and enough voice data were passed to subsequent analysis, and the total number was 7,596 (3,580 cases and 4,016 controls). All utterances were downsampled to 8 kHz for subsequent processing.

2.3 Data Analysis

2.3.1 PRS Models

We used the linear mixed model (Li and Zhu 2013) to calculate the PRS. The model can be written as follows:

$$y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{e}$$

Here, **X** is the matrix of the fixed effects, including covariates and the genetic matrix; the vector $\boldsymbol{\beta}$ is the coefficient of fixed effects; g is a random effect reflecting polygene background; and e denotes the random residual effect.

We used the p value-based thresholding (P+T) method (Wray et al, 2007) to construct the PRS model. Usually, a lower threshold than genome-wide statistical significance can be applied to increase the overall predictability, generally at the sacrifice of generalizability (Murray et al., 2021). Different p-value thresholds (PTs) were tested, ranging from 5×10^{-8} to 5×10^{-3} (10^{-3} is a conservative significance threshold of p suggested by Euesden et al., 2015). The PRS model was trained using the FaST-LMM (Lippert et al., 2011) predictor, which efficiently reduced the computational time.

To assess how the confounders affect the model's predictability and generalizability, we considered the following covariates: age, education, occupation, social class, marital status, height, weight, and 40 genetic principal components. We compared three different covariate use strategies. The first was a model ignoring the covariates (no-cov), the second was trained by and predicted on the genetic matrix along with covariates (all-cov), and the last was a model trained by a genetic matrix along with true covariates of the training samples, but made predictions on test samples whose covariate values were replaced with random numbers (random-cov).

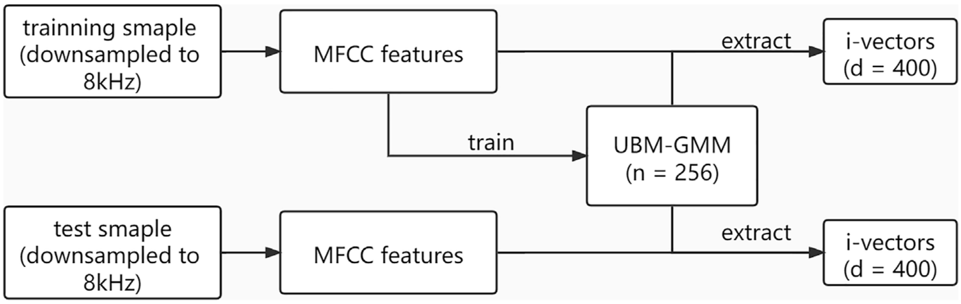


FIGURE 2 | Process of i-vector extraction. UBM-GMM is a universal background model adapted by a Gaussian mixture model. *n* = 256 means there were 256 Gaussian mixture clusters. *d* = 400 means the dimension of i-vectors is 400.

2.3.2 Voice i-Vectors

The i-vector extraction process is shown in **Figure 2**. Firstly, mel frequency cepstral coefficients (MFCCs) were extracted with a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with a coefficient of 0.97, and a sinusoidal lifter with a coefficient of 22. A filter bank with 23 filters was used, and 12 coefficients were extracted. Then, for the given voice features, we set the number of Gaussian mixtures as 256 to estimate the utterance-dependent Gaussian mixture model (GMM) parameters and adapted the UBM (Kenny et al., 2005), which represents the feature distribution of the acoustic space.

The i-vectors are low-dimensional representations of the voice features based on factor analysis (Kenny et al., 2008), onto which the acoustic space is mapped via a linear transformation while keeping the majority of the variability inherent in the acoustic space. This approach has been widely used in speaker verification. The i-vector method (Dehak et al., 2011) can be expressed as follows:

$$\boldsymbol{M} = \boldsymbol{m} + T\boldsymbol{v}$$

where $\boldsymbol{m}$ is the mean supervector of the UBM. For the purpose of depression classification, it is expected that the UBM approximately models the phonetic variability of the acoustic space. $\boldsymbol{M}$ is the mean centered supervector of the speech utterance derived using the zeroth- and first-order Baum–Welch statistics. $\boldsymbol{v}$ is the i-vector, which captures variations in this structure caused by other factors, such as depression level, speaker identity, and channel effects (Cummins et al., 2014). We used the Kaldi speech recognition toolkit (Povey et al., 2011) and extracted the 400 dimensions of i-vectors.

2.3.3 ML Models

We used a logistic regression (LR) classifier as a benchmark model, for which PRS, i-vectors, and both were used as input features. Then, we used random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP) classifiers to test whether there was an improvement compared to the benchmark. We report the sensitivity, specificity, and area under the receiver operator characteristic curve (AUC) from the fivefold cross-validation. We used scikit-learn (Pedregosa et al., 2011) for the above process.

TABLE 1 | Number of SNPs selected on different *p*-value thresholds (PTs)

PT	5E–08	1E–06	1E–5	5E–5	1E–4	5E–4	1E–3	5E–3
N	3	5	11	44	79	321	580	2,350

For the LR model, we also divided the test samples into the top 25%, middle 50%, and bottom 25% according to their PRS and calculated the accuracy on each stratification to test whether the accuracy of the biomarkers remains consistent across different genetic risk stratifications.

2.3.4 Binary Logistic Regression

To check the contribution of voice and genes separately, we built three logistic regression models using a conditional forward step. Taking MDD as the dependent variable, voice i-vectors, PRS, and the combination of both were entered into the model separately as independent variables. Nagelkerke's R^2 (Nagelkerke, 1991) was utilized as an indicator of the contributing effect of the variables.

3 RESULTS

3.1 PRS Model and Covariates

The numbers of SNPs selected using different PTs are shown in **Table 1**. The SNPs and their estimated weights in previous GWAS (CONVERGE Consortium 2015) are provided in **Supplementary Data Sheet S1**. **Figure 3** shows the detection ability of the PRS models with different PTs using different covariate use strategies. When PT = 5×10^{-8}, PRS with all-cov achieved the best AUC (0.64), while the other two were close to random guessing (0.50). When $PT > 5 \times 10^{-8}$, PRS with no-cov consistently achieved better AUC than did PRS with all-cov and random-cov, and the performances of PRS with random-cov and all-cov were close.

3.2 Prediction Results Using Different Biomarkers

Figure 4 shows the prediction results with different PTs using different biomarkers. The voice biomarkers achieved an AUC of 0.79. With the decrease in PT, the AUCs for genes only and the

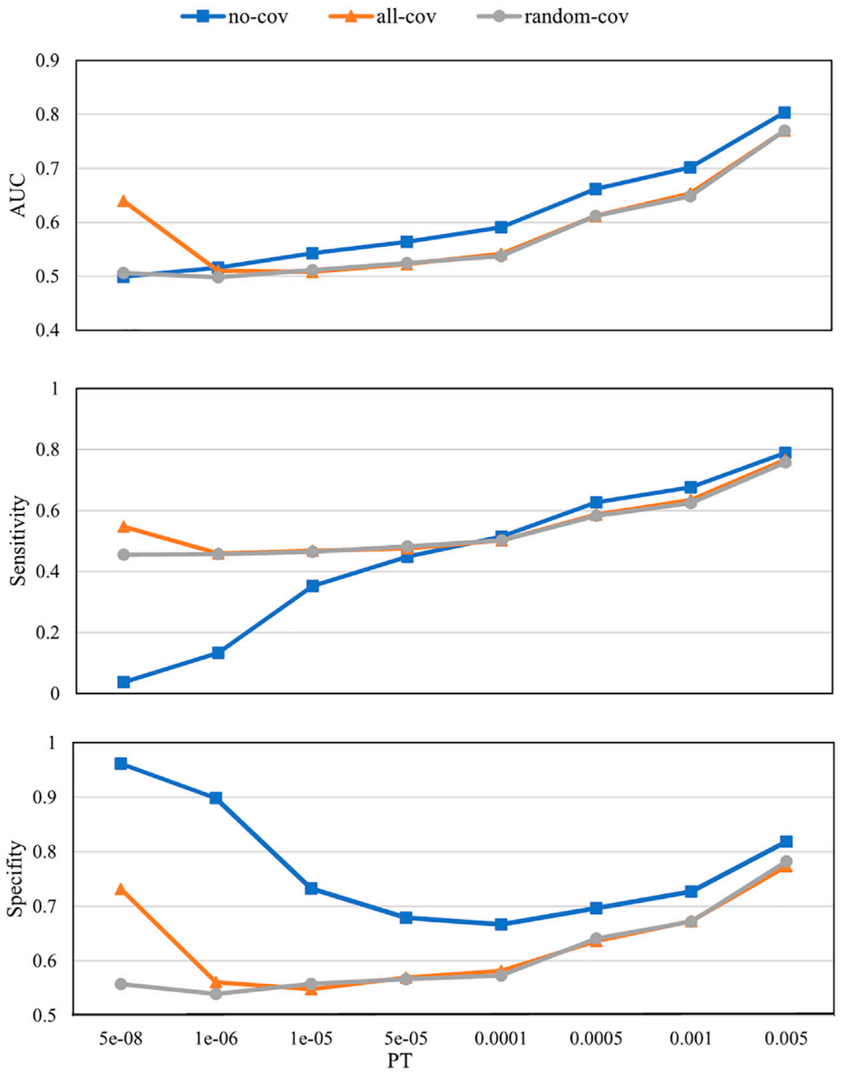


FIGURE 3 | Polygenic risk score (PRS) model prediction results with different *p*-value thresholds (PTs) under different covariate use strategies. *no-cov*, no covariates were considered during the training and prediction processes; *all-cov*, all covariates were considered during the training and prediction processes; *random-cov*, the PRS model was trained with a sample genetic matrix along with covariates, but made predictions on samples whose covariates were replaced with random numbers. AUC, Area under the receiver operating characteristic curve.

combined biomarkers both increased. Compared with genes only, the combined biomarkers always performed better. Compared with voice only, the combined biomarkers did not win until PT > 0.0005.

3.3 Binary Logistic Regression

We examined how much gene and voice contributed to MDD using Nagelkerke's R^2. For voice only and gene only, Nagelkerke's R^2 values were 0.571 and 0.829, respectively. For the combined biomarkers, Nagelkerke's R^2 was 0.902. Details of the logistic regression models are in **Supplementary Data Sheet S2**.

Figure 5 shows the stratified accuracies of the different biomarkers in predicting MDD. The voice biomarker performed consistently in the three stratifications with different genetic risks, all at 0.79. However, the accuracy of genes varied considerably between the middle (0.64) and the two ends of the population (close to 0.9). The combined biomarker performed as well as the genes in the two ends and as well as the voice in the middle.

3.4 Classification Results Using Different ML Models

The classification results using LR, SVM, RF, and MLP are shown in **Table 2**. Two PTs (0.001 and 0.005) are presented here, while the results with more PTs are shown in **Supplementary Data Sheet S3**. The AUCs of LR were 0.79 and 0.83 at the two PTs. Taking LR as a benchmark, MLP achieved better results, with AUCs of 0.81 and 0.86 at the two PTs. The performance of SVM was close to that of LR, and that of RF was worse than that of LR.

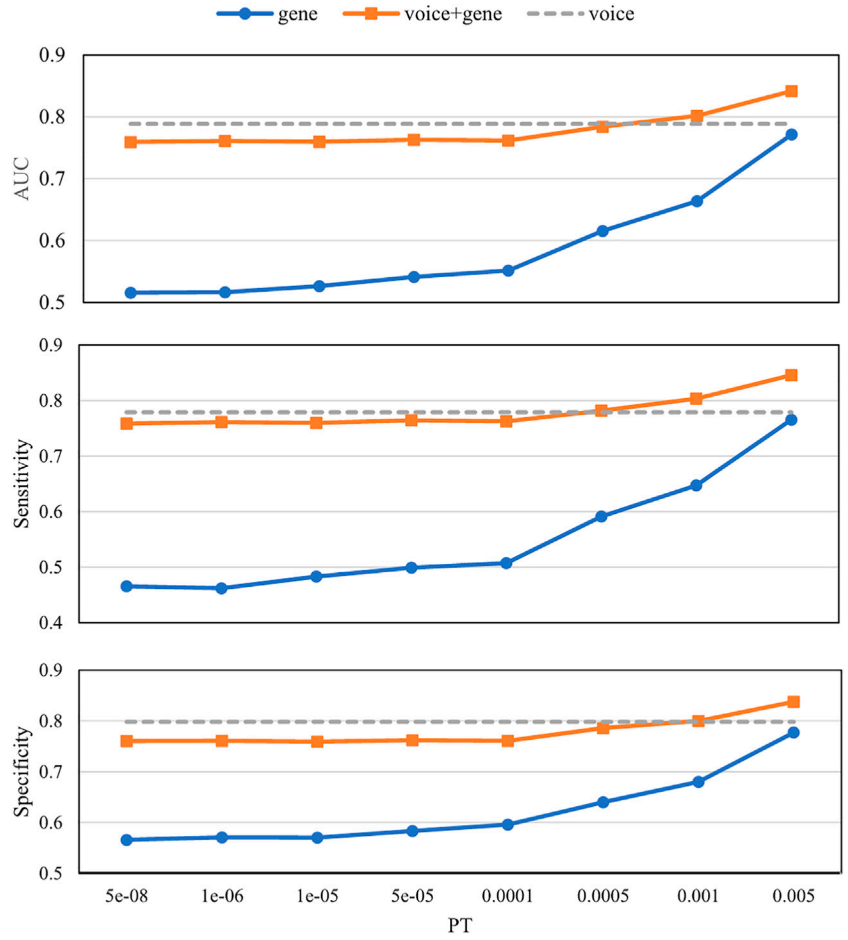


FIGURE 4 | Prediction results with different *p*-value thresholds (PTs) using different biomarkers. The *x*-axis is the *p*-value threshold (PT) used in the gene model and the combined biomarkers. Voice biomarkers are not related to PT and are indicated by a *dashed horizontal line*. AUC, Area under the receiver operating characteristic curve.

4 DISCUSSION

This study combines the PRS and voice i-vectors to evaluate their ability to detect MDD. PRSs were calculated at different PTs. Using logistic regression, we compared the abilities of single biomarkers with the combined biomarker for MDD detection. We stratified the test group by genetic risk and examined whether the detection ability differed between stratifications. We also tested various ML models to find the best model.

A good PRS model would have high predictability, contributed mainly by capturing causal genetic variants instead of confounds. The estimated genetic fixed effect may be erroneously high for a linear mixed model if the confounding effects are not estimated. Thus, similar to our data from the same cohort, the no-cov PRS model always performed better than the all-cov PRS model (**Figure 3**). We believe that the results from the no-cov model are not capable of reflecting real situations because, in practical clinical applications, the distribution of covariates for a newly arrived patient is likely to be different from the distribution of the patients in our training set.

A comparison of PRS with the all-cov and random-cov models can demonstrate how the covariates affect the final prediction results in this study. When $PT = 5 \times 10^{-8}$, the all-cov PRS achieved an AUC of 0.65, while the other two were close to random guessing, which indicated that the covariates were the main contributors to the predictor when there were few SNPs. When $PT > 5 \times 10^{-8}$, the all-cov PRS showed ability equivalent to that of the random-cov PRS, which suggested that the covariates contributed very little to the predictor when the SNP number increased. The covariate analysis suggests two conclusions. Firstly, we must consider the covariates in the training process; otherwise, the performance will be erroneously better than that in actual situations. Secondly, in practical clinical applications, covariate information is not necessary.

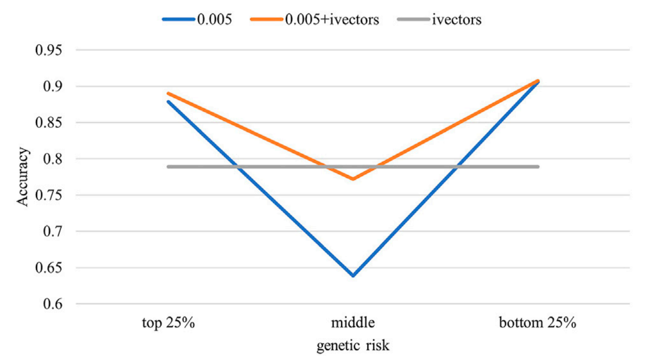


FIGURE 5 | Stratified population accuracy using different biomarkers. The test samples were divided into three groups according to their predicted polygenic risk scores (PRSs). Accuracies were calculated for the three groups separately.

The prediction results using different biomarkers demonstrated the ability of these biomarkers to detect MDD (**Figure 4**). The AUC of voice biomarkers was 0.79, which is consistent with our previous study on 1,808 clinical samples (Di et al., 2021). Since our previous study investigated the meaning of voice i-vectors, in this study, we attended to comparing its performance with the combination of PRS. Compared with genes only, the combined biomarkers can significantly improve the predictive ability at all PTs. Since the voice biomarker itself had an AUC of 0.79, only when the AUC of gene >0.65 (PT > 0.0005) can the combined biomarker perform better than voice only.

When only PRS was entered in the logistic model, it accounted for 82.9% of the variance in the dependent variable MDD (Nagelkerke's R^2). Combined with voice, the Nagelkerke's R^2 was 90.2%, indicating that the unique contribution of voice features was 8.7%. Furthermore, we illustrated how genes and voice work together to improve the predictive power by stratifying the test sample according to genetic risks and calculating the accuracies by stratifications. The results of the genes in identifying MDD for both high- and low-genetic-risk populations were consistent with the high accuracy (0.90). However, for the middle population, the accuracy of genes was poor (0.64), due mainly to the inability of genetic features to measure the effect of MDD-related environmental factors. Meanwhile, the accuracy of voice was consistent across the different genetic risk populations, suggesting that the predictive ability of voice was independent of genetic characteristics and that voice capture information was independent of genes. As a result, combining gene and voice biomarkers can effectively improve the detection ability of MDD.

We further explored whether different ML models can further improve the prediction of MDD. For the ML models, we tested the results using SVM, RF, and MLP and compared them with the results of LR. The results (**Table 2** and **Supplementary Data Sheet S3**) showed that MLP could indeed further improve the prediction of the model, improving the AUC by 2.5% with PT = 0.001 and by 3.6% with PT = 0.005.

TABLE 2 | Classification results using different machine learning (ML) models

	Gene (PT = 0.001) + voice			Gene (PT = 0.005) + voice		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
LR	0.79	0.79	0.78	0.83	0.83	0.83
SVM	0.79	0.80	0.78	0.83	0.83	0.83
RF	0.74	0.70	0.77	0.80	0.78	0.81
MLP	0.81	0.83	0.79	0.86	0.87	0.85

PT, p-value threshold; AUC, area under the receiver operating characteristic curve; LR, logistic regression; SVM, support vector machine; RF, random forest; MLP, multilayer perceptron

There are several limitations in this research. To ensure homogeneity between subjects, this study selected women with recurrent MDD as cases, and 85% of the cases met the DSM-IV criteria for melancholia, which is a severe subtype of MDD (CONVERGE Consortium 2015). Thus, our samples represent the two poles of the distribution of depression severity in natural populations. Although our experiments effectively demonstrated that the combination of genes and voice could further improve their ability to identify MDD, experimental results based on a more general population are needed before clinical application.

5 CONCLUSION

This study combines the PRS and voice i-vectors to evaluate their ability to detect MDD. PRSs are calculated at different PTs. With the *p*-value threshold at 0.005, the combined biomarker improved the AUC by 9.09% compared to genes only and 6.73% compared to voice only. Genetic risk stratification analysis showed that the ability for MDD detection of voice is genetically independent. Multilayer perceptron further improved the AUC by 3.6% compared to logistic regression. The combination of PRS and voice biomarkers in MDD detection is feasible. This study provides a foundation for exploring the clinical application of genetic and voice biomarkers in the diagnosis of MDD (Wray et al., 2018).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The data can be found here: https://www.ebi.ac.uk/ena/browser/view/PRJNA289433?show=related-records.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical Review Board of Oxford University (Oxford Tropical Research Ethics Committee). The patients/

participants provided written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the study. XL and JW organized the database. YD performed the statistical analysis. YD wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

Alexopoulos, G. S., Meyers, B. S., Young, R. C., Campbell, S., Silbersweig, D., and Charlson, M. (1997). 'Vascular Depression' Hypothesis. *Arch. Gen. Psychiatry* 54 (10), 915–922. doi:10.1001/archpsyc.1997.01830220033006

Association, A. P. (1994). *Diagnostic and Statistical Manual of Mental Disorders*. Washington, D.C: American Psychiatric Association.

Badhwar, A., McFall, G. P., Sapkota, S., Black, S. E., Chertkow, H., Duchesne, S., et al. (2020). A Multiomics Approach to Heterogeneity in Alzheimer's Disease: Focused Review and Roadmap. *Brain* 143 (May), 1315–1331. doi:10.1093/brain/awz384

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaSci* 4 (1), 7. doi:10.1186/s13742-015-0047-8

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and Evaluating Polygenic Risk Prediction Models for Stratified Disease Prevention. *Nat. Rev. Genet.* 17 (7), 392–406. doi:10.1038/nrg.2016.27

CONVERGE consortium (2015). Sparse Whole-Genome Sequencing Identifies Two Loci for Major Depressive Disorder. *Nature* 523 (7562), 588–591. doi:10.1038/nature14659

Cummins, N., Epps, J., Sethu, V., and Krajewski, J. (2014).Variability Compensation in Small Data: Oversampled Extraction of I-Vectors for the Classification of Depressed Speech, Proceeding of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, Florence, Italy. IEEE, 970–974. doi:10.1109/ICASSP.2014.6853741

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Trans. Audio Speech Lang. Process.* 19 (4), 788–798. doi:10.1109/tasl.2010.2064307

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J. P., et al. (2004). Prevalence, Severity, and Unmet Need for Treatment of Mental Disorders in the World Health Organization World Mental Health Surveys. *Jama* 291 (21), 2581–2590. doi:10.1001/jama.291.21.2581

Di, Y., Wang, J., Li, W., and Zhu, T. (2021). Using I-Vectors from Voice Features to Identify Major Depressive Disorder. *J. Affective Disord.* 288 (June), 161–166. doi:10.1016/j.jad.2021.04.004

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score Software. *Bioinformatics* 31 (9), 1466–1468. doi:10.1093/bioinformatics/btu848

Goldberg, D. (1995). Epidemiology of Mental Disorders in Primary Care Settings. *Epidemiologic Rev.* 17 (1), 182–190. doi:10.1093/oxfordjournals.epirev.a036174

Gustafsson, H., Nordstrom, A., and Nordstrom, P. (2015). Depression and Subsequent Risk of Parkinson Disease: A Nationwide Cohort Study. *Neurology* 84 (24), 2422–2429. doi:10.1212/wnl.0000000000001684

Hoang, T., Nguyen Ngoc, Q., Lee, J., Lee, E. K., Hwangbo, Y., and Kim, J. (2021). Evaluation of Modifiable Factors and Polygenic Risk Score in Thyroid Cancer. *Endocrine-Related Cancer* 28 (7), 481–494. doi:10.1530/ERC-21-0078

Howard, D. M., Adams, M. J., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., et al. (2019). Genome-Wide Meta-Analysis of Depression Identifies 102 Independent Variants and Highlights the Importance of the Prefrontal Brain Regions. *Nat. Neurosci.* 22 (3), 343–352. doi:10.1038/s41593-018-0326-7

Kapoor, P. M., Middha, P., Mavaddat, N., Choudhury, P. P., Wilcox, A. N., Lindstrom, S., et al. (2021). Combined Associations of a Polygenic Risk Score and Classical Risk Factors with Breast Cancer Risk. *Jnci-Journal Natl. Cancer Inst.* 113 (3), 329–337. doi:10.1093/jnci/djaa056

Kendler, K. S., Aggen, S. H., and Neale, M. C. (2013). Evidence for Multiple Genetic Factors Underlying DSM-IV Criteria for Major Depression. *JAMA Psychiatry* 70 (6), 599–607. doi:10.1001/jamapsychiatry.2013.751

Kendler, K. S., Gardner, C. O., Neale, M. C., and Prescott, C. A. (2001). Genetic Risk Factors for Major Depression in Men and Women: Similar or Different Heritabilities and Same or Partly Distinct Genes. *Psychol. Med.* 31 (4), 605–616. doi:10.1017/s0033291701003907

Kendler, K. S., Gatz, M., Gardner, C. O., and Pedersen, N. L. (2006). A Swedish National Twin Study of Lifetime Major Depression. *Am. J. Psychiatry* 163 (1), 109–114. doi:10.1176/appi.ajp.163.1.109

Kendler, K. S., Gatz, M., Gardner, C. O., and Pedersen, N. L. (2007). Clinical Indices of Familial Depression in the Swedish Twin Registry. *Acta Psychiatr. Scand.* 115 (3), 214–220. doi:10.1111/j.1600-0447.2006.00863.x

Kenny, P., Boulianne, G., and Dumouchel, P. (2005). Eigenvoice Modeling with Sparse Training Data. *IEEE Trans. Speech Audio Process.* 13 (3), 345–354. doi:10.1109/TSA.2004.840940

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. (2008). A Study of Interspeaker Variability in Speaker Verification. *IEEE Trans. Audio Speech Lang. Process.* 16 (5), 980–988. doi:10.1109/TASL.2008.925147

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The Epidemiology of Major Depressive Disorder. *JAMA* 289 (23), 3095–3105. doi:10.1001/jama.289.23.3095

Li, G., and Zhu, H. (2013). Genetic Studies: The Linear Mixed Models in Genome-wide Association Studies. *Open Bioinformatics J.* 7 (1), 27–33. doi:10.2174/1875036201307010027

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST Linear Mixed Models for Genome-wide Association Studies. *Nat. Methods* 8 (10), 833–835. doi:10.1038/nmeth.1681

Low, D. M., Bentley, K. H., and Ghosh, S. S. (2020). Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review. *Laryngoscope Invest. Otolaryngol.* 5 (1), 96–116. doi:10.1002/lio2.354

Masters, M. C., Morris, J. C., and Roe, C. M. (2015). "Noncognitive" Symptoms of Early Alzheimer Disease: A Longitudinal Analysis. *Neurology* 84 (6), 617–622. doi:10.1212/wnl.0000000000001238

Moldovan, A., Waldman, Y. Y., Brandes, N., and Linial, M. (2021). Body Mass Index and Birth Weight Improve Polygenic Risk Score for Type 2 Diabetes. *J. Personalized Med.* 11 (6), 582. doi:10.3390/jpm11060582

Mullins, N., Bigdeli, T. B., Børglum, A. D., Coleman, J. R. I., Demontis, D., Mehta, D., et al. (2019). GWAS of Suicide Attempt in Psychiatric Disorders and Association with Major Depression Polygenic Risk Scores. *Am. J. Psychiatry* 176 (8), 651–660. doi:10.1176/appi.ajp.2019.18080957

Murray, G. K., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., and Wray, N. R. (2021). Could Polygenic Risk Scores Be Useful in Psychiatry. *JAMA Psychiatry* 78 (2), 210. doi:10.1001/jamapsychiatry.2020.3042

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika* 78 (3), 691–692. doi:10.1093/biomet/78.3.691

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *J. Machine Learn. Res.* 12, 2825–2830. doi:10.5555/1953048.2078195

Peterson, R. E., Cai, N., Dahl, A. W., Bigdeli, T. B., Edwards, A. C., Webb, B. T., et al. (2018). Molecular Genetic Analysis Subdivided by Adversity Exposure

FUNDING

This research is funded by the Key Research Program of the Chinese Academy of Sciences (ZDRW-XH-2019-4).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.761141/full#supplementary-material

Suggests Etiologic Heterogeneity in Major Depression. *Am. J. Psychiatry* 175 (6), 545. doi:10.1176/appi.ajp.2017.17060621

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011).The Kaldi Speech Recognition Toolkit, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.

