

Prediction of Heart Disease Using Ensemble Machine Learning Techniques

Mr. Madhav Sharma¹

¹ Assistant Professor, Department of Computer Science and Applications, Mandsaur(M.P.) 458001, India

ABSTRACT: In this current world, keep hearing about heart disease problems every day and about the deaths due to them, making heart disease a major contributor to the crucial mortality rate worldwide. According to the World Health Organization (WHO), an estimated 17.9 million individuals die from cardiovascular diseases (CVDs) each year. The identification of cardiovascular disease states, including cardiac arrhythmia and coronary heart disease, based on traditional clinical data analysis is still a big challenge. The early diagnosis of cardiac disease can enable timely medical treatment and save many lives. The use of machine learning (ML) algorithms enables intelligent decision-making and accurate disease prediction by identifying complex patterns in healthcare data. This study adopted the following preprocessing methods for the UCI Heart Disease dataset: missing-value treatment, duplicate removal, noise reduction, one-hot encoding, Z-score normalization, and SMOTE data balancing. The proposed XGBoost model was developed for heart disease risk assessment and evaluated using accuracy, precision, recall, and F1-score. The proposed model achieved 99.8% accuracy, 99.7% precision, 99.9% recall, and 99.6% F1-score, demonstrating its effectiveness and reliability for accurate heart disease prediction.

KEYWORDS: Explainable Artificial Intelligence (XAI), Heart Disease, Ensemble Learning, Machine Learning (ML), Cardiovascular Risk Assessment, Clinical Decision Support, Healthcare Analytics.

1. Introduction

Healthcare is one of the most important sectors for maintaining global health and improving quality of life. The continuous advancement of medical technologies has enhanced disease diagnosis, prediction, treatment, and prevention [1][2]. Cardiovascular diseases (CVDs) are among the major global health concerns, affecting the heart and circulatory system [3]. Common risk factors include high blood pressure, high cholesterol, smoking, physical inactivity, and unhealthy lifestyles. Early identification of individuals at risk is essential for timely intervention and effective disease management [4].

Heart disease is a serious medical condition because abnormalities in the heart and blood vessels can affect normal blood circulation [5]. Symptoms such as shortness of breath, fatigue, weakness, and swelling of the feet may indicate cardiovascular problems. Coronary artery disease is one of the most common forms of heart disease and remains a major cause of morbidity and mortality worldwide [6][7]. Conventional diagnostic procedures may require considerable time, cost, specialized equipment, and

medical expertise. Therefore, reliable and efficient prediction techniques are increasingly required to support early detection and clinical decision-making [8][9][10].

The increasing availability of electronic health records and clinical datasets has created opportunities for applying computational intelligence in healthcare [11][12][13]. Artificial Intelligence (AI) can analyze large volumes of medical data and identify patterns that support disease diagnosis and risk assessment [14]. Machine Learning (ML) algorithms learn from historical patient data to predict disease outcomes and have been widely applied to heart disease prediction. These approaches can provide rapid and cost-effective risk assessment while supporting healthcare professionals in clinical decision-making [15][16][17][18].

In addition to ML, DL techniques have gained attention for their ability to automatically learn complex patterns from large, high-dimensional datasets. Ensemble Machine Learning further improves prediction by combining multiple learning models to enhance robustness and generalization [19][20][21]. Therefore, this study, titled "Prediction of Heart Disease Using

Ensemble Machine Learning Techniques,” focuses on developing an effective ensemble-based framework for accurate heart disease risk prediction [22]. The proposed approach aims to provide reliable predictions and support the early identification of patients at potential risk.

A. Motivation and Contribution

This study is motivated by the growing burden of heart disease in the world, its contribution to death, and its impact on the quality of life. Early, accurate identification of individuals at risk may enable timely treatment and/or preventive interventions. Traditional diagnosis can be time-consuming and requires a certain level of clinical experience. Therefore, this study aims to develop an effective ensemble machine learning approach for reliable and efficient heart disease risk prediction. The proposed approach can also assist healthcare professionals in making faster and data-driven clinical decisions. This research offers several key contributions as listed below:

- Developed a machine learning-based heart disease risk assessment framework using the UCI Heart Disease dataset.
- Used SMOTE to address class imbalance and improve the representation of minority heart disease classes.
- Implemented the proposed XGBoost model for accurate classification of heart disease risk.
- Compared the proposed XGBoost model with LightGBM, KNN, Logistic Regression, and CatBoost, showing its superior classification performance.
- Evaluated the proposed model using accuracy, precision, recall, and F1-score.
- Demonstrated the potential of XGBoost as an effective decision-support approach for early heart disease risk assessment.

The proposed study is justified by the rising prevalence of heart disease and the need for accurate, early risk assessment using clinical data. The traditional method can be time-consuming and heavily relies on clinical experience, which could motivate the use of machine learning for automated prediction. The novelty of this work lies in the application of systematic preprocessing, one-hot encoding, Z-score normalization, class balancing via SMOTE and risk assessment of heart disease using XGBoost. The proposed method can identify complex patterns and relationships among clinical features to improve disease classification. This framework provides a reliable and efficient machine learning-based approach that can

support early heart disease risk assessment and data-driven clinical decision-making.

