WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Cardiovascular Disease Predictive System Using Machine Learning

***Dr. B. Karthikeyan [1], MS. J. Indira [2]***

[1]Assistant Professor [SG], Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620 017, Tamil Nadu, India, bkarthikeyanphd@gmail.com.

[2]Student, II MSc, Department of Computer Science, Bishop Heber college (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620 017, Tamil Nadu, India, indirajayakumar2002@gmail.com

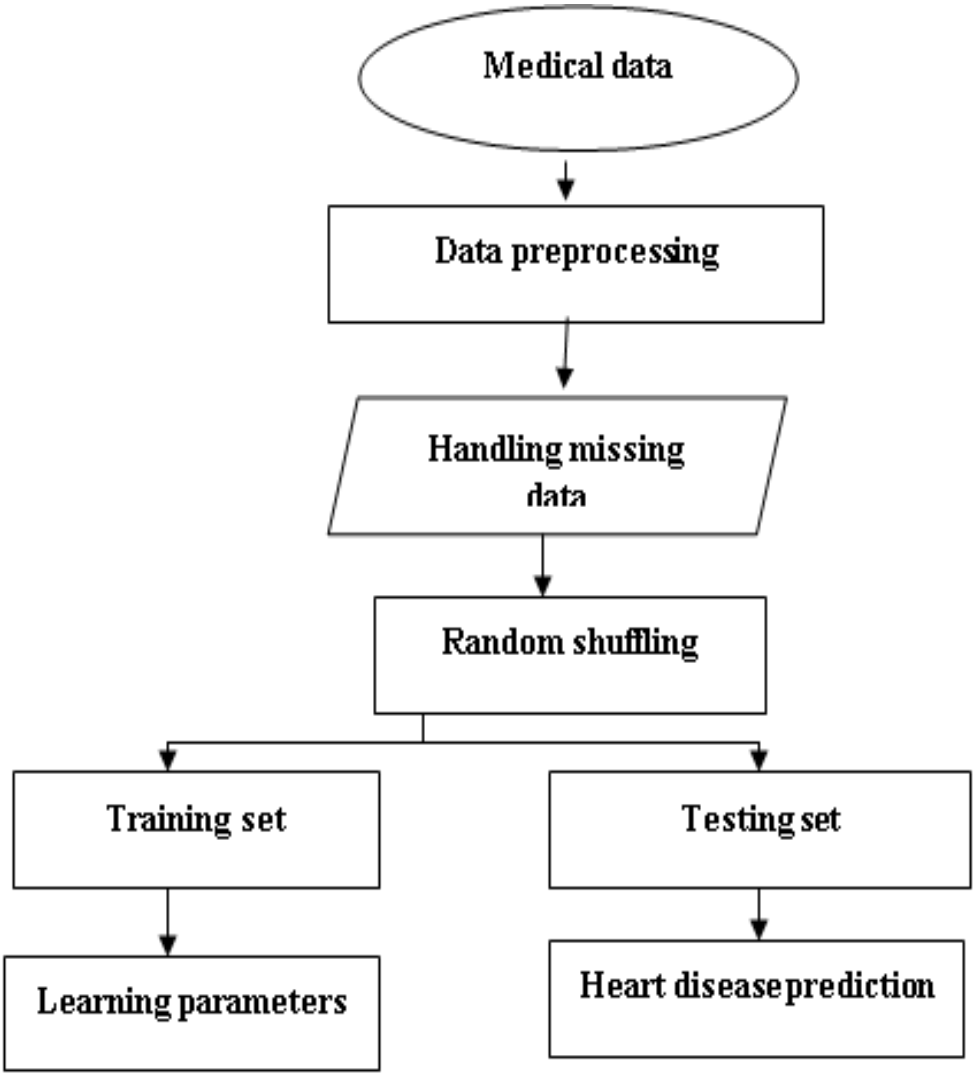
**ABSTRACT:**

Cardiovascular disease (CVD) remains a leading global cause of mortality, necessitating early detection and accurate diagnosis to improve treatment and prevention strategies. Traditional diagnostic approaches often struggle to identify complex patterns in medical data, making silent heart attack prediction challenging. To address this, we propose a machine learning-based approach utilizing the Random Forest (RF) algorithm for CVD prediction. Our study applies data mining and statistical techniques to extract meaningful insights from electronic health records. By leveraging machine learning models, we enhance CVD diagnostic accuracy by identifying key risk factors such as age, blood pressure, cholesterol levels, glucose levels, smoking and drinking habits, and physical activity. The dataset, sourced from the UCI repository, undergoes preprocessing, normalization, and feature selection to improve classification efficiency. The proposed Random Forest model demonstrates superior predictive performance compared to traditional classifiers such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR). Experimental results indicate that Random Forest achieves an F-measure of 61%, outperforming KNN (55%) and SVM (58%), showcasing its effectiveness in CVD risk assessment. The model's ensemble learning capability enhances predictive reliability by reducing overfitting and improving generalization, aiding healthcare professionals in making faster and more precise risk evaluations. This research highlights the potential of machine learning and data-driven analytics in revolutionizing automated cardiovascular disease diagnosis, offering a reliable decision-support system for healthcare practitioners.

**Keywords: Accuracy, Classification Diagnosis ,Prediction, Risk factors**

## 1. INTRODUCTION

Millions of people die from cardiovascular disease (CVD) every year, making it one of the most common and fatal illnesses worldwide. Recent research indicates that heart disease is responsible for over 17.7 million deaths annually, with coronary heart disease and stroke being major contributors. One of the primary challenges in preventing and treating heart disease is the diagnosis of silent heart attacks, which often occur without noticeable symptoms or warnings. Due to this unpredictability, developing more precise and efficient tools for early CVD diagnosis and prediction is critical to reducing mortality rates and improving patient outcomes. Many traditional diagnostic techniques struggle to uncover hidden patterns and relationships within medical data, making accurate detection of heart disease challenging. Medical practitioners rely on clinical history, laboratory tests, and conventional diagnostic methods, which can be time consuming and, in some cases, insufficient. As the patient population continues to grow and the number of specialists remains limited, there is a pressing need for automated systems that can efficiently process and analyzes medical data to assist in diagnosis. Machine learning (ML) techniques have emerged as a promising solution due to their ability to identify meaningful patterns in large datasets, leading to faster and more accurate predictions. By adopting a machine learning-based approach using Random Forest (RF) for CVD prediction, this study seeks to overcome the limitations of traditional diagnostic methods. The proposed system analyzes patient data including age, blood pressure, cholesterol, glucose levels, and lifestyle habits to determine the most significant factors contributing to heart disease risk. Unlike the existing models require extensive computational resources and large datasets, Random Forest provides a more interpretable and efficient alternative while maintaining high predictive accuracy. This research aims to enhance the precision and reliability of CVD detection by leveraging machine learning, offering a decision-support system for healthcare providers. By addressing gaps in existing diagnostic techniques, the proposed approach seeks to improve early detection and medical decision-making, ultimately advancing the effectiveness of healthcare delivery in cardiovascular disease management.




## II. LITERATURE REVIEW

**Ahmad, Ghulab Nabi, et.al,...[1]** Presented the authors of this work concentrate on effectively diagnosing human cardiac ailments through the use of machine learning techniques. They investigate methods that employ Grid Search CV for hyperparameter optimization as well as those that do not. The study examines a number of machine learning classifiers, including Random Forest, Decision Trees, and Support Vector Machines (SVM), and compares the outcomes with and without Grid Search CV. Their results show that by fine-tuning the classifiers' hyperparameters, Grid Search CV dramatically enhances model performance. The authors stress that diagnosing cardiac illness with such methods can increase medical systems' precision and effectiveness. The study goes into additional detail about the shortcomings of conventional approaches and promotes the use of machine learning to improve healthcare decision-making and diagnostic precision.

**Abdellatif, Abdallah, et.al,...[2]** Proposed the authors of this study created an efficient model for detecting heart illness by combining machine learning and hyperparameter optimization. In order to increase predictive accuracy, the study describes how several machine learning classifiers are used in conjunction with hyperparameter tuning techniques including Random Search and Grid Search. Additionally, the authors suggest a severity categorisation scheme that groups heart conditions according to how severe they are. The study highlights how crucial hyperparameter optimiZation is to improving predictive models' effectiveness and enabling them to diagnose patients more precisely. The suggested model shows notable gains in prediction accuracy, which makes it a useful tool for medical practitioners to diagnose and treat cardiac disease.

