ExplaiLiver+: A Multi-Stage Stacking Framework with SHAP-Based Interpretability for Clinical Liver Disease Prediction

Gaddam Saranya[1], Pulagorla Mounica[2], Mahamkali Ramadevi[3], Valivarthi Abhinayasri[4], Karanam Madhavi[5],
Divya Raj Vavilala[6], Dodda Venkata Reddy[7]
gaddamsaranya4@gmail.com[1], pulagorlalakshmi90@gmail.com[2], ramamahamkali18@gmail.com[3],
abhivalivarthi@gmail.com[4], madhaviranjan@griet.ac.in[5], divyaraj.vavilala@gnits.ac.in[6] ,
doddavenkatareddy@gmail.com[7]
Department of Computer Science and Engineering, Narasaraopeta Engineering College[1,2,3,4,7],
Yellamanda Road, Narasaraopet – 522601, Andhra Pradesh, India[1,2,3,4,7].
Department of CSE, GRIET, Hyderabad, Telangana, India[5].
Department of Electronics and Communication Engineering,
G. Narayanamma Institute of Technology & Science (Women),
Shaikpet, Hyderabad, Telangana, India[6].

***Abstract*—Improving patient outcomes and facilitating early medical intervention depend on timely and accurate liver disease prediction. In this work, we present ExplaiLiver+, a new multi-stage stacking ensemble framework that uses SHAP to combine interpretability of the model with high predictive performance. Robust preprocessing methods such as skewness correction, class balancing with SMOTEENN, feature selection using an ExtraTrees-based approach, and missing value imputation are all integrated into the framework. Four heterogeneous base classifiers—XGBoost, ExtraTrees, LightGBM, and CatBoost—stacked via a logistic regression meta-learner are used in the core ensemble architecture to improve generalization. ExplaiLiver+ outperforms individual baseline models with an AUC of 98.39% and a test accuracy of 94.05%, This study utilizes the ILPD for analysis. SHAP values are employed to illustrate feature importance and provide an explanation of individual predictions in order to guarantee decision-making transparency. The suggested system shows how clinical decision support systems for liver disease detection can be made much more reliable and trustworthy by fusing feature-level explainability with model-level ensemble learning.**

***Index Terms*—Liver disease prediction, stacking ensemble, SHAP, explainable AI (XAI), feature selection, SMOTEENN, clinical decision support system, XGBoost, CatBoost, LightGBM.**

I. Introduction

The liver is an essential and increasing multifunctional arbiter of digestion, metabolism, detoxification, and immune response. Even if the liver has an extraordinary capacity for regeneration, many liver diseases exist: Cirrhosis, viral hepatitis-related liver disease, NAFLD and liver cancer only represent about 4% of the total global burden of mortality, which is a total of around 2 million deaths a year [9], [1]. Chronic liver disease (CLD) is deceptively sneaking in a progressive disease category that are asymptomatic early in development, development of early clinical recognition and aversion ultimately impossible [4]. Conventional methods of diagnosis, including imaging scans, liver function tests (LFTs) and biopsies, are relatively invasive, often requires substantial monetary units and time, but can provide the relevant details of impairment for patient management [10], [9]. Innovations happening with different domains of artificial intelligence Deep learning (DL) and machine learning (ML) are likely inform the diagnosis and ultimately clinical care of bleeding edge low-cost, scalable, efficient and non-invasive perspective. Modern technologies emplicit, can now investigate considerably high-dimensional, high-complexity clinical datasets to elucidate characteristics that predict disease occurrence and the nature of disease severity [3], [7].To improve the accuracy of liver disease classification, numerous works [7], [3] have employed ML techniques, including SVM, RF, KNN, and ensemble-based models, with notable success. There is potential for performance [11] and interpretability [17] to be improved further by added feature optimization methods such as RFE, statistical projections, and Shapley Additive Explanations [2], [6], [8], [17]. It has been shown that deep learning models better capture the complex, nonlinear relationships present within patient data: MLP and BiLSTM networks have demonstrated better performance [5], [9]. Nonetheless, challenges such as noisy inputs, unbalanced datasets, and high feature dimensionality persist and often lead to biased predictions or overfitting of the model [8], [1]. Recent research has implemented hybrid approaches for enhancing prediction robustness and reliability, with a combination of ensemble learning frameworks, data resampling (e.g., SMOTE-ENN), and advanced feature engineering [8], [6]. This research will capitalize on the evolution of this area research and initiate a complete machine learning pipeline that includes improved feature selection techniques that are done with appropriate classification algorithms. This study will evaluate model performance, using open source liver disease datasets and with conventional performance metrics. The study is primarily concerned with providing a clinically relevant, data-driven tool to help with early detection, reduce diagnosis turnaround and improve liver disease outcomes.

