Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/20240642

A Random Forest-based Prediction for Liver Cirrhosis

Haitao Wei

No.999, Xi'an Road, Pidu District, Chengdu, Sichuan, China, Southwest Jiaotong University

hwei3@student.gsu.edu

Abstract. Prediction of cirrhosis through timely screening and diagnosis with effective medical and lifestyle interventions can be taken to delay the disease progression and improve treatment effectiveness. Therefore, early cirrhosis prediction helps reduce associated complications and improve quality of life. This study established a random forest-based algorithm and compared its performance with other machine learning models based on an Indian Liver Patient Records. The dataset was divided 30% of the test set and 70% of the training set. The results showed that the Accuracy, Area under the curve (AUC) and Recall values of Random Forest (RF) model were 0.73, 0.76 and 0.88, respectively, which demonstrated the proposed model can predict the results better than other models. The proposed model for predicting liver cirrhosis in the future will continue to be improved and perfected to improve the accuracy of liver cirrhosis prediction and treatment effect, thereby reducing the health risks and burdens brought by the disease.

Keywords: Liver Cirrhosis, Random Forest, Indian Liver Patient Records, Machine learning.

1. Introduction

Cirrhosis is a serious public health problem and is often diagnosed at a later stage, limiting opportunities for early intervention and treatment [1]. Traditional medical techniques play a key role in the diagnosis of cirrhosis [2], and if combined with early prediction, can better assist doctors in medical diagnosis and personalized treatment [3]. Therefore, it is an important and meaningful work to study how to predict the occurrence of liver diseases. By predicting the risk of cirrhosis in advance, early intervention and treatment can be achieved, helping to slow the progression of the disease, improve cure rates, and reduce healthcare costs [3]. Effective cirrhosis prediction can also help optimize the allocation of medical resources, ensuring that more timely and effective medical services are delivered where they are most needed. As an aid to medical decision-making, it can also help doctors better understand patient risks and disease trends. At present, the traditional method of liver puncture biopsy is used in clinical diagnosis, but it has greater trauma to patients [4] and high cost. So, if there was an easier way to predict the disease it would improve the efficiency of treatment.

Machine learning has been widely applied in disease prediction and treatment, and several studies has used machine learning methods in cirrhosis prediction. This study compares the performance of several popular machine learning models. In data processing, the label encoding method was used to convert characters into numbers. On this basis, the random forest model was proved to be the best model. Then, a feature screening experiment based on random forest was carried out to study the performance before and after feature screening. Without feature screening, random forest was proved to be the best

DOI: 10.54254/2755-2721/67/20240642

and more rigorous compared to other studies. Model training and model comparison were not discussed in the previous study as well, which are very important in the establishment of models for disease prediction and the accuracy of prediction results, and the conclusion will be more rigorous after discussion.

The feasibility of this study is based on extensive data on cirrhosis, advanced machine learning and a wide range of needs. In view of this, this study established a RF-based model to predict whether to suffer from cirrhosis. The purpose of this study was to establish a random forest model for the prediction of cirrhosis.

2. Literature Review

Automated predictive systems are introduced to allocate treatment tasks more efficiently and interventions to lower readmission rates may be made possible by early identification of high-risk patients [4]. However, in most cases, integrated classifiers are more effective, Liu et al. used techniques such as neural networks, logistic regression, and supported vector machines to predict diseases, and have improved on traditional single classifiers to improve overall diagnosis rates [5]. Random Forest (RF) is an ensemble learning method that generally provides high prediction accuracy [6]. It builds multiple decision trees, in which the RF randomly selects some features instead of considering all of them, which helps increase the diversity of each tree and effectively resizes overfitting problems [7]. Kanwal et al. evaluated three methods: gradient descent lifting, logistic regression with minimum absolute contraction and selection operator (LASSO) regularization, and logistic regression with LASSO constraints in liver cirrhosis prediction. The inclusion of predictors and model performance were cross-validated by a factor of 5. Finally, the predictors identified in the most reasonable model were fitted using the maximum likelihood estimation method, and their performance was compared with that of the liver disease sodium model [8]. Chang et al. used machine learning (ML) models including logistic regression, RF, and artificial neural networks, and used ML, FibroScan liver hardness measurements, and fibroscan-4 index to predict the histological stage of fibrosis. Finally, ML and Fibroscan-AST were used to evaluate the model performance [9].

