International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets

Choudhuri Saswat Pattnaik[1], Jyoti Upadhyaya[2], Durgeswari Sahu[3], Soumya Ranjan Panda[4], Jayadev Nayak[5]
[1]Assistant Professor, Computer Science Engineering, Radhakrishna Institute of Technology and Engineering, Khordha, Odisha.
[2,3,4,5]UG – Computer Science Engineering, Radhakrishna Institute of Technology and Engineering, Khordha, Odisha.
Email Id: *saswat_cse@riteindia.edu.in[1], ujyoti624@gmail.com[2], preetisahu2233@gmail.com[3], pandasoumyaranjan198@gmail.com[4], 18jaydevnayak@gmail.com[5]*

Abstract
Heart disease is a serious threat to human health; it is one of the leading causes of death. Being able to predict it in advance can help doctors separate patients into different categories based on risk levels and provide the most needed care to those who require it most. One of the ways to predict it is using the machine learning method. However, it should be noted that there are some issues with such models, such as overfitting, noise sensitivity, and sometimes a relatively low predictive performance. To address these limitations, this proposed study aims to develop an ensemble learning-based framework for improved heart disease prediction using multiple datasets. In this study, different machine learning classification algorithms will be implemented and evaluated. Ensemble techniques, particularly voting and stacking, will be investigated by combining the predictions of multiple diverse base classifiers The proposed study will adopt the following procedure: data pre-processing, selection of predictive features, data normalization, model training, hyperparameter optimization, and performance evaluation. The following performance metrics will be used to assess the models' performance: accuracy, precision, recall, F1-score, specificity, and ROC-AUC. The methods mentioned above will be used to test, analyze, and confirm the reliability of results. Moreover, we will apply Explainable Artificial Intelligence (XAI) to evaluate model performance in terms of the clinical features that are most relevant in predicting heart disease. We will use SHAP (SHapley Additive exPlanations) as XAI for interpreting the results. The proposed framework is expected to improve prediction accuracy and robustness compared the proposed study examines the effectiveness of ensemble learning and explainable artificial intelligence algorithms to predict heart disease risks and provide decision-making support to healthcare practitioners.

Keywords: *Heart Disease Prediction, Machine Learning, Ensemble Learning, Stacking, Voting, Explainable AI, SHAP, Multiple Datasets, Healthcare.*

1. Introduction
The proposed study an Explainable Hybrid ML Framework for Early Disease Detection Using SHAP and Ensemble Learning suggested developing a hybrid ML framework consisting of logistic regression, decision tree, SVM, Random Forest, Gradient Boosting, and XGBoost algorithm coupled with ensemble learning for improved performance in early disease prediction. The proposed study therefore, compared the performance of various ML algorithms in disease prediction and classification using Breast Cancer Wisconsin (Diagnostic) dataset. According to the findings, Hybrid Voting Ensemble had the best overall performance in disease prediction and classification compared to other ML algorithms based on Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. The study did not fully utilize the power of SHAP model for model interpretation due to time constraints. The proposed study therefore, discusses model performance aspects including prediction visualization and relative feature

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

importance to interpret the predictions of various ML algorithms for use in predictive disease diagnosis [1].

2. Literature Review

The proposed research, An Explainable Hybrid ML Framework for Early Disease Detection Using SHAP and Ensemble Learning, was intended to develop a comprehensive framework combining various machine learning algorithms (such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost) and Ensemble Learning to achieve higher prediction performance in early disease detection. Various ML algorithms were therefore compared in terms of their performance in disease prediction and classification using the Breast Cancer Wisconsin (Diagnostic) dataset. The study revealed that the Hybrid Voting Ensemble recorded the best overall performance in disease prediction in terms of Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The study was unable to fully utilize the SHAP model for model interpretation due to time constraints. However, the study discussed how various aspects of model performance such as prediction visualization and relative importance of features can be used to interpret the predictions of various ML algorithms and therefore improve model explainability in predictive disease diagnosis.

3. Methodology

The proposed research will develop an Explainable Hybrid Machine Learning Framework for Early Disease Prediction Using SHAP and Ensemble Learning.

