Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

An interpretability model for syndrome differentiation of HBV-ACLF in traditional Chinese medicine using small-sample imbalanced data

ZHOU Zhan[a], PENG Qinghua[b*], XIAO Xiaoxia[a*], ZOU Beiji[a], LIU Bin[a], GUO Shuixia[c]

a. School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

b. School of Traditional Chinese Medicine, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China

c. School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, China

ARTICLE INFO

Article history
Received 18 January 2024
Accepted 30 May 2024
Available online 25 June 2024

Keywords
Traditional Chinese medicine (TCM)
Hepatitis B-related acute-on-chronic liver failure (HBV-ACLF)
Imbalanced data
Random forest (RF)
Interpretability

ABSTRACT

Objective Clinical medical record data associated with hepatitis B-related acute-on-chronic liver failure (HBV-ACLF) generally have small sample sizes and a class imbalance. However, most machine learning models are designed based on balanced data and lack interpretability. This study aimed to propose a traditional Chinese medicine (TCM) diagnostic model for HBV-ACLF based on the TCM syndrome differentiation and treatment theory, which is clinically interpretable and highly accurate.

Methods We collected medical records from 261 patients diagnosed with HBV-ACLF, including three syndromes: Yang jaundice (214 cases), Yang-Yin jaundice (41 cases), and Yin jaundice (6 cases). To avoid overfitting of the machine learning model, we excluded the cases of Yin jaundice. After data standardization and cleaning, we obtained 255 relevant medical records of Yang jaundice and Yang-Yin jaundice. To address the class imbalance issue, we employed the oversampling method and five machine learning methods, including logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF), and extreme gradient boosting (XGBoost) to construct the syndrome diagnosis models. This study used precision, F1 score, the area under the receiver operating characteristic (ROC) curve (AUC), and accuracy as model evaluation metrics. The model with the best classification performance was selected to extract the diagnostic rule, and its clinical significance was thoroughly analyzed. Furthermore, we proposed a novel multiple-round stable rule extraction (MRSRE) method to obtain a stable rule set of features that can exhibit the model's clinical interpretability.

Results The precision of the five machine learning models built using oversampled balanced data exceeded 0.90. Among these models, the accuracy of RF classification of syndrome types was 0.92, and the mean F1 scores of the two categories of Yang jaundice and Yang-Yin jaundice were 0.93 and 0.94, respectively. Additionally, the AUC was 0.98. The extraction rules of the RF syndrome differentiation model based on the MRSRE method revealed that the common features of Yang jaundice and Yang-Yin jaundice were wiry pulse, yellowing of the urine, skin, and eyes, normal tongue body, healthy sublingual vessel, nausea, oil loathing, and poor appetite. The main features of Yang jaundice were a red tongue body and thickened sublingual vessels, whereas those of Yang-Yin jaundice were a dark tongue body, pale white tongue body, white tongue coating, lack of strength, slippery pulse, light red tongue body, slimy

*Corresponding author: XIAO Xiaoxia, E-mail: amily_x@hnucm.edu.cn. PENG Qinghua, E-mail: pqh410007@126.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2024.09.005
Citation: ZHOU Z, PENG QH, XIAO XX, et al. An interpretability model for syndrome differentiation of HBV-ACLF in traditional Chinese medicine using small-sample imbalanced data. Digital Chinese Medicine, 2024, 7(2): 137-147.

Copyright © 2024 The Authors. Production and hosting by Elsevier B.V. This is an open access article under the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

tongue coating, and abdominal distension. This is aligned with the classifications made by TCM experts based on TCM syndrome differentiation and treatment theory.

Conclusion Our model can be utilized for differentiating HBV-ACLF syndromes, which has the potential to be applied to generate other clinically interpretable models with high accuracy on clinical data characterized by small sample sizes and a class imbalance.

1 Introduction

The inability of the liver to perform its normal metabolic functions during liver failure can be lethal [1]. Four forms of liver failure are recognized: acute liver failure, subacute liver failure, acute-on-chronic liver failure (ACLF), and chronic liver failure (CLF). In China, liver failure is mostly caused by hepatitis B, mainly through ACLF and CLF forms. Hepatitis B-related ACLF (HBV-ACLF) is characterized by acute decompensation of chronic liver disease, resulting in multiple organ failure and high short-term mortality [2]. In traditional Chinese medicine (TCM), this disease is classified into Yang jaundice, Yang-Yin jaundice, and Yin jaundice. In current clinical study, it was found that Yang-Yin jaundice, which has characteristics of both Yang and Yin jaundice, is considered an intermediate form as Yang jaundice develops into Yin jaundice. And, treatment efficiency was improved to approximately 76.2%, and mortality was decreased by 10.0% when patients were treated with TCM and western medicine compared with those treated by western medicine alone [3].

In TCM, multi-dimensional information, such as self-reported symptoms, physical signs, and tongue and pulse data, is needed for effective diagnosis and syndrome differentiation. Machine learning, a method used to construct classification models of complex data, has been widely used in TCM clinics for syndrome differenti ation [4-6]. Classification models commonly used in TCM include logistic regression (LR) [7], support vector machine (SVM) [8], decision tree (DT) [9], random forest (RF) [10, 11], and extreme gradient boosting (XGBoost) [12]. These machine learning methods are conducive to improving the accuracy and efficiency of TCM diagnosis, providing new insights into the field of TCM.

There are currently two problems with the application of machine learning in TCM diagnosis. On one hand, the clinical data of HBV-ACLF in TCM are characterized by small and unbalanced sample sizes. Machine learning models tend to predict the majority class when the data are imbalanced, which makes the prediction results poor. On the other hand, in high-risk medical decision-making fields, only models with high classification accuracy and clinical interpretability can meet the needs of TCM [13]. To address the problem of class imbalance, the current solutions mainly include data sampling [14-19], feature selection [20-23], and optimization algorithms [24-32]. Oversampling, as a technique, resamples the minority class proportion to follow the majority class proportion and increases the amount of data. Therefore, data oversampling was used to alleviate the unbalanced small-sample problem in this study. For the problem of interpretability, some self-explanatory models, such as linear regression and decision trees, used in TCM, are readily interpretable. However, these models designed to improve accuracy tend to be complex, which limits their interpretability. A balance between model interpretability and accuracy is required [33]. To address this problem, Shapley Additive exPlanations (SHAP) provides a solution by calculating the Shapley value to determine the contribution of each feature to the predicted output [34] . However, this method only explains the model in a black-box way, but cannot explain the internal structure of the model. In terms of data sampling combined with interpretability, considering that each sampling method produces different datasets, resulting in different models and classification rule sets, identifying these rule sets may be challenging for TCM experts. However, there are few studies on this issue.