Rantalainen, V., Binder, E. B., Lahti-Pulkkinen, M., Czamara, D., Laivuori, H., Villa, P. M., et al. (2020). Polygenic Prediction of the Risk of Perinatal Depressive Symptoms. *Depress. Anxiety* 37 (9), 862–875. doi:10.1002/da.23066

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., et al. (2013). DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *Am. J. Psychiatry* 170 (1), 59–70. doi:10.1176/appi.ajp.2012.12070999

Reus, L. M., Shen, X., Gibson, J., Wigmore, E., Ligthart, L., Adams, M. J., et al. (2017). Association of Polygenic Risk for Major Psychiatric Illness with Subcortical Volumes and White Matter Integrity in UK Biobank. *Sci. Rep.* 7 (1), 42140. doi:10.1038/srep42140

the Major Depressive Disorder Working Group of the Psychiatric Genomics ConsortiumRipke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., et al. (2018). Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression. *Nat. Genet.* 50 (5), 668–681. doi:10.1038/s41588-018-0090-3

Schneider, B., and Prvulovic, D. (2013). Novel Biomarkers in Major Depression. *Curr. Opin. Psychiatry* 26 (1), 47–53. doi:10.1097/YCO.0b013e32835a5947

Shin, D., Cho, W. I., Park, C. H. K., Rhee, S. J., Kim, M. J., Lee, H., et al. (2021). Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning. *J. Clin. Med.* 10 (14), 3046. doi:10.3390/jcm10143046

Sullivan, P. F., Neale, M. C., and Kendler, K. S. (2000). Genetic Epidemiology of Major Depression: Review and Meta-Analysis. *Am. J. Psychiatry* 157 (10), 1552–1562. doi:10.1176/appi.ajp.157.10.1552

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The Personal and Clinical Utility of Polygenic Risk Scores. *Nat. Rev. Genet.* 19 (9), 581–590. doi:10.1038/s41576-018-0018-x

Wang, J., Zhang, L., Liu, T., Pan, W., Hu, B., and Zhu, T. (2019). Acoustic Differences between Healthy and Depressed People: A Cross-Situation Study. *BMC Psychiatry* 19 (1), 300. doi:10.1186/s12888-019-2300-7

Wells, K. B., Hays, R. D., Burnam, M. A., Rogers, W., Greenfield, S., and Ware, J. E. (1989). Detection of Depressive Disorder for Patients Receiving Prepaid or Fee-For-Service Care. *Jama* 262 (23), 3298–3302. doi:10.1001/jama.1989.03430230083030

World Health Organization (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. Available at: https://www.who.int/publications/i/item/depression-global-health-estimates (Accessed December 6, 2021).

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of Individual Genetic Risk to Disease from Genome-wide Association Studies. *Genome Res.* 17 (10), 1520–1528. doi:10.1101/gr.6665407

Zhang, L., Duvvuri, R., Chandra, K. K. L., Nguyen, T., Ghomi, R. H., and Ghomi, R. H. (2020). Automated Voice Biomarkers for Depression Symptoms Using an Online Cross-sectional Data Collection Initiative. *Depress. Anxiety* 37 (7), 657–669. doi:10.1002/da.23020

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.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 18 January 2022
doi: 10.3389/fbioe.2021.758495

Two-Stage Deep Neural Network *via* Ensemble Learning for Melanoma Classification

Jiaqi Ding[1]*, Jie Song* [1]*, Jiawei Li*[1]*, Jijun Tang*[2]** and Fei Guo*[3]**

[1]*School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China,* [2]*Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China,* [3]*School of Computer Science and Engineering, Central South University, Changsha, China*

Melanoma is a skin disease with a high fatality rate. Early diagnosis of melanoma can effectively increase the survival rate of patients. There are three types of dermoscopy images, malignant melanoma, benign nevis, and seborrheic keratosis, so using dermoscopy images to classify melanoma is an indispensable task in diagnosis. However, early melanoma classification works can only use the low-level information of images, so the melanoma cannot be classified efficiently; the recent deep learning methods mainly depend on a single network, although it can extract high-level features, the poor scale and type of the features limited the results of the classification. Therefore, we need an automatic classification method for melanoma, which can make full use of the rich and deep feature information of images for classification. In this study, we propose an ensemble method that can integrate different types of classification networks for melanoma classification. Specifically, we first use U-net to segment the lesion area of images to generate a lesion mask, thus resize images to focus on the lesion; then, we use five excellent classification models to classify dermoscopy images, and adding squeeze-excitation block (SE block) to models to emphasize the more informative features; finally, we use our proposed new ensemble network to integrate five different classification results. The experimental results prove the validity of our results. We test our method on the ISIC 2017 challenge dataset and obtain excellent results on multiple metrics; especially, we get 0.909 on accuracy. Our classification framework can provide an efficient and accurate way for melanoma classification using dermoscopy images, laying the foundation for early diagnosis and later treatment of melanoma.

Keywords: melanoma classification, ensemble learning, deep convolutional neural network, image segmentation, dermoscopy images

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Sharmin Majumder,
Texas A&M University, United States
Ravneet Kaur,
Hitachi America Ltd., United States
Stelios Fuentes,
University of Leicester,
United Kingdom

****Correspondence:***
Jijun Tang
jj.tang@siat.ac.cn
Fei Guo
guofei@csu.edu.cn

Specialty section:
This article was submitted to Biosafety and Biosecurity, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *14 August 2021*
Accepted: *06 December 2021*
Published: *18 January 2022*

Citation:
Ding J, Song J, Li J, Tang J and Guo F (2022) Two-Stage Deep Neural Network via *Ensemble Learning for Melanoma Classification. Front. Bioeng. Biotechnol. 9:758495. doi: 10.3389/fbioe.2021.758495*

1 INTRODUCTION

Skin cancer is a major public health problem, with more than 5 million new cases diagnosed annually in the United States (Siegel et al., 2016; Codella et al., 2018). Melanoma is the fastest-growing and deadliest form of skin cancer in the world; it causes many deaths each year. However, it is noticed that melanoma multiplies more slowly in the early stages, so if it is diagnosed early and treated promptly, the survival rates of patients can be greatly improved.

Pigmentation lesions occur on the skin surface, and dermoscopic technology was introduced to improve the diagnosis of skin melanoma. Dermoscopy is a non-invasive skin imaging technique that can magnify and illuminate skin areas, and then enhance visualization of deep skin by eliminating surface reflections. Compared with standard photography, dermoscopy images can greatly improve the accuracy of diagnosis (Kittler et al., 2002; Codella et al., 2018). Dermatologists usually use "ABCD" rule to evaluate skin lesions (Stolz, 1994; Moura et al., 2019). This rule analyzes asymmetry, boundary irregularities, color variations, and structures of lesions (Xie et al., 2016). However, the differentiation of skin lesions by dermatologists from dermoscopy images is often time consuming and subjective, and the diagnostic accuracy depends largely on the professional level, so inexperienced dermatologists may not be able to make accurate judgments. Therefore, we urgently need an automatic recognition method that is non-subjective and can assist dermatologists to make more accurate diagnosis.

However, there are still many challenges in automated recognition of melanoma, we show them in **Figure 1**. The first column of **Figure 1** shows malignant melanoma, the second column shows benign nevis, and the third column shows seborrheic keratosis. First, skin lesions have great inter-class similarity and intra-class variation in color, shape, and texture; the different classes of skin lesion have high visual similarity. Second, the area of skin lesions in dermoscopy images varies greatly, and the boundaries between skin lesions and normal skin are blurred in some images. Third, artifacts such as hair, rulers, and texture in dermoscopy images may make it hard to identify melanoma changes. All these factors make automatic recognition more difficult.

To solve these problems, many researches have made attempts. Generally, automatic analysis models include four steps: image preprocessing, border detection or segmentation, feature extraction, and classification. In early works, a large number of studies used shallow models to classify dermoscopy images, mainly using low-level features such as shape, color, texture, or their combination (Ganster et al., 2001; Mishra and Celebi, 2016); however, these shallow models for extracting low-level features lack high-level representation and powerful generalization capabilities. In recent years, convolutional neural network has made great breakthroughs in image analysis tasks (Krizhevsky et al., 2012; He et al., 2015; Long et al., 2015; Shin et al., 2016; Chen et al., 2017), especially the deep convolutional neural networks (DCNNs), which can extract deep features and have better discrimination ability, have achieved improved performance. So researchers started to apply DCNN to analyze medical images (Roychowdhury et al., 2015; Myronenko, 2018), including image-based melanoma classification. However, deep neural networks still face great challenges in the field of medical image analysis. DCNN requires large datasets to obtain more effective features, while medical image data are often difficult to obtain and the datasets are relatively small. If a small dataset is used directly for deep network training, it will lead to over-fitting of the model. Moreover, a single network may not be able to extract all the informative features, and it is actually difficult to train a model that performs well in all aspects. Therefore, we propose an integrated model based on transfer learning to combine the results of multiple models to get better performance.

In this paper, we propose a novel two-stage ensemble method based on deep convolutional neural networks. In the first stage, we perform the image segmentation, we use a segmentation network to generate lesion segmentation masks, and then we use these masks to resize the original images so that they are the same size. In the second stage, we implement image classification, we utilize five state-of-the-art networks to extract features, and we add Squeeze-and-Excitation Blocks (Hu et al., 2018) to the network to help emphasize more informative features. Then we construct a new neural network using local connection to integrate the classification results of these models, so that we can obtain the final classification result. We evaluate our method on ISIC 2017 challenge dataset and obtain the best results on some metrics.

2 RELATED WORKS

2.1 Traditional Methods

Traditional methods are usually based on manually extracted features to classify dermoscopy images, including features of color and texture. The "ABCD" rule is the standard used by dermatologists, and there are many automatic classification methods that are based on this rule. Barata et al. (2013) introduced two different dermoscopy image detection systems; one used a global approach to classify skin lesions and the other used local features and a bag-of-features (BoF) classifier. Ganster et al. (2001) used manual features containing shape, boundary, and radiometric features to describe lesions, and then used KNN (K-Nearest Neighbor) to classify melanoma. Celebi et al. (2007) extracted descriptors related to shape, color, and texture from dermoscopy images and used non-linear support vector machines to classify melanoma lesions. Capdehourat et al. (2011) first preprocessed the image with hair removal, then used segmentation algorithm to segment each image, and finally trained the AdaBoost classifier with descriptors containing shape and color information.

2.2 Deep CNN Models

In recent years, convolutional neural network (CNN) has been widely used in image segmentation (Roychowdhury et al., 2015; Dai et al., 2016; Myronenko, 2018) and classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016; Szegedy et al., 2016; Chollet, 2017; Szegedy et al., 2017; Huang et al., 2017), object detection (He et al., 2015; Liu et al., 2016; Redmon et al., 2016), and other scopes of computer vision (Xiao et al., 2021; Chen et al., 2021b). CNN models have multiple layers to extract features. The network extractor mainly has two parts, convolutional layers and pooling layers, and the network classifier is the fully connected layer. Convolutional layers use convolutional kernels to carry out convolution operation with input images to extract features. Kernels obtain features of the whole image by sliding on it as a window. Also, the convolution operation of each kernel is only connected to a local

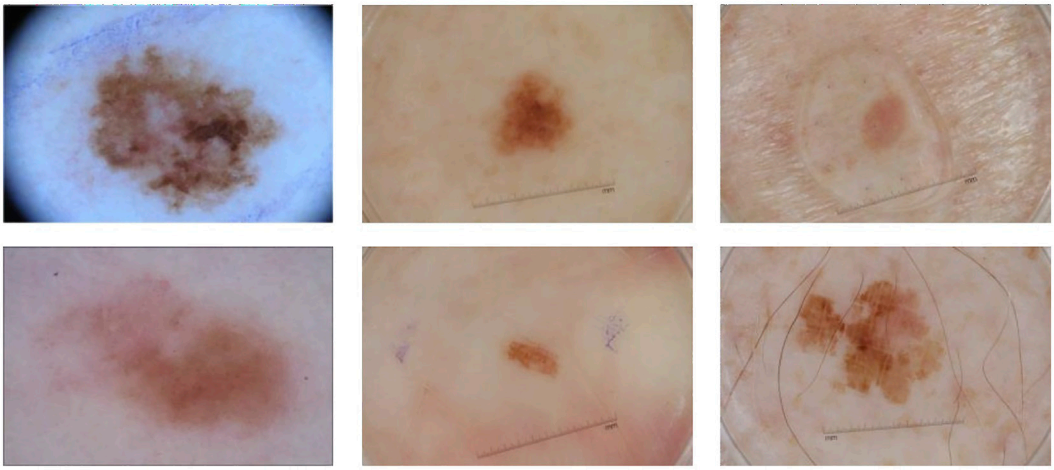


FIGURE 1 | Some samples of dermoscopy images. From left to right: malignant melanoma, benign nevis, and seborrheic keratosis.

area called receptive field of the input. Receptive field and weight sharing are important parts of convolution neural network; they can effectively change the amount of training parameters. Pooling operation is a kind of down sampling; its purpose is to reduce the training time, increase the receptive field, and prevent overfitting, including widely used max pooling and average pooling. In addition, the fully connected layer maps the learned feature representation to the label space for classification. If you need to classify the samples into *n* classes, there are *n* neurons in the last fully connected layer.

Many CNN models have great performance on computer vision tasks (Cao et al., 2021; Chen et al., 2021a; Feng et al., 2021). Studies have shown that increasing the number of layers in a network can significantly improve the performance (Simonyan and Zisserman, 2014; Szegedy et al., 2015). In recent years, deep CNN has been proposed and performed well in the field of dermoscopy recognition. Codella et al. (2015) used integrated CNN, sparse coding, and SVM for melanoma classification. Yu et al. (2016) proposed an automatic recognition method based on DCNN and residual learning, which first segmented skin lesions and identified melanoma with two classifiers. Yu et al. (2018) proposed a network based on DCNN and used feature coding strategy to generate representative features. Xie et al. (2016) processed the incomplete inclusion of lesions in dermoscopy images and proposed a new boundary feature that can describe boundary characteristics of complete and incomplete lesions. Lai and Deng (2018) combined the extracted low-level features (color, texture) with the extracted high-level features of the convolutional neural network for classification. González-Díaz (2019) proposed a CAD system called DermaKNet to help dermatologists in their diagnosis. DermaKNet was divided into four parts, first segmenting the lesions in the dermoscopic images using the Lesion Segmentation Network (LSN), then using the segmented masks to perform data augmentation on the original data, and next the Dermoscopic Structure Segmentation Network (DSSN) was used to segment the global and local features of the image; finally, the image classification is performed using the ResNet50-based network. Xie et al. (2020) proposed MB-DCNN to perform segmentation and classification of dermoscopic images. They first used a coarse segmentation network (coarse-SN) to generate a coarse lesion mask, which was used to assist the mask-guided classification network (mask-CN) to locate and classify lesions, and the localized lesion regions were fed into the enhanced segmentation network (enhanced-SN) to obtain a fine-grained lesion segmentation map. They also proposed a new rank loss to alleviate the sample class imbalance problem. Gessert et al. (2020) proposed a patch-based attention architecture to classify high-resolution dermoscopic images, which was able to provide global contextual information to improve the accuracy of classification. In addition, they proposed a new weighting loss to address the class imbalance in the data. Zunair and Hamza (2020) first performed conditional image synthesis by learning inter-class mapping and synthesizing samples of under-represented classes from over-represented classes using unpaired image-to-image translations, thereby exploiting inter-class variation in the data distribution. Then the set of these synthetic and original data was used to train a deep convolutional neural network for skin lesion classification. Bdair et al. (2021) proposed FedPerl, a semi-supervised federated learning approach, which used peer learning and ensemble averaging to build communities and encourage their members to learn from each other so that they can generate more accurate pseudo-labels. They also proposed the peer anonymization (PA) technique as a core component of FedPerl. Datta et al. (2021) explored the goal of Soft-Attention to emphasize the value of important features and to suppress features that cause noise. Then they compared the performance of VGG, ResNet, Inception ResNet v2, and DenseNet architectures for classifying skin lesions with and without the Soft-Attention mechanism. The results showed that the Soft-Attention mechanism improved the performance of the baseline networks.

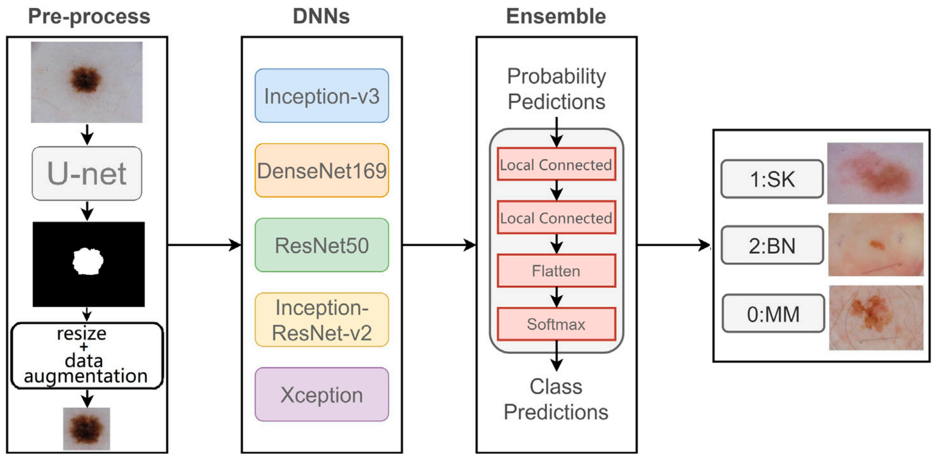


FIGURE 2 | Flowchart of our proposed model.

3 MATERIALS AND METHODS

In this section, we introduce our proposed two-stage ensemble network model. First, in the first stage, we train a segmentation network to segment skin lesions to get the lesion mask, and resize the mask area to generate lesion image with the same size. Then, in the second stage, we use five networks with good classification results on ImageNet to classify dermoscopy images, respectively. Also, we propose a new neural network to integrate the five results. The entire framework is shown in **Figure 2**.

3.1 Data Pre-Processing

The deep network model needs a large amount of training data to better fit the real data distribution, and the lack of training data may lead to over-fitting and other problems, which will seriously affect the classification ability of the model. However, most medical image datasets do not have much data, which is one of the biggest challenges of medical image analysis. Data augmentation is one of the common solutions to increase the amount of training data, and it can improve the model generalization ability. Therefore, we use different data augmentation methods on the original dataset, including rotation transform with 180°, flipping the images horizontally and vertically, and moving the image height and width direction by 10%, so that each original image generates five new samples.

3.2 Skin Lesion Segmentation

Lesion segmentation plays an important role in the automatic analysis of skin lesion. It can separate the lesion from the normal skin; therefore, the classifier can better identify the lesion features.

Unlike the classification network, which takes the images of fixed size as input and then outputs the class of each image, it gradually reduces the resolution of original images through convolution and max-pooling, and the feature maps it finally obtains are much smaller than the original image, then it classifies the feature maps through several fully connected layers. However, the output of segmentation network is the equal-sized prediction maps with input images. In the segmentation network, each pixel is a sample that needs to be classified into positive or negative. Therefore, the segmentation network needs decoder to compensate for the loss of feature resolution that is caused by max-pooling. In our experiment, we use deconvolution operation in the decoder to obtain a prediction mask with the same size as the input image.

U-net (Ronneberger et al., 2015) is an end-to-end deep convolutional neural network, which does not contain a fully connected layer, but is composed of convolution layers and up-sampling layers. U-net has an encoder and a decoder. Encoder reduces the dimension of images and extracts feature; it is composed of four blocks, each of which consists two 3 × 3 convolution layers followed by a ReLU activation function, and one max-pooling layer with stride of 2. Decoder also has four blocks, each containing a deconvolution layer, which double the size of feature maps, and two 3 × 3 convolution layers. So as for up-sampling operation in the decoder, U-net combines the output of up-sampling layer with feature map of symmetric encoder using skip-connection, so that the final output of network can consider both the shallow spatial information and deep semantic information. In this way, the outputs of the same size of the corresponding blocks in the encoder and decoder can be concatenated for segmentation and then the final prediction map is generated through a 1 × 1 convolution layer.

We train a U-net network to segment the original images and generate segmentation masks to show the lesion. These segmentation masks are used to crop the original images to help the classification network better focus on lesion features.

3.3 Skin Lesion Classification

The skin lesions have great inter-class similar visual effects; if we train our classification network to use the original images, the results will be less effective. So we divide our classification model into three stages. First, we segment skin lesions from original images using segmentation network and then resize them into a fixed size. Next, we use five classification networks with SE block to classify dermoscopy images. Finally, we construct a convolution neural network to ensemble five results.

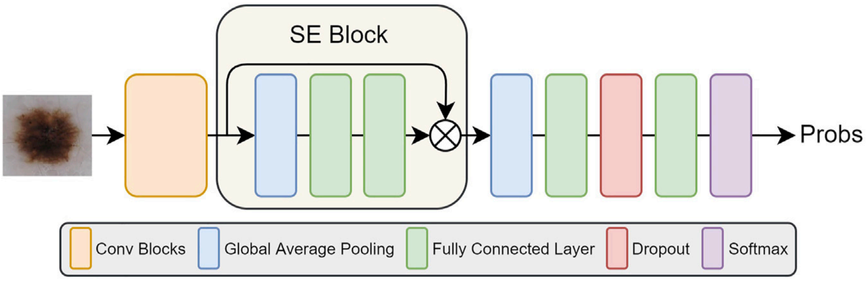


FIGURE 3 | The illustration of five network structures after adding SE Blocks.

3.4 Resize

The size of lesions varies greatly, and in most dermoscopy images, the lesion area only occupies a small part of the image, and most parts are non-lesion areas that may affect classification. In this case, if the original images are directly classified, the size of skin lesion will seriously affect the performance of network. Therefore, we first segment skin lesions from the dermoscopy images, then adjust the segmented lesion to a fixed size. Compared with the network trained on original dermoscopy images, the network trained on segmented and resized images can better extract features and has better performance.

3.5 SE Block

The features extracted by a convolutional neural network can directly affect the results of subsequent tasks, either segmentation or classification. Therefore, improving the quality of the feature representation of the network is crucial to improve the final classification results. The role of the Squeeze-and-Excitation block (Hu et al., 2018) is to further improve the classification accuracy by emphasizing the more important and informative features in the feature map. The SE block can be seen as a channel-wise attention mechanism, which emphasizes the importance of some features in the task by giving them greater weights. The specific strategy is shown in the next section that follows.

SE block is primarily concerned with the dependencies between feature channels. SE block does squeeze and excitation operation on feature maps U($H \times W \times C$). The squeeze operation includes a global average pooling; it can map feature maps to feature vectors. The *c-th* feature map can be expressed as

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

where H and W represent the height and width of feature map separately. Then the excitation operation includes two fully connected layers, a ReLU activation and sigmoid activation, so that it is able to fit complex correlations between channels by adding non-linear processing through dimensional changes. The formula can be expressed as

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where δ represents ReLU function and σ means Sigmoid, and W_1 and W_2 are the weights of the first and second fully connected layer separately. In this way, the values in this feature vector are mapped to 0, –, 1. Then the vector *s* can be multiplied as a channel descriptor with the original feature map to obtain the weighted feature map:

$$\tilde{x_c} = F_{scale}(u_c, s_c) = s_c u_c \tag{3}$$

Therefore, SE block is used to standardize feature maps according to their importance and highlight more informative feature maps, thus it can improve the network performance effectively. The schematic of adding SE Block to the five networks is shown in **Figure 3**. We add the SE Block in the same position in each network, that is, after feature extraction (orange box in **Figure 3**) and before final classification of each network.

3.6 Network Model

For ensemble problems, in addition to the ensemble method, the basic model of integration is also important. We use five state-of-the-art networks as basic network for our integration, which are Inception-v3, Densenet169, ResNet50, Inception-ResNet-v2, and Xception. These networks all have good performance on image classification tasks.

