Advanced AI Techniques for Cardiovascular Disease Forecasting Using Large-Scale Health Data

ACHUTHANANDA REDDY POLU[1], BHUMEKA NARRA[2],
DHEERAJ VARUN KUMAR REDDY BUDDULA[3], HARI HARA SUDHEER PATCHIPULUSU[4], NAVYA VATTIKONDA[5], ANUJ KUMAR GUPTA[6]

[1]Senior SDE, Cloudhub IT Solutions, USA
[2]Sr Software Developer, Statefarm, USA
[3]Software Engineer, Elevance Health Inc, USA
[4]Senior Software Engineer, Walmart, USA
[5]Business Intelligence Engineer, International Medical Group Inc, USA
[6]Senior Business Analyst, Sea Board Foods, USA

Abstract: The early and accurate prediction of CVDs is of great importance in order to make timely interventions and avoid life-threatening complications. The heart disease is forecasted using machine learning using the XGBoost (Extreme Gradient Boosting) algorithm in this study. It develops and test the model on a large heart disease dataset of 8,763 records with different features such as demographic info, medical history, clinical indicators and lifestyle habits. Missing value imputation, outlier removal, attribute reduction and feature scaling are done to extensive data preprocessing to enhance the model reliability. Key performance measures, such as accuracy, precision, recall, and F1-score, are used to compare the proposed XGBoost model to the existing classification techniques, K-Nearest Neighbors (KNN) and Naïve Bayes (NB). The results show how much better the proposed model performs than the existing models; its F1-score is 94.46%, accuracy is 92.38%, precision is 99.43%, and recall is 89.9%. The results of this confirm the ability of XGBoost for the prediction of cardiovascular disease and its promise for practical real-world healthcare applications of early diagnosis and risk assessment.

Keywords: Cardiovascular Disease Prediction, Machine Learning, XGBoost, Heart Disease Dataset.

Received: June 17, 2025. Revised: July 21, 2025. Accepted: August 8, 2025. Published: September 4, 2025.

1. Introduction

Hospitals and clinics have to offer patients premium quality of services at low costs for them, giving accurate diagnosis of medical cases and the best treatment options [1]. Early and accurate recognition of serious illnesses is necessary for preventing the outcomes of such illnesses that can lead to death in human life. Medical conditions that can become very dangerous are the cardiovascular diseases (CVDs), their mismanagement proving to be very dangerous to life. Because CVDs are caused by two or more factors such as diabetes, hypertension, and high cholesterol, plus abnormal heart rhythm, the discovery of these diseases becomes difficult [2]. Medical history analysis by healthcare providers allows to collect vital information on risk factor development of heart disease for patients.

Traditionally physicians match their clinical acumen together with electrocardiogram (ECG) and echocardiogram (ECHO) diagnostic results for cardiovascular condition predictions. The experience-based assessments of medical practitioners and their subjective judgment have proven to be dependent factors that affect the accuracy of diagnoses [3]. The expanding movement to use machine learning (ML) models represents a solution to these constraints because these models provide data-based systematic methods to predict diseases in a scalable fashion. The primary difficulty emerges from choosing an AI model which proves capable of achieving maximum accuracy predictability and reliability for patient results.

Through large dataset training machine learning systems learn to detect patterns which enable them to forecast information in unexplored data sets [4]. Training provided a system with thousands of images, enabling it to recognize photos of cats and photos not containing cats. Historical patient data can train machine learning models to determine cardiovascular disease risk in a similar fashion to identifying patterns for cats and non-cats [5]. The process of training and prediction uses specialized algorithms that learn from provided data to make accurate forecasts. The identification of CVD risk by analyzing patients' clinical indicators through NB and DT classifiers has proven successful in medical diagnosis applications.

These machine learning models achieve superior effectiveness when they use extensive health data resources that represent crucial elements in cardiovascular disease prediction. A vast array of medical data is accessible in a vast electronic format, comprising EHR records, wearable device information, MRI medical images, genetic sequences, and lifestyle information from large populations over long-term monitoring. With larger cardiovascular disease data collections, researchers can discover important patterns in the historic record of patients, along with biomarker activity and treatment response, and risk elements change [6]. Researchers are able to uncover new intricate patterns using huge and different types of healthcare data and develop precise forecasting models as well as individualized risk measurements using these. Using large-scale health data to its maximum potential becomes difficult because of data integration complexities as well as the need for standardization and patient privacy maintenance which ensures both trustworthy and ethical AI-driven prediction results.

Early diagnosis and timely intervention have become essential in reducing the impact of CVD in the current world at large and in order to improve patient outcomes. Traditional methods of diagnosing heart disease are often time-consuming and prone to human error, which highlights the need for an efficient, automated solution. The goal of this research is to increase prediction accuracy by utilizing ML techniques, namely the XGBoost classification algorithm of cardiovascular disease forecasting, enabling healthcare professionals to identify at-risk individuals with higher precision and confidence. This tactic might reduce the prevalence of cardiovascular diseases and greatly enhance clinical decision-making. The following are the study's main contributions:

- The study applies rigorous data preprocessing techniques on the heart disease dataset, enhancing data quality and ensuring more accurate predictions for cardiovascular disease forecasting.
- It identifies key features influencing cardiovascular risk, providing valuable insights for healthcare professionals to focus on critical factors for early diagnosis and prevention.
- Multiple ML models, including XGBoost, NB, and KNN, are evaluated using parameters including F1-score, recall, accuracy, and accuracy in order to select the ideal model.
- The proposed XGBoost model has better accuracy and reliability of the predicted values in forecasting cardiovascular illness, in contrast to current models.