To solve the above problems and build a high-precision TCM syndrome differentiation model, we proposed multiple-round stable rule extraction (MRSRE), which is an interpretability method based on the internal rules of the ensemble tree models. This method uses the characteristics of the decision tree structure to extract classification rules. These rules reflect the internal structure of the model and help understand which features affect the predicted output. Regarding the impact of sampling methods on classification rule sets, we assume that if the rule set obtained by the interpretability method converges, it is considered stable. To obtain stable rule sets, we applied the MRSRE method to obtain stable high-frequency feature sets. The advantage of our method is that the algorithm is relatively simple and can obtain a stable rule set after multiple rounds of oversampling, which has been verified using the HBV-ACLF dataset.

This study aims to design a clinically highly accurate and interpretable TCM diagnosis model based on the theory of TCM syndrome differentiation and treatment, which can provide the same diagnostic process and outcomes like TCM experts.

2 Materials and methods

2.1 Data source and standardization

2.1.1 Data availability (i) Data sources. This retrospective study obtained data from 261 patients with HBV-ACLF who were hospitalized between January 1, 2007 and

December 31, 2015, at the Hepatology Department of The First Hospital of Hunan University of Chinese Medicine. A systematic research and development project was supported by the "Key Special Project for the Modernization of Traditional Chinese Medicine" in the 13th Five-Year Plan of the Ministry of Science and Technology (ethical approval number: 2018-626-55-01, approved by the Institutional Review Board of Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine). All subjects filled in medical record data after informed consent.

Patients who met the diagnostic criteria for HBV-ACLF (subacute) were included. Inclusion and exclusion criteria are based on guideline for clinical diagnosis and treatment of liver failure [1].

(ii) TCM diagnostic criteria. The data contained 214, 41, and 6 cases of Yang, Yang-Yin, and Yin jaundice, respectively. The TCM syndrome differentiation standards for Yang, Yang-Yin, and Yin jaundice refer to the relevant syndrome differentiation standards in the *Internal Medicine of Traditional Chinese Medicine* [35], *Traditional Chinese Medicine Diagnostics* [36], Clinic Terminology of Traditional Chinese Medical Diagnosis and Treatment — Part 2: Syndromes/Patterns [37], and the Diagnosis and Treatment Plan for the Advantageous Diseases of the Department of TCM Hepatology of The First Hospital of Hunan University of Chinese Medicine. Detailed syndrome differentiation requirements: a patient with three main symptoms, or two main symptoms and two secondary symptoms, can be classified as having a certain syndrome. The main and secondary symptoms and signs of Yang jaundice, Yang-Yin jaundice, and Yin jaundice are as follows.

(a) Yang jaundice. The main symptoms and signs of Yang jaundice are yellow and bright skin and eyes, red or crimson tongue with ecchymosis and petechiae, yellow and greasy tongue coating, and strong and slippery pulse. The secondary symptoms and signs include dry mouth, bitter taste in the mouth, or nausea and vomiting, constipation, nose and teeth bleeding or skin ecchymosis, lack of urine, and yellowish-red urine.

(b) Yang-Yin jaundice. The main symptoms and signs of Yang-Yin jaundice are: yellow skin and eyes with bright or dark yellow, light red or slightly red tongue with ecchymosis and petechiae or teeth-printed, white and greasy tongue coating, thick greasy or light yellow coating, wiry or slippery pulse, or deep pulse. The secondary symptoms and signs are abdominal distension, loose stools, or nausea and vomiting, dry mouth or lack of desire to drink or not drinking much, lack of strength, and loss of appetite.

(c) Yin jaundice. The main symptoms and signs of Yin jaundice are dark yellow or smoky skin, pale tongue, and white and greasy tongue coating. The secondary symptoms and signs include epigastric tightness, abdominal distension, or lack of appetite, fatigue and fear of cold, tasteless mouth and lack of thirst, and deep or thready pulse.

(iii) Excluding Yin jaundice samples. Due to the limited amount of available data on Yin jaundice, using these data to build a machine learning model would result in overfitting. For example, the sample size of Yin jaundice and Yang jaundice is 220, of which there are 214 cases of Yang jaundice and only 6 cases of Yin jaundice. Even if the model misclassifies all 6 cases of Yin jaundice as Yang jaundice, the overall accuracy rate of classification would still be 97.27%. However, this accuracy rate fails to reflect the classifier's performance for the minority class. Increasing the sample size of Yin jaundice through oversampling would still lead to overfitting. In other words, oversampling the minority class cannot modify the features to accurately represent the real clinical world, thereby limiting the model's generalization ability. This is because the generation of new data in oversampling algorithms is based on the original data. These algorithms generally copy the original data or generate new samples based on the distribution of the original data. If the number of samples in the minority class is too small, the model may overfit by paying too much attention to the samples in the majority class and ignoring those in the minority class. The oversampling method, on the other hand, increases the number of samples in the minority class, which can lead to overfitting as the model becomes overly adapted to the characteristics of the minority class. Thus, we excluded 6 cases of Yin jaundice, leaving 255 cases. From these cases, 149 self-reported symptoms and 124 features of tongue and pulse data were manually extracted by the medical students from medical records. The dataset is characterized by high dimensionality, a small sample size, and imbalanced data categories.

(iv) Sample analysis. The age of the 255 patients with HBV-ACLF ranged from 14 to 75 years (mean 39.57 ± 10.96 years), and there were more men than women (233 vs. 22). However, the cases included in this study came from the hospitalization data of the Hepatology Department of The First Hospital of Hunan University of Chinese Medicine spanning eight years, and the data were not screened based on gender. Therefore, these data may indicate that more men than women suffer from the disease. Table 1 shows that there were more young and middle-aged patients than older patients. We performed statistical analysis using logistic regression and observed that the P value between gender and predicted results was 0.86, without statistically significant differences ($P >$ 0.05). This means that gender has no significant impact on the prediction results under the current sample size and the statistical methods used.

Given that Yang-Yin jaundice is an intermediate TCM syndrome between Yang and Yin jaundice, we used Formulas (1) and (2) to calculate the frequency difference of each symptom or sign under different syndrome types.