3.6.1 Inception-v3

Inception module (Szegedy et al., 2015) used 1 × 1, 3 × 3, and 5 × 5 convolution layers at the same time, then concatenated three kinds of outputs and transmitted it to the next module. In this way, it can consider information of different scales at the same time by increasing the width of the network. In addition, Inception module also can split channel-wise and spatial-wise correlation and small size of convolution kernel can greatly reduce the parameters. On the basis of Inception module, Inception-v3 (Szegedy et al. (2016)) replaced the 5 × 5 convolution layer in the original Inception network with two 3 × 3 convolution layers to further reduce the amount of parameters while maintaining the receptive field and increasing the ability of representation. Furthermore, another innovation of Inception-v3 was to decompose a large $n \times n$ convolution kernel (for example, a 7 × 7 convolution kernel) into two one-dimensional convolution kernels with the size of $n \times 1$ and $1 \times n$, respectively. This can increase the model's non-linear representation capability while reducing the risk of over-fitting.

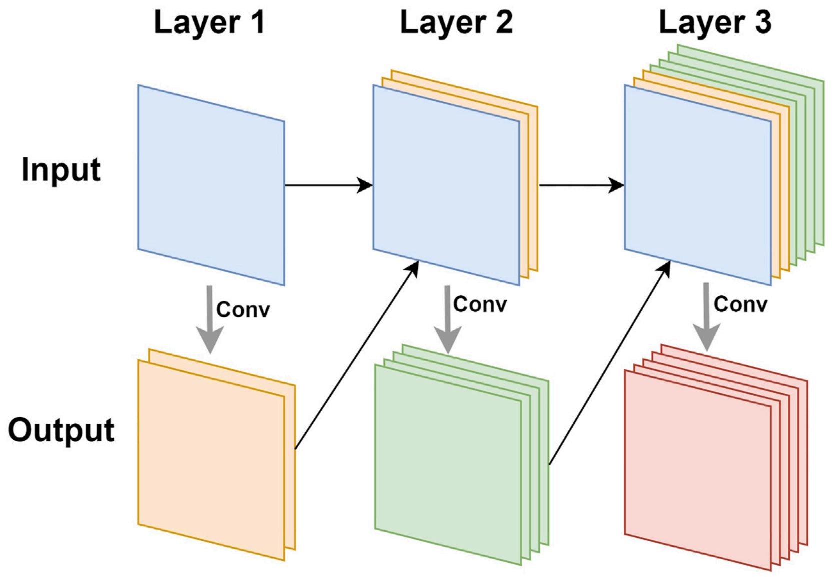


FIGURE 4 | The illustration of feature reuse of dense block.

3.6.2 ResNet-50

ResNet (He et al., 2016) appeared to alleviate the problem of vanishing/exploding gradients. ResNet was composed of a set of residual blocks, each of which is composed of several layers, including convolutional layer, ReLU layer, and batch normalization layer. Also, for each residual block, its input was directly added to its output via identity, a short connection that allowed us to perform residual learning; this is the key to solve gradient problems when training deep networks. A residual block can be formulated as

$$H_l = H_{l-1} + F(H_{l-1}) \tag{4}$$

where H_l and $H_l - 1$ are the output and input of the *l-th* residual block, respectively. F(x) represents the residual mapping function of stacked layers. It is obvious that the dimensions of $H_l - 1$ and F(H_l − 1) should be equal. However, convolution operation usually changes the dimensions, so a linear projection W_s is used to match the dimensions. So **Eq. 4** can be converted to

$$H_l = W_s H_{l-1} + F(H_{l-1}) \tag{5}$$

Therefore, ResNet-50 was obtained by stacking the residual blocks to make the final network layer count to 50.

3.6.3 Densenet169

Densenet (Huang et al., 2017) was inspired by Resnet. It also used connections to alleviate the problem of vanishing gradients, but it did not use residual blocks to achieve this goal. Densenet was composed of dense blocks. In each dense block, as shown in **Figure 4**, the input of the *n-th* layer was the result of the concatenation of all the previous *n*–1 layers. In this way, when performing related operations on the *n-th* layer, the utilization of the features of all the previous layers can be maximized. This feature reuse method can make the features work better while reducing the amount of parameters.

3.6.4 Inception-ResNet-v2

Inception-ResNet-v2 (Szegedy et al., 2017) combined Inception module with residual learning. It was based on Inception-v4, which was deeper and better than Inception-v3, but had more parameters. Inception-ResNet-v2 added residual identities to different types of Inception modules of Inception-v4, so that the network converged faster, and the training time of the network was shortened.

3.6.5 Xception

Xception (Chollet, 2017) was an improvement to Inception-v3. It mainly replaced ordinary convolution in Inception-v3 with depthwise separable convolution. The multiple convolution kernels of depthwise separable convolution only processed part of feature maps produced by the previous layer. For example, for the result of 1 × 1 convolution output from the Inception module, depthwise separable convolution referred to using three 3 × 3 convolution kernels to operate on one-third of the channel of this result, and finally three results from three 3 × 3 convolution kernels were concatenated together. In this way, the amount of parameters can be greatly reduced. Also, the author believed that Xception can decouple the channel correlation and spatial correlation of the features, thereby producing better computational results.

We use these five pre-trained networks on ImageNet as feature extractors, then add SE blocks after every extractor to emphasize more informative features. Then, a full connected layer of 128-dimension is used to generate the final feature vector, and finally we use softmax classifier to obtain class predictions.

TABLE 1 | Details of ISIC 2017 challenge dataset.

Subsets	MM	SK	BN	Total
Training	374	254	1,372	2,000
Validation	30	42	78	150
Testing	117	90	393	600

3.6.6 Ensemble Learning

There are usually two ways to ensemble multiple networks: averaging and voting. Averaging refers to the average results of multiple networks, with each network accounting for the same proportion, so that they have the same influence on the final result. However, for each class, some networks produce better results, and some have relative worse effect; taking the average directly would reduce the advantage of good networks.

For voting ensemble, we can implement it through neural networks. In detail, the neural network we build for ensemble learning is equivalent to a new classifier, whose input is the classification probabilities from five networks, and whose output is the final classification result. The reason we chose to build the classifier with locally connected layer instead of fully connected layer is that fully connected layer will be connected to all the outputs of the previous layer, while locally connected layer will only be connected to parts of the previous layer. In this case, the part of the output of the ensemble network will only be determined by a specific input, and the prediction of one class will not be influenced by the other two classes because the local connection layer extracts features for each class separately, so the network will produce more accurate classification results. This new network is used to integrate the results of the five networks, consisting of two local connected layers and a softmax layer, as shown in **Figure 2**. The result has an improvement over the averaging ensemble method.

4 RESULTS

4.1 Dataset

The dataset we use to evaluate our method was provided by ISIC 2017 challenge organized by The International Society for Digital Imaging of the Skin (Codella et al., 2018). It includes 2,750 dermoscopy images and is divided into three subsets: 2,000 for training, 150 for validation, and 600 for testing. The images in the dataset are classified as three classes: benign nevi (BN), seborrheic keratosis (SK), or melanoma (MM). The details of ISIC 2017 challenge dataset is shown in **Table 1**, MM refers to melanoma, SK refers to seborrheic keratosis, and BN refers to benign nevi. Also, we can see from **Figure 5** that the distribution of training, validation, and test sets is very uneven; the images of BN are far more than the images of the other two classes in three subsets. In addition, the ISIC 2017 dataset also provides dermoscopy images with their binary masks as their segmentation ground truth.

The ISIC 2017 challenge consists of two binary classification subtasks: melanoma or others and seborrheic keratosis or others.

4.2 Implementation

Our method is implemented with Keras on a computer with GeForce RTX 2080Ti GPU. The images with the size of 224 × 224 are taken as input of model, so all dermoscopy images are resized to 224 × 224 after segmentation. We use Adam algorithm as optimizer, and the learning rate is set as 0.0001 initially. Our epoch number is set to 100 initially. To prevent over-fitting, we use early stopping method with patience of 10 epochs.

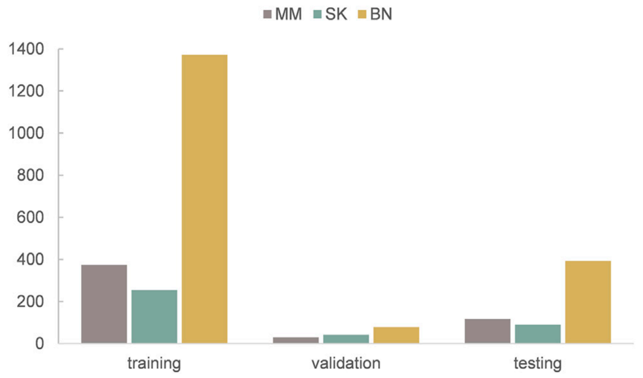


FIGURE 5 | The distribution of training, validation, and test sets of ISIC 2017 challenge dataset.

TABLE 2 | Classification results with or without segmentation.

Methods	ACC	Precision	Recall	f1 score	AUC
Without segmentation	0.698	0.598	0.622	0.592	0.781
With segmentation	0.791	0.634	0.688	0.659	0.883

4.3 Metrics

We use accuracy (ACC), recall, precision, F1-score, and AUC (area under ROC curve) as classification metrics. They are defined as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$precision = \frac{TP}{TP + FP} \tag{8}$$

$$f1score = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

where TP, TN, FP, and FN denote the number of true positive, true negative, false positive, and false negative. The number of three classes in our dataset are imbalanced, so in this case, ACC cannot well reflect the performance of our classifier; therefore, we use AUC, the same indicator as ISIC classification challenge (Codella et al., 2018), as the main metric.

4.4 Performance on Multi-Class Classification

Our method is divided into three parts. After segmenting and cropping the original dermoscopy images, five pre-trained models are used to do classification, and then the results of these models are ensembled to generate the final result. To

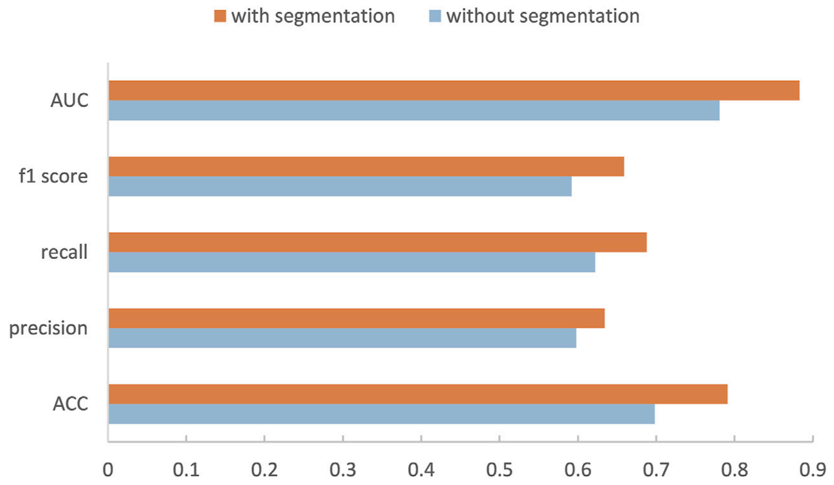


FIGURE 6 | Performance of our method with or without segmentation.

TABLE 3 | Results of different networks and two ensemble methods on multi-classification task. (The bold numbers in the table of this article are the maximum values of their columns).

Methods	ACC	Precision	Recall	f1 score	AUC
Inception-v3	0.792	0.634	0.688	0.659	0.883
Densenet169	0.800	0.739	0.727	0.722	0.881
Resnet50	0.762	0.676	0.678	0.672	0.864
Inception-Resnet-v2	0.800	0.736	0.726	0.725	0.873
Xception	0.810	0.75	**0.748**	**0.748**	0.896
Average	0.793	0.724	0.724	0.719	0.880
Ensemble	**0.851**	**0.769**	0.715	0.741	**0.913**

verify our method, in this section, we modify the dataset and convert the two binary classification tasks into a multi-classification task. Then we compare the performance with and without segmentation and resize, and the performance before and after ensemble. **Table 2** shows the experimental results with and without segmentation under one pre-trained network called Inception-v3. It can be seen that the network has better performance running on the segmented images than on the original images. As shown in **Figure 6**, especially on ACC and AUC, the results of network with segmentation get 0.791 and 0.883, respectively, which are much higher than that of network without segmentation. This is because the size of skin lesions varies greatly, and there are some interference factors such as artificial rulers in the original dermoscopy images. Segmentation can remove these interference factors to some extent, so that the network can better identify features.

In the ensemble stage, we construct a neural network model with two local connected layers with softmax classifier to fuse the results of five basic networks. Our new ensemble method can further improve the performance, and is better than the commonly used ensemble method. **Table 3** lists the results of the five pre-trained models we use and the results of averaging ensemble and our ensemble method. (The bold numbers in the table of this article are the maximum values of their columns) It can be seen that the fusion model have better performance than any single network and average method on most metrics. For the recall and f1 scores, our ensemble method is 0.033 and 0.007 lower than Xception, but it is higher than other methods in other metrics. Especially, it has a 2% improvement on AUC over the result of best network, i.e., Xception. Also, our ensemble method is better than traditional average ensemble method on all metrics except for recall.

We also compare the amount of parameters and training time of different networks (including our ensemble network). From **Table 4**, we can see that the classification networks have more parameters, especially Inception-Resnet-v2, which has up to 54.87 M. However, compared with these classification networks, our ensemble network has very few parameters, only 423. For training time, since the classification networks have been pre-trained on ImageNet, we just need to fine-tune the networks during training, and our training set is small, so we can see that the training time of each network is relatively short (when training 100 epochs). At the same time, we can also notice that the training time of the network is not entirely determined by their parameters, but is also related to the parallelism of the model and the memory access cost. In addition, these five classification networks are independent of each other, so they can be trained at the same time, which can also greatly reduce training time. Finally, our ensemble network requires very little training time, only 20 s.

4.5 Performance on Binary Classification

ISIC 2017 challenge has two binary classification tasks, melanoma or others and seborrheic keratosis or others, so we also carry out the experiment regarding challenge tasks. We show the results of melanoma classification and seborrheic keratosis classification in the form of radar diagrams, as shown in **Figure 7**. Polar

TABLE 4 | The amount of parameters and the training time of each network.

Networks	Inception-v3	Densenet169	Resnet50	Inception-resnet-v2	Xception	Ensemble
Params	22.56 M	13.22 M	24.32 M	54.87 M	21.59 M	423
Time(s)	1,900	3,200	1,900	3,000	2,700	20

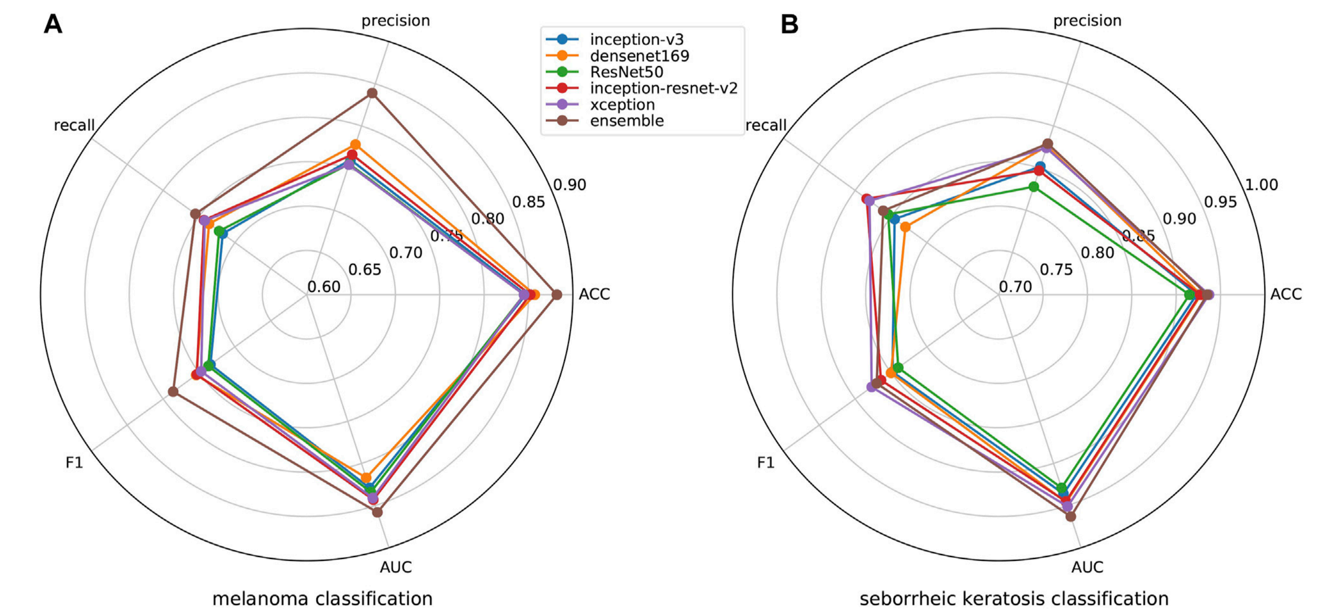


FIGURE 7 | Results of melanoma and seborrheic keratosis classification for different networks.

TABLE 5 | Average results of two skin lesion classifications of different networks.

Methods	ACC	Precision	Recall	f1 score	AUC
Inception-v3	0.885	0.806	0.781	0.791	0.883
Densenet169	0.893	0.827	0.783	0.802	0.882
Resnet50	0.88	0.792	0.788	0.789	0.882
Inception-Resnet-v2	0.89	0.807	**0.814**	0.809	0.894
Xception	0.891	0.814	0.811	0.812	0.896
SVC[1]	0.911	0.798	0.66	0.719	0.813
Random forest	**0.912**	0.802	0.664	0.721	0.816
Extra-Trees	0.911	0.805	0.65	0.716	0.809
KNN	0.908	0.782	0.657	0.709	0.81
GBDT[2]	0.91	0.808	0.644	0.71	0.807
Ensemble	0.909	**0.859**	0.808	**0.828**	**0.911**

[1] *Support Vector Classification.*
[2] *Gradient Boost Decision Tree.*

coordinates represent different metrics and each line represents a network. It can be seen that our method performs pretty well on both tasks. For the classification of melanoma, it is clear that our performance is the highest in all metrics, especially in precision, where we outperform the second highest, Densenet, by more than 10%; second, for the f1 score, which can take into account both positive and negative samples, our method also outperforms the rest of the networks by about 5%; finally, for our main metric, AUC, we also surpass the other networks by a large margin. As for the classification of seborrheic keratosis, although the advantage of our method is not as obvious as when classifying melanoma, it still performs well. First, our method still outperforms the other networks in terms of AUC, which is our main metric; second, for precision and ACC, our method leads by a small margin; and for recall and f1, we are slightly below the performance of Inception-Resnet-v2 and Xception. In general, our method is very efficient for classifying melanoma, although it is not significantly superior for classifying seborrheic keratosis, so it can improve the accuracy of classification in this task in general.

We average the performance of all networks and ensemble methods on two binary tasks and show them in **Table 5**. When compared with a single network, it can be seen that our ensemble method can effectively improve the performance; especially the AUC is 1% better than the best single network, i.e., Xception. At the same time, for precision and f1 score, our ensemble network is also the highest one. In addition, when compared with other ensemble methods, we use several machine learning classifier to do ensemble as comparison. We can see that except that ACC is 0.003 lower than Random forest, we are significantly better than machine learning methods on other metrics. We also illustrate this comparison in **Figure 8**, so we can more intuitively see the advantages of our ensemble method in various metrics.

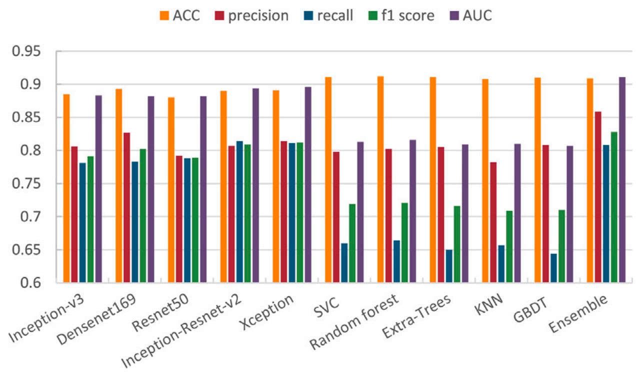


FIGURE 8 | Comparison of different methods on skin lesion classification.

TABLE 6 | Comparison among our method, some existing methods, and the top five ISIC2017 classification challenge.

Method	ACC	Precision	Recall	f1 score	AUC
Top 1	0.816	0.748	0.856	**0.851**	0.911
Top 2	0.849	0.747	0.140	0.236	0.910
Top 3	0.883	0.752	0.451	0.564	0.908
Top 4	0.888	0.732	0.508	0.600	0.896
Top 5	0.873	0.665	0.568	0.613	0.886
Zhang et al. (2019)	0.868	—	0.878	—	0.958
González-Díaz (2019)	—	—	—	—	0.917
Xie et al. (2020)	0.904	—	0.786	—	0.938
Datta et al. (2021)	0.833	—	**0.916**	—	**0.959**
Ours	**0.909**	**0.859**	0.808	0.828	0.911

4.6 Comparison of Various Predictors

In **Table 6**, we compare our method with the top five performance in the ISIC 2017 challenge skin lesion classification task (Díaz, 2017; Matsunaga et al., 2017; Bi et al., 2017; Menegola et al., 2017; Yang et al., 2017) and some excellent methods in recent years. Most of the networks participating in the challenge used external images, which we do not do. In **Table 6**, it can be seen that our method achieves 0.909 and 0.859 on ACC and precision, which are highest on these metrics. Besides, we get 0.911 on AUC, which is 0.048 lower than that of Datta et al. (2021). For f1 score, our method obtains 0.828, which is 0.023 lower than the best score. However, for recall, our model's performance is a bit unsatisfactory, which shows that our model still has some shortcomings in classifying positive samples.

5 CONCLUSION

In this paper, we have the following innovations: 1) we propose a new two-stage ensemble method that integrates five excellent classification models to classify skin melanoma; 2) we also propose a new method of segmenting the lesion area of the dermoscopy image to generate a mask of the lesion area, so that the image can be resized to focus on the lesion; 3) we propose a new ensemble network that can use local connected layers to effectively integrate the classification results from the five classification networks. We test our method on the ISIC 2017 challenge dataset and get pretty good results. In future work, we will explore more effective classification methods based on the characteristics of dermoscopy images and the association of different classes of dermoscopy images, especially in process of pre-processing, because the experimental results show that our segmented images can largely improve the accuracy of classification.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study can be accessed at https://github.com/guofei-tju/Melanoma_cls. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

JD and JS conceived and designed the experiments. JD and JL performed the experiments and analyzed the data. JD and FG wrote the article. FG and JT supervised the experiments and reviewed the article. All authors have participated in study discussion and article preparation.

FUNDING

This work is supported by a grant from the National Natural Science Foundation of China (NSFC 62 172 296, 61 972 280) and the National Key R&D Program of China (2020YFA0908400). This work was also supported by the Shenzhen Science and Technology Program (No. KQTD20200820113106007).

REFERENCES

Barata, C., Ruela, M., Francisco, M., Mendonça, T., and Marques, J. S. (2013). Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features. *IEEE Syst. J.* 8, 965–979.

Bdair, T. M., Navab, N., and Albarqouni, S. (2021). Peer Learning for Skin Lesion Classification. CoRR abs/2103.03703.

Bi, L., Kim, J., Ahn, E., and Feng, D. (2017). Automatic Skin Lesion Analysis Using Large-Scale Dermoscopy Images and Deep Residual Networks. arXiv preprint arXiv:1703.04197.

Cao, Z., Sun, C., Wang, W., Zheng, X., Wu, J., and Gao, H. (2021). Multi-modality Fusion Learning for the Automatic Diagnosis of Optic Neuropathy. *Pattern Recognition Lett.* 142, 58–64. doi:10.1016/j.patrec.2020.12.009

Capdehourat, G., Corez, A., Bazzano, A., Alonso, R., and Musé, P. (2011). Toward a Combined Tool to Assist Dermatologists in Melanoma Detection from Dermoscopic Images of Pigmented Skin Lesions. *Pattern Recognition Lett.* 32, 2187–2196. doi:10.1016/j.patrec.2011.06.015

Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., et al. (2007). A Methodological Approach to the Classification of Dermoscopy Images. *Comput. Med. Imaging graphics* 31, 362–373. doi:10.1016/j.compmedimag.2007.01.003

Chen, J., Ying, H., Liu, X., Gu, J., Feng, R., Chen, T., et al. (2020a). A Transfer Learning Based Super-resolution Microscopy for Biopsy Slice Images: The Joint Methods Perspective. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18, 1. doi:10.1109/TCBB.2020.2991173

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach Intell.* 40, 834–848. doi:10.1109/TPAMI.2017.2699184

Chen, T., Liu, X., Feng, R., Wang, W., Yuan, C., Lu, W., et al. (2021b). Discriminative Cervical Lesion Detection in Colposcopic Images with Global Class Activation and Local Bin Excitation. *IEEE J. Biomed. Health Inform.* 1, 1. doi:10.1109/JBHI.2021.3100367

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1251–1258 .

Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., and Smith, J. R. (2015).Deep Learning, Sparse Coding, and Svm for Melanoma Recognition in Dermoscopy Images. In International workshop on machine learning in medical imaging. Springer, 118–126. doi:10.1007/978-3-319-24888-2_15

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2018).Skin Lesion Analysis toward Melanoma Detection: A challenge at the 2017 International Symposium on Biomedical Imaging (Isbi), Hosted by the International Skin Imaging Collaboration (Isic). In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 168–172.

Díaz, I. G. (2017). Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions. arXiv preprint arXiv:1703.01976.

Dai, J., He, K., and Sun, J. (2016). Instance-aware Semantic Segmentation via Multi-Task Network Cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3150–3158 .

Datta, S. K., Shaikh, M. A., Srihari, S. N., and Gao, M. (2021). Soft-attention Improves Skin Cancer Classification Performance .

