Weighted Soft-Voting Ensembles for Liver Disease Prediction: A Large-Scale Comparative Study with Transparent Evaluation

Mohammad Ibraigheeth
Department of Software Engineering, Bethlehem University, Bethlehem, Palestine
mabuayyash@bethlehem.edue (corresponding author)

Suhail Odeh
Department of Software Engineering, Bethlehem University, Bethlehem, Palestine
sodeh@bethlehem.edu

Mahmoud Obaid
Computer System Engineering Department, Arab American University, Jenin, Palestine
mahmoud.obaid@aaup.edu

Received: 8 December 2025 | Revised: 25 December 2025 and 6 January 2026 | Accepted: 9 January 2026

 DOI: https://doi.org/10.48084/etasr.16808

ABSTRACT

Early detection of liver disease can significantly improve patient outcomes and reduce healthcare costs. This study presents a comparative evaluation of four traditional machine learning classifiers—Logistic Regression, Support Vector Machines, Gaussian Naïve Bayes, and a Multi-Layer Perceptron—alongside an enhanced weighted soft-voting ensemble model. Using a large, publicly available clinical dataset (~30,000 records), a fully nested, leakage-free cross-validation framework is employed to ensure robust and reliable evaluation. The proposed ensemble assigns adaptive weights based on per-fold model performance and demonstrates superior discrimination and calibration compared to individual classifiers. The results highlight the contribution of transparent ensemble modeling in achieving accurate and clinically interpretable liver disease risk detection.

Keywords-liver disease; weighted ensemble model; machine learning; traditional classifiers

I. INTRODUCTION

Liver disease is one of the critical health issues worldwide, causing high morbidity and mortality. Early detection and timely intervention and measures are essential to slow disease progression, improve patient outcomes, and reduce healthcare expenditures [1].

A. The Liver and Liver Diseases

The liver, one of the largest and most vital organs in the human body, is essential for numerous physiological processes, including protein synthesis, detoxification, and the production of biochemicals necessary for digestion. A range of conditions can impair its structure and function, posing serious—and sometimes fatal—health risks. Collectively, these conditions are referred to as liver diseases [2, 3] that can be classified into several types, such as Hepatitis, Cirrhosis, Non-Alcoholic Fatty Liver Disease (NAFLD), Metastatic Liver Cancer, and Liver Failure [4-8].

B. Machine Learning Approaches for Liver Disease Prediction

In recent years, healthcare has been changing rapidly due to advances in Machine Learning (ML) and Artificial Intelligence (AI) tools that improve disease diagnosis. ML approaches can learn from data without relying on rule-based systems [9]. Using clinical data, a variety of ML methods can be applied, including traditional techniques, such as Logistic Regression (LogReg) and Support Vector Machines (SVM), as well as more advanced ones, such as Artificial Neural Networks (ANNs), and ensemble models, such as XGBoost and Random Forest (RF). These methods have shown encouraging results in the prediction and classification of many medical conditions. For liver disease, in particular, accurate prediction models can make a real difference by helping physicians identify high-risk patients and optimizing resource allocation in the healthcare system [10]. However, selecting the most suitable ML algorithm for liver disease prediction requires a comprehensive analysis and experimentation with the strengths and limitations of each method.

In [11], the genetic factors that influence the progression of liver disease related to Hepatitis B Virus (HBV) were investigated, with a focus on polymorphisms in the Vitamin D Receptor (VDR) gene—specifically TaqI, ApaI, and BsmI—and related molecules such as GC-Globulin and CYP2R1. This study used methods such as PCR-RFLP, Sanger sequencing, SVM, and LogReg to analyze the data. The findings identified the bAt haplotype and Apa-I CC genotype as independent predictors of HBV progression, while the SVM model achieved 90% accuracy, offering valuable results for disease staging and prognosis. Similarly, in [12], predictive modeling was advanced by developing a Chronic Liver Disease (CLD) detection system using seven boosting algorithms. The results of this study reinforced the promise of ensemble techniques for clinical decision-making.

