Received: 6 November 2024 | Revised: 11 February 2025 | Accepted: 3 March 2025 | Published online: 20 March 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–13
DOI: 10.47852/bonviewMEDIN52024744

BON VIEW PUBLISHING

An Ensemble Approach for Artificial Neural Network-Based Liver Disease Identification from Optimal Features through Hybrid Modeling Integrated with Advanced Explainable AI

Safiul Haque Chowdhury[1,*], Mohammad Mamun[1], Md. Tanvir Ahmed Shaikat[1], Mohammed Ibrahim Hussain[1], MD. Sadiq Iqbal[1,2] and Muhammad Minoar Hossain[1,3]

[1]*Department of Computer Science and Engineering, Bangladesh University, Bangladesh*

[2]*Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Bangladesh*

[3]*Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Bangladesh*

Abstract: Liver disease is any condition that negatively affects the liver's function or structure, resulting in impaired liver function and various health complications. Abnormal conditions are rapidly increasing day by day. In this study, we used a dataset of key liver disease-related blood sample biomarkers to utilize various Machine learning (ML) techniques to enhance the accuracy of liver disease prediction. Specifically, we integrated the artificial neural network (ANN) model with five ML models: Stacked Generalization (Stacking), Bootstrap Aggregating (Bagging), Adaptive Boosting (AdaBoost), Gradient-Boosted Decision Tree (GBDT), and Support Vector Machine (SVM)—resulting in five distinct hybrid models: Stacking with ANN (SANN), Bagging with ANN, AdaBoost with ANN (ABANN), GBDT with ANN (GANN), and SVM with ANN (SVMANN). We tested all these hybrid models with feature selection techniques, including linear discriminant analysis (LDA), principal component analysis (PCA), recursive feature elimination (RFE), and also without feature selection. Through extensive testing, we found that these five hybrid models performed best when combined with LDA rather than PCA, RFE, or no feature selection. This discovery led us to create a max voting ensemble (MVE) of these LDA-optimized hybrid models. Remarkably, our prediction accuracy increased from 79.15% to 98.38% using the MVE. Furthermore, we employ Explainable Artificial Intelligence techniques such as Local Interpretable Model-agnostic Explanations, Shapley Additive Explanations, and Individual Conditional Expectations to analyze and enhance trust in the predictions. We also implemented 10-fold cross-validation to ensure the robustness and reliability of our results. This research underscores the significance of advancements in neural network systems and highlights the potential for hybrid models to improve predictive accuracy in liver disease diagnosis. Our findings pave the way for a new generation of computational technologies endowed with intelligence, ultimately contributing to better health outcomes and a deeper understanding of liver disease dynamics.

Keywords: liver disease, machine learning, artificial neural network, explainable artificial intelligence

1. Introduction

Liver disease is a primary global health concern, affecting millions and burdening healthcare systems. Early detection and accurate prediction can significantly improve patient outcomes [1]. Over 100 million people in the U.S. have liver disease, with 4.5 million diagnosed and an estimated 80–100 million with fatty liver disease. Untreated liver disease can lead to liver failure and cancer, resulting in 51,642 deaths in 2020 (15.7 per 100,000). Chronic liver disease/cirrhosis was the 12th leading cause of death in 2020 and the 8th for non-Hispanic African American/Black individuals aged 45–64 in 2019. Prevalence rates from a 2016 study include Japanese Americans (6.9%), Hispanic/Latino persons (6.7%), White persons (4.1%), and African American/Black and Native Hawaiian persons (3.9%). Nonalcoholic fatty liver disease is the most common cause of cirrhosis, with risk factors including heavy alcohol use, obesity, type 2 diabetes, and certain medical and lifestyle factors. Cirrhosis also increases stroke risk, with an incidence of 2.17% annually compared to 1.11% without cirrhosis. Death rates from liver cirrhosis have been higher for Black/African American men and women than for their white counterparts since the 1950s [2].

Recent research efforts have extensively explored the utilization of machine learning (ML) techniques for predicting liver disease, showcasing a variety of methodological approaches and their corresponding accuracies. All these studies focused on classification

*Corresponding author: Safiul Haque Chowdhury, Department of Computer Science and Engineering, Bangladesh University, Bangladesh. Email: safiul.haque@bu.edu.bd

for liver disease prediction. For instance, Choubey et al. [3] adopted Decision Tree (DT) algorithms and achieved an accuracy of 75.10%, while Shetty and Satyanarayana [4] enhanced Support Vector Machine (SVM) with Random Sampling for a 71% accuracy rate. Alyabis et al. [5] turned to Neural Network Analysis and obtained a 79.6% success rate, and Singh and Agarwal [6] experimented with an Extreme Learning Machine (ELM), resulting in 77.77% accuracy. Further contributions include Azam et al. [7], who integrated K-Nearest Neighbor (KNN) with Feature Selection Techniques (KNNWFST) for a 74% accuracy, and Choudhary et al. [8], who applied Logistic Regression (LR) with a 70.54% accuracy rate. Additional studies by Khan et al. [9] and Kannapiran et al. [10] utilized Random Forest (RF) and LR to achieve accuracies of 72.17% and 73.97%, respectively. Muthuselvan et al. [11] used Random Tree, and Yasmin et al. [12] studied KNN, yielding 74.2% and 76.03% accuracy, demonstrating the diverse range of ML methodologies being explored for liver disease prediction.

In our study, we critically analyzed the limitations and scopes of these previous studies, seeking to bring novelty to our research methodology. In the discussion section, we provide a comprehensive comparison of these studies with our findings to highlight the advancements and contributions of our approach.

In recent years, the emergence of advanced computational techniques, such as artificial neural network (ANNs) and Explainable Artificial Intelligence (XAI), has provided promising avenues for enhancing the predictive capabilities of liver disease diagnosis models. This research investigates the potential of ANN-based models integrated with XAI techniques for predicting liver disease from optimal features extracted from patient data. Unlike traditional statistical methods, ANNs offer the advantage of learning complex patterns and relationships from large datasets, enabling more accurate and robust predictions. Moreover, incorporating XAI methods allows for interpreting and understanding the ANN model's decision-making process, addressing the critical need for transparency and explainability in medical AI systems [13]. The primary objective of this study is to develop ANN-based models trained on a comprehensive dataset of clinical variables associated with liver disease, utilizing feature selection techniques to identify the most informative features for prediction. By leveraging XAI methods, such as Local Interpretable Model-agnostic Explanation (LIME) and Shapley Additive Explanation (SHAP), we aim to elucidate the underlying factors driving the model's predictions, enhancing its interpretability and trustworthiness.

This research is motivated by the potential of advanced computational techniques to revolutionize medical diagnostics and decision-making processes. By harnessing the power of ANNs and XAI, we aim to develop more accurate, transparent, and clinically relevant predictive models for liver disease. Specifically, this study integrates ANNs with robust ML models to enhance predictive accuracy, employs advanced XAI tools to ensure transparency in decision-making, and optimizes feature selection to target the most informative clinical variables for liver disease prediction. These advancements can potentially inform clinical practice and improve patient outcomes through early detection and personalized treatment strategies.

2. Materials and Methods

The main goal of this study is to accurately predict liver disease by employing various ANN-based hybrid models and subsequently assembling them for improved performance. The research workflow is outlined in Figure 1. Sections 2.1 to 2.9 provide a brief working structure of the study.

2.1. Dataset

We obtained a dataset from the UCI ML Repository [14] containing 583 samples of individuals, both affected and unaffected by liver disease. The dataset comprises 10 features, excluding the target variable indicating the presence or absence of liver disease. Of the 583 instances in the dataset, 416 samples are affected by the disease, while the remaining 167 are free. These 10 features contain vital information related to various blood parameters and liver conditions, including Age, Gender, Total Bilirubin (TB), Direct Bilirubin (DB), Alkaline Phosphatase (ALPH), Alanine Aminotransferase (ALAT), Aspartate Aminotransferase (ASAT), Total Proteins (TP), Albumin (AL), and Albumin and Globulin Ratio (AGR). The dataset comprehensively represents individuals' liver health features, incorporating key biochemical markers and demographic information. This diverse set of features will serve as the foundation for constructing and evaluating predictive models for liver disease diagnosis. Additionally, Table 1 provides a detailed description of all 10 features and their corresponding value types, facilitating a better understanding of the dataset's composition and characteristics.

2.2. Analysis and Visualization

Data analysis and visualization are crucial in understanding datasets, especially when applying different ML models [15]. These techniques provide valuable insights into the distribution, patterns, outliers, and relationships within the data, essential for making informed decisions during model development, feature selection, and evaluation. In our liver dataset analysis, we utilize various visualization techniques, including histograms [16], violin plots [17], and correlation heatmaps [18].

