ORIGINAL ARTICLE

A Comprehensive Approach to Heart Disease Analysis Using Machine Learning Algorithms

Shafayat Bin Shabbir Mugdha[1], Mahtab Uddin[2,*], and Hridoy Das[1]

[1]*Department of Computer Science & Engineering, United International University, Dhaka-1212, Bangladesh*

[2]*Institute of Natural Sciences, United International University, Dhaka-1212, Bangladesh*

(**Received** 05 August 2024, ***Accepted*** *30 Jan 2025,* ***Published*** *Online 17 Feb 2025*)

*Corresponding author: mahtab@ins.uiu.ac.bd

DOI: 10.5875/gkz48s35

Abstract: Heart disease, associated with cardiovascular disease, has become the leading global cause of mortality, claiming an estimated 17.9 million lives annually. Recognizing the preventable nature of nearly 90% of these cases, researchers have turned to data mining methodologies to aid healthcare professionals in early diagnosis. machine learning improves operational efficiency. Early heart disease prediction will simplify patient management, reduce hospital admissions, and minimize healthcare costs. Beyond that, machine learning technologies improve accuracy, enabling better decision-making and efficient use of medical resources. Our research seeks to establish an effective assumption framework for predicting heart disease probabilities. The ultimate goal is to pinpoint the most accurate classification algorithm, which is crucial for distinguishing between normal and abnormal cases. Notably, our analysis highlights the efficacy of the XGBoost algorithm, boasting an impressive 99.90% accuracy. Advanced feature selection and interpretability techniques, like SHAP values, to identify important risk factors for heart disease prediction. A 10-fold stratified cross-validation ensures robust performance evaluation. Unlike previous studies with limited datasets, this research validates XGBoost on a large, diverse dataset. Comparisons with recent works show its superior predictive capabilities. These enhancements make the study rigorous and relevant for heart disease prevention. This discovery promises to advance early detection strategies and reduce the global impact of heart conditions.

Keywords: Heart Disease; Machine Learning; 10-Fold Cross-Validation; Matthew's Correlation Coefficient, Data Mining; Good Health & Well-Being.

Introduction

The heart, a vital muscular organ responsible for circulating blood throughout the cardiovascular system, plays a crucial role in sustaining human life. In the contemporary hustle and bustle, numerous influential factors impact this crucial organ in diverse ways. Presently, the incidence of heart disease is on the rise, aligning with the multitude of challenges arising in our fast-paced world. In addition to smoking, obesity, lack of physical activity, familial predisposition, unhealthy dietary habits, elevated cholesterol levels, hypertension, and poorly managed diabetes, numerous risk factors have direct or indirect effects on cardiovascular health [1]. Additionally, genetic factors contribute to specific heart conditions, such as hypertrophic cardiomyopathy. This intricate interplay of factors underscores the need for proactive detection and intervention to safeguard heart health [2].

Cardiovascular diseases can impact individuals across all age groups, with no specific immunity based on age. However, the probability of developing heart disease generally increases as humans age [3]. In men, this risk typically begins to rise around the age of 45, with approximately one in every 100 men exhibiting signs of heart disease at this stage. For women, the risk of cardiovascular disease escalates with age, although symptoms tend to manifest roughly a decade later than in men. Notably, there is a concerning trend of heart disease affecting younger populations, driven by the escalating rates of diabetes and childhood obesity. The rising prevalence of obesity among youth is associated

with an increased probability of cardiovascular risk factors such as elevated cholesterol levels and high blood pressure [4]. In a recent study encompassing individuals aged 5 to 17, it was found that 70% of obese youth exhibited at least two risk factors for cardiovascular disease. Detecting and addressing these risk factors early on is crucial for effective prevention and intervention strategies.

Accurate prediction of risk factors related to cardiovascular disease is vital for effective diagnosis and treatment [5]. Biologists are increasingly leveraging cutting-edge machine learning techniques to glean pertinent information from databases [6]. It's worth noting that "data mining" doesn't extract data, but rather patterns and knowledge from large datasets [7]. Through the adoption of machine learning, biologists may examine enormous volumes of data and find hidden patterns that would not be seen employing more conventional techniques. This procedure enables a deeper understanding of the risk factors responsible for cardiovascular disease, contributing to better diagnosis and more focused strategies for treatment [8]. Additionally, the use of machine learning in data mining enables researchers to make evidence-based decisions and develop personalized interventions for patients at risk.

This procedure extracts crucial decision-making information from a repository of records, enabling analysis or prediction in subsequent instances. In the medical domain, data mining plays a significant role in forecasting and analyzing future patient states. By employing classified techniques and offering computerized training on datasets, medical data mining facilitates insights into a patient's history. This, in turn, enables clinical support through comprehensive analysis [9]. Essentially, medical data mining relies on various classifications, a crucial aspect in identifying potential risks of a heart attack before it occurs. Through the training and testing of classification algorithms, predictions can be made to discern an individual's susceptibility to cardiovascular disease or heart-related conditions. By analyzing patterns and correlations within the medical data, medical data mining can also help in identifying early warning signs of heart-related conditions [10].

This proactive approach allows healthcare professionals to intervene and provide timely preventive measures to reduce the risk of heart attacks or other cardiovascular diseases. Furthermore, the continuous refinement of classification algorithms through ongoing research and development enhances the accuracy and effectiveness of these predictions, ultimately improving patient outcomes.