The study in [13] examined inefficiencies within the U.S. healthcare system, highlighting that despite its high costs, the country does not consistently achieve better health outcomes than other developed nations. To forecast future healthcare spending as a percentage of GDP up to 2050, this study combined machine learning techniques—specifically RF and Support Vector Regression (SVR)—with traditional statistical methods such as ARIMA. RF and ARIMA delivered similar forecasting accuracy, emphasizing the benefits of integrating both data-driven and statistical approaches to support more informed long-term healthcare policy decisions. Extending the application of analytics to infrastructure optimization, in [14], energy consumption patterns were examined in an Italian hospital, with a focus on Heating, Ventilation, and Air Conditioning (HVAC) systems. In a related effort to improve healthcare operations [15], a Multistage Process Monitoring (MPM) tool was proposed to improve the early detection of workflow anomalies. With a more direct focus on liver disease detection, in [16], data from the UCI repository were used to develop a machine learning framework for Hepatitis C stage classification. In [17], the focus was on improving liver disease diagnosis using a machine-learning–based stacked ensemble framework. This study applied several classifiers, and the best accuracy was achieved by combining Extra Trees and RF as base learners with SVM as a meta-classifier [17].

This study aimed to compare various machine learning approaches and identify the most effective method for predicting liver diseases using a large clinical dataset. To achieve this, the performance of several traditional models was evaluated and, finally, an enhanced weighted ensemble method was developed. This enhanced model was built on a previous approach [18] by incorporating two key improvements: a thorough hyperparameter tuning of the individual base classifiers and a more robust weighting scheme that relies on the average performance across all data subsets. This refined ensemble approach provides superior diagnostic performance.

II. METHODOLOGY

This study focused on a comprehensive comparison of four traditional machine-learning classifiers against an enhanced custom ensemble model.

A. Data & Ethics

This study used a publicly available dataset of ~30k clinical records [19]. No identifying information was accessed. Duplicates were eliminated, and missingness patterns and variable units were reviewed.

B. Data Processing and Partitioning

Several preprocessing steps were taken to prepare the data for analysis. First, missing values were treated using suitable imputation methods to ensure the integrity of the data. Then, all features were scaled by standardization, and their range values were normalized, so that model training would not be over-influenced by these values.

All models, both base classifiers and the ensemble model, were trained and evaluated under an identical and leakage-free process on nested 5×5 Cross-Validation (CV). Hyperparameters were optimized on the "inner folds"; final measures were calculated on the "outer folds" to avoid information leakage. The ensemble was provided with the same splits in nested CVs as the baselines, thus having strict apples-to-apples comparisons. This approach trains and tests the model using different sections of the data to ensure fair performance evaluation and comparison.

C. Traditional Classifier Implementations

Four traditional machine learning classifiers were implemented, selected to capture various patterns in the dataset.

- Logistic Regression (LogReg): A type of linear model to perform binary classification.
- Support Vector Machines (SVM): A powerful algorithm used to perform classification tasks by finding the best hyperplane to separate data points.
- Multi-Layer Perceptron (MLP): An ANN capable of learning complex and non-linear relationships in the data.
- Naive Bayes (NB): A probabilistic classifier based on the Bayes theorem, which assumes a certain level of independence between features.

Before being included in the ensemble, each model was exposed to an individually optimized hyperparameter tuning process to improve its performance. This tuning process involved the systematic testing of various combinations of hyperparameters until the best configuration for each model was established (each member of the ensemble was making the best effort to operate at its peak capacity).

D. Enhanced Custom Ensemble Model

The proposed method introduces a weighted ensemble classification algorithm aimed at increasing accuracy and reliability. This novel approach combines different base classifiers in a weighted voting scheme, with dynamic weight assignment dependent on the average performance attained by each classifier when dealing with a different set of training data. The essence of this approach is to combine the exceptional prediction capability of different classifiers to make the final results more robust.

The major improvements made to the older model [18] include hyperparameter tuning. A more extensive hyperparameter search was performed for each base classifier so that each member of the ensemble operated at its best capacity. The final predictions are made using a weighted combination of the predictions from each base classifier.