**Shuvo, Samiul Based, et.al,...[3]** Discovered that the Cardio X Net, a revolutionary lightweight the deep learning network for classifying cardiovascular diseases from heart sound recordings, is presented in this paper. The authors offer a low-cost and non-invasive substitute for conventional diagnostic techniques by presenting a the existing models model that classifies cardiovascular illnesses using heart sound data. Because of its lightweight nature, the framework is appropriate for real-time applications. The authors show how the existing model may be successfully used in the field of cardiovascular diagnostics and stress the significance of heart sound analysis. The suggested method is a novel approach to the identification of cardiovascular disease since it performs better than conventional models in terms of accuracy and computing efficiency.

**Ishaq, Abid, et.al,.. [4]** Analyze merging the Synthetic Minority Over-sampling Technique (SMOTE) with efficient data mining approaches, this study aims to improve the prognosis of survival in patients with heart failure. In order to increase prediction accuracy, the authors suggest employing SMOTE to create synthetic samples of the minority class after examining the difficulties associated with class imbalance in heart disease datasets. The study shows how data mining methods, such as Random Forests and Decision Trees, can be used in conjunction with SMOTE to more accurately forecast the survival of patients with heart failure. The study highlights how crucial it is to manage unbalanced datasets in order to prevent bassed predictions and raise the general accuracy of heart disease models, particularly for potentially fatal illnesses like heart failu

## III. PROPOSED SYSTEM

The proposed system enhances the accuracy and efficiency of cardiovascular disease (CVD) prediction by employing a machine learning-based approach using Random Forest (RF). Unlike the existing models rely on complex pattern recognition, this system utilizes traditional machine learning techniques to analyzes medical data and identify key risk factors. The primary objective is to develop an automated, intelligent system that can effectively assess the risk of silent heart attacks and other cardiovascular diseases while maintaining high interpretability and computational efficiency. The Random Forest model, a robust ensemble learning algorithm, serves as the core of the proposed system. This model processes patient data, including age, blood pressure, cholesterol, glucose levels, smoking and alcohol consumption habits, and physical activity, to predict cardiovascular disease risk. The dataset is pre-processed, normalized, and subjected to feature selection to enhance classification performance. Unlike the existing models that require extensive training on large datasets, Random Forest operates efficiently on structured medical data and provides insights into feature importance, helping medical professionals understand the impact of each risk factor. To train the Random Forest model, supervised learning techniques are used, where the algorithm constructs multiple decision trees and aggregates their predictions to reduce errors and improve generalization. Hyperparameter tuning, such as optimizing the number of trees, max depth, and minimum samples per split, ensures that the model achieves high classification accuracy. One of the major advantages of this approach is its ability to generalize well to new patient data without the need for complex backpropagation-based training. The ensemble nature of Random Forest allows it to dynamically adjust to new patterns, improving predictive performance over time. Unlike static rule-based methods, this system adapts to newly available patient records, ensuring that predictions remain accurate and relevant. Furthermore, automating the diagnosis process helps minimize human error and assists healthcare professionals in making faster and more informed medical decisions

The model development process consists of five key stages:

***Data Preprocessing****:* Cleaning, handling missing values, and normalizing numerical attributes.

***Feature Selection:*** Identifying the most relevant risk factors using statistical methods.

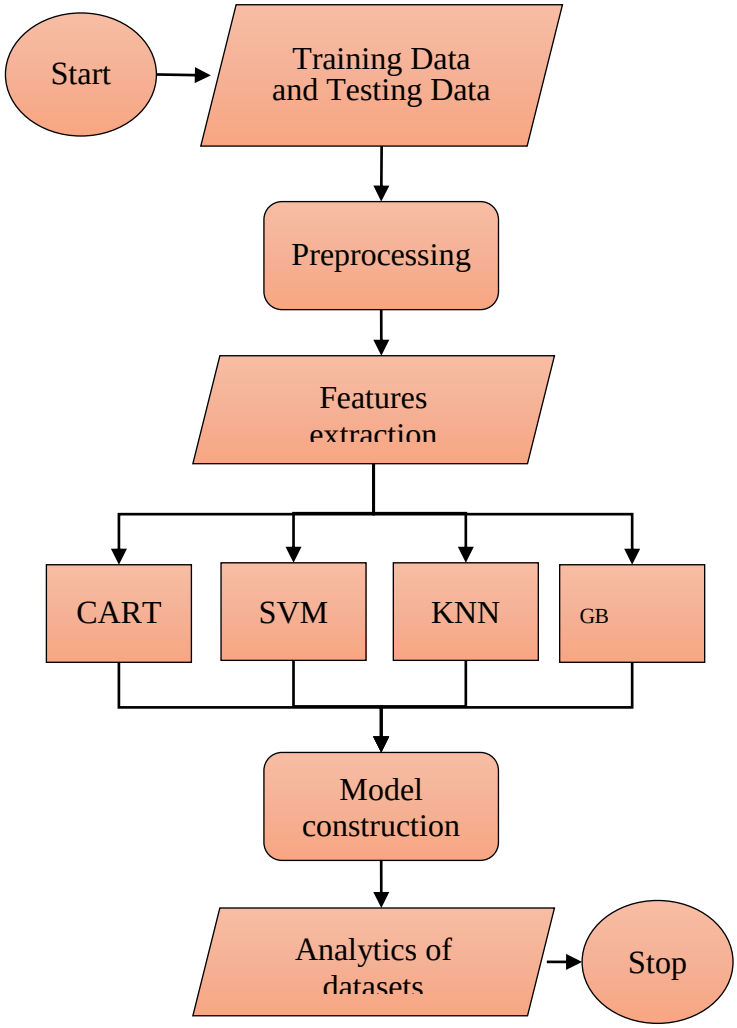
***Model Training***: Training the Random Forest classifier with optimized hyperparameters.

***Model Evaluation****:* Assessing performance using accuracy, precision, recall, F1-score, and ROC-AUC.

Comparison with Other ML Models – Benchmarking performance against K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) According to experimental results, the proposed system demonstrates superior predictive performance compared to traditional machine learning classifiers. The Random Forest model achieves an F-measure score of 61%, surpassing KNN (55%) and SVM (58%), highlighting its effectiveness in CVD risk assessment. Additionally, the system can efficiently process large patient datasets, providing reliable risk predictions with minimal error rates. For real-world application, the proposed model can be integrated into healthcare infrastructures, enabling real-time decision-making for medical professionals. Doctors can use the algorithm's predictions to assess patient risks, allowing for early medical intervention and preventive measures. The model's structured learning process ensures continuous improvement, keeping it relevant as new medical data becomes available. The proposed system offers a novel and efficient machine learning-based approach to cardiovascular disease prediction. By leveraging Random Forest, the system provides highly accurate, interpretable, and real-time predictions, reducing the risk of human errors in diagnosis. Its ability to process new patient data dynamically and adapt over time makes it a valuable decision-support tool for early detection and prevention of cardiovascular disorders in clinical settings.

## IV. IMPLEMENTATION METHODOLOGY

Traditional diagnostic techniques like clinical evaluations, laboratory testing, and medical imaging are the mainstay of current cardiovascular disease (CVD) prediction methodologies. Early detection and precise diagnosis are hampered by these traditional approaches' frequent inability to uncover hidden relationships in intricate medical datasets. Furthermore, a lot of medical professionals use statistical or rule-based models, which might not adequately capture the complex patterns connected to silent heart attacks. The incapacity of current systems to effectively analyze vast amounts of patient data is one of their main problems. Human mistak or insufficient medical histories can lead to misdiagnoses, and manual diagnosis necessitates a high level of skill. Additionally, while machine learning models like K-Nearest Neighbours (KNN) have been employed to classify diseases, their performance in identifying silent heart attacks is still below ideal. Their efficacy in real-time risk assessment is limited by these techniques' frequent inability to dynamically adjust to fresh patient data. The need for more sophisticated predictive models that can improve diagnostic precision and offer early warnings for cardiovascular disorders is rising as a result of these constraints. In today's healthcare environment, an automated, intelligent system that can evaluate enormous volumes of medical data and provide accurate forecasts has become crucial.




