Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 9 (2027) 3651

Journal of the Nigerian Society of Physical Sciences

Exploring the effect of ensemble size and technique on predictive accuracy of hepatitis B outcome

Solomon M. Emmanuel*, Moses Timothy, Ovre J. Agushaka

Department of Computer Science, Faculty of Computing, Federal University of Lafia, Nigeria

Abstract

Machine learning (ML) models for hepatitis B virus (HBV) outcome prediction must address high variance, bias, and class imbalance. Previous ensemble-learning studies have commonly combined only two models or considered a limited set of ensemble techniques. This study evaluates how ensemble size, from two to five models, affects performance across voting, bagging, boosting, and stacking approaches. Experiments evaluated ten machine learning models on the UCI Hepatitis dataset using 5-fold cross-validation, with performance assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Among individual models, the support vector machine (SVM) achieved the highest accuracy (84.52%) and an AUC of 0.873, while CatBoost and random forest (RF) achieved accuracies of 83.87% and 83.23%, respectively; K-nearest neighbors (KNN) and naïve Bayes (NB) produced comparatively high recall. The five-model soft-voting ensemble achieved 91.84% accuracy, 0.960 AUC, 92.16% F1-score, and 95.92% recall. These findings indicate that multi-model ensembles can improve outcome prediction and high-risk patient identification. The results also suggest that performance depends on strategic model selection rather than ensemble size alone. The proposed $2 + n$ framework provides a structured approach to flexible model combination across ensemble techniques. Validation using larger, independent clinical datasets is required before broader application.

DOI: 10.46481/jnsps.2027.3651

Keywords: Hepatitis B virus (HBV), Machine learning, Ensemble learning, Synthetic minority over-sampling technique (SMOTE), Predictive accuracy

Article history:
Received: 06 July 2026
Received in revised form: 29 July 2026
Accepted for publication: 09 September 2026
Available online: 12 October 2026

© 2027 The Author(s). Published by the Nigerian Society of Physical Sciences under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Communicated by: S. Folorunso

1. Introduction

Machine learning (ML) has become increasingly important for decision-making in fields such as healthcare [1]. ML enables researchers to identify and forecast disease progression in patients with hepatitis B, a viral infection that can cause severe health complications. Combining multiple ML models through ensemble learning can improve predictive performance by addressing limitations of individual methods [2]. Merging more than two models and using ensemble techniques such as bagging and boosting can further enhance predictive performance [3, 4]. This study develops and evaluates ensemble methods for forecasting outcomes in patients with hepatitis B.

Hepatitis B virus (HBV) infection affects approximately 254 million people worldwide and can be transmitted through infected blood, unprotected sexual contact, or maternal transmission [5, 6]. Although hepatitis B vaccination is an effective means of preventing hepatitis B virus (HBV) infection, many individuals with HBV infection remain undiagnosed because the infection is often asymptomatic or produces only mild symptoms in its early stages. Consequently, individ-

*Corresponding Author Tel. No.: +234-806-480-0595.
Email address: solomon.m.emmanuel@gmail.com (Solomon M. Emmanuel)

uals may remain unaware of their infection until substantial liver disease develops, highlighting the importance of early detection and screening alongside vaccination. Conventional assessment relies on blood tests and clinical evaluations [7], which may not identify early disease progression in every case. Machine-learning algorithms can extract information from patient databases to support prediction [8]. Ref. [9] reported 95% prediction accuracy using model combinations, although broader multi-model integration and ensemble techniques remain areas of research. Ensemble learning has also been investigated for early detection and treatment-related applications in hepatitis B [10].

Ref. [9] reported progress in hepatitis B outcome prediction using ML models, with 95% accuracy obtained by combining logistic regression (LR) and support vector machine (SVM). Such predictive systems may support timely treatment, patient monitoring, and lifestyle modification. The study nevertheless had three constraints: it limited the ensemble to two models, used a voting classifier without comparing bagging, boosting, or stacking, and used a dataset with substantial class imbalance (123 survival cases and 32 mortality cases), which can complicate predictive performance and identification of high-risk patients.

Ref. [11] developed an interpretable ML system for diagnosing hepatitis B. The study combined predictive performance with explanatory methods and used feature assessment and visualization to show how patient attributes affected predictions. The work emphasized interpretability as an important requirement for clinical systems [12] and provided a basis for subsequent HBV prediction research.

Harabor and colleagues used ML techniques to predict hepatitis B and C seropositivity. Their study employed logistic regression, support vector machines, and ensemble approaches to analyze demographic and laboratory data [13]. The authors reported accuracy above 90% for some methods and emphasized consideration of both precision and recall in medical testing.

Ref. [14] investigated machine-learning methods for predicting disease progression in chronic HBV infection.

Feng and Cheng (2024) developed an XGBoost-based model for predicting esophagogastric varices using ultrasound and conventional blood-test measurements.

Ref. [15] used ML to study serological HBV profiles and associated risk factors. The resulting models provided information relevant to public-health assessment of HBV distribution. Ref. [5] developed virology-informed neural networks (VINNs) to model HBV infection by incorporating biological knowledge into predictive methods.

Ref. [16] are cited in the manuscript in connection with the effect of multiple complications on survival in patients with HBV-associated cirrhosis. The cited work used multivariable ML models to examine 1-, 3-, and 5-year mortality and highlighted the importance of considering multiple comorbidities in patient assessment.

Pavani and colleagues developed ML methods for liver-disease prediction that included feature-selection techniques and hybrid ensemble models [17]. Related research also reported that appropriate feature selection and ensemble learning can improve generalization [18]. In Kazakhstan, researchers used administrative health data to forecast one-year mortality among patients with hepatitis, showing the potential value of non-clinical health-system data for prognosis and resource planning [19].

Ref. [20] conducted a comparative analysis of deep-learning and ensemble methods for time-series forecasting, demonstrating the application of ensemble approaches to forecasting tasks. The manuscript links this work to HBV and cites Ref. [21] as evidence that combining model types can improve prediction for chronic-disease applications over time.

Bai and colleagues combined ML with liver stiffness measurements (LSM) to identify liver cirrhosis in patients with chronic HBV [22]. Their approach demonstrated the potential of non-invasive measurements combined with ML. Butkevych and colleagues applied recurrent neural networks (RNNs) to forecast acute and chronic hepatitis B case counts in Ukraine [23]. Using long-term epidemiological surveillance data, the model captured seasonal patterns, outbreak trends, and other temporal dynamics. The study illustrates how deep learning may support public-health planning beyond individual diagnosis [24].

