Detecting Liver Diseases Using Advanced Machine Learning And Explainable AI

[1]Srikanth Kandula, [2]Harshavardhan Veera, [3]Harshitha Jammula, [4]Likhitha Upperla, [5]Yashwanth Kumar Burada

[1]*Assistant Professor*, [2,3,4,5] Student, *Department of Computer Science and Engineering*, *Dhanekula Institute of Engineering and Technology, Ganguru, AP, India*

[1]kandulaa1212@gmail.com, [2]harshavardhanveera2003@gmail.com, [3]harshithajammula11@gmail.com, [4]likhithaupperla@gmail.com, [5]yashwanthkumar1629@gmail.com

***Abstract*—Liver diseases, including hepatitis, fibrosis, cirrhosis, and liver failure, are serious health concerns that require early detection for timely treatment. This study presents a machine learning-based approach to identifying liver diseases by analyzing key clinical biomarkers such as bilirubin, SGOT/AST, albumin, and platelets—with additional bioclinical indicators. The model utilizes advanced algorithms like XGBoost, LightGBM, and Multilayer Perceptron (MLP) to improve accuracy in detection. To enhance transparency, Explainable AI (XAI) techniques, specifically SHAP, provide insights into how each biomarker contributes to the model's predictions. By making AI-driven diagnostics more interpretable, this approach helps build trust among healthcare professionals and supports better clinical decision-making. While the model does not suggest treatments, it plays a crucial role in early diagnosis, enabling doctors to identify liver diseases more efficiently and improve patient care.**

***Index Terms*—Liver disease detection, Hepatitis, Fibrosis, Cirrhosis, Machine learning, Explainable AI (XAI), SHAP, Clinical biomarkers, XGBoost, LightGBM, Multilayer Perceptron (MLP), Early diagnosis, Medical diagnostics, Predictive modeling, Clinical decision support.**

I. Introduction

The liver is essential for detoxification, nutrient metabolism, and protein synthesis [3]. Its multifaceted functions, however, render it vulnerable to a spectrum of disorders—from hepatitis and fibrosis to cirrhosis and liver failure—primarily driven by viral infections, chronic alcohol abuse, obesity, and metabolic imbalances [1]. Traditional diagnostic methods, including blood panels, imaging, and invasive biopsies, are widely used yet often fail to capture early, subtle changes in hepatic function [4], [5].

Our study presents a novel diagnostic framework that integrates advanced machine learning with key laboratory biomarkers—such as bilirubin, SGOT/AST, albumin, and platelet counts—alongside clinical examination data to address these challenges. In our approach, FibroScan outcomes are dichotomized as normal or abnormal, providing an efficient, non-invasive overview of liver health [6]. We employ state-of-the-art algorithms including XGB, LightGBM, and a Multilayer Perceptron to deliver robust predictive performance. Additionally, the use of SHAP (Shapley additive explanations) enhances model interpretability by elucidating the contribution of each feature, in line with recent advances in explainable AI for liver cirrhosis biomarkers [2], [9].

Our ultimate goal is to develop an AI-powered tool that improves early detection and accurate staging of liver disease, thereby enabling timely interventions and better patient outcomes [7]. This approach addresses the limitations of conventional diagnostics and paves the way for a more personalized and transparent assessment of hepatic health [8].

II. Literature Survey

The global burden of liver disease remains a critical challenge, with conditions ranging from hepatitis and non-alcoholic fatty liver disease (NAFLD) to fibrosis and cirrhosis affecting a large population worldwide [1]. Traditional diagnostic methods—such as blood panels, imaging, and biopsies—may overlook subtle pathophysiological changes, thereby delaying early intervention and effective management [3], [4]. In response, researchers have increasingly turned to machine learning (ML) and explainable AI (XAI) to enhance both predictive accuracy and interpretability.

A key study, "Explainable AI for Enhanced Interpretation of Liver Cirrhosis Biomarkers," demonstrates how combining Extreme Gradient Boosting (XGB) with Shapley additive explanations (SHAP) can illuminate the relative importance of specific biomarkers [2]. Although this base paper employed the Mayo Clinic Primary Biliary Cirrhosis (PBC) dataset, many recent efforts leverage alternative publicly available repositories to ensure reproducibility and generalizability. For instance, four distinct liver disease datasets have been sourced from references [17]–[20], allowing researchers to validate their models under various clinical conditions without providing a definitive "diagnosis" but rather a predictive assessment of potential liver disease progression.

Comparative analyses of ML models, including XGB, LightGBM, and neural networks, indicate that each algorithm has unique strengths in handling imbalanced data and complex feature interactions [8], [9], [11]. Studies show that advanced ensemble methods and hyperparameter tuning can significantly improve detection rates while maintaining robust interpretability through XAI techniques [12]. The performance metrics typically reported—accuracy, precision, recall, and F1-score—are of high clinical relevance, as small gains in these metrics can translate into more timely identification of liver abnormalities and, consequently, better patient outcomes.

Crucially, SHAP-based methods have proven especially valuable for clarifying how each feature (e.g., albumin, platelet count, AST/ALT ratio) influences a model's output. This transparency fosters clinician trust and ensures that AI-powered tools can be

integrated more seamlessly into patient care. Furthermore, because datasets in [17]–[20] are openly accessible, researchers can replicate and refine these models, promoting standardization and extending applicability to broader clinical settings.

Taken together, these advances underscore the transformative potential of XAI in hepatology. By bridging the gap between black-box ML models and practical clinical insight, modern frameworks enable earlier detection and more accurate staging of liver disease, ultimately supporting timely intervention and improving patient outcomes.

