Enhancing Liver Disease Classification Based on a Stacked Machine Learning Model

Alaa A. Almelibari
Department of Computer Science and Artificial Intelligence, College of Computing, Umm AL-Qura University, Makkah, Saudi Arabia
aamelibari@uqu.edu.sa

Mostafa Ibrahim Labib
Higher Future Institute for Specialized Technological Studies, Cairo, Egypt
Mostafa.elkhalil@fa-hists.edu.eg

Yasser Ramadan
Department of Computer Science, Faculty of Computers and Information, Suez University, Egypt
yasserfrb@gmail.com (corresponding author)

Received: 15 April 2025 | Revised: 27 May 2025 | Accepted: 1 June 2025

 DOI: https://doi.org/10.48084/etasr.11526

ABSTRACT

Liver Disease (LD) poses a serious global health issue, emphasizing the need for precise and dependable diagnostic solutions. This research introduces an enhanced Machine Learning (ML) approach based on a stacked ensemble framework to classify LD cases, leveraging a publicly accessible dataset from Kaggle comprising patient records from India. Six ML models were applied, namely Random Forest (RF), Support Vector Machine (SVM), Dummy Classifier (DC), Extra Trees classifier (ET), K-Nearest Neighbors (KNN), and Logistic Regression (LR), with ET achieving the highest accuracy at 79.82%. To improve prediction accuracy, a stacked ensemble was developed using ET and RF as base classifiers and SVM as the meta-classifier, which boosted the overall accuracy to 98.53%. The study evaluated performance using accuracy, precision, recall, F1-score, and AUC. The findings highlight the effectiveness of stacking-based ML methods in building accurate and reliable diagnostic tools for liver disease classification.

Keywords-disease classification; stacked machine learning; liver disease; liver disease classification; artificial intelligence

I. INTRODUCTION

Globally, LD causes millions of deaths annually, with chronic Hepatitis B Virus (HBV) infection affecting an estimated 257 million people. More than 40% of these cases are at risk of progressing to more severe liver complications. While early detection and antiviral treatments can effectively halt disease progression, not all patients require aggressive intervention. For instance, individuals in the early stages of liver fibrosis may only need routine monitoring, whereas those with advanced fibrosis necessitate immediate therapeutic action [1-2]. Therefore, there is a critical need for accessible, affordable, and reliable biomarkers to evaluate liver fibrosis stages using non-invasive diagnostic techniques, an area where machine learning-driven approaches can offer significant value.

Stacking-based machine learning approaches have shown great promise across various domains. In [3], the challenge of water potability classification was addressed using a stacked model, achieving an F1-score of 70.23% and an AUC of 77.62%, demonstrating how ensemble techniques can effectively enhance predictive stability. Similarly, in [4], stacking was used in an aerodynamic modeling context to predict drag and lift coefficients, where a stacked ensemble of XGBoost and RF yielded even better generalization capabilities, explaining over 99% of the variance and achieving minimal prediction error. These studies collectively underscore the versatility and effectiveness of stacking methods in improving classification and regression tasks across fields, reinforcing the rationale for applying a stacked ensemble architecture in liver disease classification to achieve high accuracy and clinical reliability.

This study uses a stacked ensemble model, employing ET and RF as base classifiers and SVM as the meta-classifier, to boost overall accuracy. The findings highlight the effectiveness of stacking-based ML methods in building accurate and reliable diagnostic tools for the classification of liver disease. The results highlight the promising role of machine learning techniques, particularly when integrated with stacking

approaches, in accurately classifying liver disease. This study contributes to the ongoing advancement of dependable and non-invasive diagnostic solutions, supporting healthcare professionals in the early detection of liver conditions and aiding informed decision-making in patient management.

Several studies have focused on the classification of LD. In [5], multiple machine learning models were explored, along with a hybrid approach, comparing performance metrics such as accuracy, precision, recall, F-measure, and AUC to predict the likelihood of liver disease. The voting classifier outperformed individual models, achieving 80.1% accuracy, recall, and F-measure, 80.4% precision, and 88.4% AUC when SMOTE was applied in conjunction with 10-fold cross-validation. In [6], an advanced machine learning pipeline was proposed for liver cirrhosis classification. This approach incorporated preprocessing techniques, such as label encoding and min-max normalization, followed by feature extraction using the ConvNeXt model on clinical attributes including age, gender, liver function tests, medical history, and comorbidities. Feature selection was performed using an improved Grasshopper Optimization Algorithm (IGOA). An optimized ensemble of Naïve Bayes and Logistic Regression (ONBLR) was used for classification, further enhanced by Harris Hawks Optimization (HHO) for hyperparameter tuning. The proposed model outperformed several benchmark techniques, achieving an accuracy of 99.18%, sensitivity of 99.12%, and specificity of 98.92%.