Ref. [9] proposed an ensemble ML approach for predicting hepatitis B patient outcomes. The authors tested algorithms including random forest (RF), Light Gradient Boosting Machine (LightGBM or LGBM), and LR and combined selected models in a voting ensemble. They also used SHAP to interpret the effects of features such as platelet count and alanine aminotransferase (ALT). The ensemble combinations were limited to two models at a time, leaving scope for larger and more diverse ensembles.

The workflow of the existing system in Ref. [9] is shown in Figure 1. It includes duplicate removal, missing-value handling, chi-square feature selection, synthetic minority oversampling technique (SMOTE) balancing, standardization, and model training. Models include SVM, decision tree, RF, naïve Bayes, LR, and boosting-based algorithms. The two highest-performing models, SVM and LR, each achieved approximately 92.5% accuracy and were combined in a voting classifier that achieved approximately 95% accuracy. Performance was evaluated with confusion matrices and receiver operating characteristic (ROC) curves, while SHAP and LIME were used for explainability.

The ensemble framework in Ref. [9] shows strong predictive performance but has methodological limitations. The train/test split was performed before dataset balancing, which may have introduced a potential risk of information leakage. It also notes that the preprocessing description is incomplete, which makes the experimental setup difficult to assess. In addition, reliance on basic voting may limit model diversity. These observations motivate evaluation of data-processing choices and aggregation approaches such as bagging and boosting.

Unlike earlier hepatitis B prediction studies that generally combined two classifiers or tested a single ensemble method, the proposed $2 + n$ framework provides a structured, incremental approach to ensemble construction. It begins with the best-performing base classifier and adds complementary mod-

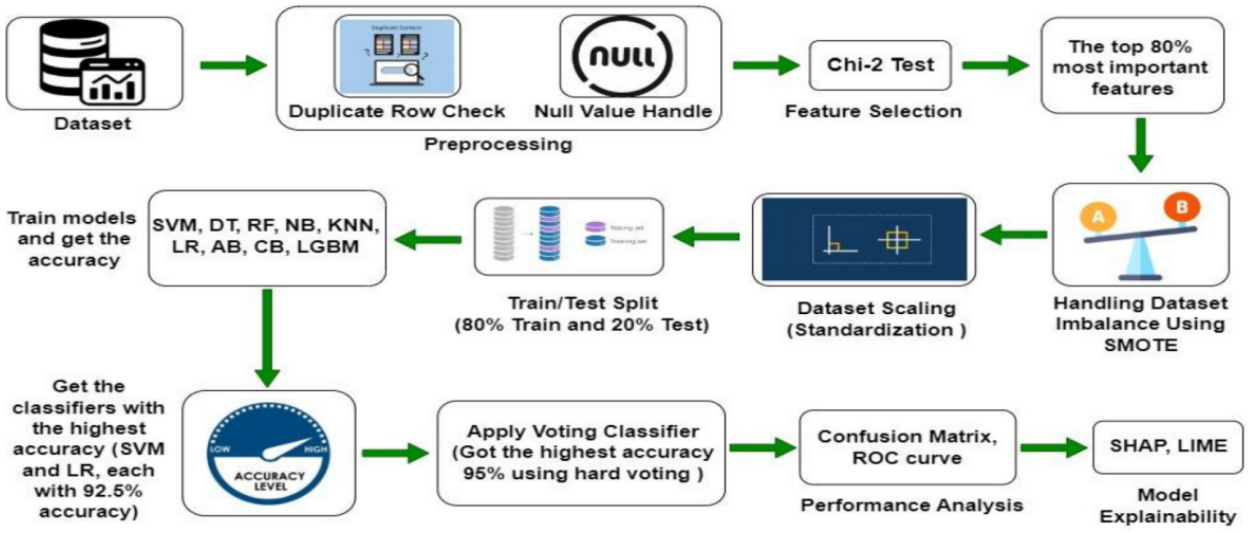

Figure 1: Architectural model of the existing system.

els using greedy forward selection. The framework evaluates increasing ensemble size across voting, bagging, boosting, and stacking. Its contribution is therefore a structured experimental procedure for examining how ensemble size and ensemble technique affect predictive performance and interpretability in hepatitis B outcome prediction.

The aim of this study is to develop and evaluate an ensemble-learning framework that combines ML models at a rate of $2 + n$, where $n \geq 0$, using different ensemble methods to predict hepatitis B patient outcomes and to interpret the results with SHAP and LIME. The objectives are to evaluate LR, decision tree classifier (DT), RF, SVM, K-nearest neighbors (KNN), naïve Bayes (NB), AdaBoost (AB), CatBoost (CB), and LGBM using clinical and demographic data; to evaluate ensembles of two to five models using voting, bagging, boosting, and stacking and compare accuracy, precision, recall, F1-score, and area under the ROC curve (AUC); and to integrate SHAP and LIME explanations to improve the interpretability of predictions for healthcare professionals.

2. Workflow of the new methodology

This study adopts an experimental research design. Figure 2 summarizes the workflow used to identify combinations of models and techniques that provide a suitable balance of accuracy and interpretability.

2.1. Dataset collection and description

The workflow begins with the UCI Hepatitis dataset, which contains 155 patient records with demographic and clinical data. The target variable represents whether a patient survived or died. Missing values occur in several clinical attributes; before preprocessing, 167 of 3,100 data cells were missing, corresponding to an overall missing-data rate of approximately 5.39%.

Table 1: Summary of the UCI Machine Learning Repository hepatitis dataset.

S/N	Features	Description
1	Dataset types	Multivariate
2	Attribute types	Categorical, integer, real
3	Total features (columns)	20
4	Total instances	155
5	Numeric features	6
6	Categorical features	14
7	Target column	Class
8	Classification categories	Die/Live
9	Number of live instances	123
10	Number of die instances	32
11	Data source	UCI ML Repo

2.2. Data preprocessing and preparation

The dataset was obtained from the University of California, Irvine (UCI) Machine Learning Repository. As summarized in Table 1, it contains 20 attributes: 19 input features comprising clinical, laboratory, and demographic information and one target variable indicating whether the patient died or lived.