B. Organization of the Paper

The paper is organized as follows: Section II provides an overview of the relevant literature. Details regarding the proposed approach and the dataset are provided in Section III, and the outcomes and comparisons are provided in Section IV in the comparative part. Finally, Section V concludes this work and provides a detailed description of future research.

2. Literature Review

A detailed examination of existing research on heart disease prediction and risk assessment was carried out to identify significant findings, methodological approaches, and research gaps relevant to the present study.

Shah and Poonia (2025) proposed that the traditional approach to prediction of heart disease using a limited set of essential physical signs from repetitive inspections is inadequate. Also, manual inspection and analysis of large data sets is time-consuming and labor-intensive. Heart disease is a major global health issue marked by high rates of death. Traditional predictive methods often fall short, making machine learning solutions more effective. This paper applies algorithms such as LR, DT, KNN, SVM, RF, naïve bias, and bagging classifier to enhance diagnostic efficiency for heart disease. The accuracy of ensemble technique, namely bagging classifier is 89% [23].

Kumar et al. (2025) study explores ML algorithms for accurately classifying patients with heart disease using a stacked ensemble model combining XGBoost, Random Forest, SVM, and Gradient Boosting. Data preprocessing included missing-value handling, feature normalization, and class balancing. Models were evaluated using standard classification metrics. Random Forest performed best individually, while the ensemble achieved a ROC-AUC of 89.56% and accuracy of 89.78%. Feature analysis identified age, ST depression, cholesterol, maximum heart rate, and exercise-induced angina as key predictors, with ROC and precision-recall curves validating model robustness [24].

Banerjee et al. (2025) introduced the Cascading Heart Disease Identifier (CHDI) approach was introduced as an advancement in predictive modeling, using ensemble learning to combine multiple classifiers and leverage preceding outputs for subsequent predictions. Various ML techniques, including voting ensemble classifiers, were used to evaluate health indicators for cardiac disease prediction. The CHDI model achieved a high accuracy of 93.4%, demonstrating its potential for early disease detection and improved healthcare outcomes [25].

Purwanto et al. (2024) primary aim of this research was to improve the accuracy and efficiency of early heart disease detection and treatment. Six ML models—Discriminant Analysis, Naïve Bayes, SVM, KNN, Ensemble, and NN—were fine-tuned using hyperparameter tuning on the Kaggle Heart Disease dataset. The dataset was evaluated using 90-10, 85-15, 80-20, 75-25, and 70-30 train-test splits. Optimized SVM outperformed other models, achieving the highest accuracy of 95% with the 85-15 split [26].

Rustamov et al. (2023) proposed a stacking ensemble classifier, which combines several single classifiers to produce an optimal predictive model. The Framingham Heart Study dataset was used to train the machine learning algorithms. The exploratory data analysis indicates that CVD was more common in males and diabetic individuals. Furthermore, individuals above the age of 65 were more susceptible to CVDs. Feature selection, missing value imputation, and data sampling were performed as part of data preprocessing. The results show that the proposed stacked ensemble classifier achieved 88.33% accuracy, 89.95% precision,

86.27% recall and 88.07% F1-score [27].

Cyriac et al. (2022) traditional heart disease prediction and diagnosis require costly medical expertise, which is limited in developing countries. Machine learning provides quick, inexpensive, and noninvasive alternatives. This work used seven ML models with Soft-Voting and Hard-Voting ensembles to improve prediction accuracy. The five-model ensemble combinations achieved the highest test accuracy of 94.2% [28].

Ambrews et al. (2022) proposed method was evaluated using UCI Heart Disease and Framingham datasets with KNN, Naïve Bayes, Decision Tree, SVM, LR, and RF classifiers. SVM achieved the highest UCI accuracy (91.09%), while LR performed best on Framingham (85.38%). RF achieved the highest sensitivity (92.52%) on UCI, while the Voting approach performed best overall, achieving 91.96% accuracy, 91.69% F1, 91.72% sensitivity, 90.77% specificity, and 92.40% precision [29].

Table I summarizes recent studies related to heart disease risk prediction, highlighting the methodologies employed, datasets analyzed, key performance outcomes, and identified limitations and future research directions.