Feng, R., Liu, X., Chen, J., Chen, D. Z., Gao, H., and Wu, J. (2021). A Deep Learning Approach for Colonoscopy Pathology Wsi Analysis: Accurate Segmentation and Classification. *IEEE J. Biomed. Health Inform.* 25, 3700–3708. doi:10.1109/JBHI.2020.3040269

Ganster, H., Pinz, P., Rohrer, R., Wildling, E., Binder, M., and Kittler, H. (2001). Automated Melanoma Recognition. *IEEE Trans. Med. Imaging* 20, 233–239. doi:10.1109/42.918473

Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., et al. (2020). Skin Lesion Classification Using Cnns with Patch-Based Attention and Diagnosis-Guided Loss Weighting. *IEEE Trans. Biomed. Eng.* 67, 495–503. doi:10.1109/TBME.2019.2915839

González-Díaz, I. (2019). Dermaknet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis. *IEEE J. Biomed. Health Inform.* 23, 547–559. doi:10.1109/JBHI.2018.2806962

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778 .

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1904–1916. doi:10.1109/tpami.2015.2389824

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7132–7141 .

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4700–4708 .

Kittler, H., Pehamberger, H., Wolff, K., and Binder, M. (2002). Diagnostic Accuracy of Dermoscopy. *Lancet Oncol.* 3, 159–165. doi:10.1016/s1470-2045(02)00679-4

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, 1097–1105.

Lai, Z., and Deng, H. (20182018). Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer Perceptron? *Comput. Intelligence Neurosci.*

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016).Ssd: Single Shot Multibox Detector. In European conference on computer vision. Springer, 21–37. doi:10.1007/978-3-319-46448-0_2

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3431–3440 .

Matsunaga, K., Hamada, A., Minagawa, A., and Koga, H. (2017). Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble. arXiv preprint arXiv:1703.03108.

Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., Avila, S., and Valle, E. (2017). Recod Titans at Isic challenge 2017. arXiv preprint arXiv:1703.04819.

Mishra, N. K., and Celebi, M. E. (2016). An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning. arXiv preprint arXiv:1601.07843.

Moura, N., Veras, R., Aires, K., Machado, V., Silva, R., Araújo, F., et al. (2019). Abcd Rule and Pre-trained Cnns for Melanoma Diagnosis. *Multimed Tools Appl.* 78, 6869–6888. doi:10.1007/s11042-018-6404-8

Myronenko, A. (2018).3d Mri Brain Tumor Segmentation Using Autoencoder Regularization. In International MICCAI Brainlesion Workshop. Springer, 311–320.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look once: Unified, Real-Time Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 779–788 .

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical image computing and computer-assisted intervention. Springer, 234–241. doi:10.1007/978-3-319-24574-4_28

Roychowdhury, S., Koozekanani, D. D., and Parhi, K. K. (2015). Iterative Vessel Segmentation of Fundus Images. *IEEE Trans. Biomed. Eng.* 62, 1738–1749. doi:10.1109/tbme.2015.2403295

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: Cnn Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* 35, 1285–1298. doi:10.1109/tmi.2016.2528162

Siegel, R. L., Miller, K. D., and Jemal, A. (2016). Cancer Statistics, 2016. *CA: a Cancer J. clinicians* 66, 7–30. doi:10.3322/caac.21332

Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

Stolz, W. (1994). Abcd Rule of Dermatoscopy: a New Practical Method for Early Recognition of Malignant Melanoma. *Eur. J. Dermatol.* 4, 521–527.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In Thirty-first AAAI conference on artificial intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going Deeper with Convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1–9 .

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2818–2826 .

Xiao, J., Xu, H., Gao, H., Bian, M., and Li, Y. (2021). A Weakly Supervised Semantic Segmentation Network by Aggregating Seed Cues: The Multi-Object Proposal Generation Perspective. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 1–19. doi:10.1145/3419842

Xie, F., Fan, H., Li, Y., Jiang, Z., Meng, R., and Bovik, A. (2016). Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model. *IEEE Trans. Med. Imaging* 36, 849–858. doi:10.1109/TMI.2016.2633551

Xie, Y., Zhang, J., Xia, Y., and Shen, C. (2020). A Mutual Bootstrapping Model for Automated Skin Lesion Segmentation and Classification. *IEEE Trans. Med. Imaging* 39, 2482–2493. doi:10.1109/TMI.2020.2972964

Yang, X., Zeng, Z., Yeo, S. Y., Tan, C., Tey, H. L., and Su, Y. (2017). A Novel Multi-Task Deep Learning Model for Skin Lesion Segmentation and Classification. arXiv preprint arXiv:1703.01025.

Yu, L., Chen, H., Dou, Q., Qin, J., and Heng, P. A. (2017). Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Trans. Med. Imaging* 36, 994–1004. doi:10.1109/TMI.2016.2642839

Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., et al. (2018). Melanoma Recognition in Dermoscopy Images via Aggregated Deep Convolutional Features. *IEEE Trans. Biomed. Eng.* 66, 1006–1016. doi:10.1109/TBME.2018.2866166

Zhang, J., Xie, Y., Xia, Y., and Shen, C. (2019). Attention Residual Learning for Skin Lesion Classification. *IEEE Trans. Med. Imaging* 38, 2092–2103. doi:10.1109/TMI.2019.2893944

Zunair, H., and Ben Hamza, A. (2020). Melanoma Detection Using Adversarial Training and Deep Transfer Learning. *Phys. Med. Biol.* 65, 135005. doi:10.1088/1361-6560/ab86d3

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.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 24 January 2022
doi: 10.3389/fbioe.2021.819991

Chronological Age Prediction: Developmental Evaluation of DNA Methylation-Based Machine Learning Models

Haoliang Fan[*†], *Qiqian Xie*[†], *Zheng Zhang, Junhao Wang, Xuncai Chen*[*] *and Pingming Qiu*[*]

Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Wenbing Zhao,
Cleveland State University,
United States

Reviewed by:
Rashed Alghafri,
Dubai Police, United Arab Emirates
Shaoqing Wen,
Fudan University, China
Cemal Gurkan,
Turkish Cypriot DNA Laboratory
(TCDL), Cyprus

***Correspondence:**
Haoliang Fan
18976431677fan@smu.edu.cn
Xuncai Chen
xche3815@smu.edu.cn
Pingming Qiu
qiupm@smu.edu.cn

[†]*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Biosafety and Biosecurity, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 22 November 2021
Accepted: 21 December 2021
Published: 24 January 2022

Citation:
Fan H, Xie Q, Zhang Z, Wang J, Chen X and Qiu P (2022) Chronological Age Prediction: Developmental Evaluation of DNA Methylation-Based Machine Learning Models. Front. Bioeng. Biotechnol. 9:819991. doi: 10.3389/fbioe.2021.819991

Epigenetic clock, a highly accurate age estimator based on DNA methylation (DNAm) level, is the basis for predicting mortality/morbidity and elucidating the molecular mechanism of aging, which is of great significance in forensics, justice, and social life. Herein, we integrated machine learning (ML) algorithms to construct blood epigenetic clock in Southern Han Chinese (CHS) for chronological age prediction. The correlation coefficient (*r*) meta-analyses of 7,084 individuals were firstly implemented to select five genes (*ELOVL2*, *C1orf132*, *TRIM59*, *FHL2*, and *KLF14*) from a candidate set of nine age-associated DNAm biomarkers. The DNAm-based profiles of the CHS cohort (240 blood samples differing in age from 1 to 81 years) were generated by the bisulfite targeted amplicon pyrosequencing (BTA-pseq) from 34 cytosine-phosphate-guanine sites (CpGs) of five selected genes, revealing that the methylation levels at different CpGs exhibit population specificity. Furthermore, we established and evaluated four chronological age prediction models using distinct ML algorithms: stepwise regression (SR), support vector regression (SVR-eps and SVR-nu), and random forest regression (RFR). The median absolute deviation (*MAD*) values increased with chronological age, especially in the 61–81 age category. No apparent gender effect was found in different ML models of the CHS cohort (all $p > 0.05$). The *MAD* values were 2.97, 2.22, 2.19, and 1.29 years for SR, SVR-eps, SVR-nu, and RFR in the CHS cohort, respectively. Eventually, compared to the *MAD* range of the meta cohort (2.53–5.07 years), a promising RFR model (*ntree* = 500 and *mtry* = 8) was optimized with an *MAD* of 1.15 years in the 1–60 age categories of the CHS cohort, which could be regarded as a robust epigenetic clock in blood for age-related issues.

Keywords: DNA methylation, CpG, chronological age prediction, machine learning, stepwise regression, support vector regression, random forest regression, epigenetic clock

INTRODUCTION

Aging is an inevitable, universal and natural phenomenon that occurs with age, characterized by progressive decline in organismal function and more susceptible to irreversible degenerative disease and even death (Sen et al., 2016). Accumulating studies have linked aging to epigenetic alterations (Grönniger et al., 2010; Sen et al., 2016; Horvath and Raj, 2018). As such, aging denotes an elementary epigenetic phenomenon, and epigenetic changes are widely considered to play a crucial role in aging (Fraga et al., 2005; Boks et al., 2009). Epigenetics is often defined by changes in gene function that do not involve any changes in DNA sequence, and epigenetic changes during aging mainly include histone modification and DNA methylation (DNAm) (Parson, 2018; Unnikrishnan et al., 2019).

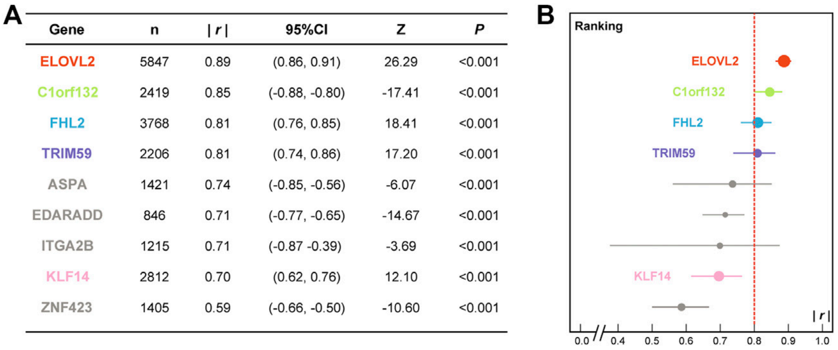
DNAm is a chemical modification that mainly occurs in cytosine-phosphate-guanine (CpG) loci, especially in the CpG islands. In fact, an initial study of age-associated methylation in normal tissue was motivated by the study of methylation in cancer (Esteller, 2002). Cancer is well recognized as a disease of aging. For example, Christensen et al. verified this by proposing that variations in age- and exposure-related methylation may significantly contribute to increased susceptibility to several diseases (Christensen et al., 2009). Emerging studies are beginning to work on the associations between methylation profiles and human tissues; however, most of them have focused on therapeutic targets for pathological tissues (Suzuki et al., 2006; Portela and Esteller, 2010; Gao et al., 2019).

In forensics, DNAm biomarkers mainly focus on normal tissues, and employing methylation levels of strongly age-related CpGs (AR-CpGs) into construction of age predictive models has become a mainstream of age-estimation strategies (i.e., epigenetic clock) (Horvath and Raj, 2018). Epigenetic clock, which measures alterations in specific CpGs, is a synonym of a highly accurate age estimator based on DNAm levels (Unnikrishnan et al., 2019). As the most promising molecular age estimator, epigenetic clock can not only accurately predict age, mortality, or morbidity but also help to disentangle the role of DNAm in the mechanisms of aging, therefore facilitating anti-aging interventions (Jylhävä et al., 2017; Horvath and Raj, 2018; Unnikrishnan et al., 2019). Moreover, the epigenetic clocks can be utilized in other non-clinical areas, such as 1) forensic DNA phenotyping, including scenes in criminal investigation or catastrophic disaster (Gršković et al., 2013; Vidaki et al., 2013; Parson, 2018); 2) potentially determination of age of criminal responsibility for judgement (Gršković et al., 2013); and 3) children and youth growth monitoring, athlete selection, and social welfare recognition in our social life (Weidner et al., 2014).

To date, even though the relationship between aging and CpG methylation is complicated (Tra et al., 2002), large series of AR-CpGs were applicable for age prediction from methylation analysis, and quite a few epigenetic clocks of different populations were generated, providing references for distinct forensic scenarios. For example, Hannum et al. (2013) identified 71 AR-CpGs using the Illumina Infinium HumanMethylation450 BeadChip assay and built an age calculator with a correlation of 96% and a median absolute deviation (*MAD*) value of 3.9 years. Naue et al. chose 15 AR-CpGs for methylation analysis using the massive parallel sequencing method and proposed a regression model with an *MAD* value of 3.21 years (Naue et al., 2017). Smeers et al. investigated 16 AR-CpGs by pyrosequencing method and constructed three statistical prediction models with *MAD* values of 3.21, 3.20, and 3.26 years, respectively (Smeers et al., 2018). Dias et al. tested 5 AR-CpGs using the multiplex SNaPshot assay and developed an age prediction model based on 4 of them, with an *MAD* value of 4.97, which explains 92.5% variation in age (Dias et al., 2020).

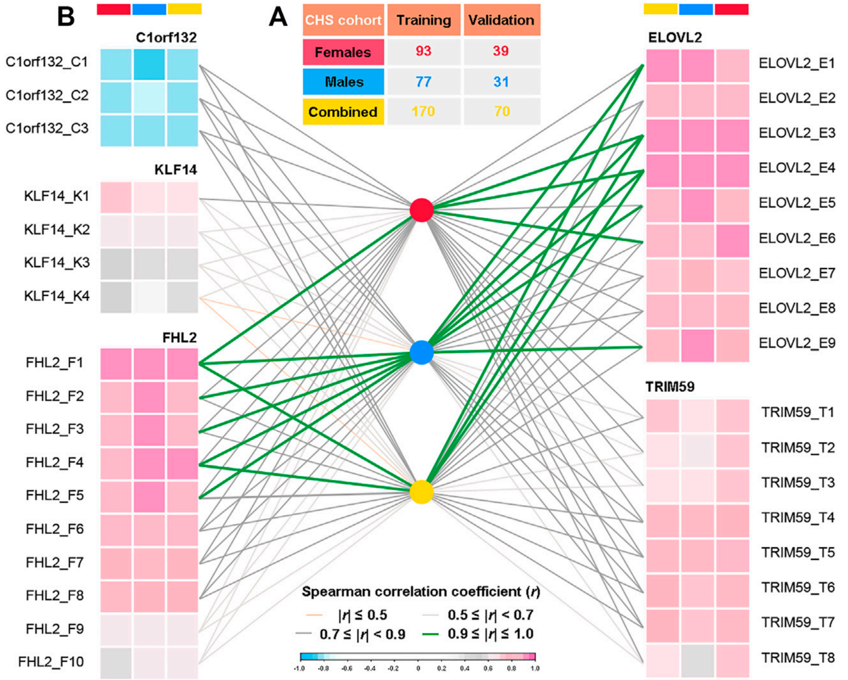
As mentioned above, the *MAD* values for most DNAm-based age prediction models were more than 3 years (Zbieć-Piekarska et al., 2015b; Cho et al., 2017; Naue et al., 2017; Vidaki et al., 2017; Aliferi et al., 2018; Smeers et al., 2018; Dias et al., 2020), and also many factors have influences on age prediction accuracy, which limited its practical application. For example, different human body fluids (blood, semen, saliva, etc.) exhibit distinct methylation patterns (Jung et al., 2019), and in different populations/genders, the same DNAm biomarkers show diverse methylation levels in the same age category (Zbieć-Piekarska et al., 2015b; Cho et al., 2017; Dias et al., 2020). In addition, there are various alternative approaches (genome-wide DNAm, Illumina BeadChip, bisulfite pyrosequencing, etc.) for DNAm detection, while the bisulfite targeted amplicon pyrosequencing (BTA-pseq) technology supports standardized and cost-effective high-throughput analysis, which is generally relatively accurate. Except for the selection of population-/gender-/tissue-specific DNAm biomarkers and detection methods, the algorithm also has an impact on the age-prediction accuracy. Aliferi et al. (2018) compared the efficiency of 17 machine learning (ML) models based on the same MPS data and suggested that multiple linear regression (MLR) models did not outperform the generalized regression neural network (GRNN) model and several non-linear approaches showed increased accuracy, especially for support vector machine polynomial (SVMp). Xu et al. (2015) found that the *MAD* values reduced in the models of nonlinear regression, BP neural network, and support vector regression (SVR) by using the same CpGs when comparing with the MLR model. Garali et al. compared six different statistical models with the MLR model of Zbiec-Pierkarska (Zbieć-Piekarska et al., 2015b), and the results suggested that multiple quadratic regression (MQR), SVM, gradient boosting regressor (GBR), and MissMDA (mMDA) models outperformed the MLR model for age prediction from *ELOVL2* (Garali et al., 2020).

Hence, in order to establish robust age prediction ML models for Southern Han Chinese (CHS), a candidate set of nine DNAm biomarkers was collected by meta-analyses of 7,084 individuals. Among them, five promising age-related genes (34 CpGs) were selected according to the correlation coefficient (r) ranking and Gene Expression Omnibus (GEO) data mining by AgeGuess (Gao et al., 2020). The DNAm-based profiles of the CHS cohort (240 blood samples with ages of 1–81 years) were generated by BTA-pseq. In addition, four different ML algorithms, stepwise regression (SR), SVR (including eps- and nu-regression), and random forest regression (RFR), were used to establish the age-prediction models based on AR-CpGs ($|r|\geq 0.7$). The samples were randomly divided into different datasets according to different genders and chronological ages, and we evaluated the model efficiencies in Training and Validation sets by *MAD* and root mean square error (*RMSE*) values, to find the best-performing ML model of CHS to estimate the chronological ages in practice.



Gene	n	|r|	95%CI	Z	P
ELOVL2	5847	0.89	(0.86, 0.91)	26.29	<0.001
C1orf132	2419	0.85	(-0.88, -0.80)	-17.41	<0.001
FHL2	3768	0.81	(0.76, 0.85)	18.41	<0.001
TRIM59	2206	0.81	(0.74, 0.86)	17.20	<0.001
ASPA	1421	0.74	(-0.85, -0.56)	-6.07	<0.001
EDARADD	846	0.71	(-0.77, -0.65)	-14.67	<0.001
ITGA2B	1215	0.71	(-0.87 -0.39)	-3.69	<0.001
KLF14	2812	0.70	(0.62, 0.76)	12.10	<0.001
ZNF423	1405	0.59	(-0.66, -0.50)	-10.60	<0.001

FIGURE 1 | Detailed meta-analysis results **(A)** and correlation coefficient ranking **(B)** of the candidate age-associated gene set. (*n*, sample size; |*r*|, absolute value of correlation coefficient; CI, confidence interval; *p*, significance of Z test.)



CHS cohort	Training	Validation
Females	93	39
Males	77	31
Combined	170	70

FIGURE 2 | Spearman correlation analyses between DNA methylation levels of 34 CpGs located at five genes and chronological ages of three different datasets in the CHS cohort (*n* = 240, blood samples). **(A)** Detailed population sizes of different datasets (in the CHS cohort, randomly 70%/30% for Training and Validation sets, detailed information in **Supplementary Table S4**). **(B)** Spearman correlations between chronological ages and DNA methylation levels at each CpG in three different gender datasets (*r*, correlation coefficient; 0.9≤|*r*|≤1.0, very high correlation; 0.7≤|*r*|<0.9, high correlation; 0.5≤|*r*|<0.7, moderate correlation; |*r*|≤0.5, low correlation, details in **Supplementary Table S5**).

MATERIALS AND METHODS

AR-CpG Selection and Sample Collection

The bibliographic search strategies were developed according to the DNAm-based age prediction studies with *MAD* values less than 5 years between 2014 and 2021, and we collected a cohort of 7,084 individuals from 16 countries or populations (Weidner et al., 2014; Bekaert et al., 2015; Xu et al., 2015; Zbieć-Piekarska et al., 2015a; Zbieć-Piekarska, et al., 2015b; Park et al., 2016; Zubakov et al., 2016; Cho et al., 2017; Feng et al., 2018; Alsaleh and Haddrill, 2019; Daunay et al., 2019; Jung et al., 2019; Li et al., 2019; Dias et al., 2020; Garali et al., 2020; Lau and Fung, 2020; Pan et al., 2020; Piniewska-Róg et al., 2021; Sukawutthiya et al., 2021; Woźniak et al., 2021; Xiao et al., 2021). The correlation coefficient (*r*) ranking of nine age-associated genes was obtained by meta-analyses (**Figure 1** and **Supplementary Figure S1**). We selected four promising DNAm biomarkers (*ELOVL2*, *C1orf132*, *FHL2*, and *TRIM59*) according to the correlation coefficient ranking (|*r*|≥0.8) and the *KLF14* gene by GEO data mining using a three-step feature selection algorithm AgeGuess (Gao et al., 2020), including a total of 34 CpGs (details in **Supplementary Table S1**). The PCR primers of five age-related DNAm biomarkers (**Supplementary Table S2**) were designed by PyroMark Assay Design Software 2.0 (Qiagen, Hilden, Germany).

A total of 240 unrelated healthy individuals were recruited from Han Chinese, who had settled in south China for at least three generations. Peripheral blood samples (2 ml) and accurate information (including age, gender, nationality) were collected from all participants of the CHS cohort. All volunteers had signed the informed consent forms (the underage children were signed by their guardians in accordance with Chinese laws and regulations), and the study was approved by the Biomedical Ethical Committee of Southern Medical University (No. 2021-015) following the standards of Declaration of Helsinki.

Sample Preparation and BTA-pseq

DNA Extraction and Quantification

Genomic DNA was extracted from 200 µl peripheral blood by QIAamp Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The extracted DNA samples were then quantified using Qubit® 4.0 Fluorometer instrument (Thermo Fisher Scientific, Waltham, MA, United States) with Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, United States) according to the manufacturer's instructions.

Bisulfite Conversion

The conversion of unmethylated cytosines to uracils in DNA samples was carried out with the EpiTect Fast DNA Bisulfite Kit (Qiagen, Hilden, Germany), following the manufacturer's instruction. With the input of 300 ng DNA, the bisulfite DNA conversion was performed using a thermal cycler that comprised: two cycles of initial denaturation at 95°C for 5 min and incubation at 60°C for 10 min followed by a hold at 20°C for up to 20 h in the thermal cycler. The converted DNA was then eluted into 15 µl of the elution buffer (EB) obtained from the same kit, normalized to 20 ng/µl as the DNA template, and subsequently stored at −20°C until use.

Targeted Amplicon PCR

After bisulfite conversion, 100 ng of each converted DNA was submitted into a multiplex polymerase chain reaction (PCR) amplification with PyroMark PCR Kit (Qiagen, Hilden, Germany). Each multiplex reaction was performed in a final volume of 25 µl containing 12.5 µl of 2× PyroMark PCR Master Mix (providing a concentration of 1.5 mM $MgCl_2$), 2.5 µl of 10× CoralLoad Concentrate, 9 µl of primer mix, and 1 µl of template DNA. The multiplex reaction was amplified under the following conditions: 1) initial PCR activation at 95°C for 15 min; 2) 45 cycles consisting of denaturation at 94°C for 30 s, annealing at 56°C for 30 s, and extension at 72°C for 30 s; and 3) final extension at 72°C for 10 min followed by a hold at 4°C. Negative control without DNA template was prepared in each PCR process.

Pyrosequencing

Following amplification, all PCR products were sequenced using PyroMark Gold Q24 Reagents (Qiagen, Hilden, Germany) in combination with PyroMark Q24 platform (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The generated pyrogram traces with sharp and distinct peaks were subsequently analyzed, and the methylation levels at different CpGs were calculated by the peak heights observed in PyroMark Q24 Advanced software v3.0.1 (Qiagen, Hilden, Germany). The missing methylation percentage values have been filled in with the median (**Supplementary Table S3**).

Statistical Analysis

Spearman Correlation

The Spearman correlation coefficient (*r*) was calculated by IBM® SPSS® Statistics 26 (IBM Corporation, Armonk, NY, United States), SAS® 9.4 software (SAS Institute Inc., Cary, NC, United States), and R (version 3.6.1). The *r* values are used to assess the strength and direction of the linear relationships between pairs of variables (predicted and chronological ages). According to Mukaka (2012), the *r* values followed the rule of thumb for interpreting size of a correlation coefficient: 1) $0.9 \le |r| \le 1.0$, very high correlation; 2) $0.7 \le |r| < 0.9$, high correlation; 3) $0.5 \le |r| < 0.7$, moderate correlation; and 4) $|r| \le 0.5$, low correlation. The AR-CpGs ($|r| \ge 0.7$) were selected to establish different ML models.

Dataset Information

As shown in **Figure 2A**, the CHS cohort was randomly divided into a Training set (70%, *n* = 170, 93 females and 77 males) and a Validation set (30%, *n* = 70, 39 females and 31 males). The obtained methylation levels of Training set and the corresponding chronological ages were used for model training. Parameter tunning was performed by leave-one-out (*k*-fold) cross-validation, during which a set of samples (a fold) is removed from the dataset as the Validation set and the remaining samples were assigned as a Training set. In addition, for the evaluation of gender differences and aging effects, both Training and Validation sets were divided into three different gender datasets (female, male, and combined datasets, details in **Figure 2A**) and four age categories (1–20, 21–40, 41–60, and 61–81 years, details in **Supplementary Table S4**), respectively.

Model Performance Comparison

Model performance was compared in terms of *MAD* and *RMSE* values, which are calculated by IBM® SPSS® Statistics 26 and R (version 3.6.1). The *MAD* value is defined as the average distance between each data value and the mean, a way to describe variation in a dataset, while the *RMSE* value is widely used to compute the error distance between the estimated values. Both of them are the main metrics used to measure the quality of the regression output models. To measure the overall performance of each model, the *MAD* and *RMSE* values were calculated for the whole CHS cohort. Subsequently, to evaluate the generalization and the actual prediction performance of the final model, and to evaluate gender or aging effects, *MAD* values for different datasets needed to be analyzed.

Machine Learning Model

Stepwise Regression Model

For multivariate linear regression analysis, the model selection procedure SR was performed using IBM® SPSS® Statistics 26 (IBM Corporation, Armonk, NY, United States) for model

TABLE 1 | Stepwise regression (SR) equations and system efficiencies in three different datasets of the CHS cohort (*n* = 240, blood samples).

