Accurate Heart Disease Prediction using Machine Learning Techniques on Clinical Data

Sunanda Budihal
Research Scholar
Karnataka State Akkamahadevi Women's University, Vijayapura-586108
Assistant Professor
Dept. of Computer Science
Government First Grade College, Bagalkot-587103
Karnataka, India

Sheetalrani R. Kawale
Assistant Professor
Dept. of Computer Science
Karnataka State Akkamahadevi Women's University, Vijayapura-586108
Karnataka, India

ABSTRACT

Heart disease is still one of the main causes of mortality in the world, and its early diagnosis represents an important part of timely treatment and prevention. The purpose of this work is to create a stable and accurate Machine Learning (ML) model for predicting the risk of heart disease with real clinical patient data. This research work relied on a clinical sample of 333 patient records of Sai Cardiac Hospital, Vijayapura, Karnataka, India (SCHV). The data set included medical parameters that included age, sex, type of chest pain, echo, test outcomes, resting Electrocardiogram (ECG), and Coronary Angiograph (CA) test. Structural pre-processing and visualization tools were employed to determine and derive meaningful predictors of heart disease. Three machine learning classifier models (Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF)) were built and evaluated using accuracy, precision, recall, F1 score, specificity, Area Under Curve (AUC), and Matthews Correlation Coefficient (MCC) for performance measurement. Out of the above ML models that were generated, the SVM classifier performed best with an accuracy of 90% and an AUC of 0.95, better than both RF and KNN models. The integrity of the proposed model was verified based on the Receiver Operating Curve (ROC) curve and confusion matrices. Comparison with existing studies showed that the developed SVM model is more reliable for prediction. The results indicate that SVM-based predictive modelling has promising prospects for medical real-time diagnosis and can serve as a candidate decision support system in healthcare practice.

General Terms

Machine Learning, Classification, Predictive Modeling, Medical Informatics, Data Analysis, Pattern Recognition

Keywords

Cardio-Vascular Disease (CVD), Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN)

1. INTRODUCTION

Cardiovascular Diseases (CVDs) represent a distinct group of diseases that cause around 18 million deaths every year all over the world, and this means 32 percent of all world deaths. Among these, 85 percent are a result of heart attacks and strokes [1]. To reduce any complications and maximize the survival rates, the diagnosis should be carried out in a timely and correct manner. This is because the growing rate of heart-related conditions in both developed and developing countries is a clear indication of a pressing need to find better diagnostic and predictive strategies that would curb the related health risks. Heart illness encompasses numerous disorders that affect the structure and operations of the heart, including coronary artery conditions, heart failure, arrhythmia, and congenital heart defects. The primary risk factors that lead to heart disease are duly recorded, and they include low dietary habits, lack of exercise, tobacco use, excessive levels of alcoholism, excessive levels of cholesterol, hypertension, diabetes, and family history [2], [3]. This results in a gradual deposition of arterial plaque, otherwise known as atherosclerosis, which blocks blood flow and can lead to myocardial infarction or stroke [4]. The heart disease diagnosis is normally comprised of both clinical assessment and diagnostic tests. Electrocardiograms (ECGs), echocardiography, stress tests, blood tests, and coronary angiography are some of the commonly used tests [5].

A distinct characteristic of the study is that real-time clinical data of 333 patient records of the SCHV in Vijayapura, Karnataka, India, were used. The health indicators set in the collection included age, gender, blood pressure, cholesterol, Electrocardiogram (ECG) readings, and lifestyle characteristics, which made models closer to the clinical reality and more applicable and relevant. The ultimate goal of this research was to come up with a strong information-based prediction approach that can identify the individuals who are at the highest risk of having a heart attack. Machine Learning (ML) systems are capable of using a number of variables and modeling their relationships, which is more extensive and customized than traditional risk assessment methods, which often use a small set of features [6]. Three machine learning algorithms, namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF), were used and compared in this study. The models were constructed in Python, and the preprocessing techniques were used to clean and standardize the data. The common evaluation metrics that were used to measure performance included F1-score, Area Under Curve (AUC), recall, accuracy, precision and specificity, and MCC. The findings suggest the possibility of employing ML-based prediction approaches as credible and scalable predictive instruments to provide an estimate of heart attack risk. These models could be of great assistance to the physicians because they are prophylactic by nature, encourage instant decision-making, and ultimately enhance the care of the patient [7]. The paper presents an ML and real-time clinical-based predictive model for heart attacks that is scalable and accurate. The suggested models will one day enhance patient outcomes by enabling healthcare providers to make quality decisions and focus on preventative medical care.

2. LITERATURE SURVEY

ML is a relatively recent technique of CVD, especially heart disease prediction, based on organized patient data in the past few years. The presented review below outlines the major trends in research, practice, and developments based on the conducted research in 2023-2025. Saha et al. [8] applied an ML model to forecast early heart disease and established that the best accuracy of 91% was achieved by the RF algorithm using the Centers for Disease Control (CDC) dataset after significant preprocessing and feature screening were done. They conducted an experiment that compared five supervised learning models, including Logistic Regression (LR), RF, K-Nearest Neighbors (KNN), Decision Tree (DT), and Extreme Gradient Boosting (XGB) in healthcare. On the same note, Chandrasekhar and Peddakrishna (2023) [9] investigated ML models with single, multiple, and mixed datasets and utilized approaches which included Naïve Bayes (NB), LR, KNN, AdaBoost (AB), Gradient Boosting (GB), Generalized Cross-Validation (GCV), RF, and Classification and Regression Trees (CART) and upheld potential of ML in early diagnosis. Building on the topic, Biswas et al. (2023) [10] compared different ML-based feature selection algorithms on the Cleveland dataset. In their work, they evaluated SVM, LR, KNN, DT, Gaussian Naïve Bayes (GNB), and RF in terms of sensitivity, specificity, Area Under Curve Receiver Operating Characteristic (AUC-ROC), and accuracy to highlight advantages and limitations of these approaches. In a different strategy, Jawalkar et al. (2023) [11] created a hybrid model where a Hidden Dirichlet Process and a DT-based RF classifier (DTRF) were used and found to be greatly superior to KNN and LR classifiers on the Cleveland dataset, with an accuracy of 87.12%. The burden of heart disease on the global population was also mentioned by Kavitha et al. [12], who provided a hybrid RF-DT model on the Cleveland dataset, which resulted in an 88.7% accuracy and recommended Deep Learning (DL) and multi-class classification. Similarly, Ahdal et al. [13] have also noted the importance of using ML to analyze medical data and identify risk factors and the clinical applicability of this tool on a more general level. Rimal et al. [14] evaluated 18 ML models (8 Automatic ML (AutoML) based models) and found AutoML to be more successful with diverse accuracies up to 88% whereas Ram Kumar revealed that DL using Convolutional Neural Network (CNN) was more successful with 83.61% accuracy, 97% recall, and AUC-ROC of 0.94 on the UCI Cleveland dataset. According to these findings, Stonier et al. [15] used a sample of 301 patients and found that RF was the most effective model with 88.52% accuracy relative to LR, NB, SVM, XGBoost (XGB), and KNN compared to a neural network. Lastly, Osei-Nkwantabisa et al. [16] found out that KNN performed better (87% accuracy, 86% precision, 90% recall, 88% F1-score) than LR and Artificial Neural Network (ANN) on the University of California Irvine (UCI) dataset and recommended that normalization, Synthetic Minority Over-sampling Technique (SMOTE), and ensemble approaches to be used in the future to improve the operation of KNN. On the one hand, these works suggest the effectiveness of ML, i.e., RF and hybrid models, in the early detection of heart diseases and the need to further increase DL models and AutoML systems to get the most out of predictions. However, certain problems, like relying on small or individual data, high complexity of its implementation because of tools like Grid Search Cross Validation (GridSearchCV) and SMOTE, poor interpretability of complex models, and the scalability issue, have not been considered [17], [18]. Hence, for addressing the following issues, this work presents a novel dataset, which is discussed in detail in the next section.