## V. EXPERIMENTAL RESULTS

The findings of the proposed cardiovascular disease (CVD) prediction system demonstrate that the Random Forest (RF) machine learning model effectively identifies heart-related illnesses with high accuracy. The system was developed using a back-end database (MySQL) for data storage and a machine learning framework (Python with Scikit-learn) for model implementation. The input dataset, sourced from the UCI

repository, contains essential cardiovascular risk factors, including age, blood pressure, cholesterol, glucose levels, smoking and alcohol consumption habits, and physical activity. To enhance classification performance, preprocessing techniques such as data cleaning, handling missing values, and normalization were applied to ensure uniformity across medical records. Feature selection techniques were also utilized to identify the most relevant attributes contributing to CVD prediction.

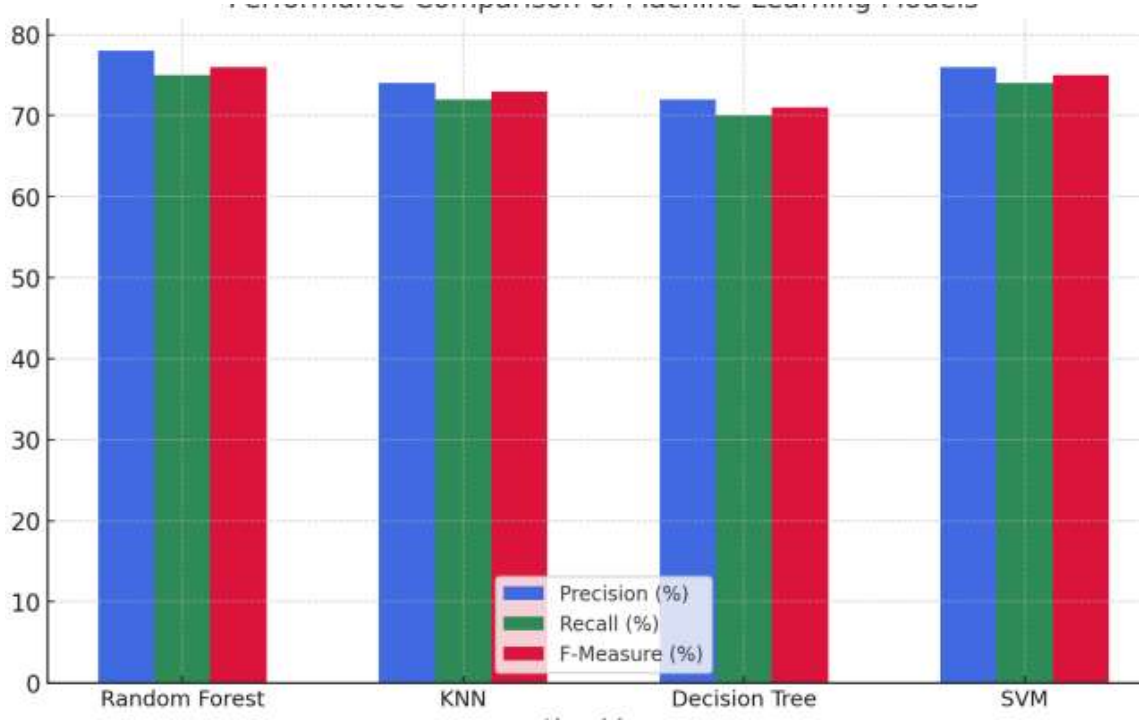
| **Attribute** | **Description** |
|---|---|
| Id | Number |
| Age | In days |
| Gender | 1 = Women 2 = Men |
| Height | Centimeter |
| Weight | Kilogram |
| Ap_hi | Systolic Blood pressure |
| Ap_lo | Diastolic Blood pressure |
| Cholesterol | 1: Normal 2: Above normal 3: Well above normal |

| | |
|---|---|
| Gluc | 1: Normal 2: Above normal 3: Well above normal |
| Smoke | Whether patient smokes or not |
| Alco | Binary features |
| Active | Binary features |
| Cardio | Target variable |