Abbreviations and Acronyms

The following abbreviations and acronyms are used in this paper:

- AI: Artificial Intelligence
- XAI: Explainable Artificial Intelligence
- SHAP: Shapley Additive Explanations
- XGBM: Extreme Gradient Boosting Model
- MLP: Multilayer Perceptron
- NAFLD: Non-Alcoholic Fatty Liver Disease
- PBC: Primary Biliary Cirrhosis
- WHO: World Health Organization
- APRI: AST to Platelet Ratio Index
- FIB-4: Fibrosis-4 Index

III. Proposed Methodology

Data for this study were obtained from four publicly available liver disease datasets [17]–[20] and integrated into a single unified dataset. This consolidated dataset comprises both original clinical features—such as Total Bilirubin, Albumin, Aspartate Aminotransferase, Alanine Aminotransferase, and Platelet Count—and derived metrics, including the FIB-4 score, AFLD_Indicator, and various ratio-based values computed from the biomarkers. The final integrated dataset includes the following features: Age, Gender, Total Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, AST/ALT Ratio, Albumin, Total Proteins, Prothrombin Time, Platelets, Albumin Globulin Ratio, Ascites, Liver Firmness, AFLD_Indicator, FIB-4 Score, and Class.The final integrated dataset contains additional features such as Total Proteins and Prothrombin Time, as well as clinical assessments like Ascites and Liver Firmness, and a target variable (Class) denoting liver disease stage

Data Acquisition and Preprocessing

Data cleaning procedures involve handling missing values via imputation or exclusion, mapping categorical variables (e.g., converting "Male" to 1 and "Female" to 0 for Gender; "Absent" to 0 and "Present" to 1 for Ascites, Liver Firmness, and AFLD_Indicator), and aligning feature names across datasets. For certain skewed variables such as Alkaline Phosphatase and Alanine Aminotransferase, a logarithmic transformation is applied to stabilize variance. Additionally, winsorization is performed on features like Albumin and Total Proteins to mitigate the effect of outliers. Finally, all continuous features are normalized using a StandardScaler to ensure that each variable contributes proportionately to the model training. The dataset is then partitioned into training, validation, and test sets using stratified sampling to preserve the distribution of liver disease classes.

Machine Learning Framework and Explainability

Our machine learning framework employs multiple algorithms to capture complex interactions among liver biomarkers. We develop models using XGB, LightGBM, and a Multilayer Perceptron (MLP). Hyperparameter tuning is conducted using RandomizedSearchCV with cross-validation on the tree-based models to optimize parameters such as learning rate, maximum depth, number of estimators, and regularization terms, ensuring that the models generalize well without overfitting. The individually tuned models are then integrated into a soft-voting ensemble classifier, which aggregates the probabilistic predictions of the base models to enhance overall performance. To increase the interpretability of our predictions, we incorporate Shapley additive explanations (SHAP).

For XGB and LightGBM, SHAP's TreeExplainer is utilized to calculate feature contributions, whereas for the MLP, KernelExplainer is applied. The resulting SHAP values are aggregated and visualized as horizontal stacked bar charts, clearly illustrating the impact of key features—such as bilirubin, albumin, and platelets—on the predicted liver disease stage. This combination of ensemble learning and explainability not only boosts predictive accuracy but also provides transparent insights that can be readily understood by clinicians.

Units

All laboratory measurements in this study are reported using SI units as the primary standard. When necessary, English units are provided in parentheses to facilitate clarity. The specific measurements used in our dataset are as follows:

- Alanine Aminotransferase (ALT): U/L
- Aspartate Aminotransferase (AST): U/L
- Alkaline Phosphatase (ALP): U/L
- Total Bilirubin: mg/dL
- Albumin: g/dL
- Total Proteins: g/dL
- Prothrombin Time: sec
- Platelets: $\times 10^3/\mu L$

All continuous values are formatted with a leading zero (e.g., 0.25 mg/dL) to avoid ambiguity. Where necessary, supplementary English units are provided in parentheses for additional context. This uniform approach to reporting units ensures that all data

inputs into our machine learning models are accurately represented and comparable, thereby enhancing the reliability and reproducibility of our predictive framework.

Equations

The proposed framework employs several key equations to derive clinically relevant features and to evaluate model performance. These equations are formatted in Times New Roman, centered, and numbered consecutively with their numbers positioned flush right, in compliance with the prescribed specifications.

Standardization ensures that all numerical features have a mean of zero and unit variance. This transformation is performed using the following formula:

$$\text{scaling\_formula} = (X - \text{mean}) / \text{std}. \quad \text{(Eq. 1)}.$$

A logarithmic transformation is applied to reduce skewness in the data. Since logarithms are undefined for zero and negative values, a small offset of 1 is added:

$$\text{Log Transformation} = \log(X+1). \quad \text{(Eq. 2)}.$$

To reduce the influence of extreme values, Winsorization caps data points below the 5th percentile and above the 95th percentile. The transformation is given by:

$$\text{winsorization\_formula} = \text{if } X < P5 \text{ then } X = P5 \text{ elif } X > P95 \text{ then } X = P95 \text{ else } X. \quad \text{(Eq. 3)}.$$

The FIB-4 score, a widely used index for assessing liver fibrosis, is calculated as follows:

$$\text{FIB-4} = (\text{Age} \times \text{AST}) / (\text{Platelets} \times \sqrt{\text{ALT}}). \quad \text{(Eq. 4)}.$$

Here, Age is in years, AST (Aspartate Aminotransferase) and ALT (Alanine Aminotransferase) are measured in U/L, and Platelet count is in $\times 10^3/\mu L$.

The AST/ALT ratio is computed by dividing the AST value by the ALT value:

$$\text{AST/ALT Ratio} = \text{AST} / \text{ALT}. \quad \text{(Eq. 5)}.$$

The Albumin/Globulin (A/G) ratio is derived as:

$$\text{A/G Ratio} = \text{Albumin} / (\text{Total Proteins} - \text{Albumin}). \quad \text{(Eq. 6)}$$

For evaluating model performance, standard metrics are defined as follows. Accuracy is given by:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN). \quad \text{(Eq. 7)}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.
Precision is defined as:

$$\text{Precision} = TP / (TP + FP). \quad \text{(Eq. 8)}$$

Recall (or Sensitivity) is calculated by:

$$\text{Recall} = TP / (TP + FN). \quad \text{(Eq. 9)}$$

The F1-score, which is the harmonic mean of precision and recall, is computed as:

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}). \quad \text{(Eq. 10)}$$

These formulations facilitate the transformation of raw clinical data into meaningful diagnostic indices and provide a standardized basis for evaluating the predictive performance of our machine learning models.