3. DATASET DESCRIPTION

The data set of actual clinical data of SCHV, Karnataka, India. Patient history has attributes like age, blood pressure, nature of chest pain, heart rate, and other vital signs. Authentic hospital-based datasets are used so that the model is more relevant and reliable for clinical situations. A complete list of these characteristics is presented in Table 1.

Table 1. Dataset description and feature details

Feature	Name	Description
Age	Age	Represents the patient's age, i.e., between 20 and 95 years
Sex	Gender	Indicates biological sex, 1 = Male, 0 = Female
SpO_2	Peripheral Capillary Oxygen Saturation	Percentage of hemoglobin saturated with oxygen. Normal range: 95–100%
HR	Maximum Heart Rate	Highest recorded heart rate, typically between 60–100 bpm
CP	Type of Chest Pain	0=No chest pain, 1=Typical chest pain, 2=Atypical chest pain, 3=Severe chest pain with sweating
BP	Resting Blood Pressure	Systolic blood pressure measured at rest (mmHg)
GRBS	Random Blood Sugar	1=Glucose >160 mg/dl, 0=Otherwise
ECG	Resting ECG Results	0=Normal, 1=Abnormal /ST wave variation
Echo	Echocardiography	0=Normal, 1=Abnormalities detected
CA	Number of Affected Coronary Vessels	0=Normal, 1=Minor plaque/single vessel disease, 2=Double vessel disease, 3=Triple vessel disease
Troponin	Cardiac Troponin Level	1=Positive (damage likely), 0=Negative
Target	Heart Disease	1=Heart disease present, 0=Heart disease absent

The collected dataset involves various medical features that are used in the diagnosis of heart diseases. The dataset consists of a combination of clinical and diagnostic characteristics that are applied to CVD risk assessment. In this dataset, the demographic variables include age (in the range of 20-95 years) and sex (1=male, 0=female). The dataset also includes physiological parameters, i.e., SpO 2 (saturation of hemoglobin in blood (the normal range was 95-100%) and Maximum Heart Rate (HR) in beats per minute (bpm), and resting Blood Pressure (BP) in mmHg. The dataset also included clinical indicators, i.e., Type of Chest Pain (CP), categorized as no chest pain (0), typical (1), atypical (2), and severe with sweating (3), and General Random Blood Sugar (GRBS), which was marked positive in case glucose was above 160 mg/dl. The results of diagnostic tests included Resting Electrocardiogram (ECG), findings (normal or abnormal/presenting with ST depression/elevation), Echocardiography (echo) findings (normal or abnormal), and the Number of Affected Coronary Vessels (CA) (between normal arteries (0) and triple vessel disease (3)). Also, the cardiac troponin levels (positive or negative) were included in the dataset as a biomarker of myocardial damage. The heart disease target variable was binary; that is, if an individual has the presence (1) or absence (0) of heart disease.

4. RESEARCH METHODOLOGY

This research includes several essential phases, such as real-world data collection as discussed in the above section, ML model training, ML model testing, and evaluating performance.

The dataset was classified into two target categories: '1' for indicating the presence of heart disease and '0' for its absence. It was divided in an 80:20 ratio, i.e., 80% of the data was used for training the models and the remaining 20% for testing. To predict CVD disease using the CVD indicator conditions, SVM, KNN, and RF were used. These ML approaches were trained on the training subset and assessed using the test subset. The entire methodological flow of the proposed system is depicted in Figure 1.

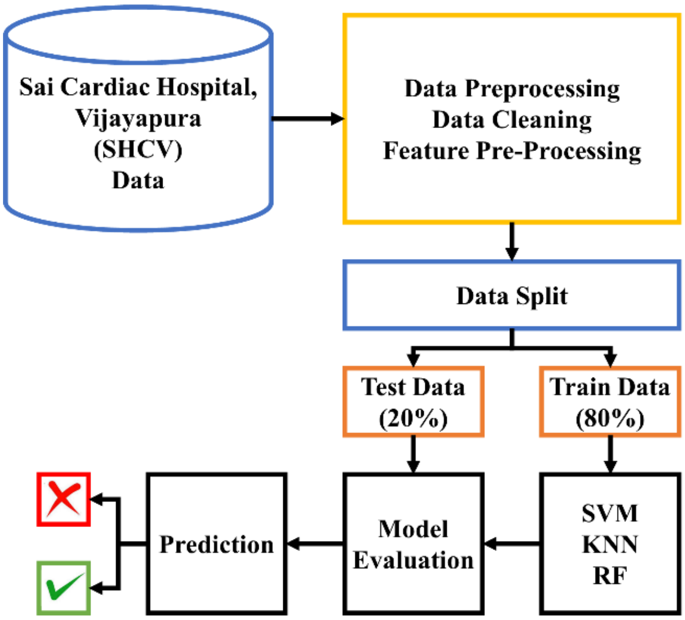

Fig 1: Overall Methodology of Proposed System

4.1 Loading Dataset

The collected dataset was first imported into an ML environment, which marks the starting point for all subsequent analytical tasks. Once loaded, the data was inspected systematically, ensuring that each attribute was recognized in its correct structure, format, and datatype. This initial stage is essential because it enables to verification of the integrity of the information before any modeling is attempted. Through structured loading, preliminary examinations like identifying missing values, detecting outliers, and assessing the distribution of medical or economic indicators become possible. These early checks help reveal inconsistencies, unexpected patterns, or abnormalities that may influence the overall accuracy and reliability of the predictive models. By establishing a clear understanding of the dataset at this stage, this work proceeded with confidence to deeper preprocessing steps, feature engineering, and model development.