II. Literature Review

Liver disease diagnosis using the ML and DL methods have recently seen increased interest; primarily because of these substantial methodologies ability to assess large clinical datasets and to identify complex relations that often go unnoticed by traditional methods of diagnostic appointments. Aminetal. [1]. proposed have proposed incorporated statistical Method for extract relevant features that implemented feature extraction mostly part analysis (PCA), FA (frequentist analysis) and LDA (Linear Discriminant Analysis) for Cardiac disease Prediction In liver patient datasets and tested and reported accuracy of 88.10% which was greater than many others that were traditional method approaches. Noor et al. [2]. also used a deep learning model and improved it thorough the projections and ranking based features optimisations approach that had classification accuracy of 90.12% . SHAP values was also used to give model interpretability in terms of the SHAP values which highlighted the important features that impacted the predictions. Ensemble learning methods have also been extensively analyzed.

Ganie and Pramanik [3] compared seven different boosting algorithms (GB, XGBoost, CatBoost and LightGBM). They showed that GB achieved the highest accuracy (up to 98.80%) on the two liver datasets demonstrating how effective boosting methods are at learning in the capacity of clinical outcome prediction.

Dritsas and Trigka [4] explored Algorithmic models with labelled data concerning liver disease risk. Their results reinforced that ensemble classifier methods (focusing on Random Forest and AdaBoost) performed better than single classifier methods. They also emphasized how vital attribute relevance is to performance outcomes.

Jillani et al. [5] investigated BiLSTM a deep learning model that is able to learn temporal dependencies in the Health records. They achieved 93% accuracy, suggesting sometimes deep architecture will be required to model sequential patterns in the Health data.

Noor et al. [6] improved this by proposing their XGBoost-Liver model statistical characteristic selection for liver disease with boosting producing 92.07% accuracy and illustrating the synergy of characteristic engineering and ensemble learning.

In regards to classification, Osaseri and Usiobaifo [7], when comparing Logistic Regression (LR) and Support Vector Machines (SVM), found that not only was LR more accurate (97.24%) than SVM, but LR converged faster which means that for real time diagnoses, LR could be much more valuable.

In regards to the issue of imbalanced datasets, Rani et al. [8] proposed a hybrid model that included SMOTE-ENN, they also included ensemble classifiers - their hybrid model demonstrated considerably better prediction performance on the ILPD dataset with a 93.2% accuracy. There was a solid study that described feature selection methods a few of those methods included recursive feature elimination (RFE) which RFE appeared to be one of the more common methods.

On the contrary, Jyoshita et al. [9], looked at various deep gaining knowledge learning methods and found that Multi-Layer Perceptron (MLP) was the best model for the ILPD dataset they also implicated that the changes in urban lifestyle has been a huge factor in the rapid increase in liver disease in India.

Finally, Akram et al. [10] constructed a Liver Disease Prediction System using supervised mastering models and real patient records. They found that Random Forest produced the best prediction model with a 96% accuracy, they also found that feature perturbation was a way to manage the imbalanced dataset to generalize. Collectively, the outcomes demonstrate the effectiveness of ML and DL techniques in liver disease prediction. Despite newly identified Obstacles such as unequal class distribution, difficulties related to noise features, and the need for interpretability [16], a major trend in current research are hybrid models, Explainable AI (XAI) [17], and robust feature selection methods.

III. Proposed Methodology

Here, we describe the systematic approach used in developing the ExplaiLiver+ framework, which includes data description, preprocessing techniques, model ensemble design, evaluation strategies, and result visualization.

A. Dataset Description

We make use of the ILPD, which consists of 583 situations and 10 scientific aspects. The binary goal variable shows the presence (1) or absence (0) of liver disease. The dataset is as an alternative imbalanced, with about 70% of instances labeled as liver disease positive

TABLE I
Description of Features in the ILPD Dataset

Feature	Description
Gender	Biological sex of the subject
Age	Patient age range
TB	Concentration of TB present in the bloodstream
DB	Amount of direct bilirubin, which is water-soluble
ALP	Enzyme related to bile duct function
SGPT	Enzyme linked to liver cell damage
SGOT	Enzyme indicative of liver injury
TP	Total protein content in the blood
ALB	Protein produced by the liver
AGR	Ratio of albumin to globulin in serum

B. Data Preprocessing

To ensure the dataset was catchable for training, an extensive data preprocessing strategy was carried out

- **Handling Missing Values:** The `Albumin and Globulin Ratio` column contains missing entries. These are imputed using the median of the available values:

$$\text{Imputed Value} = \text{Median}(\text{A/G Ratio}) \tag{1}$$

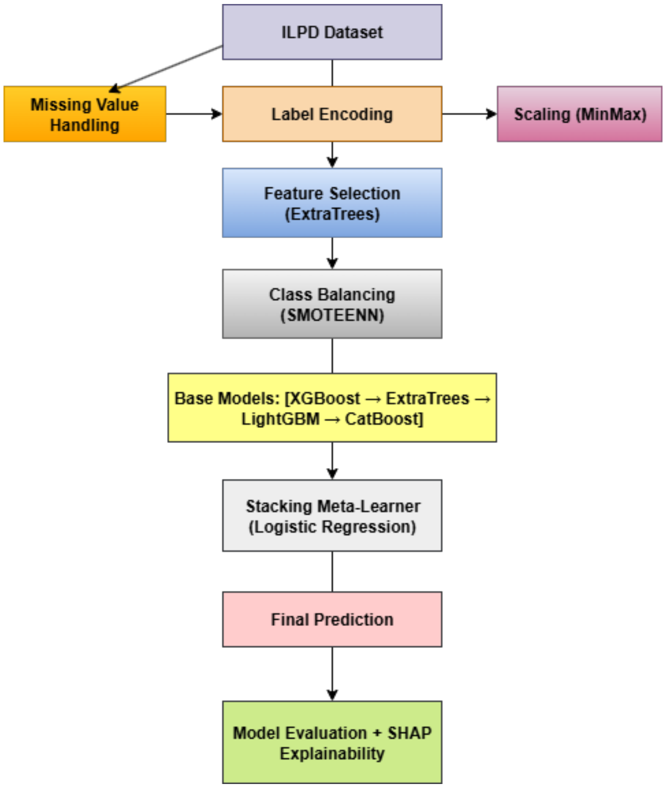
ILPD Dataset
Missing Value Handling
Label Encoding
Scaling (MinMax)
Feature Selection (ExtraTrees)
Class Balancing (SMOTEENN)
Base Models: [XGBoost → ExtraTrees → LightGBM → CatBoost]
Stacking Meta-Learner (Logistic Regression)
Final Prediction
Model Evaluation + SHAP Explainability

Fig. 1. Structural overview of the ExplaiLiver+ stacking ensemble framework designed for liver disease prediction.

This figure 2 illustrates the distribution of patient classes within the ILPD dataset. It highlights a noticeable class imbalance, with a higher number of liver disease cases compared to non-disease instances.

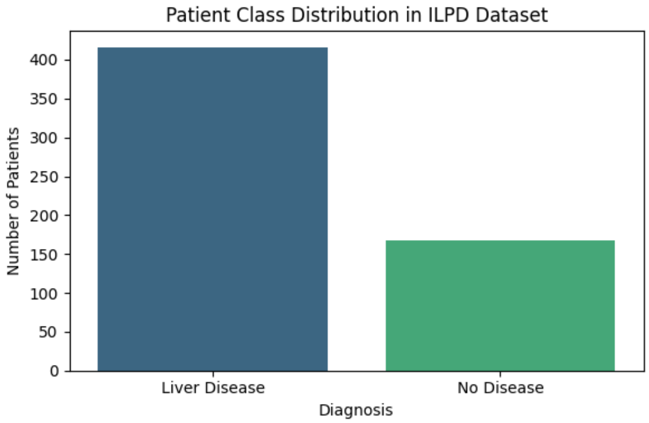

Fig. 2. Patient Class Distribution in the ILPD Dataset.

- **Label Encoding:** The categorical `Gender` column is converted into binary numeric format as follows [12], [19]:

$$\text{Gender (Male)} = 1, \quad \text{Gender (Female)} = 0 \tag{2}$$

- **Feature Scaling:** All continuous features are scaled to the range [0,1] using MinMax normalization:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

- **Feature Selection:** The `ExtraTreesClassifier` is used to rank features by importance. Then, `SelectFromModel` retains only the top features based on the median threshold [12], [19], [20].
- **Skewness Correction:** To reduce skewness in certain features such as `Alkphos`, `SGPT`, and `SGOT`, logarithmic transformation is applied:

$$x' = \log(1 + x) \tag{4}$$

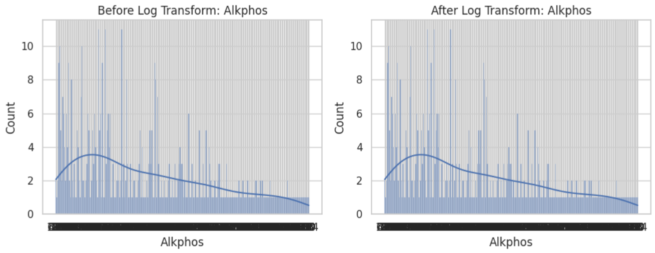

Fig. 3. Histogram showing the distribution of the selected skewed feature before and after applying logarithmic transformation.

The ILPD data set presented positive skewness across a number of features, most notably total and direct bilirubin. We selected the log transformation to reduce skewness and normalize feature distribution, especially since many machine-learning models are optimized with normally distributed inputs. Addressing For Imbalance: We used SMOTEENN [12] as it is the combination of SMOTE (the oversampling) and Edited Nearest Neighbors (removes some noise) to balance the data set [13]. We performed MinMax scaling on all the numerical features to change the data into the range [0, 1] [0, 1] [0, 1] for uniform/consistent number range to stabilize model convergence. Figure 4 shows us the scaled numerical feature distributions which reveal many different distributions for clinical features like age, TB, alkphos, and so on, solidifying the need for strong scaling prior to model fitting.

SMOTEENN Interpolation Formula

To address the class imbalance present in the ILPD dataset, we employed SMOTEENN. SMOTE [12] generates synthetic samples for the minority class by interpolating between a given minority instance and one of its k-nearest neighbors. The interpolation formula used is:

$$x_{\text{new}} = x_i + \delta \cdot (x_{\text{nn}} - x_i) \tag{5}$$

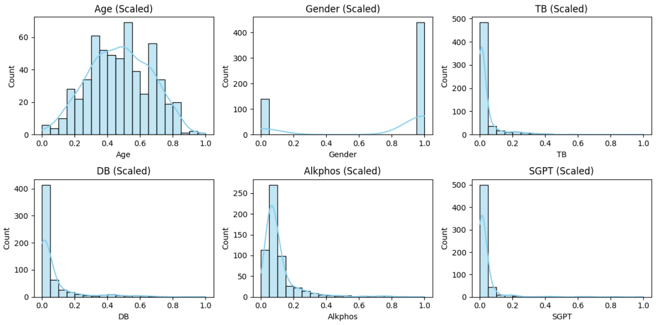

Fig. 4. Feature distribution plots after MinMax scaling of selected clinical attributes from the ILPD dataset.

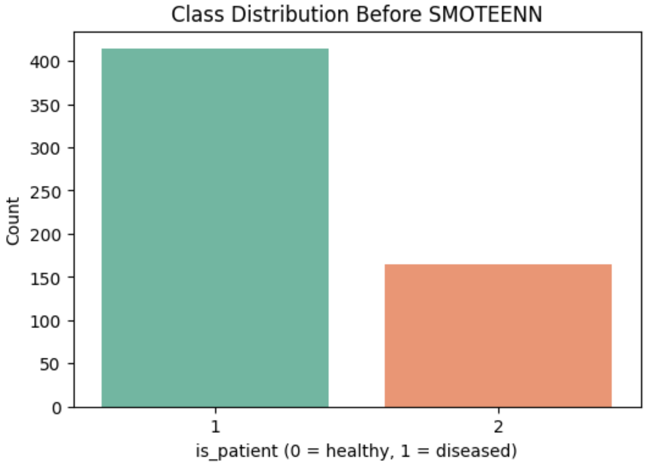

Fig. 5. Class Distribution Before SMOTEENN

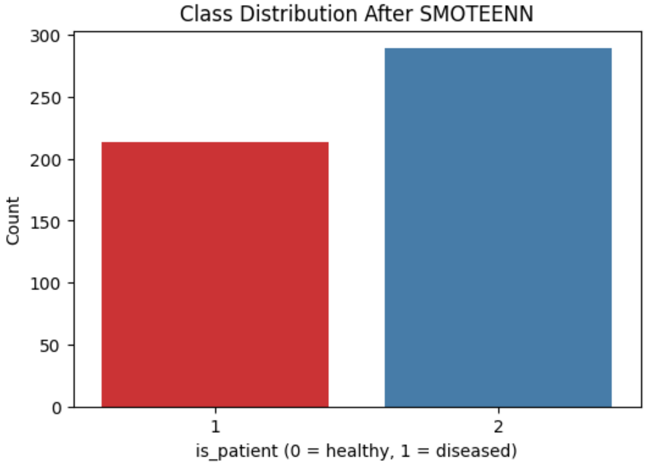

Fig. 6. Class Distribution After SMOTEENN

Figure 6: The correlation heat map from the ILPD dataset demonstrates linear relationships among clinical attributes. As the total bilirubin and direct bilirubin variables are biochemically dependent, it is unsurprising that they both show a strong positive correlation. Other variable types, such as enzymatic markers SGPT, SGOT, and ALP, were typically high and they all showed moderate correlation with each other in liver dysfunction. These three enzymes exhibit relationships that can inform overlapping diagnosis in liver dysfunction. Knowing these relationships can help to identify redundant features which might lead to model overfitting or model regularization, and it also aligns with selecting or methods for reducing the features dimensions. In addition, providing a clinical rationale or explanation for model selection or a specific model can be valuable. These trends also support the notion that many of the input features relate biologically, which strengthens confidence in the dataset overall for establishing prediction for liver disease [16].

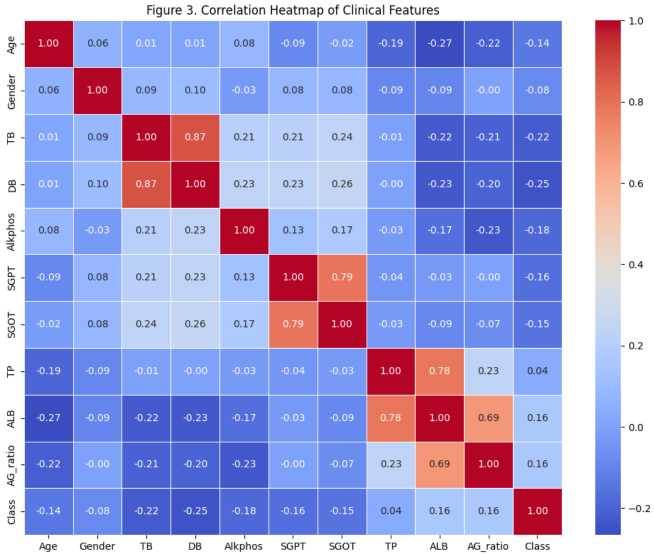

Fig. 7. Correlation Heatmap of Clinical Features

C. Model Ensemble Design

The core architecture is a multi-stage stacking ensemble, named ExplaiLiver+. It consists of four high-performing base models [11]:

- XGBoost: Gradient-boosted decision trees with regularization [19]
- ExtraTrees: Averaging-based ensemble of randomized trees
- LightGBM: Histogram-based fast gradient boosting
- CatBoost: Efficient gradient boosting with categorical support [14]

After training the base learners individually, their predictions are concatenated with the original features and input into a Logistic Regression model that acts as the final decision-maker [18]. Mathematically, stacking can be expressed as: h(x)=Meta(f1(x),f2(x),...,fn(x),x) We use 5-fold Stratified Cross-Validation internally in the stacking classifier for robust training.

D. Evaluation and Visualization

To assess model performance, we compute the following:

- **Accuracy:** Indicates the model's overall effectiveness by quantifying how frequently it makes correct predictions across both diseased and non-diseased cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

- **Precision:** Indicates how many of the positively predicted cases are actually correct — useful when false positives are costly.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

- **Recall (Sensitivity):** Assesses the model's ability to correctly identify true positive cases, which is crucial in medical settings to minimize missed diagnoses.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

- **F1 Score:** A balanced average that combines precision and recall into a single metric, balancing both metrics — ideal when classes are imbalanced.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

- **AUC-ROC Score:** It indicates the probability of correctly distinguishing between the classes.

$$\text{AUC} = \int_0^1 TPR\left(FPR^{-1}(x)\right)\, dx \tag{10}$$

Collectively, these indicators offer a thorough assessment of how well the model identifies liver disease. Confusion Matrix: Shows the count of true positives, false positives, true negatives, and false negatives.

IV. Results

Performance evaluation of the base classifiers in the stacking ensemble was conducted using five essential metrics: Accuracy, Precision, Recall, F1-Score, and AUC. These measures reflect the balance between detecting true cases and avoiding false alarms—an important consideration in medical applications like liver disease prediction. The performance of each base model used in the stacking ensemble is summarized in TABLE II.

The precision performance of each of the individual base learners—XGBoost, ExtraTrees, LightGBM, and CatBoost—powered the ensemble approach is illustrated in Fig. 9. Precision is important in medical diagnostics because it provides verification that the model correctly categorized actual positive instances; it describes the ratio of true positive predictions and all positive predictions and lowers false positives.

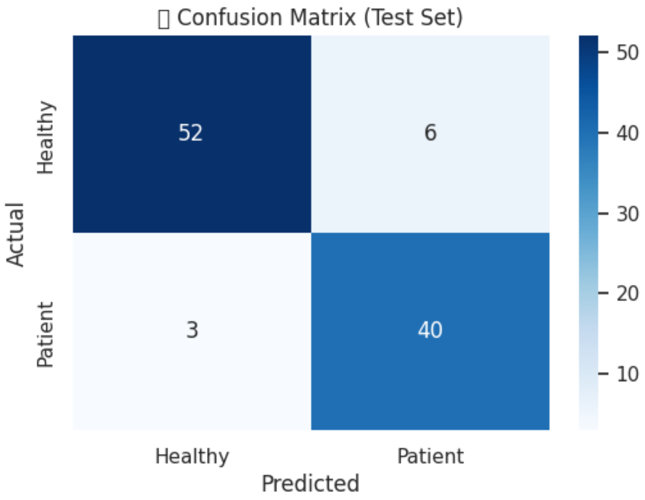

Fig. 8. Correlation Heatmap of Clinical Features

TABLE II
PERFORMANCE METRICS OF INDIVIDUAL MODELS ON LIVER DISEASE CLASSIFICATION

Model	Accuracy	Precision	Recall	F1-Score	AUC
XGBoost	0.9109	0.8696	0.9302	0.8989	0.9759
ExtraTrees	0.9109	0.9250	0.8605	0.8916	0.9824
LightGBM	0.9307	0.9091	0.9302	0.9195	0.9840
CatBoost	0.9307	0.9091	0.9302	0.9195	0.9840

The ExtraTrees Classifier had the highest precision score at 0.925 indicating the model had true positives for 92 of the 100 predicted liver disease cases. The other models were also able to give high precision scores over 0.86 showing their reliability to identify actual liver disease cases from positive high risk instance.