3. Method

3.1. Dataset

The dataset used in this study was taken from the Kaggle Website, called Indian Liver Patient Records. It had 416 records of patients with cirrhosis and 167 records of patients without cirrhosis, with a total data volume of 583 (441 men, 142 women), comprising a total of 11 characteristic dimensions. For patients over 89 years of age, they were recorded as "90." The Dataset column is a category label used to classify patients into patients with and without cirrhosis. The dataset was collected from test samples in the northeastern state of Andhra Pradesh, India.

In the dataset, there are 11 features about patients. They are Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamina Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, Albumin and Globulin Ratio and Dataset (patient with liver disease, or no disease).

3.2. Preprocessing of dataset

In the Albumin_and_Globulin_Ratio column of the dataset, there are a total of 8 missing values. If the missing values are not processed, a series of problems will occur during the experiment, so the common mode filling method is used in this study. Thirty percent of the test sets and 70% training sets were set in this dataset to implement the machine learning model.

In addition, the features in the Gender column are character classes. Since the machine learning algorithm used in this study requires the input features to be numeric, the Label Encoder method was used to digitize the character classes.

The relationship between features and features, and between features and predicted results in the dataset of the study. As can be seen from the heatmap in Figure 1, the correlation between features is

DOI: 10.54254/2755-2721/67/20240642

not high. But between the features, Total_Protiens and Albumin, Alamine_Aminotransferase and Aspartate_Aminotransferase, Direct_Bilirubin and Total_Bilirubin, there was a certain correlation between Albumin and Albumin_and_Globulin_Ratio. In addition, observing the difference between the two types of distributions in a data set is crucial for the training and application of machine learning models. This difference can help us understand the degree of differentiation between different categories, and then guide feature selection and model tuning, to improve classification performance. Figure 2 showed the line chart of the density of different features and whether the disease is present and showed the difference between the two types of distributions, from which each feature has a wireless relationship with whether the disease is present. When a certain peak value is reached, the density will begin to decline.