IV. SOFTWARE AND HARDWARE SPECIFICATIONS

Software Specifications:

- Operating System: Windows 10/11
- Programming Language: Python 3.7+

Development Tools:

- Jupyter Notebook
- Visual Studio Code

Software Stack:

- NumPy: 1.26.4
- Pandas: 2.2.2
- XGBoost: 2.1.3
- LightGBM: 4.5.0
- Joblib: 1.4.2

- SHAP: 0.46.0
- Matplotlib: 3.9.2
- Seaborn: 0.13.2
- SciPy: 1.13.1
- Scikit-Learn: 1.5.1
- Flask: 3.0.3
- Flask-CORS: 5.0.0

Front-End Technologies:

- HTML
- CSS
- JavaScript
- Access Method: Web browser

Hardware Specifications:

- Processor: Intel Core i5 or higher
- RAM: 16 GB or more
- Storage: 512 GB SSD
- Graphics: NVIDIA GeForce GTX 1050 or higher
- Display: Full HD monitor
- Input Devices: Keyboard, Mouse
- Network: Stable internet connection required for network-based functionalities

V. SYSTEM IMPLEMENTATION

The system implementation operationalizes our predictive framework through a structured, Python-based pipeline.

Algorithm and Workflow

Training Phase

The practical implementation of our diagnostic framework begins by reading the unified dataset from publicly available sources [17]–[20] using Pandas. Initial preprocessing involves mapping categorical variables (e.g., converting Gender to numerical values, and Ascites, Liver Firmness, AFLD_Indicator from text to binary), log-transforming skewed features such as Alanine Aminotransferase and Alkaline Phosphatase [3], [4], [5], and applying winsorization to Albumin and Total Proteins to mitigate outliers. The dataset is first partitioned into training, validation, and test subsets using stratified sampling to preserve class balance. Standardization is then applied using Scikit-Learn's StandardScaler, where the scaler is fitted on the training data and subsequently used to transform the validation and test sets, ensuring consistency without data leakage.

For model training, three primary classifiers—XGB, LightGBM, and a Multilayer Perceptron (MLP)—are employed. Hyperparameter tuning for the tree-based models (XGB and LightGBM) is conducted using RandomizedSearchCV, optimizing parameters such as learning rate, maximum tree depth, number of estimators, subsample ratios, and regularization factors [9], [12]. The MLP is configured with a multi-layer architecture and relies on early stopping and adaptive learning rate strategies for convergence. After individual training, the tuned models are integrated via scikit-learn's VotingClassifier with soft voting, where each model's probabilistic output is aggregated into a final, robust prediction.

Prediction Phase

Once XGB, LightGBM, and MLP have been trained and combined, the ensemble model—along with all preprocessing artifacts—is saved using Joblib for future deployment. This includes the StandardScaler and parameters for log transformation and winsorization. Storing these components ensures reproducibility, as the entire pipeline (from data normalization to model inference) can be consistently applied in production settings. Similar practices have been highlighted in machine learning–based decision support systems [11], [12].

In real-time deployment, new patient data is passed through the same transformations and then fed into the loaded ensemble. The system outputs a probabilistic estimate for each liver disease stage, which is translated into a final classification. By maintaining a unified pipeline, any updates or retraining efforts can be seamlessly integrated, facilitating incremental improvements and clinical decision support.

Explainability Integration

To enhance transparency and trust in the predictive outcomes, SHAP (Shapley Additive Explanations) is integrated into the workflow. SHAP's TreeExplainer is used for XGB and LightGBM, while KernelExplainer is applied to the MLP due to its non-tree-based architecture. Similar approaches to explainable AI in healthcare contexts have shown improved clinician acceptance and insight into model decisions [2], [10], [13], [14], [16]. The system computes SHAP values on the test set and visualizes them as horizontal stacked bar charts, illustrating how each key feature—such as Total Bilirubin, Albumin, and Platelet Count—influences the final prediction. These visualizations help clinicians understand the rationale behind each diagnostic outcome without having to interpret raw model internals.

System Design Diagrams

The system architecture is further detailed through design diagrams that map the complete data flow with system architecture and flow charts.

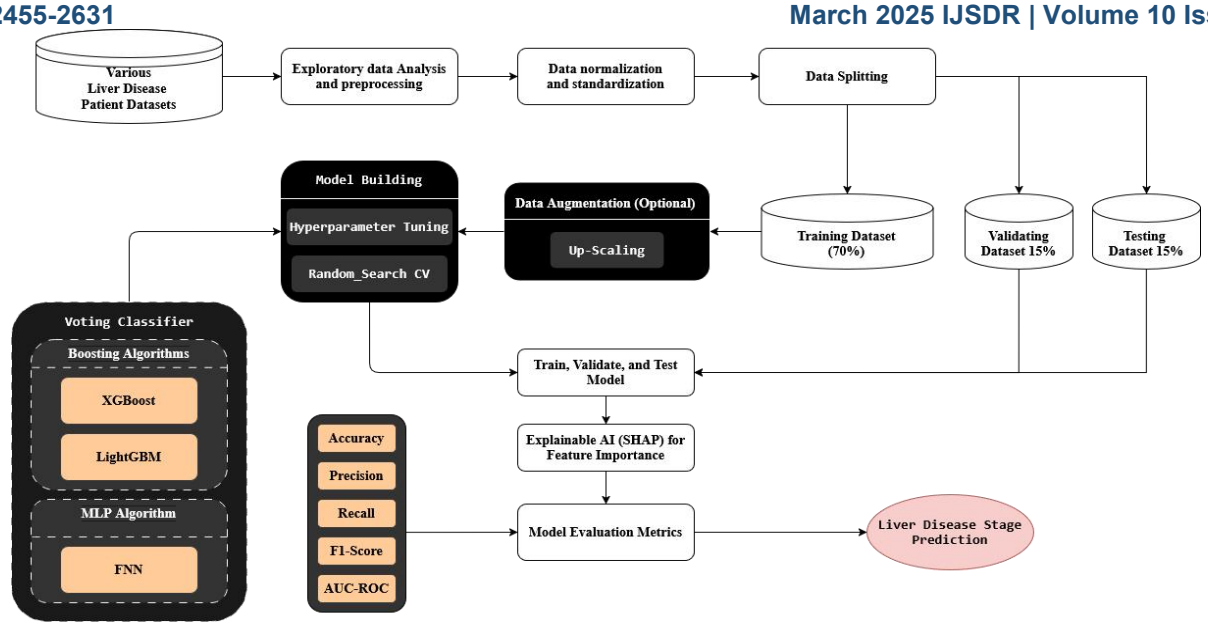
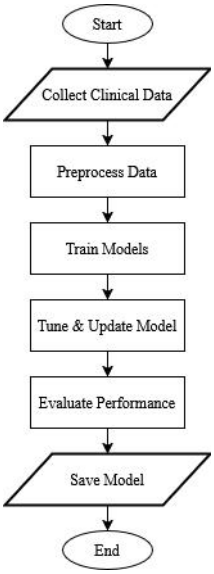