Table 1 Age distribution of patients in the study

Age (year)	Number of patients
10 – 19	2
20 – 29	48
30 – 39	84
40 – 49	76
50 – 59	33
60 – 69	10
70 – 79	2

The feature difference score of the two syndromes is less than 0.1, and the main differences are in the tongue and pulse conditions (Figure 1).

$$\text{Frequency of symptom or sign} = \frac{\text{Number of a symptom or sign}}{\text{Number of a syndrome or sign}} \tag{1}$$

$$\text{Difference of symptom or sign} = \frac{\text{Frequncy of symptom or sign of } A}{\text{Total frequency of symptoms and signs}} - \frac{\text{Frequency of symptom or sign of } B}{\text{Total frequency of symptoms and signs}} \tag{2}$$

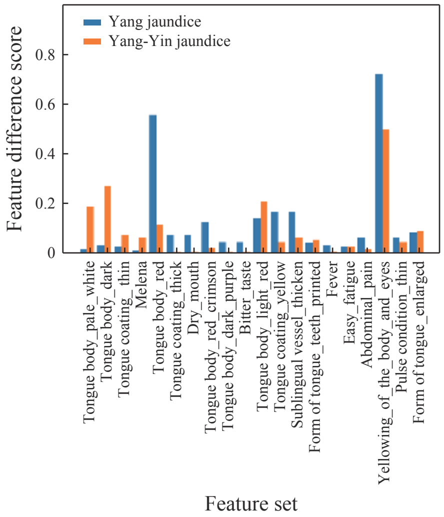

Figure 1 Comparison of the top 20 symptoms with tongue, pulse, and symptom frequency differences

2.1.2 Data standardization We did not use dimensionality reduction methods commonly used in machine learning because we could achieve high accuracy and precision simply by removing some missing values and merging linguistically ambiguous data through data standardization, while preserving all available features.

(i) Standardization references. The main standardization references were the *Internal Medicine of Traditional Chinese Medicine* [35], *Traditional Chinese Medicine Diagnostics* [36], Clinic Terminology of Traditional Chinese Medical Diagnosis and Treatment—Part 2: Syndromes/Patterns [37], Guidelines for Clinical Diagnosis and Treatment of Acute-on-Chronic Liver Failure in Traditional Chinese Medicine [38], and Guideline for Diagnosis and Treatment of Liver Failure [1].

(ii) Standardization implementation process and results. We standardized the dataset according to the standardized references, including processing missing values, compound phrases, and polysemy. The details are as follows.

(a) Missing values. Missing values refer to the absence of a particular attribute or feature for a given data point. Missing values can cause bias or instability in the model learning process. If the percentage of missing values for a feature is very high (e.g., greater than 0.90), then the feature is unlikely to provide useful information to the model. At the same time, if a sample has many missing features, retaining the sample may introduce bias. In this study, features with a missing value percentage of more than 0.90 and samples with a missing value percentage of more than 0.50 were removed.

(b) Compound phrases. A feature with compound phrases is a combination of multiple symptoms, such as "yellowing of the skin and eyes, yellow urine, and fatigue", which violates the Feature Independence Principle. Thus, we separated such features from the rest (Supplementary Table S1).

(c) Polysemy. Polysemy in TCM clinics means that a symptom may have different expression terms. We normalized multiple words with the same meaning and combined different expressions of the same symptom into one term (Supplementary Table S2).

The initial 149 entries were normalized into 57 symptoms [39] (Supplementary Table S3). Then, tongue signs were categorized into tongue, tongue body, tongue coating, and sublingual vessels, while pulse signs were categorized into pulse conditions, including eight independent pulse signs. Finally, a total of 124 tongue and pulse signs were standardized into 33 features (Supplementary Table S4).

2.2 Characteristics of the methods

Self-reported symptoms and tongue and pulse signs in the TCM records were used as feature inputs, while diagnostic syndromes were used as output labels. The syndrome differentiation problem was converted into a classification problem.

2.2.1 Data oversampling Four oversampling methods were selected: random oversampling, Synthetic Minority Over-sampling Technique (SMOTE), borderline-SMOTE, and SMOTE-D. As the samples generated by each method can be different, we applied each oversampling method six times, and the resulting datasets were used as the model inputs. The oversampling steps were as follows: (i) the entire dataset, designated as P, which included n samples after preprocessing, was utilized. This dataset encompassed both the feature sets and the associated classification labels; (ii) a sampling ratio of 1 : 1 was used to make the number of minority class samples the same as the majority class samples; (iii) four sampling techniques were used to oversample the minority class data in the dataset six times, and 24 datasets were obtained; (iv) the 24 datasets had the same sample size and comprised features and classification labels.

2.2.2 Syndrome differentiation model construction We used common classifiers (LR, SVM, and DT) and ensemble learning classifiers (RF and XGBoost). To improve classifier performance, we chose Optuna (https://github.com/optuna/optuna) for the hyperparameter selection of machine learning classifiers. To strengthen the effectiveness of small-sample training, we adopted ten-fold cross-validation to divide the oversampled medical record samples. The details are shown in Algorithm 1.

```
Algorithm 1: syndrome differentiation model construction
Input: training data set after oversampling
$S = \{S_1, S_2, S_3, S_4\}$, $S_i = \{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6\}$,
$S'_j = \{(p'_1, y'_1), (p'_2, y'_2), \dots, (p'_n, y'_n)\}$
The oversampled dataset corresponding to the selected five classifiers was iteratively tuned using Optuna for 1 000 rounds to obtain optimal parameters.
Begin
$s$ = number of training sets after oversampling
$l = \text{len}(S)$
for $i = 1$ to $l${
    for $j = 1$ to $s${
        Set training samples.
        Set the optimal parameters of the model corresponding to the training sample.
        Each dataset was divided into sample sets using ten-fold cross-validation.
        Train classifier and obtain 10 results of accuracy, precision, F1 score, and AUC.
    }
    Take the mean and variance of the results to obtain each model evaluation metrics.
}
End
Output: mean and variance of accuracy, precision, F1 score, and AUC, respectively.
```

2.2.3 Multiple round stable rule extraction Rule extraction after oversampling should account for two issues: one is the stability of the rule set, and the other is how the representation can meet the domain requirements. As the samples obtained by multiple SMOTE oversampling of the same dataset are different, the decision paths of the ensemble decision tree model are naturally different. Owing to the randomness of the ensemble decision tree model, even if the models are built multiple times using the same sample, their classification rule sets will still be different. Therefore, we need to ensure that the rule set does not change with oversampling by obtaining a stable rule set. The purpose of the model interpretability analysis is to make the model understandable to humans and to adapt to the needs of the domain. In clinical medicine, it is often challenging to distinguish between Yang jaundice and Yang-Yin jaundice. Therefore, the purpose of the interpretability of the TCM syndrome differentiation model is to obtain the basis for the model to classify diseases according to characteristics (symptoms and signs), and this basis should be consistent with the views of TCM experts to a certain extent.