Implementation of machine learning approaches enhances operational efficiency. The ability to predict heart diseases earlier will streamline patient management, reduce hospital admissions, and minimize the financial burden on both individuals and healthcare providers [11]. Heart disease prediction through machine learning holds economic significance by enabling early detection and intervention, potentially reducing healthcare costs and improving patient outcomes. The predictive power of models contributes to more targeted and efficient healthcare resource allocation, offering a promising avenue for cost-effective preventive strategies in cardiovascular health [12]. By accurately identifying individuals at high risk for heart disease, machine learning can help healthcare providers prioritize interventions and allocate resources effectively. This can lead to timely interventions, such as lifestyle modifications or medication, that can prevent or delay the onset of heart disease and its associated complications [13]. Ultimately, this approach could save lives, reduce healthcare costs, and improve population health. Additionally, the improved accuracy of machine learning approaches will lead to better-informed decision-making, ensuring that medical resources are utilized more effectively [14].

The study implements a stratified 10-fold cross-validation technique to analyze machine learning classifiers for heart disease prediction. Important measures including accuracy, precision, recall, specificity, F1-score, and Matthew's correlation coefficient are adopted for assessing classifiers. The present research is based on real medical data obtained from Kaggle, forming the cornerstone of our analysis. Through meticulous training with this dataset, we derived a classification model that has demonstrated notable results in predicting various types of heart diseases. This process involved achieving 99.90% outcomes, thereby showcasing the effectiveness of our approach in the field of heart disease prediction.

The rest of this article is organized as follows: In Section 3, we discussed previous methods related to model heart diseases, providing a comprehensive overview of the existing research in this field. Building upon this foundation, we propose a more effective and practical approach in Section 4, which leverages novel techniques to improve the accuracy and efficiency of heart disease modeling. Section 5.3 presents the results

obtained from our proposed model, highlighting its performance and comparing it with existing approaches.

Importance of Machine Learning Approaches in Heart Disease Detection and Their Economic Benefits

In this section, the importance of machine learning approaches for detecting heart disease with due examples are provided. The economic benefits of detecting heart disease through the machine learning approaches are narrated as well.

Importance of Machine Learning Approaches for Detecting Heart Disease

Machine learning (ML) has become influential in transforming healthcare, particularly in the sector of heart disease detection, by analyzing extensive medical datasets, including demographics, vital signs, and diagnostic tests like ECGs and echocardiograms. ML models like Support Vector Machines (SVMs), Random Forests, Neural Networks, and Deep Convolutional Neural Networks (CNNs) have risen as frontrunners in this field. SVMs, [15] establish optimal decision boundaries in high-dimensional data spaces, effectively differentiating healthy individuals from those with heart disease. Random Forests, [16] enhance robustness by generating diverse decision trees, leading to predictions that are both robust and resistant to overfitting. Neural Networks, [17], excel in capturing intricate relationships between features, enabling accurate predictions even in noisy data. Furthermore, in the context of medical image analysis, deep CNNs demonstrate excellence in identifying anomalies associated with heart disease [18]. These machine learning approaches, each with their unique strengths, showcase the potential to uplift heart disease detection and emphasize the importance of leveraging advanced algorithms in healthcare.

Economic Benefits Using Machine Learning in Heart Disease Detection

Using machine learning algorithms for heart disease prediction and detection has shown great potential for reducing healthcare costs. Recent studies have demonstrated that machine learning models can analyze cardiac imaging tests, patient Electronic Health Records (EHRs), and other clinical data to identify those at high risk of cardiovascular disease with a relatively high degree of accuracy [19]. For example, one group developed a deep neural network model that could predict the 5-year risk of acute coronary syndrome using EHRs with over 90% accuracy [20]. Translating these ML predictive analytics into clinical practice could enable earlier interventions in high-risk patients that prevent or delay costly adverse events like heart attacks and strokes downstream. Economic analyses have quantified the potential cost savings from the implementation of machine learning in cardiac care. A study [21] simulated the effects of an EHR-based machine learning algorithm to risk stratify patients over 5 years. They estimated it could prevent over 28,000 major adverse cardiovascular events, gaining nearly 60,000 quality-adjusted life years. At a cost of only $2.32 per patient screened, this would translate into approximately $283 million in averted healthcare costs. More widespread deployment of machine learning for proactive heart disease management may lead to billions saved in cardiovascular treatment expenditures each year [22]. Though the algorithms require substantial development, they could be highly cost-effective diagnostic and prioritization tools to improve population heart health outcomes. However, challenges remain in validating the machine learning predictions across diverse patient groups before full clinical adoption [23].

Literature Review

In our daily lives, numerous elements can influence the human heart. According to the European Public Health Alliance, 41% of all deaths are related to heart attacks, strokes, and other circulatory diseases. Heart disease presents various symptoms, posing challenges for swift and accurate diagnosis. Discomfort in the chest, shortness of breath, and weariness are all symptoms that are frequently associated with heart disease. However, it is essential to keep in mind that these symptoms can vary widely from one individual to the next. Consequently, medical professionals need to take into consideration a variety of criteria when diagnosing cardiac diseases among their patients. In addition, developments in medical technology have resulted in the creation of a variety of diagnostic tests and treatments that help diagnose heart disease and choose the treatment plan that is best suitable for the individual.