Author	Proposed Work	Dataset	Key Findings	Limitations & Future Work
Shah and Poonia (2025)	Applied LR, DT, KNN, SVM, RF, Naïve Bayes, and Bagging Classifier for heart disease prediction.	Heart Disease Dataset	Bagging achieved 89% accuracy, demonstrating the benefit of ensemble learning for early heart disease detection.	Performance could be improved using advanced boosting and hybrid ensemble techniques with larger and more diverse datasets.
Kumar et al. (2025)	Developed a stacked ensemble model combining XGBoost, RF, SVM, and Gradient Boosting.	Heart Disease Dataset	The ensemble achieved 89.78% accuracy and 89.56% ROC-AUC, outperforming the individual models.	Further validation on larger clinical datasets and optimization of ensemble parameters could improve generalization.
Banerjee et al. (2025)	Proposed the Cascading Heart Disease Identifier (CHDI) using multiple classifiers and voting-based ensemble learning.	Heart Disease Dataset	The CHDI model achieved 93.4% accuracy, demonstrating the effectiveness of cascading ensemble prediction.	Further testing with diverse datasets and advanced ensemble algorithms is required to improve robustness.
Purwanto et al. (2024)	Six ML models with hyperparameter tuning for early heart disease prediction	Kaggle Heart Disease	Optimized SVM achieved the highest accuracy of 95% with an 85–15 train-test split.	Further validation on diverse datasets and improved model generalization are needed.
Rustamov et al. (2023)	Proposed a stacking ensemble classifier using multiple machine learning models.	Framingham Heart Study Dataset	The proposed ensemble achieved 88.33% accuracy, 89.95% precision, 86.27% recall, and 88.07% F1-score.	Further evaluation with different populations and larger datasets could improve generalizability.

Cyriac et al. (2022)	Combined seven ML models using Soft-Voting and Hard-Voting ensemble approaches.	Heart Disease Dataset	The five-model ensemble combination achieved the highest test accuracy of 94.2%.	Additional advanced ensemble and boosting techniques could be evaluated on larger datasets.
Ambrews et al. (2022)	Evaluated KNN, Naïve Bayes, DT, SVM, LR, RF, and Stacking/Voting ensembles on two datasets.	UCI Heart Disease & Framingham Heart Study	Voting achieved 91.96% accuracy on the UCI dataset, while SVM and LR were the best individual classifiers on UCI and Framingham datasets, respectively.	Further optimization of ensemble combinations and validation across larger datasets is needed.

Research gaps: Existing studies demonstrate the effectiveness of machine learning and ensemble techniques for heart disease prediction; however, several limitations remain. Various methods achieve moderate predictive performance, without necessarily guaranteeing high accuracy on other datasets and patient subgroups. In some studies, only a single algorithm or simple ensemble methods are considered,

without accounting for the need for extensive preprocessing and class-balancing techniques. The applicability of existing approaches is also hindered by differences in datasets, feature selection and preprocessing steps. Therefore, there is a need for a robust and efficient machine learning framework that integrates effective preprocessing and advanced predictive modeling to improve the reliability of heart disease risk assessment.

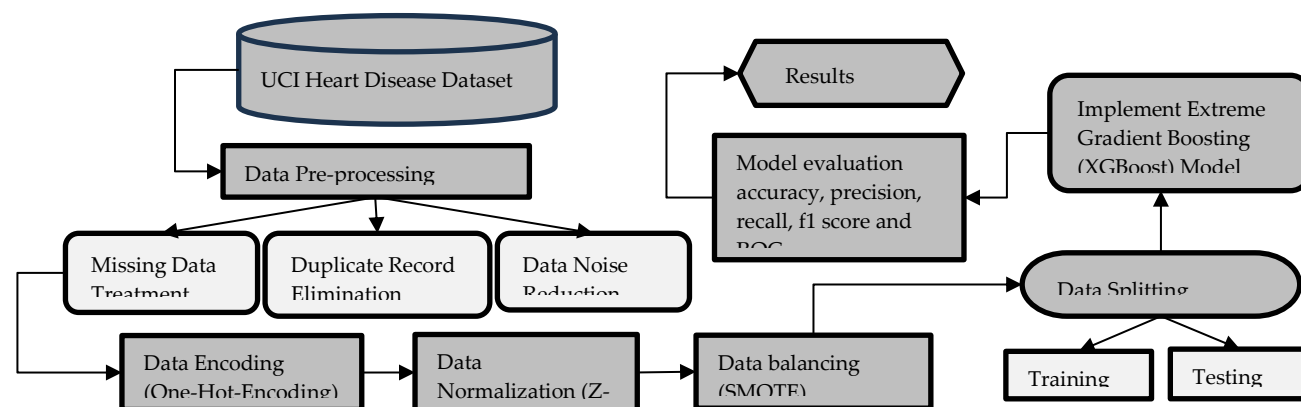


Fig.1. Proposed Workflow for Heart Disease Risk Assessment Using Machine Learning Techniques

assessment is shown in Fig. 1.