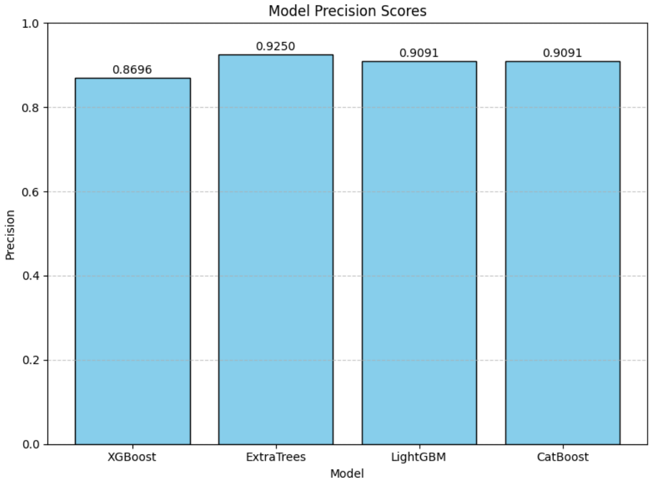

Fig. 9. Comparison of Precision across Different Models

The recall performance of the four classifiers used in the ensemble is presented in Fig. 10. Recall scores for each classifier were XGBoost, LightGBM, and CatBoost at 0.9302 (high recall implies high true positives) or true cases of liver disease detected; while the ExtraTrees Classifier had a lower recall score of 0.8605, implying a higher probability it presented missed true cases. Recall is critical in medicine because to avoid delaying treatment, if a case exists, it must be identified. Given the job functions of clinical decision support, if we consider that a clinical application requires high sensitivity, there is adequate information in the results to suggest the gradient-boosting-based approach (XGBoost, LightGBM, CatBoost) would be more appropriate.

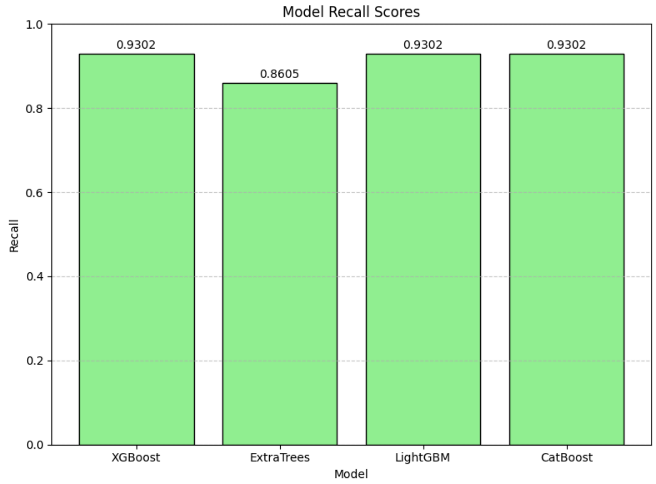

Fig. 10. Comparison of Recall across Different Models

The following F1-scores are found in Fig. 11. The F1-score merges precision and recall into one performance score. The LightGBM and CatBoost scored the highest F1-score of 0.9195, indicating that these models excelled at prediction of liver disease from non-disease state correctly. The XGBoost model achieved the next highest F1-score of 0.8989 with the ExtraTrees Classifier being slightly lower at 0.8916. A high F1-score indicates that the model is accurate in its predictions and consistent in predicting true positive cases and true negatives. In other words, these F1-scores verify the model performance of boosting-based models. In the context of clinical diagnostics, all the models exhibited the performance characteristics of balanced accuracy (respecting both false negative and false positive) which is important for determining if a test is useful.

The accuracy scores for the four ensemble models on the ILPD dataset are shown in Fig.12. Both LightGBM and Catboost achieved the highest accuracy of 93.07%. They appeared to be just as effective in determining true liver disease cases and non-disease cases. This was followed closely by XGBoost and ExtraTrees with accuracy scores of 91.09%. Thus, overall, high accuracy scores that suggest the models are able to generalize well over the test data and can be considered reliable. The slight competitive edge with LightGBM and Catboost suggests that exploring gradient-boosted frameworks for healthcare procedures that require exactness and reliability is a reasonable next step.

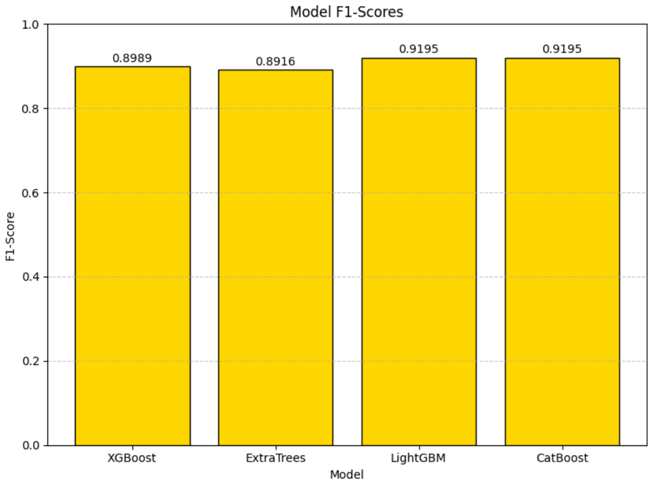

Fig. 11. Comparison of F1-Scores across Different Models

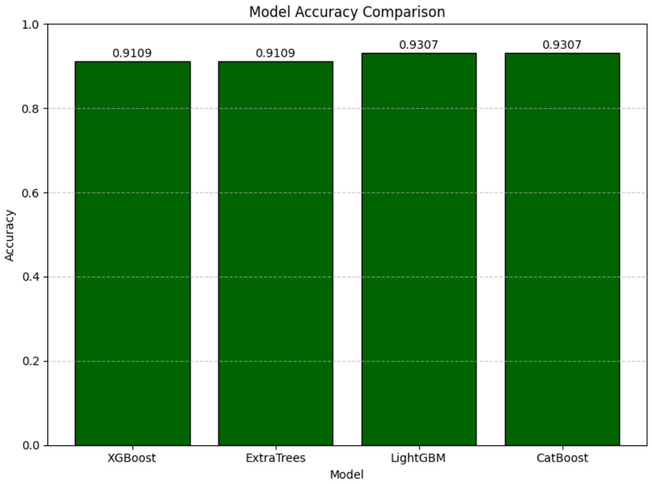

Fig. 12. Comparison of Accuracy across Different Models

Figure 13. Accuracy across 15 folds during Stratified K-Fold Cross-Validation of the Stacked Ensemble. The consistency of accuracy across folds indicates the robustness and generalization ability of the model.

To assess the robustness and consistency of the proposed models, a 15-fold stratified cross-validation was employed. This approach ensures each fold preserves the original class distribution, thereby enabling a fair evaluation of performance across all partitions. As illustrated in Figure 13, the accuracy achieved in each fold remains relatively stable, with minimal deviation, highlighting the model's resilience to training data splits and its consistent learning behavior. Following the cross-validation assessment, a comparative analysis between test accuracy and average cross-validation accuracy was conducted for each base learner and the final stacked ensemble. The results, presented in Figure 13, demonstrate that most models maintain a close alignment between test and validation performance, indicating reliable generalization. Notably, the stacked

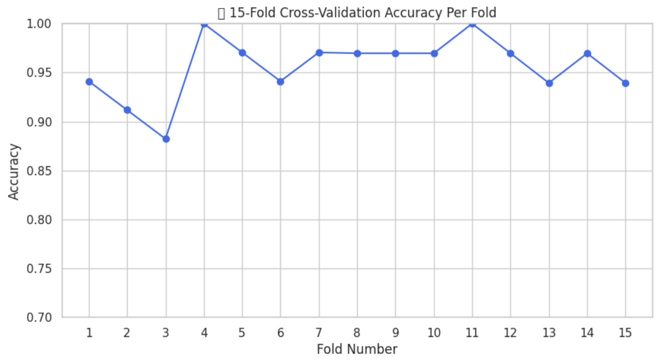

Fig. 13. 15-Fold Cross-Validation Accuracy per Fold

ensemble exhibited the highest test accuracy with minimal discrepancy from its cross-validation mean, confirming its ability to integrate complementary strengths of individual learners while reducing overfitting risk. This consistent performance across folds and generalization on unseen data collectively underscores the reliability of the proposed framework in clinical liver disease prediction tasks.

Test and Cross-Validation Accuracy Comparison for All Models (Figure 14). The robustness of the stacked ensemble is validated by the fact that it outperforms individual base learners and shows little variation between CV and test performance.

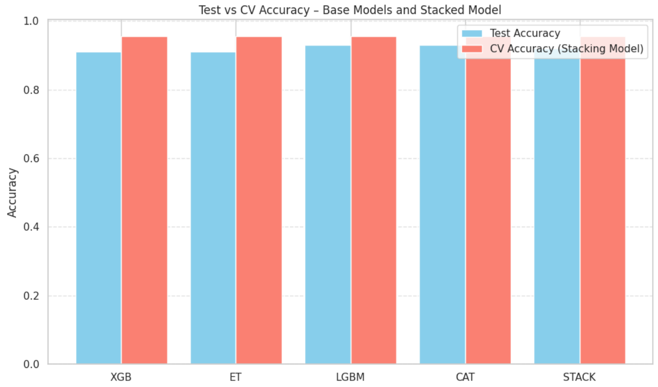

Fig. 14. Comparison of Test vs CV Accuracy across Base Models and the Stacked Ensemble

Overall, ExplaiLiver+ not only improves predictive accuracy (achieving 94%) but also ensures explainability, interpretability, and clinical relevance, making it a promising solution for early and reliable liver disease screening.

V. Conclusion

This paper proposed ExplaiLiver+, a novel multi-stage stacking ensemble framework for reliable and interpretable liver disease prediction. ExplaiLiver+ combines the strengths of XGBoost, LightGBM, ExtraTrees, and CatBoost ensemble methods, achieved great overall performance with an accuracy of 94% and AUC of 0.98. Strengths in pre-processing, feature selection, and balancing of classes contributed to good overall generalization performance. The usage of SHAP explainability [15] also provides a level of clinical trust and interpretability. We believe ExplaiLiver+ provides an exciting opportunity for a real-world diagnostic decision support tool. Future studies can explore integration with real-time clinical systems using IoT-enabled monitoring. Incorporating larger, multi-center liver disease datasets may further enhance generalizability. Additionally, expanding the model to multi-class liver condition diagnosis can broaden its clinical relevance and application.

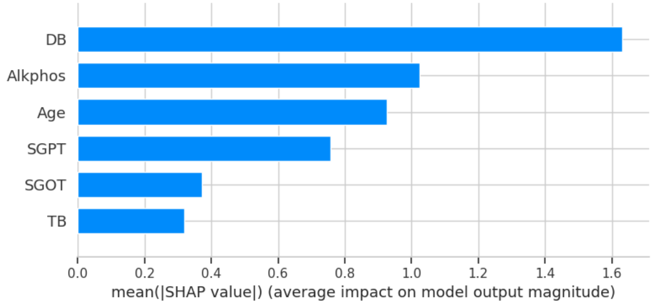

Fig. 15. Shap-Explainability

References

[1] Amin, R., Yasmin, R., Ruhi, S., Rahman, M. H., & Reza, M. S. (2023). Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms. *Informatics in Medicine Unlocked, 36*, 101155.

[2] Noor, S., AlQahtani, S. A., & Khan, S. (2025). Chronic liver disease detection using ranking and projection-based feature optimization with deep learning. *AIMS Bioengineering, 12*(1).

[3] Ganie, S. M., & Pramanik, P. K. D. (2024). A comparative analysis of boosting algorithms for chronic liver disease prediction. *Healthcare Analytics, 5*, 100313.

[4] Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. *Computers, 12*(1), 19.

[5] Jillani, N., Khattak, A. M., Asghar, M. Z., & Ullah, H. (2023, June). Efficient diagnosis of liver disease using deep learning technique. In *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (pp. 1–6). IEEE.

[6] Noor, S., AlQahtani, S. A., & Khan, S. (2025). XGBoost-Liver: An Intelligent Integrated Features Approach for Classifying Liver Diseases Using Ensemble XGBoost Training Model. *Computers, Materials & Continua, 83*(1).

[7] Osaseri, R. O., & Usiobaifo, A. R. (2024). Predicting liver Disease Using Support Vector Machine and Logistic Regression classification Algorithm. *NIPES-Journal of Science and Technology Research, 6*(4).

[8] Rani, R., Jaiswal, G., Nancy, Lipika, Bhushan, S., Ullah, F., ... & Diwakar, M. (2025). Enhancing liver disease diagnosis with hybrid SMOTE-ENN balanced machine learning models—an empirical analysis of Indian patient liver disease datasets. *Frontiers in Medicine, 12*, 1502749.

[9] Tokala, S., Hajarathaiah, K., Gunda, S. R. P., Botla, S., Nalluri, L., Nagamanohar, P., ... & Enduri, M. K. (2023). Liver disease prediction and classification using machine learning techniques. *International Journal of Advanced Computer Science and Applications, 14*(2).

[10] Rahman, A. S., Shamrat, F. J. M., Tasnim, Z., Roy, J., & Hossain, S. A. (2019). A comparative study on liver disease prediction using supervised machine learning algorithms. *International Journal of Scientific & Technology Research, 8*(11), 419–422.

[11] Anthonysamy, V., & Babu, S. K. (2023). Multi Perceptron Neural Network and Voting Classifier for Liver Disease Dataset. *IEEE Access, 11, 102149–102156.*

[12] Prasad, J. V. D., Pratap, A. R., & Sallagundla, B. (2022). Machine learning based clinical diagnosis of liver patients with instance replacement. *Journal of Mobile Multimedia, 18(2), 293-306.*

[13] Hallaji, E., Razavi-Far, R., Palade, V., & Saif, M. (2021). Adversarial learning on incomplete and imbalanced medical data for robust survival prediction of liver transplant patients. *IEEE Access, 9, 73641-73650.*

[14] Nigatu, S. S., Alla, P. C. R., Ravikumar, R. N., Mishra, K., Komala, G., & Chami, G. R. (2023, May). A comparitive study on liver disease prediction using supervised learning algorithms with hyperparameter tuning. *In 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 353-357). IEEE.*

[15] Mamun, M., Chowdhury, S. H., Hossain, M. M., Khatun, M. R., & Iqbal, S. (2025). Explainability enhanced liver disease diagnosis technique using tree selection and stacking ensemble-based random forest model. *Informatics and Health, 2(1), 17-40.*

[16] Männikkö, V., Tommola, J., Tikkanen, E., Hätinen, O. P., & Åberg, F. (2025). Large-Scale Evaluation and Liver Disease Risk Prediction in Finland's National Electronic Health Record System: Feasibility Study Using Real-World Data. *JMIR Medical Informatics, 13(1), e62978.*

[17] Kumar, D., Bakariya, B., Verma, C., & Illes, Z. (2025). LivXAI-Net: An explainable AI framework for liver disease diagnosis with IoT-based real-time monitoring support. *Computer Methods and Programs in Biomedicine, 108950.*

[18] Wang, Y., Lei, J., Jin, Z., Jiang, Y., Zhang, N., Lv, M., & Liu, T. (2025). Development and validation of a machine learning-based clinical prediction model for monitoring liver injury in patients with pan-cancer receiving immunotherapy. *International Journal of Medical Informatics, 106036.*

[19] Liu, Y., Meric, G., Havulinna, A. S., Teo, S. M., Åberg, F., Ruuskanen, M., ... & Inouye, M. (2022). Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. *Cell Metabolism, 34(5), 719-730.*

[20] Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer methods and programs in biomedicine, 170, 23-29.*