An analysis of the Correlation Between Clinical Characteristics and Staging in Patients with Liver Cirrhosis

Xinbo Zhao

College of Intelligent Science and Engineering, Yunnan Techonlogy and Business University, KunMing, 650000, China
19619500028@ytbu.edu.cn

Abstract. Cirrhosis is the terminal manifestation of chronic liver disease, and serious complications often accompany patients. The correlation analysis of clinical characteristics and stages is very important for disease management. Clarifying the correlation between clinical characteristics and stages of liver cirrhosis patients will help develop personalized treatment plans and improve patients' quality of life. This study aims to explore the clinical characteristics of patients with cirrhosis and the relationship between the clinical characteristics and disease stages. Through the analysis of relevant samples, the stage of cirrhosis can be confirmed and predicted. This study uses three prediction methods, decision tree, random forest, and logistic regression, to analyze the relationship between liver cirrhosis stage and related blood parameters. These blood test indicators include age, bilirubin, cholesterol, albumin, copper, alkaline phosphatase, SGOT, triglyceride, platelet, and prothrombin time. The study finds that there are significant differences between these indicators in different stages of cirrhosis, which provides clinicians with useful information on the progress and treatment of cirrhosis.

Keywords: Liver cirrhosis, clinical features, blood test indexes, disease staging, correlation analysis.

1 Introduction

Liver cirrhosis is a severe liver disease, often caused by liver tissue damage and fibrosis due to long-term hepatitis or alcoholism. The aggravation of liver cirrhosis leads to portal hypertension, ascites, hepatic encephalopathy, and other serious complications, increasing the risk of liver cancer and significantly shortening the life expectancy of patients [1,2]. Cirrhosis is a serious consequence of chronic liver disease, often leading to portal hypertension, ascites, hepatic encephalopathy, and other serious complications, even increasing the risk of liver cancer, significantly shortening the life expectancy of patients. Currently, the diagnosis and staging of liver cirrhosis mainly rely on liver biopsy [3]. However, liver biopsy is an invasive operation with certain risks, and it cannot monitor the changes of the condition in real

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_68

time. Therefore, finding a non-invasive and reproducible method to predict the clinical characteristics and stages of liver cirrhosis patients is of great clinical significance [4,5].

Machine learning algorithms, as an artificial intelligence technology, can realize the prediction of clinical characteristics and staging of cirrhosis patients by analyzing and learning from a large amount of data and mining the potential patterns in the data [6-8]. This method not only avoids the risk of liver biopsy but also monitors the changes of the condition in real time, which provides robust support for clinical decision-making.

This study aims to predict the clinical features and staging of cirrhosis patients using machine learning algorithms. First, a large amount of clinical data of cirrhosis patients was collected, including laboratory test results, imaging features, medical history, etc. Then, these data were pre-processed to ensure accuracy and reliability. Next, machine learning algorithms were used to train and test the data. In this paper, the information in the study sample of 25,000 cirrhosis patients was analyzed and organized, and three prediction methods, namely, decision tree, random forest, and logistic regression, were used to derive the correlation between the relevant data in the relevant samples and cirrhosis prevalence. The study in this paper can be used as supplementary information for the analysis of liver cirrhosis.

2 Data and methods

2.1 Data sources

The dataset used in this paper is from the data set named Liver Cirrhosis Stage Classification on the Kaggle platform. This dataset analyzes the correlation between different stages of liver cirrhosis and age, bilirubin, cholesterol, albumin, copper, alkaline phosphatase (AlK_Phos), and other blood test indicators.

The dataset contained clinical information and blood test results of 25,000 patients with cirrhosis. The studyDescriptive statistics were first used to analyze the data to understand the distribution of the variables, followed by analysis of variance (ANOVA).