After training, the performance of each classifier on a subset was evaluated based on a composite performance metric, i.e., P_i. This is an average of four important metrics: Accuracy (AC), F-measure (F1), Cohen's Kappa (K), and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC). The formula for P_i is given by:

$$P_i = \frac{AC_i+F_i+K_i+AUC_i}{4} \quad (1)$$

A performance vector for each classifier is constructed, which contains its P_i value for each of the k subsets. Instead of assuming a more difficult subset, the algorithm computes the average performance of all classifiers on all k subsets. This average score is a more stable and reliable basis for ranking.

E. Ranking and Weight Assignment

The classifiers are ranked as 1 to k (k is the number of classifiers) based on their average independent performance score in all subsets. The classifier with the best performance on average gets the highest rank. A normalized weight vector W is then generated, where the weight for each classifier (w_m) is calculated using its rank (R_m) according to the following equation:

$$w_m = \frac{R_m}{\sum_{m=1}^{k} R_m} \quad (2)$$

This method ensures that classifiers that are effective on average will be given a stronger influence in the final decision.

The final classification decision D is made using a weighted soft voting scheme. For each input, each classifier gives a vector of prediction probabilities D_i. The final prediction is a weighted average of these probabilities, where the weights are calculated in the earlier training step. The class having the highest weighted probability is chosen to be the final output. The prediction is calculated as:

$$D = \sum_{i=1}^{k} w_i D_i \quad (3)$$

The output of the proposed weighted ensemble model is a probability score between 0 and 1, representing the probability of a patient having liver disease. To turn this score into a binary classification (i.e., Positive for disease or Negative for no disease), a classification threshold has to be applied. In the proposed model, a typical threshold of 0.5 was set.

To provide an example of the algorithm, suppose that there are three classifiers in a simple ensemble and three subsets of data. Let $P = (P_1,\ P_2,\ P_3)$, where P_i is the performance vector of classifier i over the three data subsets. Suppose that the following composite performance scores (P_i) have been obtained after training on each subset:

- Classifier 1: $P_1 = (0.77, 0.66, 0.84)$
- Classifier 2: $P_2 = (0.78, 0.90, 0.81)$
- Classifier 3: $P_3 = (0.97, 0.60, 0.78)$

In this method, the average performance for each classifier is calculated across all three subsets to rank them.

- Average Performance of Classifier 1: $(0.77 + 0.66 + 0.84)/3 \approx 0.75$
- Average Performance of Classifier 2: $(0.78 + 0.90 + 0.81)/3 = 0.83$
- Average Performance of Classifier 3: $(0.97 + 0.60 + 0.78)/3 \approx 0.78$

Then, the classifiers are ranked based on these average performance scores. Classifier 2 has the highest average performance and receives the highest rank. The ranking is as follows: Classifier 2 (rank = 3), Classifier 3 (rank = 2), and Classifier 1 (rank = 1). The normalized weight vector is obtained as:

- $w_1 = 1/(1 + 2 + 3) \approx 0.167$
- $w_2 = 3/(1 + 2 + 3) = 0.5$
- $w_3 = 2/(1 + 2 + 3) \approx 0.333$

The final normalized weight vector is approximately $W = (0.167, 0.5, 0.333)$. The highest weight is given to Classifier 2, as it consistently demonstrated superior performance on average across all data subsets.

Finally, the classification decision for a new data instance is obtained based on these weights. Suppose the three classifiers generate probabilities for a Negative outcome: 0.66, 0.35, and 0.74, respectively. The final failure probability generated by the model is the weighted average of these probabilities:

$$D = (0.167 \cdot 0.66) + (0.5 \cdot 0.35) + (0.333 \cdot 0.74) =$$
$$0.11022 + 0.175 + 0.24642 \approx 0.53$$

This final score is then compared against a predefined decision threshold, which is usually determined to be 0.5 for a balanced binary classification. Since this probability (calculated by the model: 0.53) exceeds this value, the new data instance is classified as a Negative case by the model. This weighted procedure allows the ensemble's final prediction to be more reliable by systematically prioritizing classifiers that have proven to be the most accurate and robust on average.

