

A Two-Phase Liver Disease Classification with Stacking Ensemble Models, SHAP Analysis, and Robust Statistical Testing

Sanchita Mohanty, Raj Vikram, Adyasha Rath, Ganapati Panda

Department of Computer Science and Engineering
C. V. Raman Global University, Bhubaneswar-752054, Odisha, INDIA
Corresponding email: ganapati.panda@gmail.com

ABSTRACT:

Liver Disease (LD) is one of the most dangerous risks to human health worldwide. Early detection is essential for timely intervention and improved patient survival. The traditional diagnosis relies on manual interpretation by medical experts, a process that can be time-consuming and prone to human error. Machine Learning (ML) offers a promising alternative, delivering faster and more consistent analysis than manual methods. However, relying on a single ML algorithm is limited by its architectural constraints, and the black-box nature of Artificial Intelligence (AI) restricts its real-life adoption. To address these limitations, this work introduces a strategic, explainable AI-based ensemble approach for efficient detection of LD disorders. A two-level learning approach is employed. The original binary dataset is first converted into a multiclass severity structure using recognized medical biomarkers, giving a more clinically useful label than a simple yes/no output. SMOTE is applied to balance the classes before scaling the features. Twelve classifiers are then evaluated individually, and the three strongest models: Random Forest, XGBoost, and CatBoost are combined into a stacking ensemble through an interpretable logistic regression meta-learner. SHAP is applied throughout to explain feature contributions and confirm the meta-learner's reasoning remains transparent. In a second stage, only Moderate and Severe LD cases are considered, excluding healthy individuals, to give clinicians a sharper basis for treatment decisions. The proposed ensemble achieved an F1-Macro score of 0.8600 (± 0.0158 , 95% CI: 0.8461–0.8739), with a Wilcoxon signed-rank test confirming a statistically significant improvement over the best individual classifier ($p = 0.0312$). This presents a reliable nature of the approach. Its lightweight and interpretable nature enables ease of deployment and adoption in other domains.

Keywords: Liver Disease, Machine Learning, Artificial Intelligence, Ensemble Learning, LR, SMOTE, SHAP.

I. INTRODUCTION

The liver plays a key role in both metabolic and detoxification processes in the human body. Any functional disturbance can lead to various clinical complications. However, it usually starts silently. In many patients, the disease starts with mild biochemical

abnormalities that usually go unnoticed. Over time, these small issues progress to more severe damage. By the time patients seek medical attention, the Liver Disease (LD) may have already reached an advanced stage. LD is one of the leading health concerns worldwide. According to a study [1], It affects millions of people every year. This delay presents the importance of early detection of it.

Identifying LD requires a series of lab tests. Medical experts go through their reports to make a final judgment. However, this can be costly and requires time. In addition, the final result is also prone to errors due to misclassifications. A single high or low value of parameters may not be able to pinpoint the problem. An extensive analysis of different parameters and their synergetic effect can lead to a proper detection of LD. As a result, Machine Learning (ML) came into play.

ML models provide the ability to process large data and provide outcomes based on insights. It can detect abnormalities, learned from the datasets. They provide a third eye to doctors that can help them in decision-making. As a result, they are widely used in healthcare for different purposes. However, their clinical adoption faces various changes.

Major ML models are efficient in binary classification. However, the classification of patients as healthy or not is not enough in real-life scenarios. The treatment changes based on severity. In addition, relying on a single model for classification can lead to inaccurate answers. They are limited due to their architectural limitations. It can lead to incorrect classification and treatment. It presents the need for improved detection approaches.

Although Deep Learning (DL) and other subsets of Artificial Intelligence (AI) are termed more efficient than ML [2], the limitations of dataset and need for speed limits their use. Ensembling ML models is an effective solution to the above-mentioned problems [3]. It can improve the overall performance without increasing substantial load on the device. However, existing studies present limited analysis of ensembling ML approaches, especially on LD. Furthermore, the black-box nature of AI models hinders their real-life adoption in medical domains.

To address it, an eXplainable AI (XAI)-based stacking ensemble model is proposed to detect LD efficiently. A comparative analysis was carried out on 12 ML models. The best performing models were ensembled together with Logistic Regression (LR) as a meta-classifier. LR is

selected because of its linear structure, simplicity, and effectiveness in interpreting the results of different models. Instead of the traditional binary classification problem, this study presents LD as a multiclass problem. It identifies the severity of the LD and presents the insights behind it. SHapley Additive exPlanations (SHAP) is used for explainability [4]. Different callbacks and extensive analysis ensure proper training of the model. The proposed approach secured an accuracy of 98% and various other nearideal parameters. It is examined on a publicly available LD dataset.

The remaining sections of the paper are segmented as follows: Section II examines technological advancements in the domain. Section III explains the proposed approach in detail. Section IV analyzes the results achieved through various experiments. Section V concludes the work and discusses potential future research directions.