Research Over the Years

There has been much discussion among authors about an effort to use previous medical data to anticipate heart disease. All of the authors attempt to approach their articles from diverse points of view. Statlog is the name of the heart disease dataset that the author used in [24] to test six machine learning classification methods. Some of the authors started to use medical data from [25] by processing information like age, sex, blood pressure, and blood sugar using two algorithms: neural network and K-means clustering.

Furthermore, they discovered that K-means clustering is inferior to artificial neural networks for all the parameters. The UCI machine learning Repository is another source where the author employed two more approaches (Decision Tree and Naive Bayes) to forecast utilizing an attribute-structured clinical database [26]. Naive Bayes outperforms Decision Trees in terms of performance, according to the author's findings. An examination of heart disease expectations was conducted by Duff et al. with the participation of 533 patients who had previously had cardiac arrest [27]. The Bayesian network has primarily been relied on by them. The task was completed by Singh et al. [28] using a compiled algorithm known as K-Means. In production scenarios with large datasets, the K-means algorithm exhibits optimal efficacy, versatility, and speed of assembly. They have used the Weka data mining tool to calculate the accuracy and running time performance of k-means clustering methods. The Intelligent Heart Disease Prediction System (IHDPS) is a model developed by Palaniappan et al. [29] using a variety of data mining approaches, including Decision Trees, Naive Bayes, and Neural Networks.

Available Datasets

Over the years, various studies have utilized diverse datasets for the convenience of researchers and the feasibility of simulations. Researchers incorporated the Cleveland dataset from the UCI repository into their studies [30–32]. This dataset comprises approximately 76 attributes and 303 records [33]. However, the prior research utilized only 13 attributes. Different research endeavors have employed the Statlog dataset from the UCI repository [34, 35]. Additionally, from the UCI repository, the heart-c.arff dataset has been utilized in several studies, culminating in a notable study [36]. This study employed the data mining tool Weka to discern heart diseases employing two distinct classification methods: J48 classification was applied to the Hungarian dataset, while Naïve Bayes classification was employed on the echocardiogram database [37].

Data Collection Strategy and Methodology

Detailed attributes of the target dataset, narration of data preprocessing, and schematic methodology of this work are included in this section.

Attributes of the Dataset

The proposed model is evaluated on a publicly available dataset that comprises patient data sourced from Kaggle [38]. The dataset, described as in Table 1, includes 14 attributes related to heart disease risk factors, such as age, gender, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, and exercise-induced angina. The target attribute denotes the presence of heart disease (1 = yes, 0 = no), while the id attribute functions as the unique identifier. This dataset is widely utilized for benchmarking machine learning models in cardiovascular disease prediction.

Data Preprocessing

The preprocessing phase involved several crucial steps to prepare the dataset for machine learning. Initially, Min-Max scaling was employed to standardize continuous features such as age, cholesterol levels, and blood pressure, ensuring that all values were normalized between 0 and 1. This technique was essential to prevent any individual feature from disproportionately impacting the model's performance. Subsequently, missing and null data were systematically addressed. For continuous variables, mean imputation was utilized to fill in missing values, while categorical features were filled using the mode. Rows with excessive missing data were removed to maintain the dataset's integrity. Furthermore, the dataset underwent cleaning to identify and rectify inconsistencies or errors, including outliers, thereby enhancing data quality. Random shuffling was applied to eliminate potential biases arising from the order of entries. Ultimately, the dataset was segmented into training 80% and testing 20% subsets, ensuring a stratified split that preserved a balanced representation of target classes in both sets. These preprocessing measures collectively laid a robust and reliable foundation for model training and evaluation.

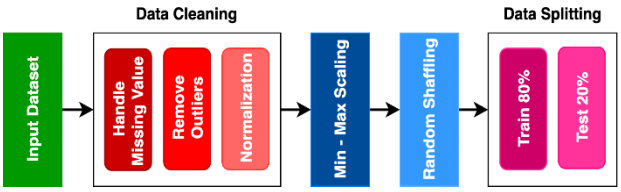

Figure 1: Applied Preprocessing Techniques

Feature Importance and Clinical Relevance

The feature importance analysis conducted using Random Forest and XGBoost models identifies key predictors of heart disease and their clinical significance. Chest pain type (15.38% for Random Forest, 24.61% for XGBoost) emerges as a strong indicator of potential cardiac complications and is the most crucial feature in both models. The number of major vessels (12.61% for Random Forest, 10.71% for XGBoost) directly correlates with heart blood supply, while ST depression induced by exercise (12.38% for Random Forest, 8.57% for XGBoost) quantifies heart response during physical exertion. The

maximum heart rate achieved (10.45% for Random Forest, 8.25% for XGBoost) reflects heart performance under stress. Thalassemia (12.27% for XGBoost) emerges as a pivotal predictor, particularly in cases associated with heart disease. These features, including chest pain type, major vessels, and ST depression, closely align with clinical diagnostic markers, underscoring their relevance in assessing cardiac risk. Figure 2 provides a visual manifestation of the clinically significant features mentioned above.