4.2 Data Preparation

Ensuring data quality and preparing the dataset for effective model training requires a series of essential preprocessing steps that act as safeguards for reliable ML outcomes. One of the first tasks is addressing missing or null values, as incomplete records can distort model behavior or reduce accuracy. Depending on the nature of the data, these gaps may be handled by removing the affected entries or by replacing them with statistically derived estimates such as the mean or median of the corresponding attribute. In this work, the dataset did not have any null values; hence, this step did not affect the ML modeling. After resolving missing values, attention shifts to encoding categorical variables. Since ML approaches operate on numerical inputs, qualitative attributes, like the type of CP, the presence of exercise-induced angina, or other clinical categories, must be transformed into numerical representations using techniques like label encoding or one-hot encoding. Hence, this work changed the complete dataset into numerical form using one-hot encoding. Finally, feature scaling was applied to ensure that all variables contribute equally during training. Methods such as StandardScaler or MinMaxScaler are usually used to normalize the range of values, which is especially important for distance-based or margin-based algorithms like KNN and SVM. Hence, this work used MinMaxScaler for feature scaling. By standardizing data in this way, the model becomes more stable, more efficient during training, and more accurate in its predictions.

4.3 Train-Test Split

In this work, as presented in Figure 1, the dataset was divided into training and testing sets using an 80:20 ratio, a step that is crucial for ensuring the model's unbiased evaluation. By training the model on a majority portion of the data and then testing it on a separate, unseen subset, the model's ability to generalize to new, unfamiliar instances can be accurately assessed. To maintain the original distribution of classes within both the training and testing sets, stratified splitting was employed. This ensured that each subset contains a representative proportion of each class, which is particularly important for imbalanced datasets and helps prevent biased model performance. Further, the total samples considered, heart disease patients, training, and testing set samples are given in Table 2. In this dataset, 143 males and 190 females were there in the total dataset.

Table 2. Total, Training and Testing Samples

Case	Training (80%)	Testing (20%)
Normal	124	31
Heart Disease	142	36
Total	**266**	**67**
Total Samples	**333**	

4.4 ML Models

In this work, three supervised ML models, i.e., SVM, KNN, and RF, were employed to identify the presence or absence of heart disease. These models were chosen because they are highly effective for binary classification tasks and are capable of handling both categorical and numerical clinical data. The dataset was divided into an 80:20 ratio, with 80% used for training the models and 20% reserved for testing, ensuring that model evaluation is performed on unseen data.

4.4.1 SVM

The SVM is a powerful supervised learning algorithm widely applied in heart disease detection due to its robustness in handling high-dimensional and non-linear medical data. The primary objective of SVM is to construct a hyperplane that maximally separates the classes, i.e., in this work, patients with heart disease and those without, based on their clinical features such as age, cholesterol, resting blood pressure, and ECG outcomes. Mathematically, the SVM attempts to solve the following optimization problem using Eq. (1).

$$\min_{w,b} \frac{1}{2} ||w||^2 \text{ subject to } y_i(w \cdot x_i + b),\ i = 1,2, \dots, n \quad (1)$$

In Eq. (1), w is the weight vector perpendicular to the hyperplane, b is the bias term, x_i represents the feature vector of the i^{th} patient, and $y_i \in \{-1, +1\}$ denotes class label (diseased or non-diseased). By maximizing the margin between classes, SVM enhances the model's generalization ability and reduces misclassification. For non-linear data relationships, kernel functions such as linear, polynomial, or Radial Basis Function (RBF) are applied using Eq. (2).

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (2)$$

In Eq. (2), ϕ maps input features into a higher-dimensional space to allow linear separation. SVM has consistently demonstrated competitive performance in predicting CVD [10].

4.4.2 RF

RF is an ensemble learning algorithm that combines multiple DTs to perform classification or regression. It has several advantages, i.e., it can handle a large number of input variables, estimate feature importance, reduce overfitting, and internally provide unbiased generalization error estimates. The principle behind RF involves two levels of randomization during the training process. First, each DT is trained on a bootstrap sample drawn randomly from the training data. Second, at each node split, a random subset of features is selected to determine the best split. For a forest with T trees, the final prediction $\hat{y}$ is obtained by majority voting for classification using Eq. (3).

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (3)$$

In Eq. (3), $h_T(x)$ is the prediction of the t^{th} tree. By aggregating the predictions of multiple trees, RF reduces the variance of individual decision trees and enhances prediction accuracy. Its ability to handle high-dimensional data and mitigate overfitting makes it a highly reliable model in CVD risk prediction [19].

4.4.3 KNN

The KNN approach is a distance-based supervised learning method that is simple yet effective for predicting heart disease. In KNN, a patient's class is determined by the majority vote among the k closest neighbors in the training set. The similarity between patients is commonly measured using Euclidean distance as presented in Eq. (4).

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \quad (4)$$

In Eq. (4), M is the number of clinical features, and x_{im} and x_{jm} are the m^{th} features of patients i and j, respectively. In this work, for efficient ML modelling, preprocessing steps as discussed in the data preparation section were considered, as KNN is sensitive to the scale of input variables. The KNN approach, classifying a patient based on similarity to previously observed cases, makes it highly interpretable for medical applications. Previous studies have demonstrated its effectiveness, with Ekong [20] reporting that KNN outperformed both RF and SVM on a Kaggle CVD dataset.