Figure 1 - System Architecture

Figure 2.a



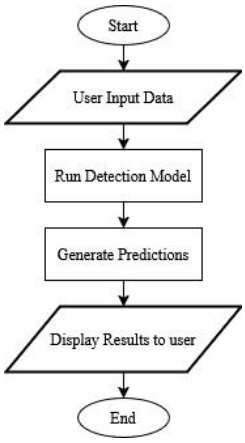


Figure 2.b

Figure 1 presents an overall system architecture that guides the entire predictive workflow, beginning with the acquisition of multiple liver disease datasets. The raw data undergoes a series of preprocessing steps—such as normalization, log transformations, and outlier management through winsorization—to ensure consistency and improve model performance. Once standardized, the data is split into training, validation, and testing subsets. If necessary, data augmentation strategies like up-scaling can be applied to address class imbalance. Each model (XGB, LightGBM, and MLP) is then tuned using RandomizedSearchCV, optimizing parameters like learning rate and regularization. These individually optimized models are subsequently combined in a soft-voting ensemble, which is evaluated against metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. To enhance transparency, SHAP is integrated to illustrate the contribution of each clinical biomarker to the final predictions.

Flow charts for both the training and prediction phases. Figure 2.a training flow chart outlines the steps, starting from data collection and preprocessing, progressing through model training and tuning, and concluding with performance evaluation and model saving. Figure 2.b prediction flow chart shows how new patient data is received, passed through the same preprocessing pipeline, and ultimately used to generate liver disease stage predictions via the saved ensemble. Together, these diagrams offer a clear, high-level view of how data moves through the system—ensuring that each step is logically connected, modular, and ready for real-world clinical integration

VI. Testing and Evaluation

To ensure the generalizability and robustness of the proposed models, each classifier was evaluated on an independent test dataset comprising 588 samples. As presented in Table 1, the evaluation metrics included accuracy, precision, recall, and F1-score for three standalone classifiers: XGBoost (XGB), LightGBM, and MLP.

Among the individual models, LightGBM achieved the highest accuracy of 96.43%, followed by XGBoost at 95.24% and MLP at 91.33%. Furthermore, LightGBM demonstrated the highest macro-averaged F1-score (96%), signifying its superior classification capability across different stages of liver disease.

While LightGBM exhibited the best standalone performance, an ensemble approach utilizing a Voting Classifier was implemented to assess the impact of integrating multiple classifiers. The Voting Classifier achieved an accuracy of 95.41%, with macro-averaged precision, recall, and F1-score values of 95%. These results underscore the effectiveness of ensemble learning, demonstrating that a combination of multiple models enhances both prediction stability and overall accuracy.

To further contextualize our findings, we compared the proposed Voting Classifier with the XGBoost model referenced in the base paper. Table 2, the Voting Classifier exhibited superior performance across all key evaluation metrics, reinforcing the advantages of ensemble learning over a single classifier approach. This comparison highlights the potential of an optimized ensemble methodology in enhancing the accuracy and reliability of liver disease classification.

To gain deeper insight into how the Voting Classifier handles each class, we generated a confusion matrix as shown in Fig. 3. The diagonal entries represent correct classifications. At the same time, off-diagonal cells indicate instances of misclassification. As the matrix illustrates, most samples fall along the diagonal, indicating high accuracy across all four classes (0, 1, 2, and 3). Specifically, the Voting Classifier correctly classified 146 out of 149 samples for Class 0, 140 out of 150 for Class 1, 139 out of 149 for Class 2, and 136 out of 140 for Class 3. This finding corroborates the high overall accuracy (95.41%) and balanced performance metrics reported in Table 1, confirming that the ensemble approach effectively distinguishes among the various liver disease stages.

Table 1. Performance of Individual Models

MODEL	ACCURACY	PRECISION (MACRO)	RECALL (MACRO)	F1-SCORE (MACRO)
XGB	95.24%	95%	95%	95%
LightGBM	96.43%	96%	96%	96%
MLP	91.33%	91%	91%	91%

Table 2. Comparison with Base Paper

EVALUATION METRIC	XGB (BASE PAPER)	VOTING CLASSIFIER (PROPOSED MODEL)
Accuracy	90.5	95.41
Recall	87	95
Precision	85	95
F1-score	89.9	95