In [7], a Decision Tree (DT) model was proposed for liver fibrosis staging using five serological biomarkers: HBV-DNA, platelet count, thrombin time, International Normalized Ratio (INR), and albumin. The model achieved high diagnostic performance across fibrosis stages (F0–F4), with AUC values ranging from 0.891 to 0.944 in the training cohort and from 0.876 to 0.933 in the external validation cohort. Risk stratification based on cutoff values demonstrated strong alignment with pathological diagnoses. This study highlights the potential of combining ML techniques with key serum markers for accurate staging of hepatic fibrosis and clinical management of Chronic Hepatitis B (CHB) patients. In [8], a hybrid model combined CNN with LSTM to improve the diagnosis and prognosis of LD. The results demonstrated that this integrated architecture outperformed standalone models such as CNN, RNN, and LSTM, achieving an accuracy of 98.73%. The model also reported a precision of 99%, a recall of 98%, an F1-score of 98%, and an AUC of 99%, confirming its effectiveness in accurately identifying liver disease and enhancing clinical decision-making.

In [9], the importance of Explainable Artificial Intelligence (XAI) in enhancing transparency and trust in AI-driven biomedical applications was emphasized, particularly for clinicians involved in diagnosis, treatment, and prognosis. Using the Indian dataset from the Andhra Pradesh region, this study developed a deep learning model using Keras-TensorFlow, achieving an accuracy of 0.81, which improved to 0.82 after hyperparameter tuning. To address class imbalance, Generative Adversarial Networks (GANs) were used for data oversampling. SHAP was integrated to interpret the model's predictions and provide clear explanations for LD detection outcomes. In [10], a machine learning model integrated RF, XGBoost, and SVM for the diagnosis of NAFLD. The model demonstrated outstanding predictive performance, achieving an accuracy of 0.99 and an AUC of 1.00, along with high precision and robustness. The findings highlighted gender as a strongly associated critical predictor of NAFLD. Compared to conventional indicators such as the Hepatic Steatosis Index (HIS), the proposed ensemble approach provided superior diagnostic accuracy and reliability, proving to be highly effective for early detection, diagnosis, and screening of NAFLD.

In [11], a two-level stacked ensemble model was proposed for LD classification using ILPD. This study compared the performance of individual and ensemble models, both with and without feature selection. Through extensive preprocessing and model optimization, the stacked model achieved an accuracy of 94.01%, along with strong precision, recall, F1-score, and AUC values. In [12], an XGBoost-based model was presented for early liver disease prediction, integrating ranking and statistical projection-based feature selection techniques. The Fisher score was used to perform global interpretability and guide optimal feature selection. The model's performance was evaluated using k-fold cross-validation, comparing both individual and hybrid feature sets, as well as benchmarking against traditional classifiers and state-of-the-art approaches. The proposed XGBoost-Liver model achieved a high average accuracy of 92.07%, outperforming existing methods.

This study presents a novel stacked ensemble framework for LD classification, combining ET and RF as base learners and SVM as the meta-classifier. This unique stacking configuration enhances prediction performance, achieving an accuracy of 98.53%, surpassing previous benchmarks. This approach contributes to the development of intelligent, non-invasive diagnostic tools by offering a more reliable and interpretable decision support model for LD detection, contributing to the ongoing development of intelligent diagnostic tools.

II. METHODOLOGY

The proposed method combines tree-based and margin-based classifiers in a stacked ensemble learning model. Standard models such as RF, ET, SVM, KNN, LR, and a DC were individually trained. The highest-performing models were selected as base learners and used to train an SVM meta-classifier. The StackingClassifier function from Scikit-learn ensured seamless integration of base and meta models. This approach achieved superior predictive results over standalone models.

A. Dataset

The publicly available liver disease dataset in [13] includes records of Indian liver patients, including clinical variables such as age, gender, total bilirubin, alkaline phosphatase, and albumin. The dataset is labeled with target attributes that indicate liver disease status [14]. Effective analysis requires a thorough preprocessing, including handling missing values, assessing data distribution, and identifying outliers to ensure data quality and model performance.