2.2 Methodology

This paper uses three data prediction methods: logic trees, logistic forests, and logistic regression. A logic tree, also known as a decision tree, is a graphical representation of the logical structure of a decision and its possible outcomes. Logic trees decompose a problem into a series of smaller, more manageable decision units, helping to identify the steps needed to solve a problem and analyzing each step's possible outcomes [9,10].

Decision nodes are bifurcation points in the tree where choices or decisions must be made. Branches are the possible directions for decision-making, with each branch from the decision node corresponding to a possible choice. Chance nodes represent

uncertain elements, usually depicted by probabilities for possible outcomes. End nodes, also known as leaf nodes, represent the final result of the problem.

Logic trees are widely used in decision analysis, risk assessment, project management, and problem-solving. All possible decision paths and their results can be visualized through logic trees, making it easier to make more reasonable and information-supported decisions. In addition, logic trees can help assess the risks and benefits of different decisions, thus providing a basis for choosing the best solution.

Logic Forest is a machine learning algorithm, an integrated learning method based on decision trees. Logistic forest is a variant of the Random Forest algorithm, which is a form of decision tree set. The main idea of logistic forest is to improve the accuracy and robustness of prediction by building multiple decision trees and combining their prediction results.

Key features of logistic forests include: Logic forest constructs each tree by randomly selecting features and sample subsets, which increases the diversity among trees and improves the overall model's generalization ability. Logic forests make final decisions by combining the predictions of multiple trees. For classification problems, this is usually done by majority voting, whereas for regression problems, it is done by calculating the average of the predictions of all trees. Since Logic Forest introduces randomness in the training process, it reduces the risk of overfitting and performs well even on datasets with many features. Additionally, logistic forests can provide an assessment of the importance of features, which is useful for understanding models and feature selection. Logistic forests have a wide range of application areas, including but not limited to classification, regression, anomaly detection, and recommender systems. Due to its powerful performance and flexibility, Logic Forest is a common tool used by data scientists and machine learning engineers.

Logistic regression is a statistical method widely used in classification problems, especially binary ones. Although its name includes "regression", it is a classification model used to estimate the probability that a sample belongs to a certain category. The core of logical regression uses the sigmoid function to simulate the probability that a sample belongs to a positive class (usually marked as 1). The sigmoid function is defined as follows:

$$S(x) = 11 + e - xS(x) = 1 + e - x1 \tag{1}$$

The output value of this function is always between 0 and 1, which is ideal for describing probability.

The basic form of a logistic regression model is as follows:

$$P(y = 1) = 11 + e - (\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn)P(y = 1) = 1 + e - (\beta 0 \quad + \beta 1 \quad X1 \quad + \beta 2 \quad X2 \quad + \cdots + \beta n \quad Xn \quad)1 \tag{2}$$

Where $P(y=1)$ represents the probability that the sample belongs to the normal class, $n\,X1, X2, ..., Xn$ are the characteristic variable, and $\beta 0, \beta 1, \ldots, \beta n$ are the model parameters.

The training process of a logistic regression model usually involves maximizing the log-likelihood function, which can be achieved by various optimization algorithms (e.g., gradient descent, Newton's method, etc.). Once training is complete, the model can be used to predict the probability that a new sample belongs to a positive class and make classification decisions based on a probability threshold.

Logistic regression models are simple in form, making them easy to explain and understand. They offer strong model interpretability, as model parameters can be directly interpreted as the effect of features on the predicted probability. Although logistic regression was originally designed for binary classification, it can be easily extended to multi-classification problems through one-to-many (One vs All) or multinomial logistic regression. Additionally, logistic regression provides an estimate of the probability that a sample belongs to each category, which is useful in applications that require probabilistic information. Logistic regression has applications in many fields, including medicine (disease diagnosis), finance (credit scoring), and marketing (customer churn prediction). Due to its simplicity and effectiveness, logistic regression is a popular data analysis tool.