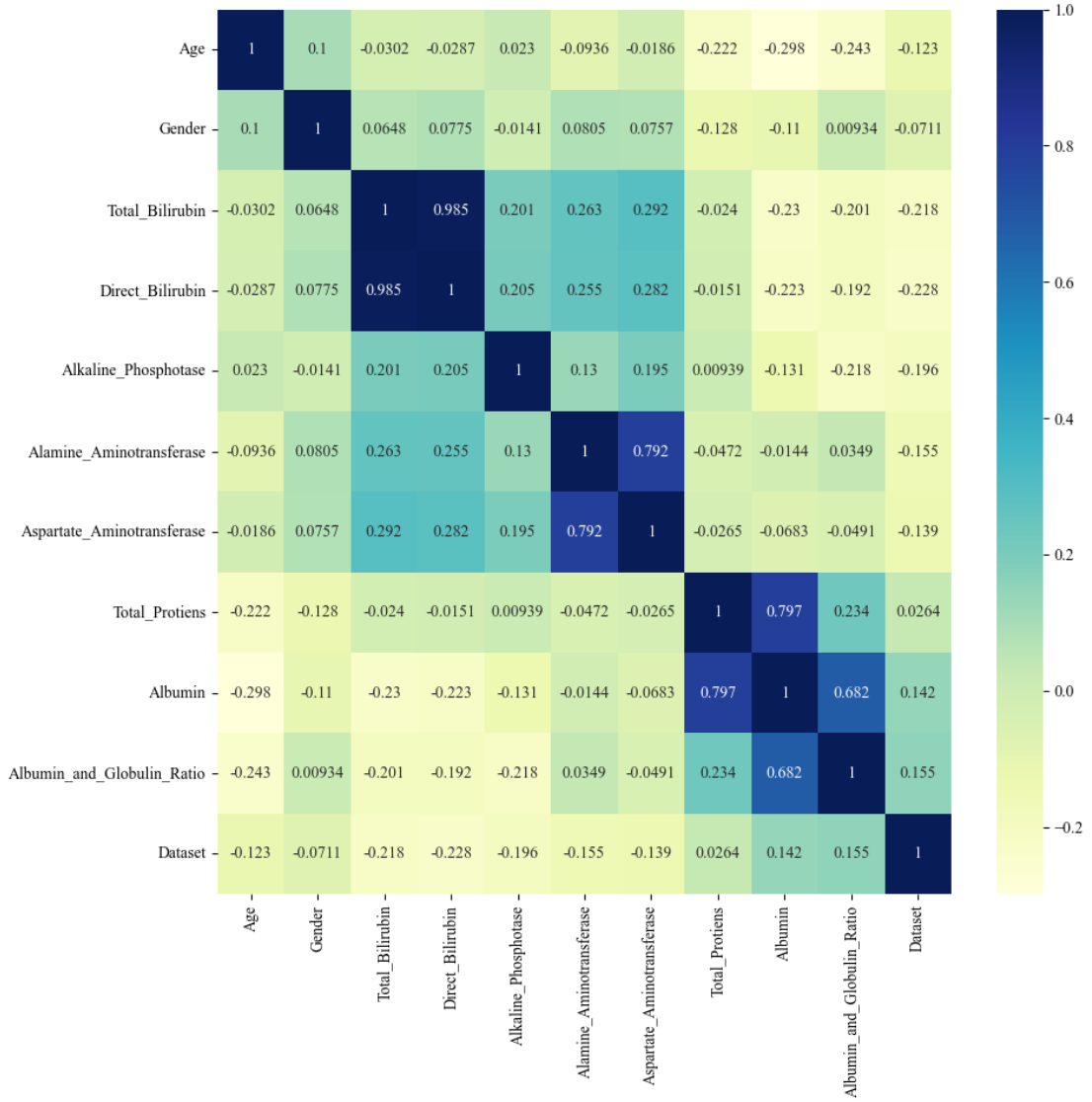

Figure 1. Correlation matrix of each feature.

DOI: 10.54254/2755-2721/67/20240642

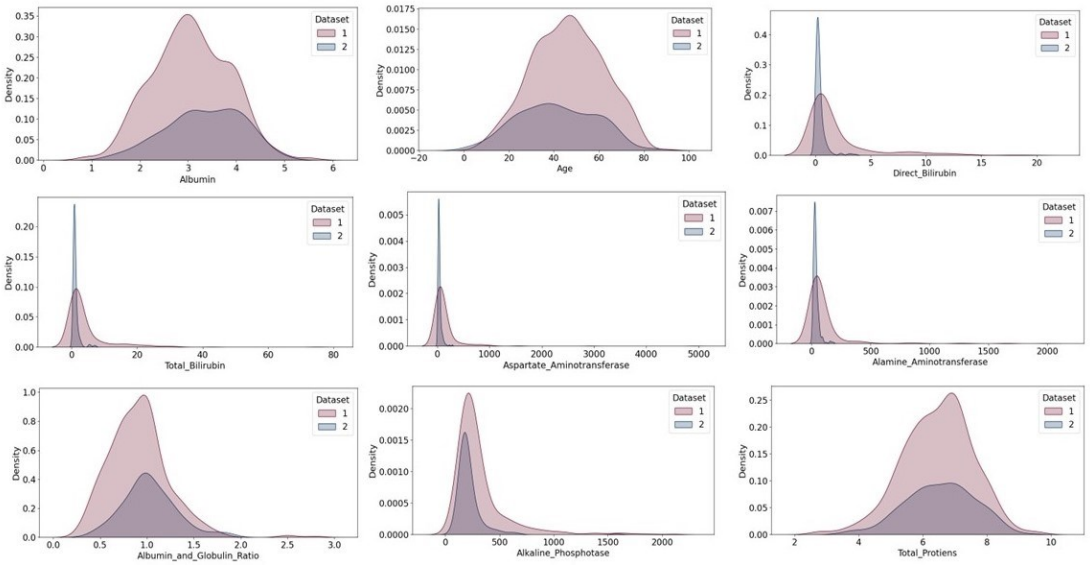

Figure 2. The density relationship between characteristics and the presence or absence of disease.