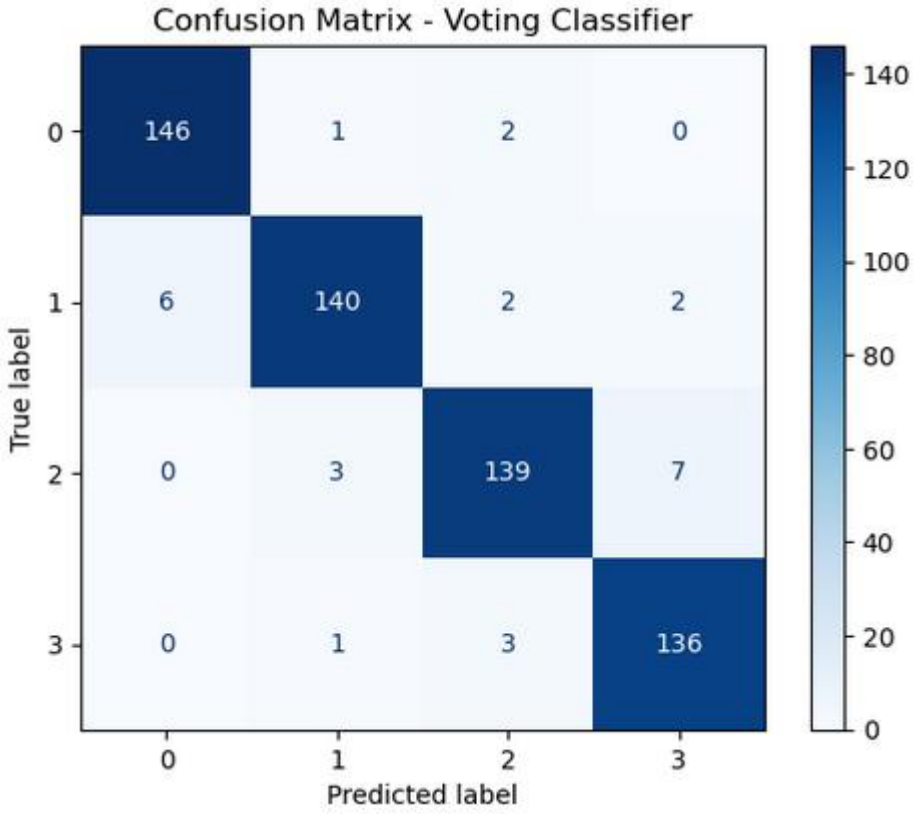


Fig. 3. Confusion matrix for the Voting Classifier on the test set

Figures 4(a), 4(b), and 4(c) present the SHAP-based feature-importance plots for XGB, LightGBM, and MLP, respectively. Each horizontal stacked bar depicts the mean absolute SHAP value across the four liver disease classes, indicating how strongly each biomarker influences the model's predictions.

In all three models, Platelets emerges as the most influential predictor. However, the relative importance of other features—such as Ascites, Alkaline Phosphatase, Aspartate Aminotransferase, and the FIB-4 Score—differs depending on the underlying algorithm. Notably, the MLP places Ascites as its second-most important feature, whereas the tree-based models (XGB and LightGBM) assign higher significance to Alanine Aminotransferase and the FIB-4 Score. These variations underscore each classifier's unique method of weighting clinical biomarkers, reflecting differences in how neural networks and gradient-boosted decision trees prioritize predictive features.