Therefore, the interpretation results should meet this clinical need of TCM. To address the above problems, we designed the MRSRE architecture with two parts: ensemble tree rule extraction and high-frequency overlapping feature extraction.

(i) Ensemble tree rule extraction. The ensemble tree model has natural interpretability because it integrates decision trees, where each path from the root to the leaf nodes represents a decision path. Each decision path represents a disease diagnosis process, which means starting from the root node of the decision tree, making decisions based on the conditions at each node (for example, whether the patient has specific symptoms or signs), and gradually traversing the tree structure downwards until reaching a leaf node (i.e., the end of the decision tree). This process simulates the thinking process of a doctor when diagnosing a disease, that is, possible diseases are gradually eliminated or confirmed based on the patient's symptoms and signs, and a diagnosis result is finally obtained. To understand the decision-making process of the model, we need to obtain the decision path set, also known as the classification rule set. In the context of disease diagnosis, a decision tree model denotes multiple possible disease diagnosis processes. By traversing each subtree, we can extract the entire subtree classification rule set. These rules are employed to reveal the decisions of the ensemble tree. However, this rule extraction method may yield an excessive number of rules, making it difficult to comprehend the model effectively. To obtain a concise rule set, we filtered the rules based on their frequency, length, and error.

The rule extraction steps of the ensemble tree model were as follows: (a) the clinical syndrome differentiation model was trained, and optimized parameters were used in the models to maintain high classification performance; (b) all decision paths were extracted from the root to the leaf nodes of all decision trees; (c) duplicate rules were removed to obtain the decision rule set of the ensemble decision tree; and (d) length, error, and frequency were adopted to measure the statistical characteristics of the rules [40]. The rules with length, error, and frequency under certain thresholds were selected as the final extraction rules. Rule length represents the number of features or the complexity of the rule. Error reflects the correctness of the rule, and frequency indicates the sample size that satisfies the rule. The error and frequency were calculated using the following Formulas (3) and (4).

$$\text{Error} = \frac{\text{A rule determines the number of wrong samples}}{\text{The number of samples that match the rule}} \tag{3}$$

$$\text{Frequency} = \frac{\text{The number of samples that match the rule}}{\text{All sample sizes}} \tag{4}$$

(ii) High-frequency overlapping feature extraction. To ensure the stability of rule sets after oversampling, we obtained multiple ensemble decision tree models and rule sets by repeating oversampling and training models multiple times. We assume that we conduct n experiments to obtain the rule set, and the rule set obtained at the rth time ($r < n$) remains basically unchanged, which we call rule set convergence. They are considered to have good stability if the rule set converges, resulting in fixed overlapping features in the model built after multiple oversampling; otherwise, the stability is poor. Distinguishing between the two types of diseases requires understanding of their similarities and differences. To meet this requirement, we can know what their main symptoms and signs are independent by finding the high-frequency overlapping features of the two syndrome types. Moreover, we can know which features can distinguish the two diseases by taking the different sets of their high-frequency overlapping features to obtain distinguishing features. The detailed steps of repeated feature extraction are shown in Algorithm 2.

```
Algorithm 2: high-frequency overlapping feature extraction
Input: the repetition number of m rounds of rule extraction, the number $k$ represents the first $k$ rules.
Begin
$S$ = []
for $i$ = 1 to $m${
    classification rules = classification rules were extracted from every ensemble decision tree train of the oversampled data according to the ensemble decision tree extraction step.
    $S_i$= classification rules
    $S$ = $S$.append($S_i$)
}
    $S'$ = []
for $i$ = 1 to $m${
        $S'_i$ = merged duplicate rules and frequency of rules in $S_i$.
        frequency = the frequency of duplicate rules in $S'_i$
        $S'_i$ = rules were sorted in descending order by frequency, and the top $k$ rules were filtered to get a total of $k \times m$ rule sets
        $S'$.append($S'_i$)
    }
overlapping_features = the overlapping feature set was obtained by finding the intersection of the features extracted from the rules in $S'_i$
difference_features = the difference set of the overlapping features was set under different category labels in $S'_i$
End
Output: overlapping_features, difference_features
```

3 Results

3.1 Classification results

3.1.1 Raw data classification results We used LR, SVM, RF, DT, and XGBoost to build disease classification models on the unsampled dataset. Accuracy, precision, F1 score, and AUC were adopted as evaluation metrics (Table 2). Based on the original data, these models were used to build classification models. The classification performance for the minority class samples, specifically Yang-Yin jaundice, was significantly compromised by the class-imbalance issue. Across all machine learning models evaluated, the F1 scores were disappointingly low, with none exceeding 0.33 (Table 2).

3.1.2 Classification results after oversampling Since the dataset featured class imbalance, we oversampled the

Table 2 Classification results of syndrome differentiation models based on raw data [mean ± standard deviation (SD)]

Model	Jaundice type	Precision	F1 score	AUC	Accuracy
LR	Yang	0.84 ± 0.01	0.91 ± 0.01	0.68 ± 0.00	0.84 ± 0.01
	Yang-Yin	0.00 ± 0.00	0.00 ± 0.00		
SVM	Yang	0.84 ± 0.01	0.91 ± 0.01	0.68 ± 0.00	0.84 ± 0.01
	Yang-Yin	0.00 ± 0.00	0.00 ± 0.00		
RF	Yang	0.86 ± 0.03	0.91 ± 0.03	0.77 ± 0.01	0.83 ± 0.04
	Yang-Yin	0.38 ± 0.43	0.25 ± 0.28		
DT	Yang	0.87 ± 0.04	0.88 ± 0.03	0.68 ± 0.04	0.80 ± 0.09
	Yang-Yin	0.33 ± 0.30	0.31 ± 0.25		
XGBoost	Yang	0.88 ± 0.04	0.89 ± 0.03	0.72 ± 0.00	0.80 ± 0.04
	Yang-Yin	0.36 ± 0.28	0.32 ± 0.22		

dataset before training the classifiers and then used five classifiers to classify the dataset to improve model performance. The final experimental results are summarized in Table 3. We applied four oversampling methods to oversample the data for six rounds as the dataset to build the classification models. Additionally, we used Optuna and 10-fold cross-validation to potentiate the classification performance, resulting in precision above 90% for each model (Table 3).