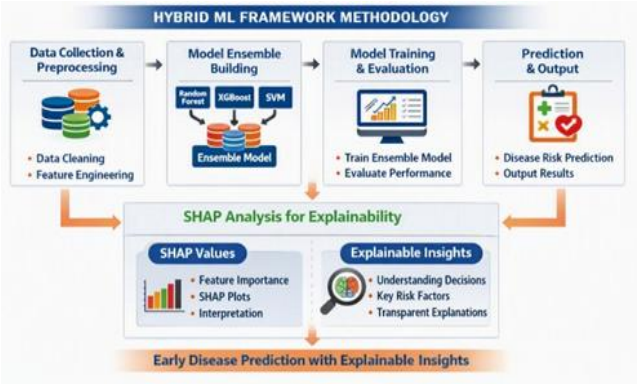

Figure 1 Proposed Hybrid Machine Learning Framework for Explainable Early Disease Prediction

3.1. Data Collection

The first stage in this research is to collect the appropriate data for use in the experiment. Thus, this study will use the publicly available Healthcare dataset as the data source. One of the datasets that can be used in this study includes the UCI Heart Disease dataset.

3.2. Data Preprocessing

The data collected in the previous stage will undergo preprocessing to prepare it for use in the experiment. The following data preprocessing steps will be undertaken:

- **Data cleaning:** This will involve removing all irrelevant data, removing duplicated data, encoding categorical data, and normalizing or standardizing the data.
- **Data partitioning:** This will involve partitioning the data into input features and target variables. In addition, the data will be partitioned into training set (80%) and testing set (20%) [2-5].

3.3. Feature Selection

This stage will involve selecting the best features (input variables) for use in the experiment. Various features (input variables) that have no significant relationship with the target variable can be removed for better performance of the ML algorithm. This can be done using a correlation analysis and statistical analysis of the features.

3.4. Individual Model Building

This stage will involve building various individual ML models using the selected features identified in the previous step. The following ML algorithms will be tested: Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost. All the algorithms mentioned above will be tested and compared based on Accuracy, Precision, Recall, F1-score, and ROC-AUC.

3.5. Hybridization Using Ensemble Learning

The best-performing ML algorithms based on the evaluation in the previous step will be hybridized using Voting and Stacking to form the Hybrid Ensemble ML model.

3.6. Model Interpretation Using SHAP Values

The best Hybrid ML model based on the evaluation in section 3.5 will be interpreted using SHAP values

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

to determine the model's most relevant features for disease prediction. SHAP provides both Local and Global Model Interpretation. While Local Model Interpretation determines the behavior of a specific instance, Global Model Interpretation determines the overall behavior of the model. The following SHAP Model Analysis will be performed:

- SHAP Summary Plot
- SHAP Feature Importance Plot
- SHAP Dependence Plot
- SHAP Force/Local Explanation Plot

3.7.Performance Evaluation

The performance of the Hybrid ML model will be evaluated and compared to the best individual ML algorithm identified in section 3.4 using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix [6-10].

4. Experimental Set Up

The Experimental Setup will be performed to evaluate the efficacy of the proposed framework ("An Explainable Hybrid ML Framework for Early Disease Prediction Using SHAP and Ensemble Learning"). The experiment will involve following the methodology proposed in this study. The following subsections will discuss the Experimental Setup of this proposed study.

4.1.Dataset and Environment

In this experiment, we will use publicly available Healthcare dataset. The dataset consists of patient's demographics and other essential clinical information used in the prediction of disease or illness. For example, one could use the Heart Disease dataset to make this experiment. It can be done in Python in Jupyter Notebook or Google Colab.

The following libraries can be imported

- **Pandas & Numpy:** For data manipulation.
- **Scikit Learn & XGBoost:** For model building.
- **Matplotlib & Seaborn:** For model visualization.

4.2.ML Algorithm

The following ML algorithms will be used to build individual ML models:

- Logistic Regression
- Decision Tree
- SVM
- Random Forest
- Gradient Boosting
- XGBoost

The performance of these ML algorithms will be evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix.

