Original Paper

Explainable Machine Learning Models for Predicting Significant Fibrosis in Steatotic Liver Disease: Model Development Study

Basile Njei[1], MD, MPH, PhD; Ulrick Sidney Kanmounye[2], MD, MPH, MSc

[1]Section of Digestive Diseases, Yale University, New Haven, CT, United States

[2]Immeuble Nziko, Rue Marie Gocker, Research Department, Association of Future African Neurosurgeons, Yaounde, Centre, Cameroon

Corresponding Author:

Basile Njei, MD, MPH, PhD
Section of Digestive Diseases
Yale University
333 Cedar Avenue
New Haven, CT
United States
Phone: 1 4752275537
Email: basilenjei@gmail.com

Abstract

Background: Steatotic liver disease (SLD) is highly prevalent, and noninvasive identification of patients with significant fibrosis remains a clinical challenge.

Objective: We conducted an exploratory evaluation of explainable machine learning models for predicting significant fibrosis in a US-based cohort of adults with SLD, with particular attention to sensitivity, false-negative risk, and the limitations of using area under the receiver operating characteristic curve (AUROC) alone to assess screening suitability.

Methods: We evaluated explainable machine learning models for predicting significant fibrosis in a US population–based dataset. A total of 43 routinely available clinical, demographic, anthropometric, laboratory, and lifestyle variables were included. After preprocessing and imputation, data were split into training (80%) and testing (20%) sets. The synthetic minority oversampling technique (SMOTE) was used to address class imbalance in the training set. Logistic regression, random forest, support vector machine, Extreme Gradient Boosting (XGBoost), k-nearest neighbors, naive Bayes, an artificial neural network, and a stacked ensemble were trained and evaluated. Performance metrics included accuracy, precision, recall, F_1-score, Cohen κ, and the AUROC. Because the intended use case was screening or triage, sensitivity was given particular interpretive importance. In addition to overall discrimination, a primary analytical objective was to examine sensitivity trade-offs, false-negative classifications, and the implications of model performance for screening suitability. Global and local model interpretability were assessed using feature importance, Shapley additive explanations (SHAP) values, individual conditional expectation (ICE) plots, and individualized explanations.

Results: Among 7056 participants, 716 (10.1%) had significant fibrosis. XGBoost achieved the highest discrimination (AUROC 0.79), followed by random forest (AUROC 0.78) and logistic regression (AUROC 0.77); however, XGBoost had low sensitivity (0.17), indicating limited utility for screening at the evaluated threshold. Logistic regression and naive Bayes demonstrated the highest sensitivity (0.65 and 0.55, respectively), suggesting that models with slightly lower discrimination may be more appropriate for triage-oriented use. Across models, waist circumference, BMI, cardiovascular comorbidity, sex, hemoglobin A_{1c} (HbA_{1c}), and liver enzymes were the most influential predictors.

Conclusions: In this exploratory model development study, moderate discrimination did not necessarily translate into clinically meaningful screening performance. The model with the highest AUROC showed poor sensitivity, indicating a substantial risk of false-negative classifications and limiting its usefulness for screening in its current form. Because imputation was performed before the train-test split, the reported performance estimates may be optimistic and should not be interpreted as definitive evidence of clinical readiness. Leakage-free preprocessing, threshold optimization, calibration assessment, decision curve analysis, external validation, and prospective evaluation are needed before clinical application.

Online J Public Health Inform 2026;18:e90677; doi: 10.2196/90677

Keywords: artificial intelligence; AI; explainability; liver fibrosis; steatosis

Introduction

Steatotic liver disease (SLD) has emerged as one of the most prevalent chronic liver diseases worldwide and is a leading cause of cirrhosis, hepatocellular carcinoma, and liver-related mortality [1-3]. Despite this growing burden, the early identification of individuals with significant fibrosis remains a persistent clinical challenge [4]. Vibration-controlled transient elastography (VCTE) offers a practical, noninvasive alternative to liver biopsy; however, its availability, cost, and technical limitations restrict its widespread application, particularly in primary care and resource-limited settings [5-8]. Traditional serum-based fibrosis scores, although inexpensive and accessible, have limited discriminatory ability and may lack sufficient sensitivity for individualized risk stratification [9-11]. In a screening or triage setting, this limitation is particularly important because false-negative results may delay further evaluation and provide false reassurance to patients at risk of progressive liver disease. Accordingly, the intended role of a predictive model in this setting is to support screening or triage rather than standalone diagnosis, which makes sensitivity especially important because false-positive results can be followed by further noninvasive testing, whereas false-negative results may lead to missed cases of clinically significant fibrosis [12].

In parallel, advances in machine learning have enabled the integration of large-scale clinical data into predictive models capable of capturing complex, nonlinear relationships among metabolic, anthropometric, biochemical, and behavioral factors that underpin fibrosis progression [13-16]. However, concerns about model transparency, reliability, and the potential for algorithmic bias have slowed the adoption of these models in clinical practice [17,18]. Established explainable AI (XAI) techniques can help characterize model behavior by providing patient-specific and population-level insights into the factors influencing predictions. In this study, XAI was used as an applied interpretability framework rather than as a methodological innovation. However, explainability alone does not ensure clinical usefulness, especially in low-prevalence settings where models with favorable area under the receiver operating characteristic curve (AUROC) values or accuracy may still perform poorly if sensitivity is inadequate.

In this context, we conducted an exploratory evaluation of machine learning models for predicting significant fibrosis in a population-based cohort using routinely collected clinical variables. Our objectives were to compare model performance in a low-prevalence screening context, identify models that may be unsuitable for screening because of low sensitivity or a high false-negative burden, and apply established XAI methods to characterize influential predictors, model behavior, and individual-level explanations. We envisioned this framework primarily as a potential noninvasive triage or risk stratification aid to help identify individuals who may benefit from further fibrosis assessment, rather than as a stand-alone diagnostic test.

Methods

Study Design and Reporting

We developed a model to predict significant fibrosis in SLD and performed internal validation. This manuscript adheres to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence (TRIPOD+AI) guidelines for clinical prediction models using machine learning, including the 27-item checklist and abstract extension [19].