Figure 1 shows the correlation matrix, highlighting several strong associations among key clinical features. Total bilirubin (*tb*) and direct bilirubin (*db*) show a high correlation of 0.87, reflecting their physiological relationship. Similarly, liver enzymes SGPT and SGOT are strongly correlated at 0.79, while total protein (*tp*) and albumin (*alb*) also exhibit a strong correlation of 0.78. The albumin/globulin ratio (*a/g_*ratio) correlates well with *alb* (0.68), indicating consistency in liver function indicators.

Figure 2 displays the distribution of the numerical features, revealing several well-structured and informative numerical features within the dataset. The age feature shows a near-normal distribution, which is beneficial for many machine learning models. The selector (the target variable) displays a clear class separation, supporting its suitability for classification. Features such as *tp* and *alb* demonstrate symmetric, bell-shaped distributions, which suggest stable measurements across samples.

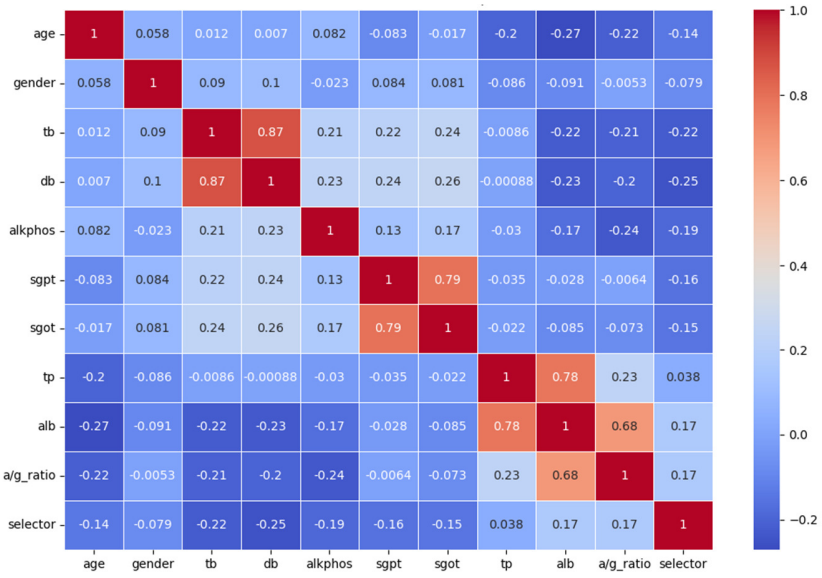

Fig. 1. Correlation matrix heatmap of the kidney dataset [13].

