Am J Cancer Res 2026;16(7):3119-3135
www.ajcr.us /ISSN:2156-6976/ajcr0171529

Original Article

An interpretable machine learning model for diagnostic classification of liver cancer using multivariable clinical data

Dianyu Wang[1,2*], Xiao Li[2,3*], Zuheng Wang[2,4*], Chunmeng Wei[2,4], Wenhao Lu[2], Rongbin Zhou[2], Leifeng Liang[5], Fubo Wang[2,3,4], Junyi Chen[1]

[1]Department of Urology, The Second Affiliated Hospital of Fujian Medical University, Quanzhou 362000, Fujian, China; [2]Center for Genomic and Personalized Medicine, Guangxi Key Laboratory for Genomic and Personalized Medicine, University Engineering Research Center of Digital Medicine and Healthcare, Guangxi Medical University, Nanning 530021, Guangxi, China; [3]School of Life Sciences, Guangxi Medical University, Nanning 530021, Guangxi, China; [4]Department of Urology, The First Affiliated Hospital of Guangxi Medical University, Guangxi Medical University, Nanning 530021, Guangxi, China; [5]Department of Oncology, The Sixth Affiliated Hospital of Guangxi Medical University, The First People's Hospital of Yulin, Yulin 537000, Guangxi, China. [*]Equal contributors.

Received January 18, 2026; Accepted July 6, 2026; Epub July 15, 2026; Published July 30, 2026

Abstract: In this retrospective analysis, we used routinely collected clinical variables to develop a machine learning model for liver cancer diagnosis and examined the variables that contributed to model performance. We studied 3,629 people who were assessed because they were suspected to have liver cancer or other liver diseases. Patients who had pathologically confirmed liver cancer, and controls comprised of healthy subjects, patients with chronic hepatitis B, chronic hepatitis C, or cirrhosis. From 87 clinical parameters, least absolute shrinkage and selection operator regression was applied to identify candidate predictors. Nine machine learning algorithms were then trained and tested through 10-fold cross-validation and external validation in an independent cohort. Performance of the models was measured based on area under the receiver operating characteristic curve, sensitivity, specificity, calibration and decision curve analysis. The variables that contributed most to discrimination included carbohydrate antigen 19-9, alpha-fetoprotein-related indicators, liver function parameters, and inflammatory markers. Among the nine algorithms, Extreme Gradient Boosting achieved the highest discriminative performance, with an AUC of 1.000 in the training cohort and 0.937 in the validation cohort. In the SHapley Additive exPlanations analysis, carbohydrate antigen 19-9 made the largest contribution to the final model. These findings suggest that machine learning algorithms, particularly Extreme Gradient Boosting, may integrate heterogeneous clinical indicators and provide a useful auxiliary tool for distinguishing liver cancer from non-liver cancer conditions in individuals undergoing clinical assessment. Further prospective validation in high-risk surveillance cohorts is required before the model can be applied to early risk prediction or population-level screening.

Keywords: Liver cancer, machine learning, multivariate clinical data, diagnostic classification

Introduction

The incidence and mortality rates of liver cancer are high around the world. Liver cancer as a primary tumor was the sixth-most frequent cancer in the world and the third-most frequent cause of cancer-related death according to the GLOBOCAN 2022 projections [1]. The most frequent histological form of primary liver cancer is hepatocellular carcinoma (HCC). Due to the lack of early-stage symptoms or minimal symptoms, it is frequently detected later when the disease has reached the intermediate or late stage. This gap restricts opportunities to treat the condition curatively and results in suboptimal clinical outcomes [2]. The outlook of HCC is also not encouraging. Survival rate in HCC is still poor and five year survival rate has been stated to be about 15% [3]. Since the majority of patients are identified at later stages and

https://doi.org/10.62347/FTNM6925

there are currently no effective therapies, more effective methods are required to diagnose and evaluate liver cancer in its initial stage.

Liver cancer is primarily detected by serological testing, imaging assessment and histopathological verification. AFP has long been routinely measured as a serum biomarker in clinical practice although it is not very effective in early detection of liver cancer [4]. About 32-59 percent of people with liver cancer could have normal AFP values [5]. However, increased serum AFP can also be found in various benign and malignant conditions other than HCC and thus decrease its specificity [6]. Even though imaging is invaluable in clinical diagnosis, small or early-stage HCC lesions can be overlooked, and interpretation accuracy can vary according to the level of experience of the radiologist [7]. Liver biopsy is the best diagnostic procedure when it comes to liver cancer, but it is invasive and associated with the risk of complications that restrict its application in daily clinical practice [8]. Such deficiencies require the formulation of non-invasive and more accurate ways to identify liver cancer in earlier stages.

Machine learning (ML) has become an increasingly used approach in medical research, with applications in disease prediction, diagnostic modeling, prognosis evaluation, and treatment strategy optimization [9]. Since ML methods are able to reflect complex relations in clinical and biological data, they have been employed to enhance diagnostic and prognostic prediction models [10, 11] and are currently applied to a broad spectrum of medical datasets [12-14]. In cancer research, ML is particularly suitable for analyzing high-dimensional clinical data and integrating multiple indicators into predictive models that may support clinical decision-making [15]. Compared with single-biomarker detection, multi-indicator models developed with ML algorithms can provide a more comprehensive assessment and may improve the stability and generalizability of prediction models [16]. Recent diagnostic studies on liver cancer using ML have revealed promising findings, suggesting that this method could contribute to the enhancement of early diagnosis [17].

Therefore, in this study, we developed and compared nine machine learning models using retrospective multivariable clinical data to establish an interpretable diagnostic classification model for liver cancer.

Materials and methods

Study participants

The medical records of 3,629 individuals treated at the First People's Hospital of Yulin City between May 2017 and December 2023 were reviewed. The cohort consisted of 2,715 patients with pathologically confirmed liver cancer and 914 non-liver cancer controls. The controls comprised healthy individuals without liver disease, and patients on chronic hepatitis B, chronic hepatitis C or cirrhosis with no indication of liver cancer during evaluation. The absence of liver cancer in controls was determined from the available clinical records, imaging examinations, laboratory tests, and pathological findings when applicable.

Eligible patients had newly diagnosed, pathologically confirmed liver cancer, had not received antitumor treatment before enrollment, and had complete clinical data available for analysis. Patients were excluded if key clinical information was missing, if they had another malignancy or severe extrahepatic disease, or if they had undergone liver cancer-directed treatment before enrollment, including hepatectomy, transarterial chemoembolization, targeted therapy, or other antitumor interventions. The study was approved by the Ethics Committee of Guangxi Medical University (2024-KY0180), and informed consent was obtained from all participants.

Study design

In this study, patients were randomly divided into a training cohort (80%, n = 2,903) and a test cohort (20%, n = 726). The training cohort was used for model development, while the test cohort was used to evaluate the model's generalization performance. Additionally, 10-fold cross-validation was applied for internal validation of the model to ensure its stability and reliability. The flowchart of the study is shown in **Figure 1**.

Study variables

Multivariate clinical characteristics of patients were retrospectively collected, including (1) demographic information: age, height, etc.; (2)

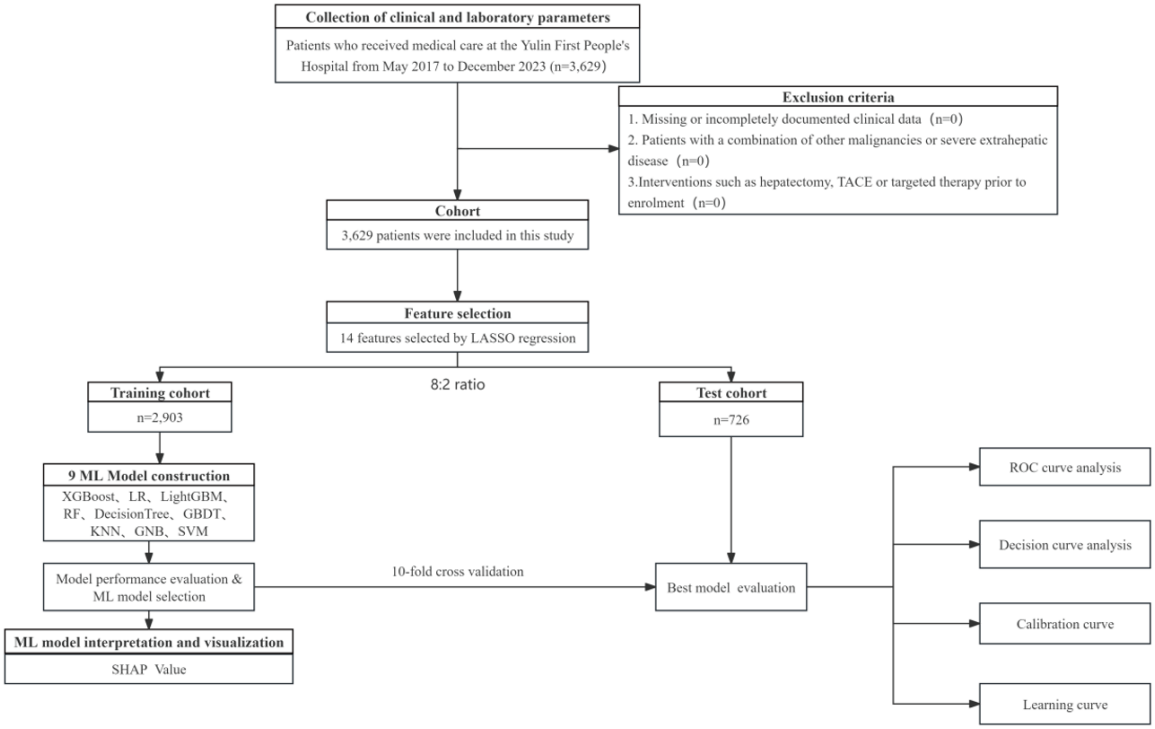

Figure 1. Workflow diagram of this study. LASSO, Least absolute shrinkage and selection operator; ML, Machine Learning; ROC, Rceiver Operating Characteristic; SHAP, Shapley Additive Explanations.