Data Source, Participants, and Setting

We analyzed a US cohort of 7056 adults with SLD from the National Health and Nutrition Examination Survey (NHANES) 2017-2020 dataset. The dataset included 43 routinely available variables spanning demographics, anthropometry, blood pressure, laboratory measurements, lifestyle, and comorbidity indicators. After preprocessing and imputation, the data were randomly divided into training (5644/7056, 80%) and test (n=1412, 20%) sets, with stratification by fibrosis status. We recognize that conducting imputation on the data before partitioning them may have introduced data leakage, as information from the test set may have influenced the training process and inflated estimates of model performance.

Outcome Definition

The primary outcome was significant fibrosis, defined as a liver stiffness measurement of ≥8 kPa obtained using VCTE [7].

Predictors

A total of 43 prespecified predictors were used, including age (years), sex (female), race, BMI (kg/m^2), waist circumference (cm), hip circumference (cm), alanine aminotransferase (ALT; U/L), aspartate aminotransferase (AST; U/L), gamma-glutamyl transferase (GGT; U/L), alkaline phosphatase (ALP; U/L), albumin (g/dL), total bilirubin (mg/dL), total protein (g/dL), hemoglobin A_{1c} (HbA_{1c}; %), fasting glucose (mg/dL), triglycerides (mg/dL), total cholesterol (mg/dL), high-density lipoprotein (HDL) cholesterol (mg/dL), low-density lipoprotein (LDL) cholesterol (mg/dL), average systolic blood pressure (mm Hg), systolic blood pressure (first oscillometric), diastolic blood pressure (first oscillometric), alcohol consumption (g/day), Healthy Eating Index (HEI), smoking status, statin use, cardiovascular comorbidity, new diagnosis of hypertension, and hip-to-waist anthropometric measures.

Handling of Missing Data

Missing data were addressed using single imputation to create a complete analytic dataset. In the original analytic workflow, imputation was performed before the data were divided into training and test subsets. We acknowledge that this approach may have introduced data leakage because information from the held-out test set could have influenced the imputed values used during model development. Consequently, the reported performance metrics may be optimistic and should be interpreted as exploratory rather than definitive estimates of out-of-sample performance. A leakage-free pipeline, in which imputation and any other learned preprocessing steps are fitted exclusively to the training data and subsequently applied to the held-out test set, is required for future validation and clinical translation.

Management of Class Imbalance

Within the training set, we applied the synthetic minority oversampling technique (SMOTE) to address class imbalance [20]. SMOTE generates synthetic minority class observations by interpolating between existing minority class samples in feature space. In this study, SMOTE was applied only to the training set after data partitioning; the held-out test set was not oversampled and retained the original class distribution. After SMOTE, the training data included 10,142 observations, with equal numbers of participants with and without significant fibrosis. This approach was intended to improve minority class learning during model development while preserving a naturally imbalanced test set for evaluation.

Model Development

We trained 8 supervised machine learning algorithms, including logistic regression, random forest, support vector machine, Extreme Gradient Boosting (XGBoost), k-nearest neighbors, naive Bayes, artificial neural network, and a stacked ensemble. Hyperparameters were set using the default values provided by the respective Python (Python Software Foundation) machine learning libraries; formal hyperparameter optimization using grid search, random search, Bayesian optimization, or cross-validation was not performed. Models were fitted to the SMOTE-balanced training data and evaluated on the held-out test set. The purpose of this comparative analysis was not solely to identify the model with the highest AUROC but also to determine which models may be unsuitable for screening because of poor sensitivity or excessive false-negative burden.

Internal Validation and Performance Measures

We assessed discrimination using the AUROC and reported accuracy, precision (positive predictive value [PPV]), recall or sensitivity, F_1-score, and Cohen κ. Threshold-dependent metrics were calculated from class labels generated by each model on the test set at the default classification threshold. Formal threshold optimization was not performed. Because the intended clinical use case was screening or triage rather than stand-alone diagnosis, particular interpretive emphasis was placed on sensitivity, false-negative classifications, and the implications of model performance for screening suitability. However, the absence of sensitivity-focused threshold optimization limits conclusions about deployment-ready screening performance. Calibration analysis and decision curve analysis were not performed in the current exploratory analysis. Given the limitations associated with imputing data before the train-test split, all reported test-set performance metrics should be interpreted cautiously as exploratory estimates rather than definitive validation results.

Explainability

Explainability analyses were performed using established model interpretability techniques. Global interpretability was assessed using random forest feature importance and Shapley additive explanations (SHAP) summary and bar plots to identify predictors with the greatest influence on model output. Local interpretability was assessed using SHAP waterfall plots for selected individuals to examine the patient-level factors driving predictions. Individual conditional expectation (ICE) plots were used to visualize how predicted risk changed with changes in the values of selected predictors. These analyses were intended to characterize model behavior and potential failure modes rather than to introduce new XAI methodology.

Subgroup and Sensitivity Analyses

We generated ICE and SHAP dependence plots for the most influential predictors to explore potential nonlinearity and interactions.

Ethical Considerations

This study analyzed publicly available, deidentified data from the National Health and Nutrition Examination Survey (NHANES), conducted by the CDC's National Center for Health Statistics (NCHS). The original NHANES protocols received institutional ethics approval, and written informed consent was obtained from participants. Because this secondary analysis used exclusively deidentified public-use data, additional institutional review board approval and informed consent were not required under 45 CFR §46.102(e)(1). Data were accessed in accordance with CDC/NCHS public-use conditions, and participant confidentiality was maintained through aggregate reporting. No restricted-use or identifiable data were accessed, and no additional recruitment, intervention, or compensation occurred. Further details are available on the NHANES Ethics Review Board website [21].

Software and Reproducibility

The analyses were conducted using machine learning and statistical packages implemented in Python, including standard machine learning libraries for model fitting and evaluation. Complete code, preprocessing details, and parameter settings are available and will be shared upon request to support reproducibility.

Results

Overview

Among the 7056 participants included in the analysis, 716 (10.1%) had significant fibrosis. After preprocessing and single imputation, all 43 prespecified demographic, anthropometric, laboratory, and lifestyle variables were complete. After the data were divided into training (n=5644) and test (n=1412) sets, SMOTE was applied to the training set, generating a balanced training dataset of 10,142 samples.