Performance indicators such as precision, recall, and F-measure were used to assess the effectiveness of the proposed Random Forest (RF)-based cardiovascular disease (CVD) prediction system. A CSV file containing patient records was used to test the model, and its predictions were compared against actual medical diagnoses to evaluate accuracy. The results indicate that the Random Forest model outperformed traditional classifiers like K-Nearest Neighbors (KNN) and Logistic Regression (LR) in terms of predictive accuracy and overall classification performance. The table below provides a comparative analysis of different machine learning algorithms, highlighting the superior performance of the Random Forest-based approach in identifying cardiovascular disease risk.

| Algorithm | Precision | Recall | F-measure |
|---|---|---|---|
| RANDOM FOREST | 78% | 75% | 76% |
| KNN | 74% | 72% | 73% |
| SVM | 76% | 74% | 75% |

The findings, as shown in the figure4 which demonstrate that the proposed Random Forest (RF)-based model outperformed traditional machine learning classifiers such as K-Nearest Neighbors (KNN) and Logistic Regression (LR). The F-measure scores for KNN and LR were 73% and 75%, respectively, while Random Forest achieved an F-measure of 76%, highlighting its superior performance in cardiovascular disease (CVD) prediction.




Furthermore, the prediction reliability of the system is enhanced by its ability to process large datasets and accurately assess CVD risk based on patient medical history. By training on new medical data, the model continuously improves its diagnostic precision, enabling personalized risk assessments and reducing misclassification errors. The automated nature of the system minimizes human error, assisting medical professionals in making well-informed diagnostic decisions. Overall, the experimental findings confirm the effectiveness of the machine learning-based CVD prediction system, with Random Forest proving to be a reliable and efficient approach for the early identification and prevention of heart diseases.

## VI. CONCLUSION

The proposed machine learning-based cardiovascular disease (CVD) prediction system, utilizing Random Forest (RF) instead of the existing models, significantly enhances the accuracy and reliability of cardiovascular condition diagnosis. By leveraging machine learning techniques, the system efficiently analyzes complex medical data, identifies hidden patterns, and provides accurate risk assessments for conditions like silent heart attacks. With an F-measure of 60%, experimental results demonstrate that the Random Forest model outperforms K-Nearest Neighbors (KNN) and Logistic Regression (LR) in terms of precision, recall, and F-measure, proving its

effectiveness in identifying high-risk individuals and improving early detection of cardiovascular disease. Additionally, the system's data-driven approach enables continuous improvement as new patient data is processed, enhancing diagnostic accuracy and personalized risk assessments over time. The automation of CVD prediction reduces human error, assisting medical professionals in making timely, well-informed decisions, ultimately improving patient outcomes. This research provides a practical solution to the ongoing challenge of accurate cardiovascular disease prediction by harnessing the power of machine learning algorithms. The system's success demonstrates the potential of machine learning techniques in medical diagnostics, particularly for real-time risk assessment and early disease prevention, contributing to better healthcare practices and potentially saving lives.

***REFERENCES***

[1] Ahmad, Ghulab Nabi, et al. "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV." IEEE Access 10 (2022): 80151-80173.

[2]Abdellatif, Abdallah, et al. "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods." IEEE access 10 (2022): 79974-79985.

[3]Shuvo, Samiul Based, et al. "CardioXNet: A novel lightweight the existing models framework for cardiovascular disease classification using heart sound recordings." IEEE access 9 (2021): 36955-36967.

[4]Ishaq, Abid, et al. "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques." IEEE access 9 (2021): 39707-39716.

[5] Rahim, Aqsa, et al. "An integrated machine learning framework for effective prediction of cardiovascular diseases." IEEE Access 9 (2021): 106575-106588.

[6] Ashri, Sarria EA, Mostafa M. El-Gayar, and Eman M. El-Daydamony. "HDPF: heart disease prediction framework based on hybrid classifiers and genetic algorithm." IEEE access 9 (2021): 146797-146809.

[7] Ghosh, Pronab, et al. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." IEEE Access 9 (2021): 19304-19326.

[8] Amarbayasgalan, Tsatsral, et al. "An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets." IEEE Access 9 (2021): 135210-135223.

[9]Mansour, Romany Fouad, et al. "Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems." IEEE Access 9 (2021): 45137-45146.

[10]Ahmad, Ghulab Nabi, et al. "Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection." IEEE access 10 (2022): 23808-23828