3. Methodology

The proposed approach for the Heart Disease Risk Assessment consists of collecting data from the UCI Heart Disease dataset, cleaning the data, removing missing data, removing duplicate data, reducing noise, one-hot encoding, and Z-normalization. The target classes are then balanced using SMOTE and the resulting dataset is split into a training set and a testing set in a 70:30 ratio. Finally, the proposed XGBoost model is trained and evaluated using accuracy, precision, recall, F1-score, and AUC-ROC to assess its classification performance. The proposed workflow for the machine learning-based heart disease risk

The following section provides a detailed description of each stage involved in the proposed methodology:

A. Data Gathering and Analysis

The UCI Heart Disease dataset was used in this study, comprising 920 instances and 16 attributes related to clinical and demographic characteristics. These features were utilized to develop and evaluate the proposed heart disease prediction model. Bar plots and heatmaps were employed to visualize disease distribution and identify relationships and correlations among the dataset attributes:

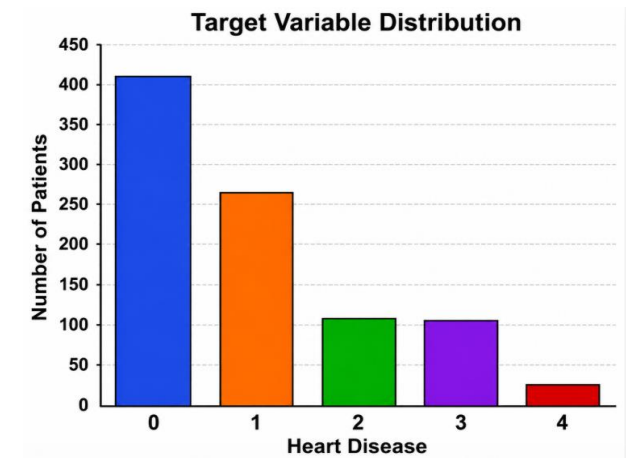


Fig.2. Target Variable Distribution

Fig. 2 shows the distribution of values of heart disease severity for the UCI Heart Disease dataset. The bar plot shows that class 0 has the highest number of patients, followed by classes 1, 2, and 3, while class 4

contains the fewest patients. This distribution indicates an imbalance among the target classes, which should be considered during data preprocessing and model development to ensure reliable heart disease prediction.

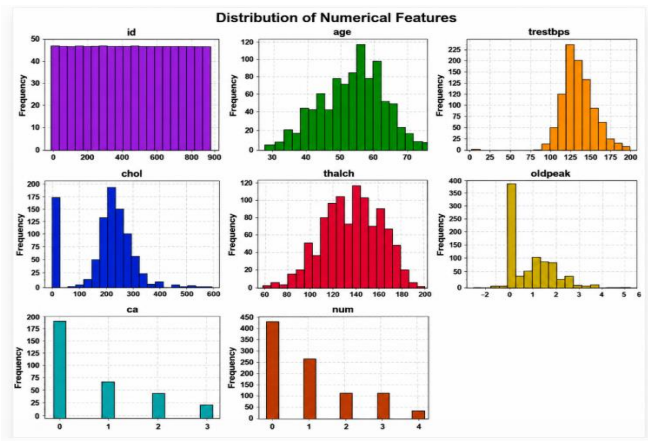


Fig.3. Distribution of Numerical Features

Fig. 3 presents the distribution of the numerical features in the UCI Heart Disease dataset using histograms. The plots illustrate the frequency patterns of age, resting blood pressure (trestbps), cholesterol

(chol), maximum heart rate (thalch), ST depression (oldpeak), number of major vessels (ca), and the target variable (num), highlighting their respective ranges, concentrations, and variations across the patient records.

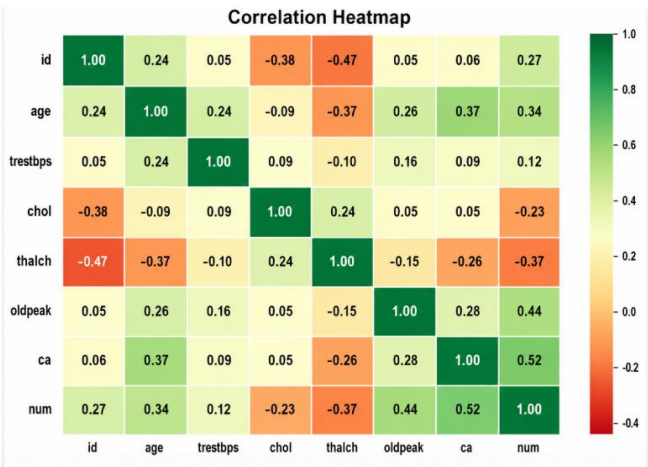


Fig.4. Correlation Matrix Heatmap

Fig. 4 presents the correlation heatmap illustrating the relationships among the numerical features of the UCI Heart Disease dataset. The heatmap shows the

strength and direction of correlations between variables, with positive correlations represented by green shades and negative correlations by red/orange shades, while

the numerical values indicate the corresponding correlation coefficients. The target variable num shows relatively stronger positive correlations with ca (0.52), oldpeak (0.44), and age (0.34), while negative correlations are observed with thalch (-0.37) and chol (-0.23).

B. Data Pre-Processing

The UCI Heart Disease dataset was utilized for data preparation, including data integration and cleansing to ensure consistency and reliability. The preprocessing phase involved handling missing values, removing duplicate and redundant records, reducing noise, and performing data labeling and normalization. The key preprocessing procedures applied to the dataset are summarized as follows:

- **Missing Data Treatment:** Incomplete observations and records with unavailable class labels were identified and removed to maintain the consistency, reliability, and quality of the dataset.
- **Duplicate Record Elimination:** Repeated and redundant observations were detected and excluded from the dataset. This ensured that duplicate information did not bias the analysis or influence the learning process.
- **Data Noise Reduction:** Unwanted, inconsistent, and erroneous values were identified during preprocessing and appropriately removed. This helped reduce data irregularities and provided cleaner input for effective machine learning model training.

