Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive Value of Nutritional and Inflammatory Factors for Liver Fibrosis Stage Classification Among HBV-Exposed Individuals

An Zhao
The First Affiliated Hospital of Zhengzhou University
Minghao Li
The First Affiliated Hospital of Zhengzhou University
Teng Li
The First Affiliated Hospital of Zhengzhou University
Huapeng Zhang
The First Affiliated Hospital of Zhengzhou University
Xiaoyi Shi
The First Affiliated Hospital of Zhengzhou University
Yuting He
The First Affiliated Hospital of Zhengzhou University
Wenzhi Guo
The First Affiliated Hospital of Zhengzhou University
Yunfeng Zhao
The People's Hospital of Zhengzhou
Peihao Wen
wenpeihao1982@163.com

The First Affiliated Hospital of Zhengzhou University

Article

Keywords: NHANES, Machine learning, Liver fibrosis, HBV

Posted Date: August 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10407842/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Liver fibrosis is a major contributor to liver-related morbidity and mortality among individuals with previous or current hepatitis B virus (HBV) exposure. Nutritional and inflammatory factors have been implicated in liver fibrosis, however their combined value for fibrosis stage classification remains insufficiently explored. This study aimed to investigate the associations of nutritional and inflammatory indicators with liver fibrosis severity and to develop machine-learning models for fibrosis stage classification.

Methods

This cross-sectional study analyzed data from the National Health and Nutrition Examination Survey (NHANES) 2017–2018 and 2019–2020 cycles. Individuals with evidence of previous or current HBV exposure (anti-HBc positive) were included. Liver fibrosis severity was assessed using transient elastography-derived liver stiffness measurements. Associations between clinical variables and fibrosis severity were evaluated using correlation analysis, univariate regression, and restricted cubic spline models. Random forest-based feature selection was performed, and six machine-learning algorithms, including Random Forest Support Vector Machine, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor, Decision Tree, and Neural Network, were developed and evaluated using repeated 10-fold cross-validation. Model performance was assessed using accuracy, multi area of under curve, F1-score, calibration analyses and .

Results

A total of 911 participants were included, comprising 769 individuals without fibrosis (84.41%), 39 with F1 fibrosis (4.28%), 52 with F2 fibrosis (5.71%), and 51 with F3 fibrosis (6.00%). Correlation and regression analyses demonstrated significant associations between liver fibrosis severity and multiple nutritional and inflammatory indicators. Restricted cubic spline analyses further revealed nonlinear relationships between fibrosis severity and age, BMI, GNRI, NPAR, and SIRI. Feature selection identified 10 key variables, including GNRI, BMI, albumin, total cholesterol, CRP, NPAR, creatinine, HDL cholesterol, waist-to-height ratio, and age. Among the six machine-learning algorithms evaluated, XGBoost achieved the best overall performance, with an accuracy of 0.940, a MAUC of 0.940, and a macro-F1 score of 0.804.

Conclusions

Nutritional and inflammatory indicators were significantly associated with liver fibrosis severity among individuals with previous or current HBV exposure. An XGBoost model constructed from routinely available clinical variables showed good performance in fibrosis stage classification and warrants further validation in independent cohorts.

1. Introduction

Liver fibrosis is a pathological wound-healing response that involves the excessive accumulation of extracellular matrix (ECM) proteins due to chronic liver injury[1]. This progressive hepatic scarring is driven by sustained inflammatory responses[2] and hepatocyte injury[3], leading to the replacement of healthy liver parenchyma with fibrotic tissue. Liver fibrosis and cirrhosis significantly impact not only individual health but also impose a considerable burden on healthcare systems and society. According to recent estimates from the Global Burden of Disease study, liver fibrosis and cirrhosis contribute to approximately 1.3 million deaths globally each year, accounting for a large proportion of liver-related morbidity and mortality[4]. Despite the high prevalence and severity potential, liver fibrosis often progresses silently in its early stages, frequently remaining asymptomatic and causing minimal functional impairment. However, as fibrosis advances, it can disrupt the liver's normal histoarchitecture and compromise its regenerative capacity[5]. Without intervention, liver fibrosis may progress to cirrhosis, characterized by extensive scarring, nodular regeneration, and functional loss of the liver[6]. Notably, the progression from fibrosis to cirrhosis is not inevitable and can be modified by various factors[7–9], including the etiology of liver injury, host characteristics, and environmental influences. Among these, viral hepatitis—especially hepatitis B (HBV)—is a major cause of liver fibrosis worldwide[10].

HBV infection is a major global health concern, with the World Health Organization (WHO) estimating that approximately 296 million individuals are living with chronic HBV infection worldwide[11]. Chronic HBV infection involves the persistent presence of the virus in the liver, leading to continuous immune-mediated damage and liver inflammation, and this process is marked by repeated cycles of hepatocyte injury, regeneration, and fibrosis. The immune response to HBV can control viral replication but also triggers chronic inflammation and hepatocyte apoptosis. This necroinflammatory activity activates hepatic stellate cells (HSCs), the main fibrogenic cells in the liver, which secrete excessive extracellular matrix proteins, leading to fibrosis[12]. Over time, the accumulation of fibrotic tissue disrupts liver architecture, forming regenerative nodules and fibrous septa, key features of cirrhosis. These pathogenic mechanisms highlight the critical importance of monitoring fibrosis progression in HBV-infected individuals, as unchecked fibrosis can eventually lead to cirrhosis, a stage with far more severe clinical consequences and challenges in management.

The progression from liver fibrosis to cirrhosis marks a pivotal stage, as cirrhosis is linked to severe clinical consequences, including an elevated risk of hepatocellular carcinoma, liver failure, and other liver-related complications. Once cirrhosis is established, the liver's regenerative ability is markedly diminished, often leading to portal hypertension, hepatic dysfunction, and potentially hepatic

decompensation[13]. Detecting liver fibrosis at an early stage remains a considerable challenge, complicating timely intervention. A primary issue in managing HBV infection-related cirrhosis is that early-stage liver fibrosis is frequently asymptomatic, delaying diagnosis until significant liver injury has already occurred. Although antiviral therapies can effectively suppress HBV replication and decrease fibrosis progression, they are not curative and do not fully reverse established fibrosis or cirrhosis. Accurate assessment of fibrosis and cirrhosis severity is another clinical challenge. While liver biopsy is the gold standard for fibrosis staging, its invasiveness and associated risks limit its practicality in routine clinical settings[14]. Non-invasive diagnostic methods, including serum biomarkers and elastography-based imaging, have become valuable tools for monitoring fibrosis in HBV patients, but these methods may have limitations in precisely distinguishing advanced fibrosis from early cirrhosis[15].

In recent years, machine learning (ML) has emerged as a powerful tool in the field of medicine, particularly in the development of clinical prediction models[16]. Unlike traditional statistical methods, which often rely on predefined assumptions and linear relationships, machine learning techniques can automatically identify complex patterns and interactions within large datasets, making them well-suited for predicting clinical outcomes in heterogeneous diseases[17]. These algorithms, including decision trees, support vector machines (SVM), random forests (RF), and neural networks, have shown promise in improving diagnostic accuracy, risk stratification, and personalized treatment strategies for patients. Despite these advances, significant gaps remain in our research, especially in comprehensive fibrosis stage prediction studies. A study has shown that machine learning-based predictive tools outperform traditional non-invasive assessment methods[18]. However, most previous studies have focused on the construction of liver fibrosis in HBV patients based on serum data, which ignores the impact of multi-dimensional data on liver fibrosis and easily leads to prediction bias[19–20]. In addition, some studies only focus on the diagnosis of liver cirrhosis, while ignoring the prediction of the progression of liver fibrosis in various stages[21–22]. Therefore, the aim of this study is to develop a robust and clinically applicable prediction model for HBV infection-related cirrhosis, leveraging advanced machine learning techniques to improve the early detection and risk assessment of liver fibrosis progression.

2. Methods

2.1 study population

National Health and Nutrition Examination Survey (NHANES) is a nationally representative cross-sectional survey conducted by the National Center for Health Statistics (NCHS), a division of the US Centers for Disease Control and Prevention (CDC) [23]. The present study extracted data from the 2017–2018 and 2019–2020 NHANES cycles because transient elastography-derived liver stiffness measurements and all nutritional and inflammatory variables required for fibrosis staging and machine-learning model development were simultaneously available only in these survey cycles[24].