F. Evaluation

All models were evaluated using the same leakage-free protocol to ensure fair comparison. A nested cross-validation approach was applied, with five outer folds for performance evaluation and five inner folds for hyperparameter tuning using only the training data. Data preprocessing steps, including imputation and standardization, were applied within each pipeline and fitted only on training folds. The performance of the model was reported as the mean with 95% Confidence Intervals (CI) for AUROC, AUPRC, F1-score, sensitivity, specificity, and Brier score. Statistical comparisons between models were conducted using DeLong's test for AUROC and McNemar's test for classification performance.

The dataset has different numbers of samples in different classes (some classes appear much more than others). To handle this problem, the learning algorithms were adjusted to give more importance (higher weight) to the minority classes by using *class_weight* = 'balanced'. This makes the model penalize errors on rare classes more than errors on frequent ones. In addition, resampling techniques (such as oversampling or undersampling) were applied only to the training data, not to the validation or test data. This prevents data leakage and ensures that performance results remain fair and reliable.

III. RESULTS AND DISCUSSION

This section compares the performance of the machine learning algorithms in predicting liver disease using the dataset in [19]. The results show how traditional models, such as SVM and LogReg, performed against the proposed implementation of the ensemble. The focus is on determining not only how accurate they are in predicting outcomes but also their robustness and practicality for early diagnosis.

Table I shows a summary of the key features of the dataset used in this study, with distribution, central tendency, and variability. These features play an important role in the interpretation of model behavior and performance. For example, the count column shows the number of non-missing entries for each feature. The Age variable consists of 30,691 complete values, proving the completeness of the dataset for reliable model training. This variable can be used to frame the clinical context of the predictions made by the models. The gender distribution shows that approximately 75% of the patients are males. This variable was binary coded, with 0 representing female and 1 representing male. The age of the patients ranges from 4 to 90 years, with a mean age of 44.11 (standard deviation of 15.98), indicating a middle-aged group with a wide age distribution. Laboratory biochemical measures are also summarized. Direct and total bilirubin levels show wide variation, indicating differing degrees of liver dysfunction among patients. Liver enzymes (Alkaline Phosphatase, SGOT, and SGPT) also exhibit high variability, reflecting heterogeneous liver activity. Protein-related measures, including albumin, total proteins, and A/G ratio, are generally within expected ranges for the studied population. Overall, the dataset is imbalanced, with roughly one-third of patients showing signs of liver disease and the majority classified as healthy.

Table II presents the comparative performance of the proposed ensemble and baseline classifiers on a stratified 20% holdout validation split. The MLP baseline (accuracy = 0.90, ROC-AUC = 0.956) is a single model with a fixed architecture and training setup. Since neural network performance strongly depends on architectural and training choices, this specific configuration was selected for consistency. All baseline results were obtained through this study's experiments, using the same dataset and evaluation protocol, ensuring a fair comparison with the proposed ensemble model.

TABLE I. DESCRIPTIVE STATISTICS OF THE DATASET

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Gender of the patient	30691	0.75	0.44	0	0	1	1	1
Age of the patient	30691	44.1	15.9	4	32	45	55	90
Direct Bilirubin	30691	1.53	2.84	0.10	0.20	0.3	1.4	19.7
Total Bilirubin	30691	3.37	6.19	0.4	0.80	1.00	2.8	75
Alkphos Alkaline Phosphatase	29895	289	238	63	175	209	298	2110
Sgot Aspartate Aminotransferase	30229	111	281	10	26.0	42	88	4929
Sgpt Alanine Aminotransferase	30153	81.4	182	10	23	35	62	2000
Albumin and Globulin Ratio (A/G Ratio)	30132	0.94	0.32	0.3	0.7	0.9	1.1	2.8
Total Proteins	30228	6.48	1.08	2.7	5.8	6.6	7.2	9.6
ALB Albumin	30197	3.13	0.79	0.9	2.6	3.1	3.8	5.5
Result (Liver Disease Indicator)	30691	0.29	0.45	0	0	0	1	1