The input variables are Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, Alkaline Phosphatase, SGOT, Albumin, Protime, and Histology; the Class attribute is the target outcome (Die or Live). The dataset was examined for duplicate samples. Missing numerical values were filled with the median using the `fillna` method. Median imputation was selected because the dataset is relatively small and contains several skewed clinical measurements; it is less sensitive to outliers than mean imputation and provides a simple, reproducible approach. Multiple imputation by chained equations (MICE) may

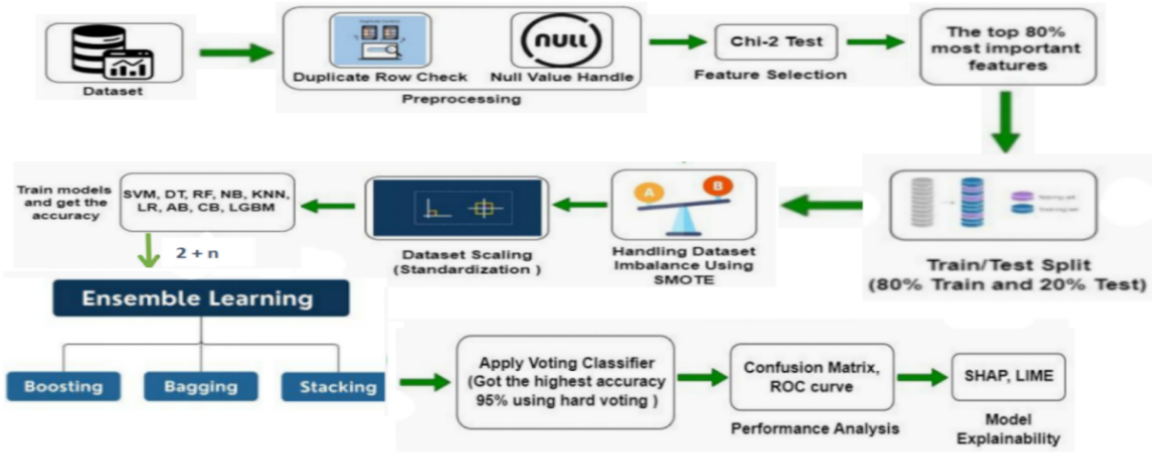

Figure 2: Architectural model of the proposed system.

be explored in future work.

After data preprocessing, the dataset was divided into training and testing sets using a stratified 80/20 split to preserve the original class distribution. Feature selection was then performed on the training data using the chi-square (χ^2) test, and the top 80% of the most significant features were selected [25].

$$Z = \frac{x_i - \mu_i}{\sigma_i}, \qquad [x_i, \mu_i, \sigma_i] \in f_i, \tag{1}$$

where x_i is the original value of the $(i)th$ continuous numerical feature f_i, μ_i is the mean of feature f_i, and σ_i is the standard deviation of feature f_i.

The selected features were subsequently standardized using Z-score scaling, with scaling applied only to continuous numerical variables, such as age and laboratory measurements, and not to categorical variables after numerical encoding. SMOTE was then applied only to the training data to address class imbalance and prevent information leakage into the testing set. The same selected feature subset and scaling procedure were applied to the testing data without applying SMOTE.

Scaling gives the features comparable influence and helps the models learn signal rather than noise.

2.3. Train/test split

Stratified sampling was used to maintain the original class distribution, with 80% of the data allocated to training and 20% to testing. The train and test subsets therefore retained the expected minority-to-majority distribution, with an overall survival-to-mortality ratio of approximately 3.8:1. The models learned patterns from the training data and were evaluated on previously unseen testing data.

2.4. Model training and evaluation

Ten models were evaluated: LR, DT, RF, SVM, gradient boosting classifier (GB), KNN, NB, AB, CB, and LGBM. For reproducibility, predefined parameter settings from scikit-learn, CatBoost, and LightGBM were used. LR was configured with `max_iter = 500`, `class_weight = "balanced"`, and `random_state = 42`. DT and RF also used `class_weight = "balanced"` and `random_state = 42`. SVM used `probability = True`, `class_weight = "balanced"`, and `random_state = 42`.

2.5. Ensemble learning (2 + n models)

Unlike earlier studies that limited ensemble construction to two-model pairings, the proposed $2 + n$ approach incrementally expands the ensemble using greedy forward selection. It begins with the best-performing classifier and adds models only when they provide measurable improvement. This procedure allows systematic evaluation of the relationship between ensemble size and predictive performance across voting, boosting, bagging, and stacking [2].

The study tests the hypothesis that combining different ML models improves performance relative to individual models. To avoid combinatorial explosion, greedy forward selection starts with the best-performing classifier and iteratively adds the model that most improves ensemble performance. Here, n denotes the number of additional base models selected through the greedy procedure, producing ensemble sizes from two to five models.

2.6. Performance analysis

Base models and ensembles were compared using accuracy, precision, recall (sensitivity), F1-score, AUC, and the standard deviation of AUC across 5-fold cross-validation. Recall was emphasized because false-negative outcomes are important in high-risk patient identification. No statistical significance tests were performed because of the limited sample size and exploratory nature of the study.

2.7. Model explainability

To support interpretability, final-model predictions were examined with SHAP (SHapley Additive exPlanations), which assigns a contribution value to each feature for each prediction

[26], and LIME (Local Interpretable Model-agnostic Explanations), which fits local surrogate models to explain individual predictions [27].

2.8. Experimental design

The experimental design is objective-driven: each objective is mapped to a specific experiment with controlled variables and standardized evaluation metrics. Figure 3 summarizes the mapping of experiments to outcomes.

For the first objective, after preprocessing, the UCI hepatitis B dataset was divided using a stratified 80/20 train/test split. Feature selection and scaling were then performed, followed by SMOTE applied only to the training data to avoid leakage, and individual models were evaluated using 5-fold cross-validation. Accuracy, precision, recall, F1-score, AUC, and standard deviation of AUC were recorded.

For the second objective, greedy forward selection was used to construct ensembles by combining the two best models and then successively adding models to obtain three-, four-, and five-model ensembles. Voting, bagging, boosting, and stacking were evaluated under the same dataset and evaluation conditions to determine how performance changed with ensemble size. For the third objective, SHAP provided global explanations and LIME provided local explanations of model predictions [1].

3. Results and discussion

After data preprocessing, the dataset was divided using a stratified 80/20 train/test split, producing 124 training and 31 testing instances. Feature selection retained 15 attributes: Age, Sex, Steroid, Fatigue, Malaise, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, Alkaline Phosphatase, SGOT, Albumin, Protime, and Histology. The selected features were standardized using Z-score scaling, and SMOTE was applied only to the training data, yielding a balanced training set of 196 samples. Model performance was evaluated using 5-fold stratified cross-validation.