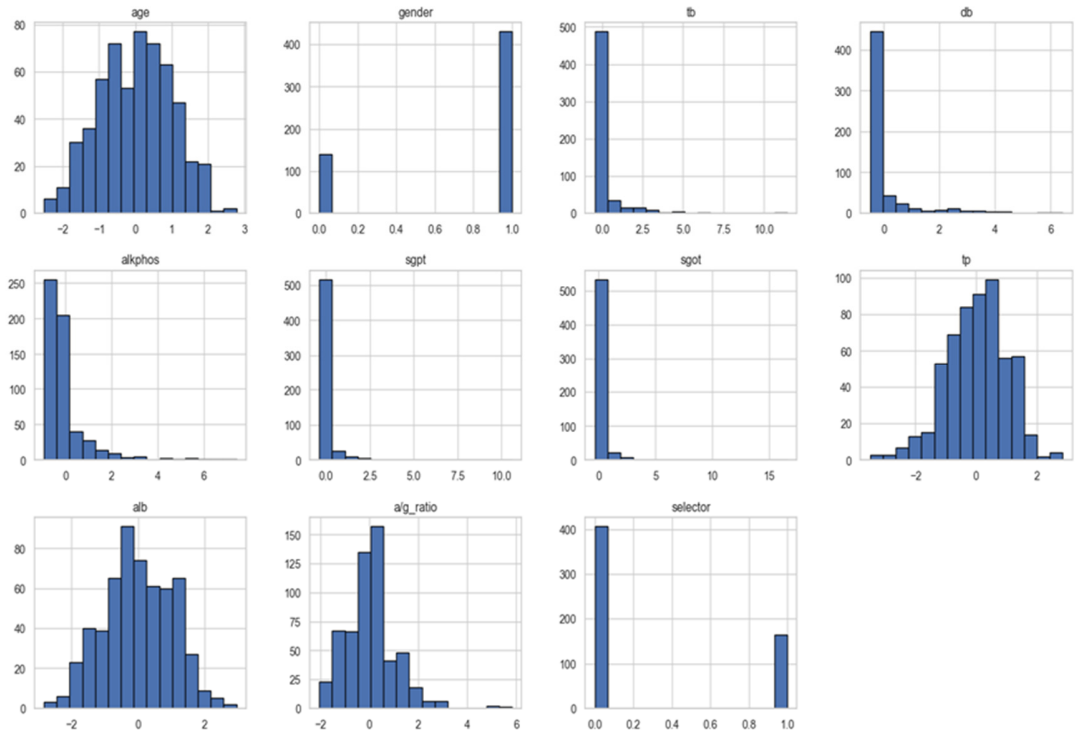

Fig. 2. The dataset's numerical features distribution.

B. Data Preparation

Several preprocessing techniques were applied to enhance the quality and readiness of the liver disease dataset for model training. Missing values were handled through imputation, min-max scaling was used for data normalization, and categorical variables, such as gender, were encoded using one-hot or label encoding methods. These steps ensure consistent feature scaling, improve dataset integrity, and contribute to better model performance.

C. Model Training and Evaluation

Model training utilized a preprocessed dataset divided into a stratified 80-20 train-test split to preserve class distributions, with further performance improvement through a stacked model. The evaluation was conducted using standard performance metrics calculated as follows [15-21]:

$$\text{Accuracy} = \frac{\text{TPos+TNeg}}{\text{TPos+FPos+FNeg+TNeg}} \quad (1)$$

$$\text{Precision} = \frac{\text{TPos}}{\text{TPos+FPos}} \quad (2)$$

$$\text{Recall} = \frac{\text{TPos}}{\text{TPos+FNeg}} \quad (3)$$

$$\text{F1} - \text{score} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall+Precision}} \quad (4)$$

$$\text{AUC} = \int_0^1 \text{TPR(FPR)}\ \text{d(FPR)} \quad (5)$$

where:

$$\text{TPR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives (FN)}}$$

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)+True Negatives (TN)}}$$

RF, SVM, DC, ET, KNN, and LR were implemented and evaluated as baselines using the preprocessed liver disease dataset. Each model was trained using Scikit-learn in Python 3.10, with min-max normalization applied to continuous features and categorical variables labeled. No synthetic balancing techniques were used to assess the raw discriminative ability of each model. Default hyperparameters were used for initial benchmarking, and evaluation metrics were computed using sklearn.metrics.

D. Lab Environment

The experiments were conducted using Python, with key libraries such as Scikit-learn, NumPy, and Pandas employed for model implementation, data processing, and performance evaluation.

E. The Proposed Framework

The proposed liver disease classification framework utilizes a structured machine learning pipeline designed to enhance prediction accuracy and model robustness. The approach emphasizes data quality through preprocessing steps, such as imputation, scaling, and encoding, followed by a stacked model composed of ET and RF as base learners and SVM as the meta-learner. This configuration aims to leverage the strengths of multiple classifiers while maintaining simplicity and interpretability. Figure 3 illustrates the overall architecture of the proposed framework.

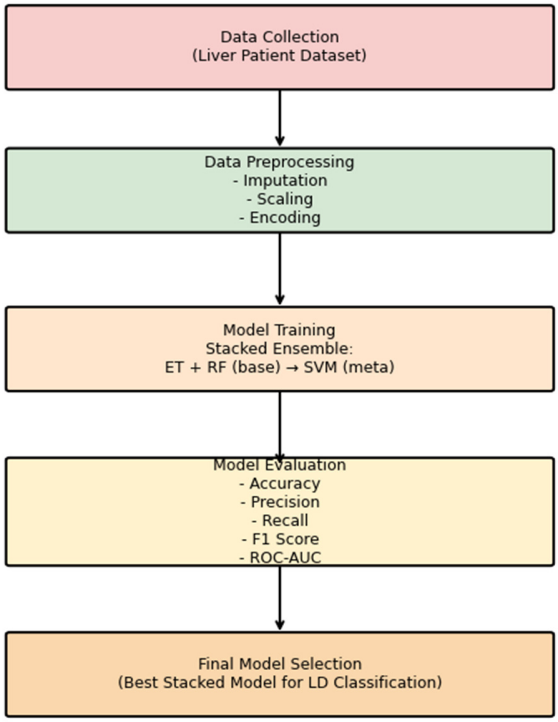

Fig. 3. The proposed framework.