3 Analysis of Results

3.1 Analysis of relevant results

Table 1 Detailed data analysis of analysis factors

	Mean	Standard Deviation	Description
Age	18.49588	37.3760	a wide age range
Bilirubin	3.40	4.71	Reflects the degree of impaired liver function
Cholesterol	372.33	193.67	Associated with cardiovascular health and liver metabolism
Albumin	3.49	0.38	Blood protein levels related to nutritional status and liver function
Copper	100.18	73.18	May be associated with the diagnosis of certain liver diseases
Alkaline phosphatase (Alk_Phos)	1995.68	1798.89	Reflects the condition of the hepatobiliary system
SGOT	123.17	47.75	Indicator of hepatocellular damage
Triglycerides	123.82	52.79	Associated with lipid metabolism

Platelets	256.01	98.68	Reflects the condition of the coagulation system
Prothrombin	10.73	0.90	Related to coagulation
Stage of liver cirrhosis	2.00	0.81	Indicates the severity of the disease

The results of the ANOVA test in Table 1 show that all the considered blood test indicators have significant differences between different stages of cirrhosis (the p value of all indicators is 0.0, far below the significance level of 0.05). These results indicate that there is a strong correlation between the different stages of cirrhosis and the patient's age, bilirubin, cholesterol, blood protein, copper, alkaline phosphatase, SGOT, triglyceride, platelet, prothrombin, and other blood test indicators. These findings also reveal a significant correlation between the different stages of cirrhosis and a series of blood test indicators. This information is crucial for the diagnosis, treatment decision-making, and patient management of cirrhosis [7]. Medical professionals can use these correlations to monitor disease progression, evaluate the treatment effect, and take appropriate interventions. In addition, these findings also provide a direction for further research to better understand the pathophysiology of cirrhosis [8] and develop new treatment strategies [10].

Currently, most research projects on the staging of liver cirrhosis primarily focus on the predictive value of individual clinical indicators for the diagnosis and assessment of cirrhosis, lacking comprehensive analysis. This has led to an inability to fully consider the interactive effects and mutual constraints of multiple indicators when evaluating disease progression. This article will conduct a correlation analysis of multiple clinical characteristic indicators.

3.2 Forecasting results

The matrix in Fig.1 shows the cases where the true value (0, 1, 2) is 0 and the predicted values are 0, 1, and 2. Specifically, at true value 0, there are 1900 predicted values of 0, 102 predicted values of 1, and 46 predicted values of 2; at true value 1, there are 1202 predicted values of 0, 1903 predicted values of 1, and 87 predicted values of 2; at true value 2, there are 49 predicted values of 0, 88 predicted values of 1, and 1903 predicted values of 2. There are 49 predictions of 0, 88 predictions of 1, and 1903 predictions of 2. From the matrix, it can be seen that when the true value is 1, the classifier predicts 1 with the highest accuracy, while when the true value is 0, the classifier predicts 0 with higher accuracy. Decision trees can exhibit high accuracy rates and relatively low error analysis rates across these three categories, demonstrating the strength of their performance. However, there is still room for relative improvement, particularly in reducing misclassification among closely related categories. This can be achieved through further analysis and model adjustments to enhance the classifier’s accuracy and recall rates across all categories.