3.1. Distribution of nominal features

Figure 4 shows the distribution of nominal features by patient outcome (Die or Live). The two outcome groups show different patterns across clinical and demographic indicators. Fatigue and Malaise occur in both groups, while treatment factors such as Steroid use and diagnostic factors such as Histology show distinct distributions. Ascites and Varices occur less frequently, reflecting their presence in fewer patients in this dataset.

3.2. Performance of individual models

The analysis compares the predictive performance of ML models on the imbalanced hepatitis B dataset. Metrics obtained through 5-fold stratified cross-validation include accuracy, precision, recall, F1-score, and AUC. AUC reflects discriminatory ability, while recall is particularly relevant to identification of high-risk patients. The standard deviation of AUC (Std AUC) was also used to assess stability across folds. Table 2 summarizes the results.

The results in Table 2 show differences in classification behavior across models. SVM achieved the highest mean AUC and the lowest variability across folds, indicating comparatively strong generalization in this experiment. CatBoost and RF also showed competitive performance. NB and KNN produced high recall at the expense of precision, reflecting more aggressive minority-class detection.

3.3. Interpretation of class-wise performance

The confusion matrices in Figure 5 show the models' performance on survival and mortality outcomes. The class imbalance makes mortality prediction more difficult and increases the importance of recall for the minority class.

NB and KNN show relatively high sensitivity for deceased patients but also produce more false positives by classifying some living patients as deceased. SVM, CatBoost, and RF provide a different balance between recall and precision. ROC analysis further evaluates the ability of the models to distinguish survival from mortality outcomes. Figure 6 presents the 5-fold cross-validation ROC curves.

As shown in Figure 6, the decision-tree model has the weakest class-separation ability, with an ROC curve closer to the diagonal. Several ensemble methods produce ROC curves that rise more rapidly toward the upper-left region. The results demonstrate the need to assess multiple metrics because increased recall for mortality prediction can be accompanied by more false-positive alerts. SVM and CatBoost provide comparatively balanced sensitivity and precision in this experiment.

The model-explainability outputs are shown in Figures 7 and 8 and are discussed in the model-explainability subsection.

3.4. Ensemble learning results

Base classifiers for the ensembles were selected using the forward greedy approach. The evaluated ensemble strategies were hard and soft voting, bagging, boosting, and stacking.

3.4.1. Voting

Ensembles containing two to five models were evaluated using hard and soft voting. Table 3 presents the results.

Voting ensembles outperformed the individual models on several metrics in Table 3. The two-model SVM–RF ensemble achieved 88.27% accuracy with hard voting and 90.31% accuracy with soft voting, with an AUC of 0.963 for the latter. Adding CatBoost produced 90.82% accuracy, 93.88% recall, and an AUC of 0.964 with soft voting. The four-model ensemble showed a small reduction in some metrics after LR was added. The five-model soft-voting ensemble achieved the highest accuracy (91.84%), with 95.92% recall, a 92.16% F1-score, and an AUC of 0.960.

3.4.2. Bagging ensemble results

The performance of the bagged classifiers was assessed using 5-fold cross-validation, with metrics including accuracy, precision, recall, F1-score, area under the curve (AUC), and

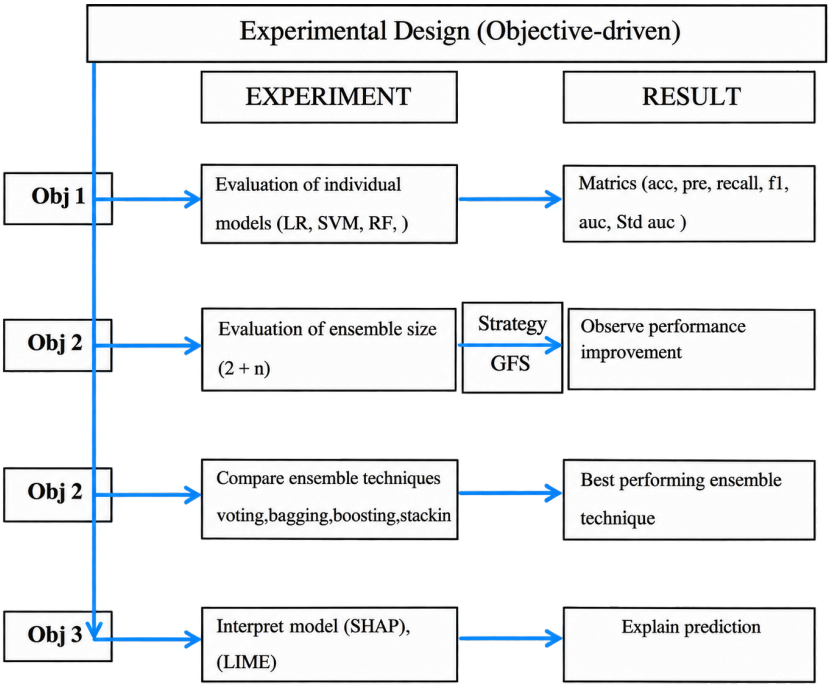

Figure 3: Objective-driven experimental design.