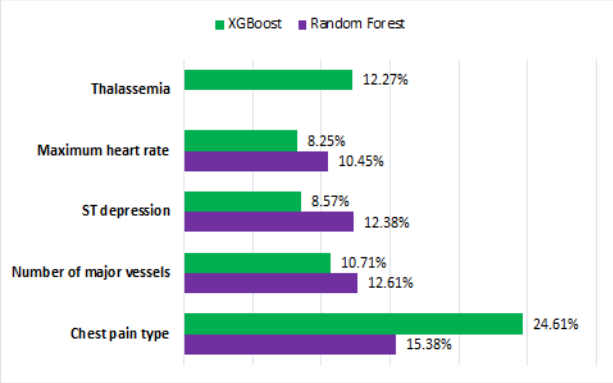

Figure 2: Comparison of Clinically Significant Features

Proposed methodology

A thorough analysis of various machine learning classifiers used to predict heart disease has been conducted. The dataset is loaded from a CSV file and processed initially through feature separation and normalization using the MinMaxScaler. A stratified 10-fold cross-validation strategy is employed to assess the classifiers' performance, ensuring an equitable distribution of the target variable across all 10 folds. Systematic testing of classifiers includes Random Forest, logistic regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), XGBoost, and Naive Bayes. Evaluation metrics such as recall, specificity, F1-score, and Matthew's correlation coefficient (MCU) are utilized to assess the performance of each classifier. This systematic approach provides a comprehensive understanding of the efficacy of each classifier, enabling informed decisions regarding the selection of the most suitable model for heart disease prediction.

Particularly, within the Random Forest classifier, we introduced feature selection techniques to enhance model interpretability and potentially improve performance. Feature selection methods within Random Forest aim to identify the most influential features, contributing to the model's decision-making process. By integrating this aspect into our analysis, we gain insights into the importance of different features in predicting heart disease, further refining the model's capabilities. We can identify the most suitable model for heart disease prediction by comparing the performance of multiple classifiers using various metrics.

We used a dual evaluation strategy a train-test split for initial performance evaluation and cross-validation for comprehensive evaluation. The train-test split provided a straightforward measure of the model's accuracy on unseen data, while cross-validation ensured the consistency and reliability of performance metrics by averaging results across multiple folds. This comprehensive evaluation helps to ensure accurate and reliable predictions, ultimately improving patient outcomes and healthcare decision- making. Following this thorough evaluation, Figure 3 visually demonstrates the methodology employed in our research, providing a comprehensive overview of the systematic approach adopted for assessing the performance of various machine learning classifiers in predicting heart disease.

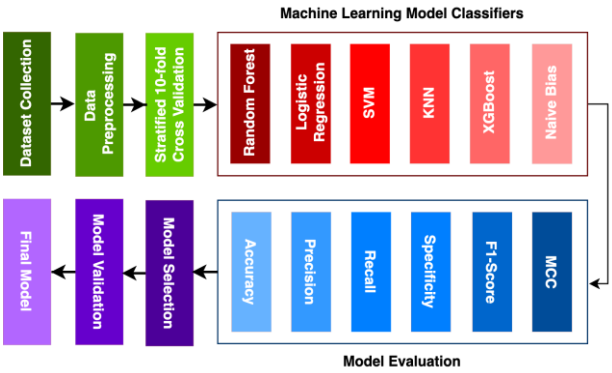

Figure 3: Flowchart of the Proposed Work

Experimental Analysis

We obtained the results after analyzing and recognizing the best classification algorithm. We conducted experiments on various scales to validate the results, utilizing the cross-validation and varying hyperparameter tuning. Table 2 shows the results of the current research.

Cross-Validation Technique

To ensure a robust evaluation of the machine learning models, a stratified 10-fold cross-validation approach was employed. This method involves dividing the dataset into 10 equal subsets, where nine subsets are used for training and one for testing, rotating the test set across all folds. Stratification ensures that each fold maintains the original class distribution, which is crucial for imbalanced datasets. The final performance metrics, including accuracy, precision, recall, and F1-score, were calculated as the average across all folds, reducing the

risk of overfitting and providing a reliable estimate of model performance.

Hyperparameter Tuning

Hyperparameter tuning was conducted to optimize Random Forest and XGBoost models. For XGBoost, the learning rate 0.01 to 0.3, maximum tree depth 3 to 10, and number of estimators 50 to 300 were tuned. Regularization parameters L1 and L2 were adjusted to mitigate overfitting. For Random Forest, n estimators 50 to 200, maximum features, and maximum tree depth were optimized. Cross-validation with grid search evaluated hyperparameter combinations. Optimal settings were selected based on performance metrics such as accuracy, precision, and F1-score, ensuring a balance between bias and variance. The tuning process substantially enhanced the predictive capabilities of both models, with XGBoost achieving the highest accuracy and demonstrating superior performance across all metrics.

Figure 4 shows the Receiver Operating Characteristic (ROC) curves, providing a visual representation of the classifier performance in terms of True Positive Rate (Sensitivity) against the False Positive Rate across various thresholds. This graphical representation enables a nuanced understanding of the trade-off between sensitivity and specificity for each machine learning classifier evaluated in the context of heart disease prediction.