II. RELATED WORKS

The researchers in [5] analyzed several ML models and explainable AI (XAI) to improve the early diagnosis of liver cirrhosis. It was found that eXtreme Gradient Boosting (XGBoost) outperformed other by achieving 91.19% after applying hyper-tuning with K-fold validation (K=10). The approach was evaluated using a public cirrhosis prediction dataset. LIME and SHAP were used to explain how different features influenced predictions.

The researchers in [6] show that routine clinical tests supported by ML contain important information to predict LD accurately. A comparative analysis among various ML models were conducted, such as Linear Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), Multi-Layer Perceptron (MLP), and Naive Bayes (NB).

The researchers in [7], [8] conducted a comparative study on different ML models to identify the best model for LD detection. These methods included NB, DT, RF, LR, Artificial Neural Networks (ANN), and Bernoulli Naive Bayes (BNB). RF was found to be the best among all in both studies.

From the above studies, it can be observed that there is a need for enhanced ML models. Standalone models have their own architectural limits that may affect the outcome. Some researchers did not handle the class imbalance. In addition, there is limited use of XAI techniques that remove the blackbox nature of AI models. To address it, the proposed XAIbased ensemble model can be used. Strategic ensembling improves performance and creates synergetic strength to reduce individual limitations.

III. PROPOSED METHODOLOGY

The methodology basically divided in two step as mentioned in “Fig. 1.” The details are discussed subsequently.

A. Data Acquisition

The publicly available LD dataset is used in this study [9]. It consists of 583 patient records. It consists of parameters from clinical reports. It consists of numerical values, segregated by 11 columns in total. The final input to the models consists of 10 features, such as total Bilirubin, direct Bilirubin, total proteins, etc. The final result is presented in binary labels: 1/2 to represent healthy and unhealthy.

B. Enhancing Prediction: Binary-to-Multiclass Conversion

The original binary labels of the dataset have been converted to multiclass classification using appropriate medical biomarkers: 0 is considered as healthy, 1 is considered as mild, 2 is considered as sever.

Various practical thresholds are used to classify patients based on the severity of liver dysfunction. A patient is considered abnormal, if any of the following criteria were met: Total Bilirubin > 3 mg/dL, Direct Bilirubin > 1 mg/dL, ALT > 150 U/L, AST > 150 U/L, Alkaline Phosphatase (ALP) > 300 U/L, Albumin < 2.8 g/dL, or an Albumin/Globulin (A/G) ratio of > 0.8. Some of these values are well supported by clinical scoring systems, such as the Child–Pugh classification, where bilirubin above 3 mg/dL and albumin below 2.8 g/dL are recognized as severe [10]. Similarly, ALT values greater than 150 U/L are in keeping with primary care guidance that recommended a specialist evaluation for determined elevations [11]. An A/G ratio of around or below 0.8 has also been highlighted in clinical trials as an indicator of poorer outcomes [12]. For the remaining markers, we selected thresholds that clearly indicate abnormal biochemical states to provide a consistent method for identifying significant liver dysfunction within our dataset [13].

C. Data Preprocessing

Preprocessing techniques were used to ensure the dataset was suitable for improving the ML model. The following techniques were used:

- **Missing Value Imputation:** A few features were observed to be missing from some samples. To handle them, appropriate values were imputed based on the data.
- **Duplicate Value Cleaning:** Duplicate samples were removed to ensure a smooth performance and a reliable analysis.
- **Data Transformer:** Numerical features were normalized using MinMaxScaler. It transforms the data within a defined range of [0,1]. It provides smooth convergence and stability

- **Class Balance:** Since the dataset had uneven samples per class; the Synthetic Minority Oversampling Technique (SMOTE) has been used to manage it. Stratified sampling has been used for it.

D. Robustness Evaluation

To ensure the reliability and robustness of the models under varying different training data conditions, a robustness evaluation framework was structured. The workflow is as follows:

- In each iteration, a random subset of data was selected from the training data, ranging from 1%

to 70% without replacement. 70% data was selected for training from the total sample size

- The data of the selected subset was used to train the model, and predictions were made on the fixed test set, which is 30% of the total sample size.

This method allowed researchers to test the data. When the training data changes repeatedly in every iteration process or is limited. With the help of this method, researchers check how stable and reliable the model's performance is.