Dataset	SR equation	R^2	Adjusted R^2	*RMSE*	*MAD*
Females	y = 35.518 + 0.679×F1–0.317×C1+0.319×T2–0.241×C2+0.438×E2+0.170×T4–0.202×F4+0.124×K1	0.94	0.93	4.07	3.00
Males	y = 21.347 + 0.488×E1–0.412×C1+0.360×F5+0.125×E7+0.320×E5	0.96	0.96	3.45	2.64
Combined	y = 24.260 + 0.348×F1–0.463×C1+0.188×E3+0.151×E1+0.088×T4+0.315×E2–0.260×F4+0.222×F2+ 0.054×E7+0.125×T5	0.95	0.94	3.89	2.97

R^2, coefficient of determination/goodness-of-fit; Adjusted R^2, adjusted coefficient of determination; RMSE, root mean square error; MAD, median absolute deviation.

building together with 0.05 significance criteria for inclusion in the final model. Specifically, by excluding all previously selected variables with a *p*-value of 0.05 or greater until no variables can be eliminated nor new variables can be introduced in the regression equation, *stepwiselm* can create a linear model and automatically add to or trim the model, thus improving the selection of important variables in relatively small datasets (Núñez et al., 2011). Overall, the essence of these steps is to establish an "Optimal" MLR equation. The accuracy of age prediction with those tested CpGs was assessed by the goodness-of-fit (R^2), which is a parameter establishing the discrepancy between the observed values (chronological ages) and the expected values (predicted ages) under an applicable model, and generally used in regression to evaluate the performance of the model. Therefore, model equations with the greatest R^2 were selected as the candidate predictors based on the multivariant regression analysis.

Support Vector Regression Model

For SVR analysis, SVR model was carried out by R (*e1071* package). As reported, support vector machine (SVM) is a powerful technique for classification, regression, and outlier detection, and a correct choice of kernel parameters is crucial for a promising result. So, we constructed and refined regression models by following methods: 1) select support vector machines with radial (SVMr) function as kernel, 2) employ eps-regression and nu-regression for comparison, and 3) adjust the parameters "cost, gamma, and epsilon" for eps-regression and "cost and nu" for nu-regression. Eventually, two optimized SVR models with best-performing parameters were obtained.

Random Forest Regression Model

For random forest regression analysis, random forest exploiting classification trees were constructed based on Breiman's random forest algorithm (on the basis of Breiman and Cutler's original Fortran code) using *randomForest* R package. Random forests represent an effective tool in prediction, and RFR algorithm that based on decision trees plays an important role in selecting the "optimal" markers for model building. To reduce bias and operate effectively in regression, optimization of the RFR model was carried out by tuning the parameters *mtry* and *ntree*. *mtry* refers to the number of variables randomly sampled as candidates at each split, and *ntree* is defined as the number of trees to grow. By multiple rounds of optimization, a final *mtry* of 8 was chosen, the *ntree* was set at 500, and the optimal RFR model was established. The value (% Var explained) represents the overall explanatory rate for the variances of the response variables by the predictive variables. We used the value (% IncMSE, increase in mean squared error) to measure the importance of predictive variables, which means that by randomly assigning a value to each predictive variable, if the predictive variable is more important, the model prediction error will increase after its value is randomly replaced.

RESULTS

AR-CpG Selection and Spearman Correlation

At first, a cohort of 7,084 individuals from 16 countries or populations related to DNAm-based age prediction studies was collected by bibliographic search to conduct meta-analyses (details in **Supplementary Figure S1**). **Figure 1A** presents the results of a meta-analysis of the detailed correlation coefficients for candidate age-associated genes in the meta cohort. The absolute values of correlation coefficients ($|r|$) for nine DNAm biomarkers ranged from 0.59 (ZNF423) to 0.89 (*ELOVL2*). There are eight of nine DNA biomarkers with $|r|≥0.7$ (**Figure 1**), and the $|r|$ ranking of the candidate genes is visualized in **Figure 1B**. According to the self-defined threshold value ($|r|≥0.8$), four promising genes (*ELOVL2*, *C1orf132*, *FHL2*, and *TRIM59*) were selected for further validation in the CHS cohort. In addition, the *KLF14* gene that was screened by a three-step feature selection algorithm AgeGuess (Gao et al., 2020) was also selected. **Supplementary Tables S1, S2** present the detailed 34 CpGs and PCR primers of five aforementioned DNAm biomarkers, respectively.

The detailed DNAm levels of 34 CpGs and the corresponding personal information (chronological ages and genders) in the CHS cohort are presented in **Supplementary Table S3**. In addition, according to gender stratification (**Figure 2A** and **Supplementary Table S4**), the Spearman correlation analyses were conducted between the DNAm levels and the chronological ages in three different datasets, which is visualized in **Figure 2B** (detailed results in **Supplementary Table S5**). Except for *C1orf132* where DNAm decreases with age, other genes have positive correlations with chronological ages. In total, we identified 25 AR-CpGs out of the 34 CpGs in the CHS cohort (29 AR-CpGs for female dataset, 24 AR-CpGs for male dataset), which are highly related ($|r|≥0.7$, $p < 0.05$) with the chronological ages of CHS. In addition, the *KLF14* has no apparent strong correlation with the chronological ages (all $r < 0.7$), except for

TABLE 2 | Model settings and system efficiencies for three different datasets of the CHS cohort (*n* = 240, blood samples) in two SVR models.

SVR	Setting				Dataset	*n*	Number of support vectors	*RMSE*	*MAD*
	cost	gamma	epsilon	nu					
SVR-eps	1	0.04	0.1	–	Females	132	90	2.84	2.09
					Males	108	69	2.93	2.12
					Combined	240	163	2.95	2.22
SVR-nu	1	–	–	0.5	Females	132	105	2.82	1.92
					Males	108	79	2.90	2.00
					Combined	240	168	2.94	2.19

SVR-eps, support vector regression eps-regression; SVR-nu, support vector regression nu-regression; RMSE, root mean square error; MAD, median absolute deviation.

KLF14_K1 in males (r = 0.7082). Meanwhile, three different AR-CpGs (ELOVL2_E3, ELOVL2_E4, and FHL2_F1) have high correlations with the chronological ages in all gender datasets of the CHS cohort. Detailed results of Spearman analyses are visualized in **Supplementary Figures S2–S6** for *ELOVL2*, *C1orf132*, *FHL2*, *TRIM59*, and *KLF14*, respectively.

Stepwise Regression Model

The AR-CpGs with $|r|$≥0.7 of different datasets were regarded as alternative stepwise variables. A stepwise variable selection was conducted to select the best possible combination of predictors from the candidate highly associated CpGs for the SR model, which guaranteed the explained variability without overfitting the data. Based on different gender datasets, we built three distinct SR equations and calculated corresponding statistics for female (*MAD* = 3.00 and *RMSE* = 4.07), male (*MAD* = 2.64 and *RMSE* = 3.45), and combined (*MAD* = 2.97 and *RMSE* = 3.89) datasets corresponding to the age prediction models (details in **Table 1**, all adjusted R^2 ≥ 0.93). There was no significant difference between females and males in the CHS cohort (t = 0.59, p = 0.61).

Furthermore, we evaluated the prediction accuracy of the SR models in Training (*MAD* = 3.04, n = 170) and Validation (*MAD* = 2.80, n = 70) sets, respectively (**Supplementary Table S6**). The *MAD* values between Training and Validation sets had no significant difference (t = −1.06, p = 0.31). In total, the *MAD* values of different CHS datasets ranged from 2.14 (1–20 age category of Training set, n = 41) to 5.12 (61–81 age category of Validation set, n = 3). In addition, in the female dataset, the *MAD* values spanned from 2.25 (1–20 age category of Training set, n = 20) to 8.39 (61–81 age category of Validation set, n = 1). In the male dataset, the *MAD* values varied from 1.91 (1–20 age category of Validation set, n = 9) to 6.73 (61–81 age category of Validation set, n = 2). For different age categories, the lowest *MAD* value (1.91) was found at male validation dataset (1–20 age category, n = 9), while the highest *MAD* value (8.39) was identified at female validation dataset (61–81 age category, n = 1). The *MAD* values between females and males had no significant difference in both Training (t = 1.06, p = 0.35) and Validation (t = 0.25, p = 0.54) sets. Apparently, the *MAD* values rise with advancing ages (especially in the 61–81 age category), which indicated that the methylation-based SR model prediction accuracy decreases due to biological and physiological changes involved in the aging process, especially for the aged.

Support Vector Regression Model

Here, we constructed SVR models with two different methods (eps- and nu-regression) using correspondingly AR-CpG loci ($|r|$≥0.7) of distinct gender groups.

SVR eps-Regression

As shown in **Table 2**, we found 163 support vectors in the CHS cohort with an *MAD* value of 2.22 (*RMSE* = 2.95). In addition, the *MAD* values were 2.09 and 2.12 for female (n = 132, *RMSE* = 2.84) and male (n = 108, *RMSE* = 2.93) datasets, respectively, with no significant difference (t = 0.51, p = 0.13). The best performance (with the lowest *MAD* value) of SVR eps-regression was obtained with the optimized parameters (cost = 1, gamma = 0.04, epsilon = 0.1). The detailed *MAD* values for Training and Validation sets are presented in **Supplementary Table S7**. The *MAD* values were 2.33 and 1.87 for Training and Validation sets, respectively, with no significant difference (t = 1.68, p = 0.12).

In different age categories, the *MAD* values ranged from 1.59 (1–20 age category of Validation set, n = 18) to 4.72 (61–81 age category of Training set, n = 12). In addition, in the female dataset, the *MAD* values spanned from 1.35 (1–20 age category of Validation set, n = 9) to 10.06 (61–81 age category of Training set, n = 4). In the male dataset, the *MAD* values varied from 1.53 (1–20 age category of Validation set, n = 9) to 5.09 (61–81 age category of Validation set, n = 2). The *MAD* values between females and males had no significant difference in both Training (t = 0.77, p = 0.07) and Validation (t = −0.38, p = 0.90) sets. Overall, except for the 61–81 age category, the *MAD* value for each dataset was no more than 2.44.

SVR nu-Regression

Besides, the SVR nu-regression model was also used to predict the chronological ages (**Table 2**). The *MAD* value of the CHS cohort was 2.19 (*RMSE* = 2.94), which was obtained at cost = 1 and nu = 0.5 (including 168 support vectors). In female and male datasets, the *MAD* values were 1.92 and 2.00 with the support vectors of 105 and 79, and the *RMSE* values were 2.82 and 2.90, respectively. However, there was no significant difference between females and males in the CHS cohort (t = 0.52, p = 0.09). The detailed *MAD* values of Training and Validation sets are presented in **Supplementary Table S7**. The *MAD* values were 2.33 and 1.84 for Training and Validation sets with no significant difference (t = 1.78, p = 0.10), respectively.

In different age categories, the *MAD* values ranged from 1.56 (1–20 age category of Validation set, n = 18) to 4.73 (61–81 age

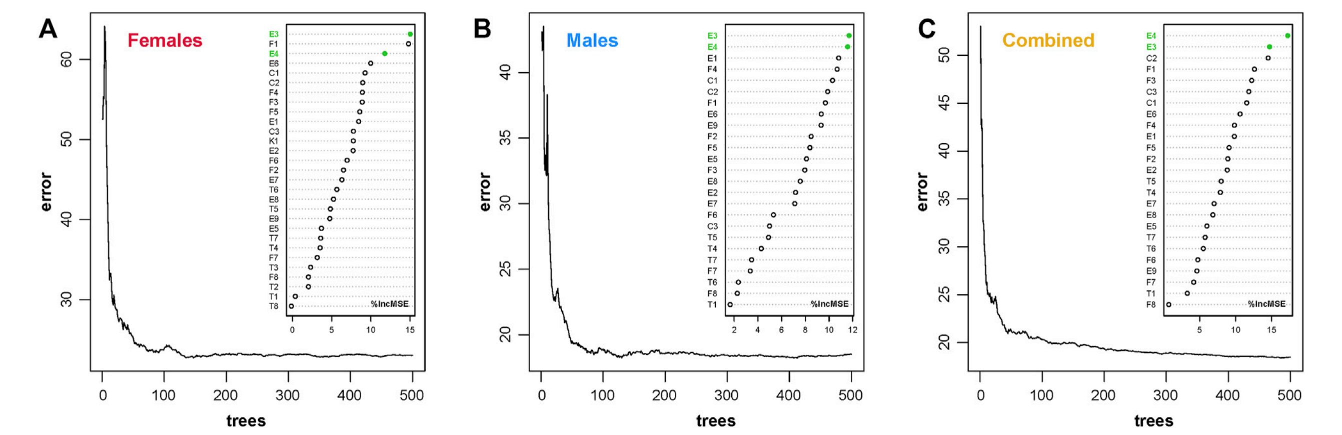


FIGURE 3 | Validation of feature selection (*ntree* = 500) and AR-CpG importance ranking in three different gender datasets of the CHS cohort (*n* = 240, blood samples). **(A)** Female dataset (*n* = 132). **(B)** Male dataset (*n* = 108). **(C)** Combined dataset (*n* = 240). (***ntree***, number of trees to grow, which should not be set to too small a number, to ensure that every input row gets predicted at least a few times; **%IncMSE**, increase in mean squared error.)

category of Training set, *n* = 12). In the female dataset, the *MAD* values spanned from 1.08 (1–20 age category of Validation set, *n* = 9) to 10.54 (61–81 age category of Training set, *n* = 4). In the male dataset, the *MAD* values varied from 1.27 (1–20 age category of Validation set, *n* = 9) to 5.18 (61–81 age category of Validation set, *n* = 2). The *MAD* values between females and males had no significant difference in both Training (*t* = 0.75, *p* = 0.07) and Validation (*t* = −0.27, *p* = 0.68) sets. The *MAD* value for each dataset was no more than 2.41 except for the 61–81 age category.

Compared with SVR-eps, the prediction capacity of the SVR-nu model was more excellent with lower *MAD* value for each dataset, while the model stability for both of them has larger fluctuations at the 61–81 age category (*MAD* values ranging from 3.42 to 10.54, details in **Supplementary Table S7**).

Random Forest Regression Model

Furthermore, the DNAm profiles of 240 CHS samples were learned by the RFR algorithm. For the *ntree* feature selection, we set six different threshold values (100, 300, 500, 1,000, 5,000, and 10,000) to find the robust limit with lower error rate (details in **Supplementary Figure S7**). In fact, the error rates tended to be stable when the *ntree* was more than 300. However, we set an *ntree* border at 500 to obtain more reliable results without regard to the hashrate for practice case handling. In addition, the feature selection (*ntree* = 500) was validated in different gender datasets, which indicated that the relatively lower and stable error rates are obtained with *ntree* of 500 (**Figure 3**). The E3 and E4 AR-CpG markers of *ELOVL2* genes (*r* > 0.9 in different gender datasets, details in **Supplementary Table S5**) ranked the top three positions in different gender datasets, which demonstrated that these biomarkers are the important predictive variables in the CHS cohort. According to different numbers of AR-CpGs for distinct gender datasets, the *mtry* values were set up at 9, 8, and 8 for female, male, and combined datasets, respectively.

With the feature selection and parameter setting as described above, the RFR model could explain 93.21% of the total variances (90.62% for females and 93.88% for males) in the CHS cohort (**Table 3**). The *MAD* values were 1.29 (*RMSE* = 1.77), 1.45 (*RMSE* = 1.95), and 1.32 (*RMSE* = 1.77) for combined, female, and male datasets, respectively. There was no significant difference between females and males in the CHS cohort (*t* = 0.98, *p* = 0.05). As shown in **Supplementary Table S8**, the *MAD* values of Training and Validation sets were 1.37 and 1.10, with no significant difference (*t* = 1.97, *p* = 0.07).

In different age categories, the *MAD* values ranged from 0.45 (1–20 age category of Validation set, *n* = 18) to 3.39 (61–81 age category of Validation set, *n* = 3). In the female dataset, the *MAD* values spanned from 0.59 (1–20 age category of Validation set, *n* = 9) to 4.47 (61–81 age category of Training set, *n* = 4). In the male dataset, the *MAD* values varied from 0.75 (1–20 age category of Validation set, *n* = 9) to 2.21 (61–81 age category of Validation set, *n* = 8). The *MAD* values between females and males had no significant difference in both Training (*t* = 0.90, *p* = 0.13) and Validation (*t* = 0.39, *p* = 0.23) sets. The detailed *MAD* values for each dataset are presented in **Supplementary Table S8**, and except for the 61–81 age category, the *MAD* values were less than 1.80.

Model Performance Comparison

Based on aforementioned ML algorithms, four different ML models have been established after multiple rounds of optimization, and the model efficiencies have been evaluated (details in **Table 4**). All R^2 values were above 0.95, and the R^2 value reached to 0.99 in the RFR model. The *MAD* values of the CHS cohort were 2.97 (*RMSE* = 3.89), 2.22 (*RMSE* = 2.95), 2.19 (*RMSE* = 2.94), and 1.29 (*RMSE* = 1.77) for SR, SVR-eps, SVR-nu, and RFR models, which are also visualized in **Figures 4A,B**. In the female dataset, the *MAD* values were 3.00 (*RMSE* = 4.07), 2.09 (*RMSE* = 2.84), 1.92 (*RMSE* = 2.82), and 1.45 (*RMSE* = 1.95) for SR, SVR-eps, SVR-nu, and RFR models, respectively. In the male dataset, the *MAD* values were 2.64 (*RMSE* = 3.45), 2.12 (*RMSE* = 2.93), 2.00 (*RMSE* = 2.90), and 1.32 (*RMSE* = 1.77) for SR, SVR-eps, SVR-nu, and RFR models, respectively. It demonstrated that no matter in female or male datasets, the RFR model had the highest predictive accuracy with an *MAD* value of 1.29.

TABLE 3 | Detailed feature selection and model efficiency information of random forest regression (RFR) models in three different gender datasets of the CHS cohort.

ML model	Dataset	n	*ntree*	*mtry*	% Var explained	*RMSE*	*MAD*
RFR	Females	132	500	9	90.62	1.95	1.45
	Males	108	500	8	93.88	1.77	1.32
	Combined	240	500	8	93.21	1.77	1.29
RFR (1–60)	Females	127	500	9	91.35	1.67	1.29
	Males	98	500	8	92.92	1.60	1.20
	Combined	225	500	8	93.13	1.54	1.15

ntree, number of trees to grow, which should not be set to too small a number, to ensure that every input row gets predicted at least a few times; mtry, number of variables randomly sampled as candidates at each split; % Var explained, the overall explanatory rate for the variances of the response variables by the predictive variables; RMSE, root mean square error; MAD, median absolute deviation.

TABLE 4 | System efficiency comparisons of different machine learning (ML) models.

ML model	R^2	*RMSE*	*MAD*
SR	0.95	3.89	2.97
SVR-eps	0.97	2.95	2.22
SVR-nu	0.97	2.94	2.19
RFR	0.99	1.77	1.29
RFR (1–60)	0.99	1.54	1.15

R^2, coefficient of determination/goodness-of-fit; RMSE, root mean square error; MAD, median absolute deviation; SR, stepwise regression; SVR-eps, support vector regression eps-regression; SVR-nu, support vector regression nu-regression; RFR, random forest regression in the CHS cohort; RFR (1–60), random forest regression at the 1–60 age categories of the CHS cohort.

In four different ML models of the CHS cohort, we definitely observed that the *MAD* values increased with the chronological ages, especially in the 61–81 age category with a rapid increase (**Figures 4C–F**). In addition, to obtain more precise prediction accuracy, we evaluated the best-performing RFR model in the age categories of 1–60 (excluding the 61–81 age category). As presented in **Supplementary Figure S8**, the *ntree* feature (*ntree* = 500) was further validated in different gender datasets, and the E3 and E4 CpGs of *ELOVL2* were also the most important predictive variables in the RFR model (1–60 age categories). The *MAD* value of all 225 CHS samples reduced to 1.15 (*RMSE* = 1.54), and the *MAD* values were 1.21 and 1.01 for Training (*n* = 158) and Validation (*n* = 67) sets, respectively (**Supplementary Table S9**). In **Table 4** and **Figures 4G,H**, the *MAD* values of the RFR (1–60) model were 1.29 in females (*RMSE* = 1.67) and 1.20 in males (*RMSE* = 1.60). Compared with the RFR model for the 1–81 age categories, both the *MAD* and *RMSE* values of RFR (1–60) have decreased, and the *MAD* values were especially less than 1.00 in the 1–20 age category (**Supplementary Table S9**), which demonstrated that the RFR (1–60) model is more suitable for the age precise prediction of youngsters. Additionally, the relationships between predicted ages and chronological ages in different ML models were conducted (**Supplementary Figure S9**), and the R^2 values of all different ML models were more than 0.94.

DISCUSSION

Forensic community has long been seeking for a molecular marker to facilitate age prediction from biological traces at crime scenes. The DNAm biomarkers served as the most promising information source for chronological age estimation, even though the aging process was impacted by both inherited genetic and environmental factors (Li et al., 2018; Morrow et al., 2020; Ryan et al., 2020; Mukherjee et al., 2021). Most of the existing studies selected their DNAm biomarkers based on these biomarkers' biological relevance to the aging process (Zubakov et al., 2016), statistically correlations with the chronological ages (Shadrina et al., 2018), or feature selection algorithms (Gao et al., 2020). In this study, the correlation coefficient ranking of nine candidate DNAm biomarkers was obtained from a cohort of 7,084 individuals using meta-analysis. Among them, we selected four top-ranking genes (*ELOVL2*, *TRIM59*, *FHL2*, and *C1orf132*) and *KLF14* chosen by a three-step feature selection algorithm AgeGuess to generate the DNAm profiles of the CHS cohort by BTA-pseq technology.

Correlation of DNAm status in five abovementioned genes with chronological age has been very well documented in different tissues and cell types (Zubakov et al., 2016; Cho et al., 2017; Jung et al., 2019; Dias et al., 2020; Anaya et al., 2021; Pfeifer et al., 2021; Woźniak et al., 2021). Our Spearman correlation analysis detected different strongly related CpG ($|r|$≥0.9) numbers in male (10 AR-CpGs) and female (4 AR-CpGs) datasets, mainly in *ELOVL2* and *FHL2*. However, the *MAD* values had no significant difference between female and male datasets in different SR (*t* = 0.59, *p* = 0.61), SVR-eps (*t* = 0.51, *p* = 0.13), SVR-nu (*t* = 0.52, *p* = 0.09), and RFR (*t* = 0.98, *p* = 0.05) models. The results indicated that the effect of gender on age prediction has not been detected in the present study (all $p >$ 0.05), which was in concordant with Koch and Wagner (2011). In contrast, some studies presented that DNAm in men changes 4% faster than that in women (Hannum et al., 2013) and the predicted age was higher in men than in women (Weidner et al., 2014; Zbieć-Piekarska et al., 2015b). The gender effect on age estimation is inconclusive; however, it is conclusive that there is no gender effect in our ML models at least.

The chosen methylomic biomarker *KLF14* has strongly age-associated relationships in Caucasians and Hispanics (Gao et al., 2020), but the age correlations were not apparent in the CHS cohort, Koreans, and Polish (**Supplementary Table S10**). In addition, we observed high *r* value of 0.798 (F7 of *FHL2*) in the CHS cohort, but the corresponding *r* value is 0.42 in Polish. In different East Asian populations, the *r* values were 0.67 and

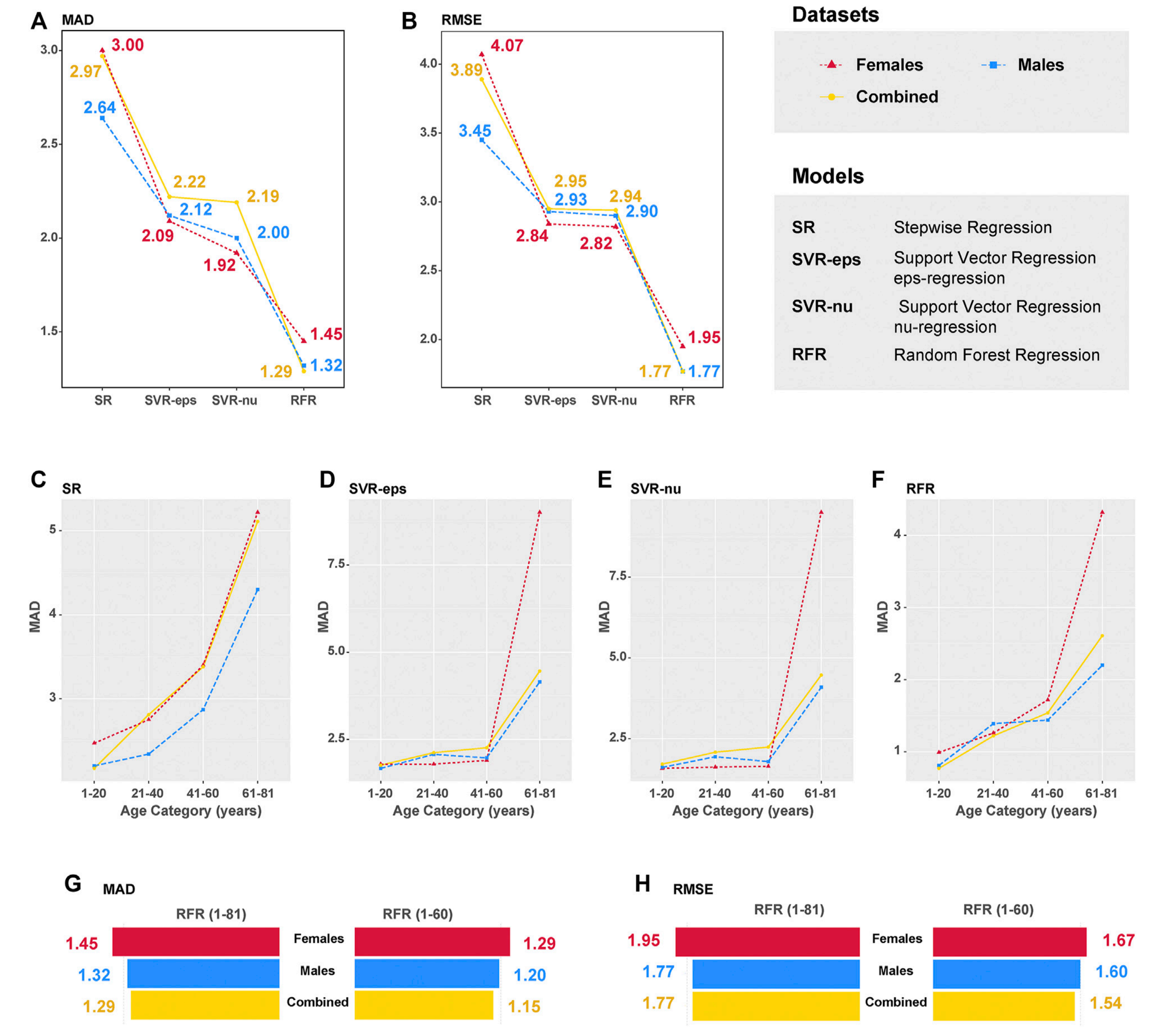


FIGURE 4 | Median absolute deviation (*MAD*) and root mean square error (*RMSE*) values of different machine learning (ML) models. **(A)** *MAD* value comparison of four different models in the CHS cohort (*n* = 240, blood samples). **(B)** *RMSE* value comparison of four different models in the CHS cohort (*n* = 240, blood samples). **(C)** *MAD* values in different age categories of the SR model. **(D)** *MAD* values in different age categories of the SVR-eps model. **(E)** *MAD* values in different age categories of the SVR-nu model. **(F)** *MAD* values in different age categories of the RFR model. **(G)** *MAD* value comparison between RFR (1–81) and RFR (1–60) models. **(H)** *RMSE* value comparison between RFR (1–81) and RFR (1–60) models.