1.1 Novelty and Justification

This study contributes to the novelty of this study through the reliability of the XGBoost classification technique is used to forecast the result, strengthening the model of cardiovascular disease utilizing a structured heart disease dataset and a robust data processing pipeline. In contrast to the conventional diagnostic methodology or the basic ML models, this study combines the attribute reduction, outlier removal, and data transformation techniques in order to enhance prediction accuracy. XGBoost is a good choice as it is proven to be effective in dealing with structured data and reducing overfitting, both of which are key in performing medical prediction tasks. This not only gives it a better diagnostic precision, but more importantly, this approach offers a scalable and interpretable framework of healthcare analytics.

1.2 Structure of the paper

The structure of the study is as follows: In Section II, the existing literature on cardiac disease prediction by the use of AI techniques is reviewed. Gathering, preprocessing, and model data development using XGBoost under the presented methodology are explained in Section III. The results and analysis from the proposed model in predicting cardiovascular disease are presented in Section IV. At last, Section V provides the conclusion and outlines potential directions for future work.

2. Literature Review

This section examines relevant research on the application of cutting-edge AI methods for Heart and cardiovascular disease prediction:

Maiga, Hungilo and Pranowo (2019) project's goal is to evaluate ML algorithms for predicting cardiovascular disease based on individuals' cardiovascular risk factors. Kaggle ML contests are the data source, and 70,000 patient records are included in the collection. The ML methods LR, KNN, NB, and RF were used in this study. The comparison's findings demonstrate that RF performs well in terms of classification accuracy, achieving 73%, 65% specificity, and 80% sensitivity. The model may be used to forecast cardiovascular disorders in the medical profession [7].

Alarsan and Younes (2019). The suggested solution is examined and verified using ML-libs and the Scala programming language on the Apache Spark framework using the baseline datasets for MIT-BIH Arrhythmia and MIT-BIH Supraventricular Arrhythmia. According to the findings, their approach obtained an overall accuracy of 97.98% and 96.75% using the GDB Tree technique with the RF approach for binary classification. GDB trees only enable binary classification, while RF achieved 98.03% accuracy for multiclass classification [8].

Haq et al. (2019) six ML classifiers and a Backpropagation Neural Network (BPNN) were evaluated for heart disease identification. Among the ML models, SVM with an RBF kernel showed the best results, achieving 88% accuracy on specific traits and 86% accuracy on all features, which further improved to 92.30% with ensemble learning. The BPNN achieved the highest classification accuracy of 93%, demonstrating that DL models, which automatically extract important features, outperform traditional ML classifiers for heart disease diagnosis [9].

Bernard et al. (2018) determine how effectively cutting-edge DL techniques can evaluate CMRI, including the ability to segment the myocardium and the two ventricles and categorize diseases. The results demonstrate that the best techniques accurately replicate the expert analysis, yielding a mean correlation score of 0.97 for automated clinical indicator extraction and an accuracy of 0.96 for autonomous diagnosis. These findings undoubtedly pave the way for completely automated and extremely accurate cardiac CMRI analysis [10].

Alić, Gurbeta and Badnjević (2017) present an overview of ML techniques in the classification of diabetes and CVD using ANN and BN. On the other hand, the NB network, a popular kind of BN, has the best accuracy rates for classifying diabetes and cardiovascular disease (CVD), with reflective accuracy rates of 99.51% and 97.92%, respectively [11].

Table I presents a summary of the literature review, highlighting the datasets, methodology, results analysis, and identified limitations for text dataset classification.

TABLE I. SUMMARY OF LITERATURE REVIEW FOR CARDIOVASCULAR DISEASE USING AI TECHNIQUES

References	Methodology	Dataset	Results Analysis	Limitations/Gaps	Future Work
Maiga, Hungilo, and Pranowo (2019)	Random Forest, Logistic Regression, k-nearest neighbor, and Naïve Bayes	Kaggle CVD dataset (70,000 patient records)	Random Forest achieved 73% accuracy, 65% specificity, and 80% sensitivity	Limited dataset scope; only traditional ML models used; no feature selection or tuning mentioned	Explore deep learning methods and ensemble models for higher accuracy