3.2 MRSRE results

The RF + SMOTE model gave the best results, so we used the MRSRE method to perform interpretability analysis on this model. To obtain compact and non-redundant classification rules, the filtering conditions for the classification rules were set as a rule length ⩽ 10, frequency > 0.03, and error < 0.05. Four rounds of rule extraction were conducted in this experiment to obtain stable features,

Table 3 Classification results of syndrome differentiation models based on oversampled data (mean ± SD)

Model	Jaundice type	Precision	F1 score	AUC	Accuracy
LR + RO	Yang	0.92 ± 0.05	0.87 ± 0.05	0.87 ± 0.00	0.86 ± 0.05
	Yang-Yin	0.85 ± 0.06	0.88 ± 0.04		
LR + SMOTE	Yang	0.94 ± 0.06	0.85 ± 0.06	0.86 ± 0.00	0.87 ± 0.05
	Yang-Yin	0.82 ± 0.06	0.88 ± 0.05		
LR + Borderline-SMOTE	Yang	0.91 ± 0.06	0.86 ± 0.05	0.89 ± 0.00	0.86±0.05
	Yang-Yin	0.84 ± 0.06	0.88 ± 0.05		
LR + SMOTE-D	Yang	0.92 ± 0.06	0.87 ± 0.05	0.92 ± 0.00	0.88 ± 0.04
	Yang-Yin	0.85 ± 0.06	0.89 ± 0.05		
SVM + RO	Yang	0.96 ± 0.05	0.91 ± 0.04	0.92 ± 0.00	0.94 ± 0.04
	Yang-Yin	0.88 ± 0.04	0.92 ± 0.04		
SVM + SMOTE	Yang	1.00±0.02	0.93 ± 0.05	0.96 ± 0.00	0.94 ± 0.04
	Yang-Yin	0.90 ± 0.06	0.94 ± 0.03		
SVM + Borderline-SMOTE	Yang	0.96 ± 0.05	0.91 ± 0.04	0.96 ± 0.00	0.93 ± 0.04
	Yang-Yin	0.88 ± 0.06	0.92 ± 0.04		
SVM + SMOTE-D	Yang	0.96 ± 0.04	0.91 ± 0.04	0.95 ± 0.00	0.91 ± 0.04
	Yang-Yin	0.88 ± 0.05	0.92 ± 0.03		
RF + RO	Yang	0.95 ± 0.05	0.92 ± 0.01	0.99 ± 0.01	0.92 ± 0.04
	Yang-Yin	0.91 ± 0.06	0.93 ± 0.01		
RF + SMOTE	Yang	0.98 ± 0.04	0.93 ± 0.05	0.98 ± 0.01	0.92 ± 0.04
	Yang-Yin	0.91 ± 0.06	0.94 ± 0.04		
RF + Borderline-SMOTE	Yang	0.94 ± 0.05	0.91 ± 0.04	0.97 ± 0.00	0.92 ± 0.04
	Yang-Yin	0.90 ± 0.05	0.92 ± 0.04		
RF + SMOTE-D	Yang	0.93 ± 0.05	0.91 ± 0.04	0.97 ± 0.00	0.91 ± 0.04
	Yang-Yin	0.90 ± 0.05	0.92 ± 0.03		
DT + RO	Yang	0.94 ± 0.05	0.90 ± 0.05	0.92 ± 0.01	0.92 ± 0.04
	Yang-Yin	0.87 ± 0.06	0.91 ± 0.04		
DT + SMOTE	Yang	0.99 ± 0.02	0.92 ± 0.04	0.92 ± 0.02	0.92 ± 0.04
	Yang-Yin	0.88 ± 0.06	0.93 ± 0.03		
DT + Borderline-SMOTE	Yang	0.93 ± 0.05	0.90 ± 0.05	0.93 ± 0.01	0.91 ± 0.04
	Yang-Yin	0.87 ± 0.05	0.90 ± 0.04		
DT + SMOTE-D	Yang	0.93 ± 0.05	0.88 ± 0.04	0.91 ± 0.01	0.89 ± 0.05
	Yang-Yin	0.86 ± 0.05	0.90 ± 0.03		
XGboost + RO	Yang	0.94 ± 0.05	0.89 ± 0.05	0.99 ± 0.00	0.92 ± 0.04
	Yang-Yin	0.87 ± 0.06	0.90 ± 0.04		
XGboost + SMOTE	Yang	0.98 ± 0.03	0.91 ± 0.05	0.93 ± 0.00	0.91 ± 0.05
	Yang-Yin	0.88 ± 0.07	0.92 ± 0.04		
XGboost + Borderline-SMOTE	Yang	0.93 ± 0.06	0.89 ± 0.05	0.92 ± 0.00	0.91 ± 0.04
	Yang-Yin	0.87 ± 0.06	0.90 ± 0.04		
XGboost + SMOTE-D	Yang	0.93 ± 0.05	0.89 ± 0.06	0.93 ± 0.00	0.90 ± 0.04
	Yang-Yin	0.87 ± 0.08	0.90 ± 0.05		

with models constructed 200, 400, 600, and 800 times in each round. The final overlapping feature sets of Yang and Yang-Yin jaundice syndromes obtained from the rule sets extracted in each round are presented in Table 4 and 5, respectively. The overlapping feature sets remained unchanged after the number of models constructed reached 400.

Table 4 RF classification of overlapping features of Yang jaundice

Feature	Time				Frequency
	200	400	600	800	
Pulse (wiry)	√	√	√	√	33
Yellowing of urine	√	√	√	√	26
Yellowing of the skin and eyes	√	√	√	√	15
Tongue body (red)	√	√	√	√	13
Form of tongue (healthy)	√	√	√	√	13
Sublingual vessel (healthy)	√	√	√	√	7
Nausea	√	/	√	√	6
Oil loathing	√	√	/	√	4
Sublingual vessel (thickening)	√	/	/	/	1

Frequency represents the number of times that the features from the first 10 rule sets in each round of iterations appear in the 10 rule sets after 200, 400, 600, and 800 iterations, following filtering, and frequency sorting. "√" represents the presence of the feature in the iteration. "/" represents the absence of the feature in the iteration.

Table 5 RF classification of overlapping features of Yang-Yin jaundice

Feature	Time				Frequency
	200	400	600	800	
Form of tongue (healthy)	√	√	√	/	15
Pulse (wiry)	√	√	√	√	11
Tongue body (dark)	√	√	√	√	10
Yellowing of skin and eyes	√	√	√	√	9
Sublingual vessel (healthy)	√	√	√	√	9
Tongue body (pale white)	√	√	√	√	9
Yellowing of urine	/	√	√	√	8
Tongue coating (white)	/	√	√	√	7
Lack of strength	/	√	√	√	6
Pulse condition (slippery)	/	√	√	√	5
Tongue body (light red)	/	√	√	√	5
Tongue coating (slimy)	/	√	/	√	4
Abdominal distension	/	√	√	√	4
Oil loathing	/	√	√	√	3
Nausea	√	√	/	/	2

"√" represents the presence of the feature in the iteration. "/" represents the absence of the feature in the iteration.

3.3 The top 20 features from the RF classification model

The important features were determined by ranking the contributions of all model features to the classification results. We selected the top 20 features that contributed the most to the classification results from a total of 90 features, including 57 symptoms and 33 tongue and pulse conditions. The results reveal that 13 out of the 20 important features extracted from the RF + SMOTE model were related to tongue and pulse signs, while the remaining 7 were symptoms (Figure 2). These results highlight the significance of inter-syndrome differences in tongue and pulse signs.

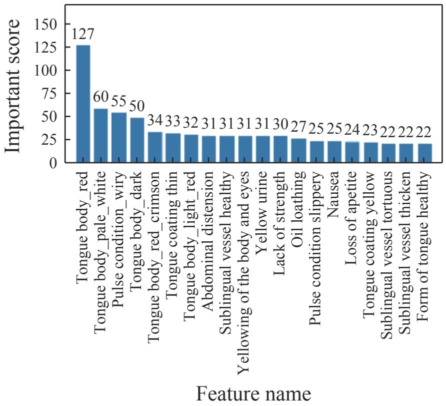

Figure 2 The top 20 most important features of the RF model

4 Discussion

4.1 Sample and important features of the RF model

Based on the classification results in Table 3, the prediction results of each model after applying SMOTE oversampling are not markedly different. The classification F1 scores of the RF + SMOTE model in the majority class and minority class are 0.94 and 0.93, respectively, and the AUC is 0.98. Comparing results with other models, it can be seen that the RF + SMOTE model has the best performance. To assess interpretability, we finally chose the RF + SMOTE model, even though the SVM + SMOTE model is comparable to the RF + SMOTE model. This decision was made based on two considerations. First, the RF + SMOTE model demonstrated superior classification performance in the minority class over the SVM + SMOTE model. Second, random forests offer advantages over linear models in many aspects, such as prediction and interpretability [41].

Sample analysis based on original data revealed that the differences in signs and symptoms between the two syndromes was not significant, with the main distinctions observed in the tongue and pulse (Figure 1). The most important classification features in the model were also reflected in the tongue and pulse (Figure 2), indicating that oversampling did not change the classification features of the data, suggesting that the model's features extracted from the classification rules can be used for clinical interpretability analysis. Furthermore, this indicated that the features after oversampling were stable and robust. Multiple rule extractions were performed on the RF model using MRSRE to obtain stable overlapping feature sets (Table 4 and 5), and it was found that most of the classified differential feature sets were still tongue and pulse. This highlights the significance of tongue and pulse features in the classification of Yang and Yang-Yin jaundice syndromes, which is consistent with clinical observations in chronic severe hepatitis B.

4.2 Clinical interpretability of the model

Chronic severe hepatitis B often has two syndromes, Yang jaundice and Yin jaundice, both of which have symptoms including yellowing of the skin and eyes and gastrointestinal discomfort. Chronic severe hepatitis often occurs on the basis of liver cirrhosis. Its pathogenesis is characterized by the necrosis of large areas of liver tissue centered on pseudolobules, which manifests as high jaundice and lasts for a long time. TCM believes that the pathogenesis of this disease has obvious particularities, which is mainly manifested in that there are many factors that cause jaundice to turn into Yin jaundice, and the proportion of non-Yang jaundice is high. The researchers analyzed the pathogenesis and characteristics of different types of jaundice. The key to distinguishing them lies in whether the jaundice color is bright or dark and whether the color of the tongue body is red or pale [42]. A study found another category of patients with spleen deficiency, characterized by pale, fat, or dentate tongue and loose stools [43]. Despite damp-heat or stagnant-heat manifestations, such as greasy or yellow tongue coating, or dry and bitter mouth and slightly red tongue, treatment is ineffective simply by identifying the dampness-heat and stagnant-heat patterns of Yang jaundice. Therefore, researchers have classified this syndrome as Yang-Yin jaundice, and medications that warm Yang and strengthen the spleen in HBV-ACLF Yang-Yin jaundice foster jaundice regression [42]. Based on the syndrome differentiation and treatment modes for Yang jaundice, Yang-Yin jaundice, and Yin jaundice, overall clinical efficiency, efficacy, and safety have improved [3]. It can be seen that the pathogenesis of Yang-Yin jaundice involves both spleen deficiency and dampness-heat. They have the manifestations of Yin jaundice syndrome such as jaundice and dark yellow, pale or fat tongue or tooth stains, and Yang jaundice syndrome such as dry or bitter mouth and light yellow tongue. Yang-Yin jaundice is an "intermediate state" between Yang jaundice and Yin jaundice. In studies of patients with severe chronic hepatitis B, Yang deficiency was an important factor in the transformation from Yang jaundice to Yin jaundice [42, 43]. In their research, it was concluded that there are notable differences in symptoms and signs between patients with Yang jaundice and Yang-Yin jaundice. Yang jaundice is mainly characterized by yellowing of the skin and eyes, gastrointestinal discomfort, bright jaundice, dry mouth and bitterness, sublingual pulse, greasy or yellow coating, and red tongue. On one hand, the cause of Yang jaundice is attributed to dampness-heat. On the other hand, the main characteristics of Yang-Yin jaundice are yellowing of the skin and eyes, gastrointestinal discomfort, pale, fat, dentate, greasy or yellow tongue coating, or dry and bitter mouth, tortuous sublingual veins, slightly red tongue, and loose stools. Yang-Yin jaundice is mostly caused by spleen deficiency and dampness-heat.

According to the results of Table 4 and 5, common features were observed in both syndromes, such as a wiry pulse, yellowing of the urine, skin and eyes, the shape of a healthy tongue, healthy sublingual vessels, nausea, and oil loathing. The main features of the two types of syndromes were obtained by calculating the difference in overlapping features. These distinctive features can be used to differentiate the two syndromes. The main features of Yang jaundice include a red tongue body and thickened sublingual vessels, which are indicative of dampness-heat. Additionally, the feature set of Yang-Yin jaundice includes a dark or pale-white tongue body, a white tongue coating, a lack of strength, a slippery pulse, a light-red tongue body, a slimy tongue coating, and abdominal distension. These features reflect the characteristics of spleen deficiency.

Therefore, the classification features extracted by MRSRE in this study are consistent with the findings of SUN et al. [42, 43], indicating that the classification rules extracted from the ensemble decision tree model can effectively explain the TCM syndrome differentiation patterns.

5 Conclusion

Consistent with models from previous TCM clinical studies, the interpretable TCM syndrome differentiation model in this study can identify 0.93 of major syndromes and 0.94 of minor syndromes, with the extracted features fully reflecting the similarities and differences between Yang-Yin jaundice and Yang jaundice. SMOTE oversampling of the minority samples did not alter their clinical features, and the constructed RF syndrome differentiation model is clinically interpretable and can be used for syndrome prediction in HBV-ACLF. This method can also be applied to other small samples of imbalanced TCM clinical data for syndrome differentiation analysis. This study provides technical support for further exploration of actual clinical syndrome differentiation in TCM and serves as a basis for constructing clinically interpretable syndrome differentiation models.

Fundings

Key research project of Hunan Provincial Administration of Traditional Chinese Medicine (A2023048), and Key

Research Foundation of Education Bureau of Hunan Province, China (23A0273).

Competing interests

The authors declare no conflict of interest.

References

[1] Liver Failure and Artificial Liver Group, Chinese Society of Infectious Diseases, Chinese Medical Association, et al. Guideline for diagnosis and treatment of liver failure. Chinese Journal of Hepatology, 2019, 27(1): 18–26.

[2] Chinese Association of Integrative Medicine. Expert consensus on the diagnosis and treatment of acute-on-chronic liver failure with integrated traditional Chinese and Western medicine. Journal of Clinical Hepatology, 2021, 37: 9.

[3] CHEN B, SUN KW, PENG J, et al. Clinical observation on the treatment of chronic severe hepatitis based on the syndrome differentiation model of Yang jaundice-Yang-Yin jaundice-Yin jaundice. Chinese Journal of Traditional Medical Science, 2012, 19: 57–58.

[4] WANG H, ZHANG XP, GONG HW, et al. Evaluation on the effects of different machine learning algorithms on the postoperative hypoproteinemia risk prediction model for elderly orthopedic patients. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2020, 22: 7.

[5] XIANG XH, PENG YH, YANG W, et al. Interpretability of Chinese medicine four examinations information of major adverse cardiovascular events in resistant hypertension: based on random forest rule extraction method. Journal of Traditional Chinese Medicine, 2022, 63: 7.

[6] ZHAO SY, ZHANG XY, LI YL. Study on diagnostic model of syndrome of deficiency of both Yin and Yang in hypertension based on decision tree and neural network. Chinese Archives of Traditional Chinese Medicine, 2019, 37(5): 1120–1123.

[7] XU L, ZHAO Y, PENG JH, et al. Binary logistic regression analysis on common syndromes characteristics of chronic hepatitis B. China Journal of Traditional Chinese Medicine and Pharmacy, 2015, 30(5): 1780–1783.

[8] HOU YM, ZHANG CY, SU YL. Risk prediction of ischemic stroke based on support vector machine. Modern Preventive Medicine, 2019, 46(15): 2692–2695, 2700.

[9] HUANG J, GUO H, KUANG YP. Preliminary research on regularity of syndrome differentiation of allergic rhinitis based on decision tree algorithm. China Journal of Traditional Chinese Medicine and Pharmacy, 2016, 31(11): 4770–4773.

[10] SHU CJ, LIANG H, WANG Y. A model for diagnosing TCM cold and hot at patterns based on random forest algorithm. Journal of Beijing University of Traditional Chinese Medicine, 2021, 44: 538–543.

[11] XU WF, GU WJ, LIU GP, et al. Study on feature selection and syndrome classification of excess syndrome in chronic gastritis based on random forest algorithm and multi-label learning. Chinese Journal of Information on Traditional Chinese Medicine, 2016, 23(8): 18–23.

[12] GONG J, DU C, ZHONG XW, et al. Researches on the illness risk of essential hypertension complicated with coronary heart disease based on machine learning algorithm. Medical Journal of Chinese People's Liberation Army, 2020, 45: 735–741.

[13] LIN SY, QU YQ, LIU C, et al. Review on the development of artificial intelligence of traditional Chinese medicine and exploration on the trend of technology integration. China Journal of Traditional Chinese Medicine and Pharmacy, 2020, 35: 6.

[14] BATISTA GEAPA, PRATI RC, MONARD MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20–29.

[15] CHAWLA NV, BOWYER KW, HALL LO, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357.

[16] ESTABROOKS A, JO T, JAPKOWICZ N. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 2004, 20(1): 18–36.

[17] HAN H, WANG WY, MAO BH. Borderline-SMOTE: a new oversampling method in imbalanced data sets learning. HUANG DS, ZHANG XP, HUANG GB, eds. Advances in Intelligent Computing: ICIC 2005. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2005, 3644: 878-887. Available from: https://doi.org/10.1007/11538059_91.

[18] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection. International Conference on Machine Learning, 1997: 179–186.

[19] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution. QUAGLINI S, BARAHONA P, ANDREASSEN S, eds. Artificial Intelligence in Medicine: AIME 2001. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2001, 2101: 63-66. Available from: https://doi.org/10.1007/3-540-48229-6_9.

[20] HAN C, TAN YK, ZHU JH, et al. Online feature selection of class imbalance via PA algorithm. Journal of Computer Science and Technology, 2016, 31(4): 673–682.

[21] MALDONADO S, MONTECINOS C. Robust classification of imbalanced data using one-class and two-class SVM-based multiclassifiers. Intelligent Data Analysis, 2014, 18(1): 95–112.

[22] VIEGAS F, ROCHA L, GONÇALVES M, et al. A Genetic Programming approach for feature selection in highly dimensional skewed data. Neurocomputing, 2018, 273: 554–569.

[23] WU QY, YE YM, ZHANG HJ, et al. ForesTexter: an efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems, 2014, 67: 105–116.