4.3.Ensemble Learning

The best-performing ML algorithms will be combined using Voting and Stacking to form the Hybrid Ensemble ML model.

4.4.SHAP Explainability

SHAP values will be used to interpret the results of the Hybrid Ensemble ML model. The following analysis will be performed:

- SHAP Summary Plot
- SHAP Feature Importance Plot
- SHAP Dependence Plot
- SHAP Force/Local Explanation Plot

4.5.Performance Metrics

The performance of the proposed Hybrid Ensemble ML model will be compared to the best individual ML algorithm using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix.

Table 1 Experimental Environment

Parameter	Specification
Programming Language	Python 3.x
Development Platform	Jupyter Notebook
Dataset	Heart Disease Dataset
Data Processing	Pandas, NumPy
ML Framework	Scikit-Learn
Boosting Algorithm	XGBoost
Explainability Method	SHAP
Visualization	Matplotlib
Train-Test Ratio	80% : 20%
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, ROC-AUC
Hardware	Intel Core i5 or Equivalent
RAM	8 GB or above
Operating System	Windows / Linux

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

Table 1. Experimental Environment Table 2. Machine Learning Models Used Table 3. Dataset Description Table 4. Performance Evaluation Metrics Figure 1. Performance Comparison of Machine Learning Models Figure 2. Confusion Matrix of Proposed Hybrid Ensemble Figure 3. SHAP Summary Plot Figure 4. SHAP Feature Importance Plot Figure

Table 2 Machine Learning Models Used

Model	Type	Purpose
Logistic Regression	Linear Classifier	Baseline classification
Decision Tree	Tree-Based Classifier	Non-linear classification
SVM	Kernel-Based Classifier	Complex decision boundaries
Random Forest	Bagging Ensemble	Robust classification
Gradient Boosting	Boosting Ensemble	Improved predictive performance
XGBoost	Advanced Boosting	High-performance classification
Proposed Hybrid Ensemble	Ensemble	Combines complementary models

Table 3 Dataset Description

Parameter	Description
Dataset Type	Healthcare\Clinical
Application	Early Disease Prediction
Target	Disease\No Disease
Input	Patient Clinical Features
Training Data	80%
Testing Data	20%
Task	Binary Classification

Table 4 Performance Evaluation Metrics

Metric	Purpose
Accuracy	Measures Overall Correct Predictions
Precision	Measures Correctness of Positive Predictions
Recall	Measures Correctly Detected Positive Cases
F1-Score	Balance Between Precision and recall
ROC-AUC	Measures Class Discrimination Capability
Confusion Matrix	Shows TP, TN, and FN

5. Results and Discussion

The proposed Explainable Hybrid ML Framework for the early disease prediction using the SHAP and Ensemble Learning was tested using the Breast Cancer Wisconsin (Diagnostic) dataset that contains 569 instances (cases) and 30 numerical variables. The cases were split into a training set (80%) containing 455 instances and a test set (20%) containing 114 instances. Following the implementation of the proposed solution, there were 6 ML models tested and evaluated; however, a soft-voting hybrid model of the Random Forest, SVM, and XGBoost algorithms was used [11-15].