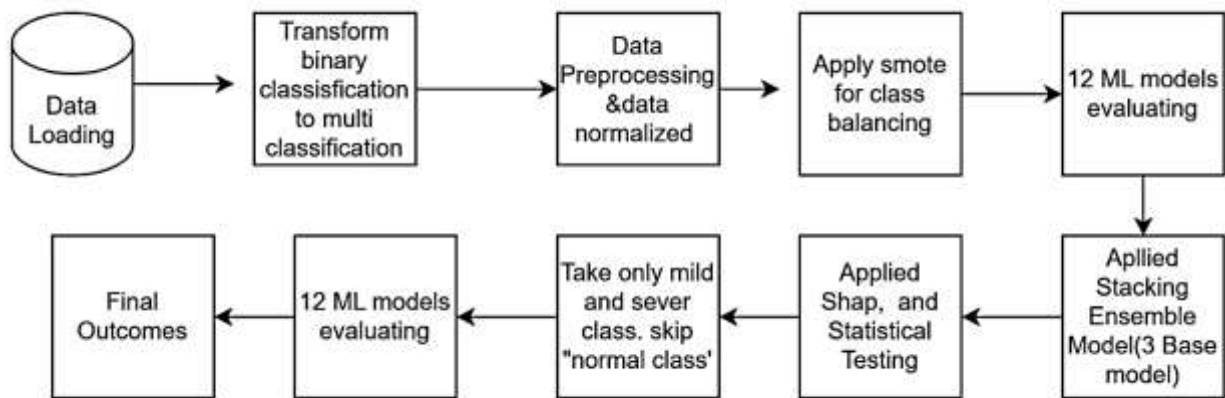


Fig. 1. Outline of proposed methodology

E. Model Development

After preprocessing, the clean and noiseless dataset was used to train the ML classifiers. The classifiers are selected based on the analysis of existing research works and their effectiveness in medical data classification tasks. Linear models, including LR and Ridge Classifier, were used to establish baseline performance. SVM, with both linear and non-linear kernels, were applied to capture more complex patterns in the data. Instance-based learning was represented by the KNN, while Decision Tree and RF models were used to explore treebased learning strategies. However, ensemble methods such as the Gradient Boosting Classifier, XGBoost, LGBM, and CatBoost were applied to leverage boosting techniques for stronger predictive performance.

F. Ensemble Learning

In our study, we selected RF, XGBoost, and CatBoost as base models because each represents a distinct and powerful ensemble strategy: bagging, gradient boosting, and an advanced boosting method with a meta-learner LR, optimized for tabular data, respectively. Their combined strengths ensure a robust and generalizable prediction model.

- **RF:** strong in small data and handles noise effectively.

- **XGBoost:** Highly effective in detecting non-linear relationships.
- **CatBoost:** strong with categorical and naturally manages imbalanced features.
- **LR:** Standard choice in stacking research because it is Lightweight, interpretable, and stable performance.

G. Model Explainability

To make our stacking model more robust and reliable, we applied SHAP to see how different features influenced. Since our problem has multiple classes, SHAP creates importance scores for each base model separately. We then subsequently combined these scores and took their average. So we could identify the features that consistently guided the model's decisions. This provides us with a more stable, clear, balanced view of which features are truly important across the whole ensemble.

We also examined the LR meta-model to understand how much each base model contributes to the overall prediction. Positive values mean the model supports that class, while negative values mean it drives the decision in the opposite direction. The LR meta-model showed that RF contributed the largest weight for class-0 in the stacked model, followed by CatBoost contributed for Class-1 and XGBoost contributed for Class-2.

Overall, this explainability setup allows us to understand two important aspects:

- Feature-Level Explainability: why the model considers certain features important (SHAP-based).
- Model-Level Explainability: how much each base classifier influences the final stacked prediction (meta-learner coefficients).

H. Classification of Moderate and Severe LD

The proposed approach focuses on patients who had already been identified as having LD or not. The objective was to gently distinguish between mild and severe cases, without adding healthy patient samples, which is important to support clinical decisions. To ensure that the models fairly learn from both classes, we first standardized the features using MixMaxScaler and applied SMOTE to balance the mild and severe classes without losing the original structure of the data. It was applied only to the training data.