[24] CHAWLA NV, LAZAREVIC A, HALL LO, et al. SMOTEBoost: improving prediction of the minority class in boosting. LAVRAČ N, GAMBERGER D, TODOROVSKI L, et al, eds. Knowledge Discovery in Databases: PKDD 2003. Berlin: Springer Berlin Heidelberg, 2003: 107–119.

[25] CHEN SL, SHEN SQ, LI DS. Ensemble learning method for imbalanced data based on sample weight updating. Computer Science, 2018, 45(7): 31–37.

[26] DHAR S, CHERKASSKY V. Development and evaluation of cost-sensitive universum-SVM. IEEE Transactions on Cybernetics, 2015, 45(4): 806–818.

[27] DUAN L, GUO H, WANG JJ. Research on identification method of equipment failure degree under unbalanced data set. Journal of Vibration and Shock, 2016, 35: 178–182.

[28] DUFRENOIS F. A one-class kernel fisher criterion for outlier detection. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(5): 982–994.

[29] GU XQ, CHUNG FL, ISHIBUCHI H, et al. Imbalanced TSK fuzzy classifier by cross-class Bayesian fuzzy clustering and imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(8): 2005–2020.

[30] MALDONADO S, WEBER R, FAMILI F. Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. Information Sciences, 2014, 286: 228–246.

[31] WANG S, YAO X. Diversity analysis on imbalanced data sets by using ensemble models. Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, 2009: 324–331.

[32] YIN S, ZHU XP, JING C. Fault detection based on a robust one class support vector machine. Neurocomputing, 2014, 145: 263–268.

[33] RIBEIRO MT, SINGH S, GUESTRIN C. “Why Should I Trust You?”: explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, 2016: 97–101.

[34] LUNDBERG S, LEE SI. A unified approach to interpreting model predictions. arXiv, 2017. doi: 10.48550/arXiv.1705.07874.

[35] WANG Y. Internal Medicine of Traditional Chinese Medicine. Beijing: Scientific Publishing & Technical Publishers, 2012.

[36] ZHU W. Traditional Chinese Medicine Diagnostics. Beijing: China Press of Traditional Chinese Medicine, 2007.

[37] State Administration for Market Regulation, National Standardization Administration. Clinic terminology of traditional Chinese medical diagnosis and treatment— part 2: syndromes/patterns. Available from: http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=C71A9DAD24CB1252F12439D1F045DA6A.

[38] China Association of Chinese Medicine. Guidelines for clinical diagnosis and treatment of acute-on-chronic liver failure in traditional Chinese medicine. Journal of Clinical Hepatology, 2019, 35: 494–503.

[39] ZOU A, ZHANG Q. Nine relationships among TCM symptoms. Journal of Beijing University of Traditional Chinese Medicine, 2013, 3: 224–226.

[40] DENG HT. Interpreting tree ensembles with inTrees. International Journal of Data Science and Analytics, 2019, 7(4): 277–287.

[41] MARCHESE ROBINSON RL, PALCZEWSKA A, PALCZEWSKI J, et al. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. Journal of Chemical Information and Modeling, 2017, 57(8): 1773–1792.

[42] SUN KW, CHEN B, HUANG YH. Clinical observation on chronic severe hepatitis B treated by principles of cooling-blood and detoxicating combined with clearing-heat and resolving-damp or combined with strengthening-Pi and warming-Yang. Chinese Journal of Integrated Traditional and Western Medicine, 2006, 26(11): 981–983.

[43] SUN KW, CHEN B, HUANG YH, et al. Clinical characteristics of jaundice in patients with chronic severe hepatitis B. Chinese Journal of Integrated Traditional and Western Medicine on Liver Diseases, 2010, 20: 8–11.

基于小样本不平衡数据构建乙肝相关慢加急性肝衰竭中医辨证分型的可解释性模型

周展[a], 彭清华[b*], 肖晓霞[a*], 邹北骥[a], 刘彬[a], 郭水霞[c]

a. 湖南中医药大学信息科学与工程学院, 湖南 长沙 410208, 中国
b. 湖南中医药大学中医学院, 湖南 长沙 410208, 中国
c. 湖南师范大学数学与统计学院, 湖南 长沙 410081, 中国

【摘要】目的 乙肝相关慢加急性肝衰竭（HBV-ACLF）临床病历数据普遍存在样本量小、类别不平衡等问题，而大部分机器学习模型是基于平衡数据设计的，缺乏可解释性。本研究旨在基于中医辨证论治理论，提出一种临床可解释、准确率高的 HBV-ACLF 中医诊断模型。**方法** 本研究收集了 261 例 HBV-ACLF 患者的病例，包括阳黄证（214 例）、阳阴黄证（41 例）和阴黄证（6 例）三种证型。为了避免机器学习模型过拟合，排除了阴黄病例。经过数据标准化和清洗，获得阳黄证和阳阴黄证相关的 255 份病历。针对类别不平衡问题，采用过采样方法和五种机器学习方法，包括逻辑回归（LR）、支持向量机（SVM）、决策树（DT）、随机森林（RF）和极端梯度提升（XGBoost），构建了证型诊断模型。本研究以精度、F1 得分、受试者工作特征曲线下面积（AUC）和准确率作为模型评价指标。选择分类结果最好的模型提取诊断规则，并深入分析其临床意义。此外，我们提出了一种新颖的多轮稳定规则提取（MRSRE）方法，以获得可以展示模型临床可解释性的稳定特征规则集。**结果** 利用过采样平衡数据构建的五种机器学习模型精度都超过了 0.90，其中 RF 证型分类准确率为 0.92，阳黄及阳阴黄两类别的 F1 均值分别为 0.93 和 0.94，AUC 值为 0.98。基于 MRSRE 方法的 RF 辨证模型提取规则显示，阳黄及阳阴黄的共同特征是脉弦，身目尿黄，舌体正常，舌下脉络正常，恶心和厌油纳差。阳黄的主要特点是舌质红、舌下脉络增粗，阳阴黄的主要特点是舌质暗、淡白、苔白、无力、脉滑、舌质淡红、舌苔腻和腹胀，该结果与中医专家依据中医辨证论治理论相一致。**结论** 本研究构建的模型可用于区分 HBV-ACLF 证型，还可用于生成其他临床可解释的模型，这些模型对样本量小且类别不平衡的临床数据具有较高的准确性。

【关键词】中医；乙肝相关慢加急性肝衰竭；不平衡数据；随机森林；可解释性