serum biomarkers: alpha-fetoprotein (AFP), carcinoembryonic antigen (CEA), etc.; (3) urine analysis indicators: urinary leukocytes (U-LEU), urobilinogen (U-URO), etc.; (4) other clinical parameters: alkaline phosphatase (ALP), triglycerides (TG), etc. Samples with significant missing data were excluded from the dataset, and samples with fewer missing values were imputed (categorical variables were filled in with 0 and 1, etc.). All preprocessing steps, including imputation and scaling, were performed within each cross-validation fold to prevent data leakage. Additionally, no patient data overlapped between training and test sets.

Feature selection

We adopt the least absolute shrinkage and selection operator (LASSO) regression to screen key features for liver cancer prediction from high-dimensional clinical data. The optimal regularization parameter λ was determined using 10-fold cross-validation to balance model complexity and predictive performance. Most feature coefficients shrank towards zero as λ varied. The λ corresponding to the minimum mean squared error (MSE) and the λ within one standard error of the minimum were selected as optimal. The correspondingly screened features were incorporated into model construction.

Model development and validation

Based on the screened clinical features, Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), Light Gradient Boosting Machine (LightGBM), Random Forest (RF), Decision Tree, Gradient Boosting Decision Tree (GBDT), k-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM) algorithms were employed to independently construct liver cancer prediction models on the training cohort. Following that, several evaluation indexes, such as area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, etc., were used to quantify the predictive ability of the nine ML models, and a comprehensive analysis was performed to determine the best-performing prediction model.

We further tested the optimal model using 10-fold cross-validation in the training set and then examined its performance in the test set to assess generalizability. Calibration curves

were used to compare the predicted probabilities with the observed classification outcomes. To explore whether the model could provide clinical benefit, decision curve analysis (DCA) was also performed by estimating the net benefit at different threshold probabilities.

Model visualization

To reduce the black-box nature of the model and assist clinical decision-makers to correctly interpret and apply the prediction results, it is required to define the level of impact of every feature on the model decision-making process. This problem has been resolved through the present study that performed the interpretable analysis of how much each feature contributes to the prediction results in the model through the application of the Shapley Additive Explanations (SHAP) tool, which increased the model transparency and clinical usability.

Statistical analysis

All statistical analyses were performed using SPSS Statistics software (version 26.0) and Python software (version 3.4.4). The normality of continuous variables was first assessed. Those that followed a normal distribution were reported in terms of mean ± standard deviation and were compared by groups through the use of Student's t-test. Variables that did not follow a normal distribution were reported as median and interquartile range (IQR) and were compared through the use of the Mann - Whitney U test. Frequencies and percentages were used to express categorical variables and the chi-square test was used for between-group comparisons. All statistical tests were two-sided, and P<0.05 was considered statistically significant.

The least absolute shrinkage and selection operator (LASSO) regression combined with 10-fold cross-validation was used as a method of feature selection in order to determine important clinical variables that can be used to classify liver cancer. Linear correlation between the chosen clinical variables was assessed by means of Pearson correlation analysis and the outcome was depicted through the correlation matrix.

Using the chosen features, nine machine learning algorithms, including XGBoost, logistic regression, LightGBM, random forest, decision tree, GBDT, KNN, Gaussian naive Bayes, and SVM were developed for liver cancer diagnostic classification. The randomization of the dataset into a training group and a held-out test group was done in a proportion of 8:2. Internal validation was performed using 10-fold cross-validation in the training cohort, and the diagnostic performance of the optimal model was tested further in the held-out test cohort.

The area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), precision, recall, F1 score, and Kappa value were used to assess model performance. Calibration curve analysis was conducted in order to determine whether the predicted probabilities and the observed outcomes of liver cancer classification could agree with each other. To assess the possible clinical net benefit of the model at various threshold probabilities, decision curve analysis was applied. SHapley Additive exPlanations (SHAP) analysis was used to interpret the contribution of each selected feature to the model output.

Results

Baseline characteristics of the study cohort

The study cohort comprised 3,629 participants, including 2,715 patients with liver cancer and 914 controls. Comparative analysis of baseline characteristics between the two groups revealed statistically significant differences (all $P<0.05$) in 61 out of 87 evaluated clinical parameters. Key variables with significant intergroup disparities included sex, aspartate aminotransferase (AST) levels, CA199, etc. **Table 1** lists the clinical variables that were statistically significantly different between groups. The remaining 26 variables with no significant differences are reported in Table S1.

Characteristic screening and correlation analysis

Based on the variables obtained from the above screening, LASSO regression analysis was used to further screen the clinical parameters that were highly correlated with the occurrence of liver cancer. The results showed (**Figure 2A**, **2B**) that when the standard error of

Table 1. Baseline characteristics with statistically significant group differences

Variables	Overall (n = 3,629)	Control (n = 914)	Liver cancer (n = 2,715)	*p*
Gender[1]				<0.001
Female	1148 (31.63)	476 (52.08)	672 (24.75)	
Male	2481 (68.37)	438 (47.92)	2043 (75.25)	
U_LEU[1]				0.005
Negative	140 (23.49)	34 (17.00)	106 (26.77)	
+	267 (44.80)	105 (52.50)	162 (40.91)	
++	78 (13.09)	20 (10.00)	58 (14.65)	
+++	73 (12.25)	23 (11.50)	50 (12.63)	
++++	38 (6.38)	18 (9.00)	20 (5.05)	
U_NIT[1]				0.022
Negative	3502 (96.50)	869 (95.08)	2633 (96.98)	
+	102 (2.81)	35 (3.83)	67 (2.47)	
++	25 (0.69)	10 (1.09)	15 (0.55)	
U_URO[1]				<0.001
Negative	1992 (82.01)	509 (89.14)	1483 (79.82)	
+	324 (13.34)	52 (9.11)	272 (14.64)	
++	93 (3.83)	7 (1.23)	86 (4.63)	
+++	16 (0.66)	3 (0.53)	13 (0.70)	
++++	4 (0.16)	0 (0.00)	4 (0.22)	
U_BIL[1]				<0.001
Negative	142 (36.22)	40 (59.70)	102 (31.38)	
+	126 (32.14)	13 (19.40)	113 (34.77)	
++	51 (13.01)	9 (13.43)	42 (12.92)	
+++	65 (16.58)	5 (7.46)	60 (18.46)	
++++	8 (2.04)	0 (0.00)	8 (2.46)	
RBC Morphology[1]				0.015
Negative	2129 (85.43)	517 (81.80)	1612 (86.67)	
+	129 (5.18)	38 (6.01)	91 (4.89)	
++	176 (7.06)	55 (8.70)	121 (6.51)	
+++	58 (2.33)	22 (3.48)	36 (1.94)	
Age[2]	56.00 [48.00, 65.00]	54.00 [45.00, 64.00]	57.00 [50.00, 66.00]	<0.001
Height[2]	163.00 [157.00, 168.00]	160.00 [155.00, 166.00]	163.00 [158.00, 168.00]	<0.001
BMI[2]	22.27 [20.03, 24.44]	22.86 [20.44, 24.97]	22.16 [19.96, 24.22]	<0.001
SG[2]	1.02 [1.01, 1.02]	1.02 [1.01, 1.03]	1.02 [1.01, 1.02]	0.022
SED_RBC[2]	6.00 [2.00, 13.00]	6.00 [3.00, 14.00]	6.00 [2.00, 12.00]	0.015
SED_WBC[2]	4.00 [1.00, 11.00]	5.00 [2.00, 20.00]	3.00 [1.00, 10.00]	<0.001
SED_EC[2]	2.00 [1.00, 6.00]	3.00 [1.00, 6.80]	2.00 [1.00, 5.00]	<0.001
SED_bacteria[2]	9.00 [2.00, 45.00]	12.00 [3.00, 71.00]	8.00 [2.00, 40.00]	<0.001
Urine conductivity[2]	15.20 [11.00, 19.70]	16.00 [11.00, 20.70]	14.90 [11.10, 19.30]	0.006
Lymph/WBC[2]	16.70 [9.00, 25.00]	18.00 [8.30, 28.90]	16.30 [9.30, 23.90]	0.001
Mono/WBC[2]	7.10 [5.40, 8.80]	6.40 [4.80, 8.00]	7.30 [5.70, 9.00]	<0.001
Eos/WBC[2]	1.30 [0.30, 2.90]	1.10 [0.10, 2.40]	1.40 [0.30, 3.10]	<0.001
Baso/WBC[2]	0.40 [0.20, 0.60]	0.30 [0.20, 0.60]	0.40 [0.20, 0.60]	0.001
Lymph[2]	1.21 [0.82, 1.68]	1.30 [0.82, 1.81]	1.19 [0.82, 1.63]	<0.001
Mono[2]	0.53 [0.38, 0.75]	0.49 [0.35, 0.69]	0.55 [0.39, 0.77]	<0.001
Eos[2]	0.09 [0.02, 0.19]	0.07 [0.01, 0.15]	0.09 [0.03, 0.21]	<0.001
Baso[2]	0.03 [0.02, 0.04]	0.03 [0.02, 0.04]	0.03 [0.02, 0.05]	0.006