Alarsan and Younes (2019)	GDB Tree, Random Forest on Apache Spark using Scala	MIT-BIH Arrhythmia and Supraventricular Arrhythmia Databases	Binary classification: 97.98% accuracy (RF); Multiclass: 98.03% accuracy (RF); GDB Tree only binary	Limited to arrhythmia datasets; GDB Tree does not support multiclass; no deep learning comparison	Integrate deep learning and support for multiclass GDB Tree classification
Haq et al. (2019)	SVM (RBF kernel), Ensemble Learning, BPNN (Backpropagation Neural Network)	Not specified, but implies a standard heart disease dataset	SVM: 88% (selected features), Ensemble: 92.30%, BPNN: 93% accuracy	Dataset source not mentioned; performance comparison lacks clinical validation	Validate results with clinical datasets; implement hybrid deep learning architectures
Bernard et al. (2018)	Deep learning for cardiac MRI segmentation and classification	Cardiac MRI data (CMRI)	Correlation of 0.97 for clinical index extraction, 0.96 diagnostic accuracy	Focused only on image data; generalization to other data types (e.g., EHR) untested	Apply to multimodal data; integrate with electronic health records
Alić, Gurbeta and Badnjević (2017)	Artificial Neural Networks, Naive Bayesian Networks	Not specified; applied to diabetes and CVD data	BN: 99.51% accuracy (diabetes), 97.92% (CVD)	Dataset source and size unclear; lacks implementation detail and validation	Test across larger and more diverse populations; evaluate real-time applicability.

3. Methodology

The purpose of this work is to employ ML techniques to create a prediction model for cardiovascular disease. Using the heart illness dataset, the algorithm seeks to precisely assess and categories the risk of heart disease. The collection of the heart disease dataset and meticulous data preparation are the initial stages in the suggested approach to heart disease prediction. In an effort to improve data quality, this stage includes operations like resolving missing values, removing outliers, reducing attributes, and transforming the data. Following refinement, the dataset is separated into test and training sets. With the help of the XGBoost classification algorithm, which is renowned for its effectiveness when working with structured data, a model for prediction has been created. The model's performance is assessed using the F1 score, recall, accuracy, and precision and determine how effective the prediction system is the suggested flowchart for conducting cardiovascular disease prediction analysis using an ML framework, on Figure 1 below, shows the dataset for cardiac disease.

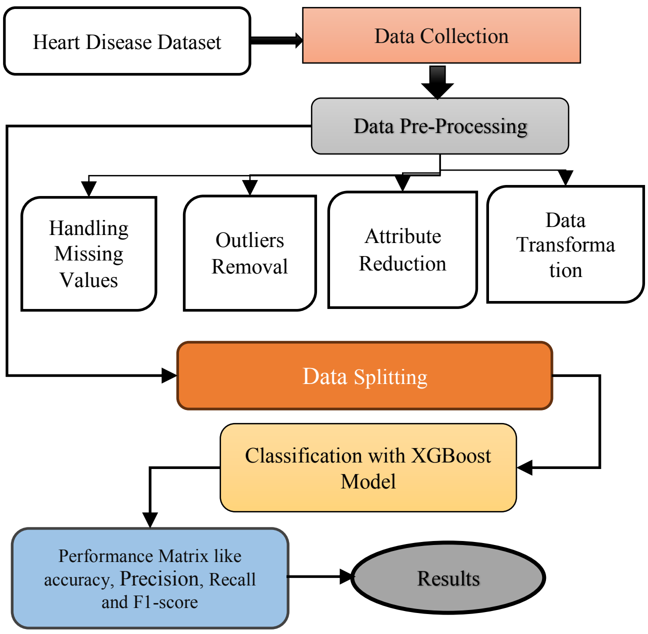

Fig. 1. Proposed Flowchart For Cardiovascular Disease Forecasting Based On ML Techniques

This paper explains how the steps of the proposed flowchart for cardiovascular disease using ML techniques on heart disease dataset:

3.1 Data Collection

The heart disease dataset includes several variables related to cardiovascular health and lifestyle choices, such as age, gender, blood pressure, cholesterol, heart rate, alcohol use, smoking, obesity, diabetes, and family history. Factors related to lifestyle, which include hours spent sedentary, stress levels, eating patterns, and exercise duration, make up the second group of variables analyzed. Triglyceride levels, drug usage, and past heart problems are among the medical considerations that are taken into consideration. In the second, income is taken into account along with other elements, including geographic characteristics, such as country, continent, and hemisphere. 8763 patient records from various global locations are included. An essential binary classification characteristic that shows whether a heart attack risk exists or not is included. Research and predictive analysis in the area of cardiovascular health can benefit greatly from this dataset. For study and prediction analysis regarding heart health, this dataset (refer to Table II) is an invaluable tool.