Participants without evidence of previous or current HBV exposure (anti-HBc negative) were excluded. Among HBV-exposed individuals, those lacking transient elastography-derived liver stiffness

measurements were removed. Participants with incomplete data for variables included in fibrosis classification and machine-learning analyses were subsequently excluded, and complete-case analysis was performed without data imputation. The detailed participant selection process is shown in Fig. 1. All participants in the final analytic sample were adults (aged ≥ 18 years).

2.2 Variables

2.2.1 Measurement of HBV and liver fibrosis

The VITROS hepatitis B core antibody (anti-HBc) assay used in NHANES was employed to identify individuals with evidence of previous or current HBV exposure. Participants were classified as HBV-exposed (anti-HBc positive) or non-exposed (anti-HBc negative) according to the test results. Transient elastography (TE) is a widely used noninvasive technique for the assessment of liver fibrosis[25]. In this study, liver stiffness measurement (LSM) obtained by TE was used to evaluate fibrosis severity[26]. Fibrosis stages were classified according to LSM values, with cutoff thresholds of 7.9, 8.8, and 11.7 kPa corresponding to fibrosis stages ≥F1, ≥F2, and F3, respectively, as previously reported[27].

2.2.2 Covariates

During the home interviews, trained interviewers collected demographic information using Computer-Assisted Personal Interview. Race/ethnicity was classified as Mexican American, Non-Hispanic White, NonHispanic Black, Other Hispanic, and other race/multiracial. Education levels were categorized into three groups: less than High school, High school, and higher than High school. Single/separated or coupled were the two marital status categories. Participants who claimed to have smoked fewer than 100 cigarettes in their lives were labeled ‘never smokers’. Former smokers were individuals who had smoked more than 100 cigarettes in their lives but had quit, while current smokers were those who were currently smoking.

For processing of CRP, latex-enhanced nephelometry with particle-enhanced assays were used for quantitation[28]. Assays were performed on a Behring Nephelometer to determine quantitative CRP levels. The primary standard used for processing was organized by Behring Diagnostics and standardized against WHO reference material. HDL-C and TC blood levels were measured with a Cobas Chemistry Analyzer (6000, Roche CO., USA). Serum and urinary creatinine were measured by the Jaffe rate methods, and urinary albumin was measured by a fluorescent immuno-assay.

Hematologic parameters were measured following the NHANES CBC Profile using the Beckman Coulter Automated Hematology Analyzer DxH 900 (Beckman-Coulter, Brea, CA, USA), which performs red and white cell counts, hemoglobin, hematocrit, and red blood cell indices. NPAR was calculated using the same blood sample and the following formula: Neutrophil percentage (in total WBC count) (%) x 100/Albumin (g/dL)[29]. SIRI is calculated by the following formula: (monocyte count x neutrophils count) / lymphocytes count[30]. BMI was calculated as weight (kg) divided by the square of the height (meter), expressed as kg/m2. WHR was calculated as WC (cm) divided by HC (cm). The specific

calculation methods of GNRI is (1.489 x serum albumin (g/dL) + 41.7 x (present weight/ideal body weight) ideal body weight = height(cm)-100-(height(cm)-150)/(4(male), 2.5(female))[31]. The fibrosis-4 (FIB-4) index was calculated using the following formula: age (years) × AST (U/L) / [platelet count (10^9/L) × √ALT (U/L)] and was used as a conventional noninvasive fibrosis marker for comparative analyses[32].

DM was defined[33] as (1) doctor diagnosed diabetes; (2) glycohemoglobin > 6.5%; (3) fasting glucose ≥ 7.0 mmol/ L; (4) random blood glucose ≥ 11.1 mmol/ L; (5) two-hour oral glucose tolerance test blood glucose ≥ 11.1 mmol/ L; and (6) use of diabetes medication or insulin. Hypertension was defined[34] as: (1) an average systolic blood pressure/average diastolic blood pressure ≥ 140/90 mmHg, (2) previous diagnosis by a doctor or health professional, or (3) currently being treated with antihypertensive medications. We defined MetS[34] according to the new International Diabetes Federation definition as: central obesity (WC ≥ 102 cm in males and ≥ 88 cm in females; if BMI > 30, central obesity can be assumed and WC did not need to be measured), plus any two of the following four factors: (1) raised TGs ≥ 150 mg/dL (1.7 mmol/L) or specific treatment for this lipid abnormality; (2) reduced HDL-C (< 40 mg/dL (1.03 mmol/L) in males, < 50 mg/dL (1.29 mmol/L) in females) or specific treatment for this lipid abnormality; (3) raised BP (systolic BP ≥ 130 or diastolic BP ≥ 85 mm Hg) or treatment of previously diagnosed hypertension; (4) raised fasting plasma glucose (FPG) ≥ 100 mg/dL (5.6 mmol/L) or previously diagnosed type 2 diabetes.

2.3 Statistics analysis

Baseline characteristics were described using means with standard error (SE) for continuous variables and frequencies with percentages for categorical variables. Baseline characteristics and disease variables were compared employing the test for continuous variables and the Rao-Scott Chi-Square test for categorical variables, and we used the Tukey HSD test to compare nutritional status at different stages of progression. Furthermore, pairwise correlation analyses were conducted to assess the interrelationships among all variables and univariate linear regression analyses were employed to evaluate the individual impact of each variable on liver cirrhosis scores. Subsequently, this study employed restricted cubic spline (RCS) to explore the underlying nonlinear association between continuous variables and fibrosis score and nonlinearity to evaluate the likelihood ratio test.

The model development procedure consisted of four steps. First, random forest-based variable importance was used to rank candidate predictors. Variables were then sequentially added according to their importance rankings, generating 19 nested feature combinations. Each feature set was evaluated using 10-fold cross-validation, and model performance was assessed using accuracy, macro-F1 score, and multiclass area under the receiver operating characteristic curve (MAUC). The final feature set was selected based on overall predictive performance and model parsimony. Second, six machine-learning algorithms, including random forest, XGBoost, k-nearest neighbors, support vector machine, neural network, and recursive partitioning and regression trees, were trained and compared using the selected feature set. The optimal algorithm was determined according to cross-validated accuracy, macro-F1 score, and MAUC. Third, the final XGBoost model was further evaluated using confusion matrix analysis,

calibration curve analysis, and ROC analysis. To assess its clinical relevance, the model's discrimination for advanced fibrosis was compared with the conventional FIB-4 index. Finally, SHAP analysis was performed to interpret the final model by quantifying the contribution of individual predictors to each fibrosis stage classification. All statistical analyses were performed using R software version 4.3.2, and two-sided P values < 0.05 were considered statistically significant.

3. Results

3.1 Demographic

This study included 911 adults, the mean age of the group was 58.46 years, and 516 participants (56.64%) were men. Among the 911 participants, the non-fibrosis rate was 84.41% (769), F1 rate was 4.28% (39), F2 rate was 5.71% (52), F3 rate was 6.00% (51). The all basic characteristics of patients are shown in Table 1 and most variables were significantly different among four groups.