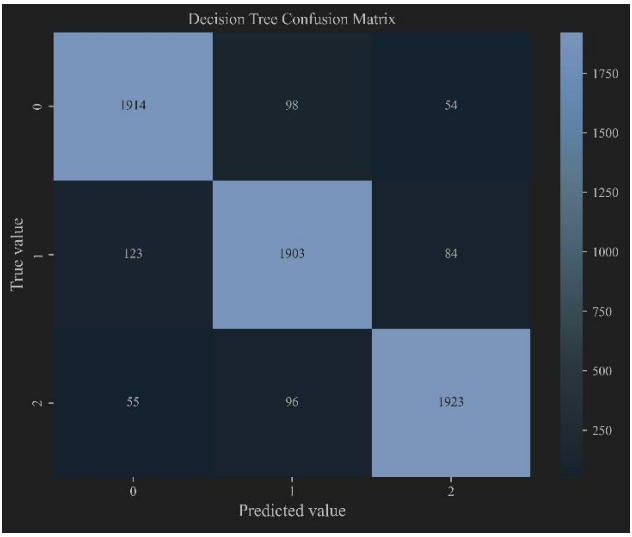

Fig.1 Logic tree confusion matrix

Fig.1 shows the performance of different classifiers on positive and negative samples. As the false positive rate (FPR) changes, the true positive rate (TPR) also changes.

There are three ROC curves, each representing a different conditions. The first yellow curve, with an area of 0.95, is the best performer. The second red curve, with an area of 0.93, is the next best performer. The third green curve, also with an area of 0.95, performs similarly to the first curve. The blue dotted line represents the reference line randomly guessed, where TPR and FPR are equal, both being 0.5. The shapes of the curves are close to the diagonal from bottom left to top right, indicating that the performance of these classifiers is quite good.

As for area indicators, generally, the larger the area below the ROC curve, the better the performance of the classifier. In this figure, the area under the three curves is very large, indicating that the classifiers for these three categories or conditions have high accuracy.

In summary, by plotting the ROC curve and calculating the AUC, the performance and behavior of decision trees can be effectively evaluated across different categories or conditions.

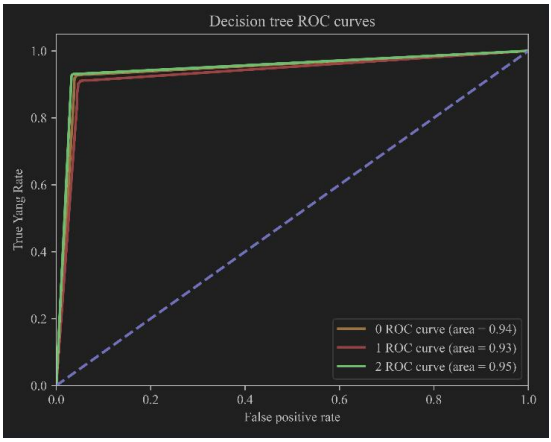

Fig.2 Screenshot of Logic Tree ROC Curve

Fig.2 shows the relationship between the true (0, 1, 2) and predicted (0, 1, 2) values. When the true value is 0, the number of predictions for 0 (20,000) is much higher than the number of predictions for 1 and 2, indicating that the model has a high accuracy in identifying samples with a true value of 0. When the true value is 1, the number of predictions for 2 (20,000) is unusually high, indicating that the model performs poorly in identifying samples with a true value of 1, and there may be classification errors. When the true value is 2, no specific data is provided in the text, but according to the description in Fig.3, the probability of predicting 2 is high, which may imply that the model performs well in identifying samples with a true value of 2.

The model may have a certain bias because it tends to predict samples with a true value of 1 as 2. This could be due to reasons such as imbalanced data, inappropriate feature selection, or model overfitting. In summary, the random forest model performs well in handling samples with a true value of 0, but there are obvious issues when dealing with samples with a true value of 1. Further optimization and adjustment of the model are needed to improve overall performance.

