Industry Science and Engineering Vol. 3 No. 2, 2026

Prediction and Validation of Liver Disease Based on Machine Learning Models

Guoyin Li, Guanghu Zhu, Zexian Lu, Zhen Wang*
Guilin University of Electronic Technology, Guilin, Guangxi, China
**Corresponding Author*

Abstract: This study aims to construct and validate machine learning models for liver disease prediction, screen the optimal model, and interpret the model using Shapley Additive exPlanations (SHAP) and Feature Permutation Method (FPM). The "Liver Disease Patient Dataset 30K train data" released on the Kaggle platform was selected, with a total of 16,308 samples included, including 11,669 patients with liver disease (71.55%), and 5-fold cross-validation was performed. Derivative processing was performed on the original features, and four algorithms were used to construct prediction models. Model performance was assessed using the area under the receiver operating characteristic curve (AUC). Meanwhile, model interpretability was illustrated using performance radar charts, decision curve analysis, feature permutation importance figures, and SHAP visualization plots. In the test set via 5-fold cross-validation, the gradient boosting decision tree (GBDT) delivered the best overall performance, with an average AUC of 0.9995 (95% CI: 0.9990–1.0000). The SHAP heatmap and feature permutation importance plot showed that ALT/ALP and ALP had the greatest impact on liver disease identification. Four machine learning models for liver disease prediction were successfully constructed and validated, among which the GBDT model performed the best, which can offer a dependable basis for the early screening of clinical patients with liver disease.

Keywords: Liver Disease; Machine Learning; SHAP Value; Feature Permutation

1. Introduction

Liver disorders have grown into a widespread health concern affecting millions of people across the globe [1]. In research on liver disease prediction, the proportion of subtype prediction is too high, while basic liver disease prediction is insufficient [2,3]. At present, most subtype studies take diagnosed liver disease patients as research objects, providing a basis for precise intervention in the progressive stage of the disease, but do not involve the early identification of basic liver disease. Therefore, it is urgent to develop a simple, non-invasive, and more efficient clinical identification model to enable early, accurate screening and identification of patients with liver disease [4].

Machine learning technology has been extensively applied in population disease prediction and diagnosis. Polat proposed a classification algorithm based on Fuzzy Logic Artificial Immune Recognition System (Fuzzy-AIRS), with an accuracy of 83.36% for classifying liver disease datasets, which had advantages over the AIRS algorithm in classification time and accuracy, and could provide technical support for in-depth research on patient diagnosis and early prevention in the general population [5]. Tripathi constructed early diagnosis models for liver disease using three machine learning algorithms: Decision Tree, Bagging Classifier, and XGBoost, all with accuracies exceeding 99%, illustrating the potential of machine learning in non-invasive, scalable, and accurate liver disease diagnosis [6]. Mounita Ghosh employed multiple machine learning approaches to diagnose liver disease, among which the best-performing random forest obtained an accuracy of 83.70% [7]. Lin Ronghe used 510 outpatients with liver disease diagnoses as a dataset, and the accuracy of CART in classifying liver disease was 92.94% with small samples [8].

Previous studies on liver disease prediction have limited reference value, with the existing prediction accuracy only about 0.85. Most conclusions are based on small sample sizes or inadequately representative data, and the interpretability of the models is also poor.

In view of this, we will introduce the Feature

Permutation Method (FPM) and Shapley Additive exPlanations (SHAP) framework to the optimal model to enhance the interpretability of the model [9].

2. Data

The data utilized in this research was derived from the "Liver Disease Patient Dataset 30K train data" made available on the Kaggle platform, and there has been no previous research that employed this dataset for the construction of a liver disease prediction model. Among 30,689 samples, 3,531 samples with missing values and 122 samples with illogical data (total bilirubin content less than direct bilirubin content) were removed, duplicate samples were screened, and 16,308 samples were finally retained.

2.1 Data Collection

Original features: age, gender, total bilirubin (TBIL), direct bilirubin (DBIL), alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total protein (TP), albumin (ALB), albumin/globulin ratio.

Derived variables:

Indirect bilirubin (IBIL):

$$IBIL = TBIL - DBIL. \quad (1)$$

Globulin (GLO):

$$GLO = TP - ALP. \quad (2)$$

ALT/AST ratio:

$$AST/ALT = \frac{AST}{ALT}. \quad (3)$$

TBIL/ALB ratio:

$$TBIL / ALB = \frac{TBIL}{ALB}. \quad (4)$$

TBIL/DBIL ratio:

$$TBIL / DBIL = \frac{TBIL}{DBIL}. \quad (5)$$

ALT/ALP ratio:

$$ALT / ALP = \frac{ALT}{ALP}. \quad (6)$$

ALB/ALP ratio:

$$ALB / ALP = \frac{ALB}{ALP}. \quad (7)$$

2.2 Statistical Methods

Excel and MATLAB 2023b were used for data sorting, and MATLAB 2023b for data analysis and visualization. After normality testing, all quantitative variables showed non-normal distributions. MATLAB 2023b software was used to construct and compare models—the test level α=0.05.

2.3 Data Preprocessing

In the current research, the 16,308 samples were randomly partitioned into five independent subsets. Four subsets were used for model training, and the remaining one subset was adopted to validate model effectiveness.

2.4 Model Construction and Validation

The model was manually adjusted during training, and 5-fold cross-validation was adopted. Model training and validation based on separate training subsets can effectively estimate the effectiveness of machine learning models, consequently decreasing the likelihood of overfitting. Finally, four prediction models were established: extremely randomized trees (Extra-Trees), random forest (RF), gradient boosting decision trees (GBDT), and back-propagation neural networks (BPNN) [10]. The receiver operating characteristic (ROC) curve was applied to estimate the model's discrimination capacity, and AUC, accuracy, recall, precision, F1 score, and other indicators were calculated. The model with the maximum AUC score was chosen as the optimum model. A risk assessment tool was established using the best-performing model to enable clinicians' efficient and accurate identification of liver disease.

2.5 Interpretability Analysis

Although machine learning models achieve good predictive performance, their results are difficult to interpret, and a single analysis has limitations. Therefore, this study used SHAP values and a feature-permutation method to analyze the influence of each variable on liver disease within the optimal model and to visualize the contribution of a single variable to the model's output. The Feature Permutation Method (FPM) is a classic method for quantifying the contribution of features to model prediction. Its core principle is to destroy the original correlation between the feature and the outcome variable by randomly perturbing the value of a single feature while keeping the others unchanged, and then observing changes in model performance (e.g., accuracy) to evaluate the feature's importance.

3. Results

3.1 Characteristics

The analysis involved 16,308 samples, of which

11,669 (71.55%) were diagnosed with liver disease. The proportion of males was high in both groups, and there were statistically significant differences in all variables except age (all P<0.05). See Table 1.

Table 1. Baseline Characteristics of the Healthy Group and the Liver Disease Group

Variable	Normal group (n=4639)	Liver disease group (n=11669)	Z/χ^2 value	*P* value
Age(years)	43.00(32.00, 50.00)	45.00(32.00, 55.00)	-0.92	0.36
Gender			0.98	<0.001
Male	3232(69.67)	8127(69.65)		
Female	1407(30.33)	3542(30.35)		
TBIL ($\mu mol/L$)	0.80(0.70, 1.10)	1.40(0.80, 3.70)	-40.53	<0.001
DBIL ($\mu mol/L$)	0.20(0.20, 0.30)	0.50(0.20, 1.80)	-39.51	<0.001
ALP(U/L)	187.00(162.00, 215.00)	228.00(186.00, 316.00)	-34.04	<0.001
ALT(U/L)	27.00(20.00, 38.00)	42.00(25.00, 79.00)	-37.57	<0.001
AST(U/L)	29.00(21.00, 43.00)	54.00(30.00, 113.00)	-40.19	<0.001
TP(g/L)	6.60(5.90, 7.30)	6.50(5.70, 7.20)	2.17	<0.001
ALB(g/L)	3.40(2.90, 4.00)	3.00(2.50, 3.60)	22.45	<0.001
ALB/GLO	1.00(0.90, 1.20)	0.90(0.70, 1.10)	24.19	<0.001
IBIL ($\mu mol/L$)	0.60(0.50, 0.70)	0.90(0.60, 1.90)	-39.71	<0.001
GLO(g/L)	3.20(2.90,3.60)	3.30(3.00, 3.90)	-16.08	<0.001
ALT/AST	1.12(0.84,1.53)	1.22(0.84, 1.84)	-11.16	<0.001
TBIL/ALB	0.26(0.20, 0.36)	0.45(0.26, 1.45)	-41.82	<0.001
TBIL/DBIL	3.50(3.00, 4.00)	2.67(2.09, 4.00)	32.77	<0.001
ALT/ALP	0.15(0.10, 0.20)	0.18(0.10, 0.28)	-16.75	<0.001
ALB/ALP	0.02(0.01, 0.02)	0.01(0.01, 0.02)	36.24	<0.001

3.2 Model Comparison

The GBDT model performed best on the test set, with the highest levels of all indicators. The average AUC value, average accuracy, average recall, average precision, and average F1 were 0.9995 (95% CI: 0.9990-1.0000), 0.9997, 0.9998, 0.9997, and 0.9998, respectively, as shown in Figure 1. The data results indicated that GBDT achieved the highest indicators and the best overall performance; in terms of clinical practicability, the GBDT model showed a substantial net benefit in a wide scope of threshold probabilities, as illustrated in Figure 2. Therefore, GBDT was selected as the best model.

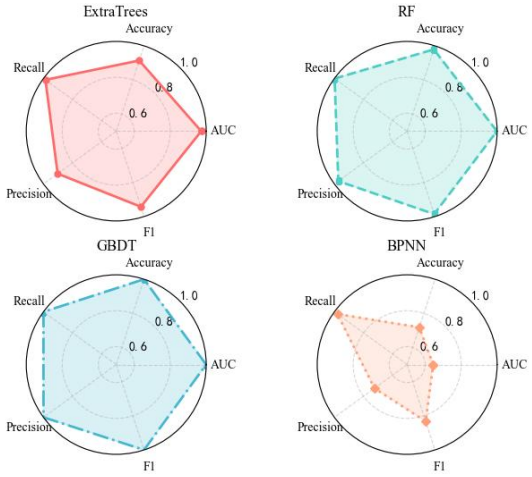
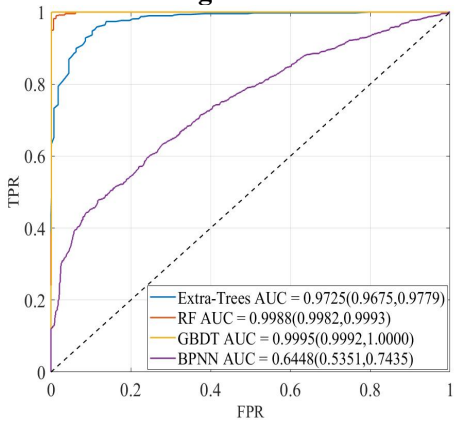
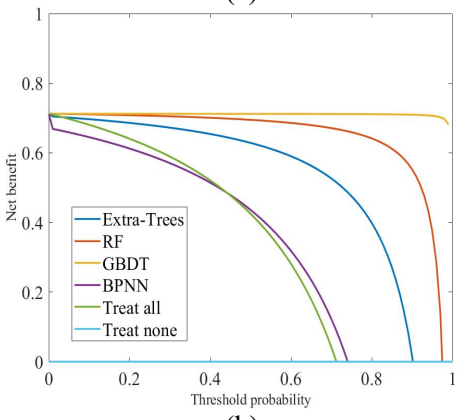


Figure 1. Performance Radar Charts of Machine Learning Models on the Test Set



(a)



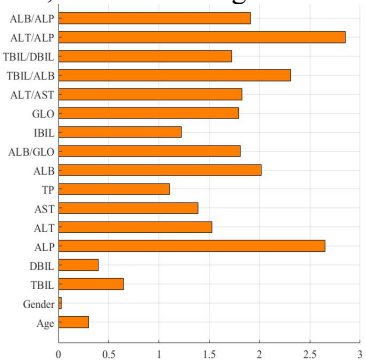
(b)

Figure 2. Model Assessment

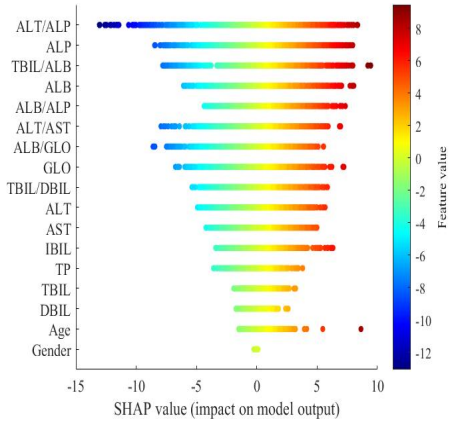
3.3 Interpretability Analysis

From the SHAP heatmap and SHAP mean absolute value bar chart, it can be seen that ALT/ALP and ALP had the greatest impact on liver disease identification, as shown in Figure 3. In the variable importance plot under the feature permutation method, ALT/ALP, ALT/AST, and

ALP had the greatest impact on liver disease identification, as shown in Figure 4.



(a) Mean Absolute SHAP Values for Different Variables



(b) SHAP Heatmap; SHAP: Shapley Additive

Figure 3. Visual Explanation of Machine Learning Models with SHAP Values

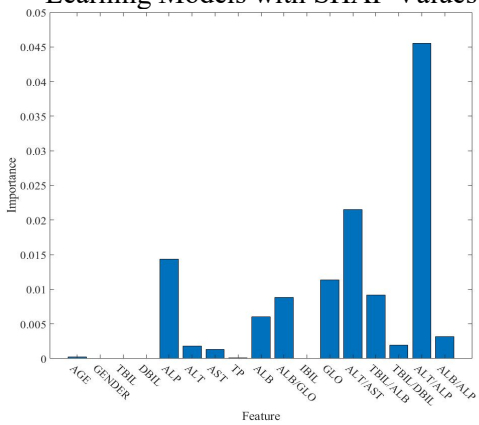


Figure 4. Variable Importance based on Feature Permutation

4. Discussion

Constructing a convenient and efficient liver disease risk prediction model based on routine clinical indicators has important practical value for primary medical institutions and large-scale epidemiological surveys. This study used the Kaggle database to construct a clinical prediction model for early screening of liver disease.

Studies have shown that models such as Extra-Trees, RF, and GBDT show excellent performance in liver disease prediction, among which GBDT has the best comprehensive performance: AUC is as high as 0.9995 (95% CI: 0.9990-1.0000), and accuracy, recall, precision and F1 score are all close to 1.0 (0.9997, 0.9998, 0.9997, 0.9998 respectively), indicating that the model has a strong ability to identify liver fibrosis and almost no deviation in positive and negative sample classification.

Compared with existing studies, the innovations of this study are as follows: (1) Focusing on the basic question of "whether to suffer from liver disease", filling the gap in current research where subtype prediction accounts for too high a proportion while basic prediction is insufficient, and can be directly applied to basic liver disease screening in healthy people; (2) Most previous studies relied on original biochemical indicators to construct models with a single feature dimension, which was difficult to reflect liver conditions fully. This study integrated 10 original indicators and 8 derived ratios. The derived indicators cover multiple dimensions such as bilirubin metabolism (IBIL, TBIL/DBIL), enzymatic balance (ALT/AST, ALT/ALP), and liver synthesis and metabolism coordination (TBIL/ALB, ALB/ALP), which can more comprehensively capture subtle changes in liver injury and provide data feature support for achieving high model performance; (3) Most previous liver disease prediction studies constructed models based on single-center small sample (<1000 cases) data, which were affected by sample representativeness, regional differences and other factors, resulting in large result bias and limited generalization ability. After strict data cleaning (removal of missing values, logical abnormalities, and duplicate samples), this study included 16,308 samples, of which 71.55% were from patients with liver disease. The sample size was 10-20 times that of previous studies and included population data from multiple regions worldwide, effectively reducing sampling error and ensuring the model's stability and applicability.