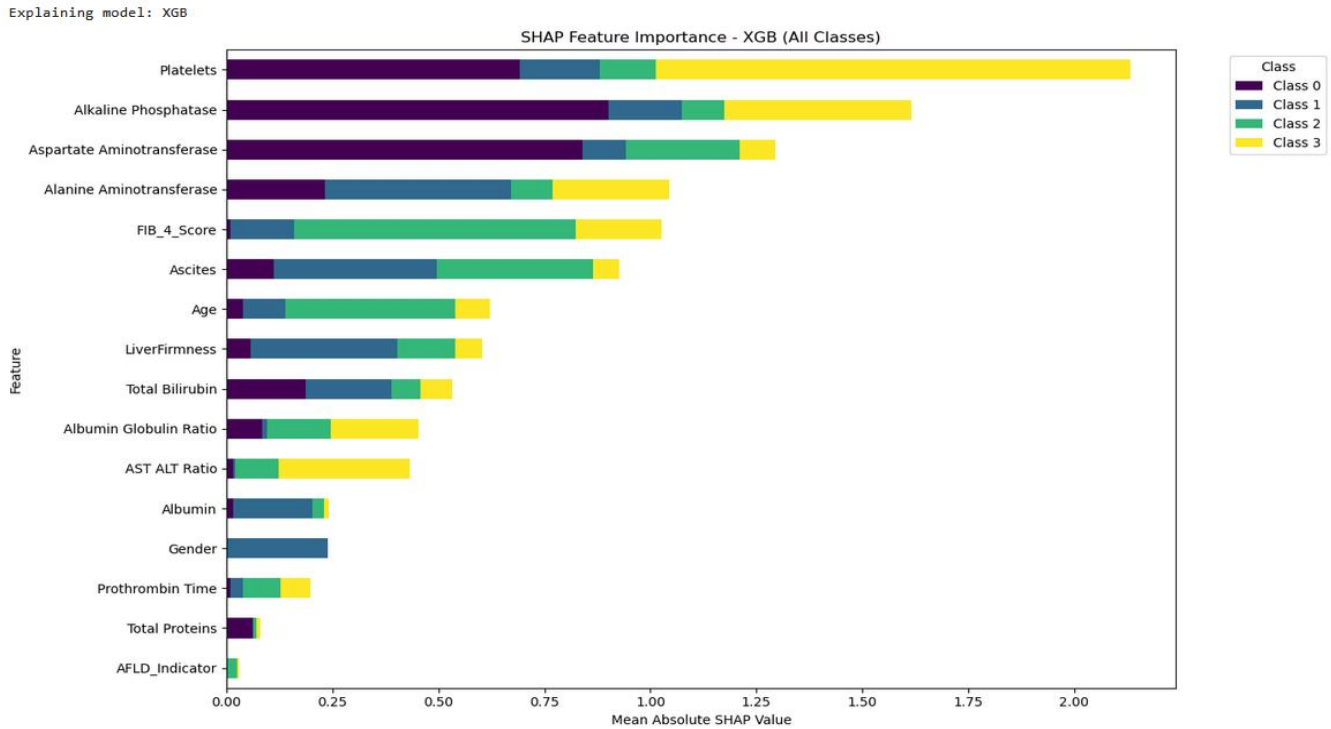


Fig. 4(a). XGB Model Feature Importance

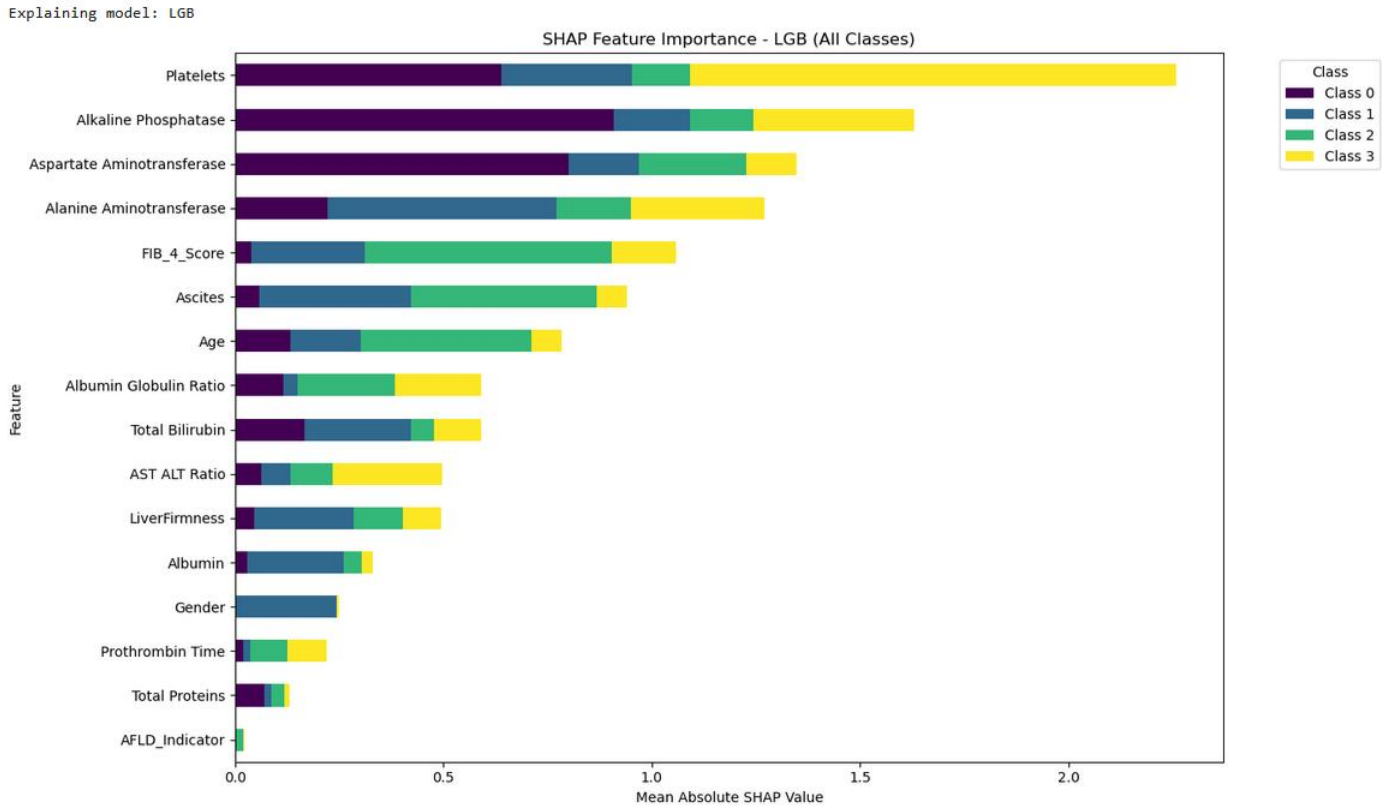


Fig. 4(b). LGBM Model Feature Importance

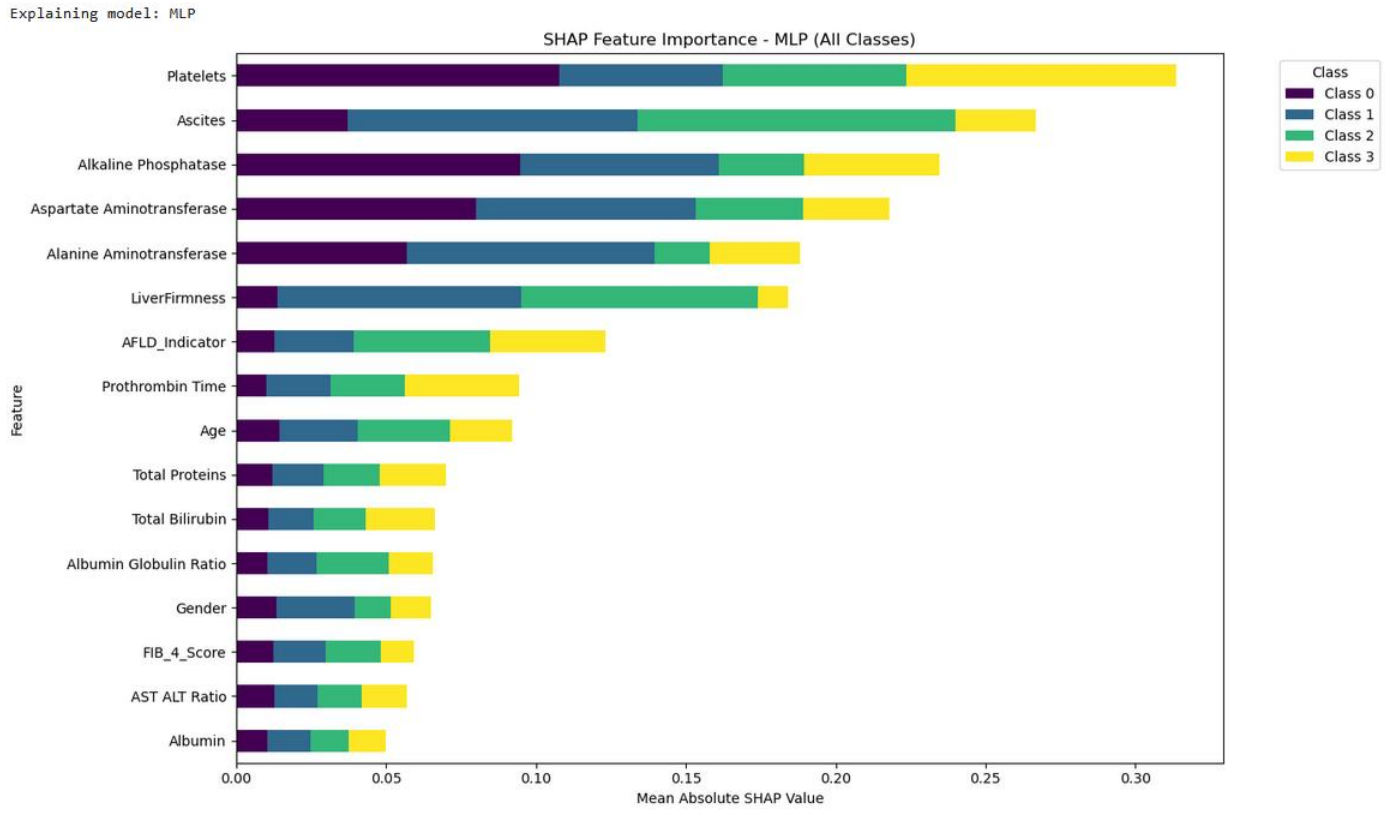


Fig. 4(c). MLP Model Feature Importance

In the base paper [2], a similar SHAP-based analysis identified Platelets, Albumin, and Age as the most influential biomarkers across multiple liver disease classes. While our study likewise confirms the critical role of Platelets, it places additional emphasis on Ascites and Alkaline Phosphatase, reflecting the differences in data composition, preprocessing methods, and modeling approaches between the two investigations. These variations underscore the context-specific nature of biomarker importance—even closely related studies may arrive at distinct feature rankings when their datasets or analytic strategies diverge.