TABLE II. PERFORMANCE COMPARISON ON A STRATIFIED 20% HOLD-OUT VALIDATION SPLIT

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Initial custom ensemble	0.77	0.77	0.77	0.77	0.858
Improved custom ensemble	0.95	0.95	0.95	0.95	0.984
MLP	0.90	0.90	0.90	0.90	0.956
LogReg	0.73	0.69	0.73	0.68	0.758
SVM	0.74	0.72	0.74	0.66	0.858
NB	0.56	0.80	0.56	0.56	0.733

Table III presents the performance of models on a stratified 20% holdout validation split. All models, including RF, the proposed ensemble, and LogReg, were trained and evaluated using the same training/testing procedure and the same data split to ensure a fair comparison. The RF model achieved the best overall performance, with perfect AUROC and AUPRC scores (1.0 each), very high accuracy (0.999), and the lowest Brier score (0.003). As also reflected in Figures 1 and 2, the Ensemble model performed competitively but consistently slightly below the RF model across all evaluation metrics. In contrast, LogReg showed notably weaker performance, particularly in AUROC (0.75) and AUPRC (0.476), indicating its limited ability to capture complex patterns compared to RF and the proposed ensemble.

Figure 1 illustrates the ROC evaluation of the ensemble and RF models on the 20% holdout validation split. Figure 2 shows the Precision-Recall curves for the ensemble and RF models, demonstrating higher true positive rates across various false positive rates. Figure 3 shows the reliability (calibration) diagram for the proposed ensemble model. The predicted probabilities generally follow the ideal calibration line, but noticeable deviations are present, particularly at medium and high probability ranges, indicating that the model tends to underestimate the true likelihood of positive outcomes.

TABLE III. PERFORMANCE COMPARISON ON A STRATIFIED 20% HOLD-OUT VALIDATION SPLIT

Model	AUROC	AUPRC	F1	Accuracy	Sensitivity	Specificity	Brier
RF	1.0	1.0	0.997	0.999	0.995	1.0	0.003
Ensemble	0.999	0.999	0.994	0.996	0.993	0.998	0.054
LR	0.75	0.476	0.561	0.628	0.83	0.547	0.204

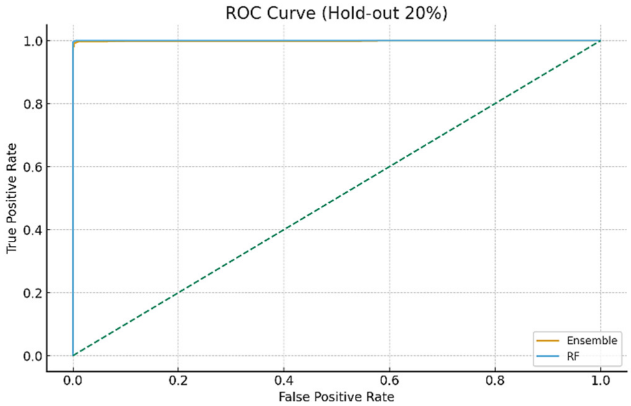

Fig. 1. ROC curve on the 20% hold-out: Ensemble vs. RF.

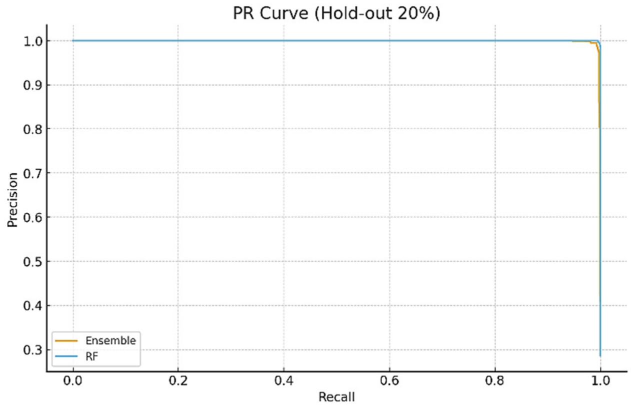

Fig. 2. Precision-Recall curve on the 20% hold-out; Ensemble vs. RF.