2.3. Preprocessing

We employed various preprocessing techniques to address missing values and transform textual values into numerical representations [19]. In our dataset, we encountered missing values in the "AGR" feature, totaling four instances. Additionally, we converted gender values, where females were represented as one and males as 0. Missing values in the dataset were addressed using data imputation techniques. Specifically, we utilized the mean imputation method to fill in the missing values of the AGR feature. The mean imputation formula is given as follows:

$$Mean(X) = \frac{\sum x}{n}$$

Here, $Mean(X)$ is the mean value that is used to fill in missing values in the dataset, $\sum x$ denotes the sum of all non-missing values in the feature, and n indicates the total number of non-missing values in the feature.

We employed one-hot encoding to convert textual values into numerical representations [20]. This technique transforms categorical variables into binary vectors, effectively representing each category as a separate feature. In our case, we encoded gender information, where females were mapped to 1 and males to 0.

2.4. Ideal feature finding

The process of selecting the most relevant and informative features from a dataset to improve the performance of ANN-based ML models. This step is essential in building efficient and

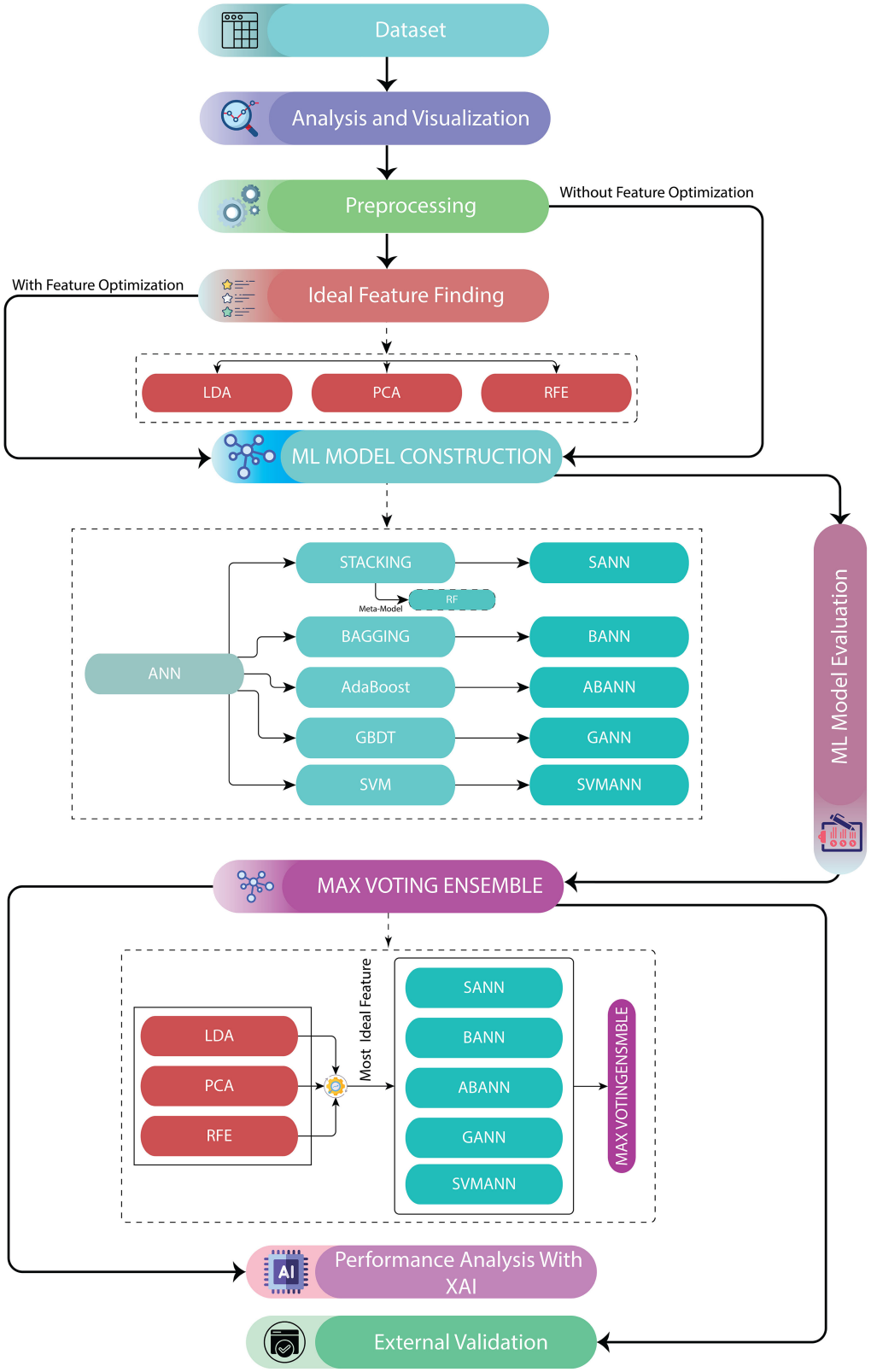

Figure 1. Working outline of the research

accurate predictive models as it helps reduce dimensionality, minimize overfitting, and enhance model interpretability.

In this study, we applied feature selection techniques such as linear discriminant analysis (LDA), principal component analysis (PCA), and recursive feature elimination (RFE) [21] to identify optimal features for ANN-based ML models, aiming to enhance predictions of liver disease outcomes. We identified the most compelling feature selection approach among the tested methods

Table 1. Analysis of different features of the dataset

Feature	Description	Value type	Unit
Age	Represents the age of the individual	Numerical	Years
Gender	Indicates the gender of the individual	Nominal	–
TB	Measures the total amount of bilirubin in the blood, indicating liver function and potential abnormalities.	Numerical	μmol/L
DB	Elevated levels of direct bilirubin may indicate obstructive liver disease	Numerical	μmol/L
ALPH	Reflects the levels of alkaline phosphatase enzyme in the blood, produced by the liver, bones, and other tissues. Elevated levels of alkaline phosphatase may indicate liver or bone disorders.	Numerical	μkat/L
ALAT	Measures the levels of alanine aminotransferase enzyme in the blood, primarily found in the liver. Elevated levels of alanine aminotransferase may indicate liver damage or disease.	Numerical	U/L
ASAT	Indicates the levels of aspartate aminotransferase enzyme in the blood, which is also predominantly found in the liver. Elevated levels of aspartate aminotransferase may suggest liver damage or inflammation.	Numerical	U/L
TP	Represents the total protein concentration in the blood, including albumin and globulin. Abnormal levels of total proteins may indicate liver disease or other underlying health conditions.	Numerical	g/dL
AL	Specifies the albumin concentration in the blood, which is synthesized by the liver and plays a crucial role in maintaining osmotic pressure and transporting various substances in the blood.	Numerical	g/dL
AGR	It provides the ratio of albumin to globulin in the blood, which can indicate liver function and overall health. Albumin and globulin ratio abnormalities may suggest liver disease or other underlying conditions.	Numerical	–
Result	Individuals with liver disease are labeled “1”, while those without liver disease are labeled “0”.	Nominal	–

and integrated it into our models to improve predictive performance. Furthermore, we employed a max voting ensemble (MVE) technique to combine multiple ANN models utilizing the best feature subset, significantly boosting accuracy and robustness.

LDA is a dimensionality reduction technique that finds linear combinations of features to best separate different classes or categories in the data. It is commonly used for classification tasks to maximize the separation between classes while minimizing the variance within each class [22].

PCA is another dimensionality reduction technique that transforms the original features into a lower-dimensional space while preserving as much variance as possible. PCA identifies the principal components that capture the most significant variation in the data, allowing for dimensionality reduction and simplification of the dataset [23].

RFE is a feature selection method that recursively removes features based on their importance from the dataset. It trains the model on the remaining features and evaluates their performance, continuing this process until the optimal subset of features is identified. RFE helps select the most informative features while discarding redundant or irrelevant ones, thereby improving model efficiency and interpretability [24].

2.5. ML model construction

Our study thoroughly examined the preprocessed dataset by integrating various ML models with ANN to enhance prediction accuracy. Our approach involved leveraging ANN as the base model and implementing five distinct algorithms. These algorithms are designed to improve predictive performance by incorporating unique methodologies and characteristics. This comprehensive analysis aims to identify the most effective model for accurately predicting liver disease outcomes. Our methodology underscores the importance of neural network systems in maximizing prediction accuracy across different ML models [25].