The study has the following limitations: (1) Although the Kaggle dataset has a large sample size and a wide range of data sources, there may be implicit biases in regional and population characteristics, and the detection standards of different medical centers and baseline

characteristics of patients are also different. The model performance in this study was based solely on internal 5-fold cross-validation, with no cross-population or cross-center external validation. (2) To ensure data quality, samples with missing values and duplicates were deleted in the preprocessing stage. Although data bias was reduced, it may lead to sample size reduction, and deleting samples with missing values may introduce selection bias; (3) Limited by the Kaggle database, this study cannot obtain specific subtype information of liver disease patients. The model can only judge "whether to suffer from liver disease" and cannot distinguish subtypes. The pathogenesis and treatment strategies of different liver disease subtypes are significantly different, and prediction based only on "whether to suffer from the disease" is difficult to meet the needs of precision medicine. For future investigations, it is essential to combine subtype annotation data to construct a hierarchical prediction model to realize two-step prediction of "first judge whether to suffer from the disease, then distinguish subtypes"; (4) This study is a cross-sectional study, which only predicts liver disease status based on baseline indicators and cannot verify the model's ability to predict "whether liver disease will occur in the future".

In conclusion, four machine learning models for predicting liver disease were constructed and validated in this study, among which the GBDT model performed the best, providing a reliable reference for the early clinical screening of liver disease patients. Future research can further improve the model's practicality by incorporating multi-center external data, supplementing subtype information, optimizing methods for handling missing values, etc., thereby providing stronger support for the precise prevention and management of liver disease.

Acknowledgments

This paper received support from the Guangxi Natural Science Foundation (Nos. 2026GXNSFAA00640927, 2024GXNSFAA010475), the National College Student Innovation and Entrepreneurship Training Program at Guilin University of Electronic Technology (S202410595262), the Course Construction Project of GUET Graduate Education (No. YKC202412), and the Innovation Project of Guangxi Graduate Education (JGY2025159).

References

[1] XIAO J, WANG F, YUAN Y, et al. Epidemiology of liver diseases: global disease burden and forecasted research trends. SCIENCE CHINA Life Sciences, 2025, 68(2). DOI: 10.1007/s11427-024-2722-2.

[2] ROSENSTENGLE C, SERPER M, ASRANI SK, et al. Variation in intention-to-treat survival by MELD subtypes: All models created for end-stage liver disease are not equal. Journal of Hepatology, 2025, 82(2). DOI:10.1016/j.jhep.2024.08.006.

[3] PATEL K, ASRANI SK, FIEL MI, et al. Accuracy of blood-based biomarkers for staging liver fibrosis in chronic liver disease: A systematic review supporting the AASLD Practice Guideline. Hepatology, 2025, 81(1):22. DOI: 10.1097/HEP.0000000000000842.

[4] YOUNES R, CAVIGLIA GP, GOVAERE O, et al. Longte-rm outcomes and predictive ability of non-invasive scoring systems in patients with non-alcoholic fatty liver disease. Journal of Hepatology, 2021(7). DOI:10.1016/j.jhep.2021.05.008.

[5] POLAT K, AHAN S, KODAZ H, et al. Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. Expert Systems with Applications, 2007, 32(1):172-183. DOI:10.1016/j.eswa.2005.11.024.

[6] TRIPATHI A, RAGIRI PR, JAIN D, et al. Machine Learning-based Predictive Models for Early Diagnosis of Liver Disease. Journal of Scientific & Industrial Research, 2025, 84(5). DOI:10.56042/jsir.v84i5.14828.

[7] GHOSH M, RAIHAN MMS, RAIHAN M, et al. A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease. Intelligent Automation and Soft Computing, 2021(3). DOI: 10.32604/IASC.2021.017989.

[8] LIN RH. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine, 2009, 47(1): 53-62. DOI: 10.1016/ j.artmed.2009.05.005.

[9] MARZOUK R, COLIN DLH. On the Tractability of SHAP Explanations under Markovian Distributions. 2024.

[10]DENG Y, ZHI P, ZHU W, et al. Prediction of PM2.5 Concentration Based on Bayesian Optimization Random Forest. 2024 43rd Chinese Control Conference (CCC), 2024:8507-8511. DOI: 10.23919/ccc63176.2024.10662123.