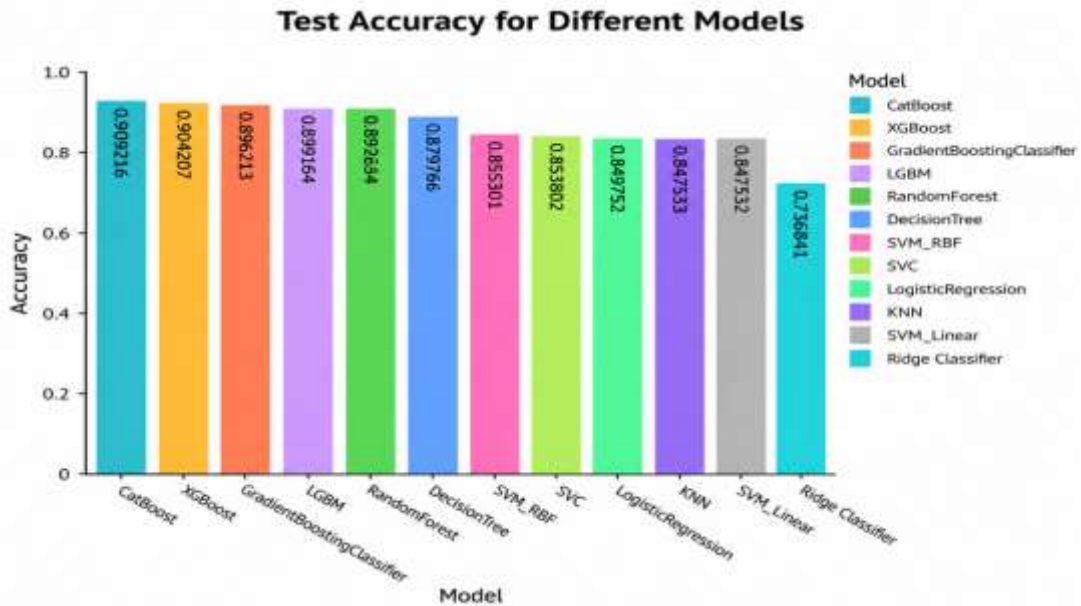


Fig. 2. Various ML models test accuracy metrics for multiclass

We trained the same ML models that were already applied before. Each model was experimented with in a separate test set, and performance was examined using accuracy, precision, recall, and F1-score. This structured, two-step process allows the system to first identify disease patients and then provide a nuanced assessment of severity, aiming to support medical professionals with reliable and interpretable predictions.

I. Model Evaluation and analysis

To evaluate the predictive performance of each model, multiple classification metrics were used. It includes accuracy, precision, F1 score, recall, and ROC AUC Curve.

The stacking meta-learner assigned a higher weight to RF for class 0, to XGBoost for class 1, and to CatBoost for class 2, indicating that each base model specialized in different severity categories, and stacking combined their strengths.

IV. RESULTS AND PERFORMANCE EVALUATION ANALYSIS

A. Experimental Setup

The experiments are conducted in Google Colab CPU in its default computational environment. It uses Python 3.10 and its libraries to design the model. During training, no GPU acceleration or specialized hardware was used. Memory consumption remained during the evaluation process, with usage approximately near 2.5 GB, confirming that the models operated efficiently.

B. Result Analysis

The proposed approach is evaluated using different types of classification to evaluate the classification and reliability of the model. The original binary labels are converted into three classes to better capture the progression of LD. Using this conversion, we trained and evaluated 12 different ML models. Their accuracies are presented in “Fig. 2,” while the corresponding ROC–AUC curves are shown in “Fig. 3.” To further improve predictive performance, we

developed a stacking ensemble model. We selected RF, XGBoost, and

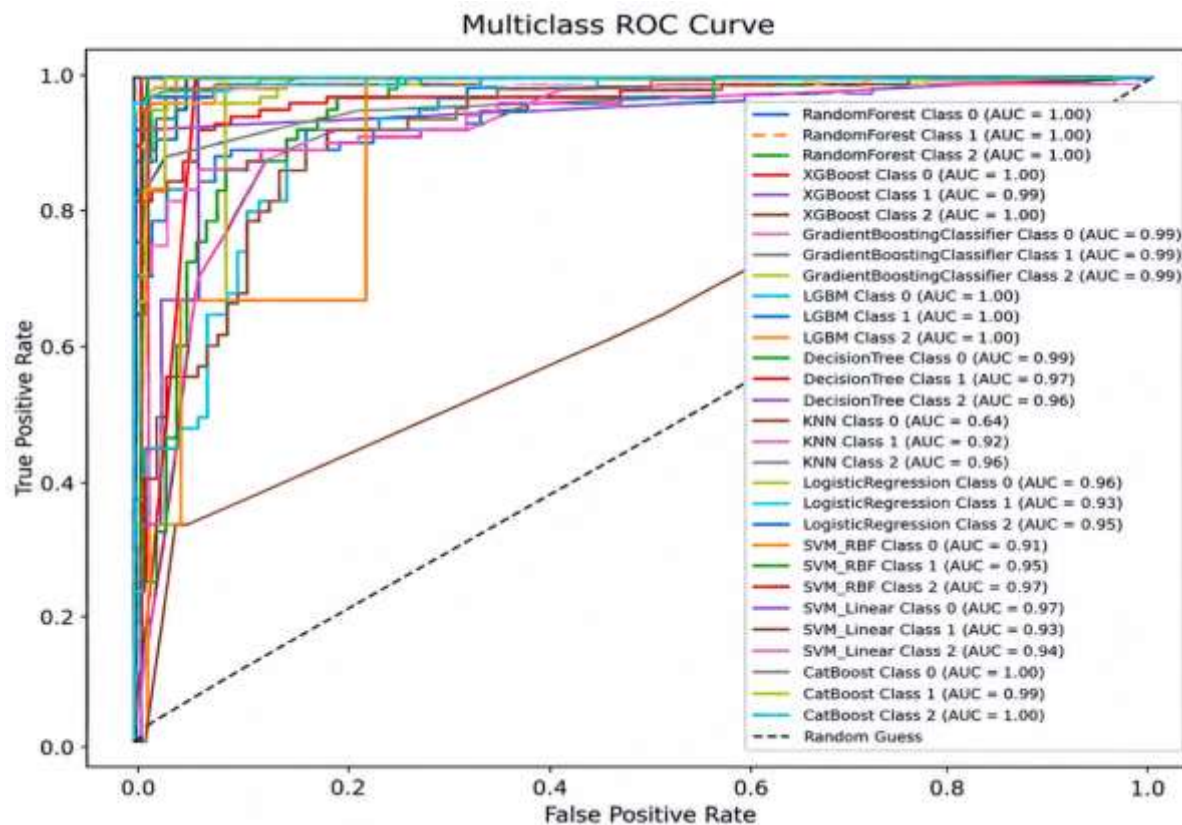


Fig. 3. ROC AUC curve for multiclass analysis.