III. RESULTS

This section evaluates the performance of six individual machine learning models developed for the classification of LD prior to the implementation of the proposed stacked model. These baseline models were evaluated using standard performance metrics to identify their strengths and limitations. Based on these insights, the most promising classifiers were integrated into a stacked ensemble framework, combining ET and RF as base learners and SVM as the meta-learner, to enhance predictive performance and improve liver disease detection reliability.

Table I presents a comparative evaluation of the machine learning models developed for LD classification. The ET model achieved the highest standalone accuracy of 0.7982, followed by RF at 0.7456. However, both models showed moderate recall and F1 scores, indicating limitations in detecting true positive cases consistently. SVM and DC showed perfect precision (1.0) but did not identify any positive cases (recall = 0), resulting in an F1 score of 0, affecting their practical effectiveness. In contrast, the proposed stacked ensemble model, combining ET and RF as base learners and SVM as the meta-learner, significantly outperformed all individual models, achieving an accuracy of 0.9853, a precision of 0.9310, a recall of 0.9474, an F1 score of 0.9391, and an AUC of 0.9853. The results demonstrate the robustness and balanced predictive power of the stacked approach, which offers high sensitivity and specificity. Strong performance in all metrics confirms the suitability of the stacked model for reliable and accurate classification of LD.

TABLE I. PERFORMANCE OF ML MODELS

Model	Accuracy	Precision	Recall	F1 score	ROC-AUC
Stacked model	0.9853	0.9310	0.9474	0.9391	0.9853
ET	0.7982	0.6364	0.4828	0.549	0.758
RF	0.7456	0.5	0.3793	0.4314	0.7365
SVM	0.7456	1	0	0	0.7014
DC	0.7456	1	0	0	0.5
KNN	0.6579	0.3611	0.4483	0.4	0.6398
LR	0.7105	0.3889	0.2414	0.2979	0.7399

Figure 4 displays the confusion matrices of the ML models. The stacked model demonstrates exceptional performance, correctly classifying 56 out of 57 instances for both the positive and negative classes, resulting in only one false positive and one false negative. This strong diagonal dominance in the matrix indicates that the stacked model has a well-balanced ability to accurately identify both liver disease and non-disease cases, which aligns with its high accuracy, precision, recall, and F1 score presented earlier. In contrast, the confusion matrices of the baseline models show a higher number of misclassifications. ET misclassified 8 negatives and 15 positives, showing weaker sensitivity. RF and LR each made 11 false positive errors, while also misclassifying 23 and 7 positive cases, respectively, indicating lower recall and overall imbalance. The KNN model performed the worst, with 23 false positives and 18 false negatives, reflecting poor ability to distinguish between the two classes. These comparisons highlight the superior predictive capability and robustness of the stacked model over traditional individual classifiers in the context of LD classification.

Figure 5 displays the ROC curves of the models. The stacked model achieved an AUC score of 1.00, indicating a perfect discriminatory ability between positive and negative classes at all thresholds. Its ROC curve closely follows the top-left corner of the graph, which reflects a high true positive rate and a low false positive rate. In comparison, baseline models demonstrate significantly lower AUC scores, indicating reduced classification performance. ET and RF achieved AUC scores of 0.78 and 0.74, respectively, reflecting moderate predictive ability with visible deviations from the ideal ROC curve. The KNN model showed the weakest performance with an AUC of 0.64, while LR yielded an AUC of 0.70, both indicating relatively poor class separability. These results emphasize the effectiveness of the stacked ensemble model in maximizing classification performance, outperforming traditional models in capturing the complex patterns associated with liver disease.

Table II provides a comparative analysis between the proposed model and a recent BPSO-optimized framework [27] that utilized the same dataset, showing that the proposed stacking approach achieved higher accuracy without relying on additional optimization techniques.