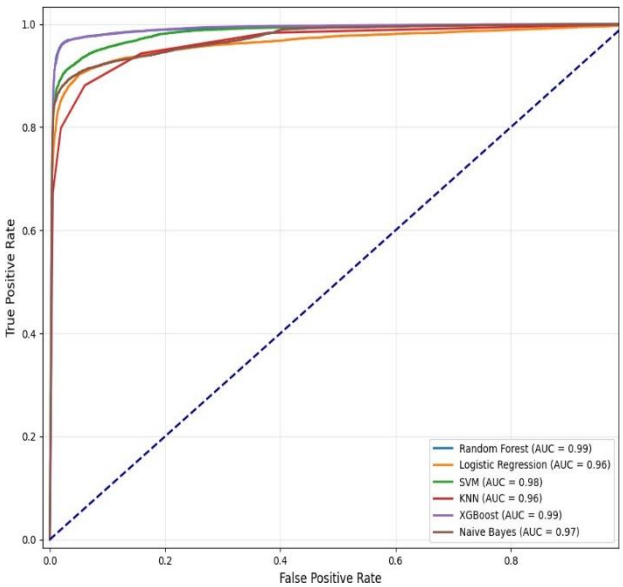

Figure 4: Comparison of ROC curves

Furthermore, Table 3 presents a visual comparison of related research, highlighting the performance of different classifiers reported in various studies. These figures collectively contribute to a clearer understanding of our research methodology and facilitate a comparative analysis with other relevant studies in the field.

Results and Discussion

Various machine learning classifiers were rigorously evaluated for heart disease prediction. The classifiers included Random Forest, Logistic Regression, SVM, KNN, XGBoost, and Naive Bayes. The performance metrics, such as accuracy, precision, recall, specificity, F1-score, and Matthew's correlation coefficient (MCC), were meticulously analyzed under a Stratified 10-fold cross-validation framework. Among these classifiers, Random Forest achieved an outstanding accuracy of 97.85%, demonstrating strong predictive capabilities. Logistic Regression exhibited commendable performance with an accuracy of 86.34%, show- casing its effectiveness in heart disease prediction. SVM, another powerful model, achieved an accuracy of 90.05%, underscoring its reliability in the context of cardiovascular health. KNN demonstrated competitive accuracy at 87.22%, suggesting its suitability for heart disease prediction. XGBoost, a boosting algorithm, excelled with a remarkable accuracy of 99.90%, making it a standout performer in the analysis. Naive Bayes, a probabilistic model, achieved an accuracy of 85.56%, showcasing its competence in the prediction task.

These results collectively provide valuable insights into the strengths of each algorithm, aiding in informed decision-making for the selection of an appropriate model in the context of heart disease prediction. We have successfully deployed our machine learning model within the Django framework, utilizing it to classify instances as either indicative of the presence or absence of heart disease. In cases where the model identifies the presence of heart disease, it is labeled as positive, while instances indicating the absence of heart disease are labeled as negative. Figures 5 and 6 show the visualization output of our machine learning model.

Limitations and Future Work

While this study shows XGBoost's potential for heart disease prediction, it has limitations. Relying on a single Kaggle dataset limits generalizability to real-world scenarios. Validation on real-world datasets was beyond the scope due to cost constraints. Future research should validate the model with diverse, external datasets.

Conclusion

The findings of this research contribute to the understanding of each algorithm's effectiveness in heart disease prediction, offering valuable guidance for

healthcare professionals and researchers. The systematic evaluation of these classifiers, considering multiple performance metrics, ensures a comprehensive assessment. The adoption of advanced machine learning techniques, as demonstrated in our study, holds promise for enhancing the accuracy and reliability of heart disease prediction models. As we move forward, these insights pave the way for informed decision-making in selecting the most suitable classifier for specific healthcare applications.

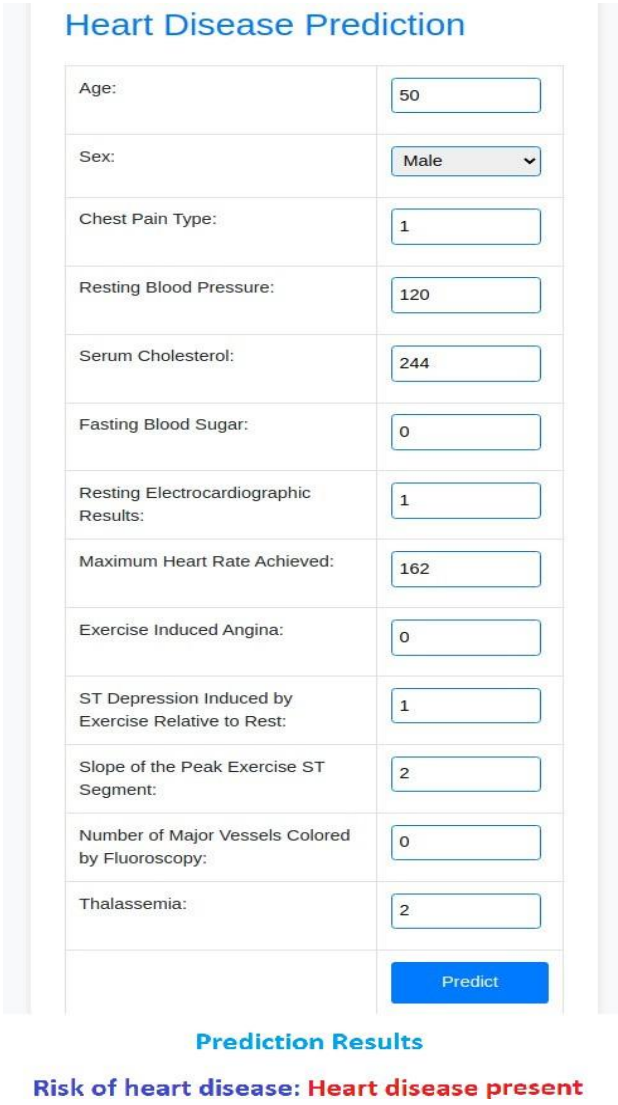

Figure 5: Model Deployed in Django (Positive)