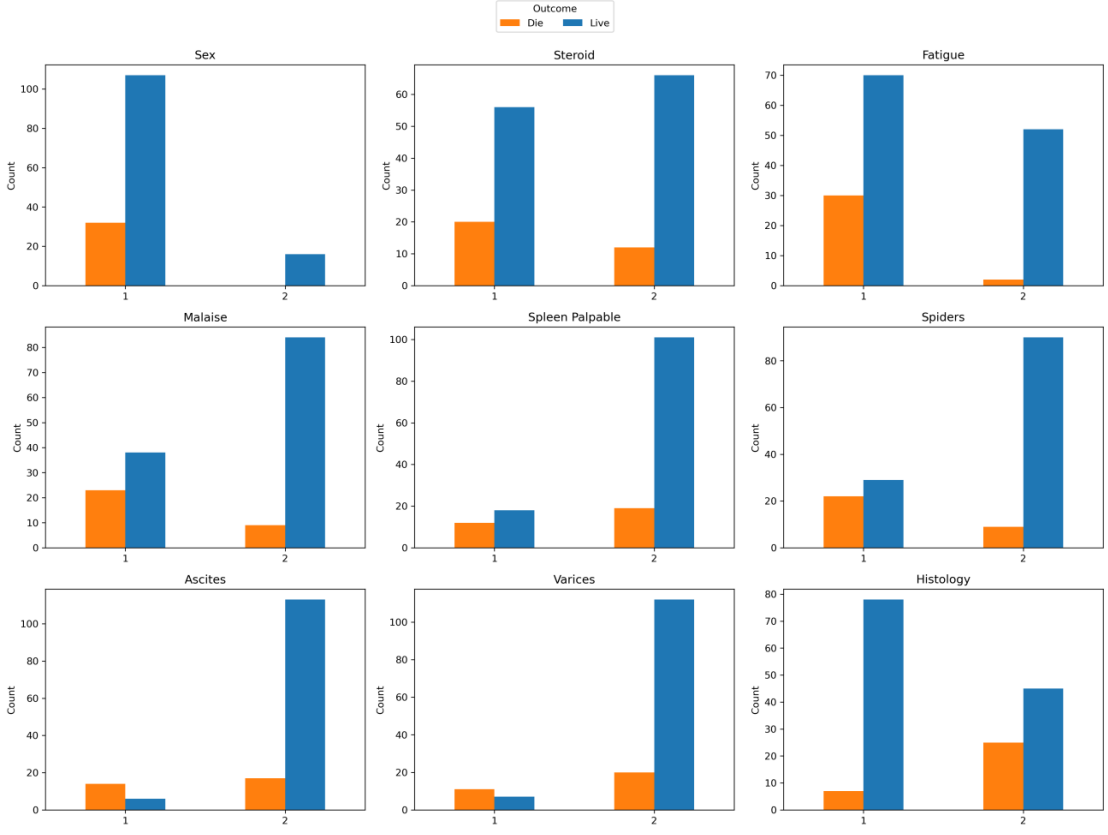

Figure 4: Distribution of nominal features.

the standard deviation of AUC (Std AUC) to measure stability across folds. The results in Table 4 show that bagging with logistic regression demonstrated the strongest overall performance across the clinically relevant evaluation metrics per-

Table 2: Comparative performance of individual models (5-fold cross-validation).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Std AUC
Support Vector Machine	84.52	65.52	57.14	59.62	0.873	0.043
CatBoost	83.87	60.95	63.33	61.95	0.843	0.062
Random Forest	83.23	64.00	50.48	55.76	0.842	0.069
Naive Bayes	65.16	37.16	84.76	51.10	0.840	0.067
K-Nearest Neighbors	76.13	46.50	81.43	58.98	0.837	0.061
Logistic Regression	80.65	53.69	66.19	58.66	0.828	0.071
Gradient Boosting	78.71	51.21	47.62	49.17	0.813	0.076
AdaBoost	80.65	56.07	50.48	52.94	0.811	0.049
LightGBM	78.71	50.50	56.19	52.47	0.807	0.060
Decision Tree	71.61	38.02	40.95	38.55	0.602	0.077

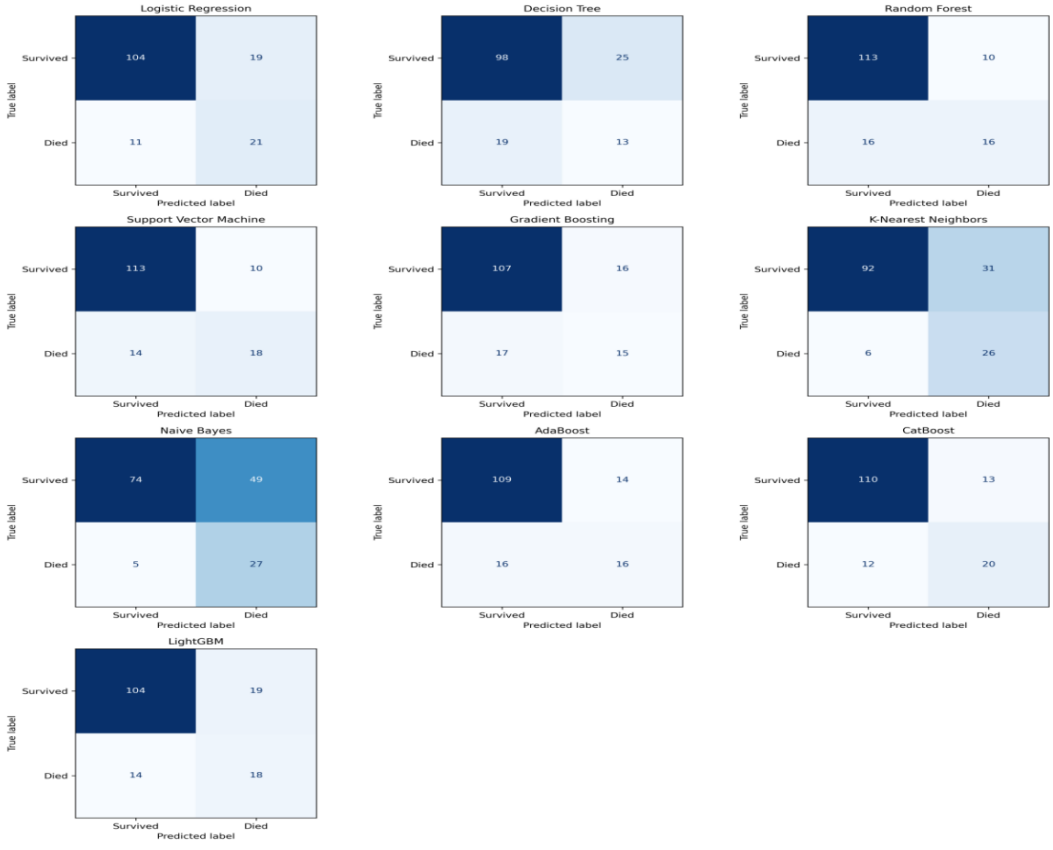

Figure 5: Confusion matrices of the evaluated classifiers.

formance among all studied ensembles because it achieved 81.29% accuracy and 68.57% recall and 0.843 AUC performance while using bootstrap resampling to decrease variance. The bagging method with decision trees achieved 83.23% accuracy but only 44.76% recall which showed its struggle to identify cases of the minority class. K-Nearest Neighbors (KNN) achieved poor performance because it attained an F1 score of 9.35% and an AUC score of 0.642 because the system lost stability when local data distributions changed.

3.4.3. Boosting ensemble results

Four boosting algorithms were assessed using 5-fold stratified cross-validation, as shown in Table 5.