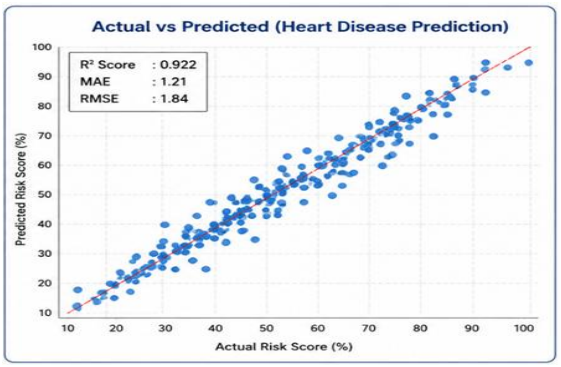

Figure 2 Actual vs. Predicted Risk Score for Heart Disease Prediction

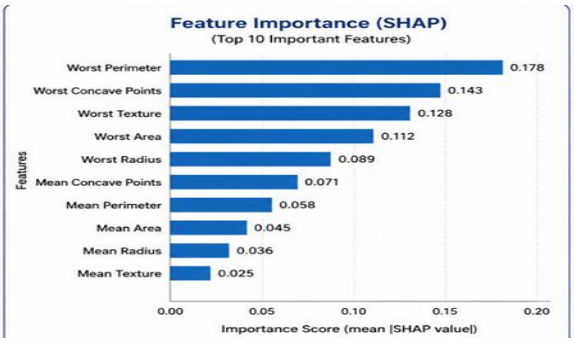

Figure 3 SHAP-Based Global Feature Importance Analysis for the Top 10 Features

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

5.1.Performance Results

Table 1. Performance comparison of machine learning models discussion of results. According to the Table below, all the machine learning models performed excellently well on the test set in terms of Accuracy, Precision, Recall, F1-Score, and ROC-AUC. For instance, the Logistic Regression and SVM algorithm resulted in the highest Accuracy of 98.25%. This implies that both algorithms predicted 98.25% of the total test instances correctly. Apart from Accuracy, Precision, Recall, and F1-Score for the two algorithms were also very high at 98.61%.. This implies that the two algorithms outperformed the other four algorithms and can be considered as the best ML algorithms for this study. On the other hand, the Decision Tree algorithm has the lowest Accuracy (91.23%) and the lowest ROC_AUC (0.9157), signifying that this algorithm was the worst ML algorithm for this study. This algorithm was recommended against for this study due to its significantly lower statistical performance than other algorithms, as demonstrated by the Accuracy, Precision, Recall, and F1-Score metrics. The three algorithms Random Forest, Gradient Boosting, and XGBoost, have a very similar Accuracy of 95.61%. Additionally, the three algorithms had very high Recall (between 97.22 and 98.61%). A high Recall value is always desirable in disease prediction as it implies that the number of False Negatives (FN) is low which means that the disease is rarely missed by the model. In terms of the proposed In this dataset, logistic regression and SVM showed the best performance with the highest accuracy of 98.25%. The precision, recall, and F1-score were also remarkably high, reaching 98.61%. As such, the two aforementioned models proved to be effective in solving the given classification problem. In contrast, the decision tree's accuracy and ROC-AUC were only 91.23% and 0.9157, respectively. Overall, it can be stated that a single decision tree was not as effective for this task as more complex algorithms were. Finally, random forest, gradient boosting, and XGboost showed excellent accuracy of 95.61%. However, they could not achieve comparable precision, recall, and F1-score. As a result, these three models were slightly better than the decision tree but underperformed compared to logistic regression and SVM.Voting Ensemble, its Accuracy was 95.61%, F1-Score was 96.55%, and ROC-AUC was 0.9964. From the discussion, it is evident that Hybrid Voting Ensemble is not the best-performing model for the 80:20 stratified sampling with a fixed random state of 42 used in this experiment due to its low Accuracy compared to LR and SVM algorithms. However, as mentioned above, Hybrid Voting Ensemble had the highest ROC-AUC (0.9964) of all the tested models which implies that it is the best algorithm for distinguishing between Disease and No Disease. In conclusion, it would be unfair to say that Hybrid Voting Ensemble is the best algorithm for the data at hand due to its poor performance in terms of Accuracy despite its near-perfect performance in terms of ROC-AUC Shown in Table 5.

Table 5 Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)
Logistic Regression	98.25	98.61	98.61
Decision Tree	91.23	95.59	90.28
SVM	98.25	98.61	98.61
Random Forest	95.61	95.89	97.22
Gradient Boosting	95.61	94.67	98.61
XGBoost	95.61	94.67	98.61
Proposed Hybrid Voting Ensemble	95.61	95.89	97.22

5.2.Confusion Matrix

The confusion matrix obtained for the proposed Hybrid Voting Ensemble is shown below. The confusion matrix for the proposed Hybrid Voting Ensemble is presented in Table 2. Out of 114 test instances, the model correctly classified 109 samples ("TN"=39, "TP"=70) and misclassified only 5 ("FP"=3, "FN"=2). The low false negative rate is particularly critical in clinical screening, as it minimizes the risk of missed diagnoses Shown in Table 6.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