5. PERFORMANCE EVALUATION

The trained models (SVM, RF, KNN) were evaluated on the test dataset using accuracy, precision, recall, and f-score performance metrics to assess their effectiveness in predicting heart disease. Accuracy represents the proportion of correct predictions made by the model out of the total predictions and provides a general measure of overall performance. However, in clinical applications, additional metrics such as precision, recall, and F1-score are crucial, as they offer a better evaluation of the model's predictive ability for each class. In this study, class 0 represents the absence of heart disease, while class 1 indicates its presence. These measures are particularly important in medical contexts because a false negative (FN), misclassifying an actual patient as healthy, can have serious consequences [21], [22], [23]. Accuracy represents the percentage of correct predictions, whether positive or negative, over total predictions made. It reveals the level of reliability of a model to classify data. The accuracy is evaluated using Eq. (5), where TP denotes true-positive, FP denotes false-positive, FN denotes false positive and TN denotes true-negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Precision is evaluated using Eq. (6), which is the ratio of the number of correct positive predictions among all those the model identified as positive, or the extent to which the positive predictions can be trusted.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall is evaluated using Eq. (7), which refers to the proportion of correct positive responses that the model has made, or what the model can identify.

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

The F1-score is evaluated using Eq. (8), which is a harmonic combination of two measures, precision and recall, that weigh both. It comes in especially when you are training models with skewed data. These indicators provide the complete picture of the categorical performance of any model [21], [24], [25], [26].

$$F - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \qquad (8)$$

Specificity measures the ability of a model to correctly identify negative cases, i.e., patients who do not have the condition. It is an important metric in medical applications because it quantifies how well the model avoids false positives, ensuring that healthy individuals are not misclassified as having the disease. Specificity is defined as the ratio of TNs to the sum of TNs and FPs, as given in Eq. (8).

$$Specificity = \frac{TN}{TN + FP} \qquad (10)$$

The final results obtained by the ML models that have undergone feature selection are further discussed in the following section. The trained final model was applied to estimate the presence (1) or absence (0) of heart disease in the newly gathered clinical cases of Sai Cardiac Hospital, Vijayapura. The prediction provide aid in supporting early intervention, which is a central objective in CVD mortality reduction [27], [28].

6. RESULTS AND DISCUSSION

In this study, supervised ML models, including SVM, KNN, and RF, were applied to a primary dataset collected from SCHV. Before model development, the dataset was thoroughly examined using various data visualization techniques, such as pie charts, heatmaps, and other graphical representations, as shown in Figures 2 to 5. The dataset consisted of 333 patient records, with each record containing vital clinical parameters such as age, sex, type of chest pain, blood pressure, and ECG findings. A preliminary EDA was conducted to better understand the underlying structure and relationships within the dataset. The correlation matrix (Figure 2) reveals that several features exhibit significant correlations with the presence of heart disease, which serves as the target variable. For instance, attributes such as type of chest pain, maximum heart rate, and exercise-induced angina display both positive and negative correlations with the target outcome. Figure 3 illustrates the distribution of individual feature values, aiding in the identification of imbalances, trends, and potential outliers in the data. Figure 4 presents the age-wise distribution of heart disease cases, highlighting the prevalence across different age groups. Finally, Figure 5 depicts the proportion of heart disease and non-heart disease cases, indicating that the dataset is relatively balanced between the two classes. This balance is critical for the development of reliable classification models, as it prevents bias toward any particular class and ensures that the trained models perform accurately across both categories.

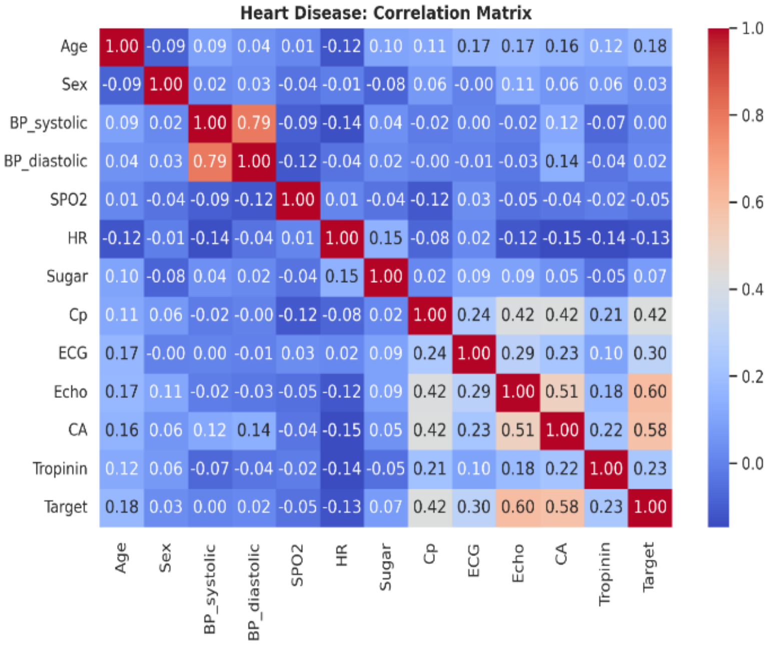

Fig 2: Correlation heat map of medical features

The correlation heatmap presented in Figure 2 illustrates the relationships between various clinical features and their predictive value for heart disease. Among the features, Echo (0.60), CA (CA, 0.58), and type of CP (CP, 0.42) emerge as the strongest predictors of heart disease. Other features, such as ECG (0.30), troponin levels (0.22), and age (0.18), show moderate influence on the target variable. Notably, systolic and diastolic blood pressure exhibit a strong correlation with each other (0.79), indicating redundancy; however, they have minimal direct association with heart disease. Meanwhile, features like sex, SpO_2, and blood sugar display very weak correlations, suggesting limited predictive utility. Overall, the correlation matrix underscores the importance of focusing on well-correlated features to enhance model performance, as prioritizing these variables can reduce redundancy, minimize noise, and improve the accuracy of predictive models.