Model Performance

Among the 8 machine learning algorithms evaluated on the test set, the XGBoost model demonstrated the highest discriminative performance (AUROC 0.79), followed by the random forest model (AUROC 0.78) and logistic regression (AUROC 0.77). Although XGBoost achieved the highest overall accuracy (0.90), its sensitivity for identifying significant fibrosis at the evaluated threshold was low (0.17). Logistic regression and k-nearest neighbors achieved the highest recall rates at the evaluated threshold, with sensitivities of 0.65 and 0.58, respectively. Naive Bayes provided a balanced performance profile, maintaining an AUROC of 0.77 and a sensitivity of 0.55. Naive Bayes achieved the highest chance-corrected agreement (Cohen κ=0.26), followed by logistic regression (Cohen κ=0.23). Model performance on the independent test set is summarized in Table 1. XGBoost and the stacked ensemble achieved the highest true-negative counts, correctly identifying 1243 and 1222 participants without fibrosis, respectively. However, both models had high false-negative rates, correctly identifying only 24 and 16 participants with fibrosis. Logistic regression and naive Bayes showed superior identification of significant fibrosis, with 93 and 78 true-positive results, respectively. This increased sensitivity was accompanied by higher false-positive rates, particularly for logistic regression, which incorrectly classified 306 participants without fibrosis. The random forest and artificial neural network models also maintained high true-negative counts (1212 and 1177, respectively) but continued to show low sensitivity, identifying only 30 and 41 true-positive cases, respectively. The support vector machine and k-nearest neighbors models demonstrated more balanced but moderate results, with k-nearest neighbors model recording 83 true-positive results alongside 384 false-positive results. The comparative confusion matrix analyses for the 8 models are illustrated in Multimedia Appendix 1.

Table 1. Comparative performance of machine learning algorithms for the prediction of significant fibrosis in the test set[a].

Models	Area under the receiver operating characteristic curve	Accuracy	Sensitivity (recall)	Precision (positive predictive value)	F_1-score	Cohen κ
Extreme Gradient Boosting	0.79	0.90	0.17	0.48	0.25	0.21
Random forest	0.78	0.88	0.21	0.34	0.26	0.20
Logistic regression	0.77	0.75	0.65	0.23	0.34	0.23
Naive Bayes	0.77	0.81	0.55	0.27	0.36	0.26
Stacked ensemble	0.76	0.88	0.11	0.25	0.16	0.10
Support vector machine	0.74	0.83	0.29	0.22	0.26	0.16
Artificial neural network	0.70	0.86	0.29	0.30	0.30	0.22
K-nearest neighbors	0.70	0.69	0.58	0.18	0.27	0.13

[a]Performance was evaluated on the independent test set (1412/7056, 10.1%; prevalence of significant fibrosis). Models were trained on the synthetic minority oversampling technique (SMOTE)–balanced training data (n=10,142) and applied to the untouched test set.

Predictors of Significant Fibrosis and Model Interpretability

Within the random forest model, waist circumference (importance score 0.07) and BMI (0.06) were identified as the leading predictors of advanced liver stiffness. The presence of cardiovascular comorbidity was the third most influential feature (0.06), followed by female sex (0.05) and HbA_{1c} level (0.04). Hip circumference (0.04) and high alcohol consumption (0.04) also contributed to the predictive model. The top 10 contributors to the prediction models are summarized in Table 2.

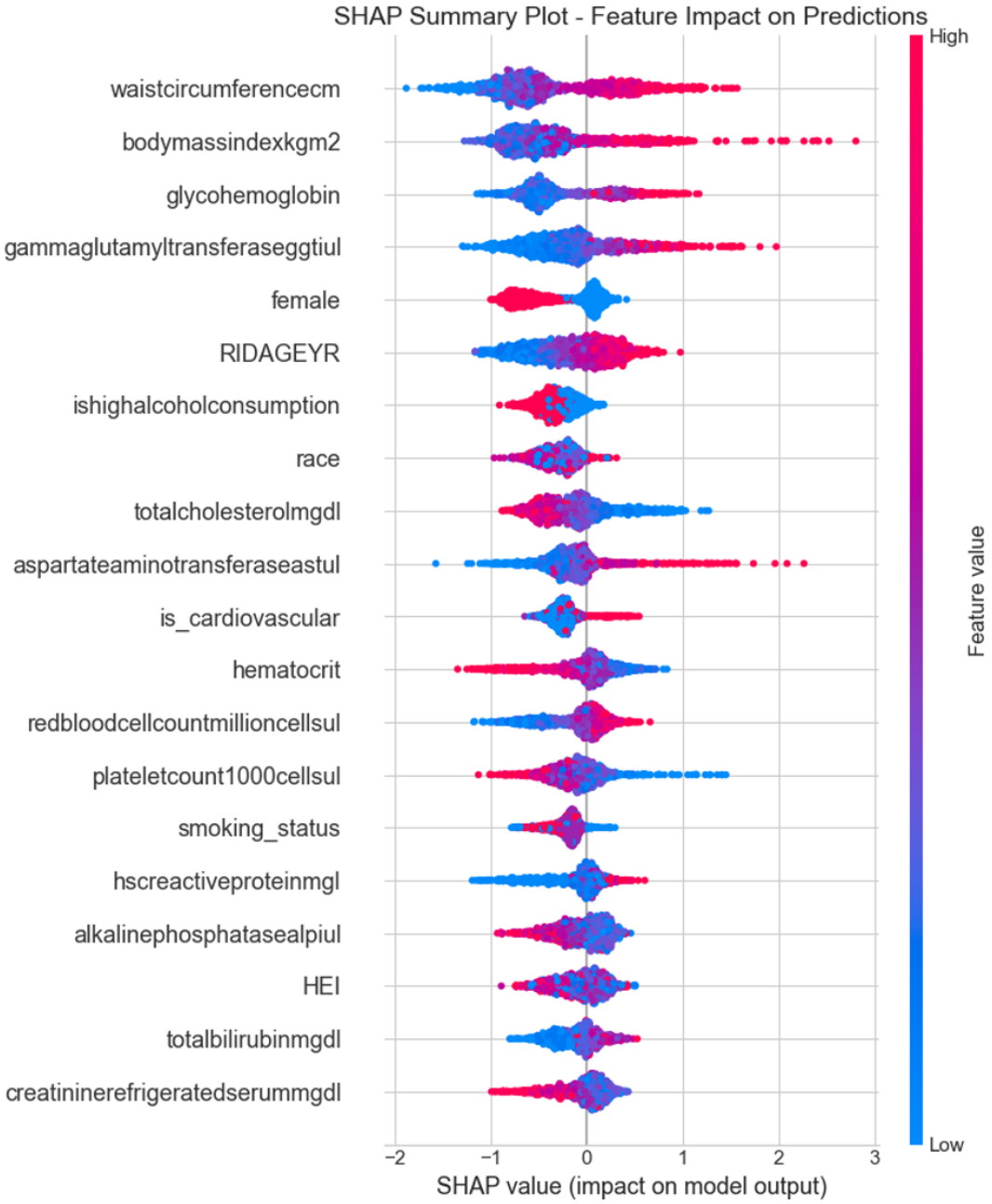
Table 2. Top 10 clinical, anthropometric, and laboratory predictors of significant fibrosis in the random forest model.