MCV[2]	89.00 [83.40, 93.20]	88.60 [83.00, 92.60]	89.20 [83.40, 93.40]	0.026
MCH[2]	29.50 [27.10, 31.30]	29.40 [27.30, 30.70]	29.60 [27.10, 31.40]	0.004
RDW[2]	13.50 [12.50, 15.40]	13.10 [12.30, 14.90]	13.70 [12.60, 15.60]	<0.001
PLT[2]	207.00 [141.00, 280.00]	223.00 [157.00, 282.00]	202.00 [139.00, 280.00]	0.003
MPV[2]	10.20 [9.60, 10.90]	10.10 [9.50, 10.80]	10.30 [9.60, 11.00]	<0.001
AST[2]	53.00 [27.00, 126.00]	38.00 [21.00, 124.00]	56.70 [30.00, 126.00]	<0.001
TBA[2]	8.40 [3.50, 22.20]	4.50 [1.90, 12.70]	9.70 [4.40, 26.50]	<0.001
TP[2]	69.10 [63.10, 74.30]	68.50 [61.50, 74.00]	69.30 [63.60, 74.40]	<0.001
ALB[2]	35.30 [31.00, 39.50]	37.10 [32.10, 41.80]	34.70 [30.70, 38.80]	<0.001
TBIL[2]	12.90 [8.40, 22.50]	10.90 [7.40, 17.40]	13.80 [9.10, 24.70]	<0.001
DBIL[2]	6.80 [4.20, 12.60]	5.40 [3.50, 9.00]	7.40 [4.50, 13.90]	<0.001
IBIL[2]	5.80 [3.60, 9.00]	5.10 [3.40, 7.90]	5.90 [3.70, 9.40]	<0.001
ALP[2]	102.00 [74.00, 161.00]	82.00 [61.00, 106.00]	114.00 [81.00, 181.00]	<0.001
GGT[2]	78.00 [34.90, 187.00]	38.60 [21.00, 91.00]	98.00 [45.00, 220.00]	<0.001
BUA[2]	282.10 [207.00, 362.40]	269.00 [188.00, 343.30]	287.90 [215.00, 368.60]	<0.001
BUN[2]	4.47 [3.49, 5.79]	4.10 [3.20, 5.30]	4.60 [3.60, 5.91]	<0.001
Scr[2]	76.00 [64.00, 91.00]	71.00 [59.90, 85.00]	77.00 [66.00, 92.60]	<0.001
beta2_MG[2]	2.24 [1.76, 2.83]	1.85 [1.50, 2.41]	2.35 [1.90, 2.96]	<0.001
CysC[2]	1.03 [0.86, 1.26]	0.89 [0.76, 1.11]	1.05 [0.90, 1.30]	<0.001
RBP[2]	23.00 [16.00, 33.00]	29.00 [19.00, 38.00]	21.00 [15.00, 31.00]	<0.001
TG[2]	1.01 [0.77, 1.45]	1.15 [0.81, 1.69]	0.99 [0.76, 1.36]	<0.001
HDL_C[2]	1.00 [0.75, 1.29]	1.12 [0.90, 1.38]	0.96 [0.73, 1.22]	<0.001
AFP[2]	13.40 [2.90, 693.00]	3.30 [2.00, 8.20]	18.00 [3.20, 1060.00]	<0.001
CEA[2]	2.69 [1.59, 5.53]	2.06 [1.23, 3.39]	2.98 [1.77, 6.48]	<0.001
CA125[2]	21.70 [11.50, 70.00]	14.10 [8.71, 23.00]	26.30 [13.00, 91.20]	<0.001
CA199[2]	26.80 [10.90, 69.80]	14.60 [7.67, 28.90]	32.40 [13.13, 88.80]	<0.001
PT[2]	12.40 [11.50, 13.50]	11.90 [11.00, 13.20]	12.50 [11.60, 13.60]	<0.001
INR[2]	1.08 [1.00, 1.19]	1.04 [0.95, 1.16]	1.09 [1.01, 1.19]	<0.001
TT[2]	19.00 [18.10, 20.30]	18.90 [17.90, 20.00]	19.10 [18.20, 20.30]	<0.001
FBG[2]	3.07 [2.29, 4.15]	2.82 [2.19, 3.64]	3.21 [2.35, 4.30]	<0.001
APTT[2]	28.10 [25.00, 32.00]	27.50 [24.20, 31.50]	28.30 [25.20, 32.00]	<0.001
D_Dim[2]	1.72 [0.54, 4.55]	1.18 [0.38, 4.92]	1.79 [0.59, 4.50]	<0.001
HBeAg[2]	0.00 [0.00, 0.04]	0.00 [0.00, 0.02]	0.00 [0.00, 0.04]	<0.001
HBeAb[2]	0.42 [0.10, 4.00]	0.14 [0.07, 1.33]	0.87 [0.13, 4.00]	<0.001
HBsAb[2]	15.06 [0.00, 170.00]	0.01 [0.00, 64.74]	150.98 [0.01, 170.00]	<0.001
HBsAg[2]	1.12 [0.13, 26.61]	7.03 [0.37, 97.88]	0.80 [0.08, 14.68]	<0.001
HBcAb[2]	9.00 [5.71, 9.00]	7.16 [0.80, 9.00]	9.00 [9.00, 9.00]	<0.001
HCV[2]	0.04 [0.04, 0.04]	0.04 [0.04, 0.04]	0.04 [0.04, 0.04]	<0.001

[1]Categorical variables as frequency (percentage); [2]Continuous variables as median [IQR]. U_LEU, urine leukocytes; U_NIT, urine nitrite; U_URO, urobilinogen; U_BIL, urine bilirubin; RBC Morphology, red blood cell morphology; BMI, body mass index; SG, urine specific gravity; SED_RBC, red blood cell count in urine sediment; SED_WBC, white blood cell in urine sediment; SED_EC, epithelial cell count in urine sediment; SED_bacteria, bacteria count in urine sediment; Lymph/WBC, lymphocyte percentage; Mono/WBC, monocyte percentage; Eos/WBC, eosinophil percentage; Baso/WBC, basophil percentage; Lymph, absolute lymphocyte count; Mono, absolute monocyte count; Eos, absolute eosinophil count; Baso, absolute basophil count; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; RDW, red cell distribution width; PLT, platelet count; MPV, mean platelet volume; AST, aspartate aminotransferase; TBA, total bile acid; TP, total protein; ALB, serum albumin; TBIL, total bilirubin; DBIL, direct bilirubin; IBIL, indirect bilirubin; ALP, alkaline phosphatase; GGT, gamma-glutamyl transferase; BUA, uric acid; BUN, blood urea nitrogen; Scr, serum creatinine; beta2_MG, serum β2-microglobulin; CysC, serum cystatin C; RBP, retinol-binding protein; TG, triglycerides; HDL_C, high density lipoprotein cholesterol; AFP, alpha-fetoprotein; CEA, carcinoembryonic antigen; CA125, carbohydrate antigen 125; CA199, carbohydrate antigen 199; PT, prothrombin time; INR, international normalized ratio; TT, thrombin time; FBG, fasting blood glucose; APTT, activated partial thromboplastin time; D_Dim, D-Dimer; HBeAg, hepatitis B e antigen; HBeAb, hepatitis B e antibody; HBsAb, hepatitis B surface antibody; HBsAg, hepatitis B surface antigen; HBcAb, hepatitis B core antibody; HCV, hepatitis C virus antibody.