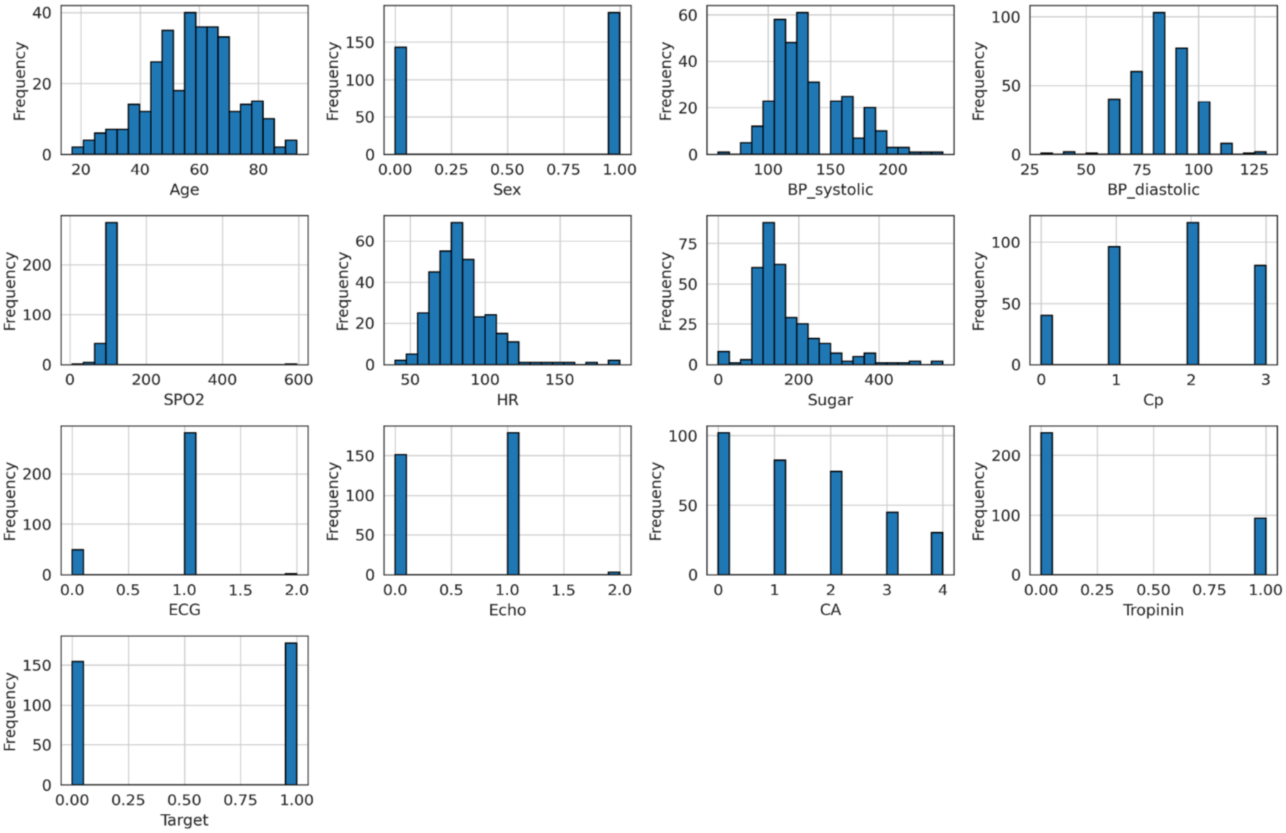

Fig 3: Data Visualization (number of counts) according to each feature in the data set

The feature distribution plots shown in Figure 3 illustrate the count and spread of each variable within the dataset. Continuous variables, including age, heart rate, and blood pressure, exhibit approximately normal distributions, suggesting a balanced spread of values across the patient population. In contrast, variables such as blood sugar and SpO₂ display skewed distributions and contain notable outliers, which may require attention during preprocessing. Categorical features, including type of chest pain, ECG results, echocardiography findings, and coronary angiography outcomes, show varying class distributions, with some categories being more prevalent than others. Features such as sex and troponin levels exhibit low variability, with values concentrated in specific categories. Importantly, the target variable, representing the presence (1) or absence (0) of heart disease, is relatively evenly distributed, which is ideal for binary classification tasks as it ensures that the model is trained on a balanced representation of both classes.

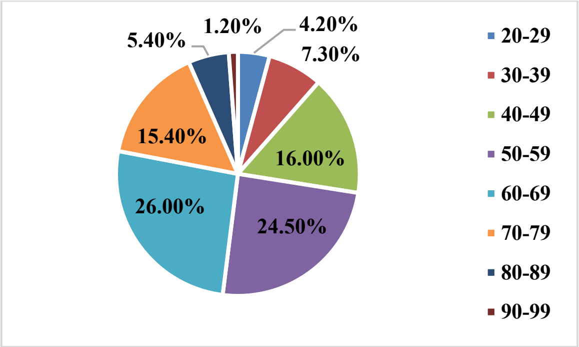

Fig 4: Heart disease distribution according to age

Figure 4 illustrates the age-wise distribution of patients in the dataset. The highest prevalence of heart disease is observed in the age groups 50–59 (24.5%), 60–69 (26%), and 70–79 (15.4%), indicating a marked increase in cardiovascular risk with advancing age. In contrast, younger age groups show a lower incidence, with less than 7% of patients aged 20–39 and approximately 16% in the 40–49 age range. These findings underscore that the risk of heart disease rises significantly after middle age, while also highlighting the occurrence of early-onset cardiovascular conditions in younger adults. This emphasizes the importance of preventive measures and early detection strategies, even in younger populations, to mitigate the long-term impact of heart disease.

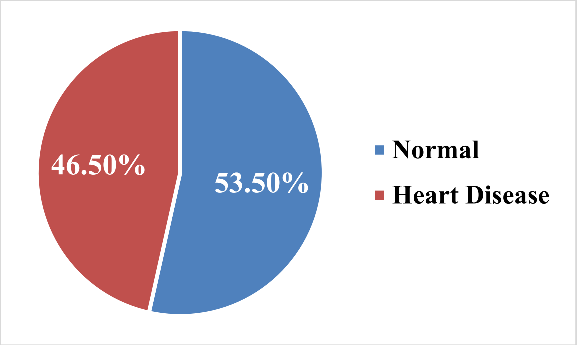

Fig 5: Percentage of heart disease and non-heart disease cases in the dataset

The distribution of the target variable in the dataset is depicted in Figure 5 and further illustrated by the accompanying pie chart. It shows that 53.5% of patients did not have heart disease (Class 0: NO), while 46.5% were diagnosed with the condition (Class 1: YES). This relatively balanced distribution ensures that the ML models are not biased toward a particular class during training, which is essential for developing fair and reliable classification models. Maintaining such a balance helps improve the generalization ability of the models and enhances their predictive accuracy across both classes.

The RF model demonstrated robust performance on the SHCV dataset, achieving an overall accuracy of 85%. It maintained a precision of 86%, indicating a high proportion of correctly identified TPs cases among all predicted positives. The RF recall was 89%, reflecting its ability to correctly detect most actual TP cases, while the F-score was 86%. Additionally, the specificity of 84% highlights RF's effectiveness in correctly identifying negative cases. Figure 6 illustrates the overall performance metrics of RF on the SHCV dataset. Figure 7 presents the AUC-ROC curve, showcasing RF discriminative capability between classes. The corresponding confusion matrix for RF, also shown in Figure 8, provides a detailed view of the TP, TN, FP, and FN predictions, offering further insight into RF prediction behavior.