Features	Importance score
Waist circumference (cm)	0.07
BMI (kg/m^2)	0.06
Cardiovascular comorbidity (yes or no)	0.06
Sex (female)	0.05
Hemoglobin A_{1c} (%)	0.04
Hip circumference (cm)	0.04

Features	Importance score
High alcohol consumption (yes or no)	0.04
Gamma-glutamyl transferase (U/L)	0.03
Alanine aminotransferase (U/L)	0.03
New diagnosis of hypertension (yes or no)	0.03

The SHAP summary analysis showed that high values of waist circumference, BMI, and HbA_{1c} were strongly associated with positive impact on model output, indicating increased likelihood of significant fibrosis. Elevated levels of GGT and AST demonstrated a positive correlation with predicted risk. Female sex generally exerted a more negative influence on model output than male sex. Older age and higher alcohol consumption also contributed positively to the risk assessment. Other factors, such as cardiovascular comorbidity, smoking status, and lipid profiles, showed more moderate or individualized effects across the study population (Figure 1).

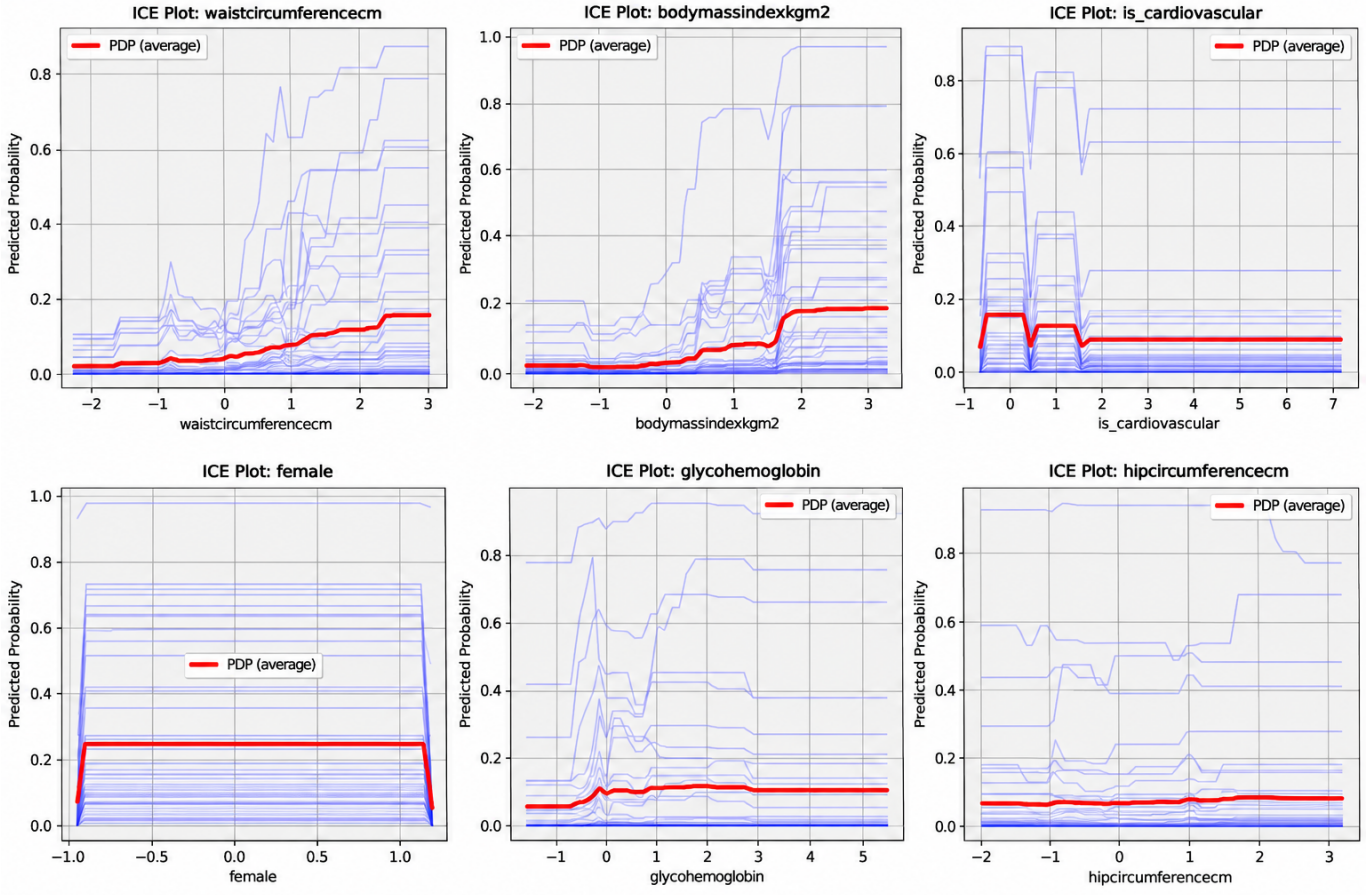
Figure 1. Shapley additive explanations (SHAP) summary plot illustrating global feature importance and the directional impact of clinical predictors on model output.



Local interpretability was assessed for 2 participants: one with fibrosis (ie, test index 30) and one without fibrosis (ie, test index 0). The model generated a predicted probability of 0.0220 for the individual with fibrosis and a predicted probability of 0.0891 for the individual without fibrosis. Patient-level model behavior was further examined using SHAP waterfall plots for the individual with fibrosis and the individual without fibrosis (Multimedia Appendices 1 and 2).