CatBoost achieved 83.87% accuracy and an AUC of 0.848. This performance is attributed to ordered boosting, which can reduce prediction error and overfitting on small datasets. AdaBoost achieved 83.23% accuracy and an AUC of 0.828. Gradient Boosting achieved 81.94% accuracy and an AUC of 0.792, while LightGBM produced the lowest AUC among the boosting models. Under the evaluated conditions, voting ensembles outperformed these boosting configurations on the principal metrics.

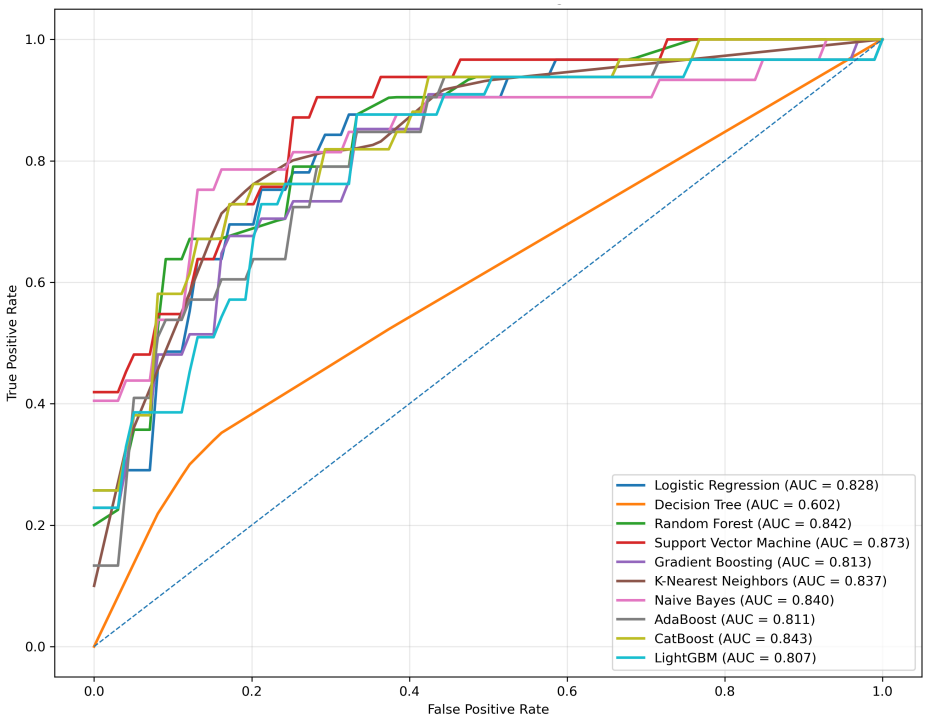

Figure 6: ROC-curve analysis (5-fold cross-validation).

Table 3: Performance of voting-ensemble models.

No. of models	Base models	Voting type	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
2	SVM, RF	Hard	88.27	87.88	88.78	88.32	N/A
2	SVM, RF	Soft	90.31	88.35	92.86	90.55	0.963
3	SVM, RF, CatBoost	Hard	90.31	86.92	94.90	90.73	N/A
3	SVM, RF, CatBoost	Soft	90.82	88.46	93.88	91.09	0.964
4	SVM, RF, CatBoost, Logistic Regression	Hard	89.80	88.24	91.84	90.00	N/A
4	SVM, RF, CatBoost, Logistic Regression	Soft	90.31	88.35	92.86	90.55	0.959
5	SVM, RF, CatBoost, Logistic Regression, KNN	Hard	90.31	87.62	93.88	90.64	N/A
5	SVM, RF, CatBoost, Logistic Regression, KNN	Soft	91.84	88.68	95.92	92.16	0.960

Table 4: Performance of bagging-based ensemble models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Std AUC
Bagging (Logistic Regression)	81.29	55.24	68.57	60.00	0.843	0.063
Bagging (Decision Tree)	83.23	65.33	44.76	52.77	0.811	0.071
Bagging (KNN)	76.77	25.00	6.19	9.35	0.642	0.109

3.4.4. Stacking ensemble results

Stacking ensembles with two to five base models used LR as the meta-learner to assess whether increasing base-model diversity improved prediction. Accuracy ranged from 79.35% to 81.29%, and AUC ranged from 0.846 to 0.852. The two-model SVM–RF configuration achieved 81.29% accuracy, a 63.51% F1-score, and an AUC of 0.852 (Table 6).

Table 5: Performance of boosting-based ensemble models (5-fold cross-validation).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Std AUC
CatBoost	83.87	67.67	47.62	55.61	0.848	0.053
AdaBoost	83.23	62.57	50.48	55.66	0.828	0.074
Gradient Boosting	81.94	63.67	50.48	55.32	0.792	0.071
LightGBM	81.29	62.79	46.67	51.70	0.768	0.054

Table 6: Performance of stacking ensembles.

No. of models	Base models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Std AUC
2	SVM, RF	81.29	56.57	75.24	63.51	0.852	0.082
3	SVM, RF, CatBoost	79.35	52.97	66.19	58.00	0.852	0.081
4	SVM, RF, CatBoost, LogReg	80.65	54.52	66.19	59.24	0.850	0.076
5	SVM, RF, CatBoost, LogReg, KNN	80.00	53.45	66.19	58.67	0.846	0.081

Table 7: Comparative performance of ensemble methods.

Ensemble method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	Std AUC
Voting (Soft, 5 models)	91.84	88.68	95.92	92.16	0.960	0.028
Bagging (Decision Tree)	83.23	65.33	44.76	52.77	0.811	0.071
Boosting (CatBoost)	83.87	67.67	47.62	55.61	0.848	0.053
Stacking (2 models)	81.29	56.57	75.24	63.51	0.852	0.082

