

<Conference abbreviation>
<Conference Series name> <Volume number and Year> <DOI Number>

Explainable Machine Learning for Non-Invasive Prediction of Liver Fibrosis in NAFLD Using Routine Clinical Biomarkers

Arpit Chhabra¹, Namrata Dhanda², Pranav Pant³, Ramsha Azra⁴, *Vandana Yadav⁵

^{1,3,4}Amity School of Engineering and Technology, Amity University, Lucknow, India

²Department of CSE, SAREC, Mirzapur, U.P., India

⁵Brainware University, Kolkata, India

Abstract. Non-alcoholic fatty liver disease (NAFLD) is a leading cause of chronic liver disease worldwide and is closely tied to metabolic dysfunction. Prognosis is governed largely by liver fibrosis, which can advance to cirrhosis and hepatocellular carcinoma. Biopsy remains the reference standard for staging but is unsuitable for large-scale screening, so accurate non-invasive alternatives are needed. This paper presents an interpretable machine learning framework for fibrosis prediction from routinely available clinical biomarkers. Logistic Regression, Random Forest, XGBoost and LightGBM were evaluated on a biopsy-confirmed dataset using accuracy, F1-score, sensitivity, specificity and ROC-AUC, benchmarked against the FIB-4 index. Threshold optimisation allowed the sensitivity-specificity trade-off to be tuned to different clinical scenarios, and SHAP (SHapley Additive exPlanations) exposed the contribution of individual features. The results indicate that machine learning can offer accurate, interpretable and scalable support for early fibrosis detection, particularly in resource-constrained settings.

Keywords. NAFLD; liver fibrosis; machine learning; clinical biomarkers; FIB-4; ROC-AUC; explainable AI; SHAP.

1. INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is among the most common chronic liver disorders worldwide, closely associated with obesity, type 2 diabetes and metabolic syndrome [1,2]; rising prevalence has made it a public health concern [3]. It ranges from hepatic steatosis to steatohepatitis, fibrosis, cirrhosis, and hepatocellular carcinoma [4]. Fibrosis is the strongest predictor of outcome, as advanced disease has a higher risk of liver failure and death [5]. Biopsy is still the standard for staging, but it is invasive and costly, which limits its use. Non-invasive scores like the AST/ALT ratio, APRI, the NAFLD Fibrosis Score, and FIB-4 are more affordable, but their accuracy varies between populations and is often unclear in intermediate stages [6–9].

Machine learning can find connections between clinical variables that traditional statistics struggle with [10]. However, two issues hinder its use. Many models provide accuracy figures without clarifying how the variables influence predictions. Additionally, many are validated on small groups. This study suggests an interpretable framework. It includes Logistic Regression, Random Forest, XGBoost, and LightGBM, all trained on a biopsy-confirmed NAFLD dataset and compared with FIB-4. The results are interpreted using SHAP.

2. RELATED WORK

Laboratory scores like the AST/ALT ratio, APRI, the NAFLD Fibrosis Score, and FIB-4 assess fibrosis severity using routine biochemistry [11]. They are inexpensive to calculate, but their performance varies among patient groups, and they often produce unclear results at intermediate stages. Logistic regression is one of the easiest classifiers to understand, as its probability estimates relate directly to individual predictors [12]. However, it does not capture interactions well. Random Forest, XGBoost [16], and LightGBM effectively model non-linear relationships [13] and often outperform traditional scores while decreasing unclear classifications [14, 15]. Their main drawback is a lack of transparency. This study

aims to combine strong predictive power with clear explanations, using FIB-4 as a benchmark.

3. METHODOLOGY

3.1. Dataset and Preprocessing

The data set is from the LiveFbr database and consists of biopsy-proven NAFLD patients; hence, the labels are generated using the reference standard. The objective is binary classification where 0 represents early fibrosis stage (S0-S2), and 1 represents advanced fibrosis stage (S3-S4). This is important due to the increased likelihood of disease progression in cases with advanced fibrosis. The features are demographics and biochemical measurements such as age, BMI, ALT, AST, platelet count, and glucose (Table 1). Missing values were imputed, continuous variables standardised with StandardScaler, and the data split into training and testing subsets.

Table 1. Class distribution of the dataset.