Early detection and treatment of heart disease will reduce healthcare costs and improve patient outcomes with machine learning. Predictive models empower target-oriented designs and efficiently allocate healthcare resources, promising lower-cost cardiovascular health prevention. Proposed machine learning models will support doctors in prioritizing interventions and allocating resources by identifying high-risk heart disease patients. Cardiovascular disease will be prevented or delayed by lifestyle changes or medication. It will save lives, lower healthcare costs, and improve population health. Future research could delve deeper into feature engineering, hyperparameter tuning, and ensemble methods to further optimize predictive models for heart disease prognosis.

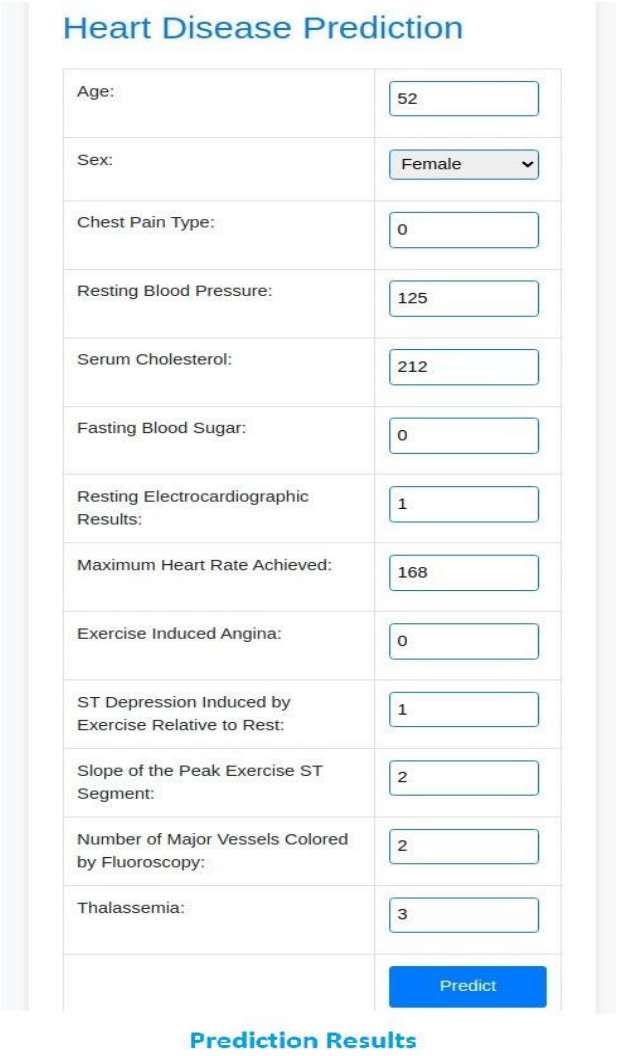

Figure 6: Model Deployed in Django (Negative)

XGBoost achieves 99.90% accuracy on a diverse dataset of 70,000 instances, surpassing previous studies. Advanced feature selection and interpretability methods, such as SHAP values, identified crucial risk factors. 10-fold stratified cross-validation ensured robustness. Unlike previous studies with limited datasets, the XGBoost model's efficacy was validated on a comprehensive dataset. These advancements enhance the study's relevance for practical heart disease prevention. Overall, our study provides a foundation for advancing the field of cardiovascular health prediction through the application of machine learning methodologies.

Table 1: Dataset Attributes

Feature	Description
age	Age of the patient
sex	Gender of the patient (1 = male; 0 = female)
cp	Chest pain type (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic)
trestbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar greater than 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0 = normal; 1 = ST-T wave abnormality; 2 = probable or definite left ventricular hypertrophy)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = downsloping)
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
target	Presence of heart disease (1 = yes; 0 = no)

Table 2: Classifier Performance

Classifier	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
Random Forest	97.85	97.87	97.85	97.85	97.85	95.72
Logistic Regression	86.34	86.65	86.34	86.34	86.29	72.94
SVM	90.05	90.27	90.05	90.05	90.02	80.29
KNN	87.22	87.31	87.22	87.22	87.21	74.50
XGBoost	99.90	99.89	99.88	99.91	99.93	99.81
Naive Bayes	85.56	85.82	85.56	85.56	85.51	71.34

Table 3: Comparison of Heart Disease Prediction

Authors	Method	Accuracy
Otoom et al. [39]	Naive Bayes	84.5%
	SVM	84.5%
	Functional trees	84.5%
Vembandasamy et al. [40]	Naive Bayes	86.419%
Chaurasia et al. [41]	J48	84.35%
	Bagging	85.03%
	SVM	94.60%
Parthiban et al. [42]	Naive Bayes	74%
Seema et al. [43]	Naive Bayes	95.556%
Kumar Dwivedi [24]	Naive Bayes	83%
	Classifcation tree	77%
	K-NN	80%
	Logistic regression	85%
	SVM	82%
	ANN	84%
proposed model	**Random Forest**	**97.85%**
	Logistic Regression	**86.34%**
	SVM	**90.05%**
	KNN	**87.22%**
	XGBoost	**99.90%**
	Naive Bayes	**85.56%**

Declarations

Funding: There is no funding or financial assistance available for this work, and it is not affiliated with any employment.