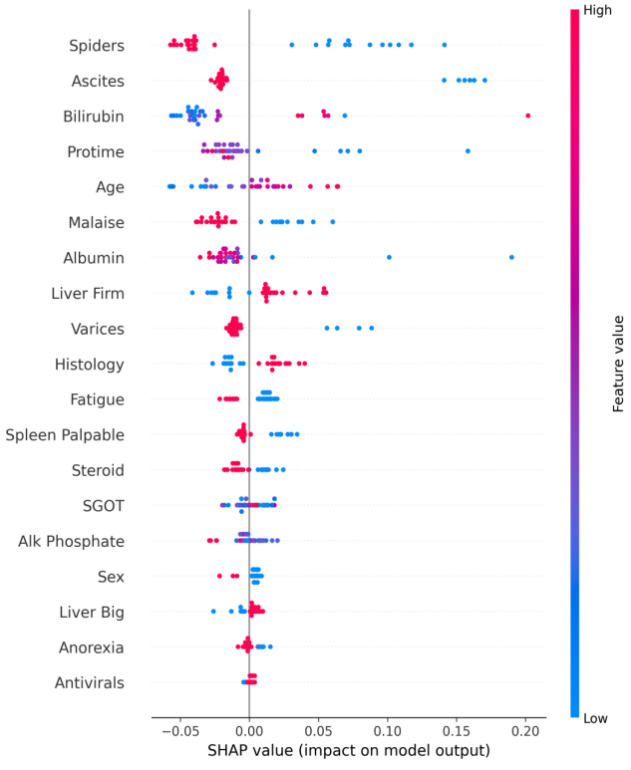

Figure 7: SHAP summary plot for the soft-voting ensemble.

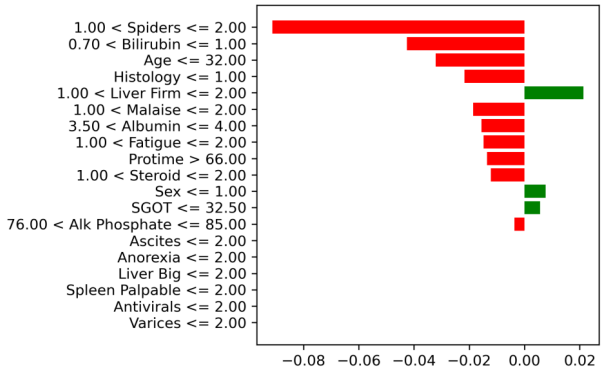

Figure 8: LIME explanation for soft voting.

3.4.5. *Comparison of voting, bagging, boosting, and stacking*

The four ensemble strategies were compared using accuracy, precision, recall, F1-score, AUC, and Std AUC. The five-model soft-voting ensemble (SVM, RF, CatBoost, LR, and KNN) achieved 91.84% accuracy, a 92.16% F1-score, and an AUC of 0.960. Table 7 summarizes the reported comparison.

3.5. *Model explainability*

The study evaluates how interpretable ML methods can support clinicians making decisions about hepatitis B prognosis. SHAP and LIME were applied to the soft-voting ensemble. As shown in Figures 7 and 8, the manuscript identifies Spiders, As-

cites, Bilirubin, and Protime as important mortality-related predictors, while higher Albumin values are associated with better survival predictions. SHAP illustrates global feature effects, including nonlinear effects of age, while LIME provides a local explanation for an individual prediction. These analyses are intended to improve model transparency; they do not establish causal relationships.

3.6. Discussion of findings

This study examines how ensemble size and ensemble technique affect ML prediction of hepatitis B outcomes. Ensemble methods produced different levels of predictive performance. Voting achieved the strongest overall metrics among the evaluated strategies, while the manuscript describes bagging as reducing variance, boosting as improving the ability to identify difficult cases, and stacking as providing balanced performance. Increasing the number of base learners did not produce monotonic improvement, indicating diminishing returns for some configurations. LR as a stacking meta-learner provided comparatively stable performance.

The study has important limitations. The dataset contains only 155 instances, so model estimates may vary substantially and complex patterns may be difficult to identify reliably. Stratified 5-fold cross-validation and SMOTE were used to address evaluation stability and class imbalance, but they do not remove the limitations of the sample size. The retrospective design also limits immediate real-world applicability, and external validation is needed. Ensemble methods increase system complexity, while SHAP and LIME provide explanations of model behavior but do not establish causal effects.

The small sample size may also limit generalizability to broader clinical populations. Stratified 5-fold cross-validation was used to obtain more stable performance estimates, and SMOTE was applied only to the training split in the experimental-design description. Nevertheless, the findings should be considered exploratory and require confirmation in larger, independent, multicenter clinical datasets before real-world healthcare deployment.

4. Conclusion

This research evaluates an interpretable ensemble-learning framework for predicting hepatitis B patient outcomes. Individual ML models showed different predictive characteristics, and SVM and RF provided useful performance while still facing challenges associated with minority-class outcomes. Voting, bagging, boosting, and stacking were evaluated within the $2+n$ framework, demonstrating that ensemble performance depends on both the selected base models and the aggregation technique. The five-model soft-voting ensemble achieved the strongest reported overall performance, while stacking with LR as the meta-learner provided a comparatively balanced alternative. SHAP and LIME were used to explain model predictions and improve transparency. Overall, the results support further investigation of ensemble learning for hepatitis B outcome prediction, with external validation on larger clinical datasets required before clinical deployment.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Funding

The authors received no specific funding from any public, commercial, or not-for-profit organisation for this research.

References

[1] M. Z. Hossain, M. M. Khan, I. Ahmad, F. Yasmin & M. S. Rahman, "Transparent and explainable AI for liver disease diagnosis using SHAP and LIME", 2025 4th International Conference on Electronics Representation and Algorithm (ICERA), Yogyakarta, Indonesia, 2025, pp. 599–604. https://doi.org/10.1109/ICERA66156.2025.11087306.

[2] T. G. Dietterich, "Ensemble methods in machine learning", in *Multiple Classifier Systems*, J. Kittler & F. Roli (Eds.), Springer, Berlin, Heidelberg, 2000, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.

[3] L. Breiman, "Bagging predictors", Machine Learning **24** (1996) 123. https://doi.org/10.1007/BF00058655.

[4] Y. Freund & R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting", Journal of Computer and System Sciences **55** (1997) 119. https://doi.org/10.1006/jcss.1997.1504.

[5] B. Das, R. Sutradhar & D. C. Dalal, "Exploration of hepatitis B virus infection dynamics through virology-informed neural network: A novel artificial intelligence approach", arXiv (2025), arXiv:2503.10708. https://doi.org/10.48550/arXiv.2503.10708.

[6] D. N. Tien, H. T. T. Bui, T. H. T. Ngoc, T. T. Pham, D. T. Nguyen, H. N. T. T. Thu, T. T. H. Vu, T. L. A. Luong, L. T. Hoang, H. C. Tu, N. Körber, T. Bauer & L. K. Ho, "A data-driven approach to assessing hepatitis B mother-to-child transmission risk prediction model: Machine learning perspective", JMIR Formative Research **9** (2025) e69838. https://doi.org/10.2196/69838.