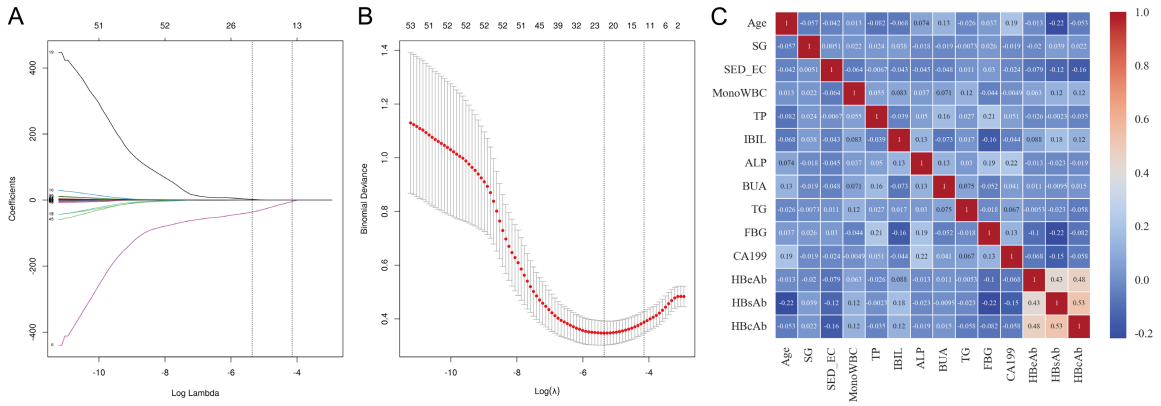

Figure 2. Selecting feature parameters and variable correlation analysis. A. LASSO coefficient path diagram. B. LASSO regularization path diagram. C. Pearson correlation analysis.

the minimum distance of λ was 0.016, the feature variables with strong predictive ability were finally screened out, including Age, SG, SED_EC, Mono/WBC, TP, IBIL, ALP, BUA, TG, CA199, FBG, HBeAb, HBsAb, and HBcAb. These variables exhibited a significant predictive ability and were used to construct an ML model for early diagnosis of liver cancer.

Furthermore, Pearson correlation analyses were performed on the key clinical variables screened to assess the linear relationship between them. The correlation matrix (**Figure 2C**) demonstrated the Pearson correlation coefficients (r) between the variables, ranging from -1 to 1. The results of the correlation analysis revealed the potential relationship between the clinical variables and provided an important reference for the construction of the subsequent ML model.

Model construction

Based on the selected key clinical features, a total of nine ML algorithms were used in this study to construct liver cancer diagnostic classification models. The results from the training cohort (**Figure 3A**; **Table 2**) showed that XGBoost, LightGBM, RF, and Decision Tree models performed exceptionally well, with their main evaluation metrics approaching optimal values. Among them, the AUC for XGBoost, LightGBM, and Decision Tree models was 1.000 (95% CI: 1.000-1.000). The performance of the RF model was also excellent, with an AUC of 1.000 (95% CI: 1.000-1.000), other metrics were slightly lower but still comparable. Additionally, in the training cohort, the AUC for GBDT was 0.985, and for KNN, it was 0.910. LR, SVM, and GNB performed relatively weakly, with AUCs of 0.794, 0.781, and 0.765, respectively.

The validation cohort results (**Figure 3B**; **Table 3**) showed that XGBoost remained the best-performing model with an AUC of 0.937 (95% CI: 0.887-0.984), indicating stable and balanced diagnostic classification performance. Furthermore, LightGBM and RF performed similarly well in the validation cohort, with AUCs of 0.935 and 0.934, respectively, slightly lower than XGBoost. GBDT and Decision Tree perform relatively poorly, with AUCs of 0.893 and 0.824, respectively. Meanwhile, the AUCs of LR, SVM, KNN, and GNB in the validation cohort further decline, with AUCs of 0.783, 0.765, 0.758, and 0.754, respectively.

In summary, although XGBoost, LightGBM, RF, and Decision Tree all performed excellently in the training cohort, the performance of the remaining three models, except XGBoost, decreased in the validation cohort, indicating a possible overfitting issue. In contrast, XGBoost maintained high consistency between the training and validation cohorts, with strong model generalization. CC further confirmed (**Figure 3C**) that XGBoost showed favorable agreement between predicted probabilities and observed liver cancer classification outcomes, while DCA (**Figure 3D**) also showed that it provided higher net clinical benefit at each risk threshold. Therefore, considering multiple assessment metrics and practical application

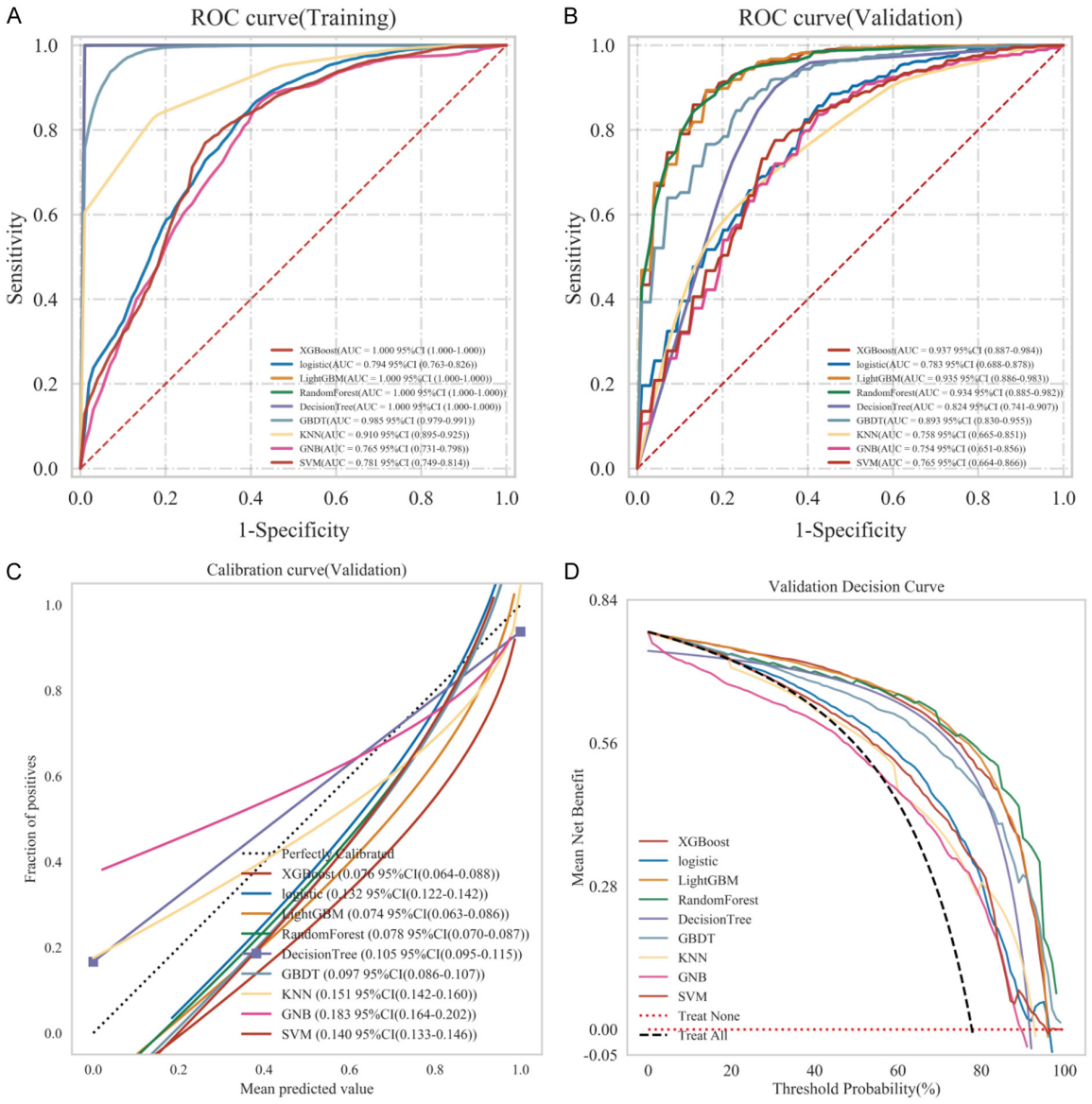

Figure 3. Comprehensive analysis of ML models. A, B. ROC curves and AUC values of the nine ML models in the training and validation cohorts; C. Calibration curves of the nine ML models; D. Decision curve analysis (DCA) of the nine ML models.

value, XGBoost was finally identified as the optimal model for liver cancer diagnostic classification in this study.

Model validation

In this study, 20% (n = 726) of the overall samples were randomly selected as the test set, and the remaining samples were used as the training set. A 10-fold cross-validation approach was employed within the training set to optimize the model's performance. The results of model evaluation (**Figure 4A-C**) demonstrated strong predictive power, with an AUC of 1.000 (95% CI: 1.000-1.000) in the training set, 0.911 (95% CI: 0.838-0.983) in the validation set, and 0.925 in the independent test set (95% CI: 0.888-0.962).