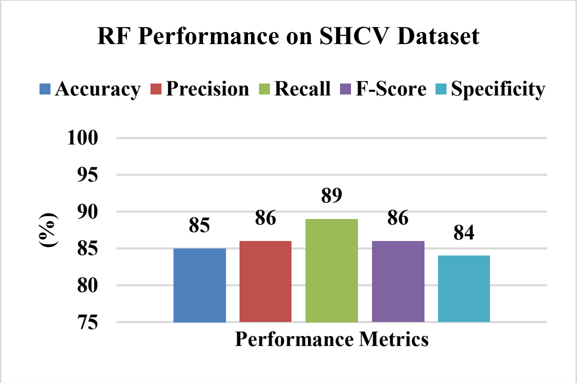

Fig 6: Performance of RF on SHCV dataset.

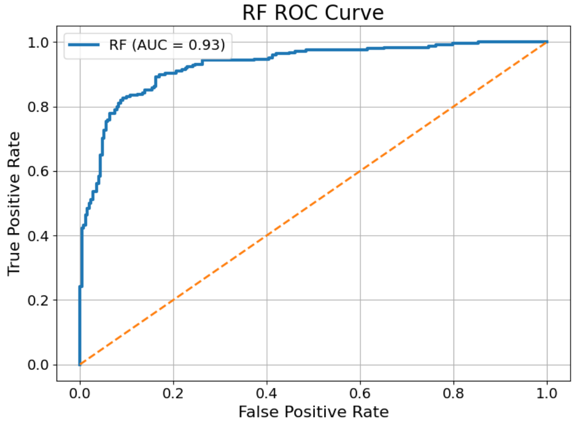

Fig 7: AUC-ROC for RF.

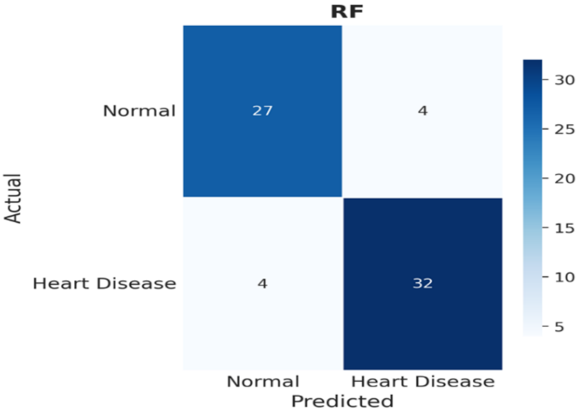

Fig 8: Confusion matrix for RF.

The KNN model demonstrated strong performance on the SHCV dataset, achieving an overall accuracy of 88%. The KNN attained a precision of 89%, indicating that the majority of the predicted TP cases were correctly identified. Its recall was 86%, reflecting its ability to detect a substantial portion of actual positive cases, while F-score was 89%. The specificity of 87% highlights KNN's effectiveness in correctly identifying negative cases. Figure 9 illustrates the overall performance metrics of the KNN model on the SHCV dataset. Figure 10 presents the AUC-ROC curve, showing the model's ability to distinguish between classes. The corresponding confusion matrix, shown in Figure 11, provides a detailed representation of the TP, TN, FP, and FN predictions, offering further insight into the classification performance of KNN.

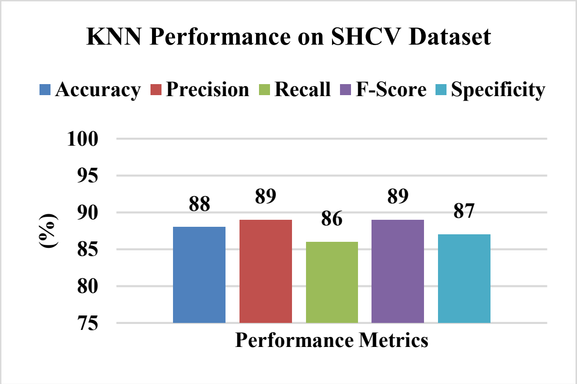

Fig 9: Performance of KNN on SHCV dataset.

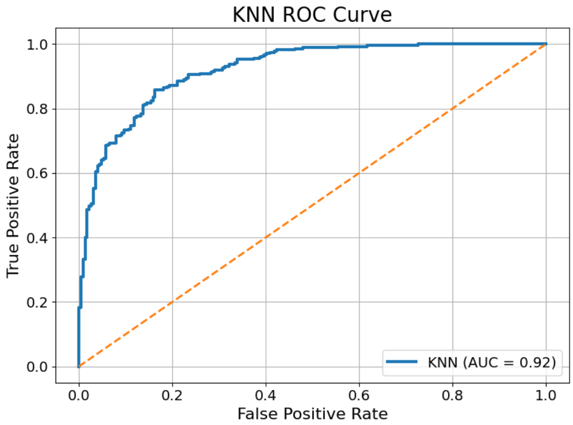

Fig 10: AUC-ROC for KNN.

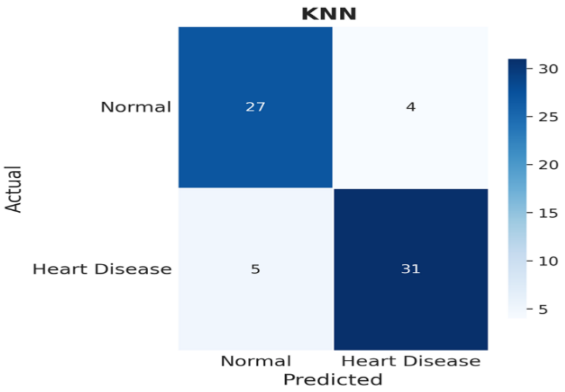

Fig 11: Confusion matrix for KNN.

The SVM model exhibited excellent performance on the SHCV dataset, achieving an overall accuracy of 90%. It attained a precision of 91%, indicating a high proportion of correctly predicted TP cases among all positive predictions. The SVM recall was 89%, reflecting its strong ability to identify actual positive cases, while the F-score, which balances precision and recall, was 90%. Additionally, the specificity of 90% demonstrates SVM's effectiveness in correctly identifying negative cases. Figure 12 illustrates the comprehensive performance metrics of the SVM model on the SHCV dataset. Figure 13 presents the AUC-ROC curve, highlighting the SVM's capability to distinguish between classes. The corresponding confusion matrix, shown in Figure 14, provides a detailed overview of TP, TN, FP, and FN predictions, offering further insight into the classification performance of SVM.