TABLE II. DESCRIPTION OF HEART DISEASE DATASET

Index	Feature	Description	Value Type
1	Age	The patient's age	Numerical
2	Sex	The patient's gender	(Male/Female)
3	Cholesterol	Patient's cholesterol levels	Numerical
4	Blood Pressure	Systolic and diastolic blood pressure measurements for the patient	Numerical
5	Heart Rate	The patient's heart rate	Numerical
6	Diabetes	If the patient has diabetes	(Yes/No)
7	Family History	A family history of cardiac issues	(1: Yes, 0: No)
8	Smoking	The patient's smoking status	(1: Smoker, 0: Non-smoker)
9	Obesity	The patient's state of obesity	(1: Obese, 0: Not obese)
10	Alcohol Consumption	The patient's level of alcohol consumption	(None/ Light/ Moderate/ Heavy)
11	Exercise Hours Per Week	The weekly number of hours spent exercising	Numerical
12	Diet	Patient's dietary practices	(Healthy/ Average/ Unhealthy)
13	Previous Heart Problems	Prior cardiac issues with the patient	(1: Yes, 0: No)
14	Medication Use	Patient use of medication	(1: Yes, 0: No)
15	Stress Level	The patient's stated level of stress	(1–10)
16	Sedentary Hours Per Day	Sedentary activity hours per day	Numerical
17	Income	The patient's income level	Numerical
18	BMI	Patient's Body Mass Index (BMI)	Numerical
19	Triglycerides	Patient's triglyceride levels	Numerical
20	Physical Activity Days Per Week	Physical exercise days per week	Numerical
21	Sleep Hours Per Day	Sleep duration in hours	Numerical
22	Country	The patient's nation	Numerical
23	Continent	The patient's residence continent	Numerical
24	Hemisphere	The patient's home hemisphere	Numerical
25	Heart Attack Risk	Risk of a heart attack	(1: Yes, 0: No)

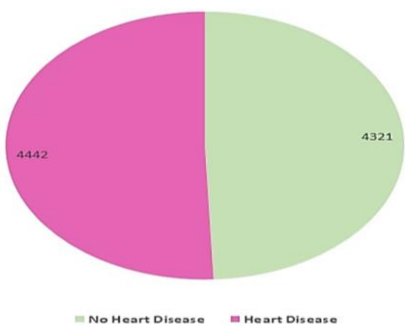

Fig. 2. Frequency Distribution of Classes in the Heart Disease Dataset.

The dataset's class distribution is seen in Figure 2 with respect to heart disease diagnosis. Out of the total instances, 4,442 individuals (represented in pink) are diagnosed with heart disease, while 4,321 individuals (shown in light green) are not. The distribution is nearly balanced, with a slight predominance of heart disease cases. This balance is crucial for training ML models, as it helps prevent bias toward either class and ensures more reliable performance in classification tasks.

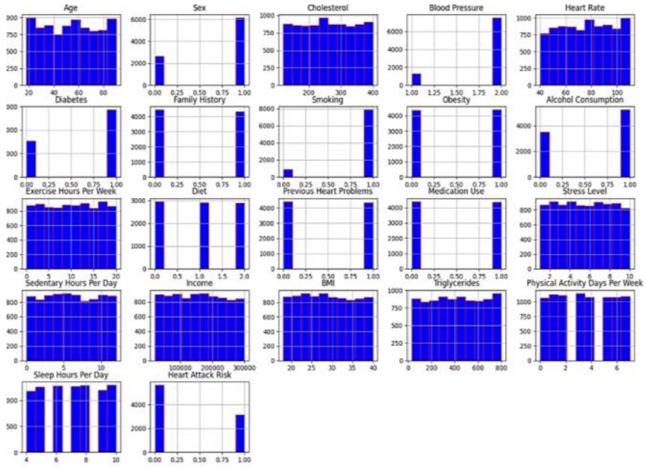

Fig. 3. The Distribution of the Heart Dataset Features.

Figure 3 displays histograms of key features related to heart disease, including demographics, clinical metrics, lifestyle habits, and medical history. It shows varied distributions some features like age and cholesterol are spread evenly, while others like smoking, diabetes, and previous heart issues are skewed. These plots help highlight the data characteristics relevant to predicting heart attack risk.

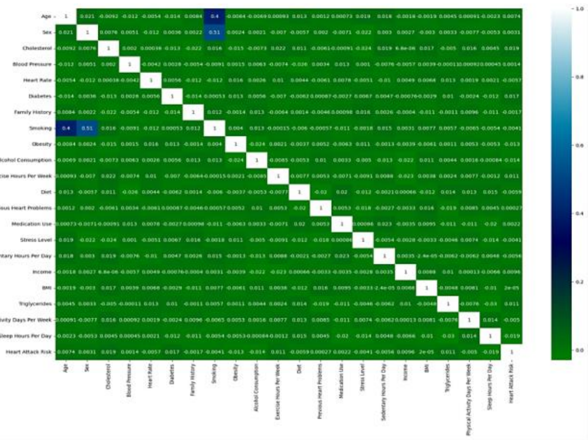

Fig. 4. The Relationships And Patterns Within The Heart Disease Dataset

The heatmap displays Pearson correlation coefficients among the dataset features seen in Figure 4, with most values ranging between –0.1 and 0.1, indicating weak or negligible linear relationships. For example, cholesterol and blood pressure show a slight positive correlation (~0.15), while variables like age and exercise hours exhibit near-zero correlation. This overall low correlation suggests minimal redundancy and highlights that most features contribute unique information[12]. It is advantageous for creating heart attack risk prediction models that work.

3.2 Data preprocessing

Pre-processing data is an essential step for enhancing model performance before applying ensemble ML techniques. The dataset used in this investigation was pre-processed using methods such as resampling and discretization to improve data quality and ensure more accurate and reliable model outcomes. The heart disease dataset was prepared using the following pre-processing techniques:

- **Handling Missing Values:** As with the data set, finding features with missing values in the data set was essential (null entries in key attributes such as employee satisfaction levels) and to solve them employing suitable imputation techniques (such mean or median imputation) to ensure the data is accurate and comprehensive.
- **Outliers Removal:** Outliers are identified in key features, i.e., values that are much higher or lower than other values in variables such as, work hours or duration of time at the company, and specific outlier treatment techniques, like capping or transformation [13] applied to minimize the impact of outliers on the analysis while retaining the key information.
- **Attribute Reduction:** Attribute reduction is the process of reducing a data set's attribute count, which allows the useful information needed for classifying or predicting the system to be obtained while dropping or removing the unnecessary or redundant attributes[14]. can say it is used generally for the reduction of model inefficiency, overfitting and generalization, especially with high-dimensional data.
- **Data Transformation using Scaling:** Data Transformation with Scaling is the procedure for changing the data values in so that they fall in the same scale. Especially when dealing with features of different magnitudes or units, scaling is very useful in a way that can speed up many ML algorithms, especially those that rely on distance metrics or optimization approaches.

3.3 Data splitting

In data mining, preparing, organizing, and cleaning raw data is known as data preprocessing so that it may be used to create and train ML models[15]. Initially, dividing it is a usual practice to create two separate subsets from the dataset: a testing set and a training set. Usually, 40% of the data is reserved for verification and 60% is utilized to train the model.

3.4 Classification of XGBoost Classifier

It uses a collection of different DT (weak learners) to independently determine similarity ratings. The method alters the gradient descent and regularization process to solve the issue of overfitting during the training phase. Overfitting may be managed with the use of regularization, which XGBoost offers. To do this, each tree's weights and biases are subject to L1/L2 penalties [12].

In order to avoid overfitting, the two distinct components of XGBoost's objective function combine to capture the regularization term and the model's deviation when optimization techniques are used. The dataset may be represented as $D=\{(xi,yi)\}$, which includes a considerable number of n samples and m characteristics. With this information, an additive model that incorporates many fundamental models can be used to describe the predictive variable [39]. The results of making sample predictions may be summed up as follows in Equations (1) (2):

$$\hat{y}_i = \sum_{k=1}^{k} f_{k(x_i), f_k} \in \varphi \tag{1}$$

$$\varphi = \{f(x) = w_s(x)\}\ (s: R^m \rightarrow T, w_s \in R^T) \tag{2}$$

When $\hat{y}\ i$ represents the label prediction and xi is one of the samples, the projected score is $fk(xi)$ or the sample that was provided. When As an added bonus, it represents the collection of regression trees that include the parameters for the tree structure of s,$f(x)$. On top of that, w stands for both the number of leaves and their approximate weight.

3.5 Performance Metrics

To assess the effectiveness of cardiovascular disease forecasting, several evaluation metrics, also known as performance metrics, are utilized. These measures give a thorough picture of how well the model predicts the risk of cardiovascular illnesses[16]. Boosting models' efficacy in identifying erroneous or accidentally labelled cardiac disease is assessed using the confusion matrix displayed below. It uses four criteria to compare anticipated values with actual results: FN, FP, TP, and TN. The following metrics are defined below:

1) Accuracy

The most often used a performance indicator called accuracy shows what proportion of forecasts were correct, taking into account both positive and negative outcomes, out of the overall quantity of projections in Equation (3).

$$\text{Accuracy} = \frac{\text{TP+TN}}{(\text{TP + TN + FP + FN})} \tag{3}$$

2) Precision

The important cases concerning the occurrences that were obtained, such as accuracy. Below is the precision Equation (4):

$$Precision = \frac{(TP)}{(TP)+(FP)} \tag{4}$$

3) Recall

Recall is a small proportion of pertinent examples that are recovered out of the entire number of examples that are pertinent. The recall Equation (5) may be found below:

$$Recall = \frac{(TP)}{(TP)+ (Fn)} \tag{5}$$

4) F1-Score

The double precision times recall provides the basis for the F-measure, which is then split by the total of accuracy and recall. The F-Measure Equation (6) is as follows:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

4. Result Analysis And Discussion

The results and evaluations of the research study are presented in this part. A computer with an Intel(R) Core i7 CPU with Windows 10 and 16GB of RAM was used for each experiment. Python was used to create the cardiovascular disease prediction models, using necessary libraries including Matplotlib, NumPy, Pandas, and Scikit-learn. Existing models like NB are contrasted with the suggested XGBoost model [17] and KNN [18], as seen in Table IV. The performance results are shown in Table III, highlighting the superiority of the XGBoost model in forecasting cardiovascular disease using the heart disease dataset.

TABLE III. RESULTS OF XGBOOST MODEL ON CARDIOVASCULAR FORECASTING USING HEART DISEASE DATASET.

Performance Matrix	XGBoost (XGB)
Accuracy	92.38
Precision	99.43
Recall	89.9
F1-score	94.46

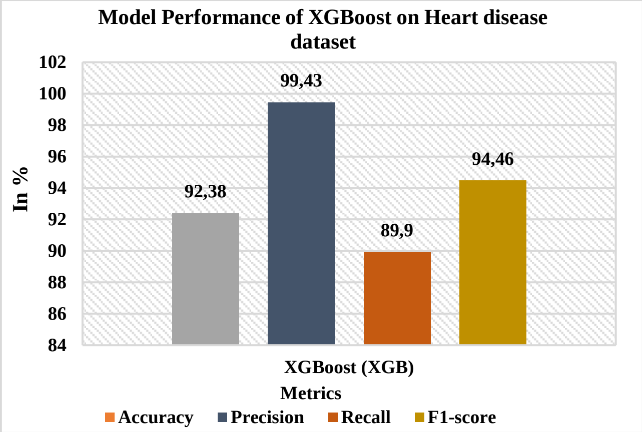

Fig. 5. Model Performance of XGBoost

The model proves highly effective for classification use and is shown in Figure 5. The system demonstrated excellent predictive accuracy of 92.38%, which proved most of its predictions to be correct. There are very few false positives since the model demonstrates an outstanding precision of 99.43%. The 89.9% recall shows that the system detects most actual positive cases, with the exception of certain cases, it fails to identify. The 94.46% balanced F1-score demonstrates that the model provides dependable results by effectively balancing precision with recall in order to support accurate predictive operations.

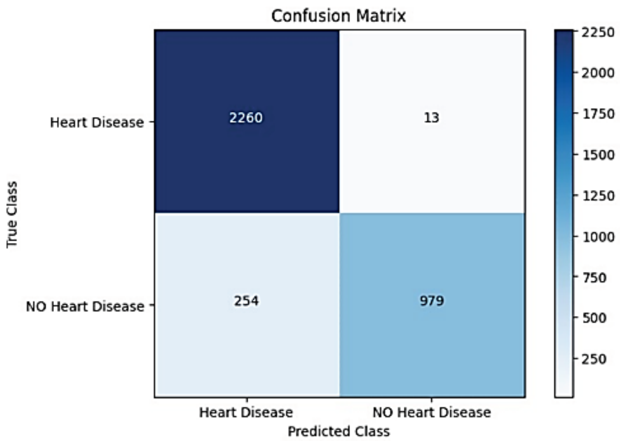

Fig. 6. Confusion matrix for XGBoost

The performance matrix, displayed in Figure 6, summarizes the XGB model's classification findings for heart disease diagnosis. The model's 92.38% accuracy rate shows that its forecasts were generally quite accurate. With a 99.43% accuracy rate, the model shows great dependability in accurately detecting cases of heart disease with little false positives. The detection percentage of 89.9% indicates the system identifies most genuine heart disease occurrences. The F1-score value of 94.46% verifies the model's capacity to manage high detection sensitivity along with accurate predictions in heart disease diagnosis.

TABLE IV. COMPARATIVE ANALYSIS OF MODEL PERFORMANCE ON THE HEART DISEASE DATASET

Measures	XGBoost	Naïve Bayes[17]	KNN[18]
Accuracy	92.38	86.40	87
Precision	99.43	86.40	93.02
Recall	89.9	86.40	80
F-measure	94.46	86.40	85.91

The comparative analysis, as presented in Table IV highlights demonstrates the proposed XGBoost model outperforms the existing NB and KNN models across all evaluation metrics. XGBoost delivered the highest accuracy of 92.38%, beyond NB 86.40% and KNN 87%. Its precision outcome of 99.43% showed better performance than the current available precision levels of other models. The proposed model achieved recall results of 89.9%, but NB reached 86.40%, and KNN reached 80%. The F-measure of XGBoost reached 94.46%, which validated its dependability and reliability when using the heart disease dataset for cardiovascular disease prediction.

5. Conclusion And Future Work

A predictive model based on the XGBoost classification algorithm forecasts cardiovascular disease risks effectively than classical methods, including NB and KNN. Typical data preprocessing methods coupled with a large heart disease dataset allowed the proposed model to reach an accuracy level of 92.38% while maintaining precision at 99.43% and recall of 89.9% with a 94.46% F1-Score. The model shows reliable performance in detecting heart disease patients, which provides useful clinical support in uncovering early disease signs.

In order to identify temporal and geographical trends in large medical datasets, future research should enhance the model by integrating DL with CNN or RNN. The prediction accuracy would increase through Data from wearable technology and EHRs for real-time patient monitoring, whereas continuous health assessment would become possible. Extending the data with more samples from diverse geographical regions and demographic populations alongside the implementation of explainable AI techniques will enhance both the general accessibility of the model to clinicians and its clinical adoption.

References

[1]. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” *Mob. Inf. Syst.*, 2018, doi: 10.1155/2018/3860146.

[2]. Y. Ruan *et al.*, “Cardiovascular disease (CVD) and associated risk factors among older adults in six low- and middle-income countries: Results from SAGE Wave 1,” *BMC Public Health*, 2018, doi: 10.1186/s12889-018-5653-9.

[3]. G. D. Flora and M. K. Nayak, “A Brief Review of Cardiovascular Diseases, Associated Risk Factors and Current Treatment Regimes,” *Curr. Pharm. Des.*, 2019, doi: 10.2174/1381612825666190925163827.

[4]. N. Kagiyama, S. Shrestha, P. D. Farjo, and P. P. Sengupta, “Artificial intelligence: practical primer for clinical research in cardiovascular disease,” *Journal of the American Heart Association*. 2019. doi: 10.1161/JAHA.119.012788.

[5]. T. R. Einarson, A. Acs, C. Ludwig, and U. H. Panton, “Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of

scientific evidence from across the world in 2007–2017," *Cardiovasc. Diabetol.*, vol. 17, no. 1, p. 83, Dec. 2018, doi: 10.1186/s12933-018-0728-6.

[6]. A. C. Dimopoulos *et al.*, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 179, Dec. 2018, doi: 10.1186/s12874-018-0644-1.

[7]. J. Maiga, G. G. Hungilo, and Pranowo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data," in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, Oct. 2019, pp. 45–48. doi: 10.1109/ICIMCIS48181.2019.8985205.

[8]. F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0244-x.

[9]. A. U. Haq *et al.*, "Identifying the Predictive Capability of Machine Learning Classifiers for Designing Heart Disease Detection System," in *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, IEEE, Dec. 2019, pp. 130–138. doi: 10.1109/ICCWAMTIP47768.2019.9067519.

[10]. O. Bernard *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?," *IEEE Trans. Med. Imaging*, 2018, doi: 10.1109/TMI.2018.2837502.

[11]. B. Alić, L. Gurbeta, and A. Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases," in *2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings*, 2017. doi: 10.1109/MECO.2017.7977152.

[12]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785.

[13]. A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier detection approaches for wireless sensor networks: A survey," *Comput. Networks*, 2017, doi: 10.1016/j.comnet.2017.10.007.

[14]. C. Wang, Y. Shi, X. Fan, and M. Shao, "Attribute reduction based on k-nearest neighborhood rough sets," *Int. J. Approx. Reason.*, vol. 106, pp. 18–31, Mar. 2019, doi: 10.1016/j.ijar.2018.12.013.

[15]. A. S. Al-Mudimigh, Z. Ullah, and F. Saleem, "A Framework of an Automated Data Mining Systems Using ERP Model," *Int. J. Comput. Electr. Eng.*, pp. 651–655, 2009, doi: 10.7763/IJCEE.2009.V1.101.

[16]. M. L. King and R. J. Thomas, "Performance Measures for Cardiac Rehabilitation—An Opportunity to Bridge an Important Treatment Gap," *US Cardiol. Rev.*, vol. 4, no. 2, pp. 60–62, Dec. 2007, doi: 10.15420/usc.2007.4.2.60.

[17]. S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," *World J. Eng. Technol.*, 2018, doi: 10.4236/wjet.2018.64057.

[18]. R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," in *2019 2nd International Conference on New Trends in Computing Sciences, ICTCS 2019 - Proceedings*, 2019. doi: 10.1109/ICTCS.2019.8923053.

[19]. R. Swamidurai, U. Kannan, A. Raglin, and L. Scott, "Smart City Internet of Things Reasoning System for Emergency Responders," in *Conference Proceedings - IEEE SOUTHEASTCON*, 2019. doi: 10.1109/SoutheastCon42311.2019.9020580.

[20]. Bodepudi, V., & Chinta, P. C. R. (2024). Enhancing Financial Predictions Based on Bitcoin Prices using Big Data and Deep Learning Approach. *Available at SSRN 5112132*.

[21]. Chinta, P. C. R. (2023). The Art of Business Analysis in Information Management Projects: Best Practices and Insights. *DOI*, *10*.

[22]. Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. *Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures*.

[23]. Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *American Journal of Computing and Engineering*, *4*(2), 35-51.

[24]. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V., & Maka, S. R. (2025). Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering. *European Journal of Applied Science, Engineering and Technology*, *3*(2), 41-54.

[25]. Moore, C. (2024). Enhancing Network Security With Artificial Intelligence Based Traffic Anomaly Detection In Big Data Systems. Available at SSRN 5103209.

[26]. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., & Bodepudi, V. (2025). Predictive Analytics for Disease Diagnosis: A Study on Healthcare Data with Machine Learning Algorithms and Big Data. *J Cancer Sci*, *10*(1), 1.

[27]. Chinta, P. C. R., Jha, K. M., Velaga, V., Moore, C., Routhu, K., & SADARAM, G. (2024). Harnessing Big Data and AI-Driven ERP Systems to Enhance Cybersecurity Resilience in Real-Time Threat Environments. *Available at SSRN 5151788*.

[28]. Chinta, P. C. R. (2023). Leveraging Machine Learning Techniques for Predictive Analysis in Merger and Acquisition (M&A). *Journal of Artificial Intelligence and Big Data*, *3*(1), 10-31586.

[29]. Chinta, P. C. R. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimisation Strategies. *Journal of Artificial Intelligence & Cloud Computing*, *1*(4), 10-47363.

[30]. Chinta, P. C. R., & Karaka, L. M. AGENTIC AI AND REINFORCEMENT LEARNING: TOWARDS MORE AUTONOMOUS AND ADAPTIVE AI SYSTEMS.

[31]. Sadaram, G., Karaka, L. M., Maka, S. R., Sakuru, M., Boppana, S. B., & Katnapally, N. (2024). AI-Powered Cyber Threat Detection: Leveraging Machine Learning for Real-Time Anomaly Identification and Threat Mitigation. MSW Management Journal, 34(2), 788-803.

[32]. Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.

[33]. Sadaram, G., Sakuru, M., Karaka, L. M., Reddy, M. S., Bodepudi, V., Boppana, S. B., & Maka, S. R. (2022). Internet of Things (IoT) Cybersecurity Enhancement through Artificial Intelligence: A Study on Intrusion Detection Systems. *Universal Library of Engineering Technology*, (2022).

[34]. Jha, K. M., Velaga, V., Routhu, K. K., Sadaram, G., & Boppana, S. B. (2025). Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector. *J Cardiobiol*, *9*(1), 1.

[35]. Maka, S. R. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Available at SSRN 5116707*.

[36]. Karaka, L. M. (2021). Optimising Product Enhancements Strategic Approaches to Managing Complexity. *Available at SSRN 5147875*.

[37]. KishanKumar Routhu, A. D. P. Risk Management in Enterprise Merger and Acquisition (M&A): A Review of Approaches and Best Practices.

[38]. Routhu, KishanKumar & Katnapally, Niharika & Sakuru, Manikanth. (2023). Machine Learning for Cyber Defense: A Comparative Analysis of Supervised and Unsupervised Learning Approaches. Journal for ReAttach Therapy and Developmental Diversities. 6. 10.53555/jrtdd.v6i10s(2).3481.

[39]. Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1)*, *191*, 2-7.

[40]. Chinta, Purna Chandra Rao & Moore, Chethan Sriharsha. (2023). Cloud-Based AI and Big Data Analytics for Real-Time Business Decision-Making. 36. 96-123. 10.47363/JAICC/2023.

[41]. Kuraku, D. S., & Kalla, D. (2023). Phishing Website URL's Detection Using NLP and Machine Learning Techniques. *Journal on Artificial Intelligence-Tech Science*.

[42]. Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.

[43]. Jha, K. M., Velaga, V., Routhu, K., Sadaram, G., Boppana, S. B., & Katnapally, N. (2025). Transforming Supply Chain Performance Based on Electronic Data Interchange (EDI) Integration: A Detailed Analysis. *European Journal of Applied Science, Engineering and Technology*, *3*(2), 25-40.

[44]. Kuraku, D. S., & Kalla, D. (2023). Phishing Website URL's Detection Using NLP and Machine Learning Techniques. *Journal on Artificial Intelligence-Tech Science*.

[45]. Jha, K. M., Velaga, V., Routhu, K. K., Sadaram, G., & Boppana, S. B. (2025). Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector. *J Cardiobiol*, *9*(1), 1.

[46]. KishanKumar Routhu, A. D. P. Risk Management in Enterprise Merger and Acquisition (M&A): A Review of Approaches and Best Practices.

[47]. Kalla, D., Mohammed, A. S., Boddapati, V. N., Jiwani, N., & Kiruthiga, T. (2024, November). Investigating the Impact of Heuristic Algorithms on Cyberthreat Detection. In *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (Vol. 1, pp. 450-455). IEEE.

[48]. Bodepudi, V. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Journal of Artificial Intelligence and Big Data*, *3*(1), 10-31586.

[49]. Kalla, D., Smith, N., & Samaah, F. (2025). Deep Learning-Based Sentiment Analysis: Enhancing IMDb Review Classification with LSTM Models. *Available at SSRN 5103558*.

[50]. Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

[51]. Kalla, D., Samaah, F., Kuraku, S., & Smith, N. (2023). Phishing detection implementation using databricks and artificial Intelligence. *International Journal of Computer Applications*, *185*(11), 1-11.

[52]. Boppana, S. B., Moore, C. S., Bodepudi, V., Jha, K. M., Maka, S. R., & Sadaram, G. AI And ML Applications In Big Data Analytics: Transforming ERP Security Models For Modern Enterprises.

[53]. Sreeramulu, M. D., Mohammed, A. S., Kalla, D., Boddapati, N., & Natarajan, Y. (2024, September). AI-driven Dynamic Workload Balancing for Real-time Applications on Cloud Infrastructure. In *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)* (Vol. 7, pp. 1660-1665). IEEE.

[54]. Kalla, D., & Samaah, F. (2023). Exploring Artificial Intelligence And Data-Driven Techniques For Anomaly Detection In Cloud Security. *Available at SSRN 5045491*.

[55]. Kalla, D., Smith, N., & Samaah, F. (2023). Satellite Image Processing Using Azure Databricks and Residual Neural Network.