C. Data Encoding (One-Hot Encoding)

Data encoding is the process of transforming categorical or non-numerical attributes into numerical representations suitable for machine learning algorithms. In this study, non-numerical attributes were converted into numerical values, with One-Hot Encoding applied to nominal variables that have no inherent order.

D. Data Normalization (Z-score)

To ensure consistency across features and prevent dominance of any single feature, the dataset underwent Z-score standardization. Each feature was rescaled by subtracting its mean and dividing by its Standard Deviation. This process standardized all numerical attributes, improving the model's training stability and convergence. Equation (1): for Z-score standardization:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

- Z = Standardized Value
- X = Original Value

- μ = mean of the feature
- σ = Standard Deviation of the feature

This transformation ensures that the mean of the data is zero and the Standard Deviation is one.

E. Data balancing (SMOTE)

Data balancing is a preprocessing technique used to reduce class imbalance by ensuring a more uniform distribution of samples across different classes. In this study, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to increase the representation of the minority class. SMOTE identifies the nearest minority-class neighbors and generates new synthetic instances by interpolating between existing minority samples. The generated samples are incorporated into the training data, improving class representation and helping the model learn minority-class patterns more effectively.

F. Data Splitting

The dataset was split into training and testing subsets using a 70:30 ratio with stratified sampling. This approach ensured that the relative proportion of each class was maintained consistently across both subsets. Consequently, the training and testing data retained a class distribution representative of the original dataset.

G. Proposed Extreme Gradient Boosting (XGBoost) Model

In this study, a supervised machine learning-based boosting algorithm, Extreme Gradient Boosting (XGBoost), is proposed for heart disease risk assessment. XGBoost is an ensemble learning technique that sequentially constructs multiple decision trees, where each subsequent tree improves the errors made by the preceding trees. The model optimizes a predefined objective function that combines the prediction loss with a regularization term to enhance predictive performance and control model complexity. The mathematical formulation of the proposed XGBoost model is expressed as Equation (2):

$$y = f(x) \quad (2)$$

where y is the predicted property price, x is the vector of input features (such as square footage, number of bedrooms, etc.), and f(x) is the XGBoost model that predicts y based on x. To compute f(x), XGBoost builds an ensemble of decision trees that are trained to minimize the mean squared error (MSE) loss function. The model combines the predictions from multiple decision trees to arrive at a final prediction. The general form of the XGBoost regression model can be expressed as Equation (3):

$$y = \sum_{k=1}^K f_k(x) \quad (3)$$

where $f(x)$ is the prediction of the k th decision tree, and K is the total number of decision trees in the ensemble. The prediction of each tree is a weighted sum of the leaf values of the tree, which are learned during training. The prediction of the XGBoost model for a given input x is obtained by summing the predictions of all the decision trees in the ensemble.

The proposed XGBoost model was configured with 200 estimators, a maximum depth of 6, a learning rate of 0.05, subsample ratio of 0.80, and column sampling ratio of 0.80. The model used $\text{min_child_weight} = 1$, $\text{gamma} = 0.10$, $\text{reg_alpha} = 0.01$, and $\text{reg_lambda} = 1.00$ to control model complexity and reduce overfitting. For multi-class heart disease classification, multi was used as the objective function, with mlogloss as the evaluation metric and $\text{random_state} = 42$ for reproducible results.

H. Evaluation Metrics

The performance metrics used in this study are essential for evaluating the effectiveness and classification capability of the proposed model. The selected metrics include accuracy, precision, recall, F1-score, and AUC-ROC, providing comprehensive insights into model performance. The classification outcomes are represented by true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Here, TP and TN denote correctly predicted positive and negative instances, respectively, while FP and FN represent incorrectly predicted positive and negative instances. The performance metrics are represented in Equations (4) to (7).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

Accuracy represents the ratio of correctly classified instances to the total number of input samples in the dataset. Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, indicating the reliability of positive predictions. Recall refers to the ratio of correctly identified positive instances to all actual positive instances, reflecting the model's ability to detect positive cases. F1-score is the harmonic mean of precision and recall, providing a balanced measure of classification performance with values ranging from 0 to 1. AUC-ROC quantifies the overall performance of a classification model across all possible decision thresholds by representing the trade-off between the true positive rate and false positive rate as the threshold varies from 0 to 1.

4. Results and Discussion

This section describes the experimental setup and highlights the performance of the proposed model across both training and testing phases, demonstrating its evaluation and efficient processing. In this work, the prediction performance of the proposed Extreme Gradient Boosting (XGBoost) model is investigated for heart disease risk assessment using a well-known UCI heart disease dataset. The model was implemented using Python and Jupyter Notebook on an Intel Core i5-8256U CPU with 8 GB RAM. The proposed XGBoost model was evaluated using accuracy, precision, recall, and F1-score during both training and testing phases. As presented in Table I, the model achieved 99.8% accuracy, 99.7% precision, 99.9% recall, and 99.6% F1-score on testing, while achieving 100% across all metrics during training.