TABLE II. COMPARATIVE ANALYSIS OF LIVER DISEASE CLASSIFICATION MODELS ON THE SAME DATASET

Study	Method	Accuracy
[27]	RF, SVM, DC, ET, KNN, LR with BPSO	85.00 (ET after BPSO)
Proposed	ET, RF (base-classifiers) + SVM (meta-classifier)	98.53

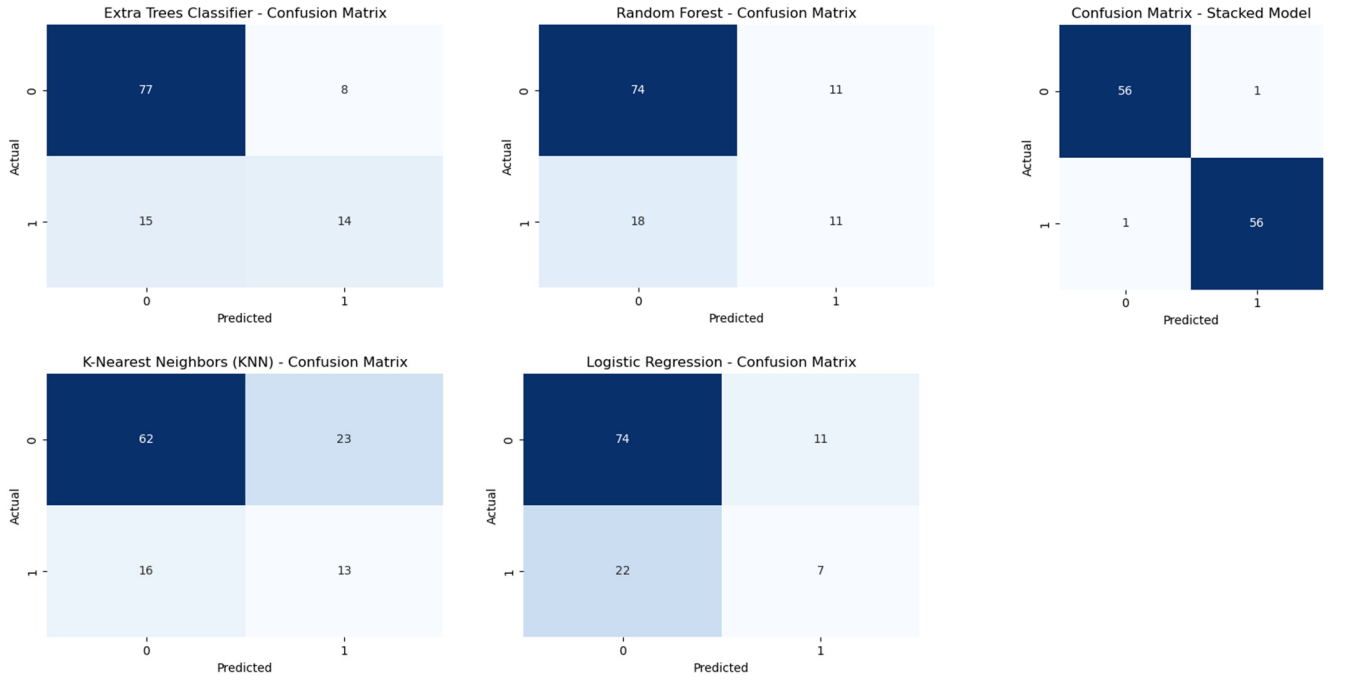

Fig. 4. Confusion matrices of the suggested ML models.