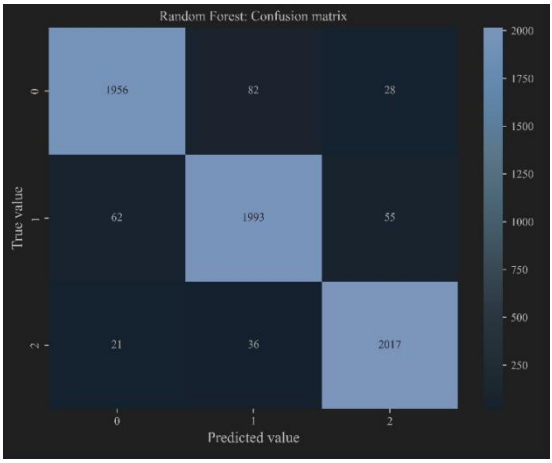

Fig.3 Random Forest Confusion Matrix

It can be seen from Fig.4 that there are three ROC curves, which are represented by different colors: The orange line represents the performance of the random forest model on the test set, with an area close to 0.99, indicating that the model performs very well with high accuracy and recall. The red line also represents the performance of the random forest model on another test set, with an area close to 0.99, indicating similarly excellent performance. The green line represents the performance of a different model or baseline with an area of 1.00, indicating that this is a perfect classifier that correctly predicts the outcome at all times.

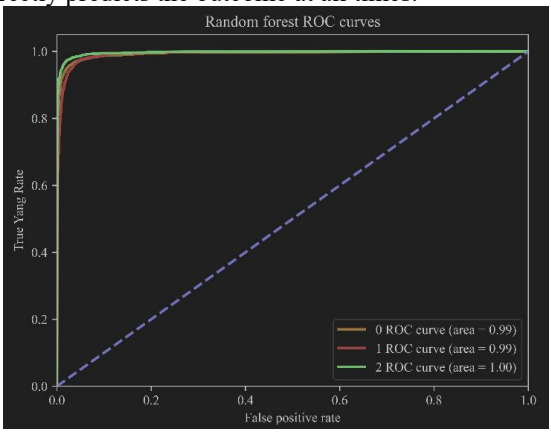

Fig.4 Screenshot of ROC curve of logistic forest

For a true value of 0, the number of forecasts with a value of 0 is 1200, and the number of forecasts with a value of 1 is 750 (Fig.5). Precision for class 0 would be 1200 / (1200 + 750), and for the classes not equal to 0, it would be (430 + 380 + 610 + 470) / (430 + 380 + 610 + 470 + 750 + 1300). However, without the total number of true positives for each class, the exact precision cannot be calculated. Compare the

logistic regression model's performance with other models, such as random forest, to determine if an alternative algorithm might yield better results.

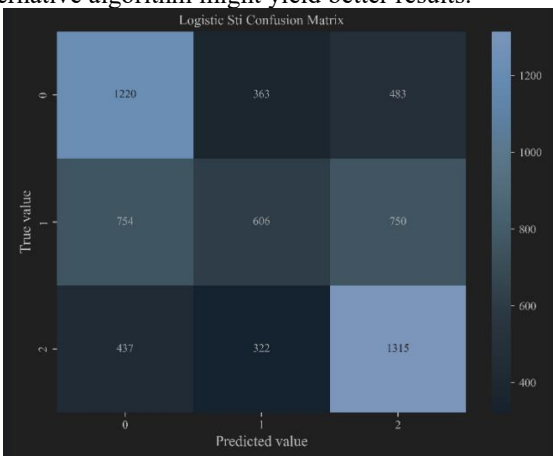

Fig.5 Screenshot of logistic regression confusion matrix.

Brown, red, and green represent 0 ROC curve (area = 0.72), 1 ROC curve (area = 0.61), and 2 ROC curve (area = 0.74) respectively.

Vertical coordinates indicate the true positive rate, with values ranging from 0.0 to 1.0. Horizontal coordinates indicate the false positive rate, with values ranging from 0.0 to 1.0.

Fig. 6 displays three ROC curves represented by brown, red, and green, corresponding to class 0 (area = 0.72), class 1 (area = 0.61), and class 2 (area = 0.74) respectively. All three curves exhibit an upward trend, with the green curve having the highest true positive rate and the brown curve having the lowest true positive rate.