TABLE II. Classification Results of the Proposed Model for Heart Disease Risk Assessment

Matrix	Testing	Training
Accuracy	99.8	100
Precision	99.7	100
Recall	99.9	100
F1-score	99.6	100

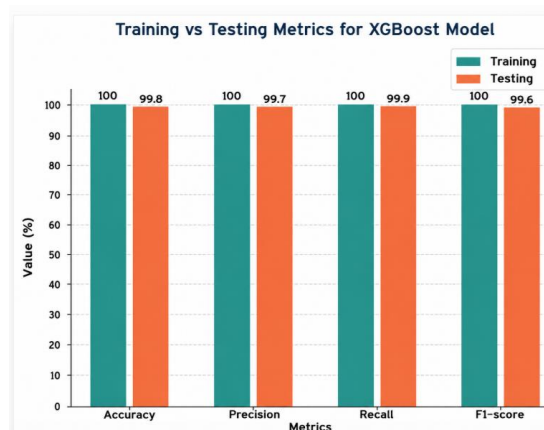


Fig.5. Train and Test performance comparison of proposed model for Heart Disease Risk Assessment

Fig. 5 illustrates the training and testing performance of the proposed XGBoost model for Heart Disease Risk Assessment. The model achieved 100% performance across all evaluation metrics during training. In testing, it obtained 99.8% accuracy, 99.7% precision, 99.9% recall, and 99.6% F1-score. These results demonstrate the model's high predictive performance and consistent classification capability.

Fig. 6 presents the confusion matrix of the proposed XGBoost model for heart disease risk assessment. The model correctly classified 102 non-CHD cases and 100 CHD cases. It produced 3 false-negative predictions and no false-positive predictions. These results demonstrate the model's high classification accuracy and reliable performance in identifying heart disease cases.

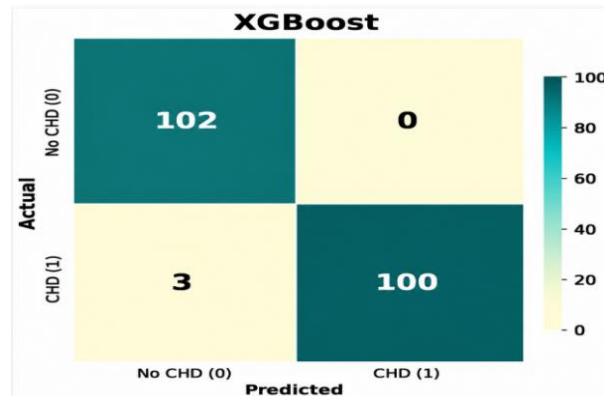


Fig.6. Confusion Matrix for the XGBoost Model

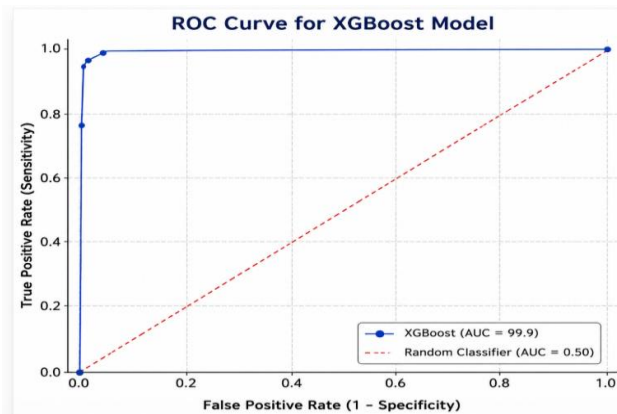


Fig.7. ROC Curve for XGBoost Model

Fig. 7 illustrates the ROC curve of the proposed XGBoost model, showing a strong ability to distinguish between positive and negative classes. The model achieves an excellent AUC of 99.9%, indicating highly accurate classification performance. The curve remains close to the top-left corner and substantially outperforms the random-classifier baseline (AUC = 0.50).

A. Comparative Analysis

To assess the effectiveness of the proposed XGBoost

model, a comparative evaluation against existing machine learning models is presented in Table I. The models are evaluated using accuracy, precision, recall, and F1-score to provide a comprehensive performance assessment. LightGBM, KNN, LR, and CatBoost achieve accuracies of 81.97%, 83%, 88.25%, and 96.61%, respectively. In comparison, the proposed XGBoost model achieves the highest accuracy of 99.8%. It also records 99.7% precision, 99.9% recall, and 99.6% F1-score. These results demonstrate that the proposed XGBoost model provides superior and reliable performance for heart disease risk prediction.