ICE analysis of the predicted probability of significant fibrosis across the 6 most influential features (ie, waist circumference, BMI, cardiovascular comorbidity, sex (female), HbA_{1c}, and hip circumference) showed that the average predicted probability remained low at lower values of waist circumference and BMI but demonstrated a marked increase as these values rose, particularly beyond the standardized mean. Similar patterns were observed for HbA_{1c}, where higher levels were associated with an elevated probability of fibrosis. Hip circumference exhibited a gradual positive relationship. Of note, cardiovascular comorbidity and female sex were associated with stable, stratified effects on predicted probability patterns across their respective categories (Figure 2).

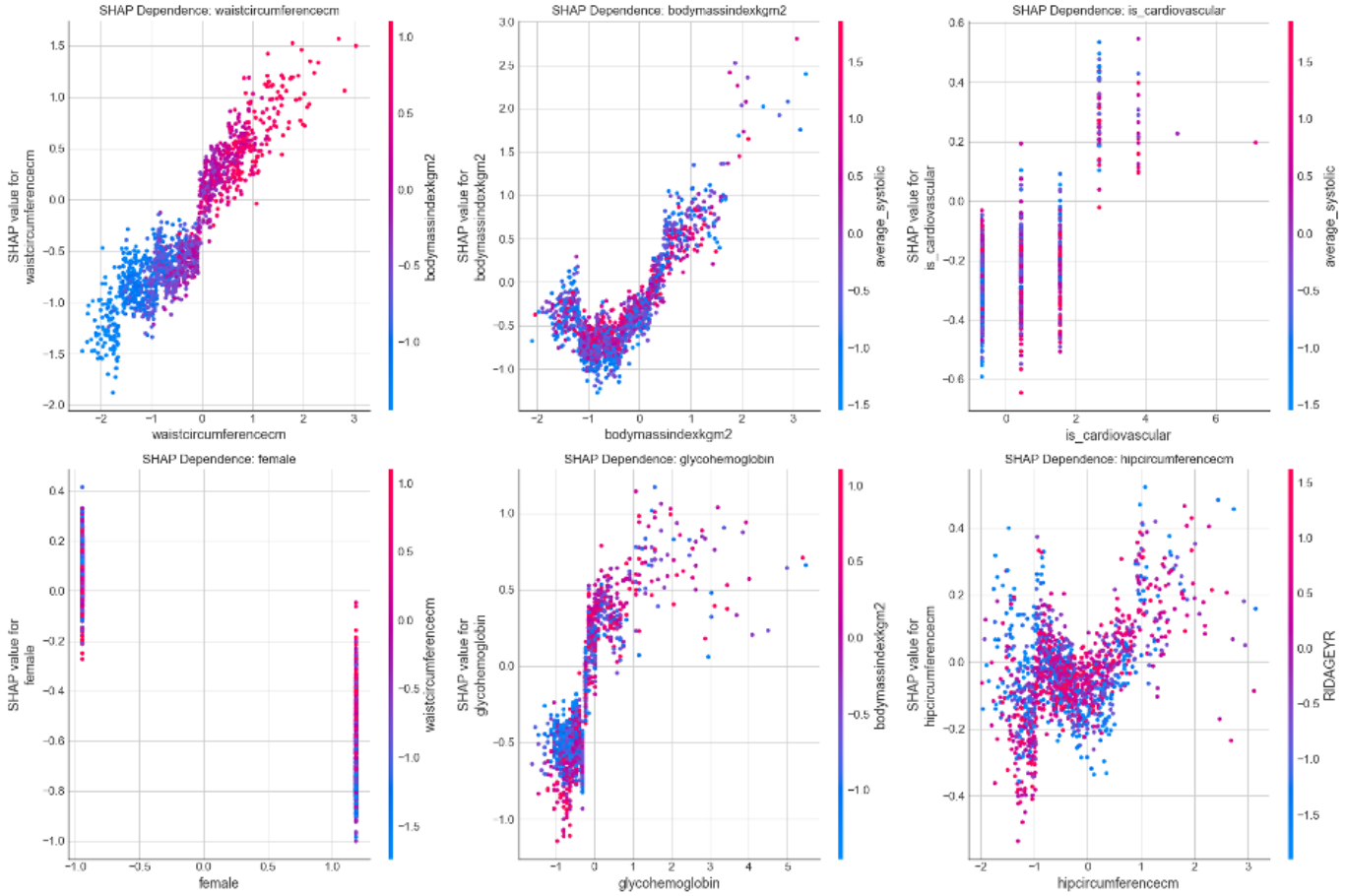
Figure 2. Individual conditional expectation (ICE) and partial dependence plots (PDPs) for the 6 most influential predictors of significant fibrosis (ie, waist circumference, BMI, cardiovascular comorbidity, sex [female], hemoglobin A_{1c}, and hip circumference).



Model Dependence and Feature Interaction

SHAP dependence analysis demonstrated distinct nonlinear relationships for waist circumference and BMI. Both features exhibited a strong positive, nearly linear relationship with predicted risk. In addition, cardiovascular comorbidity exerted a positive impact on the model's predictions across all observed values, with interactions often mediated by average systolic blood pressure. Moreover, female sex generally resulted in lower SHAP values, and the magnitude of this effect was influenced by waist circumference. Of note, HbA_{1c} displayed a steep increase in its contributions to predicted risk at lower concentrations, followed by a leveling off as values increased; this pattern often interacted with BMI. In contrast, hip circumference showed a more complex, U-shaped relationship with significant fibrosis, with its contribution to predicted risk influenced by age (Figure 3).

Figure 3. Shapley additive explanations (SHAP) dependence plots illustrating feature marginal effects and secondary interactions for the top predictors of significant fibrosis.