In addition, the learning curves (**Figure 4D**) indicated that the XGBoost model exhibited well-fitting performance and high stability on both the training and validation sets, with no obvious overfitting trend. Overall, these results supported the favorable diagnostic performance and generalization ability of the

Table 2. The performances of 9 machine learning algorithms in the training cohort

Algorithms	AUC (95% CI)	Cutoff (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score (95% CI)	Kappa (95% CI)
XGBoost	1.000 (1.000-1.000)	0.903 (0.871-0.936)	0.999 (0.998-0.999)	0.998 (0.998-0.999)	1.0 (1.000-1.000)	1.0 (1.000-1.000)	0.994 (0.993-0.996)	0.999 (0.999-0.999)	0.996 (0.996-0.997)
Logistic	0.794 (0.763-0.826)	0.717 (0.694-0.740)	0.794 (0.777-0.810)	0.842 (0.813-0.871)	0.624 (0.595-0.654)	0.887 (0.882-0.892)	0.538 (0.508-0.568)	0.863 (0.850-0.877)	0.44 (0.421-0.459)
LightGBM	1.000 (1.000-1.000)	0.909 (0.887-0.931)	0.999 (0.998-0.999)	0.998 (0.998-0.999)	1.0 (1.000-1.000)	1.0 (1.000-1.000)	0.994 (0.993-0.996)	0.999 (0.999-0.999)	0.996 (0.996-0.997)
RandomForest	1.000 (1.000-1.000)	0.625 (0.599-0.651)	0.998 (0.998-0.999)	0.998 (0.997-0.999)	1.0 (0.999-1.000)	1.0 (1.000-1.000)	0.993 (0.990-0.995)	0.999 (0.999-0.999)	0.995 (0.994-0.996)
DecisionTree	1.000 (1.000-1.000)	1.0 (1.000-1.000)	0.999 (0.998-0.999)	0.998 (0.998-0.999)	1.0 (1.000-1.000)	1.0 (1.000-1.000)	0.994 (0.993-0.996)	0.999 (0.999-0.999)	0.996 (0.996-0.997)
GBDT	0.985 (0.979-0.991)	0.717 (0.693-0.740)	0.947 (0.938-0.957)	0.954 (0.939-0.969)	0.924 (0.911-0.936)	0.978 (0.974-0.981)	0.856 (0.820-0.893)	0.966 (0.959-0.972)	0.853 (0.830-0.877)
KNN	0.910 (0.895-0.925)	0.8 (0.800-0.800)	0.832 (0.830-0.834)	0.834 (0.832-0.836)	0.826 (0.820-0.832)	0.944 (0.942-0.945)	0.587 (0.584-0.590)	0.885 (0.884-0.886)	0.576 (0.572-0.580)
GNB	0.765 (0.731-0.798)	0.428 (0.336-0.521)	0.805 (0.803-0.807)	0.868 (0.864-0.872)	0.585 (0.575-0.596)	0.88 (0.877-0.882)	0.559 (0.554-0.564)	0.874 (0.872-0.875)	0.446 (0.439-0.452)
SVM	0.781 (0.749-0.814)	0.813 (0.805-0.821)	0.762 (0.755-0.768)	0.777 (0.764-0.790)	0.707 (0.693-0.722)	0.903 (0.900-0.906)	0.477 (0.467-0.487)	0.835 (0.829-0.841)	0.413 (0.407-0.420)

AUC, area under the receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval.

Table 3. The performances of 9 machine learning algorithms in the validation cohort

Algorithms	AUC (95% CI)	Cutoff (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score (95% CI)	Kappa (95% CI)
XGBoost	0.937 (0.887-0.984)	0.903 (0.871-0.936)	0.909 (0.895-0.923)	0.936 (0.922-0.950)	0.817 (0.752-0.882)	0.948 (0.930-0.965)	0.788 (0.753-0.823)	0.941 (0.932-0.950)	0.739 (0.695-0.784)
Logistic	0.783 (0.688-0.878)	0.717 (0.694-0.740)	0.779 (0.745-0.813)	0.832 (0.793-0.872)	0.591 (0.533-0.650)	0.877 (0.860-0.893)	0.515 (0.452-0.577)	0.853 (0.828-0.878)	0.402 (0.328-0.476)
LightGBM	0.935 (0.886-0.983)	0.909 (0.887-0.931)	0.907 (0.898-0.916)	0.93 (0.916-0.945)	0.826 (0.774-0.878)	0.95 (0.936-0.964)	0.778 (0.746-0.809)	0.94 (0.934-0.945)	0.737 (0.710-0.764)
RandomForest	0.934 (0.885-0.982)	0.625 (0.599-0.651)	0.906 (0.896-0.915)	0.966 (0.955-0.976)	0.697 (0.658-0.736)	0.918 (0.908-0.928)	0.857 (0.823-0.890)	0.941 (0.935-0.947)	0.708 (0.677-0.738)
DecisionTree	0.824 (0.741-0.907)	1.0 (1.000-1.000)	0.895 (0.885-0.905)	0.952 (0.939-0.965)	0.696 (0.647-0.745)	0.917 (0.905-0.929)	0.811 (0.773-0.849)	0.934 (0.927-0.940)	0.68 (0.645-0.715)
GBDT	0.893 (0.830-0.955)	0.717 (0.693-0.740)	0.858 (0.838-0.878)	0.897 (0.869-0.925)	0.725 (0.662-0.788)	0.92 (0.904-0.936)	0.68 (0.626-0.734)	0.907 (0.893-0.922)	0.604 (0.554-0.654)
KNN	0.758 (0.665-0.851)	0.8 (0.800-0.800)	0.73 (0.705-0.756)	0.767 (0.731-0.803)	0.601 (0.563-0.640)	0.871 (0.860-0.881)	0.431 (0.400-0.463)	0.815 (0.793-0.836)	0.323 (0.284-0.362)
GNB	0.754 (0.651-0.856)	0.428 (0.336-0.521)	0.801 (0.782-0.820)	0.869 (0.843-0.894)	0.563 (0.493-0.632)	0.875 (0.860-0.891)	0.556 (0.512-0.599)	0.871 (0.858-0.885)	0.427 (0.372-0.482)
SVM	0.765 (0.664-0.866)	0.813 (0.805-0.821)	0.742 (0.717-0.766)	0.756 (0.721-0.792)	0.689 (0.638-0.740)	0.896 (0.882-0.909)	0.451 (0.422-0.481)	0.819 (0.797-0.841)	0.375 (0.333-0.418)

AUC, area under the receiver operating characteristic; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval.

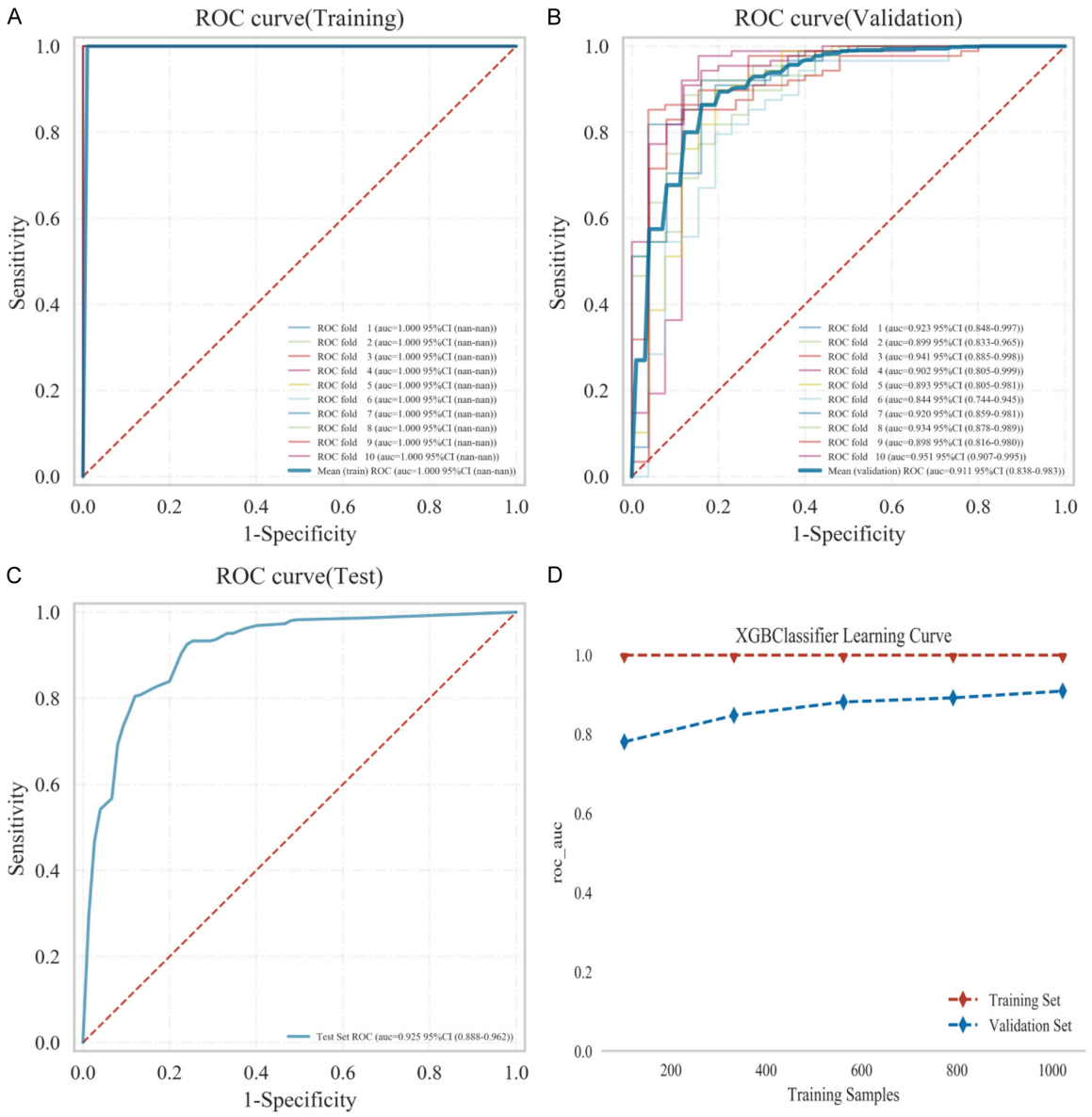

Figure 4. ROC curves of the XGBoost model. A. Training set; B. Validation set; C. Test set; D. Learning curve.

XGBoost model, suggesting its potential value as an auxiliary diagnostic tool for liver cancer.

Model clinical value and interpretability

To evaluate the calibration performance of the model, calibration curves were generated for both the training and validation cohorts. As illustrated in **Figure 5A**, the XGBoost model demonstrated strong concordance between the validated predicted values and the true values. Furthermore, the DCA (**Figure 5B**) showed that the XGBoost model had a consistent net clinical benefit at thresholds ranging from 5% to 100%.

To enhance the interpretability of the model, SHAP values were employed to visually illustrate the contribution of the screened crucial variables to the ML model. Based on the mean absolute SHAP values, we ranked the importance of the clinical features and showed (**Figure 5C**) that CA199 emerged as the most influential predictor in assessing the risk of liver cancer. Other important features included ALP, HBsAb, TG, IBIL, FBG, HBcAb, BUA, Age, Mono/WBC, SG, TP, HBeAb, and SED_EC. Among them, higher values of CA199, ALP, IBIL, FBG, BUA, Age, and Mono/WBC were associated with a higher probability of being diagnosed with liver cancer (**Figure 5D**).

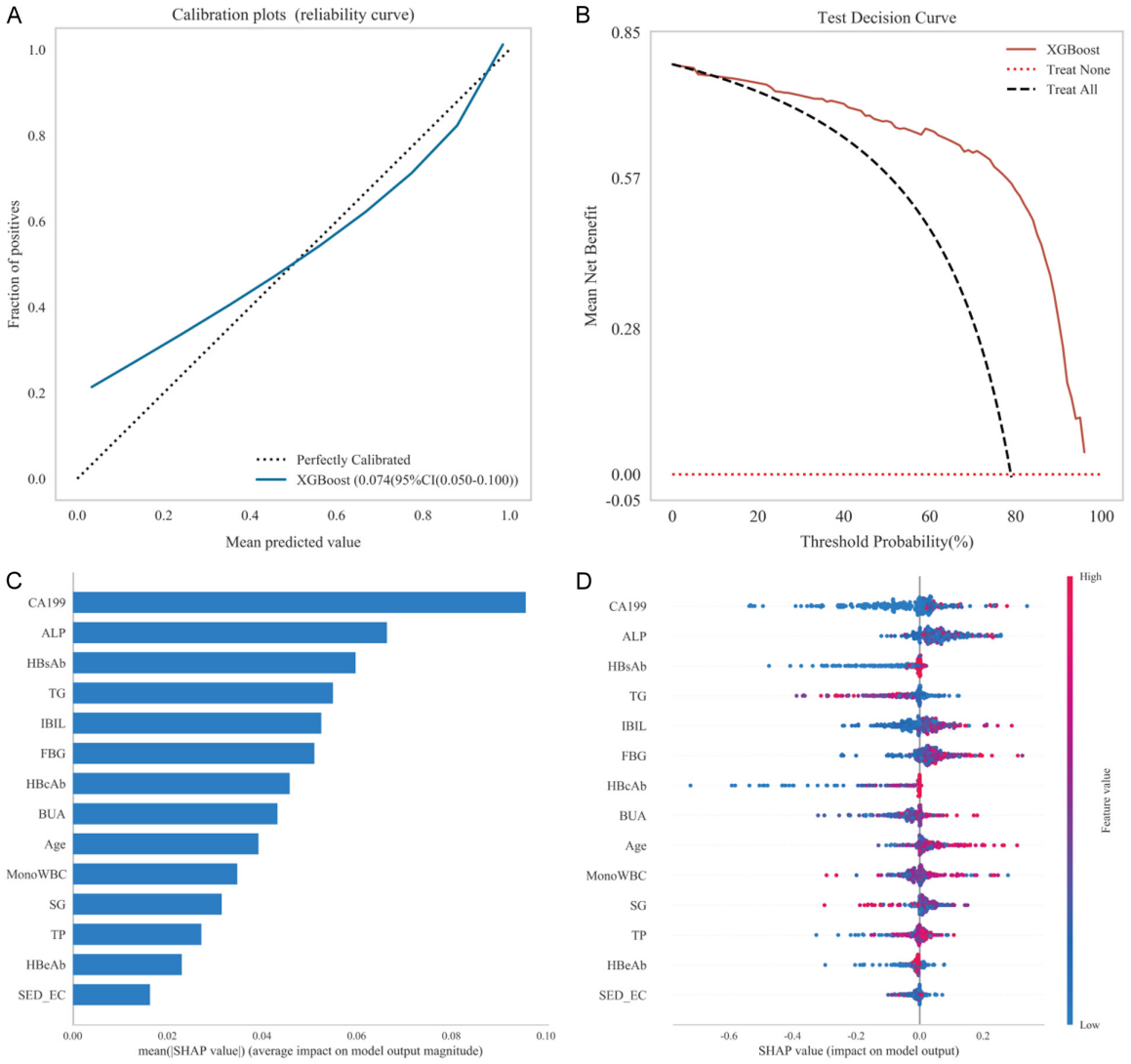

Figure 5. Clinical value and interpretability of the XGBoost model. A. Calibration curve; B. Decision curve analysis; C. Global feature importance ranking based on mean absolute SHAP values. Bar graphs show the average impact of each feature on the model output magnitude; D. Detailed view of the impact of each feature on the model output for a single sample. Scatter plots illustrate the SHAP values for each feature, with blue indicating low feature values and red indicating high feature values.

Online deployment and visualization of the diagnostic model

To improve the accessibility of the diagnostic model for clinicians, we integrated the final XGBoost model into an easy-to-use online diagnostic tool. After entering the values of the selected key variables, users can obtain the estimated probability of being diagnosed with liver cancer. For Patient 1, the values of HBcAb, HBsAb, HBeAb, FBG, CA199, TG, BUA, ALP, IBIL, TP, Mono/WBC, SED_EC, SG and Age were 8.560, 1.39, 3.700, 3.78 mmol/L, 10.01 U/mL, 0.79 mmol/L, 467.0 μmol/L, 61.0 U/L, 9.2 μmol/L, 76.30 g/L, 5.5%, 5 cells/μL, 1.034 and 62 years old, respectively (**Figure 6A**). The model estimated a 99.6% probability of being diagnosed with liver cancer, exceeding the predefined diagnostic threshold of 91.1%, indicating a high probability of liver cancer diagnosis. This result was consistent with the final clinical diagnosis. The SHAP force plot visualized the contributions of individual variables to this prediction (**Figure 6B**).

For Patient 2, the corresponding values were 6.48, 0, 0.1, 5.31 mmol/L, 57.65 U/mL, 0.6 mmol/L, 254 μmol/L, 409 U/L, 1.4 μmol/L, 74

Figure 6. Schematic diagram of the interface of the online prediction system. A, B. Liver cancer group; C, D. Control group.

g/L, 3.8%, 2 cells/μL, 1.024 and 62 years old (**Figure 6C**). The model estimated a 5.5% probability of being diagnosed with liver cancer, which was below the diagnostic threshold, indicating a low probability of liver cancer diagnosis, consistent with the final clinical diagnosis. The SHAP force plot again illustrated the individual contributions to the prediction (**Figure 6D**). This feature may assist clinicians in individualized diagnostic assessment and provide patients with a more intuitive understanding of the model output.