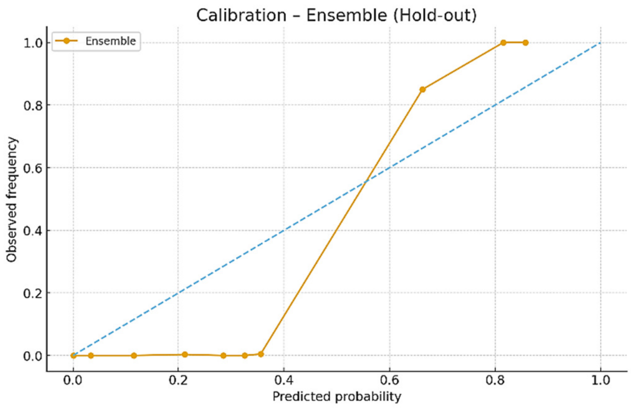

Fig. 3. Reliability (calibration) diagram for the Ensemble (holdout split).

IV. CONCLUSION AND FUTURE WORK

This study conducted a comparative analysis of traditional machine learning models and an improved custom weighted ensemble model for liver disease prediction on a large-scale dataset. The results demonstrate that although commonly used classifiers, such as LogReg, SVM, and NB, achieve baseline performance, and MLP achieves relatively good performance, none of them can achieve the accuracy and robustness of the proposed ensemble method. The proposed ensemble, combining LogReg, NB, SVM, and MLP, optimized hyperparameters, and the improved weighting scheme achieved the highest overall accuracy score (0.95), F1-score (0.95), and ROC-AUC (>0.984), demonstrating better predictions.

The superior performance of the proposed ensemble model can be attributed to the freedom to capture complementary skills among different classifiers while reducing any of their individual weaknesses, effectively addressing the bias-variance tradeoff. These results suggest the potential of carefully designed ensemble frameworks that can be used to improve the diagnostic performance of clinical decision support systems. Beyond being better suited than traditional models, the proposed ensemble is a scalable and reliable tool that could help physicians detect liver disease early, allowing for timely intervention to improve patient outcomes.

Future work may include additional clinical and demographic features to test the model on multi-center clinical datasets to determine its generalizability. In addition, investigating Explainable AI (XAI) methods can improve the interpretability of model predictions. Such extensions could further enhance practical applicability in real healthcare settings.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Deanship of Graduate Studies and Scientific Research at Bethlehem University for their continued support and encouragement. This research was conducted with the support of Research Group #RG-BU006, whose guidance and collaboration have been instrumental in advancing this work. Their contributions are deeply appreciated.

REFERENCES

[1] M. D. Leise, W. R. Kim, W. K. Kremers, J. J. Larson, J. T. Benson, and T. M. Therneau, "A Revised Model for End-Stage Liver Disease Optimizes Prediction of Mortality Among Patients Awaiting Liver Transplantation," *Gastroenterology*, vol. 140, no. 7, pp. 1952–1960, June 2011, https://doi.org/10.1053/j.gastro.2011.02.017.

[2] B. D. Ershoff *et al.*, "Improving the Prediction of Mortality in the High Model for End-Stage Liver Disease Score Liver Transplant Recipient: A Role for the Left Atrial Volume Index," *Transplantation Proceedings*, vol. 50, no. 5, pp. 1407–1412, June 2018, https://doi.org/10.1016/j.transproceed.2018.03.017.

[3] A. Singh *et al.*, "The development of the diabetes liver fibrosis score: A new prediction model to detect advanced fibrosis in diabetics with nonalcoholic fatty liver disease," *Journal of Hepatology*, vol. 68, pp. S98–S99, Apr. 2018, https://doi.org/10.1016/S0168-8278(18)30418-5.

[4] L. Saba *et al.*, "Automated stratification of liver disease in ultrasound: An online accurate feature classification paradigm," *Computer Methods and Programs in Biomedicine*, vol. 130, pp. 118–134, July 2016, https://doi.org/10.1016/j.cmpb.2016.03.016.