Table 1
Demographic and other variables of participants from the NHANES 2017–2020

Variable	Normal	F1	F2	F3	Statistic	P.Value
	(n = 769)	(n = 39)	(n = 52)	(n = 51)		
Age	57.72 ± 13.87	61.23 ± 9.30	63.62 ± 10.70	62.16 ± 11.23	8.83	**0.033**
Sex					7.729	**0.052**
Male	430 (55.92)	30 (76.92)	26 (50.00)	30 (58.82)		
Female	339 (44.08)	9 (23.08)	26 (50.00)	21 (41.18)		
GNRI	115.52 ± 12.10	116.99 ± 15.49	125.76 ± 21.25	123.23 ± 21.21	34.84	**< 0.001**
WtHR	0.94 ± 0.08	0.96 ± 0.07	0.96 ± 0.08	0.97 ± 0.06	6.82	**0.030**
TC	190.64 ± 40.24	181.59 ± 31.72	168.02 ± 41.31	174.14 ± 45.35	20.07	**< 0.001**
HDL	54.89 ± 15.40	53.82 ± 10.47	45.06 ± 9.80	50.01 ± 17.39	16.98	**< 0.001**
CRP	1.39 (0.68,3.27)	1.57 (1.05,3.75)	3.21 (1.93,7.88)	2.72 (0.88,7.06)	31.41	**< 0.001**
ALB	9.10 (4.60,19.60)	10.90 (5.80,15.80)	19.10 (6.70,35.60)	19.60 (6.30,70.65)	20.16	**< 0.001**
Creatinine	108 (63,157)	121 (89,167)	109.50 (51.25,162.75)	112 (71,174)	3.24	0.356
BMI	27.40 ± 5.75	28.71 ± 6.73	33.28 ± 8.95	32.71 ± 9.98	66.12	**< 0.001**
Ethnicity					29.26	**< 0.01**
Non-Hispanic White	112 (14.57)	5(12.82)	8 (15.38)	9 (17.65)		
Non-Hispanic Black	238 (30.95)	21(53.85)	23 (44.23)	24 (47.06)		

Variable	Normal	F1	F2	F3	Statistic	P.Value
	(n = 769)	(n = 39)	(n = 52)	(n = 51)		
Mexican American	25 (3.25)	2 (5.13)	2 (3.85)	2 (3.92)		
Other Hispanic	56 (7.28)	5 (12.82)	6 (11.54)	4 (7.84)		
Other Race - Including Multi-Racial	338 (43.95)	6 (15.38)	13 (25.00)	12 (23.53)		
NPAR	13.77 ± 2.80	13.71 ± 2.86	14.22 ± 2.77	14.83 ± 4.07	6.71	**0.009**
SIRI	0.90 (0.59,1.45)	0.80 (0.61,1.49)	0.93 (0.67,1.60)	1.17 (0.74,1.63)	4.83	0.184
Marital Status					4.61	0.201
Married	468 (60.86)	19 (48.72)	27 (51.92)	27 (52.94)		
Single	301 (39.14)	20 (51.28)	25 (48.08)	24 (47.06)		
Education					11.98	0.062
Less Than 9th Grade	281 (36.54)	19 (48.72)	17 (32.69)	25 (49.02)		
High School Grad/GED or Equivalent	395 (51.37)	16 (41.03)	26(50.00)	26 (50.98)		
College Graduate or above	93 (12.09)	4 (10.25)	9 (17.31)	0 (0)		
MET					40.71	**< 0.001**
No	568 (73.86)	30 (76.92)	22 (42.31)	23 (45.10)		
Yes	201(26.14)	9 (23.08)	30 (57.69)	28 (54.90)		

Variable	Normal	F1	F2	F3	Statistic	P.Value
	(n = 769)	(n = 39)	(n = 52)	(n = 51)		
Smoke					14.10	**0.007**
No	628 (81.66)	30 (76.92)	39 (75.00)	31 (60.78)		
Yes	141 (18.34)	9(23.08)	13 (25.00)	20 (39.22)		
DM					33.16	**< 0.001**
No	585 (76.07)	32 (82.05)	22 (42.31)	32 (62.75)		
Yes	184 (23.93)	7 (17.95)	30 (57.69)	19 (37.25)		
Hypertension					23.52	**< 0.001**
No	343 (44.60)	9 (23.07)	18 (34.62)	8 (15.69)		
Yes	426 (55.40)	30 (76.93)	34 (65.38)	43 (84.31)		

Figure 2 summarizes the relationships between candidate predictors and liver fibrosis. Figure 2A presents the pairwise correlation matrix, showing that most variables exhibited weak-to-moderate correlations, with no evidence of severe multicollinearity. Strong positive correlations were observed between GNRI and BMI (r = 0.91), as well as between NPAR and SIRI (r = 0.59). Figure 2B illustrates the chord diagram of the major variables associated with fibrosis grade, highlighting the interconnected contributions of nutritional, metabolic, and inflammatory factors. Figure 2C-H show the results of univariate linear regression and restricted cubic spline analyses. Age, GNRI, BMI, WtHR, NPAR, and SIRI were significantly associated with fibrosis severity, and several variables exhibited potential nonlinear associations with fibrosis risk, particularly GNRI and NPAR.

3.2 Feature selection

Random Forest (RF) was first used to rank the importance of all candidate variables according to the mean decrease in Gini impurity (Fig. 3A). Subsequently, variables were sequentially added according to their importance ranking to generate 19 feature combinations (M1-M19), and model performance was

evaluated using 10-fold cross-validation. As shown in Fig. 3B, model performance improved with the progressive inclusion of variables and reached an optimal balance at M10. Specifically, M10 achieved the highest macro-F1 score (0.52) and MAUC (0.93), while maintaining a high classification accuracy (0.89) (Fig. 3C-E). Further addition of variables provided minimal improvement in predictive performance. Therefore, the variables included in M10 were selected as the final feature set for subsequent model development, including GNRI, BMI, NPAR, ALB, TC, CRP, creatinine, WtHR, HDL, and SIRI.

3.3 Machine learning model development and selection

Based on the optimal feature set identified in the feature-selection stage (M10), six machine-learning algorithms, including XGBoost, RF, SVM, KNN, NNET, and RPart, were developed and evaluated using repeated 10-fold cross-validation. Model performance was assessed using ACC, MAUC, and macro-F1 score.

Among all evaluated algorithms, XGBoost achieved the best overall performance, with an ACC of 0.940, a MAUC of 0.940, and a macro-F1 score of 0.804. ACC reflects overall classification accuracy, MAUC evaluates the model's discrimination ability across all fibrosis stages, and macro-F1 score assesses the balance between precision and recall for each class, making it particularly suitable for imbalanced multiclass classification. The consistently superior performance of XGBoost across all three complementary metrics indicated not only high overall predictive accuracy but also robust multiclass discrimination and balanced classification performance. Therefore, XGBoost was selected as the final model for subsequent validation and interpretation analyses.

3.4 Performance evaluation and interpretation of the final model

The performance and interpretability of the final XGBoost model are presented in Figure 4. The confusion matrix demonstrated good classification performance across fibrosis stages, with most samples correctly classified (Figure 5A). The multiclass calibration curves showed good agreement between predicted and observed probabilities, indicating satisfactory calibration of the model (Figure 5B).

To further evaluate its clinical utility, the model was compared with the conventional FIB-4 index for the identification of advanced fibrosis (Figure 5C). XGBoost achieved a significantly higher AUC than FIB-4 (0.902 vs. 0.645), suggesting superior discrimination ability for fibrosis risk assessment.

SHAP analysis was subsequently performed to quantify the contribution of individual predictors to fibrosis stage classification (Figure 5D-G). Across different fibrosis stages, GNRI, BMI, ALB, WtHR, CRP, TC, HDL, and NPAR consistently ranked among the most influential features. Stage-specific feature profiling further demonstrated distinct patterns of nutritional, metabolic, and inflammatory alterations

across fibrosis stages (Figure 5H). In particular, higher BMI, CRP, and GNRI values were more strongly associated with advanced fibrosis stages, whereas HDL and TC showed inverse associations. These findings provide biological interpretability for the model and support the important roles of nutritional status, metabolic dysfunction, and systemic inflammation in fibrosis progression.

4. Discussion

Accurate assessment of liver fibrosis is important for identifying individuals with advanced liver disease and guiding clinical management. To our knowledge, this is the first study to investigate the associations of nutritional and inflammatory indicators with liver fibrosis severity among individuals with previous or current HBV exposure in a nationally representative NHANES cohort. Using data from the NHANES 2017–2018 and 2019–2020 cycles, we explored the relationships between nutritional status, inflammatory burden, and fibrosis severity, and developed an explainable machine-learning model for cross-sectional classification of liver fibrosis stages. The findings support the potential utility of routinely available clinical and laboratory indicators for noninvasive fibrosis assessment and provide a foundation for future validation studies in independent HBV-exposed populations.

4.1Multidimensional Clinical and Molecular Determinants of Liver Fibrosis Severity among HBV-Exposed Individuals

In the clinical setting, multidimensional factors such as basic characteristics, pathophysiological characteristics, and biochemical molecules can determine the progression of cirrhosis. Here, we comparedthe fibrosis level with almost 19 clinical variables and molecular characteristics collected from NHANES in HBV patients. According to the results of univariate analysis, factors such as age, ethnicity, smoke, GNRI, WHR, TC, HDL, BMI, MET, DM, hypertension, CRP, and NPAR are associated with the degree of liver fibrosis. These factors can be broadly categorized into four groups based on clinical experience: sociodemographic, comorbidity, nutritional, and inflammatory factors.

Correlation analyses further revealed that individuals with poorer nutritional status, greater metabolic burden, and higher levels of systemic inflammation were more likely to exhibit advanced liver fibrosis. Positive associations were observed between fibrosis severity and BMI, WtHR, CRP, and NPAR, whereas HDL and TC showed inverse relationships. However, these associations were not uniformly linear. To further characterize their nature, RCS analyses were performed for key nutritional and inflammatory indicators. The results demonstrated several non-linear associations between fibrosis severity and variables including GNRI, NPAR, SIRI, BMI, and age, suggesting that the influence of these factors may vary across different stages of disease progression rather than remaining constant throughout the course of fibrosis development.

Notably, both nutritional and inflammatory indicators exhibited dynamic stage-dependent patterns. Nutritional status appeared to improve during the early-to-intermediate stages of fibrosis before declining in advanced disease, suggesting a transient compensatory phase followed by progressive metabolic and functional deterioration as hepatic injury accumulates. Likewise, NPAR and SIRI

demonstrated complex non-linear relationships with fibrosis severity, indicating that the role of inflammation may evolve throughout disease progression. While chronic low-grade inflammation is widely recognized as a major driver of hepatic fibrogenesis, persistent activation of inflammatory pathways promotes extracellular matrix deposition and fibrosis progression. In contrast, more intense inflammatory responses may reflect distinct biological processes, including immune-mediated clearance of damaged hepatocytes or acute compensatory reactions. Although causality cannot be inferred from RCS analyses, these findings suggest that liver fibrosis progression is not simply the consequence of cumulative liver injury, but rather a dynamic process involving the interaction of nutritional status, metabolic disturbance, and systemic inflammation.

4.2 Nutritional and Inflammatory Predictive Factors of Liver Fibrosis

In this study, the RF algorithm was used to rank candidate predictors according to their importance scores, resulting in the identification of GNRI, BMI, NPAR, ALB, TC, CRP, creatinine, WtHR, and HDL as the most informative features for fibrosis prediction. Rather than selecting variables solely based on their importance rankings, we adopted a sequential feature-selection strategy in which predictors were gradually added to the model according to their ranked importance, generating a series of nested models from the highest-ranked feature to the full set of 19 candidate variables. Model performance was assessed throughout the sequential feature-selection process, and the final feature set was chosen based on overall predictive performance. This approach optimized model parsimony and predictive accuracy while reducing redundant information, allowing the identification of a feature set that achieved strong predictive performance with a limited number of clinically meaningful variables. Notably, the selected variables span multiple biological domains, including nutritional status, adiposity, lipid metabolism, renal function, and systemic inflammation, suggesting that liver fibrosis progression is influenced by a complex interplay of metabolic and inflammatory processes rather than a single pathological pathway. The integration of these complementary indicators provides a more comprehensive characterization of fibrosis risk and establishes a robust foundation for subsequent machine-learning model development.

From an interpretive perspective, we believe these nutritional and inflammatory variables are meaningfully associated with liver fibrosis and may be biologically involved in the mechanisms driving its progression. First, GNRI, a nutritional index commonly used to assess malnutrition in elderly patients, ranked highest in importance, underscoring the critical role of nutritional status in liver health. Malnutrition is associated with poorer outcomes in liver disease, as it impairs immune function and exacerbates inflammation, accelerating fibrosis progression[36]. BMI, another key nutritional indicator, further supports the importance of body composition and metabolic health in fibrosis development. Higher BMI levels, often linked to metabolic syndrome and NAFLD, are established risk factors for liver fibrosis[37], suggesting that weight management is crucial in mitigating fibrosis risk. Besides, Albumin and NPAR serve as markers of both nutritional and inflammatory status. Low albumin levels often indicate compromised liver function and malnutrition[38], which are associated with more severe

fibrosis. NPAR and SIRI, reflects systemic inflammation and is linked to various inflammatory diseases[39–41]. Its importance in predicting liver fibrosis highlights the pivotal role of inflammation in fibrosis progression, consistent with previous studies associating chronic inflammation with liver damage. Then TC and HDL were also identified as significant predictors. Dyslipidemia is common in liver disease, and lipid metabolism disturbances are implicated in the pathogenesis of NAFLD and fibrosis[42]. Higher HDL levels, typically associated with a protective cardiovascular profile, may also benefit liver health, while elevated TC suggests metabolic dysregulation, contributing to fibrosis. Last, CRP, a well-known marker of systemic inflammation, and creatinine, an indicator of kidney function, also play roles in predicting liver fibrosis. Elevated CRP levels indicate increased inflammation, driving the fibrotic process. Creatinine, which reflects renal function, underscores the interconnection between liver and kidney health. Chronic liver disease can lead to renal impairment, and reduced kidney function exacerbates liver disease outcomes. Together, these markers provide a comprehensive profile of liver fibrosis risk, integrating nutritional, metabolic, and inflammatory dimensions.

4.3 Clinical prediction model of liver fibrosis

To facilitate fibrosis risk stratification among individuals with HBV infection, we developed an explainable machine-learning model based on routinely available nutritional, metabolic, and inflammatory indicators, using TE-derived fibrosis stages as reference standards. Although TE is a reliable noninvasive tool for fibrosis assessment, its clinical application remains limited by the requirement for specialized equipment and trained personnel. Therefore, a prediction model based on accessible clinical variables may provide a practical approach for fibrosis screening and monitoring. To ensure reliable model development, multiple machine-learning algorithms were evaluated and model performance was assessed using complementary metrics, including accuracy, MAUC, and macro-F1 score, thereby capturing overall classification performance, multiclass discrimination, and classification balance. Among all evaluated algorithms, XGBoost consistently demonstrated the best overall performance and was therefore selected as the final model for fibrosis risk stratification. This superior performance may be attributable to its ability to capture complex nonlinear relationships and interactions among predictors, enabling the effective integration of nutritional, metabolic, and inflammatory information into a unified prediction framework.

In addition to its predictive performance, the model also offers a high degree of interpretability through SHAP analysis. Unlike conventional “black-box” machine-learning approaches, SHAP enables quantification of the contribution of individual predictors to model outputs and provides patient-specific explanations for fibrosis-stage classification. This transparency not only improves clinician confidence in model predictions but also helps identify the key nutritional, metabolic, and inflammatory factors driving fibrosis risk in individual patients. Interestingly, SHAP analyses also revealed that the relative importance and direction of individual predictors varied across fibrosis stages, suggesting that different biological processes may dominate at different phases of disease progression. Nutritional indicators, inflammatory markers, and metabolic variables did not contribute uniformly to all fibrosis stages, but instead exhibited stage-specific patterns of influence. These findings further support the notion that liver fibrosis

progression is a dynamic process involving evolving interactions among nutritional status, metabolic disturbance, and systemic inflammation. By identifying the key drivers of prediction at each fibrosis stage, SHAP analysis provides additional biological insight beyond overall model performance and may facilitate more personalized risk assessment and intervention strategies.

Importantly, the final model was developed using routinely available nutritional, metabolic, and inflammatory indicators, including GNRI, BMI, ALB, lipid-related markers, and inflammation-related indices. These variables are inexpensive, widely accessible, and commonly collected during routine clinical assessments, eliminating the need for specialized imaging equipment or invasive procedures. Consequently, the model may be readily implemented in a variety of clinical settings, particularly in resource-limited environments where access to transient elastography is limited. To further facilitate clinical translation, reproducibility, and external validation, the complete modeling framework and source code have been made publicly available through GitHub (https://github.com/githubliteng/HBV-Liver-Fibrosis.git). Users can directly input routinely collected clinical variables to obtain predicted probabilities for fibrosis stages as well as individualized SHAP-based explanations of feature contributions. This open-access framework not only enables independent validation and future optimization of the model but also provides a foundation for the development of user-friendly web-based platforms or clinical decision-support tools. With further external validation, such tools may assist clinicians in identifying high-risk individuals, prioritizing patients for further fibrosis assessment, and implementing earlier preventive or therapeutic interventions.

In summary, the clinical application of our predictive model not only enhances decision-making and patient management but also promotes efficiency and safety in the assessment process. Its ability to integrate seamlessly into existing workflows while providing reliable predictions positions it as a crucial asset in improving patient outcomes and optimizing healthcare delivery. Meanwhile, several limitations of this study should be acknowledged. First, this was a cross-sectional study based on NHANES data, therefore, causal relationships between nutritional or inflammatory factors and liver fibrosis severity cannot be established. Second, liver fibrosis staging was determined using transient elastography-derived liver stiffness measurements rather than histological assessment. Third, model performance was evaluated using repeated 10-fold cross-validation within the NHANES dataset only, and no independent external validation cohort was available. Therefore, the generalizability of the findings requires further confirmation in prospective studies and external populations.

5. Conclusion

Using nationally representative NHANES data, this study investigated the associations between nutritional and inflammatory indicators and liver fibrosis severity among individuals with previous or current HBV exposure. Among the evaluated algorithms, the XGBoost model demonstrated the best overall performance. These findings suggest that nutritional and inflammatory factors may provide useful information for the classification of liver fibrosis severity in HBV-exposed populations. However, given the cross-sectional design, transient elastography-based outcome definition, class imbalance, and

lack of external validation, the proposed model should be considered exploratory. Further validation in independent and prospective cohorts is necessary before clinical application can be considered.

Declarations

Ethics approval and consent to participate

The study protocol was approved by the NHANES Institutional Review Board, and was performed in accordance with the Declaration of Helsinki, with all NHANES participants providing signed informed consent.

Competing interests

We declare that we have no conflicts of interest.

Funding

This study was supported by the project of the Department of Education of Henan Province (SB201904006) and the scientific research project assisted by the Liver and Gallbladder Fund of the Charity Federation of Henan Province (GDXZ2019003).

Author Contribution

A.Z.: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing—original draft.M.L.: Conceptualization, Methodology, Software, Formal analysis, Validation, Writing—original draft.T.L.: Methodology, Data curation, Investigation, Visualization, Writing—original draft.H.Z.: Data curation, Formal analysis, Validation, Writing—review and editing.X.S.: Data curation, Writing—review and editing.Y.H.: Data curation, Visualization, Writing—review and editing.W.G.: Methodology, Supervision, Writing—review and editing.Y.Z.: Conceptualization, Supervision, Project administration, Writing—review and editing.P.W.: Conceptualization, Supervision, Project administration, Funding acquisition, Writing—review and editing.

Acknowledgments

The authors would like to acknowledge the support of all team members and the Department of Hepatobiliary and Pancreatic Surgery of the First Affiliated Hospital of Zhengzhou University.

Data Availability

The datasets analyzed in the current study are publicly available from the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. Data from the 2017–2018 and 2019–March 2020 pre-pandemic cycles were used and can be accessed through the NHANES website. The analytic code used for data processing and model development is publicly available at https://github.com/githubliteng/HBV-Liver-Fibrosis.git.

References

1. Hernandez-Gea, V. & Friedman, S. L. Pathogenesis of liver fibrosis. *Annu. Rev. Pathol.* **6**, 425–456 (2011).
2. Kisseleva, T. & Brenner, D. Molecular and cellular mechanisms of liver fibrosis and its regression. *Nat. Rev. Gastroenterol. Hepatol.* **18** (3), 151–166 (2021).
3. Friedman, S. L. Liver fibrosis – from bench to bedside. *J. Hepatol. 38 (Suppl* 1), S38–S53 (2003).
4. Lee, Y. S. et al. Exosomes derived from palmitic acid-treated hepatocytes induce fibrotic activation of hepatic stellate cells. *Sci. Rep.* **7**, 3710 (2017).
5. Collaborators, G. B. D. C. The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet Gastroenterol. Hepatol.* **5** (3), 245–266 (2020).
6. Lu, M. et al. Serum biomarkers indicate long-term reduction in liver fibrosis in patients with sustained virological response to treatment for HCV infection. *Clin. Gastroenterol. Hepatol.* **14**, 1044–1055e3 (2016).
7. Arndtz, K. & Hirschfield, G. M. The pathogenesis of autoimmune liver disease. *Dig. Dis.* **34**, 327–333 (2016).
8. Marcellin, P. & Kutala, B. K. Liver diseases: a major, neglected global public health problem requiring urgent actions and large-scale screening. *Liver Int.* **38** (Suppl. 1), 2–6 (2018).
9. Younossi, Z. et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 11–20 .
10. Bataller, R. & Brenner, D. A. *Liver Fibros. J. Clin. Invest.* **115**, 209–218 (2005).
11. GBD & Europe Hepatitis, B. & C Collaborators. Hepatitis B and C in Europe: an update from the Global Burden of Disease Study 2019. Lancet Public Health. 2023;8(9):e701-e716. (2019).
12. Jeng, W. J., Papatheodoridis, G. V. & Lok, A. S. F. *Hepat. B Lancet* ;**401**(10381):1039–1052. (2023).
13. Gong, J. et al. Hepatic SATB1 induces paracrine activation of hepatic stellate cells and is upregulated by HBx. *Sci. Rep.* **6**, 37717 (2016).
14. Michalopoulos, G. K. & Bhushan, B. Liver regeneration: biological and pathological mechanisms and implications. *Nat. Rev. Gastroenterol. Hepatol.* **18** (1), 40–55 (2021).
15. Rosselli, M. et al. Beyond scoring: a modern interpretation of disease progression in chronic liver disease. *Gut* **62**, 1234–1241 (2013).

16. Clinical Practice, E. A. S. L. A. L. E. H. & Guidelines Non-invasive tests for evaluation of liver disease severity and prognosis. *J. Hepatol.* **63**, 237–264. https://doi.org/10.1016/j. jhep.2015.04.006 (2015).
17. Efthimiou, O. et al. Developing clinical prediction models: a step-by-step guide. *BMJ* **386**, e078276 (2024).
18. Jamthikar, A. et al. Cardiovascular/stroke risk predictive calculators: a comparison between statistical and machine learning models. *Cardiovasc. Diagn. Ther.* **10** (4), 919–938 (2020).
19. Chang, D. et al. Machine learning models are superior to noninvasive tests in identifying clinically significant stages of NAFLD and NAFLD-related cirrhosis. *Hepatology* **77** (2), 546–557 (2023).
20. Zhang, C. et al. A machine learning-based model analysis for serum markers of liver fibrosis in chronic hepatitis B patients. *Sci. Rep.* **14** (1), 12081 (2024).
21. Wang, Z. et al. Clinical prediction of HBV-associated cirrhosis using machine learning based on platelet and bile acids. *Clin. Chim. Acta*. **551**, 117589 (2023).
22. Pu, X. et al. High-dimensional hepatopath data analysis by machine learning for predicting HBV-related fibrosis. *Sci. Rep.* **11** (1), 5081 (2021).
23. NHANES. National health and nutrition examination survey Homepage. (2022). Available: https://www.cdc.gov/nchs/nhanes/
24. NHANES. National health and nutrition examination survey questionnaires, datasets, and documentation. https://wwwn.cdc.gov/nchs/nhanes/
25. Castera, L., Friedrich-Rust, M. & Loomba, R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. *Gastroenterology* **156**, 1264–1281e4 (2019).
26. Eddowes, P. et al. Accuracy of FibroScan controlled attenuation parameter and liver stiffness measurement in assessing steatosis and fibrosis in patients with nonalcoholic fatty liver disease. *Gastroenterology* **156**, 1717–1730 (2019).
27. Chon, Y. E. et al. Performance of transient elastography for the staging of liver fibrosis in patients with chronic hepatitis B: a meta-analysis. *PLoS One*. **7** (9), e44930 (2012).
28. Crook, J. M. et al. Vitamin C Plasma Levels Associated with Inflammatory Biomarkers, CRP and RDW: Results from the NHANES 2003–2006 Surveys. *Nutrients* **14** (6), 1254 (2022).
29. Wu, C. C. et al. Association between neutrophil percentage-to-albumin ratio (NPAR), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR) and long-term mortality in community-dwelling adults with heart failure: evidence from US NHANES 2005–2016. *BMC Cardiovasc. Disord*. **23**, 312 (2023).
30. Wang, R. H. et al. The clinical value of neutrophil-to-lymphocyte ratio (Nlr), systemic immune-inflammation index (Sii), platelet-to-lymphocyte ratio (Plr) and systemic inflammation response index (Siri) for predicting the occurrence and severity of pneumonia in patients with intracerebral hemorrhage. *Front. Immunol.* **14**, 1115031 (2023).
31. Bouillanne, O. et al. Geriatric Nutritional Risk Index: A New Index for Evaluating at-Risk Elderly Medical Patients. *Am. J. Clin. Nutr.* **82**, 777–783 (2005).

32. Cao, Z. et al. Liver fibrosis biomarkers as a prognostic tool beyond the defining components in cardiovascular-kidney-metabolic syndrome. *Am. J. Prev. Cardiol.* **26**, 101438 (2026).
33. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2022. *Diabetes Care*. **45** (Suppl 1), S17–s38 (2022).
34. Alberti, K. G., Eckel, R. H., Grundy, S. M. & of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association. Harmonizing the metabolic syndrome: a joint interim statement ; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation. ;120(16):1640-5. (2009).
35. Clària, J., Arroyo, V. & Moreau, R. Roles of systemic inflammatory and metabolic responses in the pathophysiology of acute-on-chronic liver failure. *JHEP Rep.* **5** (9), 100807 (2023).
36. van Zutphen, T. et al. Malnutrition-associated liver steatosis and ATP depletion is caused by peroxisomal and mitochondrial dysfunction. *J. Hepatol.* **65** (6), 1198–1208 (2016).
37. Huang, D. Q. et al. Type 2 diabetes, hepatic decompensation, and hepatocellular carcinoma in patients with non-alcoholic fatty liver disease: an individual participant-level data meta-analysis. *Lancet Gastroenterol. Hepatol.* **8** (9), 829–836 (2023).
38. Allison, S. P. & Lobo, D. N. The clinical significance of hypoalbuminaemia. *Clin. Nutr.* **43** (4), 909–914 (2024).
39. Pan, J., Li, J. & Gao, Y. The value of 7 peripheral blood serum ratios in diagnosis and prediction of disease activity of patients within inflammatory bowel disease individuals. *Front. Med. (Lausanne)*. **10**, 1122005 (2023).
40. Deng, L. et al. Neutrophil percentage-to-albumin ratio is a potential marker of intravenous immunoglobulin resistance in Kawasaki disease. *Sci. Rep.* **14** (1), 15232 (2024).
41. Wang, Y. et al. Association of systemic immune biomarkers with metabolic dysfunction-associated steatotic liver disease: a cross-sectional study of NHANES 2007–2018. *Front. Nutr.* **11**, 1415484 (2024).
42. Sinha, R. A., Singh, B. K. & Yen, P. M. Direct effects of thyroid hormones on hepatic lipid metabolism. *Nat. Rev. Endocrinol.* **14** (5), 259–269 (2018).
43. Guo, C. et al. Machine-learning-based plasma metabolomic profiles for predicting long-term complications of cirrhosis. *Hepatology* Apr **17**. (2024).

Figures

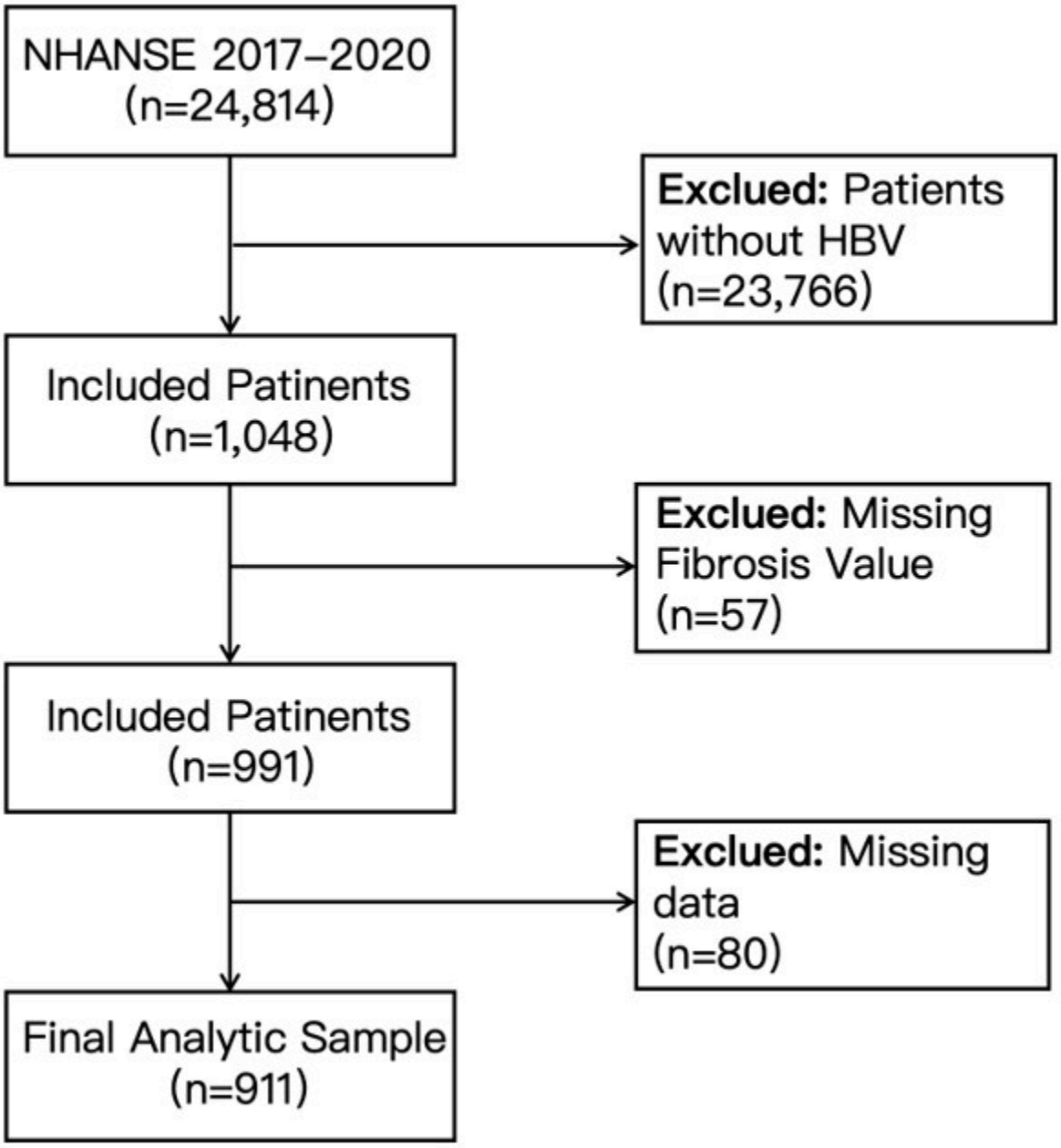


Figure 1

Flowchart of study participant selection

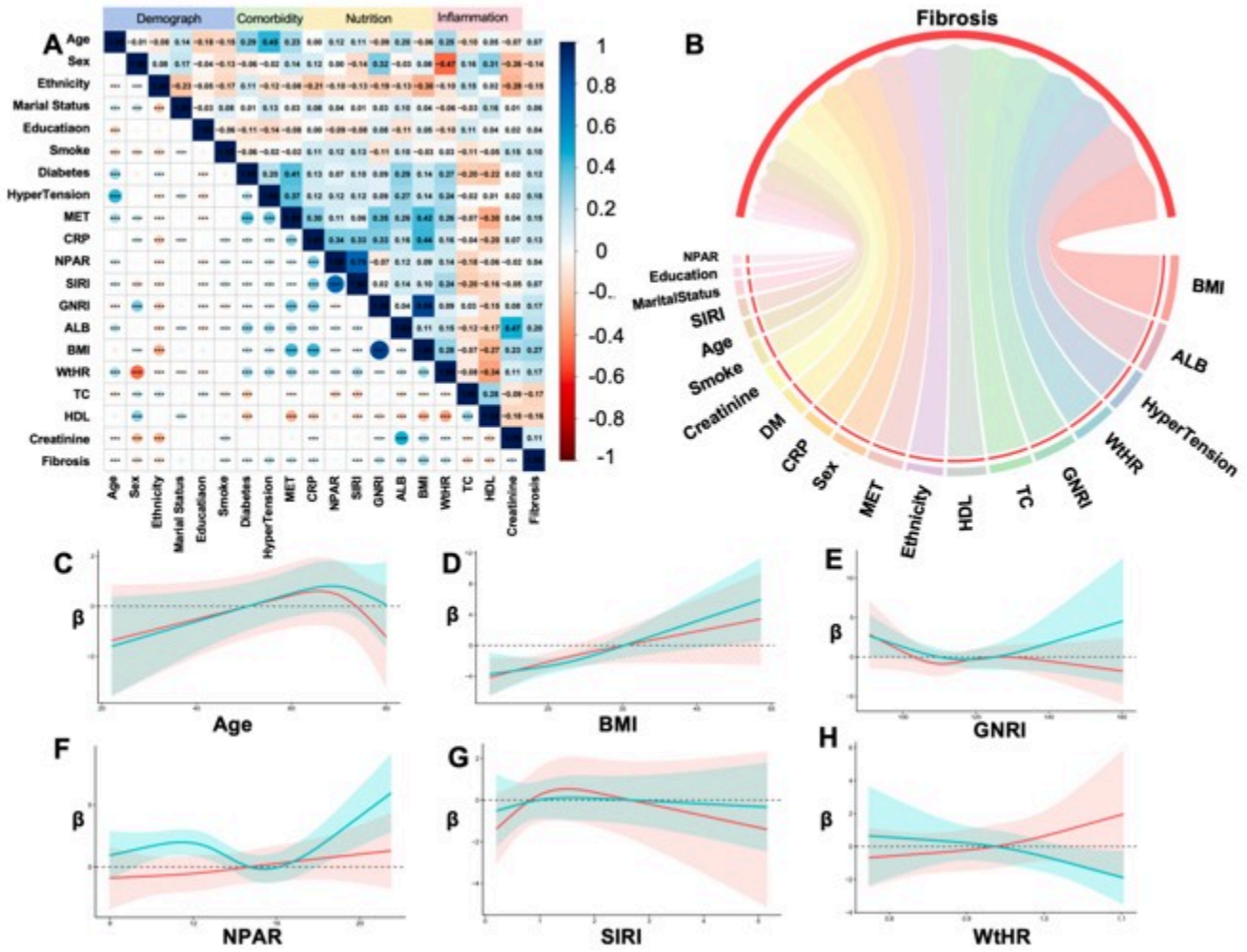


Figure 2

Correlation structure and nonlinear relationships between clinical variables and liver fibrosis. (A) Pairwise correlation matrix showing the relationships among demographic, clinical, nutritional, metabolic, and inflammatory variables. Circle size and color intensity represent the strength of the correlation.**(B)** Chord diagram illustrating the major variables associated with fibrosis severity. Wider ribbons indicate stronger associations with fibrosis.**(C-H)** Restricted cubic spline (RCS) analyses evaluating the nonlinear associations between continuous variables and liver fibrosis. Specifically, panels show the relationships for age (C), BMI (D), GNRI (E), NPAR (F), SIRI (G), and WtHR (H). Solid lines represent the estimated effect sizes (β coefficients), and shaded areas indicate 95% confidence intervals.

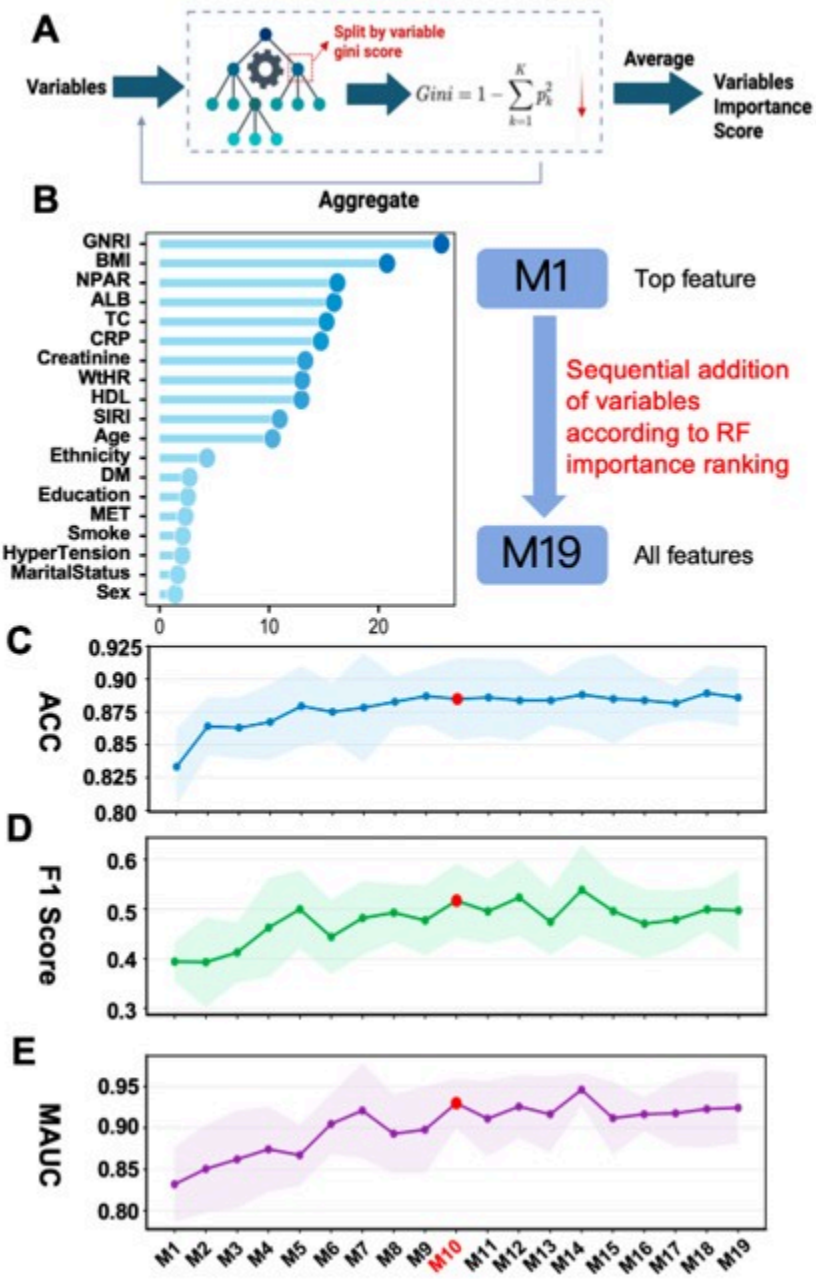


Figure 3

Variable selection and identification of the optimal feature set. (A) Schematic illustration of Random Forest (RF)-based feature importance calculation using the Gini impurity index. Variable importance scores were obtained by aggregating the reduction in Gini impurity across all trees. **(B)** Ranking of candidate predictors according to RF importance scores. Variables were sequentially added from the highest- to the lowest-ranked feature to generate 19 feature combinations (M1-M19). **(C-E)** Performance comparison of different feature combinations evaluated using 10-fold cross-validation. Model performance was assessed by classification accuracy (ACC), macro-F1 score, and multiclass area under the receiver operating characteristic curve (MAUC). The red dots indicate the optimal feature combination (M10), which achieved the best overall balance across the three evaluation metrics and was selected for subsequent model development.

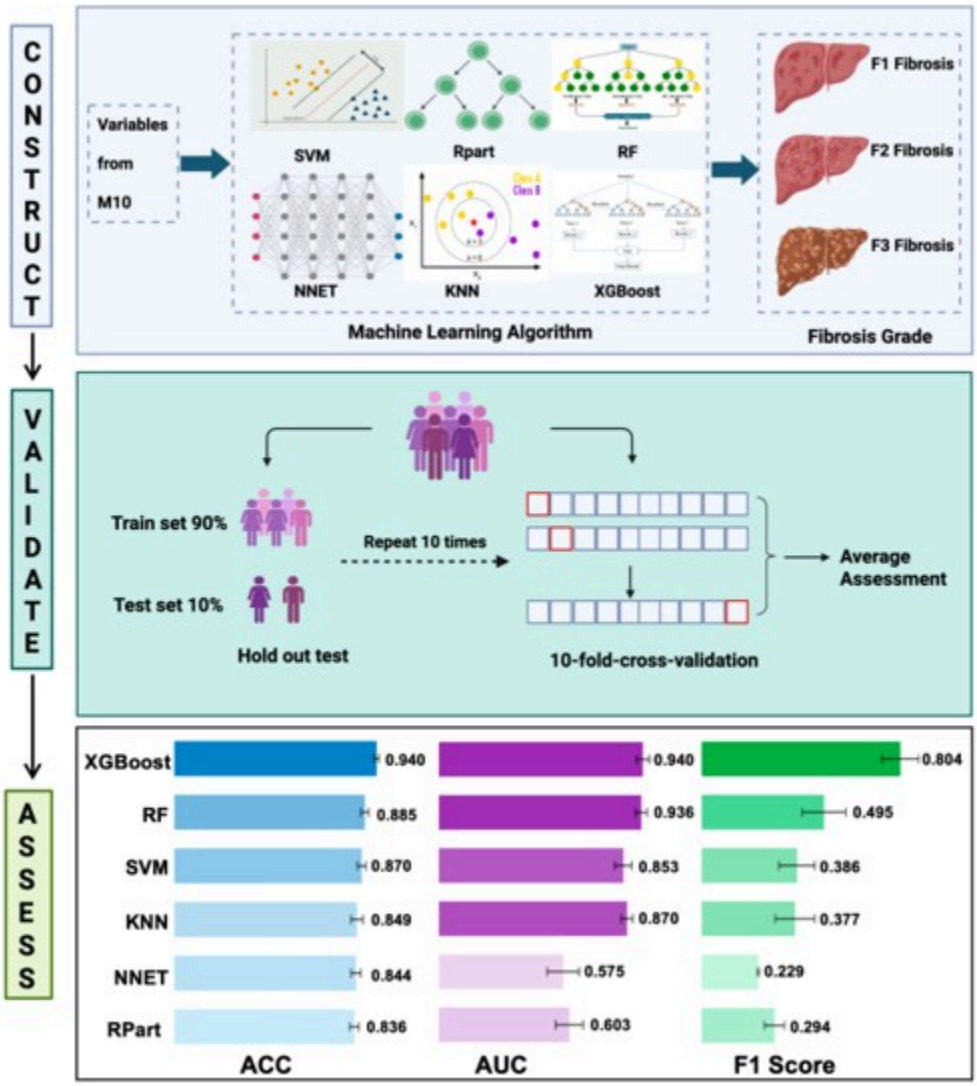


Figure 4

Development and evaluation of machine-learning models for fibrosis stage prediction. Construction: Variables from the optimal feature combination (M10) were used to develop six machine-learning algorithms (SVM, RPart, RF, NNET, KNN, and XGBoost) for fibrosis stage classification. **Validation:** Model performance was assessed using repeated 10-fold cross-validation. **Assessment:**ACC, MAUC, and macro-F1 score were calculated to compare model performance. XGBoost demonstrated the highest ACC (0.940), MAUC (0.940), and macro-F1 score (0.804), and was therefore selected as the final model.

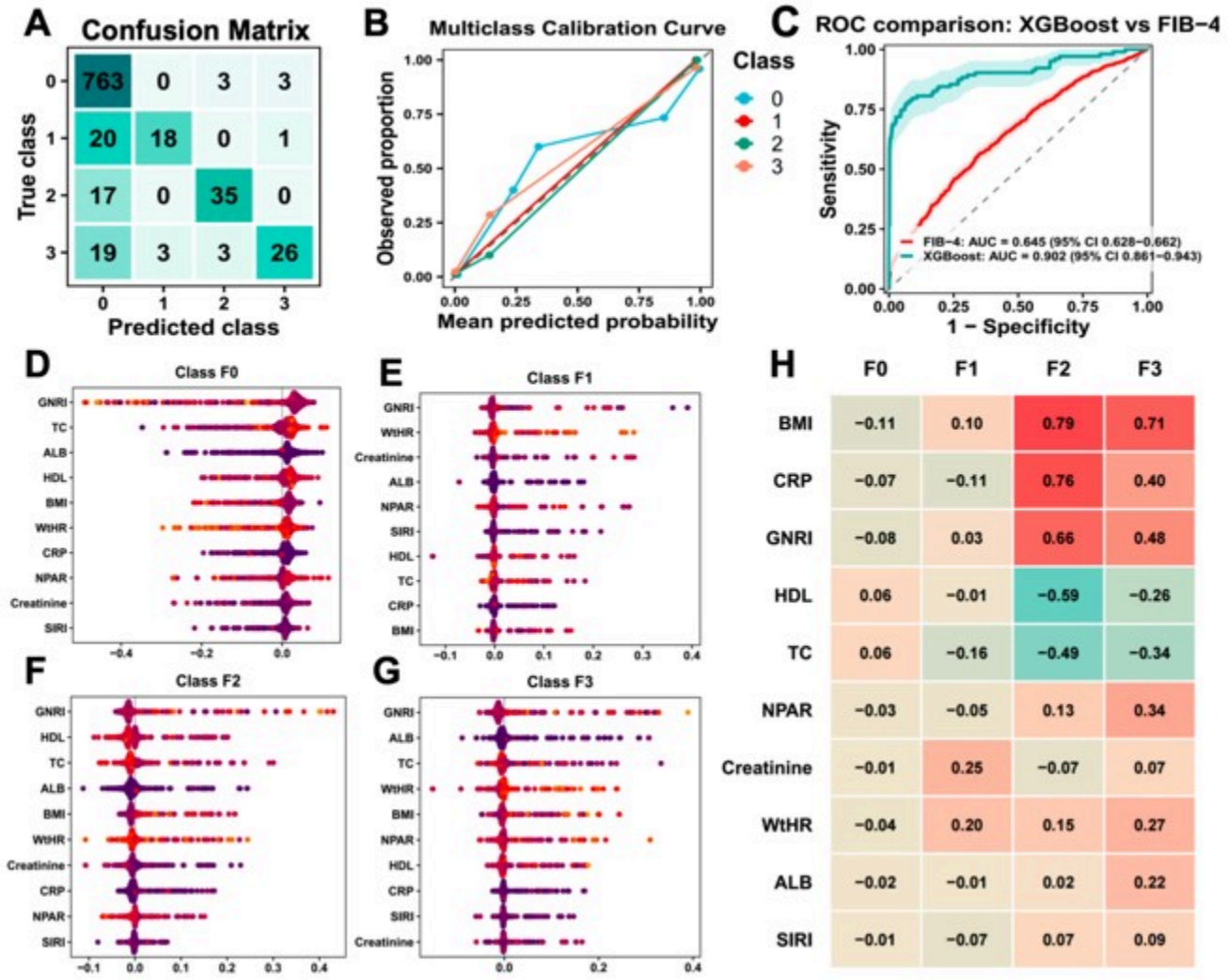


Figure 5

Performance evaluation and interpretability of the XGBoost model for liver fibrosis staging. (A) Confusion matrix showing the classification performance of the XGBoost model across the four fibrosis stages (F0–F3). **(B)** Multiclass calibration curves demonstrating the agreement between predicted probabilities and observed outcomes for each fibrosis stage. **(C)** Receiver operating characteristic (ROC) analysis comparing the XGBoost model with the conventional FIB-4 index for identifying advanced fibrosis. **(D–G)** SHAP summary plots for fibrosis stages F0, F1, F2, and F3, respectively. Each point represents an individual participant, and the color indicates the feature value. **(H)** Heatmap of mean SHAP values across fibrosis stages, highlighting the differential influence of selected predictors on model outputs.