0.87 at T8 of *TRIM59* in CHS and Koreans, respectively. Our results demonstrated that different populations have distinct methylation status under the same conditions, for both intercontinental and regional populations (termed as population-specific), which indicated that it is urgently necessary to determine the population-specific AR-CpGs available for practical application regionally.

This study further established four different ML models for chronological age prediction in the CHS cohort. Our results obtained from both Training and Validation sets are concordant in four different ML models (all $p > 0.05$), and the *MAD* values were less than 3.0 years (**Table 4**), which indicated that all ML models are robust in the CHS cohort. Based on the same five age-related genes, Zbieć-Piekarska et al. constructed the SR model in Polish with the *MAD* values of 3.4 and 3.9 in Training and Validation sets, respectively (Zbieć-Piekarska et al., 2015b). Another SR model exhibited an *MAD* value of 4.18 in 100 Korean blood samples (Cho et al., 2017). Jung et al. used multiplex methylation SNaPshot assay to establish the SR model using 150 Korean blood samples with the *MAD* values of 3.174 and 3. 478 in Training and Validation sets, respectively (Jung et al., 2019). Compared to the aforementioned SR models, the SR model of the CHS cohort showed higher prediction accuracy (*MAD* = 3.04 in Training set and *MAD* = 2.80 in Validation set). In addition, the *MAD* values of two optimized SVR models were 2.22 and 2.19 for SVR-eps and SVR-nu models (**Table 2**, **Table 4**), which were better than the SR model in the CHS cohort. Additionally, the RFR model with an *MAD* value of 1.29 was the best-performing ML model in the CHS cohort, which was confirmed at both Training (*MAD* = 1.45) and Validation (*MAD* = 1.32) sets without significant difference. Under the same condition, different ML algorithms have apparent influences on age prediction model accuracy.

In our data, we also found that the age prediction accuracy decreases with chronological age in different ML models (**Figures 4C–F**). As DNAm is a dynamic modification process, age-associated changes in DNAm have been well documented, and a previous study has identified that DNAm tends to increase with age on some CpG islands (Field et al., 2018). Moreover, the *MAD* values are affected by small sample size (only 15 individuals in the 61–81 age category of the CHS cohort), resulting in some biases for chronological age prediction. Thus, the absolute differences between predicted and chronological ages are larger in the categories of older people, which are also confirmed by previous studies (Zbieć-Piekarska, et al., 2015b; Hamano et al., 2016; Cho et al., 2017; Dias et al., 2020). Notably, the *MAD* value of the RFR model reduced to 1.15 years in the age range of 1–60. In the meta cohort, the *MAD* values ranged from 2.53 to 5.07 years. As far as we know, it is the best chronological age prediction model in Han Chinese.

In fact, the DNAm status reflects biological age rather than chronological age. However, DNAm estimated age can be considered as an "epigenetic clock," which in many cases runs parallel with chronological age (Horvath, 2013; Marioni et al., 2015). The epigenetic clock of CHS can be established by four age-related genes and different ML algorithms. From our perspectives, finding more population-specific and age-associated genes, expanding larger sample sizes (**Figures 4G,H**), and optimizing ML algorithms will contribute to generating more precise epigenetic clocks for diverse human populations.

CONCLUSION

In the present study, we conducted that 1) a candidate set of nine DNAm biomarkers was collected by meta-analysis with a number of 7,084 individuals; 2) the DNAm profiles of five promising genes were generated using BTA-pseq in the CHS cohort; and 3) four different ML models based on age-related CpGs ($|r|\geq 0.7$) were established and optimized in different datasets. In addition, we concluded that 1) gender effect has little influence on age prediction; 2) methylation levels at different CpGs exhibit population specificity; and 3) the age prediction accuracy decreases with chronological age. Eventually, an optimized RFR ML model with an *MAD* value of 1.15 has been established (*ntree* = 500 and *mtry* = 8) at the 1–60 age categories of CHS using whole blood DNAm data generated by BTA-pseq.

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

The studies involving human participants were reviewed and approved by the Biomedical Ethics Committee of Southern Medical University (No. 2021-015). 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 individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Conceptualization, HF; methodology, HF and QX; software, HF and QX; validation, HF, QX, JW, and ZZ; formal analysis, HF; investigation, HF, QX, and ZZ; resources, HF and PQ; data curation, PQ and XC; writing—original draft preparation, HF, QX, and XC; writing—review and editing, HF, QX, and XC; visualization, HF and QX; supervision, HF, XC, and PQ; project administration, PQ; funding acquisition, HF and PQ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was supported by grants from the Program of Hainan Association for Science and Technology Plans to Youth R&D Innovation (QCXM201705), Science Foundation of School of Forensic Medicine, Southern Medical University (No. 2021KY02), and National Natural Science Foundation of China (NSFC, No. 81671865 and No. 81971786).

ACKNOWLEDGMENTS

The authors sincerely thank all the volunteers who contributed samples for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2021.819991/full#supplementary-material

REFERENCES

Aliferi, A., Ballard, D., Gallidabino, M. D., Thurtle, H., Barron, L., and Syndercombe Court, D. (2018). DNA Methylation-Based Age Prediction Using Massively Parallel Sequencing Data and Multiple Machine Learning Models. *Forensic Sci. Int. Genet.* 37, 215–226. doi:10.1016/j.fsigen.2018.09.003

Alsaleh, H., and Haddrill, P. R. (2019). Identifying Blood-specific Age-Related DNA Methylation Markers on the Illumina MethylationEPIC BeadChip. *Forensic Sci. Int.* 303, 109944. doi:10.1016/j.forsciint.2019.109944

Anaya, Y., Yew, P., Roberts, K. A., and Hardy, W. R. (2021). DNA Methylation of Decedent Blood Samples to Estimate the Chronological Age of Human Remains. *Int. J. Leg. Med* 135 (6), 2163–2173. doi:10.1007/s00414-021-02650-8

Bekaert, B., Kamalandua, A., Zapico, S. C., Van de Voorde, W., and Decorte, R. (2015). Improved Age Determination of Blood and Teeth Samples Using a Selected Set of DNA Methylation Markers. *Epigenetics* 10 (10), 922–930. doi:10.1080/15592294.2015.1080413

Boks, M. P., Derks, E. M., Weisenberger, D. J., Strengman, E., Janson, E., Sommer, I. E., et al. (2009). The Relationship of DNA Methylation with Age, Gender and Genotype in Twins and Healthy Controls. *PLoS One* 4 (8), e6767. doi:10.1371/journal.pone.0006767

Cho, S., Jung, S.-E., Hong, S. R., Lee, E. H., Lee, J. H., Lee, S. D., et al. (2017). Independent Validation of DNA-Based Approaches for Age Prediction in Blood. *Forensic Sci. Int. Genet.* 29, 250–256. doi:10.1016/j.fsigen.2017.04.020

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., et al. (2009). Aging and Environmental Exposures Alter Tissue-specific DNA Methylation Dependent upon CpG Island Context. *Plos Genet.* 5 (8), e1000602. doi:10.1371/journal.pgen.1000602

Daunay, A., Baudrin, L. G., Deleuze, J.-F., and How-Kit, A. (2019). Amp; How-Kit, AEvaluation of Six Blood-Based Age Prediction Models Using DNA Methylation Analysis by Pyrosequencing. *Sci. Rep.* 9 (1), 8862. doi:10.1038/s41598-019-45197-w

Dias, H. C., Cordeiro, C., Pereira, J., Pinto, C., Real, F. C., Cunha, E., et al. (2020). DNA Methylation Age Estimation in Blood Samples of Living and Deceased Individuals Using a Multiplex SNaPshot Assay. *Forensic Sci. Int.* 311, 110267. doi:10.1016/j.forsciint.2020.110267

Esteller, M. (2002). CpG Island Hypermethylation and Tumor Suppressor Genes: a Booming Present, a Brighter Future. *Oncogene* 21 (35), 5427–5440. doi:10.1038/sj.onc.1205600

Feng, L., Peng, F., Li, S., Jiang, L., Sun, H., Ji, A., et al. (2018). Systematic Feature Selection Improves Accuracy of Methylation-Based Forensic Age Estimation in Han Chinese Males. *Forensic Sci. Int. Genet.* 35, 38–45. doi:10.1016/j.fsigen.2018.03.009

Field, A. E., Robertson, N. A., Wang, T., Havas, A., Ideker, T., and Adams, P. D. (2018). DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Mol. Cel* 71 (6), 882–895. doi:10.1016/j.molcel.2018.08.008

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., et al. (2005). From the Cover: Epigenetic Differences Arise during the Lifetime of Monozygotic Twins. *Proc. Natl. Acad. Sci.* 102 (30), 10604–10609. doi:10.1073/pnas.0500398102

Gao, X., Liu, S., Song, H., Feng, X., Duan, M., Huang, L., et al. (2020). AgeGuess, a Methylomic Prediction Model for Human Ages. *Front. Bioeng. Biotechnol.* 8, 80. doi:10.3389/fbioe.2020.00080

Gao, X., Nowak-Imialek, M., Chen, X., Chen, D., Herrmann, D., Ruan, D., et al. (2019). Establishment of Porcine and Human Expanded Potential Stem Cells. *Nat. Cel Biol* 21 (6), 687–699. doi:10.1038/s41556-019-0333-2

Garali, I., Sahbatou, M., Daunay, A., Baudrin, L. G., Renault, V., Bouyacoub, Y., et al. (2020). Improvements and Inter-laboratory Implementation and Optimization of Blood-Based Single-Locus Age Prediction Models Using DNA Methylation of the ELOVL2 Promoter. *Sci. Rep.* 10 (1), 15652. doi:10.1038/s41598-020-72567-6

Grönniger, E., Weber, B., Heil, O., Peters, N., Stäb, F., Wenck, H., et al. (2010). Aging and Chronic Sun Exposure Cause Distinct Epigenetic Changes in Human Skin. *Plos Genet.* 6 (5), e1000971. doi:10.1371/journal.pgen.1000971

Gršković, B., Zrnec, D., Vicković, S., Popović, M., and Mršić, G. (2013). DNA Methylation: the Future of Crime Scene Investigation? *Mol. Biol. Rep.* 40 (7), 4349–4360. doi:10.1007/s11033-013-2525-3

Hamano, Y., Manabe, S., Morimoto, C., Fujimoto, S., Ozeki, M., and Tamaki, K. (2016). Forensic Age Prediction for Dead or Living Samples by Use of Methylation-Sensitive High Resolution Melting. *Leg. Med.* 21, 5–10. doi:10.1016/j.legalmed.2016.05.001

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cel* 49 (2), 359–367. doi:10.1016/j.molcel.2012.10.016

Horvath, S. (2013). DNA Methylation Age of Human Tissues and Cell Types. *Genome Biol.* 14 (10), R115. doi:10.1186/gb-2013-14-10-r115

Horvath, S., and Raj, K. (2018). DNA Methylation-Based Biomarkers and the Epigenetic Clock Theory of Ageing. *Nat. Rev. Genet.* 19 (6), 371–384. doi:10.1038/s41576-018-0004-3

Jung, S.-E., Lim, S. M., Hong, S. R., Lee, E. H., Shin, K.-J., and Lee, H. Y. (2019). DNA Methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 Genes for Age Prediction from Blood, Saliva, and Buccal Swab Samples. *Forensic Sci. Int. Genet.* 38, 1–8. doi:10.1016/j.fsigen.2018.09.010

Jylhävä, J., Pedersen, N. L., and Hägg, S. (2017). Biological Age Predictors. *EBioMedicine* 21, 29–36. doi:10.1016/j.ebiom.2017.03.046

Koch, C. M., and Wagner, W. (2011). Epigenetic-aging-signature to Determine Age in Different Tissues. *Aging* 3 (10), 1018–1027. doi:10.18632/aging.100395

Lau, P. Y., and Fung, W. K. (2020). Evaluation of Marker Selection Methods and Statistical Models for Chronological Age Prediction Based on DNA Methylation. *Leg. Med.* 47, 101744. doi:10.1016/j.legalmed.2020.101744

Li, J., Zhu, X., Yu, K., Jiang, H., Zhang, Y., Wang, B., et al. (2018). Exposure to Polycyclic Aromatic Hydrocarbons and Accelerated DNA Methylation Aging. *Environ. Health Perspect.* 126 (6), 067005. doi:10.1289/ehp2773

Li, S. F., Peng, F. D., Wang, J. N., Zhong, J. J., Zhao, H., Wang, L., et al. (2019). Methylation-Based Age Estimation Model Construction and its Effectiveness Evaluation. *Fa Yi Xue Za Zhi* 35 (1), 17–22. doi:10.12116/j.issn.1004-5619.2019.01.004

Marioni, R. E., Shah, S., McRae, A. F., Chen, B. H., Colicino, E., Harris, S. E., et al. (2015). DNA Methylation Age of Blood Predicts All-Cause Mortality in Later Life. *Genome Biol.* 16 (1), 25. doi:10.1186/s13059-015-0584-6

Morrow, J. D., Make, B., Regan, E., Han, M., Hersh, C. P., Tal-Singer, R., et al. (2020). DNA Methylation Is Predictive of Mortality in Current and Former Smokers. *Am. J. Respir. Crit. Care Med.* 201 (9), 1099–1109. doi:10.1164/rccm.201902-0439OC

Mukaka, M. M. (2012). Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research. *Malawi Med. J.* 24 (3), 69–71.

Mukherjee, N., Arathimos, R., Chen, S., Kheirkhah Rahimabad, P., Han, L., Zhang, H., et al. (2021). DNA Methylation at Birth Is Associated with Lung Function Development until Age 26 Years. *Eur. Respir. J.* 57 (4), 2003505. doi:10.1183/13993003.03505-2020

Naue, J., Hoefsloot, H. C. J., Mook, O. R. F., Rijlaarsdam-Hoekstra, L., van der Zwalm, M. C. H., Henneman, P., et al. (2017). Chronological Age Prediction Based on DNA Methylation: Massive Parallel Sequencing and Random forest Regression. *Forensic Sci. Int. Genet.* 31, 19–28. doi:10.1016/j.fsigen.2017.07.015

Núñez, E., Steyerberg, E. W., and Núñez, J. (2011). Regression Modeling Strategies. *Revista Española de Cardiología (English Edition)* 64 (6), 501–507. doi:10.1016/j.recesp.2011.01.01910.1016/j.rec.2011.01.017

Pan, C., Yi, S., Xiao, C., Huang, Y., Chen, X., and Huang, D. (2020). The Evaluation of Seven Age-Related CpGs for Forensic Purpose in Blood from Chinese Han Population. *Forensic Sci. Int. Genet.* 46, 102251. doi:10.1016/j.fsigen.2020.102251

Park, J.-L., Kim, J. H., Seo, E., Bae, D. H., Kim, S.-Y., Lee, H.-C., et al. (2016). Identification and Evaluation of Age-Correlated DNA Methylation Markers for Forensic Use. *Forensic Sci. Int. Genet.* 23, 64–70. doi:10.1016/j.fsigen.2016.03.005

Parson, W. (2018). Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review. *Gerontology* 64 (4), 326–332. doi:10.1159/000486239

Pfeifer, M., Greb, A., Bajanowski, T., and Poetsch, M. (2021). Performance des PyroMark Q48 FX Age Assay auf zwei unterschiedlichen Pyrosequenzierplattformen. *Rechtsmedizin* 31 (3), 217–225. doi:10.1007/s00194-021-00491-8

Piniewska-Róg, D., Heidegger, A., Pośpiech, E., Xavier, C., Pisarek, A., Jarosz, A., et al. (2021). Impact of Excessive Alcohol Abuse on Age Prediction Using the VISAGE Enhanced Tool for Epigenetic Age Estimation in Blood. *Int. J. Leg. Med* 135 (6), 2209–2219. doi:10.1007/s00414-021-02665-1

Portela, A., and Esteller, M. (2010). Epigenetic Modifications and Human Disease. *Nat. Biotechnol.* 28 (10), 1057–1068. doi:10.1038/nbt.1685

Ryan, J., Wrigglesworth, J., Loong, J., Fransquet, P. D., and Woods, R. L. (2020). A Systematic Review and Meta-Analysis of Environmental, Lifestyle, and Health Factors Associated with DNA Methylation Age. *J. Gerontol. A. Biol. Sci. Med. Sci.* 75 (3), 481–494. doi:10.1093/gerona/glz099

Sen, P., Shah, P. P., Nativio, R., and Berger, S. L. (2016). Epigenetic Mechanisms of Longevity and Aging. *Cell* 166 (4), 822–839. doi:10.1016/j.cell.2016.07.050

Shadrina, A., Tsepilov, Y., Sokolova, E., Smetanina, M., Voronina, E., Pakhomov, E., et al. (2018). Genome-wide Association Study in Ethnic Russians Suggests an Association of the MHC Class III Genomic Region with the Risk of Primary Varicose Veins. *Gene* 659, 93–99. doi:10.1016/j.gene.2018.03.039

Smeers, I., Decorte, R., Van de Voorde, W., and Bekaert, B. (2018). Evaluation of Three Statistical Prediction Models for Forensic Age Prediction Based on DNA Methylation. *Forensic Sci. Int. Genet.* 34, 128–133. doi:10.1016/j.fsigen.2018.02.008

Sukawutthiya, P., Sathirapatya, T., and Vongpaisarnsin, K. (2021). A Minimal Number CpGs of ELOVL2 Gene for a Chronological Age Estimation Using Pyrosequencing. *Forensic Sci. Int.* 318, 110631. doi:10.1016/j.forsciint.2020.110631

Suzuki, K., Suzuki, I., Leodolter, A., Alonso, S., Horiuchi, S., Yamashita, K., et al. (2006). Global DNA Demethylation in Gastrointestinal Cancer Is Age Dependent and Precedes Genomic Damage. *Cancer Cell* 9 (3), 199–207. doi:10.1016/j.ccr.2006.02.016

Tra, J., Kondo, T., Lu, Q., Kuick, R., Hanash, S., and Richardson, B. (2002). Infrequent Occurrence of Age-dependent Changes in CpG Island Methylation as Detected by Restriction Landmark Genome Scanning. *Mech. Ageing Development* 123 (11), 1487–1503. doi:10.1016/s0047-6374(02)00080-5

Unnikrishnan, A., Freeman, W. M., Jackson, J., Wren, J. D., Porter, H., and Richardson, A. (2019). The Role of DNA Methylation in Epigenetics of Aging. *Pharmacol. Ther.* 195, 172–185. doi:10.1016/j.pharmthera.2018.11.001

Vidaki, A., Ballard, D., Aliferi, A., Miller, T. H., Barron, L. P., and Syndercombe Court, D. (2017). DNA Methylation-Based Forensic Age Prediction Using Artificial Neural Networks and Next Generation Sequencing. *Forensic Sci. Int. Genet.* 28, 225–236. doi:10.1016/j.fsigen.2017.02.009

Vidaki, A., Daniel, B., and Court, D. S. (2013). Forensic DNA Methylation Profiling-Potential Opportunities and Challenges. *Forensic Sci. Int. Genet.* 7 (5), 499–507. doi:10.1016/j.fsigen.2013.05.004

Weidner, C., Lin, Q., Koch, C., Eisele, L., Beier, F., Ziegler, P., et al. (2014). Aging of Blood Can Be Tracked by DNA Methylation Changes at Just Three CpG Sites. *Genome Biol.* 15 (2), R24. doi:10.1186/gb-2014-15-2-r24

Woźniak, A., Heidegger, A., Piniewska-Róg, D., Pośpiech, E., Xavier, C., Pisarek, A., et al. (2021). Development of the VISAGE Enhanced Tool and Statistical Models for Epigenetic Age Estimation in Blood, Buccal Cells and Bones. *Aging* 13 (5), 6459–6484. doi:10.18632/aging.202783

Xiao, C., Yi, S., and Huang, D. (2021). Genome-wide Identification of Age-related CpG Sites for Age Estimation from Blood DNA of Han Chinese Individuals. *Electrophoresis* 42 (14-15), 1488–1496. doi:10.1002/elps.202000367

Xu, C., Qu, H., Wang, G., Xie, B., Shi, Y., Yang, Y., et al. (2015). A Novel Strategy for Forensic Age Prediction by DNA Methylation and Support Vector Regression Model. *Sci. Rep.* 5, 17788. doi:10.1038/srep17788

Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T., Makowska, Ż., Spas, A., Parys-Proszek, A., et al. (2015a). Examination of DNA Methylation Status of the ELOVL2 Marker May Be Useful for Human Age Prediction in Forensic Science. *Forensic Sci. Int. Genet.* 14, 161–167. doi:10.1016/j.fsigen.2014.10.002

Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T., Parys-Proszek, A., Makowska, Ż., Pałeczka, A., et al. (2015b). Development of a Forensically Useful Age Prediction Method Based on DNA Methylation Analysis. *Forensic Sci. Int. Genet.* 17, 173–179. doi:10.1016/j.fsigen.2015.05.001

Zubakov, D., Liu, F., Kokmeijer, I., Choi, Y., van Meurs, J. B. J., van IJcken, W. F. J., et al. (2016). Human Age Estimation from Blood Using mRNA, DNA Methylation, DNA Rearrangement, and Telomere Length. *Forensic Sci. Int. Genet.* 24, 33–43. doi:10.1016/j.fsigen.2016.05.014

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.

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 28 February 2022
doi: 10.3389/fbioe.2021.811028

Classification of Diabetic Foot Ulcers Using Class Knowledge Banks

Yi Xu[1], *Kang Han*[2], *Yongming Zhou*[3]*, *Jian Wu*[1], *Xin Xie*[1] *and Wei Xiang*[4,2]

[1]Shanghai TCM-Integrated Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]College of Science and Engineering, James Cook University, Cairns, QLD, Australia, [3]Yueyang Hospital of Integrated Traditional Chinese Medicine and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [4]School of Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC, Australia

Diabetic foot ulcers (DFUs) are one of the most common complications of diabetes. Identifying the presence of infection and ischemia in DFU is important for ulcer examination and treatment planning. Recently, the computerized classification of infection and ischaemia of DFU based on deep learning methods has shown promising performance. Most state-of-the-art DFU image classification methods employ deep neural networks, especially convolutional neural networks, to extract discriminative features, and predict class probabilities from the extracted features by fully connected neural networks. In the testing, the prediction depends on an individual input image and trained parameters, where knowledge in the training data is not explicitly utilized. To better utilize the knowledge in the training data, we propose class knowledge banks (CKBs) consisting of trainable units that can effectively extract and represent class knowledge. Each unit in a CKB is used to compute similarity with a representation extracted from an input image. The averaged similarity between units in the CKB and the representation can be regarded as the logit of the considered input. In this way, the prediction depends not only on input images and trained parameters in networks but the class knowledge extracted from the training data and stored in the CKBs. Experimental results show that the proposed method can effectively improve the performance of DFU infection and ischaemia classifications.

Keywords: diabetic foot ulcer, image recongnition system, deep learning, infection and ischemia classification, knowledge learning

OPEN ACCESS

Edited by:
Honghao Gao,
Shanghai University, China

Reviewed by:
Xiaohui Tao,
University of Southern Queensland, Australia
Dezhi Li,
Harbin Institute of Technology, China
Zhong Zhang,
Tianjin Normal University, China

****Correspondence:***
Yongming Zhou
yongmingz@sohu.com

Specialty section:
This article was submitted to Biosafety and Biosecurity, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *08 November 2021*
Accepted: *31 December 2021*
Published: *28 February 2022*

Citation:
Xu Y, Han K, Zhou Y, Wu J, Xie X and Xiang W (2022) Classification of Diabetic Foot Ulcers Using Class Knowledge Banks. Front. Bioeng. Biotechnol. 9:811028. doi: 10.3389/fbioe.2021.811028

1 INTRODUCTION

The diabetic foot ulcer (DFU) is a complication of diabetes with high incidence Armstrong et al. (2017). According to the estimation of the International Diabetes Federation Atlas et al. (2015), 9.1 million to 26.1 million people with diabetes develop foot ulcers each year in the world. For people with diabetes, the presence of DFU can result in amputation and even increase the risk of death Walsh et al. (2016). Identifying whether the DFU is infection and ischaemia is important for its assessment, treatment, and management Jeffcoate and Harding (2003), where the infection is defined as bacterial soft tissue or bone infection in the DFU and ischaemia means inadequate blood supply Goyal et al. (2020). Classification of DFU infection and ischaemia by computerized methods is thus a critical research problem for automatic DFU assessment.