CatBoost for the base model because each reflects a complementary ensemble strategy: bagging, gradient boosting, and advanced boosting, respectively. The LR classifier served as the meta-learner for its transparent

behaviour. The performance of the stacking ensemble is presented in “Fig.4”.

```
*** Stacking Accuracy: 0.9766081871345029
```

	precision	recall	f1-score	support
0	0.75	1.00	0.86	3
1	0.96	1.00	0.98	65
2	1.00	0.96	0.98	103
accuracy			0.98	171
macro avg	0.90	0.99	0.94	171
weighted avg	0.98	0.98	0.98	171

```
Confusion Matrix:
[[ 3  0  0]
 [ 0 65  0]
 [ 1  3 99]]
```

Fig. 4. Classification report of the stacking model

To enhance the interpretability, we applied SHAP analysis only to the ensemble base models. This allows identification of the biochemical markers that contribute to the prediction process, as mentioned in “Fig. 5.”

We also combined SHAP scores to give a single, to see how each clinical feature influences the model’s predictions. Then examined the coefficients of the LR meta-learner to understand how each feature contributed to the stacked prediction; these values are provided in

“Fig. 6.” A statistical test is a mathematical technique used to examine whether the observed differences in experimental results are statistically

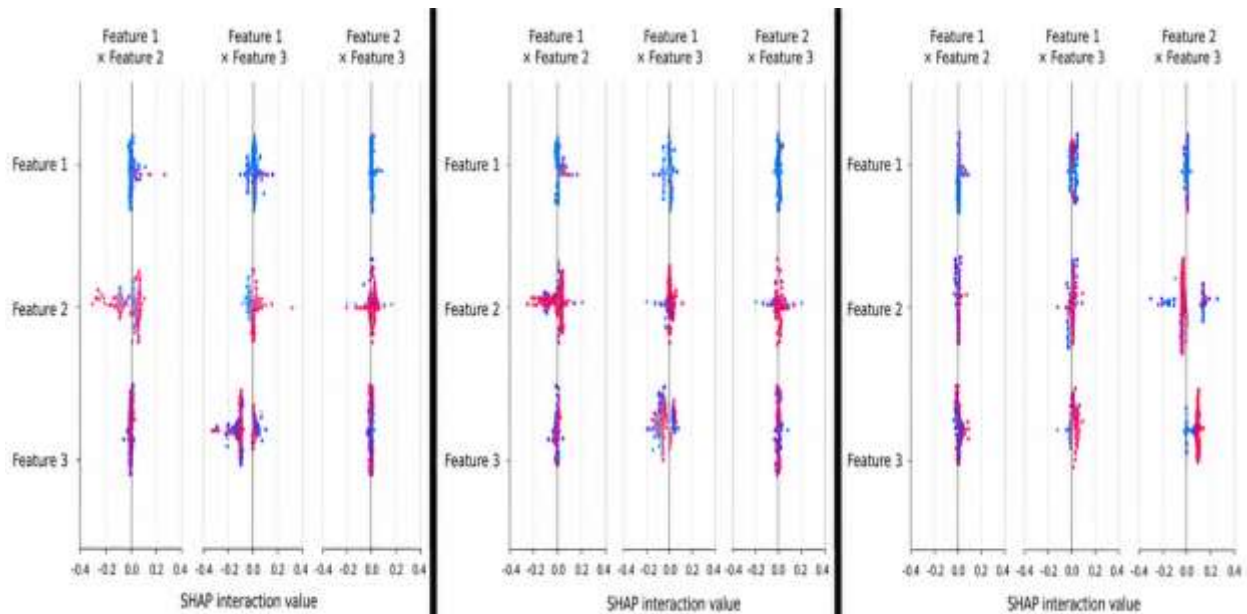


Fig. 5. Feature-wise SHAP interaction for Random Forest, CatBoost, XGBoost

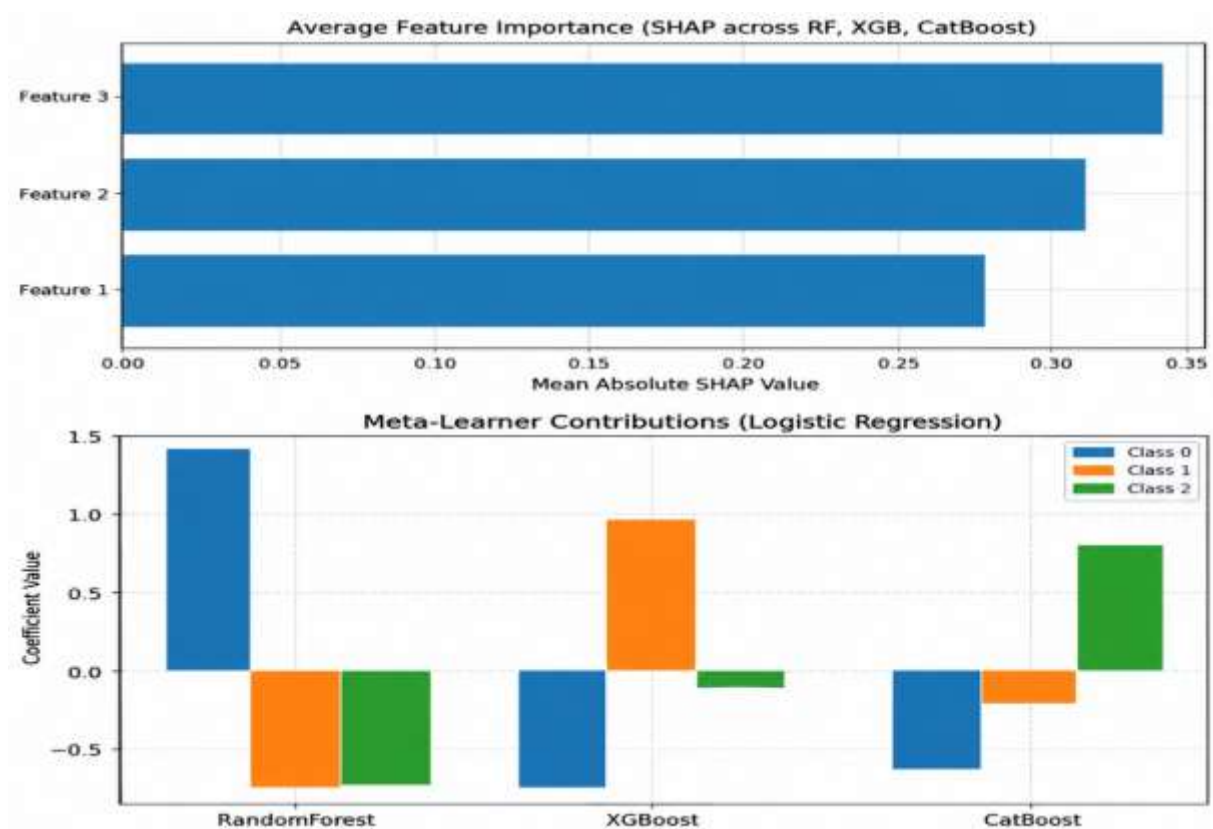


Fig. 6. Comparative SHAP based feature importance for base model, and Meta Learner.

analysis, probability calibration, and learning curve analysis that showed whether the models truly behaved differently. These results are provided in “Fig. 7.”

In an additional experiment, we simplify the problem by excluding the “normal” class and focusing only on mild and severe classes. This ensured how fairly the models were evaluated and the learning process between the two classes. Here also the same set of 12 ML models was implemented using the two-class. The

improved accuracy and ROC-AUC scores are shown in “Fig. 8” and “Fig. 9.”

Overall, these analyses demonstrate that the stacking ensemble consistently outperformed individual models, and SHAPbased interpretability provided valuable insights into the most influential biomarkers associated with the severity of LD. analysis, probability calibration, and learning curve analysis

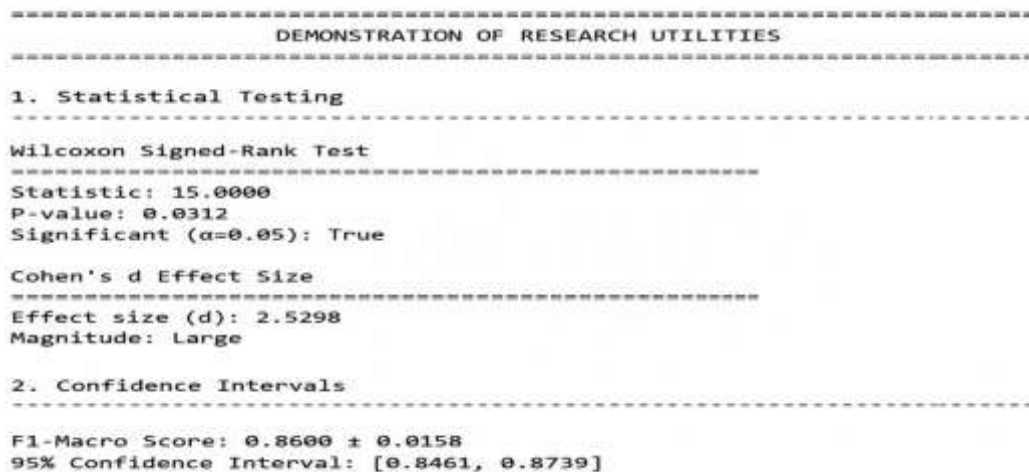


Fig. 7. Statistical testing of the Ensemble Model.

that showed whether the models truly behaved differently. These results are provided in “Fig. 7.”

In an additional experiment, we simplify the problem by excluding the “normal” class and focusing only on mild and severe classes. This ensured how fairly the models were evaluated and the learning process

between the two classes. Here also the same set of 12 ML models was implemented using the two-class. The improved accuracy and ROC-AUC scores are shown in “Fig. 8” and “Fig. 9.”

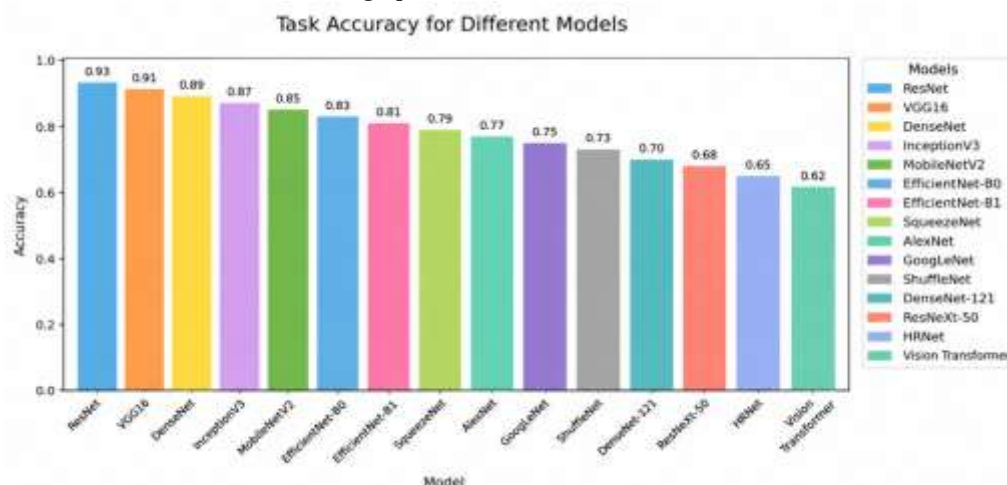


Fig. 8. Accuracy of Various ML models on Mild Vs Severe Classification.

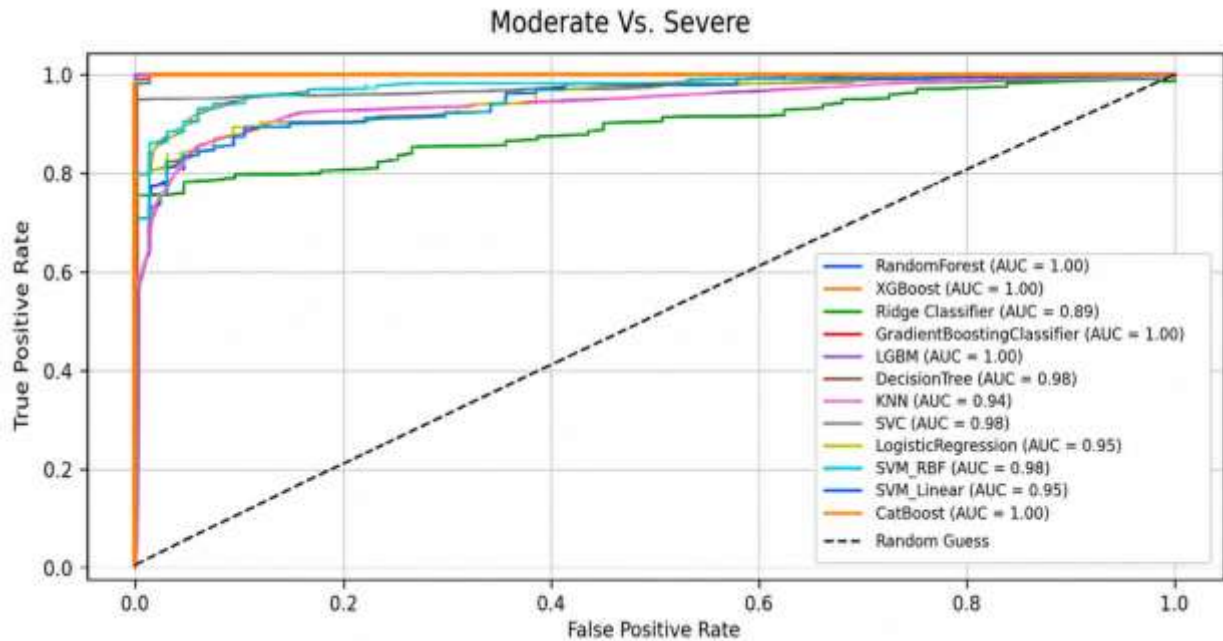


Fig. 9. ROC AUC curve analysis of various ML models

Overall, these analyses demonstrate that the stacking ensemble consistently outperformed individual models, and SHAPbased interpretability provided valuable insights into the most influential biomarkers associated with the severity of LD.

V. CONCLUSION

This study presents a XAI-based ensemble model to classify and identify the severity of LD. Performance of various individual ML models were assessed. The best ones were stacked ensembled with logistic regression as meta-learner. Instead of using class labels as input, the meta-learner was provided the probability scores, so it could weigh in on the models' confidence levels. Nested cross-validation and use of SMOTE ensured proper training procedure. A publicly available LD dataset has been used to train and evaluate the model. It achieved a test accuracy of the ensemble stacking model of 98%. This gain implies the efficacy of the proposed approach. LR-based meta-learner ensures that the results are not solely based on the biasness of one model. The explainability of SHAP provides important insights on the model's prediction. AST and ALT, followed by Albumin, were observed as the primary features used for classification. To evaluate the potential of the proposed approach, large numerical dataset is needed. The limited presence of these data hinders the real-life deployment of the approach. The evaluation of these models in diverse

biomedical scenarios with medical devices to establish an end-to-end cycle represents an important direction for future research. In addition, a multimodal explainable analysis for patients is essential for advanced clinical decision-making.

REFERENCES:

- [1] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, "Burden of liver diseases in the world," *Journal of Hepatology*, vol. 70, no. 1, pp. 151–171, 2019, pMID: 30266282.
- [2] K. Eskandar, "Artificial Intelligence in Hepatology: A Comprehensive Scoping Review of Clinical Applications, Challenges, and Future Directions," *iLIVER*, p. 100205, 2025.
- [3] S. Biswas, M. A. Islam, J. Rout, M. Mishra, and D. Muduli, "Passive Image Forgery Detection Through Ensemble Learning with ELA," in *2024 International Conference on Computer, Electronics, Electrical Engineering & their Applications (IC2E3)*, 2024, pp. 1–6.
- [4] N. A. Khan, M. F. Bin Hafiz, M. A. Pramanik, S. Hossain, S. Barman, and N. Hossain, "Machine Learning and Explainable AI for Liver Disease Prediction: An Integrated Interpretability Framework," *Biomedical Materials & Devices*, pp. 1–14, 2025.
- [5] K. Rao Makkena and K. Natarajan, "Enhancing Liver Cirrhosis Diagnosis Using Machine Learning With Explainable AI and Cross-Validated Hyperparameter Tuning Techniques," *IEEE Access*, vol. 13, pp. 93323–93335, 2025.
- [6] B. Ginter-Matuszewska, A. Adamek, M. Majchrzak, B. Rozplochowski, A. Zientarska, A. Kowala-Piaskowska,

-
- and P. Lukasiak, “FibrAIm – The machine learning approach to identify the early stage of liver fibrosis and steatosis,” *International Journal of Medical Informatics*, vol. 197, p. 105837, 2025.
- [7] S. Tokala, K. Hajarathaiah, S. R. P. Gunda, S. Botla, L. Nalluri, P. Nagamanohar, S. Anamalamudi, and M. K. Enduri, “Liver disease prediction and classification using machine learning techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, 2023.
- [8] A. E. Topcu, E. Elbasi, and Y. I. Alzoubi, “Machine Learning-Based Analysis and Prediction of Liver Cirrhosis,” in *2024 47th International Conference on Telecommunications and Signal Processing (TSP)*, 2024, pp. 191–194.
- [9] Kaggle, “Indian liver patient records dataset,” <https://www.kaggle.com/datasets/uciml/indian-liver-patient-records>, 2019, accessed: 2025-09-21.
- [10] J. T. Pugh, P. C. Murray-Lyon, J. L. Dawson, M. C. Pietroni, and R. Williams, “Transection of the oesophagus for bleeding oesophageal varices,” *British Journal of Surgery*, vol. 60, no. 8, pp. 646–649, 1973.
- [11] N. B. C. Guidelines, “Abnormal liver function tests (LFTs) in asymptomatic patient in primary care,” <https://rightdecisions.scot.nhs.uk/nhsborders-clinical-guidelines/acute-services/general-medicine/abnormalliver-function-tests-lfts-in-asymptomatic-patient-in-primary-care/>, 2023, accessed: 2025-09-20.
- [12] Y. Zhou, X. Liu, J. Zhang, Y. Xu, and X. Xu, “Clinical significance of albumin to globulin ratio in critically ill patients,” *Frontiers in Nutrition*, vol. 9, pp. 1–11, 2022.
- [13] S. Sharma and H. Nagalli, *Liver Function Tests*. Treasure Island, FL: StatPearls Publishing, 2024, available online: <https://www.ncbi.nlm.nih.gov/books/NBK482489/> (Accessed: 2025-0920).
-
- Cite this article as:*
- Sanchita, Ganapati, and et. al. "A Two-Phase Liver Disease Classification with Stacking Ensemble Models, SHAP Analysis, and Robust Statistical Testing", Proceedings of Applied Energy Systems and Computer Science, 2025, display as online February 2026.*
- Link: <http://actsoft.org/science/act2025-pro/77-esda2025.pdf>,*
- @Copyright to 'Applied Computer Technology', Kolkata, WB, India. Website: actsoft.org, Email: info@actsoft.org, ISBN: 978-81-985770-9-2*