Discussion

Liver cancer is one of the leading cause of cancer death around the world, and treatment is difficult, so early detection is the main way that can help improve the prognosis [18, 19]. The 5-year age-standardized relative survival rate of liver cancer in China is the lowest among all cancers, at 10.1% [20], and most cases are diagnosed at an intermediate to advanced stage due in large part to the absence of useful early screening methods. Thus, the objective of the present study was to develop an interpretable machine learning-based auxiliary diagnostic model for liver cancer using retrospectively collected clinical examination data.

Although univariate analysis revealed multiple significant differences among variables, ML models captured complex nonlinear and combinatorial interactions that traditional statistical methods could not. Among these, XGBoost demonstrated favorable diagnostic classification performance, achieving an AUC of 1.000 in the training cohort compared with 0.794 for LR, thereby suggesting the added value of ML modeling beyond conventional approaches. Furthermore, XGBoost performed well in predicting both the training and test sets, suggesting that it has strong stability and generalization ability. In contrast, although LightGBM, RF and Decision Tree models performed well in the training set, their performance decreased in the validation set with a certain degree of overfitting. Therefore, XGBoost is identified as the optimal diagnostic model, and its stability and practicality are further confirmed in the internal validation. The above results indicate that XGBoost can effectively distinguish between liver cancer patients and normal populations, providing strong support for clinical decision-making [17].

This study utilized the LASSO regression method to identify clinical parameters highly associated with the occurrence of liver cancer, including Age, SG, SED_EC, Mono/WBC, TP, IBIL, ALP, BUA, TG, CA199, FBG, HBeAb, HBsAb, and HBcAb. Correlation analysis further revealed the potential relationships among them, and a significant positive correlation was found between age and CA199 levels, suggesting that CA199 may increase with age, which is in line with the results of previous studies, indicating that age is one of the important risk factors for liver cancer development [21]. Simultaneously, age was significantly and negatively correlated with HBsAb level, which may reflect the phenomenon of lower HBsAb level in individuals of advanced age [22]. In addition, there was a positive correlation between Mono/WBC and TG, HBsAb and HBcAb, suggesting that these indicators may change synergistically during the pathophysiological process of liver cancer.

In the feature importance analysis, CA199 had the highest mean absolute SHAP value, indicating that it is the most critical feature for predicting liver cancer. Other important features included ALP, HBsAb, TG, etc. This result provides an intuitive basis for clinicians to understand the decision-making process of the model and helps them to better apply the model for liver cancer diagnosis and risk assessment. It also provides a direction for further research on the pathogenesis of liver cancer and the search for new biomarkers [15]. Furthermore, the analysis of the significance of these features further revealed the potential mechanisms of liver cancer development. For example, IBIL showed a significant positive correlation with HBsAb and HBcAb, suggesting that HBsAb and HBcAb may be involved in hepatocellular carcinoma by influencing the metabolism of bilirubin [22]. The positive correlation between ALP and CA199 also suggests that there may be a potential link between the two, which is in line with the view that ALP serves as a predictive marker for liver cancer in other studies [23].

The majority of HCC originates from cirrhosis, and the main causes of cirrhosis include nonalcoholic fatty liver disease, alcohol-related liver disease, or hepatitis B virus (HBV) or hepatitis C virus (HCV) infection [21]. In this study, we found significant correlations between

HBeAb, HBsAb, and HBcAb, especially between HBsAb and HBcAb, reflecting the complexity of HBV infection and its impact on hepatocarcinogenesis, which is in line with previous studies that indicated a close association between HBV infection and hepatocarcinogenesis [22, 24].

Although the results are encouraging, several limitations should be acknowledged. This was a single-center retrospective study that had a relatively limited sample size from a single center which may be subject to selection bias. Future research can increase the number of participants and carry out a multicenter prospective study in order to enhance the generalizability and reliability of the results. The control group consisted of healthy individuals as well as people with chronic hepatitis B, chronic hepatitis C, or cirrhosis. Though it is possible that this design will represent a clinically typical differential diagnostic situation, these subgroups might vary considerably in terms of liver function, inflammatory status, viral serological profiles and tumor marker levels. Since subgroup-specific modeling and stratified performance analyses were not performed in the current study, the diagnostic capability of the model in any of the control subgroups cannot be evaluated. Subsequent investigations ought to separately confirm the model in healthy controls, patients with chronic hepatitis, patients with cirrhosis, and other high-risk groups. Finally, the common clinical characteristics and serum markers were only considered in this study, whereas other potential biomarkers were not taken into account. Other biomarkers could be investigated in further studies to maximize model performance.

In conclusion, the machine learning-based diagnostic model developed in this study showed potential clinical value as an auxiliary tool for liver cancer diagnosis. However, further research is needed to validate and refine it in the actual clinical promotion so as to achieve a wider clinical application value.

Conclusions

This study conducted a retrospective analysis based on multidimensional clinical data, constructed and evaluated nine ML models, and identified XGBoost as the optimal algorithmic model through comprehensive analysis. In addition, correlation analysis and SHAP-based visualization were employed to highlight key clinical indicators strongly associated with the onset of hepatocarcinogenesis, which provided strong support for the interpretability of the model. These findings may improve the auxiliary diagnostic identification of liver cancer and provide a more reliable basis for individualized diagnostic assessment.

Acknowledgements

We sincerely thank all participating patients and the clinical personnel for their essential support and contributions to this study. This work was supported by the Science Foundation for Distinguished Young Scholars of Guangxi [Grant Number 2023GXNSFFA026003, F. W.], the Central Guidance for Local Science and Technology Development Funds Project [Grant Number ZY24212012, F. W.], the National Natural Science Foundation of China (NSFC) [Grant Numbers 82372828 and 82573878, F. W.], the Science Foundation for Distinguished Young Scholars of Guangxi Medical University [F. W.], and the Yongjiang Program of Nanning [Grant Number 2021015, F. W.].

Written informed consent was obtained from all participants prior to their enrollment.

Disclosure of conflict of interest

None.

Abbreviations

XGBoost, EXtreme Gradient Boosting; LR, Logistic Regression; LightGBM, Light Gradient Boosting Machine; RF, Random Forest; GBDT, Gradient Boosting Decision Tree; KNN, k-Nearest Neighbor; GNB, Gaussian Naive Bayes; SVM, Support Vector Machine.

Address correspondence to: Junyi Chen, Department of Urology, The Second Affiliated Hospital of Fujian Medical University, No. 34, Zhongshanbei Road, Licheng District, Quanzhou 362000, Fujian, China. Tel: +86-18259576969; E-mail: chenjunyidoctor@163.com; Fubo Wang, Center for Genomic and Personalized Medicine, Guangxi Key Laboratory for Genomic and Personalized Medicine, University Engineering Research Center of Digital Medicine and Healthcare, Guangxi Medical University, No. 22, Shuangyong Road, Qingxiu District, Nanning

530021, Guangxi, China. Tel: +86-0771-5358562; E-mail: wangfubo@gxmu.edu.cn; Leifeng Liang, Department of Oncology, The Sixth Affiliated Hospital of Guangxi Medical University, The First People's Hospital of Yulin, 495 Education Middle Road, Yulin 537000, Guangxi, China. Tel: +86-13657757927; E-mail: liangleifeng@stu.gxmu.edu.cn

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I and Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024; 74: 229-263.

[2] Rebouissou S and Nault JC. Advances in molecular classification and precision oncology in hepatocellular carcinoma. J Hepatol 2020; 72: 215-229.

[3] El-Serag HB and Kanwal F. Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? Hepatology 2014; 60: 1767-1775.

[4] Gao S, Gang J, Yu M, Xin G and Tan H. Computational analysis for identification of early diagnostic biomarkers and prognostic biomarkers of liver cancer based on GEO and TCGA databases and studies on pathways and biological functions affecting the survival time of liver cancer. BMC cancer 2021; 21: 791.

[5] Charrière B, Maulat C, Suc B and Muscari F. Contribution of alpha-fetoprotein in liver transplantation for hepatocellular carcinoma. World J Hepatol 2016; 8: 881-890.

[6] Wong RJ, Ahmed A and Gish RG. Elevated alpha-fetoprotein: differential diagnosis-hepatocellular carcinoma and other disorders. Clin Liver Dis 2015; 19: 309-323.

[7] Galle PR, Forner A, Llovet JM, Mazzaferro V, Piscaglia F, Raoul JL, Schirmacher P and Vilgrain V. EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2018; 69: 182-236.

[8] Barrea C, Ovaert C, Moniotte S, Biard JM, Steenhaut P and Bernard P. Prenatal diagnosis of abnormal cardinal systemic venous return without other heart defects: a case series. Prenat Diagn 2011; 31: 380-388.

[9] Lalla R, Mulherkar RV and Misar PV. Incomplete peripheral facial nerve palsy and ulnar neuropathy due to leprosy mistaken as faciobrachial stroke. BMJ Case Rep 2015; 2015: bcr2015210060.