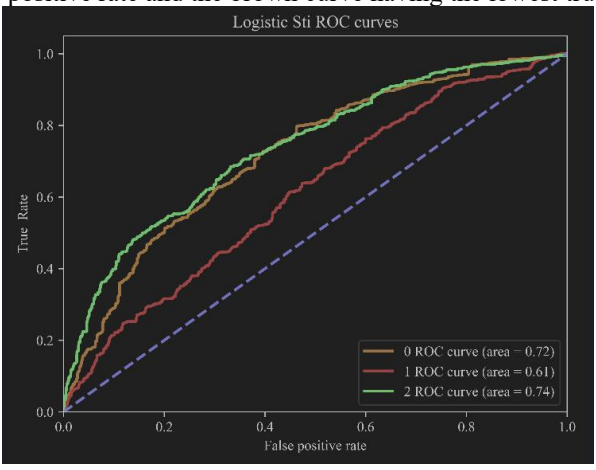

Fig.6 Logic regression ROC curve

4 Conclusion

This paper utilizes machine learning algorithms to train and test data, including decision trees, random forests, and others for comparative experiments. This paper identifies a certain correlation between clinical features and staging of cirrhosis patients during the model training process. Notably, liver function indexes from laboratory tests and liver hardness values from imaging features show certain correlations with cirrhosis staging. These features are important predictors in the machine learning algorithm, crucially influencing their predictive accuracy. Finally, a prediction model based on machine learning algorithms was established for cirrhosis patients' clinical features and staging. The model demonstrates high accuracy and reliability, providing powerful support for clinical decision-making. Utilizing this model enables quick and accurate predictions of cirrhosis patients' clinical characteristics and staging, helping to develop personalized treatment plans and improve the quality of patients' survival outcomes. However, this study still has some limitations that should be considered. First, the data was obtained from a single center, potentially introducing selection bias. Second, the sample size of the study is relatively small, which may affect the model's generalizability. Therefore, future studies should aim to expand the sample size and conduct multicenter investigations to improve the accuracy and reliability of the model.

In conclusion, leveraging machine learning algorithms for predicting cirrhosis patients' clinical characteristics and staging offers a noninvasive and replicable approach with significant implications for clinical practice.

References

1. He, Y. Research progress of hepatitis B, liver failure and liver fibrosis. APASL (2024).
2. Special News of EASL. Focus on the management of hepatitis B, cirrhosis and portal hypertension. Retrieved from https://new.qq.com/rain/a/20240610A02PC100 (2024).
3. Liver International. Research progress of hepatitis B, liver failure and liver fibrosis, APASL (2024).
4. Friedman, S. New progress in global liver fibrosis research. Journal of Clinical Hepatobiliary Diseases (2020).
5. Global Conversations. Co-compensated cirrhosis. Retrieved from https://www.thepaper.cn/newsDetail_forward_20849085 (2022).
6. Feng, G., Song, J., Ye, F., Ma, Y., Ren, Y., Zhang, Z.,: Recompensation of liver cirrhosis: Current status and challenges. The Journal of Clinical Hepatobiliary Diseases, 10, 2464-2469 (2023).
7. Huang, C., Huang, Y., Liu, Y., Han, Y., Zhang, X., Zhao, P., & Liao, H.: Correlation analysis of clinical test index and pathological stage in 54 patients with primary biliary cirrhosis. The Journal of Clinical Hepatobiliary Diseases, 02, 185-188 (2015).
8. Shan, S., Zhao, L., Ma, H., Ou, X., You, H., & Jia, J.: Definition, etiology, and epidemiology of liver cirrhosis. Clinical Journal of Hepatobiliary Diseases, 01, 14-16 (2021).

9. Wang, B., Liu, Y., & Wang, P.: Correlation study between coagulation index and liver function grade in patients with cirrhosis. Journal of Baotou Medical College, 05, 51-54+67,(2023).
10. Shao, C., & Xu, Y.: Meta-analysis of the correlation between serum CA125 levels and the degree of liver damage in cirrhotic patients. Journal of Clinical Hepatobiliary Diseases, 04, 796-800 (2019).