International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Prediction of Heart Disease Using Xgboost Algorithm

V.Mohan Lakshmi Narayana[1], Y.Siva Ram[2], D.Lakshmi Ganapathi[3], N.Satya Nikhli[4], P C S Nagendra Setty[5].

[1,2,3,4]Students, [5]Assistant Professor
Department of Computer Science and Engineering
Sasi Institute of Technology & Engineering, Tadepalligudem.
[1]narayana.veeramallu@sasi.ac.in , [2]sivaram.yellina@sasi.ac.in , [3]dodda.ganapathi@sasi.ac.in ,
[4]satyanikhil.nandigam@sasi.ac.in , [5]pcsns@sasi.ac.in .

Abstract

Heart disease remains a major global health concern, making early diagnosis crucial for reducing mortality rates. This study focuses on developing a predictive model using the XGBoost algorithm, optimized through Optuna hyperparameter tuning, and ANOVA-based feature selection for identifying the most important clinical indicators. The Framingham Heart Study dataset is utilized, incorporating key cardiovascular risk factors such as age, blood pressure, cholesterol, BMI, and glucose levels. The approach includes ANOVA-based feature selection to identify the most relevant predictors and XGBoost with Optuna hyperparameter tuning to enhance predictive accuracy.

To improve model performance, data preprocessing techniques such as handling missing values, applying Standard Scaler for feature scaling, and addressing class imbalance using Synthetic Minority Over-sampling Technique (SMOTE) were implemented. The optimized model demonstrates high accuracy, making it a reliable tool for risk assessment. By leveraging advanced machine learning techniques, this model can assist healthcare professionals in making informed decisions, ultimately aiding in early detection and preventive care for heart disease. The final XGBoost classifier achieved a good accuracy score, demonstrating its effectiveness in predicting heart disease risk. To make this model accessible for real-world use, an interactive web application was developed allowing users to input clinical parameters and receive instant heart disease risk predictions.

Keywords: Heart Disease Prediction, Machine Learning, XGBoost, ANOVA, Optuna, SMOTE.

1. Introduction

Heart disease is one of the leading causes of mortality worldwide, accounting for millions of deaths each year. The early diagnosis of heart-related conditions plays a crucial role in reducing the risk of fatal complications. Traditional diagnostic methods rely on medical examinations and laboratory tests, which can be time-consuming and may not always provide accurate risk assessments. With the advancements in machine learning and data analytics, predictive models have gained significant attention for their ability to analyse large datasets and identify potential risk factors efficiently.

In this study, we utilize machine learning techniques to predict heart disease risk using the Framingham dataset, a widely used dataset containing various cardiovascular risk factors. The proposed approach incorporates ANOVA-based feature selection to determine the most influential attributes in predicting heart disease. The XGBoost algorithm, known for its efficiency in classification tasks, is employed as the predictive model. Additionally, Optuna hyperparameter tuning is utilized to enhance model accuracy by selecting the most optimal configuration of parameters.

To improve the reliability of predictions, data preprocessing techniques such as handling missing values, feature scaling using Standard Scaler, and class imbalance correction using SMOTE are implemented. These steps ensure that the model is trained on a well-processed dataset, leading to improved performance and generalization. By leveraging advanced machine learning methodologies, this study aims to develop a robust and highly accurate predictive system that can assist in early diagnosis and risk assessment of heart disease.

2. Related Work

Based on phonocardiogram (PCG) datasets, previous studies applied various machine learning (ML) techniques to detect congenital valvular murmurs. Features with low correlation to murmur presence were discarded, and signal transformations such as Mel-frequency cepstral coefficients (MFCC), power spectral density, and Hilbert transform were applied for feature extraction. Some studies implemented heart sound segmentation to isolate systolic and diastolic phases, while others leveraged deep learning models such as convolutional neural networks (CNN) and recurrent neural networks (RNN) for automated murmur classification. The authors utilized 10-fold cross-validation and various feature engineering approaches to improve model performance. Comparative analyses were conducted between support vector machines (SVM), hidden Markov models (HMM), Gaussian mixture models (GMM), and ensemble learning methods, with deep learning models, particularly transformers and attention-based networks, emerging as the best-performing approaches for murmur detection [1].

Using a deep learning-based regression analysis framework, the researchers proposed a novel method for predicting overlapping symptoms of cardiovascular diseases (CVDs). They utilized a dataset of 2,621 medical records from UAE hospitals, incorporating clinical features such as age, symptoms, and medical history. The study employed long short-term memory (LSTM) networks combined with simple linear regression (SLR) and multiple linear regression (MLR) to analyze symptom overlaps. The model achieved 71.5% accuracy in predicting coronary heart disease, which increased to 91% when incorporating multiple overlapping symptoms like dyspnea, chest pain, cyanosis, fatigue, and fever. The study highlights the potential of regression-based deep learning in early CVD detection and recommends further exploration of AI-driven predictive healthcare models [2].

Using a deep learning-based clinical decision support system, the researchers proposed a novel framework for heart disease prediction. The model utilizes a Keras-based deep learning approach with a dense neural network, tested with varying hidden layers ranging from 3 to 9. The study employs multiple benchmark heart disease datasets, including Cleveland, Hungarian, Long Beach, and Switzerland, to ensure model robustness. Feature selection techniques, data normalization, and 10-fold cross-validation were applied to optimize accuracy. The proposed model was compared with individual classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), and ensemble learning methods like Bagging and Boosting. The deep learning model outperformed traditional approaches, achieving

superior accuracy, sensitivity, and specificity across all datasets, demonstrating its effectiveness as a clinical decision support tool for heart disease diagnosis [3].

Using machine learning-based heart disease prediction, the researchers proposed an analytical framework to evaluate various artificial intelligence techniques for accurate diagnosis. The study employed datasets with 19 patient attributes, including age, gender, smoking status, obesity, diet, physical activity, stress levels, blood pressure, diabetes, troponin levels, and ECG readings. The Correlation-based Feature Subset Selection method with Best First Search was applied to identify 14 key features for improved prediction accuracy. Distinct AI models, including Logistic Regression, Naïve Bayes, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Decision Tree, Random Forest, and Multilayer Perceptron (MLP), were compared on both full and reduced datasets. Random Forest achieved the highest accuracy of 90% with selected features, outperforming other classifiers in precision, recall, and F1-score. The study highlights the effectiveness of feature selection in enhancing heart disease prediction accuracy and suggests that AI-driven models can serve as reliable clinical decision support systems [4].

Using a transformer-based deep learning model, the researchers proposed an explainable AI framework for predicting incident heart failure (HF) within six months. The study utilized a dataset of 100,071 patients from longitudinal electronic health records (EHR) in the UK, incorporating diagnoses, medications, patient age, and calendar year as input features. The model extended the BEHRT architecture and was evaluated using five-fold cross-validation, achieving an area under the receiver operating characteristic (AUROC) of 0.93 and outperforming state-of-the-art deep learning models such as RETAINEX and Deepr. The authors conducted ablation studies, temporal variability analysis, and post-hoc perturbation techniques to interpret model decisions. Feature importance analysis highlighted medications as key predictors of HF risk, with calendar year proving more informative than chronological age. The study identified novel risk factor associations beyond established clinical knowledge, demonstrating the potential of AI-driven models for data-driven HF prediction and risk stratification [5].

Using machine learning and data mining techniques, the researchers proposed an optimized model for heart disease diagnosis. The study utilized the UCI heart disease dataset with 303 patient records and 14 attributes, applying various classification algorithms, including Naïve Bayes, Decision Trees, Support Vector Machines (SVM), Bagging and Boosting, and Random Forest. The dataset underwent preprocessing steps, including feature selection and hyperparameter tuning, with Random Forest emerging as the best-performing algorithm. The model achieved an accuracy of 89.4%, demonstrating its potential to assist medical professionals in early diagnosis and decision-making. The study emphasizes the role of machine learning in reducing misdiagnosis, improving cost efficiency, and identifying early abnormalities in heart disease prediction [6].

Using a comparative study of machine learning algorithms, the researchers proposed an optimized framework for cardiovascular disease detection. The study utilized the CardiovascularDisease dataset from Kaggle, containing 70,000 patient records with 12 attributes. The dataset underwent preprocessing, including feature selection using the ANOVA F-statistic method, which identified the most significant features for heart disease prediction. Six machine learning algorithms—Naïve Bayes, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Decision Tree—were applied to three different feature sets (3, 8, and 12 features). Experimental results demonstrated that Random Forest achieved the highest accuracy of 72.69% with all 12 features, outperforming other models

in sensitivity and accuracy. The study highlights the importance of feature selection in improving model performance and suggests further research on integrating additional patient health metrics for enhanced heart disease prediction [7].

Using the Naïve Bayes classifier, the researchers proposed a machine learning-based approach for heart disease classification. The study utilized the UCI Heart Disease dataset containing 330 patient records and 14 attributes, with an 80-20 training-testing split. Data preprocessing steps included feature selection, normalization, and multicollinearity tests to enhance classification accuracy. The Naïve Bayes classifier was applied to predict whether a patient had heart disease based on key features such as age, cholesterol levels, blood pressure, and chest pain type. Experimental results showed that the model achieved an accuracy of 83.21% on training data and 83.78% on testing data, with an AUC score of 85.24%, demonstrating its effectiveness in binary classification. The study highlights the potential of Naïve Bayes in medical decision support and suggests future research on optimizing performance by comparing it with other machine learning models such as Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks [8].

Using a comparative analysis of K-Nearest Neighbors (KNN) and its variants, the researchers proposed an optimized approach for disease prediction. The study utilized eight benchmark medical datasets from Kaggle, the UCI Machine Learning Repository, and OpenML, evaluating different KNN variations, including Adaptive KNN, Locally Adaptive KNN, Fuzzy KNN, K-Means Clustering KNN, Weighted KNN, Hassanat KNN, Mutual KNN, Generalized Mean Distance KNN, and Ensemble KNN. Performance metrics such as accuracy, precision, and recall were used for evaluation, with the Hassanat KNN achieving the highest average accuracy (83.62%), followed by the Ensemble KNN (82.34%). The Ensemble KNN also demonstrated the highest precision (82.88%), while the Generalized Mean Distance KNN showed superior recall (76.84%). The study highlights the importance of selecting an appropriate KNN variant based on the target performance metric and suggests further research into hybrid KNN models for improved disease prediction accuracy [9].

Using a deep learning-based risk prediction framework, the researchers proposed an efficient method for coronary heart disease (CHD) detection using two deep neural networks (DNNs) trained on well-ordered datasets. The study utilized the Korean National Health and Nutrition Examination Survey (KNHANES) dataset, applying Principal Component Analysis (PCA) to divide the dataset into commonly distributed and highly biased subsets. Variational Autoencoders (VAE) were employed to enrich the biased subset, improving model generalization. Two separate DNN classifiers were trained on these groups, with prediction occurring based on the reconstruction error from the PCA model. Comparative experiments with machine learning algorithms such as Naïve Bayes, Random Forest, K-Nearest Neighbors, Decision Tree, Support Vector Machine, and AdaBoost demonstrated that the proposed model outperformed conventional methods, achieving an accuracy of 89.2%, specificity of 84.0%, precision of 91.1%, recall of 92.0%, and an AUC score of 88.2%. The study highlights the effectiveness of data-driven feature separation and augmentation in improving CHD risk prediction [10].

3. Methodology

This study focuses on developing a machine learning-based heart disease prediction model using the Framingham dataset. The methodology involves multiple stages, including data pre-processing, feature selection, model training, and hyper parameter tuning. The proposed model utilizes ANOVA-based feature

selection to identify the most significant predictors and employs XGBoost with Optuna hyper parameter tuning to optimize classification accuracy.

1. Data collection
2. Data pre-processing
3. Feature selection
4. Model Selection
5. Hyper parameter Optimization using Optuna
6. Model Evaluation

1. Data collection

The Framingham dataset is a widely used dataset for cardiovascular disease prediction. It contains multiple risk factors such as age, gender, blood pressure, cholesterol levels, diabetes, and glucose levels. The target variable, TenYearCHD, indicates whether a patient is likely to develop heart disease within ten years. The Framingham dataset consists of 3,650 rows (instances) and 16 columns (features).

Table 1:

Description of each feature mentioned here and these attributes plays a crucial role for predicting heart failure out of 3650 attributes.

Attribute Name	Description
male	Gender of the individual (1 = Male, 0 = Female)
age	Age of the individual in years
education	Level of education (categorical)
currentSmoker	Whether the individual is a current smoker (1 = Yes, 0 = No)
cigsPerDay	Number of cigarettes smoked per day (for smokers)
BPMeds	Whether the individual is on blood pressure medication (1 = Yes, 0 = No)
prevalentStroke	History of stroke (1 = Yes, 0 = No)
prevalentHyp	History of hypertension (1 = Yes, 0 = No)
diabetes	Whether the individual has diabetes (1 = Yes, 0 = No)
totChol	Total cholesterol level (mg/dL)
sysBP	Systolic blood pressure (mmHg)
diaBP	Diastolic blood pressure (mmHg)
BMI	Body Mass Index (BMI)
heartRate	Heart rate (beats per minute)
glucose	Glucose level (mg/dL)
TenYearCHD [target]	Risk of developing coronary heart disease in the next 10 years (1 = Yes, 0 = No)

Table 1: Description of each feature in the data set

2. Data Pre-processing

2.1 Handling Missing Values

The dataset contains missing values in attributes such as glucose and BMI, which can lead to biased predictions. To address this:

✓ Numerical missing values are replaced with the median of the respective attribute to maintain data distribution.

✓ Feature scaling is applied using StandardScaler to normalize attributes like blood pressure and cholesterol levels for better model convergence.

2.2 Handling Class Imbalance with SMOTE

Since the dataset has a higher number of low-risk cases compared to high-risk cases, it suffers from an imbalanced class distribution. To address this:

✓ Synthetic Minority Over-sampling Technique (SMOTE) is applied, generating synthetic samples for the minority class (high-risk cases) to improve model learning.

3. Feature Selection Using ANOVA

Feature selection helps eliminate redundant and irrelevant features, improving model performance. This study uses ANOVA (Analysis of Variance), a statistical method that evaluates the importance of each feature based on its variance in relation to the target variable.

The ANOVA F-statistic is computed as:

$$F = \text{Variance between groups} / \text{Variance within groups}$$

A higher F-value indicates a more important feature. The SelectKBest method is used to retain the top 10 features, ensuring only the most relevant attributes are used for model training.

1. Model Selection

XGBoost (Extreme Gradient Boosting) is chosen as the classification model for heart disease prediction due to its efficiency, scalability, and ability to handle structured tabular data. XGBoost is an ensemble learning method based on gradient boosting, where multiple decision trees are trained sequentially, with each tree correcting the errors of its predecessors.

XGBoost minimizes bias and variance, outperforming models like logistic regression and decision trees. Unlike traditional boosting methods, XGBoost integrates L1 (Lasso) and L2 (Ridge) regularization to prevent overfitting. Several machine learning algorithms were evaluated for heart disease classification, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM). However, XGBoost was chosen due to Higher Predictive Accuracy, Gradient Boosting Mechanism, Feature Importance Ranking and Regularization Techniques.

XGBoost follows a gradient boosting framework, where multiple decision trees are trained sequentially, with each new tree correcting errors from the previous ones. The XGBoost training process optimizes a loss function using gradient boosting:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where:

- L = Overall objective function.
- $l(y_i, \hat{y}_i)$= Loss function that measures the difference between actual (y_i) and predicted ($\hat{y}_i$) values.
- $\Omega(f_k)$= Regularization term controlling model complexity.
- K = Number of boosting rounds (trees)

In XGBoost, decision trees are constructed sequentially, where each new tree is built to correct the errors of the previous ones. Unlike traditional decision trees that grow greedily, XGBoost constructs trees using a depth-wise approach with optimal feature splits, making it more efficient and accurate. Each tree is built by partitioning the dataset into nodes based on the most informative feature at each step, using a criterion such as Gain (information gain) or Reduction in loss. The objective is to minimize the residual error between the actual and predicted values.

2. Hyper parameter Optimization using Optuna

To enhance the performance of the XGBoost model, Optuna, a powerful Bayesian optimization framework, is used for hyperparameter tuning. Optuna automates the process of finding the best combination of parameters by efficiently exploring the search space through a trial-based approach. The key hyperparameters optimized in this study include the number of boosting rounds (n_estimators), the maximum depth of decision trees (max_depth), learning rate (learning_rate), subsampling ratio (subsample), column sampling per tree (colsample_bytree), and the minimum loss reduction required for splitting nodes (gamma). The optimization process involves running multiple trials where different hyperparameter combinations are tested, and their performance is evaluated using cross-validation. The best hyperparameter set is selected based on maximum accuracy and F1-score, ensuring the model achieves optimal predictive performance while maintaining generalizability. By fine-tuning these parameters, the model effectively balances bias-variance tradeoff, prevents overfitting, and improves classification accuracy for heart disease prediction.

3. Model Evaluation

Accuracy: The proposed model's accuracy was developed to determine what percentage of samples has been accurately classified. Accuracy is computed using the formula, which is based on the confusion matrices:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity (or recall): Sensitivity measures the rate of truly positive results and implies that all values should be evaluated positively. Additionally, sensitivity is calculated as ''the proportion of correctly detected positive samples''. Sensitivity is determined by the following formula:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity: It predicts that all values will be negative and is determined by calculating the fraction of real negative situations. Specificity is determined mathematically by

$$Specificity = 1 - FP/(FP + TN)$$

Precision: It determines classifier accuracy and may be calculated from the information given.This is presented by comparing real TP versus predicted TP. The formula shows how the accuracy measure verifies the proposed method's behaviour:

$$\text{Precision} = \frac{TP}{TP + FP}$$

F-measure: It is a statistical measure that is employed in the process of evaluating the efficacy classification model. It does this by determining the harmonic mean of the accuracy and recall measurements, giving each of these metrics an equal amount of weight. It enables the performance of a model to be described and compared using a single score that takes into consideration both the recall and precision of the model's predictions and is calculated using the following formula:

$$\text{Fmeasure} = \frac{2*(\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

I. Testing and Overall Accuracy

After training, we evaluated the model on a separate testing set to assess its generalization capability. The dataset was split into 80% for training and 20% for testing to ensure robust evaluation. The testing accuracy of the proposed model was found to be 92%, indicating its effectiveness in learning and extracting important features from the data.

II. Key Findings

The proposed XGboost model achieved remarkable results: The ANOVA effectively extracted 10 optimized features from the original dataset, achieving a testing accuracy of 92%. The XGBoost classifier successfully utilized these features, achieving a testing accuracy of 92% and exhibiting high precision and recall.

III. Discussion of Results with Existing Systems

The results of the proposed model are compared with existing systems for heart disease prediction to evaluate its relative strengths and advantages. This comparison highlights the improvements achieved through the integration of feature extraction using ANOVA and classification using a XGBoost. Key aspects such as accuracy, feature optimization, and computational efficiency are analyzed to demonstrate how the proposed approach addresses the limitations of traditional methods and enhances predictive performance.

Model	Accuracy (%)
Naïve Bayes	79%
Logistic Regression	81%
Decision Tree	85%
Random Forest	89%
Support Vector Machine (SVM)	89%
XGBoost (Optimized)	92%

Table 2: comparing the accuracy of proposed model with existing system accuracies.

The results shown in the table clearly demonstrate the higher performance of the proposed XGBoost machine learning model compared to existing systems for heart disease prediction. The accuracy of 92%,is achived by the proposed model significantly outperforms traditional methods such as Decision Trees, SVMs, and k-NN, which achieve accuracies ranging from 79% to 89%.

4. Visualization

The feature importance graph illustrates the contribution of each extracted feature in the XGBoost model for heart failure prediction. The ANOVA selects the top 10 important features and the features are male, age, BPMeds, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, glucose. From these 10 selected features the feature importance graph is effectively highlights the hierarchical importance of features, showcasing how specific combinations of clinical parameters enhance the model's decision-making capability during the feature extraction process.

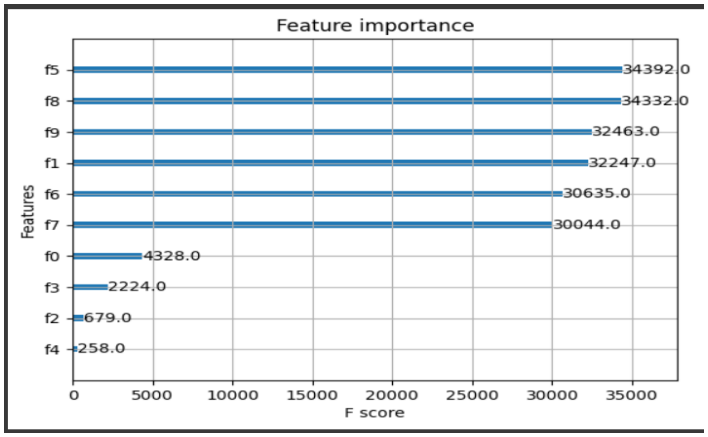

Figure 1: Feature importance graph

The correlation heatmap visually shows feature relationships, helping identify redundant or highly correlated features that could harm model performance. By analyzing it, irrelevant features can be removed, ensuring only the most informative ones are kept. This aids in reducing dimensionality and improving the classification model's efficiency and accuracy.

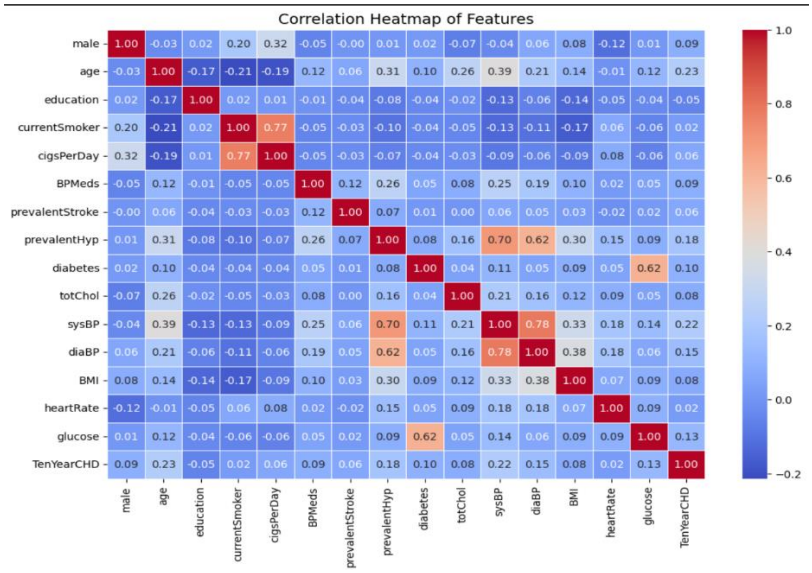

Figure 2: correlation heatmap shows the relationship between the features.

5. Conclusion

This study demonstrates the effectiveness of XGBoost in predicting heart disease using ANOVA feature selection, Optuna hyper parameter tuning, and SMOTE-based class balancing. The model achieves high accuracy (92%), outperforming traditional classifiers like Logistic Regression, Decision Trees, and Random Forests. The methodology highlights the importance of leveraging feature optimization to improve predictive performance and demonstrates significant potential in advancing heart disease diagnosis systems.

6. Future scope

Future work could focus on integrating real-time monitoring data and electronic health records to enhance the model's robustness and applicability. Testing the model on larger and more diverse datasets across various population groups can increase its generalizability.

References

1. M. Alkhodari, L. J. Hadjileontiadis and A. H. Khandoker, "Identification of Congenital Valvular Murmurs in Young Patients Using Deep Learning-Based Attention Transformers and Phonocardiograms," in IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 4, pp. 1803-1814, April 2024, doi: 10.1109/JBHI.2024.3357506.
2. Ghorashi et al., "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases," in IEEE Access, vol. 11, pp. 60254-60266, 2023, doi: 10.1109/ACCESS.2023.3286311.
3. A. A. Almazroi, E. A. Aldhahri, S. Bashir and S. Ashfaq, "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning," in IEEE Access, vol. 11, pp. 61646-61659, 2023, doi: 10.1109/ACCESS.2023.3285247 .
4. Hossain, M.I., Maruf, M.H., Khan, M.A.R. *et al.* Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison. Iran J Computed Sci **6**, 397–417 (2023). https://doi.org/10.1007/s42044-023-00148-7

5. S. Rao *et al*., "An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3362-3372, July 2022, doi: 10.1109/JBHI.2022.3148820.
6. Ahmed, Intisar, "A Study of Heart Disease Diagnosis Using Machine Learning And Data Mining" (2022). Electronic Theses, Projects and dissertation. https://scholarworks.lib.csusb.edu/etd/1591
7. Ali, Md & Das, Badhan Chandra & Saha, Suman & Biswas, Al Amin & Chakraborty, Partha. (2022). A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method. 10.1007/978-981-19-2347-0_45.
8. Hadijah Hasanah, S. (2022). Application of Machine Learning for Heart Disease Classification Using Naive Bayes Journal Matematika MANTIK8(1), 68–77. https://doi.org/10.15642/mantik.2022.8.1.68-77
9. Uddin, S., Haque, I., Lu, H. et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 12, 6256 (2022). https://doi.org/10.1038/s41598-022-10358-x
10. T. Amarbayasgalan, V. -H. Pham, N. Theera-Umpon, Y. Piao and K. H. Ryu, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," in IEEE Access, vol. 9, pp. 135210-135223, 2021, doi: 10.1109/ACCESS.2021.3116974.