Traditional methods for diagnosis of DFU employ hand-crafted features followed by a classifier Veredas et al. (2009); Wannous et al. (2010); Wang et al. (2016). However, research in literature has shown that learned features by deep neural networks are more effective than traditional hand-crafted features LeCun et al. (2015). Extensive research has been done to increase the performance of computerized automatic medical image classification Litjens et al. (2017), where methods based on deep learning LeCun et al. (2015) are very popular in this field because they perform significantly better than other techniques Li et al. (2014); Kumar et al. (2016); Goyal et al. (2020); Wang et al. (2020); Cao et al. (2021).

The most widely used deep learning method in medical image classification is the convolutional neural network (CNN) Gulshan et al. (2016); Albawi et al. (2017). CNNs can effectively extract useful features for image classification He et al. (2016); Tan and Le (2019), object detection Redmon et al. (2016); Zhao et al. (2019), image segmentation Chen et al. (2018); Hesamian et al. (2019); Xiao et al. (2021) and many other vision tasks LeCun et al. (2015). With the availability of large-scale training data and high-performance modern GPUs and ASICs, methods based on CNNs have greatly improved the accuracy of image classification. Popular CNNs for general image classification tasks include AlexNet Krizhevsky et al. (2012), VGG Simonyan and Zisserman (2014), ResNet He et al. (2016), and EfficientNet Tan and Le (2019). These networks usually serve as the backbone of a medical image classification network, or directly apply to medical image classification by transfer learning with pre-trained parameters on large-scale datasets, e.g., ImageNet Deng et al. (2009). In practice, collecting and labeling medical images are costly. Transfer learning is thus an effective way to solve the problem of the lack of medical training data Shie et al. (2015); Shin et al. (2016); Cheplygina et al. (2019); Chen et al. (2020a).

Other emerging techniques for image classification include vision transformer Dosovitskiy et al. (2020); Touvron et al. (2020); Liu et al. (2021) and contrastive learning Wang et al. (2020); Jaiswal et al. (2021). Vision transformer methods are based on the attention mechanism, where an input image is split into small patches and the vision transformer can learn to focus on the most important regions for classification. Contrastive learning usually performs in an unsupervised way, where the network learns to minimize intra-class distance and maximize inter-class distance. The networks trained by contrastive learning perform well on the subsequent tasks like image segmentation but their classification accuracies are still inferior to those of state-of-the-art supervised methods.

However, existing medical image classification networks do not explicitly consider class knowledge in the training data when performing prediction. The training of existing networks involves the optimization of network parameters, where the class knowledge in the training data is extracted implicitly. In the testing, the trained networks process an input image into a high dimensional representation through trained parameters, where the class knowledge in the training data is not explicitly involved in the pipeline. To better utilize the class knowledge in the training data, we propose class knowledge banks (CKBs) that can effectively extract class knowledge from the training data, and the extracted class knowledge can directly participate in the prediction process. A CKB consists of many trainable units that can represent class knowledge from different perspectives. The average similarity between a representation extracted from an input image and knowledge units in the CKB can be used as a class probability. In this way, the class knowledge in the training data is explicitly utilized. Besides, the proposed CKB method can handle class imbalance as each class is given the same importance in CKBs. As a result, the network with the CKB is able to achieve state-of-the-art classification performance in the DFU image dataset (Goyal et al. (2020)). In summary, we make the following contributions:

- We propose a class knowledge bank (CKB) method that can explicitly and efficiently extract and utilize class knowledge in the training data.
- We show that the proposed CKB is good at handling class imbalance in the DFU image classification dataset.

The remainder of the paper is organized as follows. We first briefly review the related work in **Section 2**. Then we describe the proposed method in detail in **Section 3**. **Sections 4**, **5** present experimental results and discussions. We conclude the paper in **Section 6**.

2 RELATED WORK

In this section, we briefly review the related work on image classification, including convolutional neural networks, vision transformers, and contrastive learning.

2.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) are the most widely used technique for image classification LeCun et al. (2015). CNNs utilize multiple convolutional kernels in each layer and multiple layers to extract and process features from low levels to high levels. Since the success of AlexNet Krizhevsky et al. (2012) in image classification in 2012, a lot of methods based on CNN have been proposed to tackle this problem, and the performance of image classification on large datasets, e.g., ImageNet Deng et al. (2009), has been significantly improved. Many medical image classification methods are based on CNNs Li et al. (2014); Kumar et al. (2016); Anwar et al. (2018); Yadav and Jadhav (2019); Feng et al. (2020). Typical networks for image classification include VGG Simonyan and Zisserman (2014), ResNet He et al. (2016), Densenet Huang et al. (2017), Efficientnet Tan and Le (2019), and RegNet Radosavovic et al. (2020). These networks follow the structure of deep CNNs (to extract feature) and fully-connected (FC) layers (to predict classes). After training, the prediction depends on the input image and parameters in CNNs and FC layers, without explicit use of the class knowledge in the training data.

2.2 Vision Transformers

The transformer model was firstly proposed for natural language processing Vaswani et al. (2017). The model uses an attention mechanism Gao et al. (2021); Wang et al. (2017) to capture the correlation within tokens and learns to focus on important tokens. Dosovitskiy et al. (2020) first applied the transformer to image classification and achieved an even better performance than CNNs on the ImageNet dataset. Such a model is called vision transformer, where an input image is divided into patches and these patches are regarded as tokens to feed into the network. The vision transformer can learn to focus on important regions by the attention mechanism to predict class labels. Vision transformers have also been applied in medical image classification Dai et al. (2021). Furthermore, knowledge distillation Hinton et al. (2015); Wei et al. (2020) from the model based on CNN is shown to be effective in improving the performance of the vision transformer Touvron et al. (2020). Instead of simply regarding image patches as tokens, Yuan *et al.* proposed a tokens-to-token (T2T) method to better tokenize patches with the consideration of image structure Yuan et al. (2021). The T2T method achieves better accuracy using fewer parameters compared with the vanilla vision transformer Dosovitskiy et al. (2020).

2.3 Contrastive Learning

Contrastive learning aims at learning effective representations by maximizing the similarity between positive pairs and minimizing the similarity between negative pairs Jaiswal et al. (2021). It usually performs in a self-supervised manner, where positive pairs are from different augmentations of the same sample and negative pairs are simply different samples. SimCLR constructs contrastive loss by a large batch size, e.g., 4,096, to fully explore the similarity in negative pairs Chen et al. (2020b). Followed by SimCLR, SimCLRv2 achieves a better performance than SimCLR by leveraging bigger models and deeper projection head Chen et al. (2020c). Since contrastive learning requires a large number of representations of negative pairs, He *et al.* utilized a queue to store the representations of samples and updated the queue via a momentum mechanism He et al. (2020), which is shown to be more effective than sampling representations from the last epoch Wu et al. (2018). Although these contrastive learning methods achieve good performance on image classification by fine-tuning with few labeled samples, their performances are still inferior to those of state-of-the-art supervised methods.

Existing image classification networks do not explicitly take the class knowledge in the training data into account when performing prediction. To explicitly leverage class knowledge in the training data, we propose the so-called class knowledge bank method that is able to extract class knowledge from the training data, and the extracted class knowledge can directly participate in the prediction process.

3 METHODS

3.1 Proposed Network Structure

Given an input medical image **x**, the goal of image classification is to produce its class $y \in \{0, 1, \ldots, N-1\}$, where N is the number of classes. Existing deep neural networks for image classification usually extract discriminative features (representations) through a layer-by-layer structure, and directly yield the class from extracted features by multilayer perceptrons (MLPs). We introduce the class knowledge bank into the traditional pipeline to enable explicit utilization of class knowledge in the training data. As shown in **Figure 1**, the input image **x** is first fed to an encoder and then a projection head to extract a high-level representation $\mathbf{r} \in \mathbb{R}^D$

$$\begin{aligned} \mathbf{r}_0 &= Encoder(\mathbf{x}) \\ \mathbf{r} &= Projection(\mathbf{r}_0) \end{aligned} \tag{1}$$

where D is the dimension of the extracted representation. The projection head is introduced for two main purposes. Firstly, it can transfer the representation extracted by the encoder to a space that is suitable for contrastive learning, and thus improves the quality of the representation. Secondly, the representation from the encoder does not contain information specific to diabetic foot images as the encoder is pre-trained on a large-scale natural image dataset and froze when training the proposed network. The projection head can learn specific useful information from the diabetic foot dataset to improve classification performance. The extracted representation **r** is then used to compute the similarity with units in CKBs. The computed similarity can be regarded as the logits and used to build a contrastive loss to train the network.

3.2 Class Knowledge Bank

Each class of images has its properties, such as color and structure, which can be used to distinguish them from other classes of images. The knowledge of a class should contain these properties from different perspectives to comprehensively describe the class. The CKB method is proposed to achieve this goal and one CKB is designed to represent the knowledge of one class. A CKB consists of a number of units that can represent class knowledge from different perspectives. Each unit in the CKB is of the same size as the extracted representation **r** and there are M units in a CKB. The size of a CKB is thus $M \times D$. For image classification with N classes, we need N CKBs to store all the class knowledge. A CKB for class i can be represented as $\mathbf{C}_i$

$$\mathbf{C}_i = \{\mathbf{u}_i^1, \mathbf{u}_i^2, \mathbf{u}_i^3, \ldots, \mathbf{u}_i^M\} \tag{2}$$

where **u** denotes the unit in the CKB. The average similarity s_i between the $\mathbf{C}_i$ and the extracted representation **r** can be measured by the mean similarity across units

$$s_i = \frac{1}{M} \sum_j cos(\mathbf{u}_i^j, \mathbf{r}) \tag{3}$$

where cos $(\cdot, \cdot)$ is the cosine similarity

$$cos(\mathbf{u}_i^j, \mathbf{r}) = \frac{\mathbf{u}_i^j \cdot \mathbf{r}}{\| \mathbf{u}_i^j \| \| \mathbf{r} \|}. \tag{4}$$

A large s_i indicates the representation is close to the class i, which means the input **x** has a high probability of class i.

The measured similarities between the representation and the CKBs can be seen as the logits of the input image, and thus can be used to compute probabilities of classes through softmax function

$$p_i = \frac{exp(s_i)}{\sum_{k=0}^{N-1} exp(s_k)} \tag{5}$$

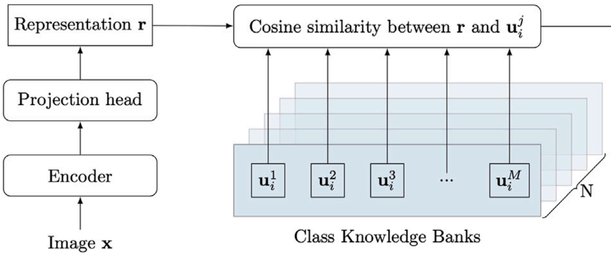


FIGURE 1 | Overview of the proposed network. The encoder and projection head embed the input image **x** into a representation **r**. The average similarity between the representation **r** and the unit $\mathbf{u}_i^j$ in the class knowledge bank (CKB) $\mathbf{C}_i$ is measured as the logit of **x** for class i. The CKBs are parameterized by a matrix of dimension $N \times M \times D$, where N is the number of classes, M is the number of units in each CKB, and D denotes the dimension of a unit. The CKBs are randomly initialized and can be trained through back-propagation.

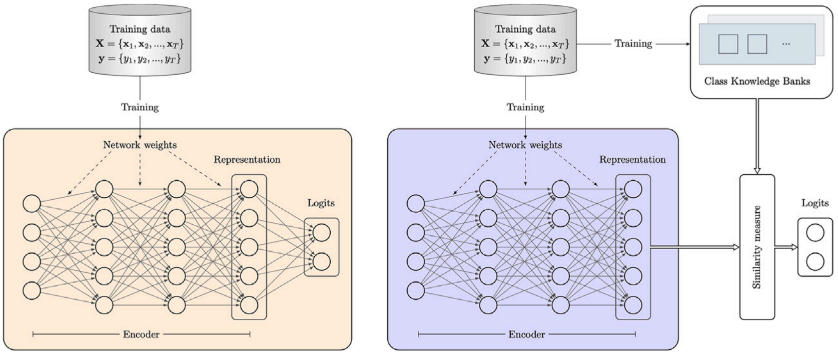

Training data
$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$
$\mathbf{y} = \{y_1, y_2, ..., y_T\}$
Training
Network weights
Representation
Logits
Encoder
Training data
$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$
$\mathbf{y} = \{y_1, y_2, ..., y_T\}$
Training
...
Class Knowledge Banks
Training
Network weights
Representation
Similarity measure
Logits
Encoder

FIGURE 2 | Comparison between common classification networks and the network with the proposed class knowledge banks. A common image classification network on the left is trained on the training data. After training, class knowledge in the training data only functions through the network weights. By contrast, except for the network weights, the proposed class knowledge banks can learn class knowledge from the training data and store the knowledge. The learned knowledge in the CKBs participates in the classification process by measuring the similarity between the CKBs and the representation of the input.

where p_i is the probability of class i. Based on the similarities, we define the following contrastive loss

$$\mathcal{L}_{CON}(\mathbf{C}, \mathbf{r}, y) = -\frac{1}{M}\sum_{j} cos(\mathbf{u}_y^j, \mathbf{r}) + \frac{1}{N-1}\sum_{i \neq y}\frac{1}{M}\sum_{j} cos(\mathbf{u}_i^j, \mathbf{r}). \tag{6}$$

Label y serves as the index of the correct class. Minimizing this contrastive loss is equivalent to maximizing the similarity between **r** and units in the correct CKB, and to minimizing the similarity between **r** and units in other CKBs. The final training loss is a combination of the above contrastive loss and cross-entropy loss $\mathcal{L}_{CEL}$:

$$\mathcal{L} = \mathcal{L}_{CON}(\mathbf{C}, \mathbf{r}, y) + \mathcal{L}_{CEL}(\mathbf{s}, y) \tag{7}$$

where $\mathbf{s} = \{s_0, s_1, \ldots, s_{N-1}\}$ are logits (represented by averaged similarities in (**3**)) of the input.

The units in CKBs are randomly initialized and then optimized through back-propagation. The network will try to extract the class knowledge into the CKBs with the objective of minimizing the designed contrastive loss in **(6)**. In this way, the proposed CKB method is more effective in utilizing knowledge in the training data than existing contrastive learning methods, e.g., end-to-end mechanism Oord et al. (2018), memory bank Wu et al. (2018) and momentum contrast He et al. (2020). This effectiveness is mainly derived from two aspects. Firstly, the proposed CKBs do not rely on a large number of specific samples. Instead, CKBs can learn to extract class knowledge and represent them by the units in the CKBs. Since the CKBs are optimized on the whole training dataset, they contain more comprehensive knowledge than some specific samples. Secondly, a small number of, e.g., 64, units in a CKB can represent the knowledge of one class very well, which can greatly reduce the computational complexity and memory usage compared with existing contrastive learning methods that usually require thousands of samples in one training iteration.

Figure 2 compares the proposed method and existing popular image classification methods. In existing image classification networks, parameters are mainly weights that are trained via back-propagation using the training data. The classification is achieved by directly predicting class logits from the discriminative representation extracted from the encoder. In such process, class knowledge in the training data is not

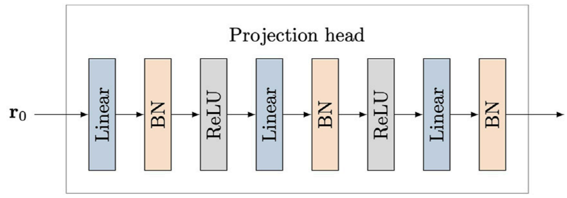


FIGURE 3 | Structure of projection head. Three linear layers are used and the first two linear layers are followed with batch normalization (BN) and ReLU activation. The input of the projection head is the representation $\mathbf{r}_0$ extracted from the encoder, and its output is the transformed representation $\mathbf{r}$ used for contrastive learning.

explicitly utilized as the network is trained to focus on extracting more discriminative representation from the input. As shown in **Figure 2**, the network with the proposed CKBs has a different pipeline of producing class logits. The CKBs learn and represent class knowledge through units parameterized by vectors. Then the learned class knowledge in the CKBs explicitly participates in the classification process by measuring the similarity between the units in the CKBs and the representation of the input. In this way, the class knowledge in the training data not only implicitly functions through network weights but explicitly works in the form of class similarity.

3.3 Encoder and Projection Head Structures

The encoder is concerned with extracting a discriminative representation from an input image. Training the network with an encoder from scratch on medical image datasets is not effective since medical datasets are usually comparatively small. Thus, we employ a pre-trained image classification network Touvron et al. (2020) that is trained on ImageNet as the encoder in our proposed network. This strategy is shown to be very effective for many medical image processing tasks when training datasets are small Shin et al. (2016); Goyal et al. (2020); Chen et al. (2021). We further introduce a multilayer perceptron (MLP) projection head as in Chen et al. (2020c) to transform the output representation from the encoder into a suitable space for contrastive learning. As shown in **Figure 3**, the input of the projection head is the representation $\mathbf{r}_0$ extracted from the encoder, and its output is the transformed representation $\mathbf{r}$ used for the following contrastive learning. The MLP projection head includes three linear layers and the first two linear layers are followed by batch normalization (BN) Ioffe and Szegedy (2015) and ReLU activation Nair and Hinton (2010). The output of the last linear layer is only processed by batch normalization without ReLU activation.

4 RESULTS

We use the diabetic foot ulcer (DFU) dataset in Goyal et al. (2020) to evaluate the performance of the proposed method. The DFU dataset includes ischaemia and infection parts that were collected from the Lancashire Teaching Hospitals. There are 628 non-infection and 831 infection cases, and 1,249 non-ischaemia and 210 ischaemia cases in the dataset. It can be observed that class imbalance exists in this dataset. The collected images were labeled by two healthcare professionals and augmented by the natural data augmentation method which extracts region of interest (ROI) ulcers by a learn-based ROI localization method Goyal et al. (2018). After augmentation, the ischaemia and infection parts include 9,870 and 5,892 augmented image patches, respectively. **Figure 4** shows samples of infection and ischaemia images from this dataset. We use 5-fold cross-validation and report on average performance and standard deviation.

The proposed method is implemented by the deep learning library Pytorch Paszke et al. (2017). We utilize the AdamW Loshchilov and Hutter (2017) algorithm as the optimizer to train models. The AdamW improves the generalization performance of the commonly used Adam algorithm Kingma and Ba (2014). The learning rate and weight decay are initialized to be 5e-4 and 0.01, respectively. The step learning rate scheduler is employed with the step size of 2 and the decay factor of 0.6. We use the batch size of 64 and train models in 20 epochs. Several popular image classification models are used for performance comparisons, including CNN-based ResNet He et al. (2016), RegNetY Radosavovic et al. (2020), EfficientNet Tan and Le (2019), contrastive learning-based MoCo He et al. (2020), and vision transformer-based DeiT Touvron et al. (2020). The input images are resized to the resolution of 224 × 224 for all methods for fair comparison.

To investigate whether larger models can lead to better performance, we evaluate the performance of the above models with different layers. Small and base DeiT models are denoted as DeiT-S and DeiT-B. For fair comparisons, all the competing methods use three linear layers with dimension 512 (first and second layers are followed by ReLU) as their classifiers, where the objective of the classifiers is to yield the logits. The objective of the projection head in our method is to produce discriminative representations. The number of units in each CKB is 64. Batch normalization (BN) is not applied in the MLP classifiers for comparison methods, since BN degrades these networks' performances. For all methods, we use the models pre-trained on ImageNet and freeze their parameters except for the parameters in the MLP classifiers, MLP projection head, and CKBs. We find freezing the pre-trained parameters leads to better performance than fine-tuning the whole network. For DeiT with knowledge distillation, there are two classifiers or projection heads that process the class token and distillation token, and the final prediction is the sum of two logits. We use accuracy, sensitivity, precision, specificity, F-measure, and area under the ROC curve (AUC) to measure the performance of the classification models.

Table 1 presents the DFU infection classification performances of various methods. As shown in **Table 1**, larger CNN models usually produce better results. The F-measure and AUC score of ResNet-152 are superior to those of ResNet-101. Similar results are also observed for RegNetY, where RegNetY-16GF achieves better performances than RegNetY-4GF and RegNetY-8GF. However, the performance differences for EfficientNet with different sizes are not significant, and the

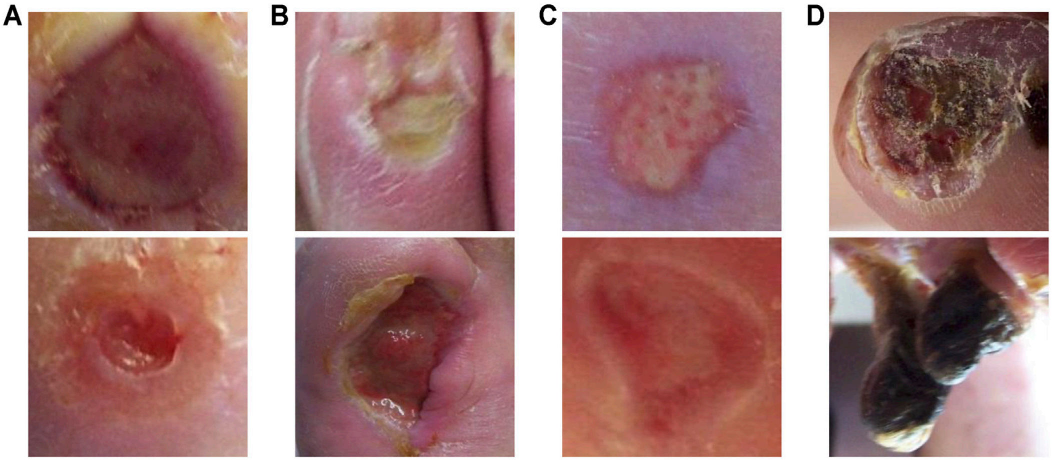


FIGURE 4 | Sample images from the DFU dataset Goyal et al. (2020). **(A)** are non-infection images, **(B)** are infection images, **(C)** are non-ischaemia images and **(D)** are ischaemia images.

TABLE 1 | Performance of binary classification on the DFU infection dataset.

Network	Accuracy	Sensitivity	Precision	Specificity	F-measure	AUC score
ResNet-18	74.20 ± 1.25	76.66 ± 2.82	73.08 ± 4.13	71.98 ± 2.98	74.72 ± 2.08	82.23 ± 1.26
ResNet-50	73.79 ± 1.31	76.96 ± 3.13	72.33 ± 2.93	70.75 ± 0.92	74.50 ± 1.98	81.44 ± 1.60
ResNet-101	74.63 ± 0.98	75.97 ± 0.88	73.88 ± 2.65	73.26 ± 1.37	74.89 ± 1.63	82.52 ± 0.81
ResNet-152	74.82 ± 1.02	76.80 ± 0.67	73.82 ± 2.42	72.82 ± 1.95	75.25 ± 1.17	82.78 ± 0.85
RegNetY-4GF	73.63 ± 1.50	75.92 ± 1.97	72.57 ± 2.72	71.37 ± 2.30	74.16 ± 1.53	81.33 ± 1.54
RegNetY-8GF	74.85 ± 1.51	76.72 ± 2.18	73.91 ± 3.18	73.00 ± 2.66	75.24 ± 1.89	81.90 ± 1.60
RegNetY-16GF	75.41 ± 0.95	77.42 ± 1.48	74.44 ± 2.61	73.51 ± 1.82	75.85 ± 0.75	83.02 ± 1.28
EfficientNet-B0	73.95 ± 1.06	77.81 ± 2.48	72.23 ± 3.51	70.19 ± 3.00	74.83 ± 1.75	81.77 ± 1.01
EfficientNet-B2	73.85 ± 1.21	77.64 ± 1.15	72.13 ± 3.34	70.09 ± 2.88	74.73 ± 1.70	81.58 ± 0.90
EfficientNet-B4	73.61 ± 1.17	77.24 ± 2.28	72.05 ± 3.78	70.16 ± 3.07	74.46 ± 1.66	81.70 ± 1.08
EfficientNet-B6	73.43 ± 0.54	75.77 ± 1.85	72.40 ± 3.30	71.29 ± 1.69	73.97 ± 1.14	80.98 ± 1.10
EfficientNet-B7	72.79 ± 1.53	72.14 ± 3.10	73.10 ± 2.60	73.43 ± 3.16	72.54 ± 1.67	80.08 ± 1.26
MoCo	74.97 ± 2.01	74.06 ± 1.75	75.47 ± 4.37	75.96 ± 4.08	74.68 ± 2.22	82.77 ± 1.46
DeiT-S	73.65 ± 0.64	77.22 ± 1.79	72.02 ± 3.18	70.14 ± 1.89	74.47 ± 1.62	80.90 ± 0.95
DeiT-B	73.97 ± 0.83	78.09 ± 2.23	72.12 ± 3.40	69.97 ± 2.42	74.91 ± 1.63	81.58 ± 1.35
DeiT-S-D	73.98 ± 1.97	78.06 ± 2.07	72.21 ± 4.28	70.09 ± 3.81	74.93 ± 2.23	81.51 ± 1.81
DeiT-B-D	75.82 ± 1.96	**79.96 ± 2.88**	73.86 ± 3.33	71.88 ± 2.14	76.72 ± 2.11	83.26 ± 2.36
CKB-DeiT-S-D	75.18 ± 1.27	76.91 ± 2.15	74.36 ± 3.79	73.54 ± 3.44	75.53 ± 1.78	82.66 ± 1.28
CKB-DeiT-B-D	**78.00 ± 0.93**	79.16 ± 1.74	**77.38 ± 2.68**	**77.00 ± 1.51**	**78.20 ± 0.94**	**84.78 ± 1.30**

large model even performs slightly worse than small models. MoCo with the backbone of ResNet-50 performs better than the vanilla ResNet-50 for infection classification, showing that the contrastive learning method helps the network learns more discriminative representations for image classifications. Vision transformer-based DeiT models trained with knowledge distillation (denoted as DeiT-S-D and DeiT-B-D) perform better than CNN models. This is reasonable as DeiT-B-D is shown to perform better than the comparison CNN models on ImageNet classification task Touvron et al. (2020). The superior performance of DeiT-B-D when transferred for the task of diabetic foot infection classification demonstrates its robustness. We also observe a phenomenon similar with Touvron et al. (2020) that knowledge distillation can significantly improve the performance of DeiT. For example, the F-measure and AUC score of DeiT-B-D are 76.72 and 83.26, which are better than those of DeiT-B (F-measure 74.91 and AUC 81.58) by large margins. The performance improvements of knowledge distillation for DeiT may be due to the inherited inductive bias from the CNN-based teacher model, e.g., RegNet Radosavovic et al. (2020), where DeiT mainly consists of multilayer perceptrons and attention modules.