Competing Interests: The authors state that they have no competing interests.

Compliance with Ethical Standards: The authors state that there are no issues to demand compliance with ethical standards.

Research Data Policy and Data Availability Statements: The manuscript contains data with permission to use due to the open-access policy. Research data used in this work is available at the site: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

Authors' Contributions: ***Shafayat Bin Shabbir Mugdha*** conceptualized this work, wrote the codes, generated the figures, and wrote the main manuscript. ***Mahtab Uddin*** configured the method- ology, investigated this work, modified the figures, and finalized the manuscript. ***Hridoy Das*** checked the results, analyzed the figures, and modified the manuscript.

References

[1] A. S. Update, "Heart disease and stroke statistics—2020 update: a report from the american heart association," Circulation, vol. 141, no. 9, pp. e139–e596, 2020. https://doi.org/10.1161/CIR.0000000000000075

[2] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das et al., "Heart disease and stroke statistics—2019 update: a report from the american heart association," Circulation, vol. 139, no. 10, pp. e56–e528, 2019. https://doi.org/10.1161/CIR.0000000000000659

[3] D. M. Lloyd-Jones, Y. Hong, D. Labarthe, D. Mozaffarian, L. J. Appel, L. Van Horn, K. Greenlund, S. Daniels, G. Nichol, G. F. Tomaselli et al., "Defining and setting national goals for cardiovascular health promotion and disease reduction: the american heart association's strategic impact goal through 2020 and beyond," Circulation, vol. 121, no. 4, pp. 586–613, 2010. https://doi.org/10.1161/CIRCULATIONAHA.109.192703

[4] S. Yusuf, S. Hawken, S. Ounpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos et al., "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study," The lancet, vol. 364, no. 9438, pp. 937–952, 2004. https://doi.org/10.1016/S0140-6736(04)17018-9

[5] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature medicine, vol. 25, no. 1, pp. 44–56, 2019. https://doi.org/10.1038/s41591-018-0300-7

[6] R. Guo, R. Zhang, R. Liu, Y. Liu, H. Li, L. Ma, M. He, C. You, and R. Tian, "Machine learning-based approaches for prediction of patients' functional outcome and mortality after spontaneous intracerebral hemorrhage," Journal of Personalized Medicine, vol. 12, no. 1, p. 112, 2022. https://doi.org/10.3390/jpm12010112

[7] S. B. S. Mugdha, S. M. Ferdous, and A. Fahmin, "Evaluating machine learning algorithms for Bengali fake news detection," in 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020, pp. 1–6. https://doi.org/10.1109/ICCIT51783.2020.9392662

[8] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger et al., "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," Nature medicine, vol. 25, no. 1, pp. 70–74, 2019. https://doi.org/10.1038/s41591-018-0240-2

[9] K. W. Johnson, J. Torres Soto, B. S. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley, and J. T. Dudley, "Artificial intelligence in cardiology," Journal of the American College of Cardiology, vol. 71, no. 23, pp. 2668–2679, 2018. https://www.doi.org/10.1016/j.jacc.2018.03.521

[10] S. B. S. Mugdha, M. B. M. M. Kuddus, L. Salsabil, A. Anika, P. P. Marma, Z. Hossain, and S. Shatabda, "A gaussian naive Bayesian classifier for fake news detection in Bengali," in Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2. Springer, 2021, pp. 283–291. https://doi.org/10.1007/978-981-33-4367-2_28

[11] A. Mehmood, M. Iqbal, Z. Mehmood, A. Irtaza, M. Nawaz, T. Nazir, and M. Masood, "Prediction of heart disease using deep convolutional neural networks," Arabian Journal for Science and Engineering, vol. 46, no. 4, pp. 3409–3422, 2021. https://doi.org/10.1007/s13369-020-05105-1

[12] N. Gautam, P. Saluja, A. Malkawi, M. G. Rabbat, M. H. Al-Mallah, G. Pontone, Y. Zhang, B. C. Lee, and S. J.

Al'Aref, "Current and future applications of artificial intelligence in coronary artery disease," in Healthcare, vol. 10, no. 2. MDPI, 2022, p. 232. https://doi.org/10.3390/healthcare10020232

[13] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," Journal of the American Medical Informatics Association, vol. 24, no. 2, pp. 361–370, 2017. https://doi.org/10.1093/jamia/ocw112

[14] S. Sanchez-Martinez, O. Camara, G. Piella, M. Cikes, M. A´. Gonz´alez-Ballester, M. Miron, A. Vellido, E. G´omez, A. G. Fraser, and B. Bijnens, "Machine learning for clinical decision-making: challenges and opportunities in cardiovascular imaging," Frontiers in Cardiovascular Medicine, vol. 8, p. 765693, 2022. https://doi.org/10.3389/fcvm.2021.765693

[15] N. V. Chawla, "Machine learning for heart disease prediction: A review of existing approaches and future research directions," Expert Systems with Applications, vol. 150, p. 113214, 2020. https://doi.org/10.1016/j.eswa.2020.113214