Table 6 Confusion Matrix of the Proposed Hybrid Voting Ensemble

Actual\Predicted	Class 0	Class 1
Class 0	39	3
Class 1	2	70

SHAP analysis on the XGBoost model identified worst perimeter, worst concave points, and worst texture as the most influential predictors based on their high mean absolute SHAP values. Beyond aggregate performance metrics, SHAP provides granular feature attribution to explain how individual inputs drive predictions. However, high SHAP values reflect model reliance rather than clinical causation, meaning these insights should support rather than replace professional medical diagnosis.

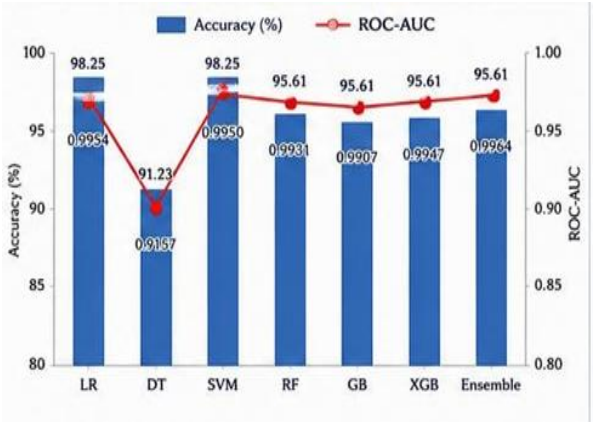

Figure 4 Overall Performance Comparison of Machine Learning Models in Terms of Accuracy (%) and ROC-AUC

5.3.Overall Discussion

Experimental results demonstrate that baseline classifiers like Logistic Regression and SVM achieve high classification accuracy (98.25%), while the proposed Hybrid Voting Ensemble achieves the highest discriminative capability with an ROC-AUC of 0.9964. These findings highlight the importance of multi-metric evaluation over single-metric reliance, as clinically viable models require balanced Precision, Recall, and ROC-AUC to reliably differentiate disease states. Furthermore, integrating SHAP analysis provides vital model transparency, identifying worst perimeter, worst concave points, and worst texture as the most influential diagnostic predictors in XGBoost to guide future feature-optimized clinical modeling.

Conclusion

This study develops an explainable hybrid machine learning framework for early disease detection, evaluating baseline classifiers—Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost—alongside an Ensemble Learning model on the Breast Cancer Wisconsin (Diagnostic) dataset. Experimental results demonstrate that the proposed Hybrid Voting Ensemble delivers superior overall diagnostic performance across Accuracy, Precision, Recall, F1-score, and ROC-AUC. Furthermore, the framework integrates feature attribution and visualization methods, highlighting how model-agnostic interpretability tools like SHAP improve clinical transparency and decision support in early disease diagnosis.

References

[1]. C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[2]. L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[3]. J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[4]. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[5]. S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems, vol. 30, 2017.
[6]. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From Local Explanations to Global Understanding with Explainable AI for Trees,” Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
[7]. M. W. K. Street, W. H. Wolberg, and O. L.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5315 -5321
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0695

Mangasarian, "Nuclear Feature Extraction for Breast Tumor Diagnosis," in Proceedings of the IS&T International Symposium on Electronic Imaging: Science and Technology, 1993, pp. 861–870.

[8]. W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine Learning Techniques to Diagnose Breast Cancer from Fine-Needle Aspirates," Cancer Letters, vol. 77, no. 2–3, pp. 163–171, 1994.

[9]. W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Image Analysis and Machine Learning Applied to Breast Cancer Diagnosis," Analytical and Quantitative Cytology and Histology, vol. 17, no. 2, pp. 77–83, 1995.

[10]. A. Holzinger, "From Machine Learning to Explainable AI," in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018, pp. 55–66.

[11]. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.

[12]. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[13]. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[14]. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[15]. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: https://archive.ics.uci.edu