VII. Results and Discussion

We evaluated our models on a hold-out set of liver disease cases, finding that the Voting Classifier consistently outperformed individual algorithms (XGB, LightGBM, and MLP) in accuracy, precision, recall, and F1-score. Across all liver disease stages, only a few misclassifications were observed, highlighting robust diagnostic performance. To enhance interpretability, we employed SHAP (Shapley Additive Explanations), which illustrated how clinical features—such as bilirubin, albumin, and platelet count—contributed to each prediction. Compared to a baseline XGB model from prior literature [9], our ensemble method demonstrated superior results, reinforcing the benefits of combining multiple well-tuned classifiers.

For real-time prediction, we deployed the final ensemble using Flask (with Flask-CORS for cross-origin requests) and a lightweight HTML/CSS/JavaScript front end. The system applies the same preprocessing steps used during training to new patient data and then provides a stage classification along with detailed feature explanations. Figure 5 showcases an example interface where the model predicts "Fibrosis (Scarring of the Liver)." Although many values fall within expected ranges, certain indicators (e.g., Total Bilirubin, AST/ALT Ratio, and Albumin/Globulin Ratio) suggest a moderate degree of liver injury or scarring. The system also identifies Ascites as present while noting Liver Firmness as absent, reflecting a mixed clinical picture that warrants further evaluation. These feature-level insights offer clinicians a transparent rationale for the predicted stage, ensuring more informed decision-making in monitoring and managing liver health.

Overall, these findings validate the robustness of our ensemble-based framework in detecting various stages of liver disease. By combining high predictive accuracy with detailed feature-level insights, the system offers clinicians an actionable tool for early intervention and ongoing patient management. In the next section, we summarize our contributions, discuss limitations, and outline future directions for real-world clinical adoption.

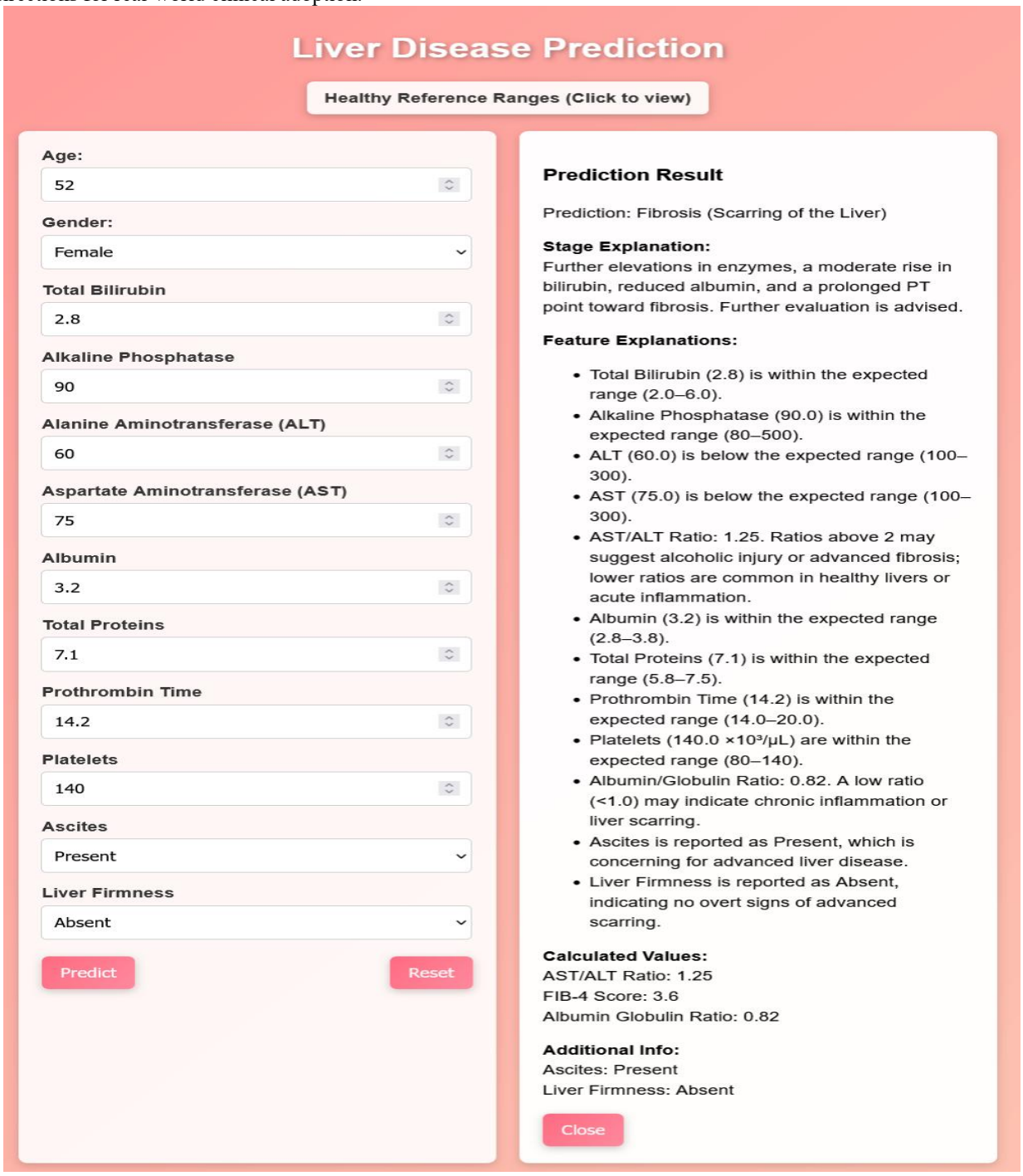


Fig. 5. Liver Disease Prediction

VIII. Conclusion

This paper introduced an ensemble framework for liver disease detection, uniting XGBoost, LightGBM, and MLP models in a Voting Classifier. The ensemble consistently outperformed individual algorithms in accuracy, precision, recall, and F1-score. By integrating SHAP, we provided clear feature-level explanations—highlighting biomarkers such as bilirubin, albumin, and platelet count—that enhance interpretability and clinical relevance. Despite these promising findings, the system's robustness depends on dataset quality and diversity, necessitating further external validation.