Dataset	Early fibrosis (S0–S2)	Advanced fibrosis (S3–S4)	Total samples	% advanced
Discovery set	391	149	540	27.59
Validation set	116	31	147	21.09

3.2. Baseline, Models and Evaluation

The FIB-4 index was implemented as the baseline, computed as

$$FIB-4 = (Age \times AST) / (Platelets [10^9/L] \times \sqrt{ALT})$$

FIB-4 is widely used because it is simple and needs only routine biomarkers. Four algorithms were then investigated: Logistic Regression, Random Forest, XGBoost [16] and LightGBM, each trained on the same data under identical conditions. Accuracy, precision, recall, F1-score, sensitivity, specificity and ROC-AUC [17] were computed for every model, and SHAP [18] estimated each feature’s influence on individual predictions.

4. RESULTS

The four methods worked reliably on the validation set (Table 2). The tree-based ensemble techniques yielded the best results and logistic regression remained in the competition, which means that biomarkers contain sufficient information to distinguish early and advanced stages of fibrosis.

Table 2. Performance Evaluation of Models

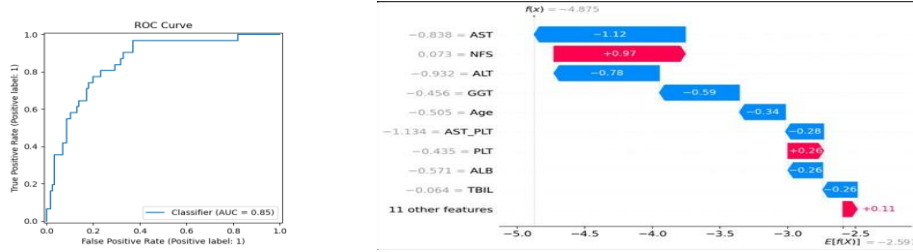
Model	ACC	PRE	AUROC	REC	F1	SPEC	SENS
Logistic Regression	0.8299	0.6154	0.8732	0.5161	0.5614	0.9138	0.5161
Random Forest	0.8299	0.6154	0.8760	0.5161	0.5614	0.9138	0.5161
XGBoost	0.8299	0.5938	0.8526	0.6129	0.6032	0.8879	0.6129
LightGBM	0.8163	0.5625	0.8529	0.5806	0.5714	0.8793	0.5806
FIB-4	0.7959	0.6000	0.8493	0.0968	0.1667	0.9828	0.0968

Table 3 depicts a comparison between the selected model and FIB-4 at the selected operating point. While FIB-4 maintains high specificity, its sensitivity remains very low; hence it fails to detect most positive cases. The selected model is much more balanced than FIB-4.

Table 3. Machine learning model compared with FIB-4.

Classifier	ACC	AUROC	PRE	REC	F1	SPEC	SENS
ML model	0.8981	0.9289	0.9000	0.6667	0.7660	0.9753	0.6667
FIB-4	0.7500	0.5203	0.0000	0.0000	0.0000	1.0000	0.0000

ROC curves in Figure 1 show that there is a sharp increase in the curve at low false positive rates; therefore, high-risk patients will be identified with fewer false positives. SHAP attributions are shown for the representative prediction in Figure 2, where the significant contributors are AST, NAFLD, ALT, GGT, and age.



Figs. 1, 2. Left: ROC curve of the proposed model. Right: SHAP attribution for a representative prediction.

The confusion matrix (Table 4) shows most early fibrosis cases classified correctly, though some advanced cases are missed, as is common where feature distributions overlap.

Table 4. Confusion matrix.

	Predicted 0	Predicted 1
Actual 0	102	14
Actual 1	13	18

Threshold analysis (Table 5) exposes the trade-off: at 0.3 recall reaches 0.8889, suiting screening, at the cost of precision; at 0.5 accuracy and F1 are highest; at 0.7 precision and specificity reach 1.0000 but recall falls to 0.4444. Threshold choice should follow clinical priority [19].

Table 5. Effect of the decision threshold on model performance.