[5] M. E. Haas *et al.*, "Machine learning enables new insights into genetic contributions to liver fat accumulation," *Cell Genomics*, vol. 1, no. 3, Dec. 2021, https://doi.org/10.1016/j.xgen.2021.100066.

[6] Y. S. Park, Y. J. Moon, I. G. Jun, J. G. Song, and G. S. Hwang, "Application of the Revised Cardiac Risk Index to the Model for End-Stage Liver Disease Score Improves the Prediction of Cardiac Events in Patients Undergoing Liver Transplantation," *Transplantation Proceedings*, vol. 50, no. 4, pp. 1108–1113, May 2018, https://doi.org/10.1016/j.transproceed.2018.01.024.

[7] R. Masuzaki *et al.*, "Noninvasive Assessment of Liver Fibrosis: Current and Future Clinical and Molecular Perspectives," *International Journal of Molecular Sciences*, vol. 21, no. 14, July 2020, https://doi.org/10.3390/ijms21144906.

[8] J. Singh, S. Bagga, and R. Kaur, "Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques," *Procedia Computer Science*, vol. 167, pp. 1970–1980, Jan. 2020, https://doi.org/10.1016/j.procs.2020.03.226.

[9] A. A. Almelibari, M. I. Labib, and Y. Ramadan, "Enhancing Liver Disease Classification Based on a Stacked Machine Learning Model," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26403–26409, Oct. 2025, https://doi.org/10.48084/etasr.11526.

[10] H. U. Janjua, F. Andleeb, S. Aftab, F. Hussain, and G. Gilanie, "Classification of Liver Cirrhosis with Statistical Analysis of Texture Parameters," *International Journal of Optical Sciences*, vol. 3, no. 2, pp. 18–25, 2017.

[11] M. J. Kalita *et al.*, "Vitamin-d receptor (VDR) polymorphism and types of HBV related liver disease along with an SVM based disease prediction model," *Human Gene*, vol. 37, Sept. 2023, Art. no. 201211, https://doi.org/10.1016/j.humgen.2023.201211.

[12] S. M. Ganie and P. K. Dutta Pramanik, "A comparative analysis of boosting algorithms for chronic liver disease prediction," *Healthcare Analytics*, vol. 5, June 2024, Art. no. 100313, https://doi.org/10.1016/j.health.2024.100313.

[13] J. Wang, Z. Qin, J. Hsu, and B. Zhou, "A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure," *Healthcare Analytics*, vol. 5, June 2024, Art. no. 100312, https://doi.org/10.1016/j.health.2024.100312.

[14] M. Zini and C. Carcasci, "Machine learning-based energy monitoring method applied to the HVAC systems electricity demand of an Italian healthcare facility," *Smart Energy*, vol. 14, May 2024, Art. no. 100137, https://doi.org/10.1016/j.segy.2024.100137.

[15] A. Yeganeh, A. Johannssen, N. Chukhrova, and M. Rasouli, "Monitoring multistage healthcare processes using state space models and a machine learning based framework," *Artificial Intelligence in Medicine*, vol. 151, May 2024, Art. no. 102826, https://doi.org/10.1016/j.artmed.2024.102826.

[16] A. A. Ahad, B. Das, M. R. Khan, N. Saha, A. Zahid, and M. Ahmad, "Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling," *Results in Engineering*, vol. 22, June 2024, Art. no. 102059, https://doi.org/10.1016/j.rineng.2024.102059.

[17] M. Ibraigheeth, "Software project risk assessment using machine learning approaches," *American Journal of Multidisciplinary Research & Development*, vol. 4, no. 2, pp. 35–41, Feb. 2022.

[18] M. A. Ibraigheeth, A. I. A. Eid, Y. A. Alsariera, W. F. Awwad, and M. Nawaz, “A New Weighted Ensemble Model to Improve the Performance of Software Project Failure Prediction,” *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, 2024, https://doi.org/10.14569/IJACSA.2024.0150238.

[19] "Liver Disease Patient Dataset 30K train data." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset.