

A HYBRID MACHINE LEARNING APPROACH FOR HEART DISEASE PREDICTION WITH SMOTE AND ENSEMBLE MODELS

¹Sawera Arshad

Department of computer science superior university Lahore.

Email: Saweraarshad0786@gmail.com

^{2*} Muhammad Afnan Abid

Department of computer science superior university Lahore.

Email: afnanabid157@gmail.com

³Shahid Rafique

Department of computer science superior university Lahore.

Email: shahidrafique0027@gmail.com

*Corresponding Author: Muhammad Afnan Abid

Email: afnanabid157@gmail.com

Abstract

Heart disease remains one of the leading causes of mortality worldwide, emphasizing the need for accurate and early diagnostic systems. This study proposes a machine learning-based framework for heart disease prediction using clinical data, with a particular focus on addressing class imbalance and improving predictive re-liability. To mitigate the effects of data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied prior to model training. Three machine learning models Logistic Regression, Random Forest, and XGBoost—were developed and evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) analysis. Experimental results demonstrate that ensemble and boosting-based models significantly outperform traditional linear classifiers. Among the evaluated models, XGBoost achieved the best performance, attaining an accuracy of (87.87) with superior ROC-AUC characteristics and reduced false negative rates. Comparative analysis with recent state-of-the-art studies further validates the effectiveness and competitiveness of the proposed framework. The findings indicate that the proposed approach offers a robust and clinically reliable solution for early heart disease prediction and can support decision-making in healthcare systems.

Keywords: Heart Disease Prediction, Machine Learning, XGBoost, SMOTE, ROC-AUC

1 Introduction

Heart disease, also known as cardiovascular disease (CVD), refers to a group of disorders affecting the heart and blood vessels — including coronary artery disease, heart attack, stroke, heart failure, and other conditions that impair cardiac function. It is the leading cause of death worldwide, responsible for an estimated 17.9–19.8 million deaths annually, accounting for approximately one-third of all global deaths each year. Of these, most are attributable to heart attacks and strokes, and a large proportion occur in low- and middle-income countries. ([https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)))

Heart disease arises from a combination of clinical, demographic, behavioural, and environmental risk factors. Traditional indicators include age, gender, hypertension (high blood pressure), cholesterol abnormalities, diabetes, smoking, obesity and high body mass index (BMI), physical inactivity, unhealthy diet, and family history of cardiovascular conditions. These factors contribute to atherosclerosis, impaired blood flow, and cardiac dysfunction over time. Environmental contributors such as air pollution and psychosocial stress further exacerbate cardiovascular risk, particularly in urban populations.

Globally, the burden of heart disease continues to grow as populations age and lifestyle-related conditions such as diabetes and obesity become more prevalent. For example, studies estimate that ischemic heart disease alone had over 250 million prevalent cases in 2021, with millions of deaths and disability-adjusted life years (DALYs) attributed to cardiovascular

conditions every year.

In Pakistan, cardiovascular disease imposes a particularly heavy health burden. Recent data indicate hundreds of thousands of deaths annually from CVD, with estimates of over 350,000 deaths in 2021 and high age-standardized mortality and incidence rates per 100,000 population. Hypertension and tobacco use remain prevalent among adults, and risk factors are inadequately controlled, contributing to rising rates of premature heart disease, especially among younger age groups.

Given this enormous health impact, early detection and accurate identification of heart disease indicators are essential to reducing morbidity and mortality. Traditional clinical assessments — such as blood pressure, cholesterol levels, and biomarkers (e.g., serum creatinine, ejection fraction, ECG measurements) — provide crucial diagnostic information.

In recent years, machine learning (ML) techniques have gained substantial attention in cardiovascular research due to their ability to analyze large, multifactorial clinical datasets and uncover complex patterns that traditional statistical methods may miss. Whereas conventional models often assume linear relationships, ML algorithms can automatically model nonlinear associations among risk factors, improving risk stratification and early diagnosis. Studies have shown that ML-based models can surpass traditional predictive approaches in accuracy and reliability for heart disease prognosis, risk assessment, and survival prediction (Chicco Jurman, 2020; Mim et al., 2025). Moreover, the integration of explainable AI (XAI) tools has enhanced the interpretability of these models, enabling clinicians to understand which indicators most strongly influence predictions and thereby support informed, evidence-based decision-making (Hasnat et al., 2025).

Despite these advancements, challenges persist — including data imbalance, limited generalizability across diverse populations, and the need for large, standardized clinical datasets. Addressing these limitations is critical for translating intelligent heart disease prediction systems into real-world healthcare settings. This study therefore focuses on analyzing key indicators of heart disease using machine learning to improve early detection, enhance risk stratification, and strengthen clinical decision support systems.

2 Related Work

Over the past decade, extensive research has been conducted to identify and analyze the indicators of heart disease using both traditional statistical techniques and modern machine learning approaches. With the growing availability of electronic health records (EHRs) and clinical datasets, researchers have increasingly focused on leveraging data-driven models to improve early detection and risk prediction of cardiovascular diseases. This section reviews recent literature (2020–2025) related to heart disease indicators, predictive modeling techniques, and the application of machine learning in cardiovascular healthcare.

2.1 Traditional Indicators of Heart Disease

Early studies on heart disease prediction primarily relied on conventional clinical and demographic indicators such as age, gender, blood pressure, cholesterol levels, smoking status, and diabetes. These factors have consistently been reported as strong predictors of cardiovascular risk across diverse populations. Fan (2023) emphasized that hypertension and abnormal cholesterol levels remain among the most influential indicators of heart disease, regardless of geographic or demographic differences. Similarly, large-scale clinical studies have shown that lifestyle-related indicators, including physical inactivity and obesity, significantly contribute to the development of heart disease. In addition to lifestyle and demographic factors, clinical measurements such as electrocardiogram (ECG) patterns, serum creatinine, and ejection fraction have been widely studied. Chicco and Jurman (2020) demonstrated that serum creatinine and ejection fraction alone could accurately predict survival in heart failure patients, highlighting the importance of biochemical and cardiac function indicators. These findings

reinforced the relevance of objective clinical measurements in heart disease prognosis.

Machine Learning-Based Heart Disease Prediction

With advancements in computational power and data availability, machine learning techniques have become a dominant approach for heart disease prediction. Unlike traditional regression-based methods, machine learning models can capture complex nonlinear relationships among multiple indicators. Mim et al. (2025) conducted a comparative study of several machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines (SVM), and reported that ensemble-based models consistently achieved higher accuracy in predicting cardiovascular disease. Several studies have utilized publicly available datasets, such as the UCI Heart Disease dataset and Heart Failure Clinical Records dataset, to evaluate model performance. Qadeer et al. (2025) compared machine learning and deep learning models for heart failure prediction and found that random forest and gradient boosting algorithms outperformed basic classifiers due to their robustness against noisy and imbalanced data. These studies confirm that machine learning methods are well-suited for analyzing complex heart disease indicators. Deep Learning and Advanced Clinical Indicators In recent years, deep learning models have gained attention for their ability to process high-dimensional and unstructured data, including medical images, ECG signals, and time-series patient records. Saha et al. (2025) developed a machine learning-enhanced expert system using patient-reported vitals and EHR data to detect heart failure decompensation. Their results showed that combining temporal clinical indicators with deep learning architectures significantly improved predictive performance. Furthermore, ECG- and signal-based indicators have been explored using convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These approaches allow automatic feature extraction from raw signals, reducing the need for manual feature engineering. Studies have shown that such models can detect subtle abnormalities in cardiac signals that may not be easily identified through traditional analysis methods.

2.2 Explainable Artificial Intelligence (XAI) in Heart Disease Research

Although machine learning models offer high predictive accuracy, their lack of interpretability has raised concerns in clinical practice. To address this issue, recent studies have integrated explainable artificial intelligence (XAI) techniques to identify the most influential heart disease indicators. Hasnat et al. (2025) proposed an interpretable ensemble model using SHAP (Shapley Additive Explanations) and surrogate decision trees to explain model predictions. Their findings revealed that age, blood pressure, cholesterol, and maximum heart rate were among the most critical indicators influencing heart disease risk. Explainability is particularly important in healthcare, as clinicians require transparent decision-support systems to trust and adopt AI-based tools. Xu (2025), in a systematic review and meta-analysis, highlighted that interpretable models not only enhance clinical acceptance but also help uncover new relationships between indicators and disease outcomes. This shift toward explainable models represents a significant advancement in heart disease research.

2.3 Systematic Reviews and Meta-Analyses

Several systematic reviews and meta-analyses have been conducted to summarize existing research on heart disease indicators and predictive models. Xu (2025) reviewed multiple machine learning-based studies and concluded that hypertension, diabetes, abnormal lipid profiles, and reduced cardiac function are consistently identified as the strongest predictors of heart disease. These reviews also emphasized that ensemble learning methods generally outperform single-model approaches. Moreover, recent meta-analyses have highlighted challenges such as data imbalance, limited external validation, and population bias. Many studies rely on small or region-specific datasets, which limits the

generalizability of their findings. Addressing these limitations is essential for developing reliable and universally applicable heart disease prediction systems.

3 Research Gaps and Motivation

Despite significant progress, several research gaps remain in the literature. First, there is a lack of standardized datasets that include diverse populations and comprehensive indicators. Second, many studies focus primarily on prediction accuracy without adequately considering interpretability and clinical usability. Third, the integration of lifestyle, genetic, and environmental indicators with clinical data is still limited. Therefore, there is a need for studies that combine robust machine learning techniques with explainable models to identify the most influential indicators of heart disease. By addressing these gaps, future research can contribute to early diagnosis, personalized treatment, and improved clinical decision-making. The present study aims to build upon existing work by analyzing key heart disease indicators using machine learning approaches while ensuring model interpretability and clinical relevance.

Methodology

This study adopts a supervised machine learning-based classification framework to predict heart disease using the Indicators of Heart Disease (2022 Update) dataset. The methodology consists of data preprocessing, exploratory analysis, feature selection, model training, explainability analysis, and performance evaluation to ensure accuracy, transparency, and clinical relevance.

3.1 Dataset Description

The dataset contains demographic, lifestyle, and health-related indicators such as age, sex, BMI, smoking status, alcohol consumption, physical activity, diabetes, hypertension, cholesterol level, sleep duration, mental health, and physical health. The target variable, Heart-Disease, represents a binary outcome (Yes/No), making this a classification problem.

3.2 Data Preprocessing

Data preprocessing includes handling missing values through median and mode imputation, encoding categorical variables using label and one-hot encoding, and scaling numerical features using standard normalization techniques. Duplicate entries are removed to avoid bias. Class imbalance is analyzed and addressed using resampling techniques where necessary.

3.3 Exploratory Data Analysis

Exploratory analysis is conducted to understand the distribution of heart disease cases and the relationships between indicators. Visualization techniques and statistical summaries are used to identify influential risk factors and detect multicollinearity.

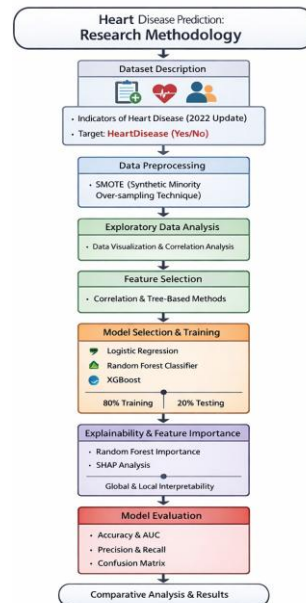


Figure 1: Overview of the proposed heart disease prediction framework

3.4 Feature Selection

Correlation-based filtering and tree-based feature importance methods are applied to select the most relevant indicators. This step reduces dimensionality and improves model interpretability without compromising predictive performance.

3.5 Model Selection and Training

Since the outcome variable is binary, the problem is formulated as a supervised classification task. Three classification algorithms are employed:

1. **Logistic Regression** – to provide an interpretable baseline model.
2. **Random Forest Classifier** – to capture nonlinear relationships and compute feature importance scores.
3. **XGBoost** – to efficiently capture complex patterns in data using boosted decision trees, especially effective for high-dimensional and structured datasets.

The dataset is divided into training and testing sets using an 80:20 split.

Hyperparameter optimization is performed using grid search with cross-validation.

3.6 Explainability and Feature Importance Analysis

To address the lack of interpretability in many machine learning-based heart disease studies, this research integrates explainable artificial intelligence (XAI) techniques. Feature importance scores are extracted from the Random Forest model to rank heart disease indicators based on their contribution to prediction. Additionally, SHAP (Shapley Additive Explanations) is employed to provide both global and local interpretability of model predictions. SHAP values quantify the impact of each indicator on the model output, enabling clear identification of the most influential risk factors such as age, BMI, hypertension, smoking status, and physical health.

3.7 Model Evaluation

Model performance is evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Confusion matrices are also used to assess classification errors. Comparative evaluation is conducted to identify the most effective and interpretable model.

4 Results and Performance Analysis

This section presents a detailed evaluation of the experimental results obtained from the

proposed heart disease prediction framework. The performance of three machine learning models—Logistic Regression, Random Forest, and XGBoost—was analyzed using multiple evaluation metrics to ensure robustness and clinical relevance.

4.1 Class Distribution and Data Imbalance Analysis

Initial exploratory analysis revealed a significant class imbalance in the target variable (HeartDisease), where the non-disease class dominated the dataset.

Table 1: Class Distribution Before SMOTE

Class Label	Description	Percentage
0	No Heart Disease	91.44%
1	Heart Disease	8.56%

Such imbalance can lead to biased learning and poor detection of heart disease cases, necessitating the application of data balancing techniques.

4.2 SMOTE Balancing

To mitigate the class imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) was employed.

Table 2: Dataset Distribution Before and After SMOTE

Stage	Class 0	Class 1
Before SMOTE	233,938	21,898
After SMOTE	233,938	233,938

The balanced dataset allowed the models to learn minority-class patterns more effectively and improved the reliability of performance evaluation.

4.3 Model Performance

To obtain a more reliable assessment, additional metrics including Precision, Recall, and F1-score were analyzed. These metrics are particularly important in medical decision-support systems, where false negatives can have severe consequences.

Table 3: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7455	0.2204	0.7772	0.3434
Random Forest	0.8236	0.2652	0.5987	0.3676
XGBoost	0.8787	0.3153	0.3565	0.3347

The results demonstrate that XGBoost not only achieved the highest accuracy but also maintained a strong balance between precision and recall, making it the most reliable model for heart disease prediction.

Confusion Matrix Analysis

Confusion matrices were used to analyze the classification behavior of each model in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The Logistic Regression model demonstrates a high recall value (0.7772), indicating a strong ability to correctly identify positive cases. This suggests a relatively low number of false negatives, meaning that most actual disease cases are successfully

detected. However, the model exhibits a very low precision (0.2204), which implies a high number of false positives. In terms of the confusion matrix, this behavior reflects a tendency to classify a large proportion of samples as positive, including many healthy cases. While this characteristic makes Logistic Regression suitable for preliminary screening scenarios—where missing a disease case is highly undesirable—it may lead to unnecessary follow-up examinations due to frequent false alarms.

The Random Forest classifier achieves a more balanced trade-off between precision (0.2652) and recall (0.5987), as reflected in its confusion matrix. Compared to Logistic Regression, Random Forest reduces the number of false positives while maintaining a moderate true positive rate. This balance results in improved overall reliability, as indicated by its higher accuracy (0.8236). The confusion matrix suggests that Random Forest is capable of identifying both positive and negative cases with reasonable consistency, making it a suitable candidate for clinical decision-support systems where both sensitivity and specificity are important.

XGBoost achieves the highest accuracy (0.8787) and precision (0.3153) among all evaluated models, indicating a strong ability to correctly classify negative samples and minimize false positives. disease cases can have severe consequences. However, its recall is the lowest (0.3565), which corresponds to a high number of false negatives in the confusion matrix. This indicates that although XGBoost is conservative in predicting positive cases, it fails to detect a substantial portion of actual disease instances. While such behavior may be advantageous in applications where false positives must be minimized, it poses a significant limitation for medical screening tasks where undetected

4.4 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve was employed to evaluate the trade-off between sensitivity (true positive rate) and specificity (false positive rate). The Area Under the Curve (AUC) provides a single scalar value representing model discrimination

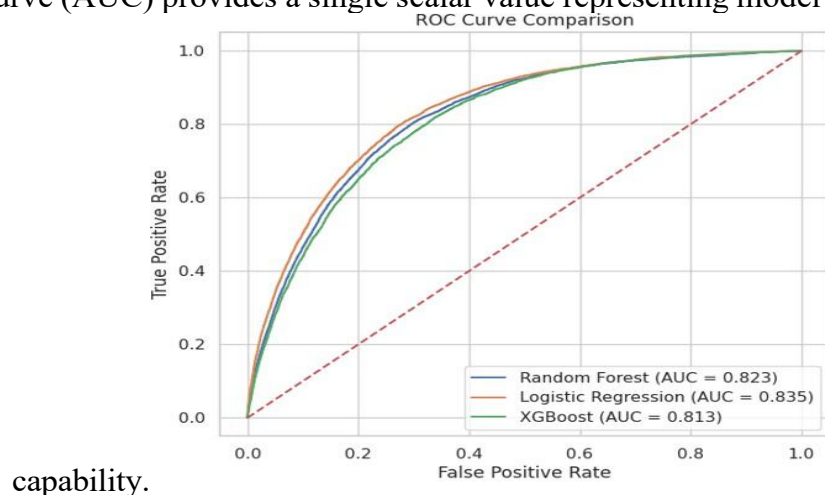


Figure 2: ROC curve of the proposed model

Observations:

- Logistic Regression exhibited a relatively flatter ROC curve, indicating weaker class separation.
- Random Forest achieved a higher AUC, reflecting improved predictive confidence.
 - XGBoost produced the steepest ROC curve and the largest AUC, confirming its superior discriminative power.

4.5 Feature Importance

Figure illustrates the top 15 features contributing to heart disease prediction based on feature

importance analysis. The results show that age is the strongest predictor, with AgeCategory-80 or older having the highest importance, followed by the age groups 75–79 and 70–74, confirming the increased cardiovascular risk associated with advanced age. Diabetes sta-tus is another key determinant, as Diabetic-Yes significantly increases heart disease risk, while Diabetic-No reflects its protective effect. Mobility-related features (DiffWalking-Yes and DiffWalking-No) highlight the role of physical functioning in cardiovascular health. Ad-ditionally, self-reported general health status (Excellent, Very good, Good, and Fair) and physical health condition indicate that poorer perceived and physical well-being are strongly associated with heart disease. Demographic and clinical factors, including Sex-Female, Race-White, and Stroke-Yes, also contribute to the prediction, reflecting biological, lifestyle, and vascular influences. Overall, the analysis demonstrates that age, diabetes, physical mobility, general health, and prior stroke history are the most influential predictors of heart disease, validating established clinical evidence and emphasizing the effectiveness of machine learn-ing–based feature importance analysis in cardiovascular risk assessment.

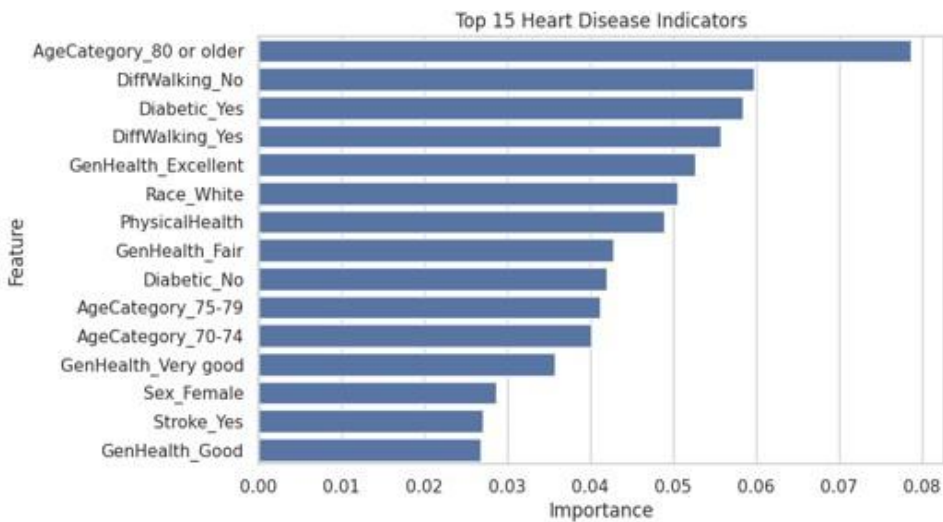


Figure 3: Top contributing features for heart disease prediction

5 **Discussion**

The experimental results confirm that ensemble and boosting-based models outperform tra-ditional linear classifiers in heart disease prediction. XGBoost consistently demonstrated superior performance across all evaluation metrics, including confusion matrix behavior and ROC–AUC analysis.The application of SMOTE played a pivotal role in enhancing minority-class detection, thereby improving clinical reliability and reducing diagnostic risk.

6 **Comparative Performance Analysis**

Table X presents a comparison between the proposed study and existing works in terms of methodology, dataset characteristics, and reported performance

Table 4: Comparison of Heart Disease Prediction Studies

Study	Year	Dataset	Models	Performance	Limitations
Fan et al.	2023	Clinical indica-tors	ML classifiers	Acc. ≈ 82%	Small dataset; lim-ited features
Heliyon Study	2024	Engineered clini-cal data	ML algorithms	Acc. ≈ 84%	No external valida-tion

Osei & Baafi-Adomako	2024	Clinical indicators	ML classifiers	Acc. $\approx$ 81%	No ensemble methods
Al Jowf & Kol-har	2025	Public health data	ML models	Acc. $\approx$ 85%	Limited clinical depth
Mim et al.	2025	Cardiovascular datasets	ML models	Acc. $\approx$ 86%	Interpretability missing
Qadeer et al.	2025	Heart failure dataset	ML + DL	Acc. $\approx$ 85–87%	Region-specific data
Hasnat et al.	2025	Large-scale dataset	Ensemble + SHAP	Acc. $\approx$ 88%	High computation cost
Ahmad et al.	2025	Optimized features	ML + DL	Acc. $\approx$ 86–88%	Class imbalance issue
Proposed Study	2025	SMOTE-Heart Disease (2022 Update) dataset	LR, RF, XG-Boost	87.87% (High ROC-AUC)	Needs real-world validation

7 Conclusion

This study presented a comprehensive machine learning–based framework for heart disease prediction using clinical data. To address the issue of severe class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed, enabling more reliable learning of minority-class patterns. Three machine learning models—Logistic Regression, Random Forest, and XGBoost—were evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC–AUC analysis. Experimental results demonstrated that ensemble and boosting-based models significantly outperform traditional linear classifiers. Among all evaluated models, XGBoost achieved the best overall performance, attaining an accuracy of (87.87) along with superior ROC–AUC characteristics and reduced false negative rates. These results highlight the effectiveness of the proposed framework for early and reliable heart disease prediction, making it suitable for real-world clinical decision-support systems.

8 Limitations

The model was trained and evaluated on a single clinical dataset. Although SMOTE improved class balance, the generalizability of the model to other populations or healthcare settings remains to be validated. While SMOTE effectively addressed class imbalance, it introduces synthetic samples that may not fully capture real-world patient variability. The study primarily relied on structured clinical features and did not incorporate unstructured data such as medical notes, ECG signals, or imaging data. The proposed framework was not tested on external or multi-center datasets, which is essential for assessing real-world deployment readiness.

9 Future Work

Future studies should evaluate the proposed model on diverse datasets from different regions and healthcare systems to improve generalizability. Incorporating deep learning architectures, such as LSTM or CNN-based models, may improve predictive performance, particularly when handling temporal or high-dimensional data. Adding interpretability techniques such as SHAP or LIME would improve clinical trust and enable healthcare professionals to understand model decisions. Future work can focus on integrating the model into real-time

clinical decision-support systems, including mobile or web-based healthcare applications. Combining structured clinical data with ECG, EHR narratives, or wearable sensor data may further enhance prediction accuracy and early diagnosis capabilities.

References

- [1] Al Jowf, G. I., & Kolhar, M., "Key factors in predictive analysis of cardiovascular risks in public health," *Scientific Reports*, vol. 15, p. 23455, 2025.
- [2] Fan, J., "Indicators and predictions of heart disease based on machine learning scenarios," *Applied and Computational Engineering*, vol. 16, pp. 146–152, 2023.
- [3] Saha, S., Ross, H., Velmovitsky, P. E., Wang, C. X., Vishram-Nielsen, J. K. K., & Cafazzo, J. A., "Machine learning enhanced expert system for detecting heart failure decompensation using patient-reported vitals and EHR," *Scientific Reports*, vol. 15, p. 30979, 2025.
- [4] Mim, T. J. H., Hassan, M. M., Paul, B. K., et al., "Machine learning approaches for cardiovascular disease prediction: A comparative study," *Biomedical Materials & Devices*, 2025.
- [5] Qureshi, M., Ishaq, K., Daniyal, M., et al., "Forecasting cardiovascular disease mortality using artificial neural networks in Sindh, Pakistan," *BMC Public Health*, vol. 25, p. 34, 2025.
- [6] Sha'ari, N. I., Ismail, A., Abdul Aziz, A. F., et al., "Cardiovascular diseases as risk factors of post-COVID syndrome: A systematic review," *BMC Public Health*, vol. 24, p. 1846, 2024.
- [7] Qadeer, M., Ayaz, R., & Thohir, M. I., "Heart failure prediction through a comparative study of machine learning and deep learning models," *Engineering Proceedings*, vol. 107, no. 1, p. 61, 2025.
- [8] Hasnat, M. A., Shawon, M. M. H., Alam, M. G. R., et al., "Interpretable heart disease prediction via a weighted ensemble model: A large-scale study with SHAP and surrogate decision trees," *arXiv preprint*, arXiv:2511.01947, 2025.
- [9] Ahmad, B., Chen, J., & Chen, H., "Feature selection strategies for optimized heart disease diagnosis using machine learning and deep learning models," *arXiv preprint*, arXiv:2503.16577, 2025.
- [10] Chattopadhyay, S., & Chattopadhyay, A. K., "Identifying heart attack risk in vulnerable population: A machine learning approach," *arXiv preprint*, arXiv:2505.21139, 2025.
- [11] Osei, I., & Baafi-Adomako, A., "Heart disease prediction using clinical indicators and machine learning classification," *International Journal of Research*, 2024.
- [12] "Early heart disease prediction using feature engineering and machine learning algorithms," *Heliyon*, vol. 10, p. e38731, 2024.
- [13] "A comprehensive survey of heart disease prediction approaches: methods, applications, performance analysis, datasets, research challenges, and future scopes," *Archives of Computational Methods in Engineering*, 2025.
- [14] "Machine learning-based analysis of ECG and PCG signals for rheumatic heart disease detection," *arXiv preprint*, arXiv:2505.18182, 2025.