TABLE III. Comparison of Machine Learning Models for Heart Disease Risk Prediction

Model	Accuracy	Precision	Recall	F1-score
LightGBM[30]	81.97	82	82	82
KNN[31]	83	82	75	78
LR[32]	88.25	87.91	88.25	88.65
CatBoost[33]	96.61	95.06	98.09	96.55
Proposed XGBoost Model	99.8	99.7	99.9	99.6

The proposed XGBoost model achieves an outstanding 99.8% accuracy for heart disease risk prediction. This accuracy is higher than the existing models considered in the comparative evaluation. The high accuracy indicates that the model can classify heart disease risk with greater reliability. It demonstrates improved predictive capability and effective learning of complex patterns in the dataset. Therefore, the proposed XGBoost model is highly suitable for accurate heart disease risk assessment.

5. Conclusion and Future Study

Heart disease remains a significant healthcare challenge, making accurate and timely risk assessment important for supporting early intervention and effective clinical decision-making. Machine learning provides promising capabilities for analyzing clinical data and identifying patterns associated with heart disease risk. In this study, an XGBoost-based framework was developed using data preprocessing, One-Hot Encoding, Z-score normalization, SMOTE-based class balancing, and stratified data splitting. The proposed model achieved 99.8% accuracy, 99.7% precision, 99.9% recall, and 99.6% F1-score on the testing data, while achieving 100% across all metrics during training. Comparative analysis showed that XGBoost outperformed LightGBM, KNN, Logistic Regression, and CatBoost. These findings demonstrate the effectiveness of the proposed approach for heart disease risk prediction and highlight its potential to support reliable early risk assessment and data-driven clinical decision-making.

Future work will focus on validating the proposed framework using larger, independent, and diverse clinical datasets to improve generalizability. Advanced ensemble and deep learning techniques can also be explored with explainable AI methods. Further research may investigate feature optimization, external clinical validation, and real-time deployment to enhance practical healthcare applications.

REFERENCES

1. S. Akinola, R. Leelakrishna, and V. Varadarajan, "Enhancing cardiovascular disease prediction: A hybrid machine learning approach integrating oversampling and adaptive boosting techniques," *AIMS Med. Sci.*, vol. 11, no. 2, pp. 58–71, 2024, doi: 10.3934/medsci.2024005.
2. S. C. Patra, B. U. Maheswari, and P. B. Pati, "Forecasting Coronary Heart Disease Risk With a 2-Step Hybrid Ensemble Learning Method and Forward Feature Selection Algorithm," *IEEE Access*, vol. 11, pp. 136758–136769, 2023, doi: 10.1109/ACCESS.2023.3338369.
3. A. Jaiswal, M. Singh, and N. Sachdeva, "Empirical Analysis of Heart Disease Prediction Using Deep Learning," in *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, IEEE, May 2023, pp. 1–9. doi: 10.1109/ACCAI58221.2023.10201235.
4. P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
5. M. Beri, K. S. Gill, and N. Sharma, "Advanced Machine Learning Techniques for Early Prediction of Heart Disease: A Comprehensive Analysis," in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, IEEE, Oct. 2024, pp. 1219–1223. doi: 10.1109/ICSES63445.2024.10763007.
6. M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Med.*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022, doi: 10.1515/med-2022-0508.
7. K. P. Joshi, M. L. Prasad, R. Natchadalingam, P. C. S. Reddy, S. Mukherjee, and G. C. Babu, "An Accurate Prediction of Coronary Heart Disease Using Ensemble Algorithms," in *IEEE 1st International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics, AIKIIIE 2023*, 2023. doi: 10.1109/AIKIIE60097.2023.10390172.
8. H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Sci. Rep.*, vol. 14, no. 1, p. 23277, Oct. 2024, doi: 10.1038/s41598-024-74656-2.
9. M. Mahalakshmi and R. Rajakumari, "Delving into Heart Disease Prediction: A Comprehensive Review," in *2023 2nd International Conference on Ambient Intelligence in Health Care (ICAHC)*, IEEE, Nov. 2023, pp. 01–07. doi: 10.1109/ICAHC59020.2023.10430484.
10. K. Sravani et al., "Early Detection of Lung Cancer Using Convolutional Neural Networks and DNA Methylation Biomarker Analysis," in *Next-Generation Computing and Information Systems*, London, India: CRC Press, 2026, pp. 250–258. doi: 10.1201/9781003648987-34.
11. H. J. Suleiman, I. R. A. Hamid, and O. R. Olaniran, "Smart Health Monitoring for Predicting Heart Disease using IoT-Fog-Cloud Computing Model," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 3, pp. 22565–22572, 2025, doi: 10.48084/etasr.10048.