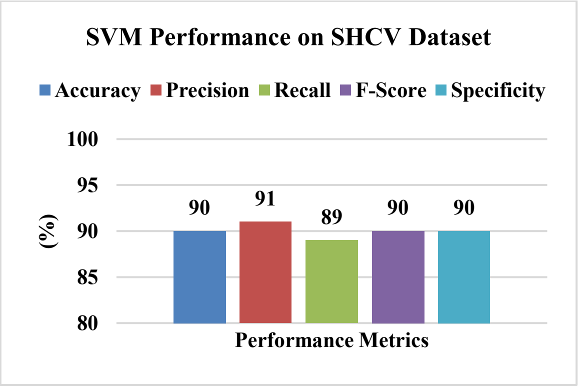

Fig 12: Performance of SVM on SHCV dataset.

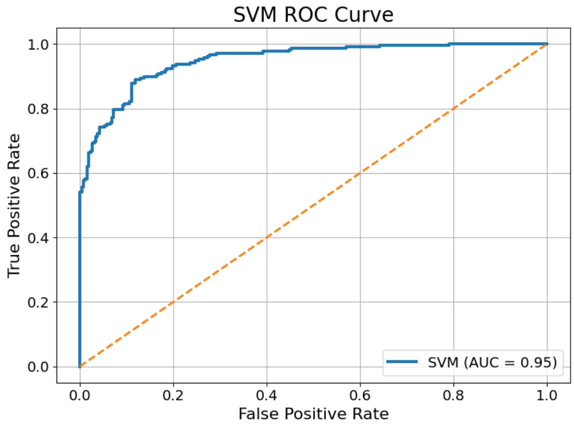

Fig 13: AUC-ROC for SVM.

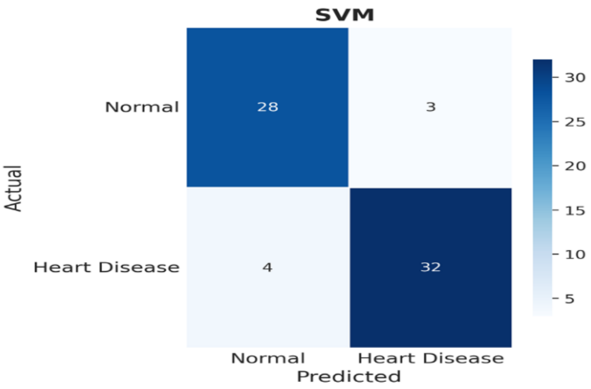

Fig 14: Confusion matrix for SVM.

Table 3 presents a comparison of RF, KNN, and SVM models based on MCC, Misclassification Rate (MCR), and computational time (in seconds). The SVM model achieved the highest MCC of 0.79, indicating the strongest overall correlation between predicted and actual class labels, while RF and KNN obtained MCC values of 0.70 and 0.76, respectively. In terms of misclassification, SVM also performed best, with the lowest MCR of 0.10, followed by KNN at 0.12 and RF at 0.15, highlighting SVM's superior accuracy in minimizing prediction errors. Regarding computational efficiency, KNN was the fastest model, requiring only 0.01 seconds, whereas SVM and RF took slightly longer, 0.02 and 0.20 seconds, respectively. Overall, Table 3 indicates that while SVM provides the most reliable classification performance, KNN offers the advantage of faster execution time, and RF demonstrates moderate performance across both accuracy and speed metrics.

Table 3. Comparative Performance of RF, KNN, and SVM Based on MCC, MCR, and Computational Time

Model	MCC	MCR	Time (s)
RF	0.70	0.15	0.20
KNN	0.76	0.12	0.01
SVM	**0.79**	**0.10**	**0.02**

Table 4 presents a comparative analysis of the proposed models against previously established methodologies. The earlier studies, using the UCI dataset, reported accuracies of 83.60% and 87.00%, with CNN and KNN identified as the best-performing classifiers, respectively. In comparison, the proposed approach, evaluated on the SHCV dataset, achieved a superior accuracy of 90.00%, with the SVM model emerging as the most effective classifier. This improvement highlights the enhanced predictive capability of the proposed models, particularly SVM, in accurately classifying the dataset. The comparison demonstrates that the suggested methodology not only surpasses the performance of conventional models but also confirms the robustness and reliability of the approach in handling real-world SHCV data.

Table 4. Comparative analysis of a suggested model against established methodologies

Ref	Models	Dataset	Accuracy	Best Classifier
[17]	KNN, SVM, ANN, CNN	UCI	83.60	CNN
[16]	LR, KNN, SVM, ANN	UCI	87.00	KNN
Proposed	**RF, KNN, SVM**	**SCHV**	**90.00**	**SVM**

7. CONCLUSION AND FUTURE SCOPE

This study focused on developing and validating ML models to predict heart disorders using a dataset of 333 patient instances from SCHV. Through thorough preprocessing and EDA, several meaningful patterns and relationships between features were identified, which contributed to enhancing model performance. Among the models evaluated, SVM demonstrated superior performance, achieving a high accuracy of 90%, an AUC value of 0.95, and a low misclassification rate. The robustness of the SVM model was further validated using confusion matrices and ROC curves. Comparative analysis with existing literature indicated that the proposed approach outperformed prior methodologies, emphasizing the effectiveness of SVM in balancing accuracy, recall, and computational efficiency for heart disease classification. For future work, the work will be continued by exploring additional datasets and testing alternative ML models to further improve prediction performance. Expansion of the current dataset will also be undertaken to enhance model generalizability. Moreover, future research could investigate the integration of AutoML systems, ensemble learning techniques, and the development of in situ diagnostic decision support systems, aiming to improve predictive accuracy and provide real-time clinical support. These efforts are expected to strengthen the applicability and reliability of machine learning models in cardiovascular disease diagnosis.

8. REFERENCES

[1] M. Di Cesare *et al.*, “The Heart of the World,” *Global Heart*, vol. 19, no. 1, Jan. 2024, doi: 10.5334/gh.1288.

[2] C. Antza *et al.*, “Prevention of cardiovascular disease in young adults: Focus on gender differences. A collaborative review from the EAS Young Fellows,” *Atherosclerosis*, vol. 384, p. 117272, Sep. 2023, doi: 10.1016/j.atherosclerosis.2023.117272.

[3] G. K. Ghodeshwar, A. Dube, and D. Khobragade, “Impact of lifestyle modifications on cardiovascular health: A narrative review,” *Cureus*, vol. 15, no. 7, pp. 1–8, Jul. 2023, doi: 10.7759/cureus.42616.