Threshold	Accuracy	Precision	AUROC	Recall	F1	Specificity	Sensitivity
0.3	0.7963	0.5581	0.9289	0.8889	0.6857	0.7654	0.8889
0.5	0.8981	0.9000	0.9289	0.6667	0.7660	0.9753	0.6667
0.7	0.8611	1.0000	0.9289	0.4444	0.6154	1.0000	0.4444

5. DISCUSSION

Use of routine biomarkers in non-invasive prediction of fibrosis proved effective. Ensemble methods demonstrated the highest efficiency due to their ability to model interactions among patient's characteristics and use of SHAP allowed every prediction to be explainable. In comparison with FIB-4 the proposed framework achieved higher discrimination (ROC-AUC 0.8529 compared to 0.8493) and accuracy; the difference is slight but significant in the context of mass screening and in accordance with previous researches that have achieved ROC-AUC between 0.82 and 0.89. Prediction based on SHAP scores conforms to the

clinical intuition, since AST/PLT ratio, age, BMI and fasting glucose were found to have the highest impact. The web-based prediction system providing real-time predictions allows using the model for clinical practice [20]. There are two drawbacks: lack of representativeness of the data set in terms of NAFLD patients and oversimplification of the disease through binary classification.

6. CONCLUSION

In this paper, we have provided an interpretable machine learning model to predict non-invasive liver fibrosis based on biomarkers. Ensemble models performed better than traditional statistical models and even slightly better than FIB-4, while SHAP explained the reasons for each prediction. This framework is very useful in developing countries like Southeast Asia, where there is no access to advanced diagnostics tests.

REFERENCES

- [1] B. J. Perumpail, et al., ‘Clinical epidemiology and disease burden of NAFLD’, *World J. Gastroenterol.*, vol. 23, pp. 8263–8276, 2017.
- [2] Z. M. Younossi, et al., ‘The economic and clinical burden of NAFLD in the United States and Europe’, *Hepatology*, vol. 64, pp. 1577–1586, 2016.
- [3] Z. M. Younossi, et al., ‘Global epidemiology of NAFLD’, *Hepatology*, vol. 64, pp. 73–84, 2016.
- [4] S. A. Polyzos, J. Kountouras, C. S. Mantzoros, ‘Obesity and NAFLD: from pathophysiology to therapeutics’, *Metabolism*, vol. 92, pp. 82–97, 2019.
- [5] F. Nassir, ‘NAFLD: mechanisms, treatments, and biomarkers’, *Biomolecules*, vol. 12, art. 824, 2022.
- [6] M. Charlton, ‘Non-alcoholic fatty liver disease: current understanding and future impact’, *Clin. Gastroenterol. Hepatol.*, vol. 2, pp. 1048–1058, 2004.
- [7] S. Y. Park, et al., ‘AST/ALT ratio and platelet counts as predictors of fibrosis’, *Korean J. Gastroenterol.*, vol. 43, pp. 246–251, 2004.
- [8] R. K. Sterling, et al., ‘A simple non-invasive index to predict significant fibrosis in HIV/HCV coinfection’, *Hepatology*, vol. 43, pp. 1317–1325, 2006.
- [9] A. Robinson, R. J. Wong, ‘Noninvasive methods for evaluating hepatic fibrosis in NAFLD’, *Clin. Liver Dis.*, vol. 15, no. 4, pp. 157–161, 2020.
- [10] A. Esteva, et al., ‘A guide to deep learning in healthcare’, *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019.
- [11] M. S. Alam, et al., ‘Comparison of FIB-4, APRI and AST/ALT ratio with FibroScan in NAFLD’, *Egypt. Liver J.*, vol. 15, art. 38, 2025.
- [12] C. Sang, et al., ‘Diagnosis of fibrosis using blood markers and logistic regression in NAFLD’, *Front. Med.*, vol. 8, art. 637652, 2021.
- [13] H. Zamanian, et al., ‘AI techniques for NAFLD diagnosis: a systematic review’, *Comput. Methods Programs Biomed.*, vol. 244, art. 107932, 2024.
- [14] G. Feng, et al., ‘Machine learning outperforms fibrosis markers in biopsy-confirmed NAFLD’, *J. Hepatobiliary Pancreat. Sci.*, vol. 28, pp. 593–603, 2021.
- [15] Z. Zhang, et al., ‘XGBoost-based explainable model for predicting significant fibrosis in autoimmune hepatitis’, *QJM*, vol. 119, pp. 34–41, 2026.
- [16] T. Chen, C. Guestrin, ‘XGBoost: a scalable tree boosting system’, *Proc. KDD*, 2016.
- [17] K. Hajian-Tilaki, ‘ROC curve analysis for medical diagnostic test evaluation’, *Caspian J. Intern. Med.*, vol. 4, no. 2, pp. 627–635, 2013.
- [18] S. M. Lundberg, S. I. Lee, ‘A unified approach to interpreting model predictions’, *Proc. NeurIPS*, 2017.