12. R. Lingam, "Agentic AI and Federated Learning for Privacy-Preserving Dental Imaging," in 2026 Innovations in Machine, Engineering, and Digital Conference (IMED), IEEE, Mar. 2026, pp. 1–6. doi: 10.1109/IMED68921.2026.11484887.
13. K. K. Mohammed, "A Survey on Digital Health Care Data Analysis Techniques for Developing Machine Learning Models," *Int. J. Sci. Eng. Technol.*, vol. 13, no. 5, pp. 1–6, 2025.
14. D. P. Ramaraj and M. Mittal, "Deep Learning Models for Automated Medical Imaging: Accuracy, Ethics, And Deployment Challenges," *Int. J. Artif. Intell. Deep Learn.*, vol. 2, no. 1, pp. 11–31, June, 2025, doi: 10.34218/IJAIDL_02_01_002.
15. M. Swathy and K. Saruladha, "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques," *ICT Express*, vol. 8, no. 1, pp. 109–116, Mar. 2022, doi: 10.1016/j.icte.2021.08.021.
16. M. Swathy and K. Saruladha, "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques," *ICT Express*, vol. 8, no. 1, pp. 109–116, Mar. 2022, doi: 10.1016/j.icte.2021.08.021.
17. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
18. S. N, P. K, C. A, and P. P, "Artificial Intelligence and Pharmaceutical Discovery: A Comprehensive Review of Methods, Applications, and Future Directions," *Int. J. Sci. Res.*, vol. 14, no. 8, pp. 679–681, Aug. 2025, doi: 10.21275/SR25808215949.
19. M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telemat. Informatics*, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.
20. E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, "Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India," *Med. J. Armed Forces India*, vol. 77, no. 3, pp. 302–311, 2021, doi: 10.1016/j.mjafi.2020.10.013.
21. A. Najafi, A. Nemati, M. Ashrafzadeh, and S. Hashemkhani Zolfani, "Multiple-criteria decision making, feature selection, and deep learning: A golden triangle for heart disease identification," *Eng. Appl. Artif. Intell.*, vol. 125, p. 106662, Oct. 2023, doi: 10.1016/j.engappai.2023.106662.
22. A. V. S. R. Dantuluri, "A Scalable Deep Learning Analytics Pipeline for Converting Longitudinal Real-World Data Into Predictive Disease Trajectories," *IEEE Access*, vol. 14, pp. 24871–24878, 2026, doi: 10.1109/ACCESS.2026.3663571.
23. N. Shah and A. Poonia, "Heart Disease Prediction System using Ensemble Machine Learning Technique," in 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA), IEEE, Aug. 2025, pp. 945–949. doi: 10.1109/ICSCSA66339.2025.11171051.
24. S. Kumar, A. Saundik, S. Sindhu, A. Chattraaj, R. Pathania, and A. P. Singh, "Predicting Heart Disease using Machine Learning Approach," in 2025 8th International Conference on Circuit, Power & Computing Technologies (ICCPCT), IEEE, Aug. 2025, pp. 1119–1124. doi: 10.1109/ICCPCT65132.2025.11176573.
25. P. Banerjee, N. T M, M. G. Paul, N. G. S. Krishna, and S. Agarwal, "A Novel Approach for Heart Disease Prediction Using Ensemble Machine Learning Techniques," in 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), IEEE, Feb. 2025, pp. 520–526. doi: 10.1109/ISACC65211.2025.10969413.
26. D. Purwanto, S. C. Hidayati, D. I. Ricoida, and K. A. Putri, "Optimized Machine Learning Models for Heart Disease Prediction: A Performance Analysis," in 2024 International Seminar on Application for Technology of Information and Communication (iSemantic), IEEE, Sep. 2024, pp. 559–562. doi: 10.1109/iSemantic63362.2024.10762500.
27. Z. Rustamov, J. Rustamov, M. S. Sultana, J. Ywei, V. Balakrishnan, and N. Zaki, "Cardiovascular Disease Prediction using Ensemble Learning Techniques: A Stacking Approach," in 2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), IEEE, Mar. 2023, pp. 93–98. doi: 10.1109/CSPA57446.2023.10087730.
28. S. Cyriac, S. R, N. Raju, and Y. Woon Kim, "Heart Disease Prediction Using Ensemble Voting Methods in Machine Learning," in 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), IEEE, Oct. 2022, pp. 1326–1331. doi: 10.1109/ICTC55196.2022.9952655.
29. A. B. Ambrews, E. G. Mounq, A. Farzamnia, F. Yahya, S. Omatu, and L. Angeline, "Ensemble Based Machine Learning Model for Heart Disease

- Prediction,” in 2022 International Conference on Communications, Information, Electronic and Energy Systems (CIEES), IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/CIEES55704.2022.9990665.
30. J. Deshmukh, “Efficient Model for Early Prediction of Heart Disease Using Ensemble Technique,” *Proc. Eng. Technol. Innov.*, vol. 30, pp. 66–78, Jun. 2025, doi: 10.46604/peti.2024.14787.
31. J. S. Mishra, N. K. Gupta, and A. Sharma, “Enhanced Heart Disease Prediction Using Machine Learning Techniques,” *J. Intell. Syst. Internet Things*, vol. 12, no. 2, pp. 19–33, 2024, doi: 10.54216/JISIoT.120202.
32. C. A. ul Hassan et al., “Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers,” *Sensors*, vol. 22, no. 19, p. 7227, Sep. 2022, doi: 10.3390/s22197227.
33. D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, “Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization,” *Algorithms*, vol. 16, no. 6, p. 308, Jun. 2023, doi: 10.3390/a16060308.