Discussion

Summary of Findings

In this exploratory, population-based model development study, machine learning models for predicting significant fibrosis demonstrated a clinically important trade-off between discrimination and sensitivity, although all performance estimates should be interpreted cautiously because imputation before the train-test split may have introduced data leakage. Accordingly, the purpose of the comparative analysis was not simply to identify the model with the highest AUROC but also to determine which models may be clinically unsuitable or potentially unsafe for screening because of the high number of false-negative classifications. Although ensemble methods such as XGBoost demonstrated higher overall discrimination, they were notably less effective at identifying individuals with significant fibrosis at the evaluated threshold than simpler probabilistic models, such as logistic regression and naive Bayes, which showed higher sensitivity. This finding should not be interpreted as support for using the model with the highest AUROC for screening; rather, it illustrates that models with favorable discrimination may still be unsuitable for triage if they miss a large proportion of true cases. This distinction is important because a model intended for screening or triage should prioritize the identification of true cases rather than rely on discrimination alone. Anthropometric measures related to central adiposity, specifically waist circumference and BMI, were the most influential predictors, showing a strong positive correlation with increased predicted risk. Interpretability analyses further showed that metabolic indicators, liver enzymes, cardiovascular comorbidities, and lifestyle factors, such as alcohol consumption, also contributed meaningfully to model predictions. These explainability analyses should be interpreted as an application of established methods for examining model behavior and error patterns rather than as a methodological advance in XAI. These findings highlight the multifactorial clinical risk profile underlying advanced liver stiffness. Furthermore, the findings should be interpreted in light of the study's internal validity limitations, particularly the risk of data leakage introduced by performing imputation before the train-test split, which may have inflated discrimination and classification metrics [22].

Accuracy vs Sensitivity

The principal finding was a clear trade-off between discrimination and sensitivity across algorithms. These results highlight the importance of selecting both model families and operating points that align with clinical priorities, particularly in a screening or triage context, where missed disease may delay further evaluation. In such settings, a model with a modestly lower AUROC but higher sensitivity may be more clinically appropriate than a model with superior

discrimination but a substantial false-negative burden. The performance profile observed across models likely reflects the characteristics of the underlying predictive signals in SLD.

Anthropometric and metabolic predictors consistently dominated global importance rankings and SHAP summaries, reinforcing the central role of adiposity and glycemic dysregulation in fibrosis risk [23-26]. From an implementation perspective, the contrast between high-AUROC, low-sensitivity models and simpler, more sensitive alternatives is clinically relevant. In many care pathways, initial risk assessment must favor sensitivity to avoid false reassurance, with subsequent specialized testing (eg, VCTE) reserved for individuals most likely to benefit. This concern was reinforced by the local SHAP example of a false-negative prediction, in which a participant with confirmed fibrosis was assigned a predicted probability of only 2.2%, illustrating how an apparently interpretable model may still generate clinically misleading reassurance at the individual level. In our analysis, logistic regression and naive Bayes identified more fibrosis cases than complex ensembles. Although these models had higher false-positive rates, such trade-offs may be acceptable in stepped triage strategies, particularly when downstream tests are noninvasive, and the baseline disease prevalence is modest. The use of SMOTE-balanced training data may also have influenced precision-recall behavior when models were evaluated on the naturally imbalanced test set, contributing to lower PPV and higher false-positive counts in models with greater sensitivity. Explainability may add value by illuminating model behavior, identifying influential predictors, and exposing error patterns. However, explainability does not, by itself, establish reliability or clinical usefulness. The local explanation showing that a participant with confirmed fibrosis received a very low predicted probability illustrates that interpretable outputs may still be clinically misleading when probability estimates are poorly calibrated or when minority class cases are missed [27]. The clinical implications are pragmatic. In health care systems seeking to expand noninvasive fibrosis detection, a sensitive, explainable model could serve as a front-end screening tool to prioritize elastography, specialist referral, and targeted lifestyle or metabolic interventions [28-31]. Although the framework suggests potential utility for noninvasive fibrosis risk stratification, its current performance and methodological limitations preclude clinical deployment at this stage. Any implementation-oriented interpretation should await reanalysis using a leakage-free pipeline, sensitivity-focused threshold optimization, probability calibration, decision curve analysis, and external validation. We specifically note the importance of validating and adapting models across geographies and health care settings, where population risk profiles, measurement pathways, and access constraints differ. Embedding explainability throughout the workflow can help ensure transparency, support clinician trust, and identify context-specific biases at an early stage.

Future Research

Future work should prioritize (1) threshold optimization at clinically meaningful operating points, including fixed-sensitivity thresholds appropriate for screening or triage and reporting the corresponding specificity, PPV, negative predictive value (NPV), false-negative rate, and false-positive burden; (2) probability calibration and decision curve analysis to assess the reliability of predicted probabilities and the net clinical benefit; (3) external validation across multisite, multinational cohorts, with subgroup and fairness reporting; (4) head-to-head comparisons with parsimonious fibrosis scores, such as fibrosis-4 index (FIB-4) and the nonalcoholic fatty liver disease (NAFLD) fibrosis score, to quantify incremental value; and (5) prospective evaluation of workflow integration and clinical impact.

Limitations and Strengths

This study has several limitations. The most important limitation is that missing data imputation was performed on the full dataset before the train-test split. This creates a risk of data leakage because information from the held-out test set may have influenced the imputation process used during model development. Consequently, the reported discrimination and classification metrics may be optimistic, and comparisons across models should be interpreted cautiously. This limitation weakens internal validity and means that the results should be viewed as exploratory and hypothesis-generating rather than confirmatory. Future work must use a leakage-free pipeline in which imputation, scaling, encoding, and other learned preprocessing steps are fitted exclusively to the training data and then applied to the held-out test set.

Furthermore, model performance was not fully aligned with the intended screening use case at the evaluated operating points. Although XGBoost achieved the highest AUROC, its sensitivity was only 0.17 at the evaluated threshold, meaning that most true fibrosis cases were missed. This limits its clinical utility as a front-end screening or triage tool in its current form and highlights the danger of relying on discrimination metrics alone. In addition, thresholds were not optimized according to a prespecified clinical objective, and fixed-sensitivity analyses were not performed. Accordingly, the present results should be interpreted as exploratory, and future work should define clinically meaningful operating points based on the intended use case, including high-sensitivity thresholds for screening and the corresponding trade-offs in specificity, PPV, NPV, false-negative rate, false-positive burden, and downstream testing demand.

Calibration and clinical usefulness were also not fully assessed. We did not evaluate calibration plots, the Brier score, the calibration slope or intercept, or decision curve analysis, limiting the interpretation of individual predicted probabilities and the assessment of net clinical benefit. This limitation is particularly important because the local explanation analysis showed that a participant with confirmed fibrosis received a very low predicted probability, whereas a participant without fibrosis received a higher predicted probability. This inconsistency suggests potential miscalibration or minority class prediction failure and reinforces that the model is not ready for individual-level clinical decision-making. Future work should incorporate probability calibration

methods, such as Platt scaling or isotonic regression, and decision curve analysis before considering clinical use.