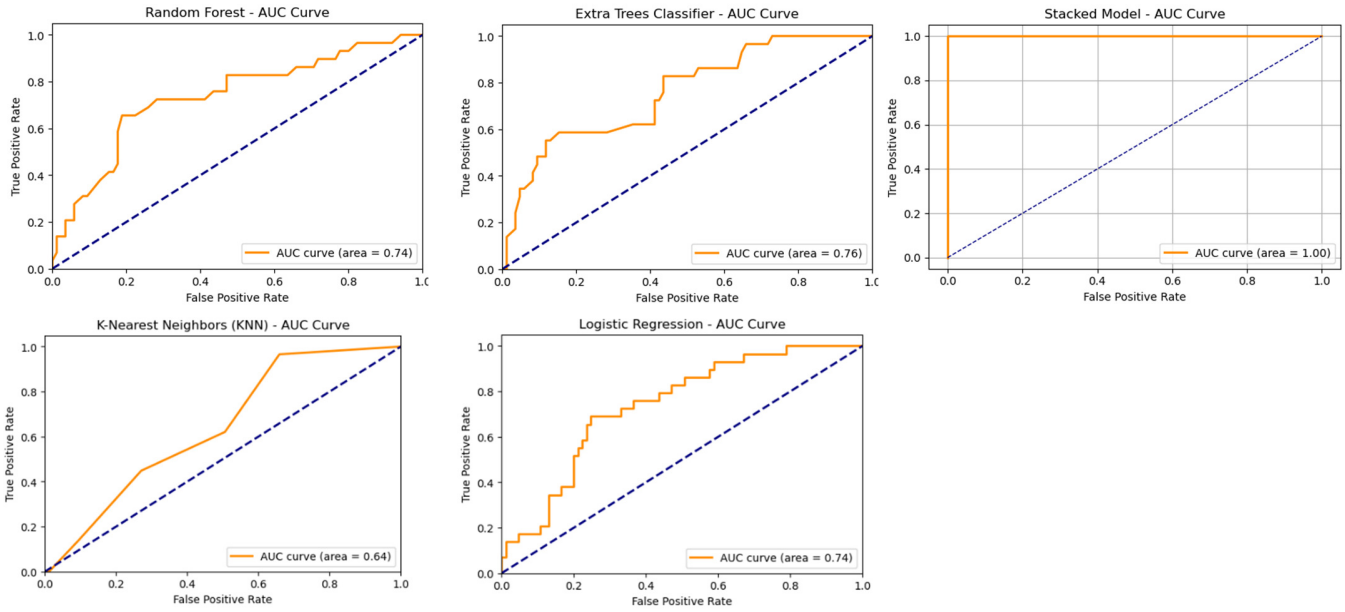

Fig. 5. ROC curves and ROC-AUC scores of the suggested ML models.

IV. CONCLUSIONS AND FUTURE WORK

This study presents a stacked ensemble machine learning model for LD classification, integrating ET and RF as base learners and SVM as the meta-learner. The ensemble model demonstrated superior classification performance compared to individuals, achieving a high accuracy of 98.53% along with strong precision, recall, F1 score, and AUC values. The results emphasize the effectiveness of the stacked learning approach in handling complex medical datasets, such as those involving liver patient records, while ensuring accurate and reliable predictions. The proposed framework highlights the potential of ML techniques to support non-invasive, data-driven diagnostic systems for liver disease, contributing to early detection and improved clinical decision-making. Future work will aim to assess the generalizability of the model on larger and more diverse datasets from various geographic regions. Additionally, solutions to data imbalance will be explored through oversampling techniques or synthetic data generation methods such as SMOTE. Further enhancements to classification performance may also be achieved by incorporating composite or ensemble learning strategies. Real-time implementation in clinical settings will be considered to evaluate the model's practical applicability. Future research may also investigate the integration of the current stacked framework with deep learning architectures to better manage high-dimensional medical data and further improve diagnostic robustness.

REFERENCES

[1] A. Schweitzer, J. Horn, R. T. Mikolajczyk, G. Krause, and J. J. Ott, "Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013," *The Lancet*, vol. 386, no. 10003, pp. 1546–1555, Oct. 2015, https://doi.org/10.1016/S0140-6736(15)61412-X.

[2] M. A. Quadir, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," *Biomedicines*, vol. 11, no. 2, Feb. 2023, Art. no. 581, https://doi.org/10.3390/biomedicines11020581.

[3] A. M. Elshewey, R. Y. Youssef, H. M. El-Bakry, and A. M. Osman, "Water potability classification based on hybrid stacked model and feature selection," *Environmental Science and Pollution Research*, vol. 32, no. 13, pp. 7933–7949, Mar. 2025, https://doi.org/10.1007/s11356-025-36120-0.

[4] A. M. Elshewey, M. A. Aziz, O. A. Gaheen, M. S. Sawah, A. Abd ELhamid, and A. M. Osman, "Prediction of aerodynamic coefficients based on machine learning models," *Modeling Earth Systems and Environment*, vol. 11, no. 3, Mar. 2025, Art. no. 184, https://doi.org/10.1007/s40808-025-02355-6.

[5] E. Dritsas and M. Trigka, "Supervised Machine Learning Models for Liver Disease Risk Prediction," *Computers*, vol. 12, no. 1, Jan. 2023, Art. no. 19, https://doi.org/10.3390/computers12010019.