[16] R. C. Deo, "Machine learning in medicine," Cardiovascular Diagnosis and Therapy, vol. 9, no. 4, pp. 495–502, 2019. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[17] K. A. Bhavsar, J. Singla, Y. D. A. Otaibi, O. Y. Song, Y. B. Zikria, A. K. Bashir., "Medical Diagnosis Using Machine Learning: A Statistical Review," Computers, Materials & Continua, vol. 67, no. 1, p. 107-125, 2021. https://doi.org/10.32604/cmc.2021.014604

[18] S. X. a. Z. Y. Shen, Wenqi, "Deep learning in medical image analysis," Annual Review of Biomedical Engineering, vol. 23, no. 1, pp. 289–316, 2021. https://doi.org/10.1146/annurev-bioeng-071516-044442

[19] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?" Heart, vol. 104, no. 14, pp. 1156–1164, 2018. https://doi.org/10.1136/heartjnl-2017-311198

[20] Z. Huang, W. Dong, H. Duan, J. Liu, "A Regularized Deep Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records," IEEE Transactions on Biomedical Engineering, vol. 65, no. 5, pp. 956-968, 2018. https://doi.org/10.1109/TBME.2017.2731158

[21] V. Patel, S. Azad, J. Keraenen, B. J. Hoogwerf, and R. Joshi, "Cost-effectiveness of machine learning for cardiovascular disease," Journal of the American College of Cardiology, vol. 77, no. 4, pp. 490– 499, 2021. https://doi.org/10.1016/j.jacc.2020.11.041

[22] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" PloS one, vol. 12, no. 4, p. e0174944, 2017. https://doi.org/10.1371/journal.pone.0174944

[23] M. Panahiazar, V. Taslimitehrani, N. Pereira, and J. Pathak, "Empowering personalized medicine with machine learning and big data: promises and challenges," in Advances in Personalized Medicine. Springer, 2020, pp. 235–271. https://doi.org/10.1109/BigData.2014.7004307

[24] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Computing and Applications, vol. 29, pp. 685–693, 2018. https://doi.org/10.1007/s00521-016-2604-1

[25] A. D'Souza, "Heart disease prediction using data mining techniques," International Journal of Research in Engineering and Science (IJRES) ISSN (Online), pp. 2320–9364, 2015.

[26] B. Venkatalakshmi and M. Shivsankar, "Heart disease diagnosis using predictive data mining," International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 3, pp. 1873–1877, 2014.

[27] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," AI magazine, vol. 13, no. 3, pp. 57–57, 1992. https://doi.org/10.1609/aimag.v13i3.1011

[28] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS international conference on computer systems and applications. IEEE, 2008, pp. 108–115. https://doi.org/10.1109/AICCSA.2008.4493524

[29] Y. Xing, J. Wang, Z. Zhao et al., "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," in 2007 International Conference on Convergence Information Technology (ICCIT 2007). IEEE, 2007, pp. 868–872. https://doi.org/10.1109/ICCIT.2007.146

[30] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," Heart Disease, vol. 7, no. 1, pp. 129–137, 2015. https://doi.org/10.090592/IJCSC.2016.018

[31] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," Health and Technology, vol. 10, pp. 1137–1144, 2020.

https://doi.org/10.1007/s12553-020-00440-7

[32] D. Chaki, A. Das, and M. Zaber, "A comparison of three discrete methods for classification of heart disease data," Bangladesh Journal of Scientific and Industrial Research, vol. 50, no. 4, pp. 293–296, 2015. https://doi.org/10.3329/bjsir.v50i4.25839

[33] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, 2019. https://doi.org/10.14569/IJACSA.2019.0100637

[34] H. B. F. David and S. A. Belcy, "Heart disease prediction using data mining techniques." ICTACT Journal on Soft Computing, vol. 9, no. 1, 2018. https://doi.org/10.21917/ijsc.2018.0254

[35] N. Bhatla and K. Jyoti, "An analysis of heart disease prediction using different data mining techniques," International Journal of Engineering, vol. 1, no. 8, pp. 1–4, 2012.

[36] Y. Gultepe and S. Rashed, "The use of data mining techniques in heart disease prediction," International Journal of Computer Science and Mobile Computing, vol. 8, no. 4, pp. 136–141, 2019.

[37] D. Dua, C. Graff et al., "Uci machine learning repository, 2017," URL http://archive.ics.uci.edu/ml, vol. 7, no. 1, 2019.

[38] S. Angirekula, "Heart.csv dataset," 2019, accessed: 2019-07-22. [Online]. Available: https://www.kaggle.com/datasets/srikalaangirekula/heartcsv/data

[39] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 143–156, 2015. https://dx.doi.org/10.14257/ijseia.2015.9.1.12

[40] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart diseases detection using Naive Bayes algorithm," International Journal of Innovative Science, Engineering & Technology, vol. 2, no. 9, pp. 441–444, 2015.

[41] V. Chaurasia and S. Pal, "Data mining approach to detect heart diseases," International Journal of Advanced Computer Science and Information Technology (IJACSIT), vol. 2, pp. 56–66, 2014.

[42] G. Parthiban and S. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25–30, 2012. https://doi.org/10.5120/ijais12-450593

[43] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," in 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). IEEE, 2016, pp. 381–386. https://doi.org/10.1109/ICATCCT.2016.7912028