[10] Zou Y, Shi Y, Liu J, Cui G, Yang Z, Liu M and Sun F. A comparative analysis of six machine learning models based on ultrasound to distinguish the possibility of central cervical lymph node metastasis in patients with papillary thyroid carcinoma. Front Oncol 2021; 11: 656127.

[11] Waljee AK and Higgins PD. Machine learning in medicine: a primer for physicians. Am J Gastroenterol 2010; 105: 1224-1226.

[12] Zhu J, Zheng J, Li L, Huang R, Ren H, Wang D, Dai Z and Su X. Application of machine learning algorithms to predict central lymph node metastasis in T1-T2, non-invasive, and clinically node negative papillary thyroid carcinoma. Front Med (Lausanne) 2021; 8: 635771.

[13] Wu Y, Rao K, Liu J, Han C, Gong L, Chong Y, Liu Z and Xu X. Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer. Front Endocrinol (Lausanne) 2020; 11: 577537.

[14] Singal AG, Mukherjee A, Elmunzer JB, Higgins PD, Lok AS, Zhu J, Marrero JA and Waljee AK. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol 2013; 108: 1723-1730.

[15] Capobianco E. High-dimensional role of AI and machine learning in cancer research. Br J Cancer 2022; 126: 523-532.

[16] Zhang M, Zheng Y, Maidaiti X, Liang B, Wei Y and Sun F. Integrating machine learning into statistical methods in disease risk prediction modeling: a systematic review. Health Data Sci 2024; 4: 0165.

[17] Martinino A, Aloulou M, Chatterjee S, Scarano Pereira JP, Singhal S, Patel T, Kirchgesner TP, Agnes S, Annunziata S, Treglia G and Giovinazzo F. Artificial intelligence in the diagnosis of hepatocellular carcinoma: a systematic review. J Clin Med 2022; 11: 6368.

[18] Johnson P, Zhou Q, Dao DY and Lo YMD. Circulating biomarkers in the diagnosis and management of hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 2022; 19: 670-681.

[19] Wang Z, Qin H, Liu S, Sheng J and Zhang X. Precision diagnosis of hepatocellular carcinoma. Chin Med J (Engl) 2023; 136: 1155-1165.

[20] Zeng H, Zheng R, Guo Y, Zhang S, Zou X, Wang N, Zhang L, Tang J, Chen J, Wei K, Huang S, Wang J, Yu L, Zhao D, Song G, Chen J, Shen Y, Yang X, Gu X, Jin F, Li Q, Li Y, Ge H, Zhu F, Dong J, Guo G, Wu M, Du L, Sun X, He Y, Coleman MP, Baade P, Chen W and Yu XQ. Cancer survival in China, 2003-2005: a population-based study. Int J Cancer 2015; 136: 1921-1930.

[21] Calderaro J, Seraphin TP, Luedde T and Simon TG. Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma. J Hepatol 2022; 76: 1348-1361.

[22] Jin ZZ, Jin FF, Liu X, Liu N, Wen F and Lou JL. Coexistence of low levels of HBsAg and high

levels of anti-HBs may increase risk of hepatocellular carcinoma in chronic hepatitis B patients with high HBV load. Braz J Infect Dis 2019; 23: 343-351.

[23] Li Z, Yuan L, Zhang C, Sun J, Wang Z, Wang Y, Hao X, Gao F and Jiang X. A novel prognostic scoring system of intrahepatic cholangiocarcinoma with machine learning basing on real-world data. Front Oncol 2021; 10: 576901.

[24] Lin YJ, Lee MH, Yang HI, Jen CL, You SL, Wang LY, Lu SN, Liu J and Chen CJ. Predictability of liver-related seromarkers for the risk of hepatocellular carcinoma in chronic hepatitis B patients. PLoS One 2013; 8: e61448.

Table S1. Baseline Variables (*P*>0.05)

Variables	Overall (n=3,629)	Control (n=914)	Liver cancer (n=2,715)	*p*
U_PRO[1]				0.319
Negative	128 (19.31)	30 (24.00)	98 (18.22)	
+	421 (63.50)	69 (55.20)	352 (65.43)	
++	91 (13.73)	21 (16.80)	70 (13.01)	
+++	15 (2.26)	3 (2.40)	12 (2.23)	
++++	8 (1.21)	2 (1.60)	6 (1.12)	
U_GLU[1]				0.191
Negative	155 (33.92)	37 (29.60)	118 (35.54)	
+	95 (20.79)	20 (16.00)	75 (22.59)	
++	41 (8.97)	13 (10.40)	28 (8.43)	
+++	70 (15.32)	24 (19.20)	46 (13.86)	
++++	96 (21.01)	31 (24.80)	65 (19.58)	
KET[1]				0.14
Negative	156 (51.32)	39 (50.65)	117 (51.54)	
+	112 (36.84)	27 (35.06)	85 (37.44)	
++	28 (9.21)	11 (14.29)	17 (7.49)	
+++	8 (2.63)	0 (0.00)	8 (3.52)	
BLD[1]				0.426
Negative	129 (21.04)	34 (18.58)	95 (22.09)	
+	239 (38.99)	80 (43.72)	159 (36.98)	
++	115 (18.76)	34 (18.58)	81 (18.84)	
+++	130 (21.21)	35 (19.13)	95 (22.09)	
Stool_color[1]				0.629
Negative	2889 (98.03)	717 (97.82)	2172 (98.10)	
Positive	58 (1.97)	16 (2.18)	42 (1.90)	
Weight[2]	59.00 [52.00, 66.00]	58.50 [50.00, 67.00]	59.00 [52.00, 66.00]	0.382
PH[2]	6.00 [5.50, 6.50]	6.00 [5.50, 6.50]	6.00 [5.50, 6.50]	0.335
USCC[2]	0.00 [0.00, 0.40]	0.00 [0.00, 0.50]	0.00 [0.00, 0.40]	0.11
USSRCC[2]	1.30 [0.50, 3.30]	1.30 [0.40, 3.00]	1.30 [0.50, 3.50]	0.588
USYCC[2]	0.10 [0.00, 1.00]	0.00 [0.00, 1.00]	0.10 [0.00, 1.00]	0.272
USMFC[2]	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.837
SED_casts[2]	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.22
WBC[2]	7.64 [5.62, 11.06]	7.64 [5.53, 12.19]	7.66 [5.64, 10.82]	0.168
Neut/WBC[2]	72.50 [62.50, 83.30]	72.30 [60.10, 85.90]	72.50 [64.00, 82.30]	0.458
Neut[2]	5.45 [3.54, 8.89]	5.23 [3.32, 10.09]	5.48 [3.67, 8.61]	0.934
NRBC[2]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.196
NRBC_[2]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.186
RBC[2]	4.20 [3.65, 4.73]	4.26 [3.73, 4.75]	4.18 [3.64, 4.72]	0.06
HGB[2]	121.00 [105.00, 136.00]	121.00 [107.00, 134.00]	120.00 [104.00, 136.00]	0.988
HCT[2]	0.37 [0.32, 0.41]	0.37 [0.33, 0.40]	0.37 [0.32, 0.41]	0.824
MCHC[2]	330.00 [319.00, 340.00]	330.00 [319.00, 338.00]	330.00 [320.00, 340.00]	0.064
PCT[2]	0.22 [0.15, 0.29]	0.23 [0.17, 0.28]	0.21 [0.15, 0.29]	0.072
ALT[2]	38.00 [20.00, 83.00]	38.00 [17.00, 111.00]	39.00 [21.00, 76.00]	0.523
TCHO[2]	4.12 [3.44, 4.94]	4.27 [3.52, 5.01]	4.09 [3.41, 4.90]	0.057
LDL-C[2]	2.58 [1.97, 3.27]	2.64 [1.98, 3.31]	2.56 [1.97, 3.26]	0.199
FDP[2]	5.50 [1.80, 16.60]	5.40 [1.60, 17.90]	5.60 [2.00, 16.00]	0.78

[1]Categorical variables as frequency (percentage); [2]Continuous variables as median [IQR]. U_PRO, urine protein; U_GLU, urine glucose; KET, ketone body; BLD, urine occult blood; USCC, urinary sediment crystal count; USSRCC, urinary sediment small round epithelial cell count; USYCC, urinary sediment yeast-like cell count; USMFC, urinary sediment mucus filament count; SED_casts, urine sediment microscopy for casts; WBC, white blood cell count; Neut/WBC, neutrophil percentage; Neut, neutrophil; NRBC, nucleated red blood cell; NRBC_, nucleated red blood cell percentage; RBC, red blood cell; HGB, hemoglobin; HCT, hematocrit; MCHC, mean corpuscular hemoglobin concentration; PCT, plateletocrit; ALT, alanine aminotransferase; TCHO, total cholesterol; LDL-C, low-density lipoprotein cholesterol; FDP, fibrinogen and fibrin degradation products.