[7] A. Ali, A. Ashraf & K. Rahouma, *Exploring explainable machine learning in early liver disease detection: Insights from fatty liver and hepatitis B*, 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2024, pp. 5–10. https://doi.org/10.1109/NILES63360.2024.10753181.

[8] B. I. Ajuwon, A. Richardson, K. Roper & B. A. Lidbury, "Clinical validity of a machine learning decision support system for early detection of hepatitis B virus: A binational external validation study", Viruses **15** (2023) 1735. https://doi.org/10.3390/v15081735.

[9] A. B. Ahosan, F. Islam, K. M. M. Uddin, N. Hasan & M. A. Uddin, "Improving hepatitis B outcome prediction with ensemble machine learning: A study on predictive models and interpretability", Digital Health **11** (2025) 20552076251350755. https://doi.org/10.1177/20552076251350755.

[10] B. Adepu & T. Archana, "Multi-model stack ensemble deep learning approach for multi-disease prediction in healthcare application", Journal of Mechanics of Continua and Mathematical Sciences **20** (2025) 135. https://doi.org/10.26782/jmcms.2025.03.00009.

[11] G. Obaido, B. Ogbuokiri, T. G. Swart, N. Ayawei, S. M. Kasongo, K. Aruleba, I. D. Mienye, I. Aruleba, W. Chukwu, F. Osaye, O. F. Egbelowo, S. Simelane & E. Esenogho, "An interpretable machine learning approach for hepatitis B diagnosis", Applied Sciences **12** (2022) 11127. https://doi.org/10.3390/app122111127.

[12] M. Li, S. Han, F. Liang, C. Hu, B. Zhang, Q. Hou & S. Zhao, "Machine learning for predicting risk and prognosis of acute kidney disease in critically ill elderly patients during hospitalization: Internet-based and interpretable model study", Journal of Medical Internet Research **26** (2024) e51354. https://doi.org/10.2196/51354.

[13] V. Harabor, R. Mogoş, A. Nechita, A.-M. Adam, G. Adam, A.-S. Melinte-Popescu, M. Melinte-Popescu, M. Stuparu-Creţu, I.-A. Vasilache, E. Mihălceanu, A. Cărăuleanu, A. R. Bivoleanu & A. Harabor, "Machine learning approaches for the prediction of hepatitis B and C seropositivity", International Journal of Environmental Research and Public Health **20** (2023) 2380. https://doi.org/10.3390/ijerph20032380.

[14] H. Desalegn, X. Yang, Y.-S. Yen, N. Berhe, B. Kenney, G. H. Siwo, W. Tang, J. Zhu, A. K. Waljee & A. Johannessen, "Machine-learning methodologies to predict disease progression in chronic hepatitis B in Africa", Hepatology Communications **8** (2024) e0584. https://doi.org/10.1097/HC9.0000000000000584.

[15] N. Yao, Y. Liu, J. Xu, Q. Wang, Q. Zhou, Y. Wang, D. Yi & Y. Wu, "Identification of associated risk factors for serological distribution of hepatitis B virus via machine learning models", BMC Infectious Diseases **24** (2024) 66. https://doi.org/10.1186/s12879-023-08911-8.

[16] D. Shen, L. Sha, L. Yang & X. Gu, "Identification of multiple complications as independent risk factors associated with 1-, 3-, and 5-year mortality in hepatitis B-associated cirrhosis patients", BMC Infectious Diseases **25** (2025) 151. https://doi.org/10.1186/s12879-025-10566-6.

[17] P. Pavani, P. Sravanthi, C. Suneetha, M. Azhar, L. H. Supriya Reddy & T. Maddineni, *Advanced machine learning approaches for liver disease prediction: A focus on feature selection and hybrid ensemble model*, 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2025, pp. 1003–1010. https://doi.org/10.1109/ICMCSI64620.2025.10883633.

[18] G. Rekha, M. H. Teja, S. Priyadarshani, K. S. Babu, R. Brindha & R. K. Pongiannan, *Optimized machine learning framework for liver disease prediction using ensemble learning and feature selection*, 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS), Prawet, Thailand, 2025, pp. 476–481. https://doi.org/10.1109/ICMLAS64557.2025.10968566.

[19] I. Arupzhanov, D. Syssoyev, A. Alimbayev, G. Zhakhina, Y. Sakko, S. Yerdessov, A. Zollanvari & A. Gaipov, "One-year mortality prediction of patients with hepatitis in Kazakhstan based on administrative health data: A machine learning approach", Electronic Journal of General Medicine **21** (2024) em618. https://doi.org/10.29333/ejgm/15747.

[20] S. Dhulipala, B. Maram, R. R. R. Rao, M. Naved, J. K. S. Seles & M. S. Nidhya, *A comparative analysis of deep learning and ensemble methods for time series forecasting*, 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2025, pp. 1713–1716. https://doi.org/10.1109/ICMCSI64620.2025.10883177.

[21] S. Agrawal, M. I. Khan, H. K. Pushpat & S. K. Jain, *Optimized stacked classifier to predict hepatitis B disease*, 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT), 2024, pp. 383–386. https://doi.org/10.1109/IC-SIT63503.2024.10862853.

[22] X. Bai, C. Pu, W. Zhen, Y. Huang, Q. Zhang, Z. Li, Y. Zhang, R. Xu, Z. Yao, W. Wu, M. Sun & X. Li, "Identifying liver cirrhosis in patients with chronic hepatitis B: An interpretable machine learning algorithm based on LSM", Annals of Medicine **57** (2025) 2477294. https://doi.org/10.1080/07853890.2025.2477294.

[23] M. Butkevych, S. Yakovlev & D. Chumachenko, "Data-driven forecasting of acute and chronic hepatitis B in Ukraine with recurrent neural networks", Applied Sciences **15** (2025) 7573. https://doi.org/10.3390/app15137573.

[24] V. C. Chiawa & K. I. Chibueze, "Predictive modeling of hepatitis using machine learning algorithms and mathematical formulations", International Journal of Development Mathematics **2** (2025) 278. https://doi.org/10.62054/ijdm/0202.16.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research **16** (2002) 321. https://doi.org/10.1613/jair.953.

[26] S. M. Lundberg & S.-I. Lee, "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems **30** (2017) 4765. https://doi.org/10.48550/arXiv.1705.07874.

[27] M. T. Ribeiro, S. Singh & C. Guestrin, *Why should I trust you? Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.