Figure 3 showed the importance degree of each feature in the RF model, which was obtained by feature screening based on the RF model. From here, it showed that Age, Alkaline_Phosphotase, Alamine_Aminotransferase and Aspartate_Aminotransferase were relatively important in the RF model. The feature screening here would be used in feature screening experiment.

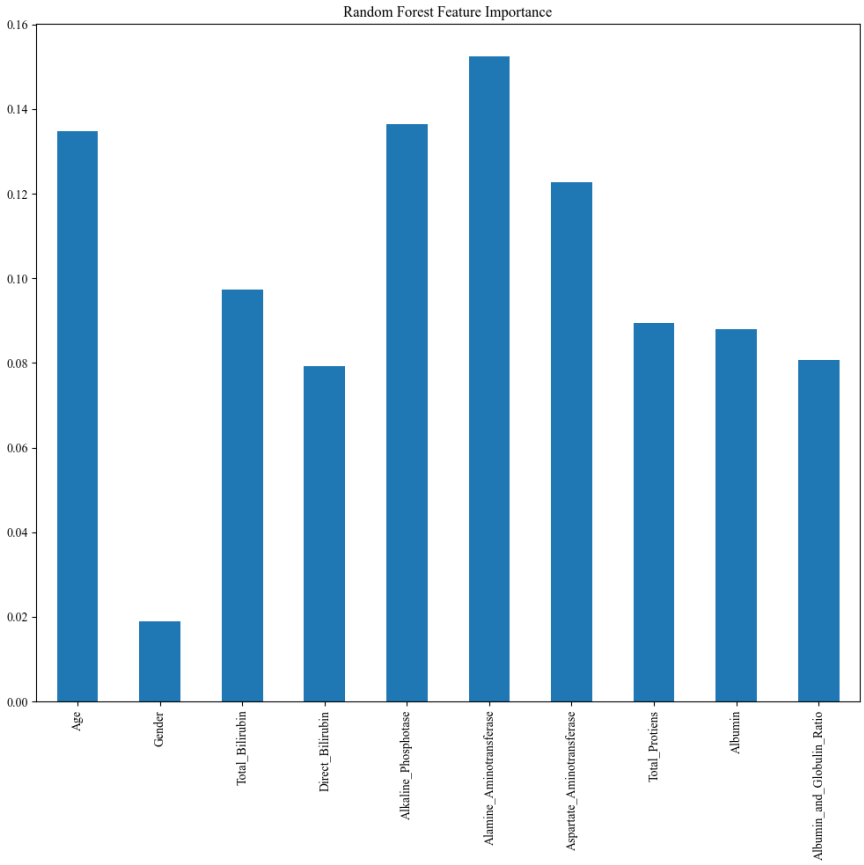

Figure 3. The importance of features in Random Forest model.

DOI: 10.54254/2755-2721/67/20240642

3.3. RF model

RF is a powerful machine learning algorithm that builds a forest in a random way by integrating multiple decision trees, most of which have high accuracy and relatively weak correlation, and then combines to form a predictive model. Each tree randomly selects variables and observations, then creates a classifier and votes for the result. In general, logN features are chosen for each tree, where N is the number of features. If every tree selects all features, then a RF can be considered a Bagging algorithm. Bagging is an ensemble learning method that generates multiple sub-datasets by sampling the training dataset, and then trains a basic learner based on each sub-dataset. Finally, the prediction results of the integrated model can be obtained by regression or classification of the prediction results of all the basic learners. The following diagram shows the algorithm flow of RF (Figure 4).