[4] E. Młynarska *et al.*, “From atherosclerotic plaque to myocardial infarction—the leading cause of coronary artery occlusion,” *International journal of molecular sciences*, vol. 25, no. 13, pp. 7295–7295, Jul. 2024, doi: 10.3390/ijms25137295.

[5] M. F. Costantino, F. Cortese, Gianpaolo D’Addeo, A. Nicoletti, S. Mancino, and L. Stolfi, “Diagnostic modalities for ischemic heart disease: evaluating the role of stress echocardiography, cardiac CT, and myocardial perfusion scintigraphy in guiding coronary angiography,” *Exploration of Cardiology*, Jan. 2025, doi: 10.37349/ec.2025.101243.

[6] H. Kamal *et al.*, "Heart Disease Prediction Using Machine Learning," *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, Wardha, India, 2024, pp. 1-6, doi: 10.1109/IDICAIEI61867.2024.10842908.

[7] A. Gnanavelu, C. Venkataramu, and R.Chintakunta, “Cardiovascular Disease Prediction Using Machine Learning Metrics,” *Journal of Young Pharmacists*, vol. 17, no. 1, pp. 226–233, Jan. 2025, doi: 10.5530/jyp.20251231.

[8] S. Saha, M. M. Rahman, T. T. Suki, M. M. Alam, M. S. Alam and M. A. S. Haque, "Heart Disease Prediction Using Machine Learning Algorithms: Performance Analysis," *2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, Gazipur, Bangladesh, 2024, pp. 1-6, doi: 10.1109/ICAEEE62219.2024.10561820.

[9] N. Chandrasekhar and Samineni Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” *Processes*, vol. 11, no. 4, pp. 1210–1210, Apr. 2023, doi: 10.3390/pr11041210.

[10] N. Biswas *et al.*, “Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques,” *BioMed Research International*, pp. 1–15, May 2023, doi: 10.1155/2023/6864343.

[11] A. P. Jawalkar *et al.*, “Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting,” *Journal of Engineering and Applied Science*, vol. 70, no. 1, Oct. 2023, doi: 10.1186/s44147-023-00280-y.

[12] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[13] A. A. Ahdal, M. Rakhra, S. Badotra and T. Fadhaeel, "An integrated Machine Learning Techniques for Accurate Heart Disease Prediction," *2022 International Mobile and Embedded Technology Conference (MECON)*, Noida, India, 2022, pp. 594-598, doi: 10.1109/MECON53876.2022.9752342.

[14] Y. Rimal, Siddhartha Paudel, N. Sharma, and Abeer Alsadoon, “Machine learning model matters its accuracy: a comparative study of ensemble learning and AutoML using heart disease prediction,” *Multimedia Tools and*

Applications, vol. 83, no. 12, pp. 35025–35042, Sep. 2023, doi: 10.1007/s11042-023-16380-z.

[15] A. A. Stonier, R. K. Gorantla, and K Manoj, "Cardiac disease risk prediction using machine learning algorithms," *Healthcare technology letters*, Nov. 2023, doi: 10.1049/htl2.12053.

[16] Osei-Nkwantabisa, Akua Sekyiwaa and R. Ntumy, "Classification and Prediction of Heart Diseases using Machine Learning Algorithms," *arXiv.org*, 2024, doi: 10.48550/arXiv.2409.03697.

[17] P. Ram *et al.*, "Investigations on cardiovascular diseases and predicting using machine learning algorithms," *Cogent Engineering*, vol. 11, no. 1, Aug. 2024, doi: 10.1080/23311916.2024.2386381.

[18] S. R. Kawale and Pooja Kallappagol, "AI-Driven Neural Networks for Early-Stage Diabetes Prediction," *Journal of Computational Analysis and Applications (JoCAAA)*, vol. 33, no. 07, pp. 740–751, Aug. 2024, Accessed: Dec. 08, 2025. [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/1135

[19] V. E. Ekong, "Evaluation of machine learning techniques towards early detection of cardiovascular diseases", *American Journal of Artificial Intelligence*, vol. 7, no. 1, pp. 7–14, 2023.

[20] P. Pachiyannan, M. Alsulami, D. Alsadie, A. K. J. Saudagar, M. AlKhathami, and R. C. Poonia, "A Novel Machine Learning-Based Prediction Method for Early Detection and Diagnosis of Congenital Heart Disease Using ECG Signal Processing," *Technologies*, vol. 12, no. 1, p. 4, Jan. 2024, doi: 10.3390/technologies12010004.

[21] H. A. Al-Alshaikh *et al.*, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports*, vol. 14, no. 1, p. 7819, Apr. 2024, doi: 10.1038/s41598-024-58489-7.

[22] A. Mir *et al.*, "A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques," *ESC heart failure*, Jul. 2024, doi: 10.1002/ehf2.14942.

[23] A. Hussain and A. Aslam, "Cardiovascular Disease Prediction Using Risk Factors: A Comparative Performance Analysis of Machine Learning Models," *Journal on Artificial Intelligence*, vol. 6, no. 1, pp. 129–152, 2024, doi: 10.32604/jai.2024.050277.

[24] S. Mariettou, C. Koutsojannis, and V. Triantafillou, "Predicting Coronary Heart Disease Through Machine Learning Algorithms," *Lecture notes in networks and systems*, pp. 652–659, Jan. 2024, doi: 10.1007/978-3-031-65522-7_56.

[25] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," *Health & Social Care in the Community*, vol. 2023, p. e1406060, Feb. 2023, doi: 10.1155/2023/1406060.

[26] N. Narayanan and Jayashree, "Implementation of Efficient Machine Learning Techniques for Prediction of Cardiac Disease using SMOTE," *Procedia Computer Science*, vol. 233, pp. 558–569, Jan. 2024, doi: 10.1016/j.procs.2024.03.245.

[27] R. Hoque, M. Billah, A. Debnath, S. Hossain, and N. Bin, "Heart Disease Prediction using SVM," *International Journal of Science and Research Archive*, vol. 11, no. 2, pp. 412–420, Mar. 2024, doi: 10.30574/ijsra.2024.11.2.0435.

[28] D. M. K. Selvi, J. Aswini, C. Balakrishnan, K. Suganya, B. G. Sheena and S. S. R, "Revolutionizing Cardiovascular Care: The Role of AI, ML, and DL in Early Heart Disease Prediction and Treatment," *2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)*, Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/ICECCC61767.2024.10593941.