Furthermore, when distilled DeiT is used in conjunction with the proposed CKB, denoted by CKB-DeiT-B-D, further performance improvements are obtained, leading to the best

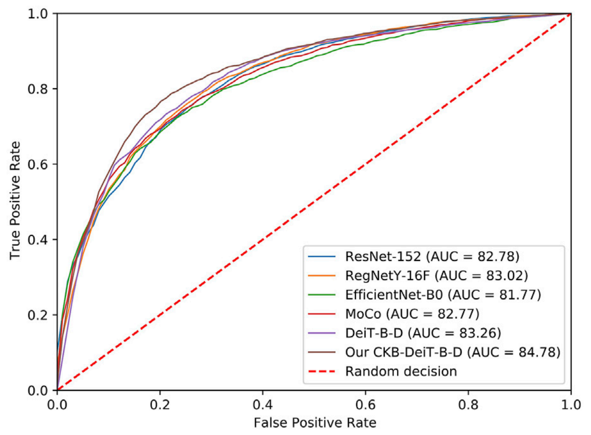


FIGURE 5 | ROC curves of different methods for the DFU infection classification.

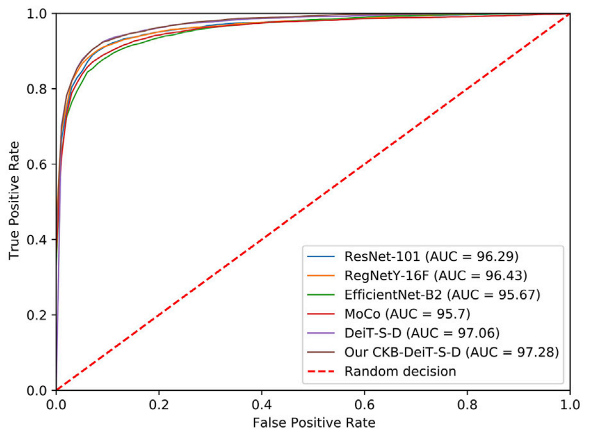


FIGURE 6 | ROC curves of different methods for the DFU ischaemia classification.

TABLE 2 | Performance of binary classification on the DFU ischaemia dataset.

Network	Accuracy	Sensitivity	Precision	Specificity	F-measure	AUC score
ResNet-18	88.30 ± 1.36	82.40 ± 3.48	93.23 ± 1.73	94.16 ± 1.03	87.43 ± 2.03	95.48 ± 0.89
ResNet-50	88.13 ± 1.77	81.46 ± 3.49	93.80 ± 1.42	94.76 ± 0.59	87.16 ± 2.26	95.09 ± 1.24
ResNet-101	89.95 ± 1.29	85.17 ± 2.50	94.10 ± 1.25	94.78 ± 0.55	89.38 ± 1.37	96.29 ± 1.17
ResNet-152	88.62 ± 2.18	82.45 ± 3.16	93.92 ± 2.22	94.84 ± 1.55	87.79 ± 2.42	95.58 ± 1.15
RegNetY-4GF	89.55 ± 0.89	83.64 ± 1.53	94.66 ± 1.53	95.41 ± 1.05	88.80 ± 1.34	95.92 ± 1.33
RegNetY-8GF	89.36 ± 1.23	83.64 ± 1.61	94.45 ± 0.60	95.13 ± 0.45	88.70 ± 0.80	95.59 ± 1.07
RegNetY-16GF	90.48 ± 1.01	85.54 ± 2.00	94.79 ± 1.54	95.40 ± 1.19	89.91 ± 1.25	96.43 ± 1.03
EfficientNet-B0	87.26 ± 1.74	79.07 ± 3.89	94.38 ± 1.11	95.38 ± 0.68	85.99 ± 2.29	94.81 ± 0.90
EfficientNet-B2	88.23 ± 0.71	81.17 ± 2.42	94.37 ± 1.46	95.20 ± 1.23	87.24 ± 1.21	95.67 ± 0.92
EfficientNet-B4	87.24 ± 1.77	79.11 ± 3.75	94.27 ± 1.30	95.31 ± 0.67	85.98 ± 2.33	94.46 ± 1.22
EfficientNet-B6	87.40 ± 2.38	80.57 ± 4.96	93.41 ± 0.95	94.34 ± 0.95	86.41 ± 2.50	94.53 ± 1.21
EfficientNet-B7	86.41 ± 2.40	78.57 ± 4.95	93.06 ± 0.90	94.21 ± 0.58	85.11 ± 2.94	94.64 ± 1.62
MoCo	89.74 ± 1.29	86.01 ± 3.24	92.92 ± 1.86	93.56 ± 1.53	89.28 ± 1.41	95.70 ± 1.01
DeiT-S	88.89 ± 2.13	82.35 ± 4.19	94.58 ± 1.08	95.38 ± 0.64	87.99 ± 2.54	96.45 ± 0.95
DeiT-B	89.10 ± 2.32	81.76 ± 5.02	95.84 ± 1.13	96.50 ± 0.80	88.13 ± 2.68	96.19 ± 1.49
DeiT-S-D	89.96 ± 1.88	83.44 ± 3.65	95.96 ± 1.06	96.58 ± 0.63	89.21 ± 1.96	97.06 ± 1.07
DeiT-B-D	89.69 ± 1.93	82.51 ± 3.48	**96.29** ± **0.63**	**96.88** ± **0.17**	88.83 ± 2.12	96.61 ± 1.01
CKB-DeiT-S-D	90.27 ± 1.90	84.09 ± 4.00	95.97 ± 1.41	96.59 ± 0.86	89.57 ± 2.04	**97.28** ± **0.91**
CKB-DeiT-B-D	**90.90** ± **1.74**	**86.09** ± **2.98**	95.00 ± 1.29	95.59 ± 0.71	**90.30** ± **1.83**	96.80 ± 1.16

performance for infection classification on all performance metrics except sensitivity. As can be seen from **Table 1**, the proposed CKB-DeiT-B-D performs better than the latest vision transformer DeiT-B-D, and significantly better than other comparison CNN-based methods in terms of all the reported metrics except sensitivity. For instance, the proposed CKB-DeiT-B-D achieves the best F-measure of 78.20 and the best AUC score of 84.78, which are better than the results of 76.72 and 83.26 achieved by the second-best DeiT-B-D, and significantly better than the results of 75.85 and 83.02 achieved by the CNN-based RegNetY-16GF. The proposed CKB significantly improves the precision and specificity of DeiT-B-D, e.g., improving precision from 73.86 to 77.38 and specificity from 71.88 to 77.00. Also, CKB-DeiT-S-D that combines the CKB with the small DeiT with knowledge distillation performs better than the vanilla DeiT-S-D. Although the proposed CKB-DeiT-B-D performs slightly worse in terms of sensitivity, the performance improvements on all the other metrics demonstrate the superiority of the proposed method. In **Figure 5**, we compare the ROC curves of the comparison methods. The methods that achieve the best AUC score over the networks with the same architecture but different layers are selected for comparison. It can be observed from **Figure 5** that our proposed CKB-DeiT-B-D produces a better ROC curve than the comparison methods.

The proposed method also achieves the best accuracy, sensitivity, F-measure, and AUC score on the DFU ischaemia

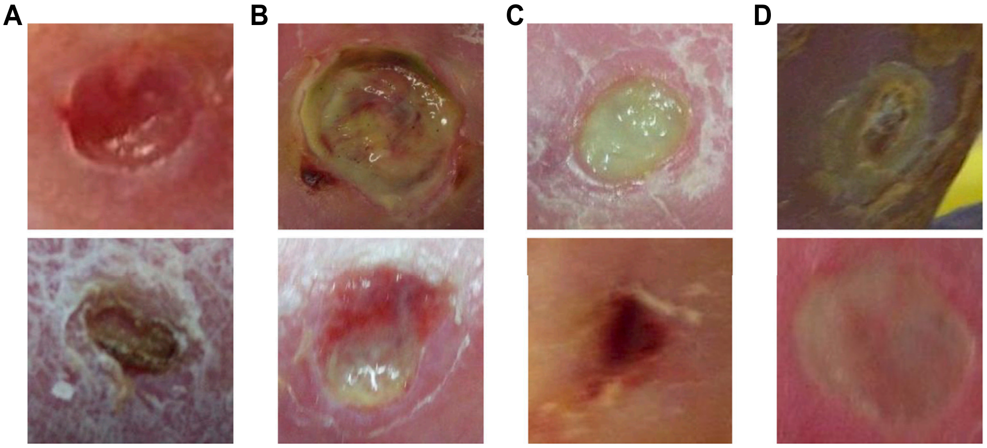


FIGURE 7 | Examples of classification results of the proposed method on the infection dataset. **(A)** true negative cases, **(B)** true positive cases, **(C)** false negative cases and **(D)** false positive cases.

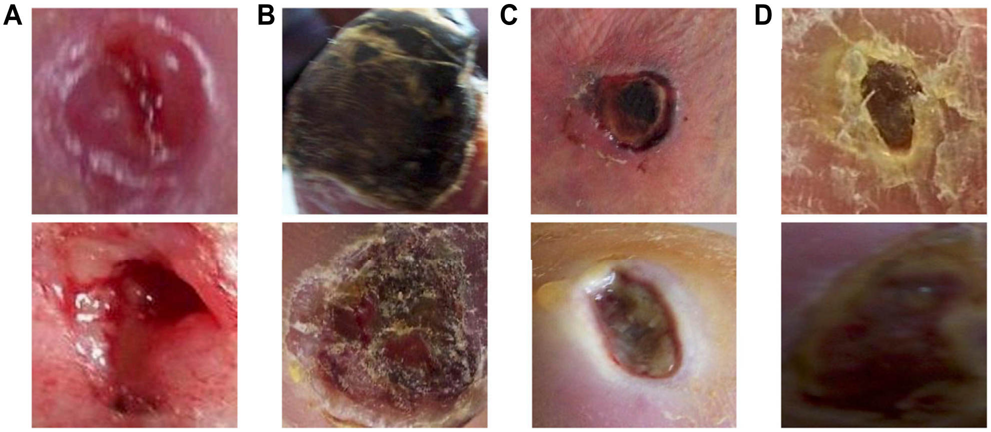


FIGURE 8 | Examples of classification results of the proposed method on the ischaemia dataset. **(A)** true negative cases, **(B)** true positive cases, **(C)** false negative cases and **(D)** false positive cases.

dataset. As shown in **Table 2**, the performances of different methods on the DFU ischaemia dataset are better than their performances on the DFU infection dataset since the characteristics of ischaemia are more discriminative as shown in **Figure 4**. The precision and specificity of the proposed method are better than the CNN-based methods (ResNet, RegNetY, and EfficientNet) and contrastive learning method (MoCo) but inferior to the DeiT-B-D. The comparison methods all seem to produce high precision and specificity but significantly lower accuracy, sensitivity, and F-measure. The proposed CKB-DeiT-B-D produces more balanced results across all the reported metrics. The proposed CKB-DeiT-S-D achieves the best AUC score but the improvement of the ROC curve of our method shown in **Figure 6** is not significant compared with DeiT-S-D. Overall, the proposed CKB using DeiT Touvron et al. (2020) as the encoder achieves the best infection and ischaemia classification performances in terms of most metrics.

5 DISCUSSIONS

The main finding of this research is that better utilization of class knowledge in the training data can improve the performance of DFU image classifications. We have proposed an approach called class knowledge bank which can explicitly and effectively extract class knowledge from the training data and participate in prediction process in the testing. Experimental results have demonstrated the effectiveness of the proposed method in improving classification performances on both DFU infection and ischaemia datasets.

Examples of classification results by the proposed method on the infection and ischaemia datasets are presented in **Figures 7**, **8**, respectively. Correctly classified ulcer images (true negative and true positive) are shown to have discriminative visual characteristics, which are useful for image-based classifications. For instance, true negative non-infection cases in **Figure 7A** are clean and dry, while true positive infection cases in **Figure 7B** are full of yellow secretion. For ischaemia classification examples, it is observed from **Figures 8A,B** that color characteristic is very different between true negative and true positive cases. A close inspection of incorrectly classified cases in **Figures 7C,D**; **Figures 8C,D** suggests that many factors including lighting condition, size of ulcers, secretion, subtle characteristic and model's ability can all affect classification results. This observation means that one needs to carefully consider these factors in real applications.

The proposed method is good at handling class imbalance than the comparison methods. As can be observed from **Tables 1**, **2**, the specificity of the comparison methods is significantly worse than the sensitivity caused by the class imbalance on the infection dataset, while the proposed method can achieve high sensitivity and specificity simultaneously. Also, the proposed method produces more balanced sensitivity and specificity than the comparison methods on the ischaemia dataset. The advantage of the proposed method in handling imbalance data is derived from the structure of the class knowledge banks, where different CKBs have the same units which give the same importance to different classes.

The proposed classification network is based on a pre-trained powerful encoder as training a network from scratch on a relatively small medical image dataset is not efficient. This is a limitation of the proposed network because its performance relies on the pre-trained encoder. We believe that one can achieve better DFU classification performances without relying on a pre-trained encoder when more training data are available. Another limitation is that the proposed method does not consider the contrastive idea in samples in the training data and units in class knowledge banks. Incorporating this idea into the proposed method may further improve DFU classification performances. This paper verifies the performance of the proposed method on the DFU infection and ischaemia datasets. It will be interesting to extend this research to wider areas such as other medical image classification tasks, including binary or multi-class classification problems. The proposed method also has the potential to work as an incremental learning method as we can train additional class knowledge banks for incremental classes. Its performance and characteristics for incremental learning remain further investigation in the future.

6 CONCLUSION

In this paper, we proposed the method called the class knowledge banks (CKBs) which can effectively extract class knowledge from the training data and explicitly leverage the class knowledge in the testing. The proposed method is an alternative means to produce the logits instead of the usual linear classifiers in the literature. The CKBs leverage their units to extract and represent class knowledge from different perspectives and the similarities between the representation of the input and the corresponding CKBs can be regarded as the logits of the input. The CKB can be trained through back-propagation and be easily embedded into existing image classification models. Experimental results on the DFU infection and ischaemia datasets demonstrate the effectiveness of the proposed CKB in DFU image classifications.

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

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported in part by Shanghai University of TCM under Grants 2019LK049 and RCPY0039, in part by Shanghai TCM-Integrated Hospital under Grants 18-01-04 and RCPY0028, and in part by Shanxi S&T under Grant 2021KW-07.

REFERENCES

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). "Understanding of a Convolutional Neural Network," in Proceeding of the 2017 International Conference on Engineering and Technology (ICET), Aug. 2017 (IEEE), 1–6. doi:10.1109/icengtechnol.2017.8308186

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., and Khan, M. K. (2018). Medical Image Analysis Using Convolutional Neural Networks: a Review. *J. Med. Syst.* 42, 226–313. doi:10.1007/s10916-018-1088-1

Armstrong, D. G., Boulton, A. J. M., and Bus, S. A. (2017). Diabetic Foot Ulcers and Their Recurrence. *N. Engl. J. Med.* 376, 2367–2375. doi:10.1056/nejmra1615439

Atlas, D., et al. (2015). *International Diabetes Federation. IDF Diabetes Atlas*. 7th edn.. Brussels, Belgium: International Diabetes Federation.

Cao, Z., Sun, C., Wang, W., Zheng, X., Wu, J., and Gao, H. (2021). Multi-modality Fusion Learning for the Automatic Diagnosis of Optic Neuropathy. *Pattern Recognition Lett.* 142, 58–64. doi:10.1016/j.patrec.2020.12.009

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D. (2018). Drinet for Medical Image Segmentation. *IEEE Trans. Med. Imaging* 37, 2453–2462. doi:10.1109/tmi.2018.2835303

Chen, J., Ying, H., Liu, X., Gu, J., Feng, R., Chen, T., et al. (2020a). A Transfer Learning Based Super-resolution Microscopy for Biopsy Slice Images: the Joint Methods Perspective. *Ieee/acm Trans. Comput. Biol. Bioinform* 18, 103–113. doi:10.1109/TCBB.2020.2991173

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). "A Simple Framework for Contrastive Learning of Visual Representations," in Proceeding of the International Conference on Machine Learning (PMLR) (IEEE), 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. (2020c). *Big Self-Supervised Models Are strong Semi-supervised Learners*. arXiv preprint arXiv: 2006.10029.

Chen, T., Liu, X., Feng, R., Wang, W., Yuan, C., Lu, W., et al. (2021). "Discriminative Cervical Lesion Detection in Colposcopic Images with Global Class Activation and Local Bin Excitation," in Proceeding of the IEEE Journal of Biomedical and Health Informatics, July 2021 (IEEE), 1. doi:10.1109/jbhi.2021.3100367

Cheplygina, V., de Bruijne, M., and Pluim, J. P. W. (2019). Not-so-supervised: a Survey of Semi-supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis. *Med. Image Anal.* 54, 280–296. doi:10.1016/j.media.2019.03.009

Dai, Y., Gao, Y., and Liu, F. (2021). Transmed: Transformers advance Multi-Modal Medical Image Classification. *Diagnostics* 11, 1384. doi:10.3390/diagnostics11081384

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). “Imagenet: A Large-Scale Hierarchical Image Database,” in Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June 2009 (IEEE), 248–255. doi:10.1109/cvpr.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). *An Image Is worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv preprint arXiv:2010.11929.

Feng, R., Liu, X., Chen, J., Chen, D. Z., Gao, H., and Wu, J. (2020). A Deep Learning Approach for Colonoscopy Pathology Wsi Analysis: Accurate Segmentation and Classification. *IEEE J. Biomed. Health Inform.* 25, 3700–3708. doi:10.1109/jbhi.2020.3040269

Gao, H., Xu, K., Cao, M., Xiao, J., Xu, Q., and Yin, Y. (2021). “The Deep Features and Attention Mechanism-Based Method to Dish Healthcare under Social Iot Systems: an Empirical Study with a Hand-Deep Local-Global Net,” in Proceeding of the IEEE Transactions on Computational Social Systems, August 2021 (IEEE), 1–12. doi:10.1109/tcss.2021.3102591

Goyal, M., Hassanpour, S., and Yap, M. H. (2018). *Region of Interest Detection in Dermoscopic Images for Natural Data-Augmentation*. arXiv preprint arXiv:1807.10711.

Goyal, M., Reeves, N. D., Rajbhandari, S., Ahmad, N., Wang, C., and Yap, M. H. (2020). Recognition of Ischaemia and Infection in Diabetic Foot Ulcers: Dataset and Techniques. *Comput. Biol. Med.* 117, 103616. doi:10.1016/j.compbiomed.2020.103616

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316, 2402–2410. doi:10.1001/jama.2016.17216

He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep Residual Learning for Image Recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016 (IEEE), 770–778. doi:10.1109/cvpr.2016.90

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). “Momentum Contrast for Unsupervised Visual Representation Learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 2020 (IEEE), 9729–9738. doi:10.1109/cvpr42600.2020.00975

Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J. Digit Imaging* 32, 582–596. doi:10.1007/s10278-019-00227-x

Hinton, G., Vinyals, O., and Dean, J. (2015). *Distilling the Knowledge in a Neural Network*. arXiv preprint arXiv:1503.02531.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). “Densely Connected Convolutional Networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017 (IEEE), 4700–4708. doi:10.1109/cvpr.2017.243

Ioffe, S., and Szegedy, C. (2015). “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in Proceedings of the International Conference on Machine Learning (PMLR), July 2015 (IEEE), 448–456.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2021). A Survey on Contrastive Self-Supervised Learning. *Technologies* 9, 2. doi:10.3390/technologies9010002

Jeffcoate, W. J., and Harding, K. G. (2003). Diabetic Foot Ulcers. *The Lancet* 361, 1545–1551. doi:10.1016/s0140-6736(03)13169-8

Kingma, D. P., and Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.

Kumar, A., Kim, J., Lyndon, D., Fulham, M., and Feng, D. (2016). An Ensemble of fine-tuned Convolutional Neural Networks for Medical Image Classification. *IEEE J. Biomed. Health Inform.* 21, 31–40. doi:10.1109/JBHI.2016.2635663

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. (2014). “Medical Image Classification with Convolutional Neural Network,” in Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, Dec. 2014 (IEEE), 844–848. doi:10.1109/icarcv.2014.7064414

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* 42, 60–88. doi:10.1016/j.media.2017.07.005

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. arXiv preprint arXiv:2103.14030.

Loshchilov, I., and Hutter, F. (2017). *Decoupled Weight Decay Regularization*. arXiv preprint arXiv:1711.05101.

Nair, V., and Hinton, G. E. (2010). “Rectified Linear Units Improve Restricted Boltzmann Machines,” in Proceedings of the 27 th International Conference on Machine Learning, Haifa, Israel, June 2010, 807–814.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). *Representation Learning with Contrastive Predictive Coding*. arXiv preprint arXiv:1807.03748.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). *Automatic Differentiation in Pytorch*.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). “Designing Network Design Spaces,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 2020 (IEEE), 10428–10436. doi:10.1109/cvpr42600.2020.01044

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). “You Only Look once: Unified, Real-Time Object Detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016 (IEEE), 779–788. doi:10.1109/cvpr.2016.91

Shie, C.-K., Chuang, C.-H., Chou, C.-N., Wu, M.-H., and Chang, E. Y. (2015). “Transfer Representation Learning for Medical Image Analysis,” in Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, Aug. 2015 (IEEE), 711–714. doi:10.1109/EMBC.2015.7318461

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* 35, 1285–1298. doi:10.1109/tmi.2016.2528162

Simonyan, K., and Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv preprint arXiv:1409.1556.

Tan, M., and Le, Q. (2019). “Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks,” in Proceedings of the International Conference on Machine Learning (PMLR), 6105–6114.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2020). *Training Data-Efficient Image Transformers & Distillation through Attention*. arXiv preprint arXiv:2012.12877.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention Is All You Need*. arXiv preprint arXiv:1706.03762.

Veredas, F., Mesa, H., and Morente, L. (2009). Binary Tissue Classification on Wound Images with Neural Networks and Bayesian Classifiers. *IEEE Trans. Med. Imaging* 29, 410–427. doi:10.1109/TMI.2009.2033595

Walsh, J. W., Hoffstad, O. J., Sullivan, M. O., and Margolis, D. J. (2016). Association of Diabetic Foot Ulcer and Death in a Population-Based Cohort from the united kingdom. *Diabet. Med.* 33, 1493–1498. doi:10.1111/dme.13054

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., et al. (2017). “Residual Attention Network for Image Classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017 (IEEE), 3156–3164. doi:10.1109/cvpr.2017.683

Wang, L., Pedersen, P. C., Agu, E., Strong, D. M., and Tulu, B. (2017). Area Determination of Diabetic Foot Ulcer Images Using a Cascaded Two-Stage Svm-Based Classification. *IEEE Trans. Biomed. Eng.* 64, 2098–2109. doi:10.1109/TBME.2016.2632522

Wang, Z., Liu, Q., and Dou, Q. (2020). Contrastive Cross-Site Learning with Redesigned Net for Covid-19 Ct Classification. *IEEE J. Biomed. Health Inform.* 24, 2806–2813. doi:10.1109/jbhi.2020.3023246

Wannous, H., Lucas, Y., and Treuillet, S. (2010). Enhanced Assessment of the Wound-Healing Process by Accurate Multiview Tissue Classification. *IEEE Trans. Med. Imaging* 30, 315–326. doi:10.1109/TMI.2010.2077739

Wei, L., Xiao, A., Xie, L., Chen, X., Zhang, X., and Tian, Q. (2020). *Circumventing Outliers of Autoaugment with Knowledge Distillation*. arXiv preprint arXiv:2003.11342 2.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). "Unsupervised Feature Learning via Non-parametric Instance Discrimination," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018 (IEEE), 3733–3742. doi:10.1109/cvpr.2018.00393

Xiao, J., Xu, H., Gao, H., Bian, M., and Li, Y. (2021). A Weakly Supervised Semantic Segmentation Network by Aggregating Seed Cues: The Multi-Object Proposal Generation Perspective. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 1–19. doi:10.1145/3419842

Yadav, S. S., and Jadhav, S. M. (2019). Deep Convolutional Neural Network Based Medical Image Classification for Disease Diagnosis. *J. Big Data* 6, 1–18. doi:10.1186/s40537-019-0276-2

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Tay, F. E., et al. (2021). *Tokens-to-token Vit: Training Vision Transformers from Scratch on Imagenet*. arXiv preprint arXiv:2101.11986.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. (2019). Object Detection with Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 3212–3232. doi:10.1109/tnnls.2018.2876865

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.

Advantages of publishing in Frontiers

OPEN ACCESS

Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION

Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices

FOLLOW US

@frontiersin

IMPACT METRICS

Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK

Our network
increases your
article's readership