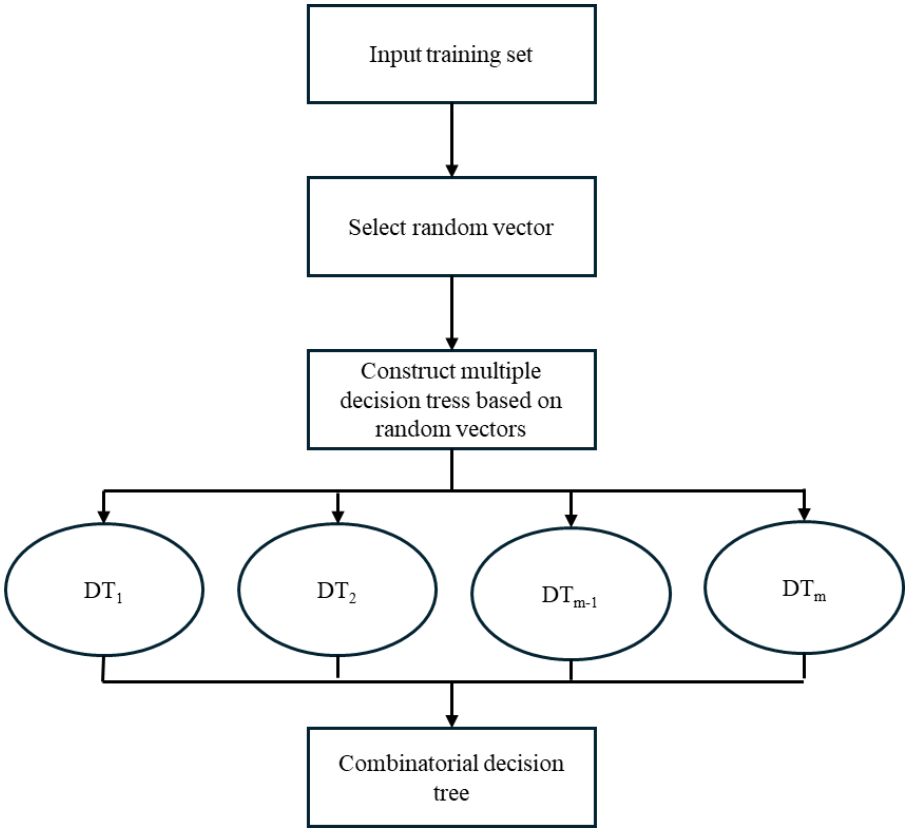

Figure 4. Random Forest model algorithm.

In this study, the following machine learning models besides RF were also compared, they are Logistic Regression (LR) [10], Decision Tree (DT) [11], Gradient Boosting Decision Trees (GBDT) [12], AdaBoost [13], XGBoost [14], Deep Neural network (DNN) [15]. And the RF, GBDT, Adaboost and XGBoost are ensemble models. The performance of RT before and after screening was compared.

3.4. Training process

In the whole model training process, the dataset was initially processed, and then the dataset was divided into 30% test set and 70% training set, and then the code was used to establish LR, DT, RF, GBDT, AdaBoost and XGBoost models. For random forest, the final prediction is obtained by constructing multiple decision trees and integrating their prediction results. Logistic regression can minimize the loss function to find the best model parameters to fit the training data. The decision tree divides the sample by recursively selecting the features that can minimize the loss function and constructs the subtree until the stop condition is met. By adjusting the weights of misclassified samples, AdaBoost iteratively trains several weak classifiers and combines them into strong classifiers. XGBoost is based on the integrated

DOI: 10.54254/2755-2721/67/20240642

method of decision tree and improves the accuracy and generalization ability of the model through gradient lifting technology. In addition, GBDT builds multiple decision trees through iteration, each tree attempts to correct the prediction error of the previous tree and forms a strong learner through multiple iterations to improve the prediction accuracy of the model. These model training procedures are all aimed at finding the model parameters and structures that best fit the training data and predict the unknown data. To predict the performance better, the parameters of DT, RF, GBDT, AdaBoost and XGBoost models were trained by Bayesian optimization. Table 1 shows the tuning parameters which will set in different models in this study during the training process and uses the tuned model for subsequent integration training and validation.

Table 1. Parameters of Different Models.

Model	Parameters
DT	Criterion: 'gini' ; Spliter: 'best'; Max_depth: 'none'; Min_samples_split: '2'
RF	N_estimators: '100'; Criterion: 'gini'; Max_depth: 'none'; Min_sanples_split: '2'; Min_samples_leaf: '1'
GBDT	N_estimators: '100'; Learning_rate: '0.1'; Max_depth: '3'; Min_sanples_split: '2'; Min_samples_leaf: '1'
AdaBoost	N_estimators: '50'; Max_depth: '1'; Learning_rate: '1'; Random_state: '50'
XGBoost	Learning_rate: '0.3'; Max_depth: '6'; Subsample: '1'; Colsample_bytree: '1'; Gamma: '0'; Min_child_weight: '1'; Lambda and alpha: '1 and 0'; Num_boost_round: '100'

In addition, when building the DNN model, the data set was also set to 30% test set and 70% training set, using the Adam (Adaptive Moment Estimation) optimizer to minimize the loss function, and to optimize network parameters by back propagation. Then, for different parameters of the DNN model, the epochs were set to 50 and the batch size was set to 80. To ascertain the network convergence, we computed the loss at the conclusion of each training epoch. Finally, the performance indexes of these models are obtained and compared. The Figure 5 below shows the visualization of the DNN training. The horizontal axis is the number of iterations, and the vertical axis is Accuracy (right picture) and Loss (left picture). In the figure, it can be concluded that the effect of the neural network model will become better as the number of iterations increases. However, when the number of iterations reaches a certain level, the effect begins to deteriorate, and the image begins to become non-convergent.

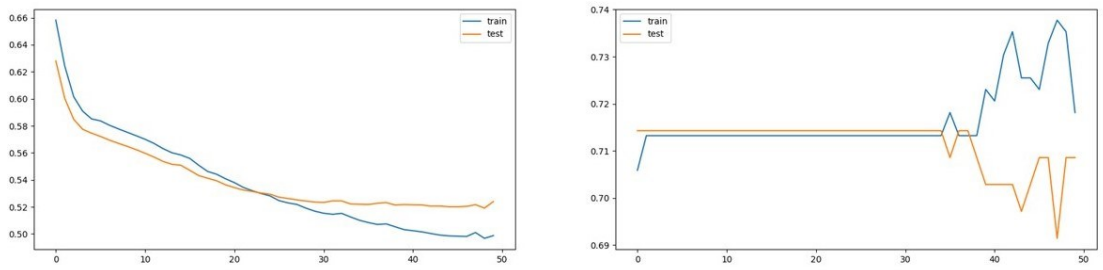

Figure 5. DNN training visualization.

When the feature screening was carried out, the correlation analysis based on Pearson coefficient was first carried out, and the features with a correlation greater than 0.1 were screened out. It was found that the model performance was reduced, so the importance feature screening based on the model itself was used.

3.5. Evaluation metrics

Each of them has different advantages, but in this study, the actual research problem was about binary classification, that was, the characteristics of the dataset to determine whether there was cirrhosis. So, for the diagnosis of diseases, Recall is especially important, because missing judgment will cause serious

DOI: 10.54254/2755-2721/67/20240642

consequences for patients. Therefore, Recall is an important indicator in model comparison. In addition, Accuracy and Area under the curve (AUC) are both important indicators. Among them, Accuracy is expressed by the ratio of the number of samples correctly classified to the number of samples for diagnosis, which represents the accuracy of the model. AUC is usually used as an indicator to measure the generalization ability of the model.

4. Results

4.1. Comparison results

Table 2 showed the performance results of different models. According to the important evaluation indicators, the ensemble learning model can achieve better performance. In these integrated learning models, the performance of RF model is better than other models according to various important indicators. In Table 1, the Accuracy value of the RF model is 0.726, the Recall_0 value is 0.88, and the AUC value is 0.759. Therefore, compared with other models, the RF model had better performance and is more stable.

Table 2. Comparison of models based on Indian Liver Patient Records[a].

	LR	DT	GBDT	AdaBoost	XGBoost	DNN	RF
Accuracy	0.72	0.71	0.69	0.69	0.70	0.70	0.73
Precision_1	0.67	0.50	0.43	0.43	0.47	0.25	0.53
Recall_1	0.04	0.50	0.32	0.32	0.34	0.02	0.34
F1_1	0.08	0.50	0.37	0.37	0.40	0.04	0.42
Precision_0	0.72	0.80	0.75	0.75	0.76	0.71	0.77
Recall_0	0.99	0.80	0.83	0.83	0.85	0.98	0.88
F1_0	0.84	0.80	0.79	0.79	0.80	0.82	0.82
AUC	0.69	0.65	0.75	0.75	0.76	0.71	0.76

[a] 0 for diseased and 1 for not diseased.

4.2. Comparative results of feature screening

Table 3 showed the performance comparison of the RF model before and after feature screening. It can be found that the performance of the model after feature screening was worse than that before feature screening.

Table 3. Comparison results before and after feature screening in Random Forest model[a].

Model	Random Forest (not screened)		Random Forest (Screening)	
Indicator	Precision	Recall	Precision	Recall
1	0.77	0.92	0.75	0.82
2	0.66	0.37	0.45	0.35
Accuracy	0.64		0.58	

[a] 1 means diseased and 2 means not diseased.

5. Discussion

In this study, the data were first processed, and the performance of different models was compared, and then conducted the feature screening. In the data analysis, the density curves of Gender and Dataset were not shown in Figure 3 because of the relationship between the two characteristics and whether the dataset was diseased.

The performance of ensemble learning model was generally higher than other models, and the reason was that in this dataset, the correlation between features and disease was not high, and each feature was not linear with disease, so the prediction effect of the linear model might be worse than that of the integrated model. When training DNN, the performance of the model deteriorates after a certain number

DOI: 10.54254/2755-2721/67/20240642

of iterations. Reasons for DNN overfitting may include the excessive complexity of the model, causing it to overlearn specific details on the training set and fail to capture the general laws of the data. At the same time, if the training data set is limited or underrepresented, DNNS may not be able to learn the true distribution of the data, resulting in memory of noise and details. In addition, overtraining or poor choice of model architecture can also exacerbate the overfitting phenomenon. The RT was the best model to predict the liver cirrhosis based on this study. The advantages of RF include being able to handle large datasets with high accuracy and robustness, being able to handle high-dimensional data and not easily overfit, being robust to missing data, being suitable for various types of data, and being easy to implement and tune. In this study, RF can effectively capture potential nonlinear relationships and interactions, which improves the prediction performance. During the training process, random forest uses randomly selected features and samples for modeling, reducing the risk of overfitting and being able to process large amounts of feature data. In addition, Random Forest provides an assessment of the importance of each feature (feature screening) to help identify the factors most influential in predicting cirrhosis.

After the RF model was selected as the optimal model, the feature screening experiment was carried out, and then compared the performance between the previous RF model and the one after feature screening. The results showed that the performance of the screened model was inferior to that of the pre-screened model. The reason might be that after feature screening, there was a lack of information, and clinical patients might also show other characteristics, unable to make full use of all the characteristics, and the patient's condition should be combined with a variety of clinical characteristics to judge.

Compared with the existing literature, this study focuses on various aspects of comparison, comparing the performance of different common machine learning models. In the aspect of data processing, label encoding method is used to convert characters into numbers. The random forest model is confirmed as the best model based on this study. After that, a feature screening experiment based on random forest was conducted to further explore the performance before and after feature screening. It is confirmed that random forest has the best performance without feature screening. Therefore, the random forest model was established to predict cirrhosis in this study. This research can be applied to the prediction of clinical patients' condition and can help doctors to accurately grasp various data indicators of patients. Therefore, prediction based on random forest model can play a key role in determining whether patients have cirrhosis in the future. Besides, the novelty of this study lies in the prediction of cirrhosis by random forest algorithm. This method not only improves the stability and accuracy of prediction, especially in the processing of complex medical data, but also can handle unbalanced data and has good interpretability. These properties make the random forest model uniquely useful in the field of cirrhosis prediction, providing clinicians with new tools to understand and predict cirrhosis risk to optimize treatment and improve patient management.