[6] D. Badvath, A. safali Miriyala, S. chaitanya K. Gunupudi, and P. V. K. Kuricheti, "ONBLR: An effective optimized ensemble ML approach for classifying liver cirrhosis disease," *Biomedical Signal Processing and Control*, vol. 89, Mar. 2024, Art. no. 105882, https://doi.org/10.1016/j.bspc.2023.105882.

[7] C. Zhang *et al.*, "A machine learning-based model analysis for serum markers of liver fibrosis in chronic hepatitis B patients," *Scientific Reports*, vol. 14, no. 1, May 2024, Art. no. 12081, https://doi.org/10.1038/s41598-024-63095-8.

[8] A. M. Hendi, M. A. Hossain, N. A. Majrashi, S. Limkar, B. M. Elamin, and M. Rahman, "Adaptive Method for Exploring Deep Learning Techniques for Subtyping and Prediction of Liver Disease," *Applied Sciences*, vol. 14, no. 4, Jan. 2024, Art. no. 1488, https://doi.org/10.3390/app14041488.

[9] E. Agbozo and D. M. Balungu, "Liver Disease Classification - An XAI Approach to Biomedical AI," *Informatica*, vol. 48, no. 1, Jan. 2024, https://doi.org/10.31449/inf.v48i1.4611.

[10] A. Alizargar, Y. L. Chang, M. Alkhaleefah, and T. H. Tan, "Precision Non-Alcoholic Fatty Liver Disease (NAFLD) Diagnosis: Leveraging Ensemble Machine Learning and Gender Insights for Cost-Effective Detection," *Bioengineering*, vol. 11, no. 6, Jun. 2024, Art. no. 600, https://doi.org/10.3390/bioengineering11060600.

[11] H. M. Marghany *et al.*, "Toward an Accurate Liver Disease Prediction Based on Two-Level Ensemble Stacking Model," *IEEE Access*, vol. 12,

pp. 180210–180237, 2024, https://doi.org/10.1109/ACCESS.2024.3459429.

[12] S. Noor, S. A. AlQahtani, and S. Khan, "XGBoost-Liver: An Intelligent Integrated Features Approach for Classifying Liver Diseases Using Ensemble XGBoost Training Model," *Computers, Materials & Continua*, vol. 83, no. 1, pp. 1435–1450, 2025, https://doi.org/10.32604/cmc.2025.061700.

[13] F. Mehrparvar, "liver disorders." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/fatemehmehrparvar/liver-disorders.

[14] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, no. 12, Jan. 2023, Art. no. 1808, https://doi.org/10.3390/healthcare11121808.

[15] E. Sivari, E. Bostanci, M. S. Guzel, K. Acici, T. Asuroglu, and T. Ercelebi Ayyildiz, "A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models," *Diagnostics*, vol. 13, no. 4, Jan. 2023, Art. no. 720, https://doi.org/10.3390/diagnostics13040720.

[16] J. Liu, X. Dong, H. Zhao, and Y. Tian, "Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion," *Processes*, vol. 10, no. 4, Apr. 2022, Art. no. 749, https://doi.org/10.3390/pr10040749.

[17] M. O. Edeh *et al.*, "Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease," *Frontiers in Public Health*, vol. 10, Apr. 2022, https://doi.org/10.3389/fpubh.2022.892371.

[18] S. R. Velu, V. Ravi, and K. Tabianan, "Data mining in predicting liver patients using classification model," *Health and Technology*, vol. 12, no. 6, pp. 1211–1235, Nov. 2022, https://doi.org/10.1007/s12553-022-00713-3.

[19] A. M. Elshewey and A. M. Osman, "Orthopedic disease classification based on breadth-first search algorithm," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, Art. no. 23368, https://doi.org/10.1038/s41598-024-73559-6.

[20] M. Kaddes, Y. M. Ayid, A. M. Elshewey, and Y. Fouad, "Breast cancer classification based on hybrid CNN with LSTM model," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 4409, https://doi.org/10.1038/s41598-025-88459-6.

[21] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Applied Nanoscience*, vol. 13, no. 3, pp. 1829–1840, Mar. 2023, https://doi.org/10.1007/s13204-021-02063-4.

[22] E. S. M. El-Kenawy, N. Khodadadi, A. Ibrahim, M. M. Eid, A. M. Osman, and A. M. Elshewey, "An optimized model for Liver disease classification based on BPSO Using Machine learning models," *Mesopotamian Journal of Computer Science*, vol. 2024, pp. 214–223, Dec. 2024, https://doi.org/10.58496/mjcsc/2024/017.