2.5.1. Stacking with ANN (SANN)

Model stacking, or stacked generalization, is an ML technique that combines multiple models to enhance predictive performance. It trains several base models and uses their predictions as input features for a meta-model, which learns to refine and integrate these predictions. The meta-model addresses errors and biases of individual models, yielding a more robust and accurate prediction [26]. Mathematically, the stacking process with an ANN is as follows:

$$\mathbf{X}_{meta} = [\mathrm{ANN}_1(\mathbf{X}) \quad \mathrm{ANN}_2(\mathbf{X}) \quad \ldots \quad \mathrm{ANN}_n(\mathbf{X}), \ \mathbf{y}]$$

Here, $\mathbf{X}$ represents the input features, $\mathbf{y}$ represents the target variable, ANN$i(\mathbf{X})$ represents the prediction made by ANN model i, and RF($\mathbf{X}_{meta}$) represents the prediction made by the RF as a meta-model. Then, the RF meta-model is trained on $\mathbf{X}_{meta}$:

$$\mathrm{RF}(\mathbf{X}_{meta}) = ([\mathrm{ANN}_1(\mathbf{X}), \ \mathrm{ANN}_2(\mathbf{X}), \ldots, \mathrm{ANN}_n(\mathbf{X})])$$

The performance metrics are then calculated based on the predictions of the RF meta-model.

2.5.2. Bagging with ANN

Bagging is an ensemble method that enhances ML stability and accuracy by bootstrap sampling to create multiple training subsets. Each subset trains a base model, and their predictions are aggregated for the final output [27]. By introducing model diversity, bagging reduces overfitting and improves generalization. This study employs Bagging with ANNs as base models to mitigate prediction variance. While ANNs excel at capturing complex data patterns, they are sensitive to training subsets. Bagging reduces this sensitivity, enhancing prediction stability.

The Bagging Classifier aggregates predictions from multiple ANN base models through averaging. For N base model, predictions are the average of all base model predictions. Mathematically, this is represented as:

$$\mathrm{BC}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{ANN}_i(\mathbf{X})$$

Here, $\mathrm{ANN}_i(\mathbf{X})$ represent the prediction made by the i-th ANN model on the input features $\mathbf{X}$, $\mathrm{BC}(\mathbf{X})$ represent the prediction made by the

Bagging Classifier on the input features. $\mathbf{X}$. N represents the number of ANN models used in the ensemble method.

2.5.3. ABANN

Adaptive Boosting (AdaBoost) enhances classification performance by combining weak learners, typically shallow DTs, through iterative training that assigns higher weights to misclassified samples [28]. In AdaBoost with ANNs (ABANN), ANNs replace traditional weak learners. Multiple ANNs are trained sequentially, with each focusing more on previous misclassifications. The final ABANN prediction is a weighted sum of individual ANN predictions, with weights based on their accuracy during training. Mathematically, this process is expressed as:

$$\mathrm{ABANN}(\mathbf{X}) = \mathrm{sign}\left(\sum_{i=1}^{N} \alpha_i \cdot \mathrm{ANN}_i(\mathbf{X})\right)$$

Here, $\mathrm{ANN}_i(\mathbf{X})$ represent the prediction made by the i-th ANN model on the input feature. $\mathbf{X}$, $\mathrm{ABANN}(\mathbf{X})$ represent the prediction made by the AdaBoost model on the input features. $\mathbf{X}$, AdaBoost model combines predictions from multiple base models ANN through a weighted sum. Considering N base models and α_i Represents the weight assigned to the i-th base model. The sign function ensures the final prediction is binary, typically $\{-1, 1\}$ in classification tasks. The weights α_i are determined during the training process, favoring models with better performance. This iterative approach of combining multiple ANNs with AdaBoost enhances the model's overall predictive accuracy and robustness.

2.5.4. GBDT with ANN (GANN)

Gradient Boosting, mainly represented by the Gradient Boosting Classifier in scikit-learn, is a powerful ensemble learning technique that builds a strong predictive model by sequentially adding weak learners (typically DT) to an ensemble [29]. Each subsequent weak learner corrects the errors made by the previous ones, leading to a final strong learner that combines the predictions of all weak learners.

In this implementation, Gradient Boosting with ANNs is used as a base learner, and GANN utilizes ANNs instead of DT as weak learners. Mathematically, the GANN model is represented as follows:

$$\mathrm{GANN}(\mathbf{X}) = \sum_{i=1}^{N} \mathrm{ANN}_i(\mathbf{X})$$

Here, $\mathrm{ANN}_i(\mathbf{X})$ represent the prediction made by the i-th ANN model on the input features. $\mathbf{X}$, $\mathrm{GANN}(\mathbf{X})$ represent the prediction made by the Gradient Boosting model on the input features. $\mathbf{X}$ and N represent the number of ANN models.

2.5.5. SVM with ANN (SVMANN)

SVM is a supervised algorithm for classification and regression that identifies the optimal hyperplane to maximize class separation, using support vectors to define the margin [30]. It solves a convex optimization problem to minimize errors and employs kernel functions for non-linearly separable data. In the hybrid SVMANN model, SVM's high-dimensional handling is combined with ANN's ability to capture complex patterns, enhancing classification accuracy.

Let's denote the output of the SVM model as $f_{\mathrm{SVM}}(\mathrm{X})$ and the output of the ANN model as $f_{\mathrm{ANN}}(\mathrm{X})$. Then, the combined prediction $\mathcal{Y}$ SVMANN can be obtained by applying the outputs of both models to a decision function, which could be a simple sum or another function, depending on the specific implementation. Here's the mathematical representation:

$$\mathcal{Y} = \mathrm{decision\_function}(f_{\mathrm{SVM}}(\mathrm{X}) + f_{\mathrm{ANN}}(\mathrm{X}))$$

This equation encapsulates the idea of integrating the predictions from both the SVM and ANN models to form the output of the SVMANN hybrid model.

2.6. ML model evaluation

We evaluate our hybrid ANN-based ML models for liver disease prediction using an 80:20 train-test split, ensuring robust training and a realistic performance assessment. Predictions are analyzed through a confusion matrix (CM), which categorizes predictions into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), forming the basis for calculating key performance metrics [31].

Accuracy is the proportion of correctly classified instances and is calculated as:

$$\frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Precision measures the accuracy of optimistic predictions, which is crucial for minimizing FPs in medical contexts. It is calculated as:

$$\frac{TP}{TP + FP} \times 100$$

Recall reflects the model's ability to identify actual positive cases, ensuring minimal missed diagnoses. It is computed as:

$$\frac{TP}{TP + FN} \times 100$$

$F1$ Score balances precision and recall, comprehensively evaluating the model's performance. It is given by:

$$2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100$$

Based on these performance metrics, we selected the models with the highest Accuracy, Precision, Recall, and $F1$ Score to determine which should be taken forward for the MVE. This approach ensures that only the most reliable models, demonstrating strong diagnostic capabilities in liver disease prediction, contribute to the ensemble, enhancing our final model's robustness and overall predictive accuracy.

To evaluate the stability of the Max Voting model, we generated the standard deviation (SD), 95% confidence interval (CI), and receiver operating characteristic (ROC) curve.

The SD indicates the variability of the model's performance across different runs. A lower SD suggests the model's performance is more stable and consistent [32].

The 95% CI gives us a range of values within which we can be 95% confident that the true model performance lies [33]. This helps us understand the potential variability of the model's effectiveness, ensuring it performs reliably.

The ROC curve was generated to visualize the model's ability to distinguish between classes at various threshold values. A higher area under the curve (AUC) indicates that the model is more stable and reliable in its discrimination between classes [34].

2.7. Max Voting

After evaluating various feature optimization techniques, we found that LDA outperformed PCA and RFE. We then applied

MVE [35] to combine predictions from five hybrid ANN models, leveraging their diverse strengths to improve accuracy and robustness. This approach enhances prediction using ANN's learning capabilities and LDA's discriminative power.

$$\hat{y} = \text{MajorityVote}\left(y_{SANN+LDA}, y_{BANN+LDA}, y_{ABANN+LDA}, y_{GANN+LDA}, y_{SVMANN+LDA}\right)$$

2.8. Performance analysis with XAI

We analyze our best model's predictions using XAI techniques, such as SHAP, LIME, and Individual Conditional Expectation (ICE) plots, to gain transparency into its decision-making process. SHAP attributes prediction contributions to individual features, while LIME provides local explanations for specific predictions. ICE plots reveal feature effects across instances. These techniques enhance the interpretability of our ANN-based Max Voting model, improving its transparency for clinical applications [36–38].

2.9. External validation

To further evaluate the performance of our MVE model, we test it in various ways. We gather real-time patient information from multiple internet sources [39, 40], collecting three patient data sets representing diverse demographic and health conditions. These datasets are then tested against the pre-trained model, developed using a well-established dataset, allowing us to assess how well the model generalizes to unseen real-time data. Additionally, we apply the model to a multiclass classification dataset instead of the original binary classification task to examine its performance with more complex classification problems. This approach helps us evaluate the model's adaptability and scalability across a broader range of potential outcomes. The results from both the real-time patient data and the multiclass dataset provide valuable insights into the model's capabilities and highlight areas for future improvement.

3. Results and Discussion

After preprocessing the liver dataset, we evaluated ANN-based models and applied feature reduction techniques, finding LDA to be the most effective. We then implemented a MVE model with LDA, addressed outliers through scalarization, and used 10-fold cross-validation for results. Finally, XAI techniques were applied to enhance the interpretability and trustworthiness of the predictions.

We gain valuable insights into the dataset's structure and feature relationships through comprehensive data analysis using visualizations such as histograms, violin plots, and correlation heatmaps. Histograms reveal that Age, TPs, and AL follow near-normal distributions, while features like TB, ALPH, and ASAT exhibit right-skewed distributions with notable outliers, indicating the presence of extreme values that could impact model performance. Violin plots further confirm that bilirubin and enzyme levels are highly skewed. In contrast, protein levels and Age maintain more symmetric distributions, providing a clearer view of data spread and potential anomalies. Additionally, the correlation heatmap highlights strong positive relationships, such as between TB and DB and Alamine Aminotransferase and ASAT, suggesting collinearity among liver function markers. Moderate negative correlations, like the inverse relationship between AL and Age, also emerge, offering insights into potential dependencies. These analyses are crucial in understanding data characteristics, guiding feature selection, and optimizing model performance.

Table 2 consolidates the performance metrics for six models across different feature optimization scenarios—no optimization, LDA, PCA, and RFE—providing a comprehensive comparison of accuracy, precision, recall, and *F*1 score. Without feature reduction, SVMANN leads with an accuracy of 78.03%, while SANN trails at 76.32%, setting the baseline for model effectiveness. With LDA applied, overall performance improves, with SANN achieving the highest accuracy of 79.15% and SVMANN recording the lowest at 75.72%, underscoring the nuanced impact of LDA on these models. When PCA is used, SVMANN emerges as the top performer with a 77.44% accuracy, contrasting with ABANN's lower accuracy of 74.70%, while corresponding precision, recall, and *F*1 scores further delineate these differences. Finally, under RFE, SVMANN again attains the highest accuracy at 78.03%, whereas GANN shows the lowest at 75.81%. This table highlights how various feature optimization techniques distinctly influence model performance, offering detailed insights into their relative strengths and weaknesses across multiple evaluation metrics.

Table 3 summarizes the LDA model's feature importance rankings and coefficients within the Max Voting framework across ten cross-validation folds and the Final Optimal Feature Set (FOFS), derived as the union of top features across all folds.

Table 2. Model performances with and without feature reduction

Feature optimization	Models	ANN	SANN	BANN	ABANN	GANN	SVMANN
No Optimization	Accuracy	76.92	76.32	75.73	76.41	76.32	78.03
	Precision	70.94	49.17	67.66	68.76	49.17	39.02
	Recall	58.98	51.48	58.84	67.23	51.48	50
	*F*1 Score	58.17	47.57	59.23	67.65	47.57	43.83
LDA	Accuracy	78.03	**79.15**	78.55	78.21	78.55	75.72
	Precision	39.02	39.64	39.28	39.1	39.28	37.86
	Recall	50	49.89	50	50	50	50
	*F*1 Score	43.83	44.18	43.99	43.88	43.99	43.09
PCA	Accuracy	76.49	75.81	76.07	74.7	76.75	77.44
	Precision	70.98	56.81	68.33	64.88	67.23	38.75
	Recall	57.18	52.64	57.35	60	61.27	49.95
	*F*1 Score	56.2	50.16	56.86	60.42	62.33	43.64
RFE	Accuracy	77.35	77.44	77.26	77.01	75.81	78.03
	Precision	47.02	55.19	46.08	69.12	66.89	39.02
	Recall	50.52	50.95	50.38	61.62	58.22	50
	*F*1 Score	45.63	46.22	45.3	62.77	58.42	43.83

Table 3. LDA feature importance sets and coefficients for the Max Voting Model

Fold	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
1	AL (0.6101)	ALAT (0.5493)	TP (0.5429)	Age (0.4106)	DB (0.3990)	ALPH (0.2983)	ASAT (0.1353)	Gender (0.0952)	AGR (0.0525)	TB (0.0328)
2	AL (0.6748)	TP (0.5036)	Age (0.4102)	DB (0.3646)	ALPH (0.3164)	ASAT (0.2267)	ALAT (0.2087)	AGR (0.0369)	TB (0.0233)	Gender (0.0033)
3	AL (0.6895)	TP (0.5345)	ALAT (0.5043)	DB (0.4687)	ALPH (0.3130)	Age (0.3116)	Gender (0.1841)	ASAT (0.1548)	AGR (0.1255)	TB (0.0251)
4	AL (0.6904)	TP (0.6365)	ALAT (0.5186)	DB (0.4095)	Age (0.3715)	ALPH (0.3164)	ASAT (0.1823)	Gender (0.1533)	AGR (0.0447)	TB (0.0305)
5	AL (0.6444)	TP (0.6181)	ALAT (0.4810)	DB (0.4431)	Age (0.3211)	ALPH (0.2995)	Gender (0.1462)	ASAT (0.1423)	AGR (0.0348)	TB (0.0257)
6	AL (0.6692)	ALAT (0.4887)	TP (0.4548)	DB (0.4103)	ALPH (0.3241)	Age (0.2472)	Gender (0.2022)	ASAT (0.1475)	TB (0.0263)	AGR (0.0141)
7	DB (1.0238)	AL (0.6993)	TB (0.6883)	TP (0.3867)	ASAT (0.3472)	Age (0.3014)	ALPH (0.2984)	ALAT (0.1607)	AGR (0.1386)	Gender (0.1017)
8	AL (0.5628)	ALAT (0.4848)	TP (0.4481)	ALPH (0.4084)	DB (0.4010)	Age (0.3340)	Gender (0.2337)	ASAT (0.1741)	TB (0.0118)	AGR (0.0053)
9	AL (0.6611)	TP (0.4712)	ALAT (0.4439)	DB (0.3973)	Age (0.3622)	Gender (0.2879)	ALPH (0.2864)	ASAT (0.1313)	AGR (0.0801)	TB (0.0281)
10	DB (0.9085)	AL (0.9073)	TP (0.6697)	ALAT (0.5307)	TB (0.4987)	Age (0.3817)	ALPH (0.3109)	AGR (0.2219)	ASAT (0.1835)	Gender (0.1746)
FOFS	Gender	ASAT	ALPH	AL	TP	DB	ALAT	Age	TB	AGR

Table 4. Metrics across 10 folds for Max Voting Model

Fold	Accuracy	Precision	Recall	*F*1 score
1	98.40	98.28	98.24	98.25
2	98.35	98.25	98.30	98.28
3	98.42	98.35	98.27	98.38
4	98.36	98.27	98.28	98.30
5	98.39	98.29	98.31	98.30
6	98.37	98.28	98.29	98.31
7	98.38	98.24	98.27	98.28
8	98.36	98.31	98.28	98.30
9	98.40	98.25	98.28	98.32
10	98.37	98.30	98.26	98.36
Mean	98.38	98.28	98.28	98.31
Standard Deviation	0.0221	0.0329	0.0199	0.0382
95% Confidence Interval	±0.0158	±0.0236	±0.0142	±0.0274

AL consistently ranks as the most influential feature, followed by TP and DB, highlighting their critical role in classification. Mid-ranked features such as ALAT, Age, and ALPH exhibit moderate predictive significance, while Gender, AGR, and TB contribute less but ensure comprehensive feature representation. The coefficients, shown in parentheses, indicate each feature's weight in the classification model, with fold-specific variations, such as the prominence of DB in Fold 7 and TB in Fold 10, showcasing the adaptability of the LDA model in leveraging diverse markers for liver disease diagnosis. The FOFS integrates the top-ranked features across all folds, ensuring a comprehensive feature set for effective classification: {'Gender,' 'AST,' 'ALPH,' 'AL,' 'TP,' 'DB,' 'ALAT,' 'Age,' 'TB,' 'AGR'}.

The MVE method demonstrates exceptional performance, achieving an accuracy of 98.38% along with precision, recall, and *F*1 score values of 98.28%, 98.28%, and 98.31%, respectively. Additionally, LDA emerges as the most compelling feature optimization method, and the ensemble approach leverages the strengths of individual models with LDA to enhance predictive performance.

The metrics provided in Table 4 demonstrate the consistency of the Max Voting model across different folds. These include accuracy, precision, recall, and *F*1 score, with averages calculated for each metric. Additionally, the SD and 95% CI quantify variability and highlight the reliability and stability of the model's performance across multiple data subsets.

Figure 2 compares the performance metrics of the top-performing models across various feature optimization arrangements and without feature optimizations for liver disease prediction. The MVE method, integrating multiple ANN models with LDA, achieves the highest accuracy (98.38%) and *F*1 score (81.8%), highlighting its superiority over individual models.

Figure 3(a) presents the CM for the Max Voting model. It shows only one misclassification between disease and non-disease cases, indicating strong predictive performance. Meanwhile, Figure 3(b) displays the ROC curve, where the model achieves an AUC of 1.00, signifying perfect classification and excellent discriminative ability.

Figure 4 provides a detailed examination of the features influencing the model's predictions. Figure 4(a) displays the SHAP summary plot, showing that features like DB and ALPH strongly contribute to prediction accuracy, while lower values of AL and TB negatively affect the predictions. In Figure 4(b), the SHAP waterfall plot illustrates the individual feature impacts. It highlights that ALPH reduces the likelihood of liver disease while TPs, along with Age and AL, increase it. Figure 4(c) presents the

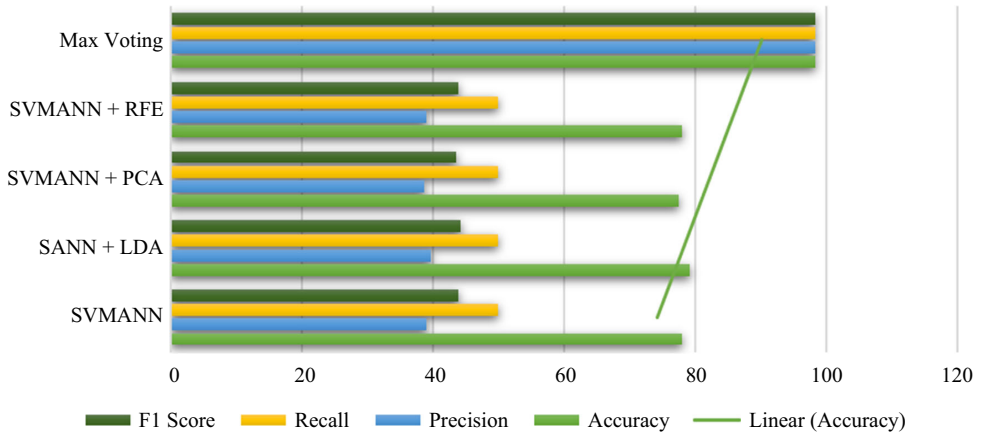

Figure 2. Comparison of the best-performed models with and without feature optimization and Max Voting

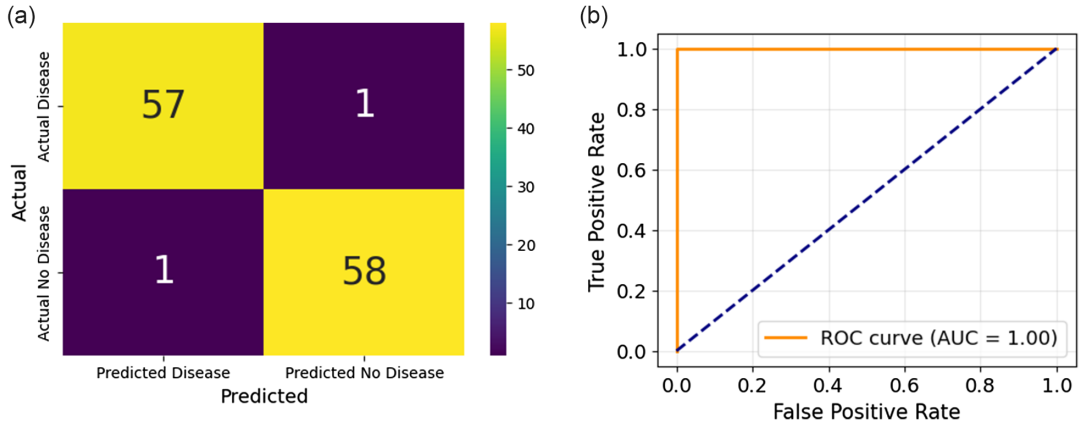

Figure 3. Performance evaluation of the MVE model: (a) CM for classification accuracy and (b) ROC curve for model discrimination

LIME explanation for class 0 (no liver disease), where Alamine Aminotransferase and DB contribute negatively, while TPs have a small positive impact. Figure 4(d) shows the LIME explanation for class 1 (liver disease), where Alamine Aminotransferase and ALPH contribute negatively. Finally, Figure 4(e) illustrates the SHAP FORCE plot, offering a more granular view of the force and direction of each feature's influence on the final prediction, emphasizing their relative contribution in a visual format. These visualizations comprehensively understand the feature impacts driving the Max Voting model's decisions.

Figure 5(a) presents the ICE plots for each feature, showing how prediction values change with varying feature values. Features like Age and TB exhibit a more substantial influence on predictions, while Gender and TPs have minimal impact. Figure 5(b) shows the SHAP dependence analysis, revealing that Age and TB contribute positively to predictions. At the same time, ALPH and AL have varying impacts, suggesting their effects are more context-dependent. These analyses provide a deeper understanding of how individual features drive the model's decision-making process.

Table 5 compares feature prioritization between various XAI methods, such as SHAP, LIME, FORCE, ICE, and clinical experts. DB and TB are consistently high-priority features. This alignment between the model's decisions and expert judgment underscores the model's interpretability and potential for real-world clinical applications.

The performance of Deep Learning models, including Long-Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network-Long-Short-Term Memory (CNN-LSTM) Ensemble models, was evaluated on a 583-instance dataset. Among these models, LSTM achieved the highest accuracy at 68.38%, closely followed by GRU at 68.12%. However, neither LSTM nor GRU outperformed the Max Voting Model. Regarding precision, recall, and *F*1 score, LSTM achieved 52.47%, 52.45%, and 47.56%, respectively, while GRU showed better precision and recall at 59.01% and 55.47% but slightly lower *F*1 at 53.88%. The CNN-LSTM ensemble model, on the other hand, had the lowest performance across all metrics, with an accuracy of 50.1%, precision of 51.15%, recall of 50.78%, and *F*1 score of 45.62%. Despite the strengths of these deep learning models, they do not surpass the MVE in predictive accuracy.

We further evaluated the performance of our Max Voting model using real-time patient data and a different dataset to assess its accuracy across various contexts. The validation with real-time sample data involved testing the model on new patient samples, including data from Mr. Akash (23, male) from Dr. Lal's Pathology Lab, Mrs. Sushila (53, female) from House of Diagnostics, and Mr. Wasif (30, male) from Chughtai Lab. Key