There were some limitations in the present study. At first, the sample size of the dataset was small, which can increase the risk of overfitting and then lead to poor generalization of unknown data. In addition, small datasets can introduce bias and inaccuracies into predictions. Therefore, future studies can combine image sets or improved RF model to further improve the performance and reliability of cirrhosis prediction models. Besides, in the future, more models can be tried to predict cirrhosis and perform performance comparison and analysis. In this experiment, the parameters set by different models are not necessarily the best scheme, so the parameters of different models can also be set to improve the performance of the models.

6. Conclusion

In this study, by comparing the performance of different machine learning models, we determined the optimal model RF for liver disease prediction based on cirrhosis patient datasets. Compared with other machine learning models, it can more comprehensively and accurately judge whether patients have cirrhosis. Compared with the traditional diagnostic method of liver biopsy, the RF model was used in this study to predict, which can effectively reduce medical costs. And it may help optimize the allocation of medical resources, and doctors can better understand patient risk and disease development trends

DOI: 10.54254/2755-2721/67/20240642

through data. The development of this model provides strong support for the early diagnosis and treatment of cirrhosis, which is of great significance for most people to predict cirrhosis.

References

[1] Iredale J P 2003 Cirrhosis: new research provides a basis for rational and targeted treatments *BMJ* 327 143

[2] Smith A, Baumgartner K and Bositis C 2019 Cirrhosis: Diagnosis and Management. Am. Fam. Physician *Am Fam Physician* 100 759–770

[3] Yeom SK, Lee CH, Cha SH and Park CM 2015 Prediction of liver cirrhosis, using diagnostic imaging tools *World J Hepatol* 7 2069-79

[4] Liu J, Zhang H, Liu Y, Zhang H and Liu Y 2019 Cirrhosis diagnosis prediction research based on Random Forest *J comput sci app* 9 1928-38

[5] Singal AG, Rahimi RS, Clark C, et al 2013 An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission *Clin Gastroenterol Hepatol* 11 1335-41

[6] Steven J Rigatti 2017 Random Forest. *J Insur Med* 47 31–9

[7] Dou Z 2019 Study on correlation between symptoms, syndrome elements and instantaneous elastic imaging values in patients with hepatitis cirrhosis *(Beijing, Beijing University of Chinese Medicine)*

[8] Kanwal F, Taylor TJ, Kramer JR, Cao Y, Smith D, Gifford AL, El-Serag HB, Naik AD and Asch SM 2020 Development, Validation, and Evaluation of a Simple Machine Learning Model to Predict Cirrhosis Mortality *JAMA Netw Open* 2 e2023780

[9] Chang D et al 2023 Machine learning models are superior to noninvasive tests in identifying clinically significant stages of NAFLD and NAFLD-related cirrhosis. *Hepatology* 77 546-57

[10] Stoltzfus JC 2011 Logistic regression: a brief primer Acad Emerg Med 18 1099-1104

[11] Banihashem S Y and Shishehchi S 2023 Ontology-Based decision tree model for prediction of fatty liver diseases Comput Methods Biomech Bio*med Engin* 26 639-49

[12] Zhang Z and Jung C 2021 GBDT-MO: Gradient-Boosted Decision Trees for Multiple Outputs *IEEE Trans Neural Netw Learn Syst* 32 3156-67

[13] Gubbala K, Kumar MN and Sowjanya AM 2023 AdaBoost based Random forest model for Emotion classification of Facial images *MethodsX* 14 102422

[14] Li F, Xin H, Zhang J, Fu M, Zhou J and Lian Z 2021 Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database *BMJ Open* 11 e044779

[15] Kriegeskorte N and Golan T 2019 Neural network models and deep learning *Curr Biol* 29 R231-6