In addition, the analysis was based on NHANES 2017-2020 data, which may not fully reflect contemporary definitions of SLD or current metabolic risk profiles. This may limit the generalizability of the findings to contemporary populations. Another limitation is that the machine learning models were not directly compared with established clinical fibrosis scores such as FIB-4 or the NAFLD fibrosis score, within the same analytic cohort. Consequently, the incremental value of the machine learning approach over simpler, widely used clinical tools cannot be determined from the present analysis. Future work should include head-to-head comparisons with established fibrosis scores using the same cohort, thresholds, and clinically relevant performance metrics, including sensitivity, specificity, PPV, NPV, calibration, and decision curve analysis.

Model optimization was limited. Hyperparameters were set to default library values, and formal tuning using grid search, random search, Bayesian optimization, or nested cross-validation was not performed. This may have affected comparative model performance and disadvantaged algorithms that are more sensitive to hyperparameter specification. In addition, the use of SMOTE-balanced training data may have influenced calibration, PPV, and false-positive burden when models were evaluated on the naturally imbalanced test set. Future work should evaluate systematic hyperparameter tuning and alternative approaches to class imbalance, such as cost-sensitive learning, adaptive synthetic sampling (ADASYN), or hybrid resampling methods, within a leakage-free pipeline.

Despite these limitations, the study possesses significant strengths in its breadth of features and commitment to transparency. The analysis successfully used 43 diverse clinical, anthropometric, and lifestyle variables to identify key predictors of liver stiffness. To address the identified limitations in future iterations, we will implement a rigorous pipeline in which imputation, scaling, and encoding are fitted strictly to the training data and then applied to the test set. To improve clinical utility, decision thresholds should be optimized specifically for sensitivity to ensure that the model functions effectively as a triage tool. Addressing the miscalibration issues through probability calibration techniques and refining local explanation models will be essential to ensure that individual patient risk assessments are both accurate and trustworthy for bedside decision-making.

Conclusions

This exploratory study applied established machine learning and XAI methods to evaluate the prediction of significant fibrosis in SLD and to characterize model behavior, sensitivity limitations, and influential clinical predictors. The findings must be interpreted cautiously because imputation before the train-test split may have introduced data leakage and inflated performance estimates. The results should therefore be considered hypothesis-generating rather than evidence that the model is ready for clinical use. Future work should prioritize leakage-free preprocessing, sensitivity-focused threshold optimization, probability calibration, decision curve analysis, external validation, and prospective evaluation before clinical application or individual-level risk estimation.

Generative AI Use

The authors used ChatGPT (GPT-5; OpenAI) to assist with language editing and to improve the clarity of selected portions of the manuscript. All generated content was reviewed, revised, and verified by the authors, who take full responsibility for the final manuscript.

Funding

The authors declare that this study received no external funding.

Data Availability

Data generated for this study are available upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Confusion matrices for the 8 machine learning models predicting significant liver fibrosis in the test set.
[PNG File (Portable Network Graphics File), 74 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Shapley additive explanations (SHAP) breakdown plot for the individualized risk assessment for a patient with significant liver fibrosis.
[PNG File (Portable Network Graphics File), 151 KB-Multimedia Appendix 2]

References

1. Åberg F, Jiang ZG, Cortez-Pinto H, Männistö V. Alcohol-associated liver disease-global epidemiology. Hepatology. Dec 1, 2024;80(6):1307-1322. [doi: 10.1097/HEP.0000000000000899] [Medline: 38640041]

2. Kan C, Zhang K, Wang Y, et al. Global burden and future trends of metabolic dysfunction-associated steatotic liver disease: 1990-2021 to 2045. Ann Hepatol. 2025;30(2):101898. [doi: 10.1016/j.aohep.2025.101898] [Medline: 40057034]
3. European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO clinical practice guidelines on the management of metabolic dysfunction-associated steatotic liver disease (MASLD). J Hepatol. Sep 2024;81(3):492-542. [doi: 10.1016/j.jhep.2024.04.031] [Medline: 38851997]
4. Angulo P, Kleiner DE, Dam-Larsen S, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology. Aug 2015;149(2):389-397.e10. [doi: 10.1053/j.gastro.2015.04.043] [Medline: 25935633]
5. Meng F, Zheng Y, Zhang Q, et al. Noninvasive evaluation of liver fibrosis using real-time tissue elastography and transient elastography (FibroScan). J Ultrasound Med. Mar 2015;34(3):403-410. [doi: 10.7863/ultra.34.3.403] [Medline: 25715361]
6. Alkhouri N, Cheuk-Fung Yip T, Castera L, et al. ALADDIN: a machine learning approach to enhance the prediction of significant fibrosis or higher in metabolic dysfunction-associated steatotic liver disease. Am J Gastroenterol. Feb 1, 2026;121(2):362-374. [doi: 10.14309/ajg.0000000000003432] [Medline: 40146016]
7. Sterling RK, Duarte-Rojo A, Patel K, et al. AASLD practice guideline on imaging-based noninvasive liver disease assessment of hepatic fibrosis and steatosis. Hepatology. Feb 1, 2025;81(2):672-724. [doi: 10.1097/HEP.0000000000000843] [Medline: 38489518]
8. Hudson D, Afzaal T, Bualbanat H, et al. Modernizing metabolic dysfunction-associated steatotic liver disease diagnostics: the progressive shift from liver biopsy to noninvasive techniques. Ther Adv Gastroenterol. 2024;17:17562848241276334. [doi: 10.1177/17562848241276334] [Medline: 39553445]
9. Boursier J, de Ledinghen V, Zarski JP, et al. Comparison of eight diagnostic algorithms for liver fibrosis in hepatitis C: new algorithms are more precise and entirely noninvasive. Hepatology. Jan 2012;55(1):58-67. [doi: 10.1002/hep.24654] [Medline: 21898504]
10. Liu S, Chen X, Jiang X, et al. LiverRisk score: an accurate, cost-effective tool to predict fibrosis, liver-related, and diabetes-related mortality in the general population. Med. Jun 14, 2024;5(6):570-582.e4. [doi: 10.1016/j.medj.2024.03.003] [Medline: 38554711]
11. McPherson S, Hardy T, Dufour JF, et al. Age as a confounding factor for the accurate non-invasive diagnosis of advanced NAFLD fibrosis. Am J Gastroenterol. May 2017;112(5):740-751. [doi: 10.1038/ajg.2016.453] [Medline: 27725647]
12. Ekstedt M, Hagström H, Nasr P, et al. Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up. Hepatology. May 2015;61(5):1547-1554. [doi: 10.1002/hep.27368] [Medline: 25125077]
13. Decharatanachart P, Chaiteerakij R, Tiyarattanachai T, Treeprasertsuk S. Application of artificial intelligence in non-alcoholic fatty liver disease and liver fibrosis: a systematic review and meta-analysis. Ther Adv Gastroenterol. 2021;14:17562848211062807. [doi: 10.1177/17562848211062807] [Medline: 34987607]
14. Singh Y, Hathaway QA, Vera-Garcia DV, et al. Artificial intelligence-based agents in chronic liver disease: transforming diagnostic and therapeutic workflows through clinical decision-making. npj Gut Liver. 2025;2(1):34. [doi: 10.1038/s44355-025-00049-5]
15. Njei B, Al-Ajlouni YA, Lemos SY, et al. AI-based models for risk prediction in MASLD: a systematic review. Dig Dis Sci. May 2026;71(5):1987-2014. [doi: 10.1007/s10620-025-09499-6] [Medline: 41231418]
16. Dimala CA, Al-Ajlouni YA, Nso N, Njei B. Artificial intelligence-enabled electrocardiography for risk prediction in chronic liver disease: a systematic review. Int J Cardiol. Sep 23, 2025;443:133926. [doi: 10.1016/j.ijcard.2025.133926] [Medline: 40997956]
17. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Oct 29, 2019;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
18. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. Mar 2019;28(3):231-237. [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
19. Collins GS, Moons KG, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
20. Zhu T, Lin Y, Liu Y. Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognit. Dec 2017;72:327-340. [doi: 10.1016/j.patcog.2017.07.024]
21. Ethics review board approval. CDC National Center for Health Statistics. URL: https://www.cdc.gov/nchs/nhanes/about/erb.html [Accessed 2026-11-15]

22. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. [doi: 10.1016/j.patter.2023.100804] [Medline: 37720327]
23. Julián MT, Ballesta S, Pera G, et al. Abdominal obesity and dsyglycemia are risk factors for liver fibrosis progression in NAFLD subjects: a population-based study. Front Endocrinol (Lausanne). 2023;13:1051958. [doi: 10.3389/fendo.2022.1051958] [Medline: 36714592]
24. Abdennour M, Reggio S, Le Naour G, et al. Association of adipose tissue and liver fibrosis with tissue stiffness in morbid obesity: links with diabetes and BMI loss after gastric bypass. J Clin Endocrinol Metab. Mar 2014;99(3):898-907. [doi: 10.1210/jc.2013-3253] [Medline: 24423338]
25. Gaggini M, Carli F, Rosso C, et al. Altered metabolic profile and adipocyte insulin resistance mark severe liver fibrosis in patients with chronic liver disease. Int J Mol Sci. Dec 16, 2019;20(24):6333. [doi: 10.3390/ijms20246333] [Medline: 31888144]
26. Rusu E, Jinga M, Cursaru R, et al. Adipose tissue dysfunction and hepatic steatosis in new-onset diabetes. Diabetology. 2025;6(7):70. [doi: 10.3390/diabetology6070070]
27. McGrath MJ, Lack O, Tisch J, Duenser A. Measuring trust in artificial intelligence: validation of an established scale and its short form. Front Artif Intell. 2025;8:1582880. [doi: 10.3389/frai.2025.1582880] [Medline: 40416549]
28. Njei B, Osta E, Njei N, Al-Ajlouni YA, Lim JK. An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis. Sci Rep. Apr 13, 2024;14(1):8589. [doi: 10.1038/s41598-024-59183-4] [Medline: 38615137]
29. Njei B, Osta E, Ilagan-Ying YC, et al. 534: Fibrox: an explainable AI model for accurate prediction of advanced liver fibrosis and cardiovascular mortality in MASLD. Gastroenterology. May 2025;169(1):S131-S132. [doi: 10.1016/S0016-5085(25)01246-6]
30. Yip TC, Lee HW, Lin H, et al. Prognostic performance of the two-step clinical care pathway in metabolic dysfunction-associated steatotic liver disease. J Hepatol. Aug 2025;83(2):304-314. [doi: 10.1016/j.jhep.2025.01.014] [Medline: 39863175]
31. Srivastava A, Gailer R, Tanwar S, et al. Prospective evaluation of a primary care referral pathway for patients with non-alcoholic fatty liver disease. J Hepatol. Aug 2019;71(2):371-378. [doi: 10.1016/j.jhep.2019.03.033] [Medline: 30965069]

Abbreviations

ADASYN: adaptive synthetic sampling
ALP: alkaline phosphatase
ALT: alanine aminotransferase
AST: aspartate aminotransferase
AUROC: area under the receiver operating characteristic curve
FIB-4: fibrosis-4 index
GGT: gamma-glutamyl transferase
HbA_{1c}: hemoglobin A_{1c}
HDL: high-density lipoprotein
HEI: Healthy Eating Index
ICE: individual conditional expectation
LDL: low-density lipoprotein
NAFLD: nonalcoholic fatty liver disease
NHANES: National Health and Nutrition Examination Survey
NPV: negative predictive value
PPV: positive predictive value
SHAP: Shapley additive explanations
SLD: steatotic liver disease
SMOTE: synthetic minority oversampling technique
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence
VCTE: vibration-controlled transient elastography
XAI: explainable AI
XGBoost: Extreme Gradient Boosting

Edited by Edward Mensah; peer-reviewed by Mahmoud Badee Rokaya Mahmoud, Yijue Zhang; submitted 01.Jan.2026; final revised version received 08.Jun.2026; accepted 19.Jul.2026; published 18.Nov.2026

Please cite as:
Njei B, Kanmounye US
Explainable Machine Learning Models for Predicting Significant Fibrosis in Steatotic Liver Disease: Model Development Study
Online J Public Health Inform 2026;18:e90677
URL: https://ojphi.jmir.org/2026/1/e90677
doi: 10.2196/90677

© Basile Njei, Ulrick Sidney Kanmounye. Originally published in the Online Journal of Public Health Informatics (https://ojphi.jmir.org/), 18.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Online Journal of Public Health Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://ojphi.jmir.org/, as well as this copyright and license information must be included.