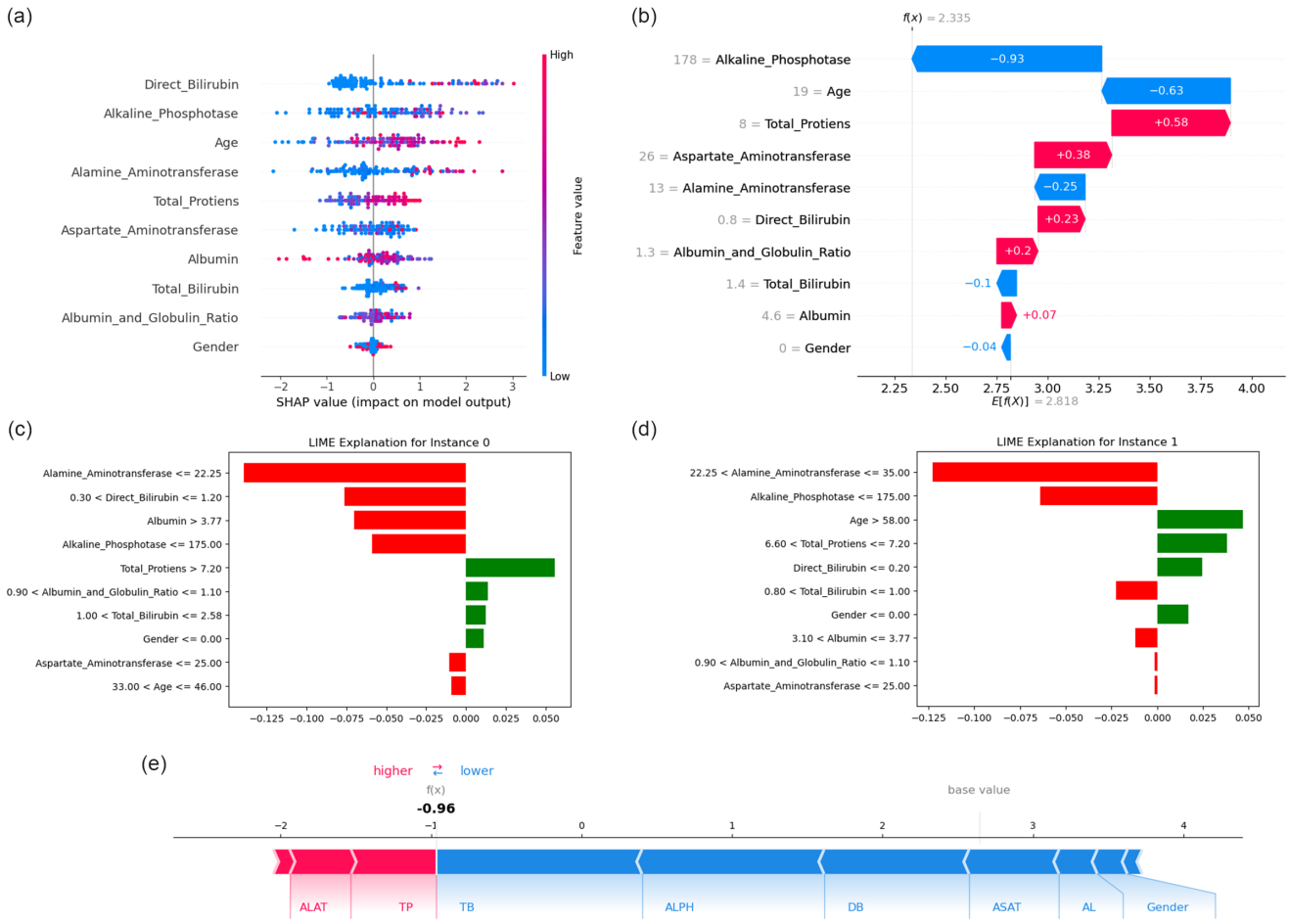

Figure 4. XAI feature impact analysis in the Max Voting Model: (a) SHAP summary plot, (b) SHAP waterfall plot, (c) LIME explanation for no liver disease prediction, (d) LIME explanation for liver disease prediction, and (e) FORCE plot

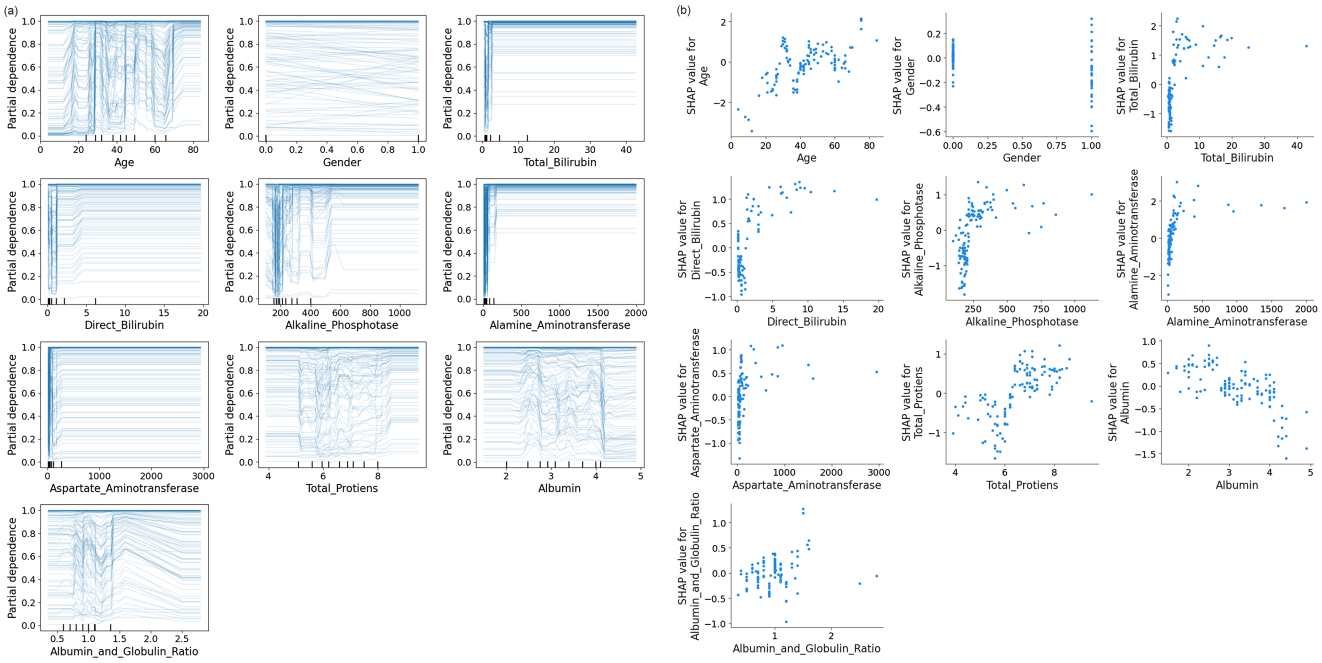

Figure 5. Feature impact analysis for liver health assessment in the Max Voting Model: (a) ICE plots for each feature and (b) SHAP dependence analysis of predictive features

Table 5. Comparison of feature importance rankings in liver disease prediction between XAI methods and clinical expert judgment

	XAI decision						Experts decision	
		LIME		SHAP				
Priority	SHAP	Disease	No Disease	WATERFALL	FORCE	ICE	Expert 1	Expert 2
First	DB	ALAT	ALAT	ALPH	TB	TB	Both TB and DB are the most important	TB/DB
Second	ALPH	ALPH	DB	Age	ALPH	ALPH		ASAT, ALPH
Third	Age	Age	AL	TP	DB	ASAT		AL, AGR
Fourth	ALAT	TP	ALPH	ASAT	ASAT	DB		Age, Gender
Fifth	TP	DB	TP	ALAT	TP	TP		–
Sixth	ASAT	TB	AGR	DB	ALAT	AGR		–

Table 6. Comparison of our study with previous studies

Existing literatures	Dataset	Accuracy	Feature optimization	Cross validation	XAI
Choubey et al.	583 Sample	75.10%	✓	✕	✕
Shetty and Satyanarayana.	583 Sample	71%	✕	✕	✕
Alyabis et al.	583 Sample	79.6%	✓	✕	✕
Singh and Agarwal.	583 Sample	77.77%	✕	✕	✕
Azam et al.	583 Sample	74%	✓	✕	✕
Choudhary et al.	583 Sample	70.54%	✕	✓	✕
Khan et al.	583 Sample	72.17%	✕	✕	✕
Kannapiran et al.	583 Sample	73.97%	✕	✕	✕
Muthuselvan et al.	583 Sample	74.2%	✕	✓	✕
Yasmin et al.	583 Sample	76.03%	✓	✕	✕
Our Study	583 Sample	98.38%	✓	✓	✓

health indicators such as TB, DB, ALPH, ALAT, ASAT, TP, AL, and AGR were used to evaluate the model's accuracy. The results of this validation were compared with the 583-sample dataset, showcasing the model's ability to accurately assess and predict patient health metrics.

The MVE model was tested on an external liver disease dataset of 30,691 patients [41] for validation. The model demonstrates strong performance with an average accuracy of 88.35%, highlighting its ability to generalize effectively to external samples and confirming its robustness in predicting liver disease outcomes. The model's precision, recall, and *F*1 score also reflect solid performance, with mean values of 92.99%, 79.48%, and 83.26%, respectively. The standard deviation for accuracy, precision, recall, and *F*1 score is 1.94, 1.30, 2.77, and 1.43, respectively, indicating relatively stable performance across the folds. The 95% confidence intervals for these metrics are ±1.20 for accuracy, ±1.30 for precision, ±2.77 for recall, and ±1.43 for *F*1 score, further validating the model's effectiveness in liver disease prediction.

We collect a Maternal Health Risk (MHR) dataset from Kaggle [42], which contains 1,014 samples and seven features divided into three classes: low, mid, and high risk. The results of the Max Voting model across 10 folds are evaluated using performance metrics such as accuracy, precision, recall, and *F*1 score. The model demonstrates strong performance, achieving an average accuracy of 93.07%, precision of 92.71%, recall at 93.06%, and an *F*1 score of 92.85%. For each fold, the standard deviation and 95% confidence intervals are calculated, showing minor variability, with confidence intervals of ±0.74 for accuracy, ±0.80 for precision, ±1.72 for recall, and ±0.89 for *F*1 score. The CM reveals that the model accurately classifies most samples across the three classes, correctly predicting 60 out of 67 Low-risk cases, 67 out of 71 Mid-risk cases, and 61 out of 64 High-risk cases. Some misclassifications occur, such as 5 Low-risk cases misclassified as Mid and 2 as High, along with a few misclassifications in the Mid and High-risk categories, but overall, the model effectively distinguishes between different risk levels, reflecting its robust predictive capabilities on the MHR dataset.

Finally, Table 6 compares with existing literature and reveals the superior accuracy of 98.38% achieved by the proposed approach, significantly outperforming the 70.54% to 79.6% range reported in previous studies. Unlike earlier research, this study incorporates feature optimization, cross-validation, and XAI techniques, addressing existing gaps. The ensemble model, built on ANN-based hybrid approaches, enhances predictive accuracy and interpretability, distinguishing it from prior work.

The lower accuracies and other performance metrics presented in Table 4 can be attributed mainly to the limited size of the dataset, which consists of only 583 samples and 11 features. This small dataset restricts the ability of individual models to generalize effectively, especially when it comes to capturing complex patterns. As a result, the models exhibit lower precision, recall, and *F*1 scores. When trained on such limited datasets with few features, models are more prone to overfitting or underfitting, as they lack sufficient information to identify intricate relationships. This ultimately leads to reduced accuracy and other performance metrics [43].

However, the MVE method with LDA achieves higher accuracy despite the limitations of individual models. By combining predictions from multiple models through Max Voting, this approach mitigates the weaknesses of each model, enhancing overall performance. LDA's role in reducing dimensionality allows each model to focus on the most relevant features, improving their performance within the ensemble. The ensemble capitalizes on the strengths of each model. At the same time, LDA's feature optimization provides a more transparent, more robust representation of the data, leading to improved accuracy and other metrics in the combined outcome.

Even though our dataset contains only 11 features, we used feature optimization techniques because they enhance model accuracy by refining the dataset to its most informative aspects.

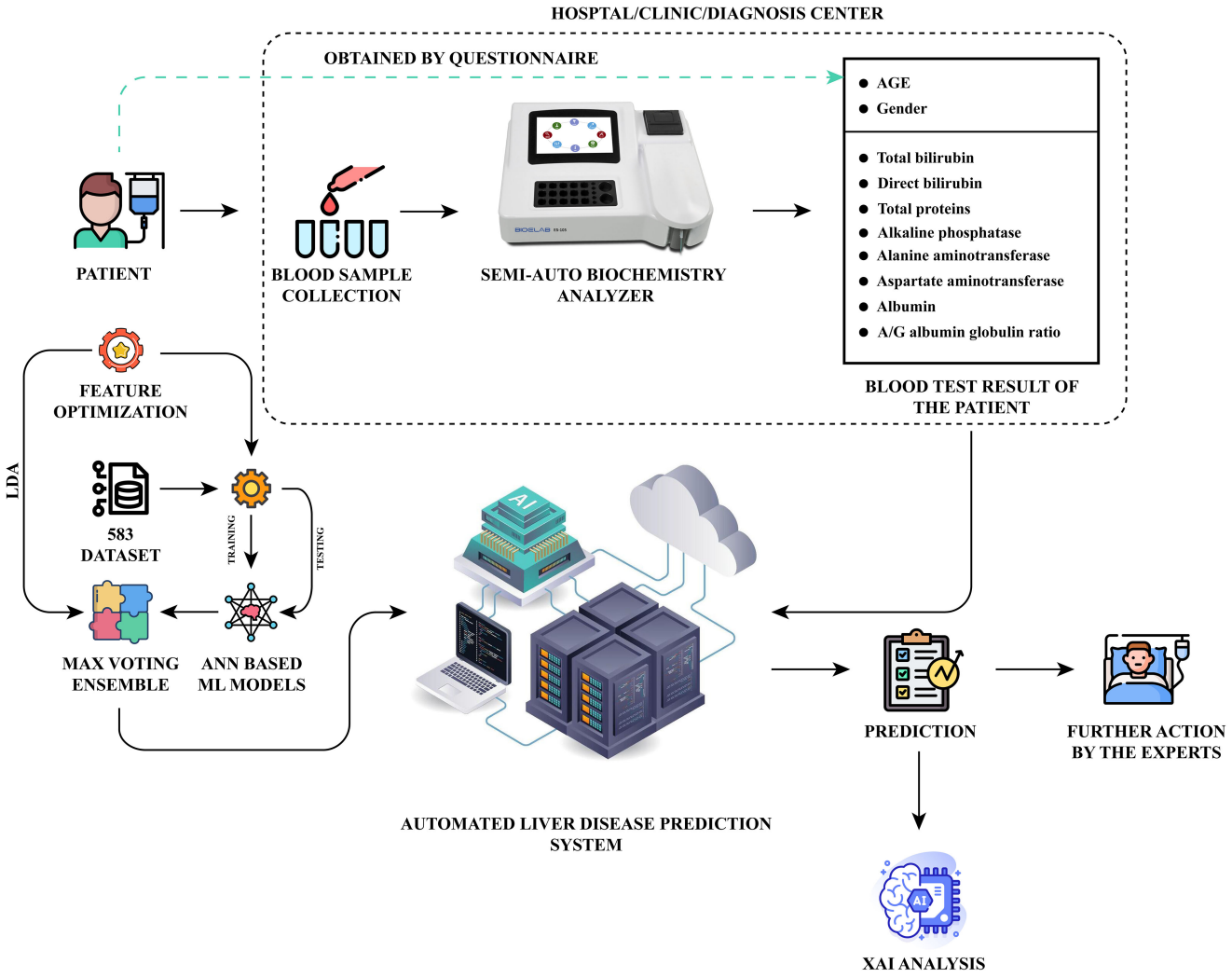

Figure 6. Real-time framework for liver disease prediction using semi-auto biochemistry analysis, ANN-based MVE models, and XAI-driven interpretation

These techniques, like LDA, PCA, and RFE, help the model focus on the features that most significantly contribute to identifying patterns and improving predictive reliability. By reducing noise and minimizing irrelevant data, feature optimization allows for more effective learning and generalization, increasing stability and reducing computational complexity. This approach is also valuable in small datasets, where maximizing the signal-to-noise ratio is critical for robust performance [44].

LDA proved the most compelling feature selection method because it maximizes class separation, making it ideal for classification tasks where distinguishing between classes is crucial. Unlike PCA, which reduces dimensionality based on variance without considering class labels, or RFE, which does not directly optimize for class discrimination, LDA enhances class separability. Additionally, LDA handles class imbalances better by considering the ratio of between-class to within-class variance, ensuring that selected features are most relevant for distinguishing between classes, even in imbalanced datasets [45].

Figure 6 depicts our liver disease prediction framework in the real-time scenario. Patient data is collected via questionnaires and blood samples, which are then analyzed in a semi-auto biochemistry analyzer to measure liver function indicators. This data is tested against an existing, optimized dataset of 583 samples using LDA, chosen for its effective feature selection in ANN-based models. Afterward, predictions are generated using the ANN-based MVE for improved accuracy. Finally, XAI enables users, including non-experts, to understand the projections and confidently take further medical actions in consultation with experts.

4. Conclusion

In conclusion, this study demonstrates a robust approach to enhancing liver disease prediction by integrating ANN with five distinct ML models—Stacking, Bagging, AdaBoost, Gradient-Boosted Decision Tree, and SVM—to create five hybrid models optimized through LDA. Combined into a MVE, these LDA-optimized hybrids achieve a significant accuracy increase from 79.15% to 98.38%. XAI techniques, such as LIME, SHAP, and ICE, further support the transparency of the model's decision-making process. We validate the ensemble model's effectiveness by comparing its predictions with doctors' decisions and testing it on samples from external sources and a multiclass MHR dataset, confirming its adaptability beyond the initial dataset. A real-time demonstration of our model underscores its practical utility, though the study notes limitations, particularly in applying the model to clinical settings due to data constraints. Future work will

address these limitations by implementing Differential Privacy and Clinical Servers to protect patient data, with plans to extend the model to support multi-disease prediction. Additionally, we aim to construct a web server that would enhance the accessibility and value of this tool for the broader community and end users.

Acknowledgment

Our heartfelt gratitude goes to two distinguished physicians who generously shared their expertise, greatly enriching this study. Special thanks are extended to Dr. Shahriar Shafiq, a Higher Specialty Registrar in Diabetes and Endocrinology at the Royal College of Physicians of Edinburgh, England, referenced as Expert 1 in Table 5. Dr. Shafiq's insightful guidance was instrumental in navigating the complexities of liver disease research. Appreciation is also due to Dr. Talha Sami Anik, Assistant Surgeon with the Government of the People's Republic of Bangladesh, listed as Expert 2 in Table 5. Dr. Anik's extensive experience, including his work at Birdem General Hospital and Dhaka Medical College & Hospital, brought valuable perspectives to this research. Despite their demanding schedules, both doctors demonstrated exceptional commitment and professionalism, providing critical support that significantly contributed to the advancement of this study.

Ethical Statement

This study contains no studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data used in this study will be accessible upon request to the corresponding author.

Author Contribution Statement

Safiul Haque Chowdhury: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Mohammad Mamun:** Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Md. Tanvir Ahmed Shaikat:** Visualization, Project administration. **Mohammed Ibrahim Hussain:** Writing – review & editing, Supervision. **MD. Sadiq Iqbal:** Writing – review & editing, Visualization, Supervision. **Muhammad Minoar Hossain:** Writing – review & editing, Supervision.

References

[1] Williams, R. (2006). Global challenges in liver disease. *Hepatology*, *44*(3), 521–526. https://doi.org/10.1002/hep.21347

[2] American Liver Foundation. (2024). *How many people have liver disease?*. Retrieved from: https://liverfoundation.org/about-your-liver/facts-about-liver-disease/how-many-people-have-liver-disease/

[3] Choubey, D. K., Dubey, P., Tewari, B. P., Ojha, M., & Kumar, J. (2023). Prediction of liver disease using soft computing and data science approaches. In *6G enabled fog computing in IoT: Applications and opportunities* (pp. 183–213). Switzerland: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-30101-8_8

[4] Shetty, P. J. (2023). Prediction performance of classification models for imbalanced liver disease data. *International Journal of Statistics and Applied Mathematics*, *8*(5), 58–62.

[5] Alyabis, M. A. S., Howaimil, B. M., Alyabes, A. M. S., Alrabiah, A. A. H., Alrabiah, A. S. H., Aljumayi, I. M., . . . , & Binshaheen, H. S. (2022). Prediction of liver diseases using neural network analysis. *International Journal of Pharmaceutical and Bio Medical Science*, *2*(08), 314–320. https://doi.org/10.47191/ijpbms/v2-i8-08

[6] Singh, G., & Agarwal, C. (2023). Prediction and analysis of liver disease using extreme learning machine. In *Sentiment analysis and deep learning: Proceedings of ICSADL 2022* (pp. 679–690). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5443-6_52

[7] Azam, M. S., Rahman, A., Iqbal, S. M. H. S., & Ahmed, M. T. (2020). Prediction of liver diseases by using few machine learning based approaches. *Australian Journal of Engineering and Innovative Technology*, *2*(5), 85–90. https://doi.org/10.34104/ajeit.020.085090

[8] Choudhary, R., Gopalakrishnan, T., Ruby, D., Gayathri, A., Murthy, V. S., & Shekhar, R. (2021). An efficient model for predicting liver disease using machine learning. In R. Satpathy, T. Choudhury, S. Satpathy, S. N. Mohanty & X. Zhang (Eds.), *Data analytics in bioinformatics: A machine learning perspective* (pp. 443–457). Wiley. https://doi.org/10.1002/9781119785620.ch18

[9] Khan, B., Naseem, R., Ali, M., Arshad, M., & Jan, N. (2019). Machine learning approaches for liver disease diagnosing. *International Journal of Data Science and Advanced Analytics*, *1*(1), 27–31. https://doi.org/10.69511/ijdsaa.v1i1.71

[10] Singh, A. S., Irfan, M., & Chowdhury, A. (2018). Prediction of liver disease using classification algorithms. In *2018 4th International Conference on Computing Communication and Automation*, 1–3. https://doi.org/10.1109/CCAA.2018.8777655

[11] Muthuselvan, S., Rajapraksh, S., Somasundaram, K., & Karthik, K. (2018). Classification of liver patient dataset using machine learning algorithms. *International Journal of Engineering & Technology*, *7*(3.34), 323. https://doi.org/10.14419/ijet.v7i3.34.19217

[12] Yasmin, R., Amin, R., & Reza, M. S. (2023). Design of novel feature union for prediction of liver disease patients: A machine learning approach. In *The fourth industrial revolution and beyond: Select proceedings of IC4IR+* (pp. 515–526). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8032-9_36

[13] Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., . . . , & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics*, *13*(15), 2582. https://doi.org/10.3390/diagnostics13152582

[14] Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2012). A critical comparative study of liver patients from USA and INDIA: An exploratory analysis. *International Journal of Computer Science Issues*, *9*(3), 506.

[15] Shinde, B. G., & Shivthare, S. (2024). Impact of data visualization in data analysis to improve the efficiency of machine learning models. *Journal of Advanced Zoology*, *45*, 107–112.

[16] Roy, S., Bhalla, K., & Patel, R. (2024). Mathematical analysis of histogram equalization techniques for medical image enhancement: A tutorial from the perspective of data loss. *Multimedia Tools and Applications*, *83*(5), 14363–14392. https://doi.org/10.1007/s11042-023-15799-8

[17] Hu, K. (2020). Become competent within one day in generating boxplots and violin plots for a novice without prior R experience. *Methods and Protocols*, *3*(4), 64. https://doi.org/10.3390/mps3040064

[18] Gu, Z. (2022). Complex heatmap visualization. *Imeta*, *1*(3), e43. https://doi.org/10.1002/imt2.43

[19] Ismail, A. R., Abidin, N. Z., & Maen, M. K. (2022). Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control*, *3*(2), 143–152. https://doi.org/10.18196/jrc.v3i2.13133

[20] Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation?. *Emerging Markets Finance and Trade*, *58*(2), 472–482. https://doi.org/10.1080/1540496X.2020.1825935

[21] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli & S. B. V. d. T. Vieira (Eds.), *Machine learning methods and applications to brain disorders* (pp. 101–121). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

[22] Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. *Nature Reviews Methods Primers*, *4*(1), 70. https://doi.org/10.1038/s43586-024-00346-y

[23] Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

[24] Chen, X. W., & Jeong, J. C. (2007). Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications*, 429–435. https://doi.org/10.1109/ICMLA.2007.35

[25] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

[26] Pavlyshenko, B. (2018). Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing*, 255–258. https://doi.org/10.1109/DSMP.2018.8478522

[27] González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007

[28] Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, *39*(6), 745–758. https://doi.org/10.1016/S1874-1029(13)60052-X

[29] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[30] Vishwanathan, S. V. M., & Murty, M. N. (2002). SSVM: A simple SVM algorithm. In *Proceedings of the 2002 International Joint Conference on Neural Networks (Cat. No. 02CH37290)*, *3*, 2393–2398. https://doi.org/10.1109/IJCNN.2002.1007516

[31] Amin, F., & Mahmoud, M. (2022). Confusion matrix in binary classification problems: A step-by-step tutorial. *Journal of Engineering Research*, *6*(5), Article 1.

[32] Lee, D. K., In, J., & Lee, S. (2015). Standard deviation and standard error of the mean. *Korean Journal of Anesthesiology*, *68*(3), 220–223. https://doi.org/10.4097/kjae.2015.68.3.220

[33] Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*, *9*(10), 4125. https://doi.org/10.21037/jtd.2017.09.14

[34] Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve?. *Emergency Medicine Journal*, *34*(6), 357–359. https://doi.org/10.1136/emermed-2017-206735

[35] Tian, T., & Zhu, J. (2015). Max-margin majority voting for learning from crowds. *Advances in Neural Information Processing Systems*, *28*, 1–9.

[36] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . , & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[37] Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, *74*, 851–886. https://doi.org/10.1613/jair.1.13283

[38] Kawakura, S., Hirafuji, M., Ninomiya, S., & Shibasaki, R. (2022). Analyses of diverse agricultural worker data with explainable artificial intelligence: Xai based on shap, lime, and lightgbm. *European Journal of Agriculture and Food Sciences*, *4*(6), 11–19. https://doi.org/10.24018/ejfood.2022.4.6.348

[39] Vikas1055. (2019). Delhi Public School, R.K. Puram CS R-501 lab report. *Course Hero*. Retrieved from: https://www.coursehero.com/file/42005265/labreportnewpdf/

[40] Asking for Self. (n.d.). Talk to liver on liver function test. *Marham*. Retrieved from: https://www.marham.pk/forum/liver-specialist/liver-function-test

[41] Velu, S. R., Ravi, V., & Tabianan, K. (2022). Data mining in predicting liver patients using classification model. *Health and Technology*, *12*(6), 1211–1235. https://doi.org/10.1007/s12553-022-00713-3

[42] Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020). Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT). In *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering*, 357–365. https://doi.org/10.1007/978-981-15-2317-5_30

[43] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

[44] Ali, M. Z., Abdullah, A., Zaki, A. M., Rizk, F. H., Eid, M. M., & El-Kenway, E. M. (2024). Advances and challenges in feature selection methods: A comprehensive review. *Journal of Artificial Intelligence and Metaheuristics*, *7*(1), 67–77. https://doi.org/10.54216/JAIM.070105

[45] Kim, A. K., & Chung, H. (2021). The effect of rebalancing on LDA in imbalanced classification. *Stat*, *10*(1), e384. https://doi.org/10.1002/sta4.384

How to Cite: Chowdhury, S. H., Mamun, M., Shaikat, T. A., Hussain, M. I., Iqbal, S., & Hossain, M. M. (2025). An Ensemble Approach for Artificial Neural Network-Based Liver Disease Identification from Optimal Features through Hybrid Modeling Integrated with Advanced Explainable AI. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52024744