Looking ahead, several avenues for future research can expand both the depth and breadth of this approach:

- Predicting Disease Progression: Extend the model to track the transition from early conditions (e.g., hepatitis) to more advanced stages like fibrosis or cirrhosis.
- Incorporating Additional Biomarkers and Imaging: Integrate data from liver ultrasound, MRI, and other advanced diagnostics to capture a more comprehensive clinical picture.
- Personalized Models: Include patient-specific factors—such as genetics, lifestyle, and comorbidities—to develop truly individualized risk assessments and treatment pathways.

IX. Acknowledgment

The authors gratefully acknowledge the UCI Machine Learning Repository for providing free access to critical liver disease datasets, which served as the basis for this study. We also extend our thanks to the open-source community for developing and maintaining essential Python libraries (NumPy, Pandas, Scikit-learn, etc.), enabling seamless data preprocessing, model development, and evaluation. Additionally, we express our gratitude to the authors of the base paper [2], whose pioneering work offered a valuable comparative benchmark for our approach. Finally, we appreciate the support and feedback from colleagues and mentors, whose insights were instrumental in refining this research.

References

[1] World Health Organization, "Global burden of liver disease," Journal of Hepatology, vol. 78, no. 4, pp. 774–785, 2023. [Online]. Available: https://www.journal-of-hepatology.eu/article/S0168-8278%2823%2900194-0/fulltext.

[2] G. Arya et al., "Explainable AI for enhanced interpretation of liver cirrhosis biomarkers," IEEE Access, vol. 11, pp. 123729–123741, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3329759.

[3] National Health Service (NHS), UK, "Assessing liver function and interpreting liver blood tests," Specialist Pharmacy Service, 2022. [Online]. Available: https://www.sps.nhs.uk/articles/assessing-liver-function-and-interpreting-liver-blood-tests/. Accessed: Feb. 20, 2025.

[4] MedlinePlus, "Total protein and albumin/globulin (A/G) ratio," U.S. National Library of Medicine, Jan. 2024. [Online]. Available: https://medlineplus.gov/lab-tests/total-protein-and-albumin-globulin-a-g-ratio/. Accessed: Feb. 20, 2025.

[5] "Liver function tests: Indication and interpretation," The Pharmaceutical Journal, PJ, January 2022, Vol. 308, No. 7957; [Online]. Available: https://doi.org/10.1211/PJ.2022.1.124202. Accessed: Feb. 20, 2025.

[6] W. Ghaniyya, I. Humairah, and U. Kholili, "Evaluating the accuracy of APRI and FIB-4 scores in chronic HBV-related liver fibrosis: A literature review," World Journal of Advanced Research and Reviews, vol. 24, no. 3, pp. 2680–2684, 2024. [Online]. Available: https://doi.org/10.30574/wjarr.2024.24.3.3937.

[7] Sterling, R. K. et al., "Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection," Hepatology, vol. 43, no. 6, pp. 1317–1325, June 2006. [Online]. Available: https://doi.org/10.1002/hep.21178.

[8] Beom Kyung Kim et al., "Validation of FIB-4 and comparison with other simple noninvasive indices for predicting liver fibrosis and cirrhosis in hepatitis B virus-infected patients," Liver International, vol. 30, no. 4, pp. 546–553, first published: 22 February 2010. [Online]. Available: https://doi.org/10.1111/j.1478-3231.2009.02192.x.

[9] S. Dalal, E. M. Onyema, and A. Malik, "Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy," World Journal of Gastroenterology, vol. 28, no. 46, pp. 6551–6563, Dec. 2022. [Online]. Available: https://doi.org/10.3748/wjg.v28.i46.6551.

[10] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 4765–4774. [Online]. Available: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.

[11] Dritsas, E., & Trigka, M. (2023). "Supervised Machine Learning Models for Liver Disease Risk Prediction." Computers, 12(1), 19. [Online]. Available: https://doi.org/10.3390/computers12010019.

[12] S. Ardchir, Y. Ouassit, S. Ounacer, M. Y. El Ghoumari, and M. Azzouazi, "An integrated ensemble learning framework for predicting liver disease," International Journal of Online Engineering (iJOE), vol. 19, no. 13, pp. 138–152, Sep. 2023. [Online]. Available: https://doi.org/10.3991/ijoe.v19i13.41871.

[13] Fuliang Yi et al., "XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease," BMC Medical Informatics and Decision Making, vol. 23, no. 3, p. 129, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10369804/. PMID: 37491248.

[14] Lu, S., Chen, R., Wei, W., Belovsky, M., & Lu, X. (2022). "Understanding Heart Failure Patients' EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions." Journal of Biomedical Informatics, 129, 102136. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8861751/. PMID: 35308970.

[15] Praveen Sharma, "Value of Liver Function Tests in Cirrhosis," Journal of Clinical and Experimental Hepatology, vol. 12, no. 3, pp. 948–964, Published online Nov. 12, 2021. [Online]. Available: https://doi.org/10.1016/j.jceh.2021.11.004.

[16] Chukwuebuka Joseph Ejiyi et al., "Polynomial-SHAP analysis of liver disease markers for capturing of complex feature interactions in machine learning models," Computers in Biology and Medicine, vol. 182, 2024, Article 109168. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2024.109168. ISSN 0010-4825.

[17] Rutweek, R. (2020). "Liver Fibrosis Stage Prediction." GitHub Repository. [Online]. Available: https://github.com/rutweek/Liver-Fibrosis-Stage-Prediction/tree/master.

[18] Ramana, B. & Venkateswarlu, N. (2022). "ILPD (Indian Liver Patient Dataset)." UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C5D02C.

[19] Lichtinghagen, R., Klawonn, F., & Hoffmann, G. (2020). "HCV Data." UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C5D612.

[20] Gong, G. (1988). "Hepatitis Dataset." UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Hepatitis.